The Development of Comparative Information Yield Curves for Application to Subsurface Characterization

by

Felipe Pereira Jorge de Barros

M.S. (Universidade Federal do Rio de Janeiro, Brazil) 2004
Engen (Universidade Federal do Rio de Janeiro, Brazil) 2003

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Engineering-Civil and Environmental Engineering

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Yoram Rubin, Chair
Professor Fotini Katopodes Chow
Professor Fraydoun Rezakhanlou
Professor Reed Maxwell

Spring 2009

The dissertation of Felipe Pereira Jorge de Barros is approved:

_____

Chair Date

_____

Date

_____

Date

_____

Date

University of California at Berkeley

Spring 2009

The Development of Comparative Information Yield Curves for Application to Subsurface Characterization

by

Felipe Pereira Jorge de Barros

Abstract

The Development of Comparative Information Yield Curves for Application to Subsurface Characterization

by

Felipe Pereira Jorge de Barros

Doctor of Philosophy in Engineering-Civil and Environmental Engineering

University of California at Berkeley

Professor Yoram Rubin, Chair

Defining rational and effective hydrogeological data acquisition strategies is of crucial importance in subsurface contamination as such efforts are always resource limited. Usually strategies are developed with the goal of reducing uncertainty, but less often they are developed in the context of their impacts on uncertainty. This work presents an approach for determining subsurface site characterization needs based on human health risk. The main challenge is in striking a balance between reduction in uncertainty in hydrogeological, behavioral and physiological parameters. Striking this balance can provide clear guidance on setting priorities for data acquisition and for better estimating adverse health effects in humans. This challenge is addressed through theoretical developments and numerical simulation. A wide range of factors that affect site characterization needs are investigated, including the dimensions of the contaminant plume and additional length scales that characterize the transport problem, as well as the model of human health risk. With the

proposed approach, conditions are investigated where reduction of uncertainties from flow physics, human physiology and exposure related parameters might contribute to a better understanding of human health risk assessment. The concept of comparative information yield curves is used for investigating the relative impact of hydrogeological and health-related parameters in risk. Results show that characterization needs are dependent on the ratios between flow and transport scales within a risk-driven context. Additionally these results indicate that human health risk becomes less sensitive to hydrogeological measurements for large plumes. This indicates that under near-ergodic conditions, uncertainty reduction in human health risk may benefit from better understanding of the physiological component as opposed to a more detailed hydrogeological characterization. Other results show that the worth of hydrogeological characterization in human health risk depends on the interplay between the characteristic time the contaminant plume takes to cross an environmentally sensitive target and on the exposure duration of the population. Finally, the role of geostatistical model uncertainty in defining sampling networks and its relevance in human health risk is also addressed.

Professor Yoram Rubin
Dissertation Committee Chair

In memory of my grandmother, Esther.

To my family,

And to the most important Chapter of my PhD.

Contents

List of Figures

List of Tables

Acknowledgements

I arrived in Berkeley at the age of 23 and now, at the age of 28, I find it hard to believe that I am writing these words. So much time has passed and so much have I learned. Berkeley truly has been an unique and rewarding experience in my life. To begin, I would like to acknowledge my advisor Professor Yoram Rubin. It has been pleasure to work with him. His dedication, intelligence and guidance was crucial for the development of this thesis. I will always have fond memories of his teachings, our conversations and his jokes. His high academic standards helped me construct my self-criticism and become a better researcher. He taught me persistence and how to better express my work. I learned from him how to extract important results using simple, yet elegant, approaches. Yoram was more than an advisor: He became a close friend of mine that followed all my daily Brazilian life drama in Berkeley! I will never forget the day he picked me up from the airport - I knew we would get along. Thank you Yoram for guidance, dedication, inspiration and friendship.

Secondly, I would like to thank Professor Reed Maxwell at the Colorado School of Mines for his support and friendship. Reed not only provided thoughtful comments on my dissertation work but also helped me in many aspects. He suggested a large body of literature and offered his numerical codes that aided in the construction of this thesis. His work in the late nineties served as an inspiration for this dissertation. My acknowledgments also goes to Professor Wolfgang Nowak, in Germany. Wolfgang came to Berkeley in 2007 as a one-year post-doctoral fellow and we became good friends. Wolfgang opened my mind to the world of optimal design and I learned a great deal from our discussions. I will remember the good times we had over dinner and drinks together with his wife Erika.

I would also like to thank other faculty members: Professor Mark Stacey, Professor Fotini

Katopodes Chow, Professor John Dracup and Professor Fraydoun Rezakhanlou. They were part of my preliminary exam, qualifying exam and dissertation committees. I greatly appreciate their input. Also, I would like to thank Souheil Ezzedine for helpful discussions on my work. Of course, throughout these years I had to opportunity to meet great people. I would like to thank my friends that are (and were) part of the research group: Gretchen Miller, Xingyuan Chen, Zhangshuan Hou, Haruko Murakami, Hang Bai, Pirjo Isosaari, Jenny Druhan, Newsha Ajami, Chuanhui Gu and finally, Federico Maggi. I would like to especially thank Federico for the hilarious times sitting near the fountain in front of Café Strada enjoying the occasional visit of the sun, cooking lessons in College Avenue (*lesson 14* is unforgettable) and for making Davis Hall a louder place. My very special thanks goes to the wonderful Brazilian community here in UC Berkeley: Ram Rajagopal, Gregorio and Maria Carolina Caetano, Andres and Priscila Donangelo, Mauricio Mancio, Henrique Barbosa and Cecilia Dantas. These people brought the very characteristic Brazilian optimism in Berkeley by adding a bit of *bossa* and *samba* rhythm to my PhD. I am thankful for all the friendships I acquired during these years. This acknowledgment would not be complete without mentioning two fantastic Brazilians that have known me long before I became a graduate student: Dr. Elizabeth May from the *Comissão Nacional de Energia Nuclear* and Professor Renato Cotta from the *Universidade Federal do Rio de Janeiro*, Brazil. I thank them for the support and encouragement throughout these years. Also, many thanks to my long time friend, Rafael Laufer, who did not let me forget the *carioca* lifestyle.

I am especially grateful for Simonetta Rubol. Her presence, love and support made my life much more colorful. Not only did she add an Italian dimension to my life but she also has the power to bring a smile to my face every day. I am extremely thankful to my family back in

Rio de Janeiro, Brazil. Their daily telephone calls gave me strength, encouragement, love, serenity and peace. They have kept me balanced throughout these years. I thank my sister, Mariana, for friendship, continuous encouragement and for keeping the telephone calls filled with fireworks, laughter and tears of *saudade*. My gratitude to my beautiful mother Icléa who injected endless faith and optimism in my daily routine. The sound of her joyful voice always brought me back to Rio de Janeiro despite that fact that I was kilometers away. Finally, my deepest gratitude to my dear father Francisco Claudio, who has always been my source of inspiration and balance. It was his motivation that brought me to Berkeley. A true spiritual, academic and human role model. A man who taught me how to find happiness by keeping things simple.

Finally, I would like to thank the CAPES fellowship program from the Brazilian government for the financial support.

Chapter 1

Introduction

1.1 Motivation

Groundwater is one of the major sources of drinking water and it is widely used in the agricultural and the industrial sector. Understanding how contaminants are transported in the subsurface and evaluating the risks they pose to humans are important environmental issues. For example, an accidental oil spill or the occurrence of leaking hazardous waste storage may severely affect groundwater quality. Thus characterization of the subsurface is vital in order to predict the magnitude of contaminant concentrations and associated human health risks.

Quantifying flow processes in natural porous formations is a challenge given the underlying heterogeneity in the subsurface. Soil properties, such as the hydraulic conductivity and porosity, exhibit a high degree of spatial variability at all length scales [*Gelhar*, 1993; *Dagan*, 1989; *Dagan and Neuman*, 1997; *Rubin*, 2003]. Ignoring the existence of spatial heterogeneity in groundwater quality modeling can result erroneous decision making. Often, such decisions are motivated by cleaning up a contaminated site to reach compliance criteria established by a drinking water stan-

dard. In most cases, such compliance criteria are often associated with an acceptable risk determined by an environmental regulation agency [*USEPA*, 1989].

Given the aforementioned reasons, it is of importance to characterize the subsurface adequately in order to capture the spatial patterns of heterogeneity. Given that financial resources are limited, hydrogeologists face the challenge of data scarcity. The combination of subsurface spatial heterogeneity and the scarcity of data leads to uncertainty in model predictions (for example, flow and contaminant transport models) and consequently, uncertainty in evaluating human health risk. Hence, incorporating hydrogeological data helps reduce the involved uncertainties. Tighter confidence bounds of concentration estimates at an environmentally sensitive target will allow for a more reliable prediction of human health risk [*Rubin et al.*, 1994; *Andricevic and Cvetkovic*, 1996; *Maxwell et al.*, 1999].

Besides the uncertainty stemming from insufficient hydrogeological data, the incomplete knowledge of how humans metabolize certain contaminants and how they are exposed to such chemicals also contributes to the total uncertainty in the estimates in human health risk. Thus, human health risk assessment consists of two main uncertain components: Hydrogeological and physiological. Therefore, it is natural to formulate human health risk in terms of a probabilistic framework that accounts for the uncertainty in both health-related and hydrogeological components [*Andricevic and Cvetkovic*, 1996; *Maxwell et al.*, 1999]. In fact, modern environmental regulations recommend application of such probabilistic tools to evaluate human health risk [*USEPA*, 2001].

The aforementioned sources of uncertainty lead to the following fundamental questions associated with subsurface site characterization:

1. Given the multi-component characteristics of health risk assessment and their corresponding

uncertainties, how much effort should one invest in data acquisition?

2. Under what conditions does the uncertainty in human physiological response overwhelms the uncertainty arising from subsurface characterization?

3. Are we able to identify conditions where a detailed geological site characterization is justified?

To answer these questions, one must be able to define rational and effective hydrogeological data acquisition strategies guidelines. Such an approach is of crucial importance in subsurface contamination as such sampling efforts are always resource limited. In most cases, strategies are developed with the goal of reducing uncertainty, but less often they are developed in the context of their impacts on uncertainty. This work presents an approach for determining task-oriented subsurface site characterization needs. More precisely, characterization needs will be addressed within a human health risk context.

1.2 Scope of the Dissertation

In the current dissertation, a stochastic human health risk framework that accounts for the uncertainties arising from physical (flow and transport) and health-related (physiological) components is developed. The intention is to have a rational approach that illustrates conditions in which characterization efforts, when counter-balanced with the uncertainty present in the health risk-related parameters, can be reduced. This is achieved by considering the scales of the contaminant plumes, geostatistical correlation lengths, travel distances, scales of capture-zones induced by the action of pumping wells and finally, pore-scale dispersion. Chapters 2-4 contain an individual

introduction, literature review, test-case, discussions and summary. In the following paragraphs, an overview of each chapter is given.

Chapter 2 introduces a general probabilistic framework that will serve as the underlying foundation for the subsequent chapters. A human health risk cumulative distribution function (CDF) is analytically developed to account for both uncertainty and variability in hydrogeological as well as human physiological parameters. Flow and transport are quantified using analyical solutions based on Lagrangian formulations. Results in this chapter indicate how the human health risk cumulative distribution function becomes less sensitive to uncertainty in physiological parameters at lower risk values associated with longer travel times. An information entropy-based graphical tool is introduced that allows investigating the relative impact of hydrogeological and physiological parameters in human health risk. A metric α that relates hydrogeological uncertainty to physiological uncertainty is developed. Other results in Chapter 2 show that the worth of hydrogeological characterization in human health risk depends on the time the contaminant plume takes to cross the control plane and on the exposure duration of the population to certain chemicals.

In Chapter 3, issues concerning the significance of physical scales in defining characterization needs are investigated. Such physical scales are: (i) subsurface heterogeneity correlation scales, (ii) pore-scale dispersion, (iii) contaminant source dimensions, (iv) capture-zones and finally (v) concentration sampling scales. Also, the role of alternative risk models in defining characterization needs is addressed. Just as in Chapter 2, the concept of information entropy is invoked to evaluate uncertainty trade-offs between hydrogeological and health-related parameters. The development of *comparative information yield curves* is introduced to investigate the relative impact of uncertainty in human health risk. The framework introduced in Chapter 2 is also extended in

Chapter 3 to account for geostatistical model uncertainty. A procedure to estimate the worth of data prior to sampling in the information yield curves is also developed in order to aid decision makers in setting priorities towards data acquisition. In this chapter, the governing equations for flow and transport are solved numerically. Concentration statistics conditional on hydrogeological data at an environmentally sensitive location are obtained through Monte Carlo simulation.

The results in Chapters 2 and 3 indicate that risk uncertainty reduction benefits more from hydrogeological sampling under certain conditions (for example, non-ergodic transport). To evaluate concentrations with less uncertainty, one must be able to capture the geostatistical description of the subsurface (i.e., the mean, trends, covariance models and their parameters). In general, geostatistical models that describe the spatial correlation patterns in the subsurface are considered to be known and given *a priori* [*Dagan and Neuman*, 1997]. This contradicts the fact that only few or even no data at all offer support for such assumptions prior to exploration effort. The objective of Chapter 4 is to relax the assumption of considering parametric uncertainty only within a single covariance model. For this reason, uncertainty within the geostatistical model is also considered and its impact in subsurface characterization is shown. In Chapter 4, the Matérn family of covariance functions is used to describe the geostatistical model. The Matérn family of covariance functions has an additional parameter that controls shape of the model. Controlling model shape by a parameter converts covariance model selection to parameter identification and resembles Bayesian Model Averaging [*Hoeting et al.*, 1999; *Neuman*, 2003] over a *continuous* spectrum of covariance models. A series of synthetic test cases are simulated in this chapter to show the importance of geostatistical model uncertainty in the sampling design. The prediction variance of contaminant concentration or arrival time at an environmentally sensitive location is minimized by optimal placement of hy-

draulic head and hydraulic conductivity measurements. A variation of the information yield curves, presented in Chapter 3, are again used to illustrate the data worth. It is shown how the uncertainty arising from the lack of full knowledge in the subsurface formation affects the sampling patterns. In addition, Chapter 4 shows how it is important to consider task-oriented minimization to define characterization needs. The relevance and applicability of the results in human health risk is also addressed.

Finally, a summary of the results and main findings are given in Chapter 5. Appendices describing in detail derivations and the numerical tools used here are included. Although every mathematical symbol used is carefully explained throughout the thesis, a notation list is added in the end of Chapters 2-4.

Chapter 2

A Risk-Driven Approach for Subsurface Site Characterization

2.1 Introduction

There are two main contributors to human health risk assessment due to groundwater contamination: the first is contaminant transport and the second is human physiology and exposure parameters [*Maxwell et al.*, 1999]. Understanding the interactions between the uncertainty and variability present in each of these elements is a challenge when managing the remediation of contaminated sites due to the high costs associated with site remediation [*James and Gorelick*, 1994].

Flow and transport in the subsurface are complicated processes to model since natural geologic media are both heterogeneous and uncertain. Uncertainty present in hydrogeology arises

[1]This chapter is based on a published article in Water Resources Research, 2008. (Vol.44, N.1, W01414, doi:10.1029/2007WR006081)

from scarcity of data, measurement errors and spatial variability of flow properties such as the hydraulic conductivity $\mathbf{K}(\mathbf{x})$ and the porosity $\phi(\mathbf{x})$ [*Dagan*, 1984, 1987; *Rubin*, 2003]. The heterogeneity patterns are sometimes difficult to capture and require large quantity of data to properly map the aquifer's flow properties. These uncertainties and variabilities lead to uncertainty in predicting the spatial distribution of contaminants in groundwater propagating uncertainty in assessing concentration of contaminants in drinking water supplies.

The uncertainty and variability present in the human physiology and exposure parameters are also important factors to consider. Individuals consuming contaminated subsurface water also add uncertainty in risk assessment since human physiology and toxicology are not fully understood and each individual may have different exposure and response to a certain chemical [*McKone and Bogen*, 1991; *Maxwell and Kastenberg*, 1999].

Previous work on human health risk assessment tied risk assessment to the contaminant transport problem [*Andricevic and Cvetkovic*, 1996; *Maxwell and Kastenberg*, 1999; *Maxwell et al.*, 1999]. In these references, general methodologies were proposed to relate hydrogeological characterization with risk assessment and to identify the important parameters affecting human health risk evaluation. Researchers analyzed the effect of hydrological data acquisition in human health risk error reduction for a wide range of hydrogeological conditions [*Maxwell and Kastenberg*, 1999; *Maxwell et al.*, 1999].

Additional efforts are needed to understand and identify the conditions under which each of the risk contributors can lead to a significant risk reduction through data acquisition. This chapter presents a theoretical framework to investigate the benefits of better sampling of hydraulic conductivities and exposure parameters on risk assessment. The relationship between human exposure and

hydrogeological characterization is also investigated. The statistical moments of risk are explicitly and analytically derived accounting for parametric uncertainty in physiology and flow.

The probabilistic formulation present in this chapter provides tools to answer the following questions: What variables are significant in human risk assessment? How does information translate into risk assessment? When do human behavioral variability and uncertainty in physiology become important compared to hydrogeological uncertainty? What impact do environmental regulations have on hydrogeological characterization? When is a detailed site characterization justified?

2.2 Problem Formulation

We consider a groundwater contamination problem that can potentially cause adverse health effects to a certain population. To quantify these adverse health effects, the concept of risk is defined. In this paper, risk, denoted by r, is the increased individual probability of developing cancer during a lifetime due to exposure to a certain contaminant (or sometimes denoted as increased cancer risk).

In a deterministic situation, r can be defined without uncertainty. Let us define $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ respectively as vector of parameters needed for hydrogeological characterization and for physiological characteristics, respectively. The vector $\boldsymbol{\theta}_H$ is used to solve the flow problem, while $\boldsymbol{\theta}_P$ is used to analyze the outcome of the flow and transport problem on humans, in terms of adverse health effects.

A deterministic $\boldsymbol{\theta}_H$ can include a detailed spatial distribution of the hydrogeological parameters, such as the hydraulic conductivity, $\boldsymbol{\theta}_H$=$\{K_1, K_2, K_3,...,K_N\}$ where $K_i = K(\mathbf{x}_i)$ and N

is the total number of hydraulic conductivities needed.

However, in general, we do not have the vector $\{K_1, K_2, K_3,..., K_N\}$ and as a result we model the hydraulic conductivity and/or other parameters as *Space Random Functions*, SRF [*Dagan*, 1984, 1987; *Rubin*, 2003]. In this case, $\boldsymbol{\theta}_H$ will include parameters of the SRF.

As for the physiological and exposure parameters, we have $\boldsymbol{\theta}_P = \{\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3,...,\mathbf{P}_I\}$ where $\mathbf{P}_i$ represents a vector of physiological and behavioral parameters for the i^{th} individual within an environmentally impacted population. Here, I is the total number of individuals present in the target population. It is clear that it is impossible to have a deterministic knowledge of all physiological and behavioral parameters for each individual within I. For these reasons, we resort to a statistical characterization of these parameters.

The increased cancer risk cumulative distribution function is denoted by $F_R\,(r|\,\boldsymbol{\theta}_H, \boldsymbol{\theta}_P)$, where the subscript R denotes the random variable for risk. Here the vectors $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ contains parameters of the statistical models for the uncertain parameters. Using a cumulative distribution function (CDF) to describe human health risk allows one to determine the probability of risk being below or above a certain regulatory standard denoted by r^* [*USEPA*, 1989]. The increased cancer risk probability density function (PDF) is denoted by $f_R(r)$.

In many cases, measurements, denoted here by the set $\{m\}$, may be available. These measurements could be hydrogeological or physiological. Our first goal is to derive an expression for the risk CDF conditioned on a set of measurements: $F_R^c(r|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P) = F_R(r|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\})$. Here, the superscript c indicates that the CDF is conditioned on measurements. The conditioned CDF for risk given vectors of hydrogeological parameters, $\boldsymbol{\theta}_H$, and physiological parameters, $\boldsymbol{\theta}_P$, can be written mathematically as follows:

$$F_R^c(r^*|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P) = \text{Prob}[R \leq r^*|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P] \ , \tag{2.1}$$

The major challenge that drives this study is in deriving an explicit expression for equation (2.1) that allows us to identify, for a range of risk values r, the effect of increased sampling of hydraulic conductivities and physiological and exposure parameters on $F_R^c(r|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P)$, while translating correctly the uncertainty in hydrogeology, exposure and physiology into uncertainty of r. With a solution for equation (2.1), we can address questions such as: How to allocate resources between $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ estimation for optimal reduction in uncertainty? How does this allocation depend on travel time and exposure duration?

2.3 Solution Methodology

In this section, a human health risk model is presented and a closed-form expression for the risk CDF given in equation (2.1) is derived.

2.3.1 Human Health Risk Formulation

Risk can be defined by an exponential model for high carcinogenic risk levels, see *USEPA* [1989]. This model is used when a target population is exposed to high doses of a certain chemical. Increased cancer risk due to groundwater contamination, denoted by r, is given as:

$$r = 1 - \text{Exp}\left[-ADD_M \times CPF_M\right] \ , \tag{2.2}$$

where CPF_M [kg-d/mg] is the metabolized cancer potency factor and ADD_M [mg/(kg-day)] is the average daily dose metabolized. It is given as:

$$ADD_M = f_{mo} \times ADD_G + f_{mr} \times ADD_H + f_{mr} \times ADD_D, \tag{2.3}$$

where ADD_G, ADD_H and ADD_D are the average daily exposure from ingestion, inhalation and dermal sorption, whereas f_{mo} and f_{mr} are the metabolized fraction of a certain carcinogenic contaminant from ingestion and inhalation (or dermal exposures) respectively. Note that the metabolized fractions are dimensionless [*Maxwell and Kastenberg*, 1999]. For this work, we will consider only risk due to ingestion of tap water. *USEPA* [1989] illustrates how to evaluate risk for other pathways and mathematical expressions for ADD_H and ADD_G are given in *Maxwell et al.* [1998].

The average daily dose for the groundwater pathway is dependent on the flux-averaged concentration at some specific location, C_f [*Kreft and Zuber*, 1978; *Cvetkovic et al.*, 1992; *Dagan et al.*, 1992], and on human behavioral and exposure parameters [*USEPA*, 1989; *McKone and Bogen*, 1991; *Maxwell and Kastenberg*, 1999]. The flux-averaged concentration is the link between the contaminant source and the receptor. The average daily exposure is defined by *USEPA* [1989]:

$$ADD_G = C_f \times \frac{IR}{BW} \frac{ED \times EF}{AT}, \tag{2.4}$$

where IR is the ingestion rate of water (l/d), BW body weight (kg), AT is the average time of the expected lifetime (d), ED is the exposure duration (y) and EF is the daily exposure frequency (d/y). All of these parameters are based on EPA guidelines [*USEPA*, 1989].

The flux-averaged concentration at any point in space ($\mathbf{x}$) and time (t) is given by the ratio between the total mass flux $Q(\mathbf{x}, t)$ and the measured water flux $Q_w(\mathbf{x})$:

$$C_f(\mathbf{x}, t) = \frac{Q(\mathbf{x}, t)}{Q_w(\mathbf{x})} \ . \tag{2.5}$$

Due to uncertainty and spatial variability of the hydraulic conductivity, $Q(\mathbf{x}, t)$ is regarded here as a SRF [*Dagan*, 1984, 1987]. Expressions for the mean of the total solute mass flux, $\langle Q(\mathbf{x}, t)\rangle$, and its variance, $\sigma_Q^2(\mathbf{x}, t)$, are available depending on how fast the contaminant is released into the aquifer and its source geometry [*Cvetkovic et al.*, 1992; *Dagan et al.*, 1992; *Andricevic and Cvetkovic*, 1996, 1998; *Rubin*, 2003].

We will assume that C_f is the actual concentration of the chemical present in tap water. Other authors have incorporated a larger number of wells to their simulations and presented a methodology to average the concentration of these numerous wells [*Maxwell et al.*, 1999].

Increased cancer risk can be estimated using an average concentration over the exposure duration or the peak concentration:

$$C_f(\mathbf{x}) = \frac{1}{Q_w(\mathbf{x})}\left[\frac{1}{ED}\int_{T_i}^{T_i+ED} Q(\mathbf{x}, t)dt\right] , \tag{2.6}$$

with T_i being the time where exposure begins. If the peak concentration is used to estimate increased cancer risk, C_f will have the following form:

$$C_f(\mathbf{x}) = \frac{1}{Q_w(\mathbf{x})}\max_t\{Q(\mathbf{x}, t)\} . \tag{2.7}$$

Here, C_f lumps all transport and flow related variables ($\boldsymbol{\theta}_H$) while the parameter vector $\boldsymbol{\theta}_P$ includes IR, BW, ED, AT, f_{mo} and CPF_M for each (or an average) individual. This allows us to use the approach described in the previous section to investigate the relation between uncertainties from hydrogeological variables and health related parameters.

2.3.2 Mathematical statement of $F_R(r)$

To obtain a closed-form expression for equation (2.1), we need the first two moments of risk, assuming its distribution is normal or lognormal. As seen previously, increased cancer risk is a function of C_f. A deterministic increased cancer risk was denoted by r and the random variable for increased cancer risk is denoted by R. Since the uncertainty and variability of the hydrogeological media is lumped within C_f, we may write the moments of R as follows:

$$\langle R \rangle = \int_0^\infty R(c_f) f_c(c_f) dc_f \tag{2.8}$$

$$\langle R^2 \rangle = \int_0^\infty R^2(c_f) f_c(c_f) dc_f, \tag{2.9}$$

where $f_c(c_f)$ is the PDF of C_f.

To derive explicit expressions for equations (2.8) and (2.9) we make use of the travel time approach to evaluate solute fluxes at a plane in space perpendicular to the mean flow direction [*Shapiro and Cvetkovic*, 1988; *Dagan and Nguyen*, 1989; *Dagan et al.*, 1992; *Cvetkovic et al.*, 1992; *Rubin and Dagan*, 1992; *Andricevic et al.*, 1994; *Andricevic and Cvetkovic*, 1996; *Rubin*, 2003]. Travel time is the time that solute particle released at time t_o takes to travel from an initial location **a** within the release source domain denoted by Ω, to a control plane situated a distance L from the source as shown in Figure 2.1.

Due to the spatial variability and uncertainty of the aquifer's properties (i.e.,the hydraulic conductivity), the travel time, denoted here by τ, is also random and a PDF conditioned on measurements, $g_1\ (\tau|L,\mathbf{a}, t_o, \boldsymbol{\theta}_H, \{m\})$, can be used to describe the travel time distribution at a control plane situated at a distance L [*Shapiro and Cvetkovic*, 1988; *Dagan and Nguyen*, 1989; *Rubin and Dagan*, 1992]. A Lagrangian formulation allows one to compute the solute mass flux at a con-

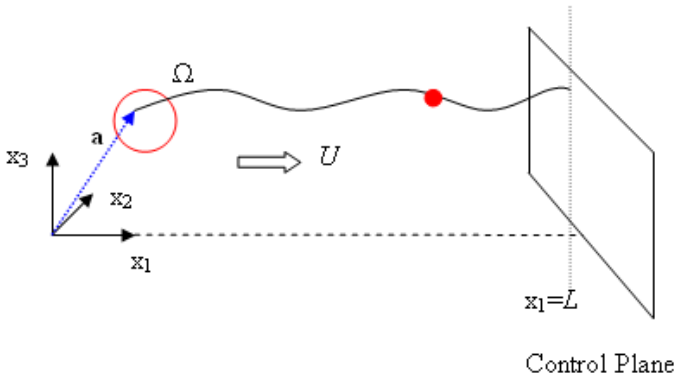


Figure 2.1: Problem configuration for an uniform-in-the-average flow with mean velocity U.

trol plane (used to denote an environmentally sensitive target) while accounting for variability and uncertainty of the subsurface [*Dagan et al.*, 1992; *Cvetkovic et al.*, 1992]. Details of derivations, closed-form solutions for $g_1(\tau|L, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \{m\})$ and theoretical aspects of this methodology can be found in the literature [*Dagan et al.*, 1992; *Cvetkovic et al.*, 1992; *Rubin*, 2003]. Now we may write the flux-averaged concentration as a function of the following variables, $C_f \equiv f(\tau,t,L,\mathbf{a},t_o,$ $\boldsymbol{\theta}_H,\{m\})$, and equations (2.8) and (2.9) may be rewritten in terms of the travel time PDF.

The random variable function $R(C_f)$ is now written as $R(t, \tau|L, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\})$ in order to explicit its dependence on travel time and other related parameters. The expected value of risk at any time t, given $\boldsymbol{\theta}_H$, $\boldsymbol{\theta}_P$ and $\{m\}$ at a control plane is as follows:

$$\langle R(t, \tau|L, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\})\rangle = \int_0^\infty R(t, \tau|L, \boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\}) g_1(\tau|L, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \{m\}) d\tau \quad (2.10)$$

For the second temporal moment of risk we make use of the two-particle travel time PDF, $g_2(\tau, \tau'|L, \mathbf{a}, \mathbf{a}', t_o, \boldsymbol{\theta}_H, \{m\})$. It is the probability density of the event that two particles originating at position vectors $\mathbf{a}$ and $\mathbf{a}'$, released at time t_o, will cross the control plane at time τ and τ'

respectively. The two-particle travel time PDF can be obtained analytically assuming a bivariate lognormal form [*Cvetkovic et al.*, 1992] or numerically [*Hassan et al.*, 2001, 2002]. The second moment of increased cancer risk is:

$$\begin{aligned}\langle R^2(t,\tau|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle &= \int_0^\infty\int_0^\infty R(t,\tau|L,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})R(t,\tau'|L,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) \\ &\quad\times g_2(\tau,\tau'|L,\mathbf{a},\mathbf{a}',t_o,\boldsymbol{\theta}_H,\{m\})d\tau d\tau'. \qquad (2.11)\end{aligned}$$

Assuming a Gaussian or lognormal distribution for risk, equations (2.10) and (11) allows us to explicitly write the cancer health risk CDF $F_R^c(r|\boldsymbol{\theta}_H,\boldsymbol{\theta}_P)$ at a given time t. Equation (2.10) is the mean value of risk while the variance of risk is given as:

$$\begin{aligned}\sigma_R^2(t|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) &\equiv \langle R^2(t,\tau|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle \\ &\quad -\langle R(t,\tau|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle^2. \qquad (2.12)\end{aligned}$$

Below we have the risk CDF conditioned on measurements assuming a lognormal form. The lognormal assumption relies on the fact that risk has to be positive and that it is a product of several parameters.

$$F_R^c(r|\boldsymbol{\theta}_H,\boldsymbol{\theta}_P) = \frac{1}{2}+\frac{1}{2}\text{Erf}\left[\frac{\ln(r)-\mu_R^*(t|\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})}{\sigma_R^*(t|\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\sqrt{2}}\right], \qquad (2.13)$$

where the mean and standard deviation of the variable's logarithm are given as:

$$\begin{aligned}&\mu_R^*(t|\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) = \\ &\quad \ln[\langle R(t,\tau|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle] - \frac{1}{2}\ln\left[1+\frac{\sigma_R^2(t|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})}{\langle R(t,\tau|L,\mathbf{a},t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle^2}\right] \qquad (2.14)\end{aligned}$$

$$\sigma_R^*(t|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\}) = \sqrt{1 + \frac{\sigma_R^2(t|L, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\})}{\langle R(t, \tau|L, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \boldsymbol{\theta}_P, \{m\})\rangle^2}}, \tag{2.15}$$

here the conditional information of L, $\mathbf{a}$ and t_o on the left hand side of equations (2.13)-(2.15) has been neglected to simplify the notation.

To consider uncertainty and/or variability in the human health risk parameters and hydrogeological parameters, we use Bayes' theorem given that the PDF for the physiological parameters, $f_P(\boldsymbol{\theta}_P)$, and hydrogeological parameters $f_H(\boldsymbol{\theta}_H)$ are known and independent:

$$F_R^c(r) = \int_0^r \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_R^c(\tilde{r}|\boldsymbol{\theta}_H, \boldsymbol{\theta}_P) f_P(\boldsymbol{\theta}_P) f_H(\boldsymbol{\theta}_H) d\boldsymbol{\theta}_P d\boldsymbol{\theta}_H d\tilde{r}. \tag{2.16}$$

Equation (2.16) provides the risk CDF considering the largest possible ensembles for the values of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_H$. Equation (2.16) is a useful tool allowing one to investigate the effect of parametric uncertainty in human physiology or exposure parameters in the final increased cancer risk distribution.

2.4 Solution of $F_R(r)$ for a Finite Duration Point Source Using a Lagrangian Stochastic Approach

The purpose of this section is to illustrate an example using the theoretical framework developed. Contamination occurs in an aquifer with spatially variable hydraulic conductivity, $\mathbf{K}(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, x_3)$ being the Cartesian coordinate system. The logconductivity is defined as $Y(\mathbf{x}) = \ln[\mathbf{K}(\mathbf{x})]$. The aquifer is characterized by the mean and variance of the natural logarithm of the hydraulic conductivity, m_Y and σ_Y^2 respectively, spatial covariance, C_Y, and the integral

scale, I_Y. Let us denote by $\mathbf{V} = (V_1, V_2, V_3)$ the groundwater velocity vector. The groundwater velocity satisfies the continuity equation $\nabla \cdot (\phi \mathbf{V}) = 0$, where ϕ is the porosity. The velocity field is heterogenous but considered stationary from a statistical perspective. Flow is at steady state and occurs at a low Reynolds number allowing the use of Darcy's Law.

Fate and transport is quantified by the flux-averaged concentration, $C_f(\mathbf{x}, t)$ expressed in equation (2.5) [*Kreft and Zuber*, 1978; *Cvetkovic et al.*, 1992; *Dagan et al.*, 1992]. It varies spatially and temporally and is a function of flow and transport parameters. A high Peclet condition is considered and transport is assumed to be advective and reactive.

2.4.1 Moments of the total solute mass flux

In the present subsection, expressions for the moments of the total solute mass flux are derived for a particular uniform-in-the-average flow. The mean flow is taken in the x_1 direction such that the longitudinal velocity is expressed as a sum of a mean value and its fluctuation, $V_1 = U + u'$. This implies that the streamlines are, in an average sense, parallel to the mean groundwater flow, where $U \equiv \langle V_1 \rangle$. The aquifer is confined, fully saturated and the variance of the logconductivity, σ_Y^2, is small allowing the use of the low-order approximation for the travel time moments given in previous works [*Cvetkovic et al.*, 1992; *Dagan et al.*, 1992]. These low-order approximations for the travel time moments were verified numerically in *Bellin et al.* [1993] and will be used to derive the statistical temporal moments of the mass flux.

The case of continuous contaminant release will be used as a starting point to derive the moments of the total solute mass flux. The contaminant source strength is given by the function $\dot{m}(\mathbf{a}, \widetilde{t})$. Chemical reactions are accounted by the pulse reaction function $\gamma(t - \widetilde{t}, \tau)$ [*Cvetkovic and Dagan*, 1994; *Andricevic and Cvetkovic*, 1996; *Cushey and Rubin*, 1997]. This function depends on

the chemical process occurring during solute transport.

The mass flux is quantified at a control plane situated a distance L from the source domain Ω (see Figure 2.1). All particles will be transported along a streamline and will cross the control plane in a given finite time. For a continuous source release we have:

$$Q(t,\tau|L,\Omega,t_o) = \int_{\Omega}\int_{t_o}^{t} \dot{m}(\mathbf{a},\widetilde{t})\gamma(t-\widetilde{t},\tau)\, d\widetilde{t}\, d^{\chi}\mathbf{a}, \tag{2.17}$$

with χ the dimensionality of the physical domain of the contaminant source. Equation (2.17) gives us the contribution at the control plane of a solute mass, $\dot{m}(\mathbf{a},\widetilde{t})d\widetilde{t}\, d^{\chi}\mathbf{a}$, released at a location $\mathbf{a} \in \Omega$ at $\widetilde{t}$.

Let M_o be the mass injected during a period T_o in a single location $\mathbf{a}_o \in \Omega$ at a time t_o. In this case, the source strength function is given by:

$$\dot{m}(\mathbf{a},\widetilde{t}\,|\mathbf{a}_o,T_o,t_o,M_o) = \frac{M_o}{T_o}\,\delta(\mathbf{a}-\mathbf{a}_o)\{H[\widetilde{t}-t_o] - H[\widetilde{t}-t_o-T_o]\}, \tag{2.18}$$

where $H[\cdot]$ is the Heaviside function. The total solute is:

$$Q(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o) = \int_{t_o}^{t} \frac{M_o}{T_o}\{H[\widetilde{t}-t_o] - H[\widetilde{t}-t_o-T_o]\}\gamma(t-\widetilde{t},\tau)d\widetilde{t}. \tag{2.19}$$

If non-reactive transport occurs, we have $\gamma(t-\widetilde{t},\tau) = \delta(t-\widetilde{t}-\tau)$ [*Cvetkovic and Dagan*, 1994]. Under the assumption of linear equilibrium, the γ pulse reaction function becomes a function of the retardation coefficient R_f [*Cvetkovic and Dagan*, 1994; *Cvetkovic et al.*, 1998]:

$$\gamma(t-\widetilde{t},\tau) = \delta(t-\widetilde{t}-\tau R_f) \tag{2.20}$$

Substituting equation (2.20) into equation (2.19) we obtain:

$$Q(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f)=\frac{M_o}{T_o}\{H[t-R_f\,\tau-t_o]-H[t-R_f\,\tau-t_o-T_o]\} \tag{2.21}$$

Note that in the above equation, R_f is deterministic. Taking the expected value for all possible travel time values, we obtain the first moment of the total solute mass flux at the control plane:

$$\langle Q(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})=\frac{M_o}{T_o}\Delta G_\tau(t|L,\mathbf{a}_o,T_o,t_o,R_f,\boldsymbol{\theta}_H,\{m\}), \tag{2.22}$$

with:

$$\Delta G_\tau(t|L,\mathbf{a}_o,T_o,t_o,R_f,\boldsymbol{\theta}_H,\{m\})=G_\tau\left(B|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right)-G_\tau\left(A|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right), \tag{2.23}$$

$$G_\tau(t|L)=\frac{1}{2}\text{Erfc}\left\{\frac{L-U\,t}{\sqrt{2\,X_{11}(t)}}\right\}, \tag{2.24}$$

and

$$A\equiv A(t|t_o,T_o,R_f)=\frac{t-t_o-T_o}{R_f}, \tag{2.25}$$

$$B\equiv B(t|t_o,R_f)=\frac{t-t_o}{R_f}, \tag{2.26}$$

where the functions $G_\tau\left(B|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right)$ and $G_\tau\left(A|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right)$ are the travel time CDF conditioned on measurements and closed-form expressions are found in the literature [*Ru-*

bin and Dagan, 1992; *Rubin*, 2003]. $X_{11}(t)$ is the displacement covariance, whose closed-form expressions are available [*Rubin*, 2003]. For this particular case, $X_{11}(t)$ is given as:

$$\frac{X_{11}(t)}{\sigma_Y^2 I_Y^2} = 2\frac{tU}{I_Y} + \frac{3}{2} - 3E + 3\left[Ei\left(-\frac{tU}{I_Y}\right) + \frac{e^{-\frac{tU}{I_Y}}\left(1+\frac{tU}{I_Y}\right)-1}{(\frac{tU}{I_Y})^2}\right], \tag{2.27}$$

where $Ei(\cdot)$ is the exponential integral and E is the Euler constant equal to 0.5777 [*Rubin*, 2003].

For the second temporal moment of the total solute mass flux at L, we make use of the following expression:

$$\lim_{\mathbf{a}\to\mathbf{a}_o} g_2(\tau,\tau'|L,\mathbf{a}_o,\mathbf{a},t_o,\boldsymbol{\theta}_H,\{m\}) = g_1(\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})\delta(\tau-\tau'), \tag{2.28}$$

since we are dealing with a point source $\mathbf{a}$=$\mathbf{a}_o$. So as a limiting case given in the literature [*Cvetkovic et al.*, 1992; *Andricevic and Cvetkovic*, 1996] we may write the two particle travel time PDF as in equation (2.28). So, we have:

$$\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \frac{M_o^2}{T_o^2}\Delta G_\tau(t|L,\mathbf{a}_o,T_o,t_o,R_f,\boldsymbol{\theta}_H,\{m\}), \tag{2.29}$$

with A and B defined in equations (2.25)-(2.26) and ΔG_τ is given in equation (2.23). An appendix is included with a detailed derivation of equation (2.29) (see Appendix A). In summary, expressions (2.22) and (2.29) are the temporal moments of the total solute mass flux for a point source of finite duration T_o. The variance of the total solute mass flux is given by:

$$\sigma_Q^2(t|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\}) \equiv \langle Q(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})^2\rangle - \langle Q(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle^2 \tag{2.30}$$

hence,

$$\begin{aligned}\sigma_Q^2(t|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\}) &= \frac{M_o^2}{T_o^2}\Delta G_\tau(t|L,\mathbf{a}_o,T_o,t_o,R_f,\boldsymbol{\theta}_H,\{m\}) \\ &- \frac{M_o^2}{T_o^2}[\Delta G_\tau(t|L,\mathbf{a}_o,T_o,t_o,R_f,\boldsymbol{\theta}_H,\{m\})]^2,\end{aligned} \tag{2.31}$$

where ΔG_τ is given in equation (2.23). The resulting equations derived here will be used in the next subsection to obtain the risk CDF conditioned on measurements.

2.4.2 Linear risk model

For small doses, a linear increased cancer risk model can be used instead of equation (2.2). A simplified equation for increased human cancer risk due to groundwater ingestion for any given time t is as follows:

$$r(t,\tau|L,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) = f_{mo} \times ADD_G(t,\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) \times CPF_M, \tag{2.32}$$

where ADD_G depends on total solute mass flux, see equations (2.4) and (2.5) and is written in terms of τ.

In order to separate which variables are health related and which ones are hydrogeological, we define a function that depends only on health related parameters, $\beta(\boldsymbol{\theta}_P)$:

$$\beta(\boldsymbol{\theta}_P) = f_{mo} \times CPF_M \times \frac{IR}{BW}\frac{ED \times EF}{AT}, \tag{2.33}$$

with $\boldsymbol{\theta}_P$=$\{f_{mo}, CPF_M, IR, BW, ED, EF, AT\}$. The variable $\beta(\boldsymbol{\theta}_P)$ incorporates all behavioral and physiological parameters (i.e., body weight, tap water intake, exposure duration, etc).

Re-arranging the terms in equation (2.32), the linearized increased cancer risk model becomes:

$$r(t,\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) = \beta(\boldsymbol{\theta}_P)\, C_f(t,\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}), \tag{2.34}$$

where the flux-averaged concentration, as shown in equation (2.5), is re-written in terms of travel time and is conditioned on hydrogeological parameters and measurements to allow the use of the theoretical framework of the previous section, $C_f(L,t) \equiv C_f(t,\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})$. The equation above involves two variables: $C_f(t,\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})$ and $\beta(\boldsymbol{\theta}_P)$. Through parametric uncertainty, see equation (2.16), we can account for uncertainty in $\boldsymbol{\theta}_P$.

Substituting equation (2.5) into equation (2.34), the first and second statistical moments of the increased cancer risk are given in equations below:

$$\langle R(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle = \frac{\beta(\boldsymbol{\theta}_P)}{Q_w(L)}\langle Q(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle \tag{2.35}$$

$$\sigma^2_R(t|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) = \left[\frac{\beta(\boldsymbol{\theta}_P)}{Q_w(L)}\right]^2 \sigma^2_Q(t|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\}) \tag{2.36}$$

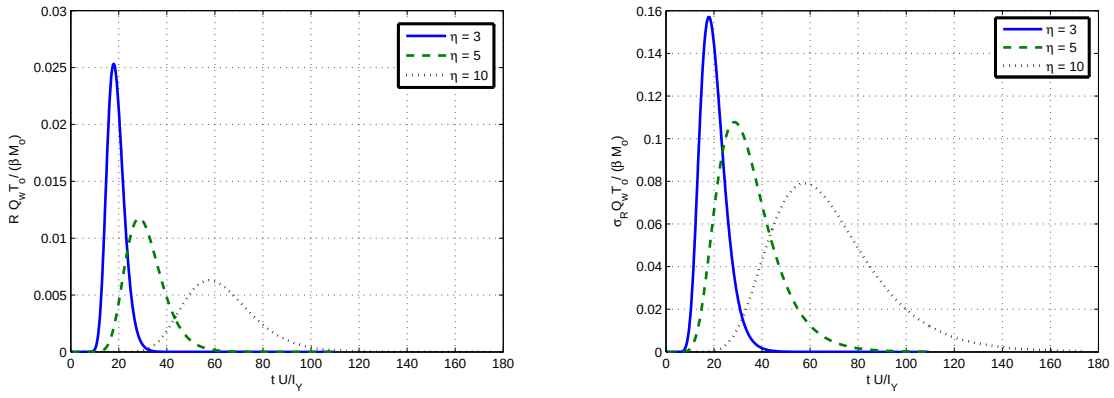
The first two temporal moments of the total solute mass flux are given in equations (2.22) and (2.29) for the particular case of a finite duration point source. Substituting equations (2.22) and (2.31) into (2.35) and (2.36), respectively, we get:

$$\langle R(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\})\rangle = \frac{\beta(\boldsymbol{\theta}_P)M_o}{Q_w(L)T_o}\left[G_\tau\left(B|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right) - G_\tau\left(A|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right)\right] \tag{2.37}$$

and

$$\begin{aligned}\sigma_R^2(t|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\boldsymbol{\theta}_P,\{m\}) = \\ \left[\frac{\beta(\boldsymbol{\theta}_P)M_o}{Q_w(L)T_o}\right]^2 \{G_\tau\left(B|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right) - G_\tau\left(A|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right) \\ - \left[G_\tau\left(B|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right) - G_\tau\left(A|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right)\right]^2\},\end{aligned} \tag{2.38}$$

with A and B defined in equations (2.25) and (2.26).



(a) **(b)**

Figure 2.2: Moments of the normalized increased cancer risk evaluated for three values of $\eta = 3$, 5 and 10 with η=L/I_Y. (a) Mean of R. (b) Standard deviation of R.

Plots illustrating how the mean and standard deviation of risk varies with time are given in

Figures 2.2.a and 2.2.b for different values of the ratio η=L/I_Y. The moments of risk were normalized by $\beta(\boldsymbol{\theta}_P)M_o/[Q_w(L)T_o]$ such that we may observe its behavior as a function of travel time. Note that the risk moments obtained in this section allow the use of several published expressions for the moments of the solute flux to account other physical scenarios [*Dagan et al.*, 1992; *Cvetkovic et al.*, 1992; *Andricevic and Cvetkovic*, 1996, 1998].

2.5 Relative contribution of information

This section illustrates an application of the theoretical framework developed to identify conditions where hydrogeological and physiological uncertainty will lead to a better understanding of risk. We wish to quantify the value of information gained in human health risk through uncertainty reduction in physiological and hydrogeological parameters. The effectiveness of conditioning $F_R(r)$ on measurements of hydraulic conductivity is also investigated. This section consists of three subsections. The data used in the simulations is presented in the first subsection; the impact of parametric uncertainty is investigated in the second subsection and finally the impact of measurements of hydraulic conductivity in human health risk is shown for the present model.

2.5.1 Data used in simulation

We consider the case of a finite duration point source detailed in the previous section. We selected the data given in Tables 2.1 and 2.2 to perform our simulations. Increased cancer risk due to ingestion of contaminated groundwater is evaluated with physiological data based from previous literature [*McKone and Bogen*, 1991; *Maxwell and Kastenberg*, 1999]. For our work, we simulated a PCE contamination case [*Maxwell and Kastenberg*, 1999; *Maxwell et al.*, 1999].

Flow and transport parameters	
L	1500, 7500 and 12500 m
U	10 m/d
I_Y	500 m
σ_Y^2	1.5
m_Y	0
R_f	2.5
T_o	1 y
t_o	0
ϕ	0.2

Table 2.1: Input data used for the flow and transport problem

The impact of uncertainty reduction in $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ on $F_R(r)$ is accounted for by making use of equation (2.16) and the PDF from Table 2.2. We have selected U as the parameter to investigate uncertainty from $\boldsymbol{\theta}_H$. It is assumed that the PDF for U is lognormal [*Andricevic et al.*, 1994; *Andricevic and Cvetkovic*, 1996].

From the human physiology component, we have selected f_{mo} and CPF_M since the pharmokinetic models, from which these parameters were derived, are uncertain [*Maxwell and Kastenberg*, 1999]. From the exposure and behavioral parameters, we have selected ED to be uncertain. Other risk-related parameters have the following values: IR/BW=0.033 L/d-kg, AT=22550 d and EF=350 d/y [*USEPA*, 1989; *McKone and Bogen*, 1991; *Dawoud and Purucker*, 1996].

The statistical distributions adopted for f_{mo}, CPF_M, ED and U are in Table 2.2. The

Uncertain parameters	
f_{mo}	Uniform$[0.2, 0.7]$
CPF_M	Uniform$[0.045, 0.175]$
ED	Uniform$[20, 40]$
U	Lognormal $(10, 3.16)$

Table 2.2: Statistical distributions for $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_H$

distribution bounds for f_{mo} and CPF_M were based, although not exactly the same, from previous published works [*McKone and Bogen*, 1991; *Maxwell et al.*, 1998]. We have assumed that both physiological and behavioral parameters were uniformly distributed. The numbers inside the parenthesis for the lognormal PDF are the arithmetic mean and standard deviation. For the uniform PDF, we specified the lower and upper bound of the distribution. These statistical distributions are assumptions and are used here for illustration purpose of the methodology proposed.

In order to investigate the value of information we will use the concept of entropy to quantify uncertainty for both physiological and hydrogeological components [*Christakos*, 1992; *Rubin*, 2003]. The entropy for $\boldsymbol{\theta}_P$ is denoted by E_P. We may write E_P as follows [*Christakos*, 1992]:

$$E_P = -\int_0^\infty f_P(\boldsymbol{\theta}_P) \ln[f_P(\boldsymbol{\theta}_P)] d\boldsymbol{\theta}_P. \tag{2.39}$$

For illustration purposes, we have assumed that f_{mo}, CPF_M and ED are independent such that the total entropy for physiology (E_P) is the sum of the individual entropies for each of these variables [*Christakos*, 1992]. Note that the proposed framework is not restricted by this

assumption. For the hydrogeological entropy, denoted by E_H, we write as follows:

$$E_H = -\int_0^\infty f_H(\boldsymbol{\theta}_H)\ln[f_H(\boldsymbol{\theta}_H)]d\boldsymbol{\theta}_H. \tag{2.40}$$

For each increase of E_H and E_P we have a corresponding entropy increase in the risk CDF relative to a risk CDF with less parametric uncertainty. This risk CDF with less parametric uncertainty is evaluated using reference entropies for both hydrogeology and physiology. These reference entropies are denoted by $E_{H,O}$ and $E_{P,O}$ and are obtained by making use of the distributions in Table 2.2 together with equations (2.39) and (2.40). In our definition, $E_H \geq E_{H,O}$ and $E_P \geq E_{P,O}$. To evaluate the impact of increased parametric uncertainty of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_H$ in $F_R(r)$ we will use two different metrics.

The first one is the relative entropy in risk, denoted by RE_R. The relative entropy is defined as [*Christakos*, 1992; *Rubin*, 2003]:

$$RE_R = \int_0^\infty f_R^o(r)\ln\left[\frac{f_R^o(r)}{f_R(r)}\right]dr, \tag{2.41}$$

where $f_R(r)$ is the risk PDF evaluated with E_H and E_P and $f_R^o(r)$ is the risk PDF evaluated with $E_{H,O}$ and $E_{P,O}$ corresponding to the statistical distributions in Table 2.2.

The second metric used is the percentage difference between the coefficient of variation for $F_R(r)$ evaluated with E_H and E_P (CV_R) and the coefficient of variation with $E_{H,O}$ and $E_{P,O}$ (CV_R^o). This is mathematically equivalent to:

$$\Delta CV_R = \frac{CV_R - CV_R^o}{CV_R} \tag{2.42}$$

The coefficient of variation for risk is given by: $CV_R{=}\sigma_R/\langle R\rangle$.

2.5.2 Hydrogeological uncertainty versus physiological and behavioral uncertainty

In the following, we focus on understanding how uncertainty reduction from $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ impacts $F_R(r)$. Afterwards, the second part of this subsection is dedicated to illustrate a graphical tool that can be used to investigate whether the trade-off between uncertainty reduction from $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ exists. All plots presented in this subsection were evaluated using peak concentration, equation (2.7).

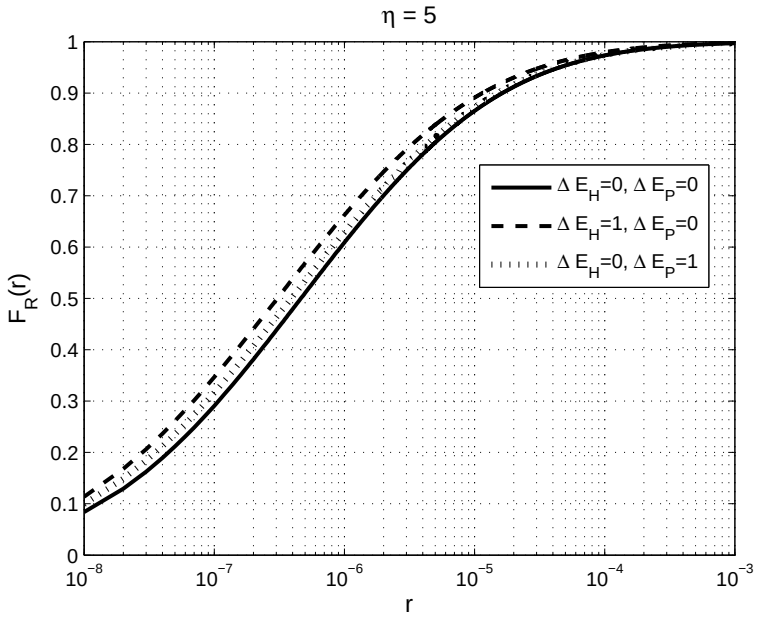


Figure 2.3: Impact of increased uncertainty of $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ in F_R at $\eta = 5$. Curves for (i) ΔE_P=ΔE_H= 0 ; (ii)ΔE_P= 0 and ΔE_H= 1; (iii)ΔE_P=1 and ΔE_H= 0. Where η=L/I_Y, ΔE_H=$E_H - E_{H,O}$ and ΔE_P=$E_P - E_{P,O}$.

Figure 2.3 depicts the comparison of $F_R(r)$ for three scenarios. The first one is evaluating $F_R(r)$ with the distributions in Table 2.2 and corresponding entropies $E_{H,O}$ and $E_{P,O}$. In the second case, we increase the entropy in $\boldsymbol{\theta}_H$ while keeping fixed the entropy in $\boldsymbol{\theta}_P$ at $E_{P,O}$. The last case is the opposite: We increase the entropy in $\boldsymbol{\theta}_P$ while keeping fixed the entropy in $\boldsymbol{\theta}_H$ at $E_{H,O}$. For Figure 2.3 and the subsequent plots we have defined ΔE_H=$E_H - E_{H,O}$ and ΔE_P=$E_P -$

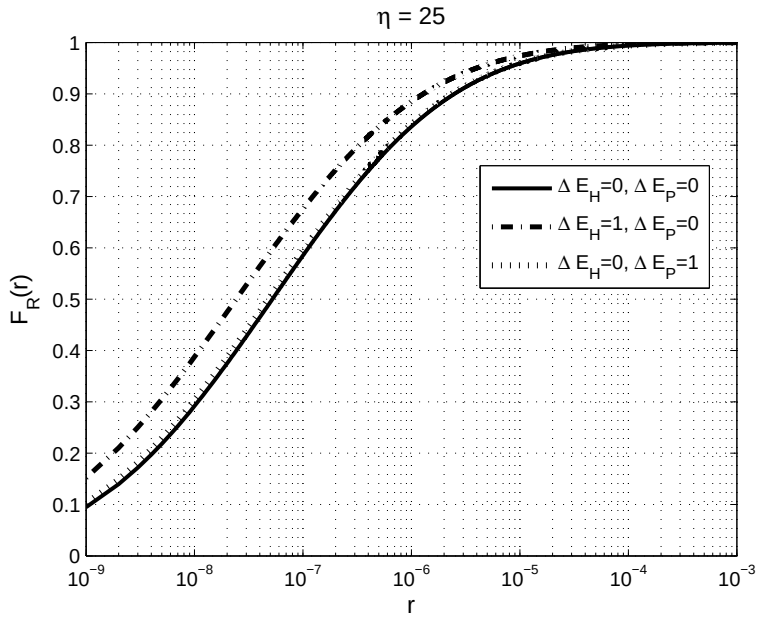


Figure 2.4: Impact of increased uncertainty of $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ in F_R at $\eta = 25$. Curves for (i) ΔE_P=ΔE_H= 0 ; (ii)ΔE_P= 0 and ΔE_H= 1; (iii)ΔE_P=1 and ΔE_H= 0. Where η=L/I_Y, ΔE_H=$E_H - E_{H,O}$ and ΔE_P=$E_P - E_{P,O}$.

$E_{P,O}$ to quantify change in entropy. Results were obtained for η=5 and we note that the effect of parametric uncertainty in $F_R(r)$ is not that strong for the selected data. From this plot we observe that $F_R(r)$ is more sensitive towards parametric uncertainty in hydrogeology, shown by the dashed curve, than uncertainty in human physiology and behavioral habits, represented here by the dotted curve.

Figure 2.4 shows results for η=25. Comparing Figures 2.3 and 2.4 we observe that increased parametric uncertainty in $\boldsymbol{\theta}_H$ has a stronger impact in $F_R(r)$ at larger η. Similar results for parametric uncertainty in flow and transport were found in *Andricevic and Cvetkovic* [1996]. One possible explanation for this phenomenon at larger η may be due to the fact that the solute plume has to travel through many more integral scales, thus suffering from variability leading to lower solute flux peaks (since chemical reaction occurs) and smoother breakthrough curves. Based on this

plot, reducing uncertainty in flow and transport parameters at larger η through site characterization improves risk estimates.

We also found that the impact of uncertainty reduction in $\boldsymbol{\theta}_P$ diminishes as η increases. This opposite behavior in parametric uncertainty reduction on physiological parameters, as opposed to uncertainty reduction in flow and transport, occurs because the concentration of the plume is much smaller at longer travel times (large η) than at early travel times (small η), thus posing less risk when compared to the case shown in Figure 2.3. With smaller concentrations, humans are at less or at almost no risk independent of how much of that contaminant is being metabolized. From a management point of view, reducing uncertainty from physiological parameters of the population may not contribute to better risk estimates at large values of η, which are associated with longer travel times.

Next we present a graphical tool that allows decision makers to set priorities in contaminated site remediation. In order to investigate trade-offs between hydrogeological and physiological uncertainties we need to define a metric that relates to the amount of information from these two components. We will make use of the concept of entropy, previously shown, to derive this metric that can relate to uncertainties from $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ and its impact in $F_R(r)$. Let α be defined as follows:

$$\alpha = \frac{10^{\Delta E_H}}{10^{\Delta E_P}}, \tag{2.43}$$

where ΔE_H=$E_H - E_{H,O}$ and ΔE_P=$E_P - E_{P,O}$.

Loss of information in $\boldsymbol{\theta}_H$ means α increasing to values greater than one. This is done by increasing ΔE_H and keeping ΔE_P equal to zero. If uncertainty increases in $\boldsymbol{\theta}_P$ while keeping ΔE_H=0, α decreases to values less than one with a lower bound equal to 0.

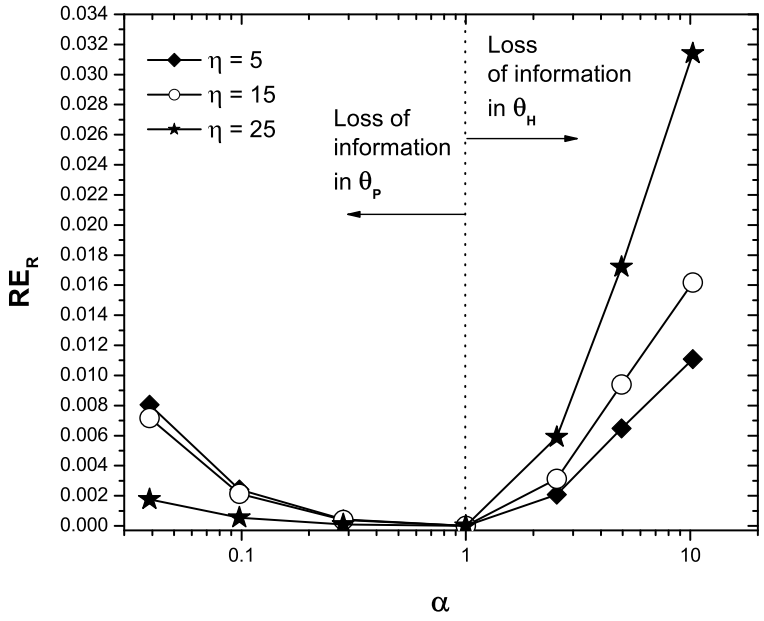


Figure 2.5: RE_R as a function of α for η=5, 15 and 25. Loss of information in $\boldsymbol{\theta}_H$ means increasing ΔE_H and keeping ΔE_P= 0 such that $\alpha > 1$. Loss of information in $\boldsymbol{\theta}_P$ means increasing ΔE_P and keeping ΔE_H= 0 such that $0 < \alpha < 1$.

When α equals to one, we have ΔE_H=ΔE_P=0 (with the following constraint: $\Delta E_H \neq \Delta E_P$ for values different than zero). This means the amount of information present in $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ at α = 1 are E_H=$E_{H,O}$ and E_P=$E_{P,O}$ respectively. From a management point of view, the choice of values for $E_{H,O}$ and $E_{P,O}$ that corresponds to α = 1 can be associated with an acceptable risk variance given by regulation. For example, one may determine $E_{H,O}$ and $E_{P,O}$ by relating the risk variance as a function of the solute flux variance, see equation 2.31, as well as the variances of other risk related parameters. In this case, uncertainty reduction in both $E_{H,O}$ and $E_{P,O}$ is constrained by risk regulation. As explained before, we will assume $E_{H,O}$ and $E_{P,O}$ are determined from the

distributions present in Table 2.2 for illustration purposes. The point α = 1 represents the baseline case, serving as a reference to the other levels of uncertainty and will allow one to investigate the relative contribution of information from each component essential to risk management. For example, one may wish to see by how much the coefficient of variation from $F_R(r)$ will change for a change in α. The intention is to relate changes in RE_R or ΔCV_R, equations (2.41) and (2.42), to α. Figures 2.5 and 2.6 illustrates the idea for different values of η.

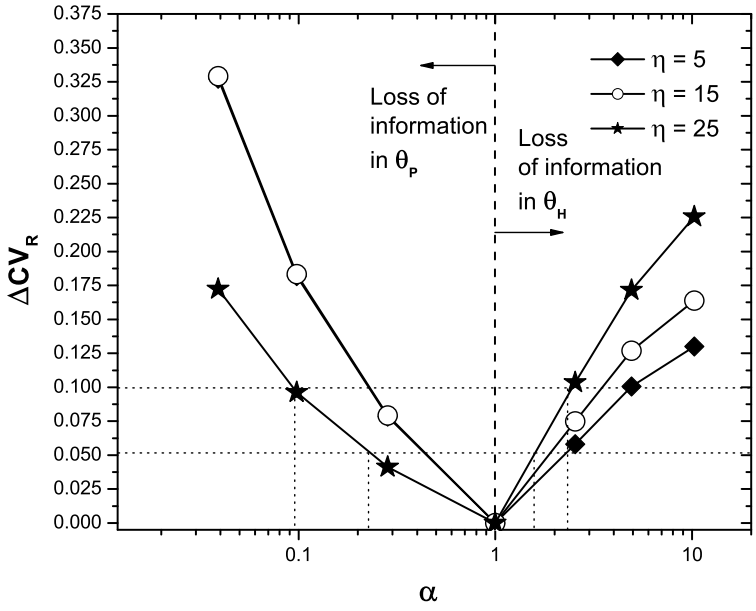


Figure 2.6: ΔCV_R as a function of α for η=5, 15 and 25. Loss of information in $\boldsymbol{\theta}_H$ means increasing ΔE_H and keeping ΔE_P= 0 such that $\alpha > 1$. Loss of information in $\boldsymbol{\theta}_P$ means increasing ΔE_P and keeping ΔE_H= 0 such that $0 < \alpha < 1$.

As seen in Figure 2.5, we have RE_R as a function of α. The behavior presented in Figures 2.3 and 2.4 is also reproduced in this plot. As η increases, we have larger values of RE_R (for $\alpha > 1$)

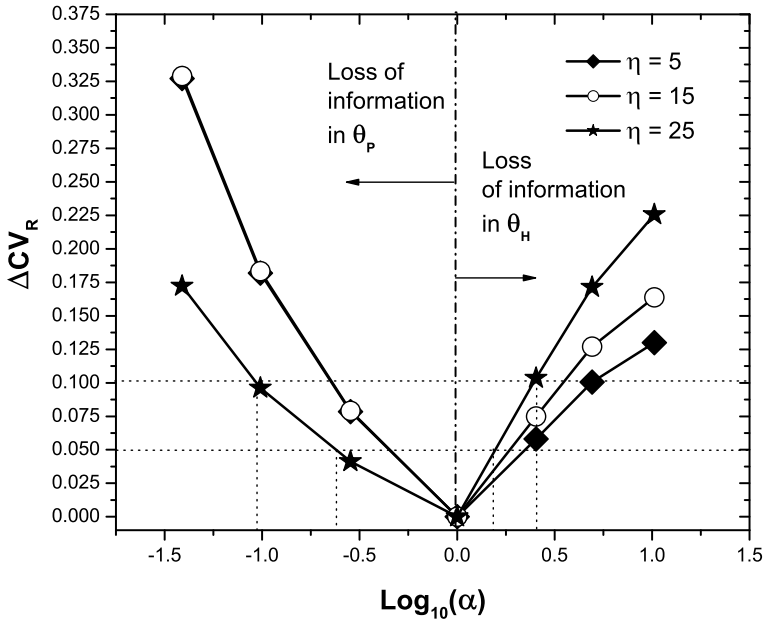


Figure 2.7: ΔCV_R as a function of $\log(\alpha)$ for η=5, 15 and 25. Loss of information in $\boldsymbol{\theta}_H$ means increasing ΔE_H and keeping ΔE_P= 0 such that $\log(\alpha) > 0$. Loss of information in $\boldsymbol{\theta}_P$ means increasing ΔE_P and keeping ΔE_H= 0 such that $\log(\alpha) < 0$.

while the opposite behavior occurs for $0 < \alpha < 1$. Parametric uncertainty reduction in hydrogeology becomes more important for larger η while parametric uncertainty reduction in physiology becomes less pronounced.

An alternative way of plotting these results is presented in Figure 2.6. Figure 2.6 depicts ΔCV_R as a function of α and illustrates how this graphical information could be used for decision making. For an instance, say we wanted to reduce ΔCV_R from 0.1 to 0.05, see pointed lines in Figure 2.6. Where should we set priorities towards uncertainty reduction? Should we invest in understanding human physiology or in reducing uncertainty in flow and transport parameters? By

calculating the slope of the curve corresponding to ΔE_P (located left of α=1) and the slope for ΔE_H (located right of α=1), one may evaluate the relative impact of each component in $F_R(r)$ as well as the efforts necessary to reduce E_P and E_H at ΔCV_R = 0.1 to their respective values at ΔCV_R=0.05.

It is also possible to graphically relate ΔCV_R to the actual values of ΔE_P and ΔE_H by plotting ΔCV_R versus $\log(\alpha)$. This is illustrated in Figure 2.7. For $\alpha > 1$ ($\log \alpha > 0$), we have $\log(\alpha) = \Delta E_H$ while for $0 < \alpha < 1$ ($\log \alpha < 0$) we have $\log(\alpha) = -\Delta E_P$.

In summary, the proposed analysis, presented in Figures 2.5- 2.7, permits decision makers to investigate whether it is cheaper to reduce ΔCV_R (or RE_R) via hydrogeology or physiology.

2.5.3 Link between site characterization, environmental regulation and human exposure duration

Lastly, we investigate the interplay between exposure duration and site characterization. Previous works studied the influence of ED on risk [*Maxwell and Kastenberg*, 1999; *Hassan et al.*, 2001]. Here we will further investigate how ED may influence site characterization based on health risk regulation. Environmental regulatory agencies [*USEPA*, 1989] suggests that the contaminant concentration should be averaged over the exposure duration period, shown in equation (2.6). In general, a 30 or 70 year period is used for exposure duration [*USEPA*, 1989; *Maxwell et al.*, 1998]. Depending on the type of contaminant and the characteristics of the target population, some regulations require the use peak concentrations to evaluate increased cancer risk [*Andricevic and Cvetkovic*, 1996]. The peak concentration criteria is also used to develop concentration guidelines for decommissioning processes, such as shown in *Taylor et al.* [2003], to account for the worst case scenario.

To evaluate the effect of measurements of hydraulic conductivity in human health risk we make use of the data available from the literature [*Rubin and Dagan*, 1992]. In Figure 2.8 we have the travel time CDF for three different hydraulic conductivity sampling scenarios: unconditioned on measurements $\{m_1\}$, conditioned on a sparse grid of measurements $\{m_2\}$, and finally conditioned on a dense grid of measurements $\{m_3\}$. The travel time CDF present in Figure 2.8 were obtained from *Rubin and Dagan* [1992] through interpolation of their published data. These interpolated curves were used as input for the derived risk CDF. It is possible to notice a bias of measurements in favor of low hydraulic conductivity measurements. A uniform-in-the-average flow condition is the underlying assumption of the travel time CDF shown in Figure 2.8. These travel time CDF will be used to illustrate how measurements may affect site characterization decisions. The travel time CDF unconditioned on measurements $\{m_1\}$ and conditional on a dense grid of hydraulic conductivity measurements $\{m_3\}$, shown in Figure 2.8, were chosen to obtain the results in Figures 2.9.a and 2.9.b.

Figure 2.9.a depicts how increased sampling of the hydraulic conductivity plays a role in site characterization when risk is evaluated using the peak concentration, equation (2.7), instead of the averaged concentration over the breakthrough curve, equation (2.6). For this particular example, we selected parameters such that the travel time distribution represents a case in which all of the solute mass arrives at the control plane within the 30 year exposure duration period for both conditional and unconditional cases. We assume exposure begins when the first solute particles arrive at the control plane. When averaging the breakthrough curve over the 30 year exposure period for this case, all solute mass is included in the averaging procedure for both conditional and unconditional cases. This means C_f of equation (2.6) and its expected values are the same for $\{m_1\}$ and $\{m_3\}$.

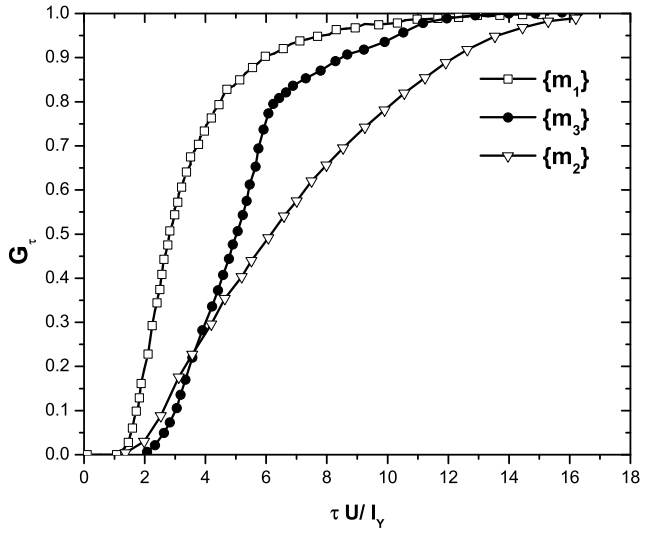


Figure 2.8: Travel time distributions for uniform-in-the-average flow. Travel time distribution unconditioned of measurements $\{m_1\}$; travel time conditioned on a sparse grid of hydraulic conductivity measurements $\{m_2\}$; and travel time conditioned on a dense grid of hydraulic conductivity measurements $\{m_3\}$ [*Rubin and Dagan*, 1992].

From this we conclude that $F_R(r)$ becomes more sensitive towards increased sampling of hydraulic conductivity when the duration of the breakthrough is larger than ED. If conditioning causes the contaminant mass within the averaging period to change, the impact of measurements on $F_R(r)$ is larger.

The sensitivity of $F_R(r)$ toward sampling may be generalized by its dependence on residence time of the contaminant plume at the control plane. This residence time is defined as the time period that the contaminant plume takes to cross the control plane (in other words, the time window where the concentration is greater than zero). If conditioning travel time to measurements

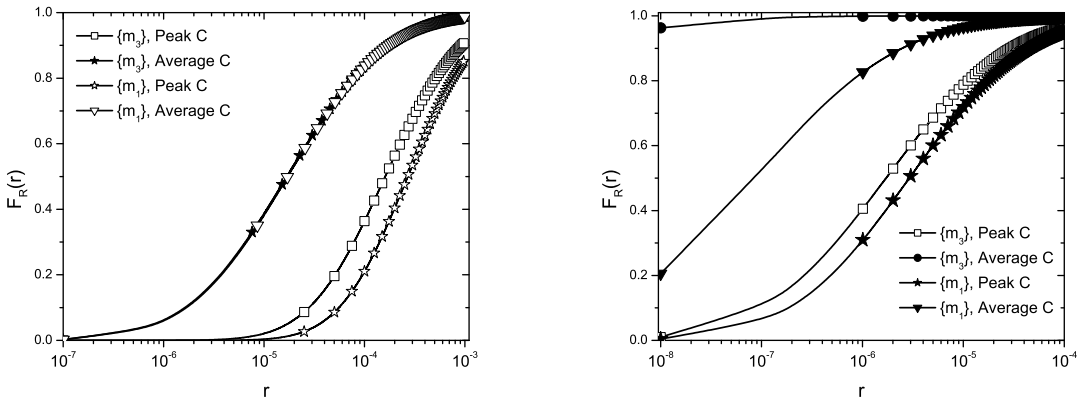


Figure 2.9: Increased cancer risk cumulative distribution function evaluated with average and peak concentrations using $\{m_1\}$ and $\{m_3\}$. (a) All mass arrives at the control plane before ED. (b) Not all mass arrives at the control plane before ED.

of hydraulic conductivity causes the breakthrough curve to change such that the plume's residence time on the control plane increases to values larger than ED, then $F_R(r)$ becomes sensitive to increased sampling even if an averaged breakthrough curve is used to evaluate risk. Figure 2.9.b illustrates $F_R(r)$ for the case in which not all contaminant mass arrives at the control plane during the averaging exposure time ED for both conditioned and unconditioned scenarios. In this case, hydrogeological characterization becomes important independent of the adopted concentration criteria (peak versus average).

2.6 Summary and conclusions

Human health risk was addressed using a stochastic framework to account for uncertainties and variabilities present in hydrogeology, human behavioral parameters and human physiology. A closed-form CDF for increased cancer health risk was derived and the dependence of hydrogeological and human health information is explicit. The temporal moments of total solute mass flux were analytically obtained by making use of the Lagrangian formulation. The impact of additional measurements of hydraulic conductivity on the increased cancer risk CDF was investigated. We also examined conditions in which reduced or increased uncertainty in risk related parameters, such as f_{mo} and CPF_M, are most likely to affect our understanding of risk. The developed methodology also investigated the trade-off between parametric uncertainty reduction from $\boldsymbol{\theta}_H$ and $\boldsymbol{\theta}_P$ in the final risk CDF through a graphical approach. The present results are obtained by making a few simplifying assumptions: (i) A steady state and uniform-in-the-average flow, (ii) the logconductivity variance σ_Y^2 is not large, (iii) human health risk is assumed to be lognormally distributed and (iv) physiological parameters are independent.

Uncertainty in human health risk parameters and hydrogeological parameters were incorporated through parametric uncertainty. For the model, data and simulation presented in this work, uncertainty reduction in $\boldsymbol{\theta}_P$ and in $\boldsymbol{\theta}_H$ have small impact on $F_R(r)$ as previously shown. The effect of parametric uncertainty in $\boldsymbol{\theta}_P$ decreases as the distance between the contaminant source and the control plane increases. The low risk values are normally associated with longer travel time values and, consequently, low concentrations. For low concentration values, $F_R(r)$ becomes less sensitive to how much an individual metabolizes the contaminated dose. High concentrations of the chemical present in the groundwater implies that the individual is most likely to be at risk and uncertainty

reduction in the physiological parameters at early travel times may become important. The link between physiology and early times suggests that further analysis should be done in understanding how other types of chemical reactions affects uncertainty of risk related parameters on the human health risk CDF. In this study we used a linear equilibrium reaction. The impact of parametric uncertainty in $\boldsymbol{\theta}_H$ increases for larger distances between the control plane and contaminant source.

The strongest contribution of this chapter is the introduction of a graphical tool that allows one to investigate the relative impact of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_H$ on $F_R(r)$. It consisted of developing a metric, α, that accounts for a ratio between information from physiology and hydrogeology. We developed this metric in such a way that we could analyze the increase of uncertainty in risk around a reference point α = 1. We used the concept of entropy in order to quantify the total amount of information contained in $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_H$. By plotting ΔCV_R or RE_R versus α (or $\log \alpha$) we are able to investigate the relative value of information from hydrogeology and risk related parameters. Values of α larger than one implies uncertainty increase in $\boldsymbol{\theta}_H$ while α less than one means uncertainty increase in $\boldsymbol{\theta}_P$. By quantifying the slopes of these curves, located to the left and right of α = 1, and the corresponding entropies one may observe where uncertainty reduction through data acquisition, via flow physics or via physiology, will lead to a better risk estimate. The ratio α is a promising complementary tool that may assist decision makers in setting priorities in site characterization.

The interplay between exposure duration and hydrogeological site characterization on $F_R(r)$ was investigated. Hydrogeological site characterization becomes dependent on the time the contaminant plume takes to cross the control plane if the concentration averaged over the exposure duration period is used to evaluate $F_R(r)$. If conditioning travel time causes the mass contained in the ED averaged breakthrough curve to change then site characterization improves our understand-

ing of risk. Assuming the exposure begins when the first solute particles arrive at the control plane we can state the following: If the residence time, defined here as the time period that the contaminant plume takes to cross the control plane, of both conditional and unconditional plume is less than or equal to the exposure duration period, hydrogeological site characterization does not greatly impact $F_R(r)$. This is true when using an averaged concentration breakthrough curve to estimate risk. However, if the residence time of both conditional and unconditional contaminant plume in the control plane is larger than the averaging period (i.e. exposure duration) then hydrogeological site characterization may become important. When using peak concentration to assess adverse health effects, the influence of additional sampling is more pronounced and affects $F_R(r)$ independent of ED. This result also shows us that exposure duration based on regulation plays an important role in estimating adverse health effects with strong management implications and should be well quantified.

To illustrate the theoretical framework, results were obtained with the classical absolute dispersion theory developed in *Dagan et al.* [1992] and *Cvetkovic et al.* [1992]. However, different results can be obtained when using the relative dispersion framework [*Andricevic and Cvetkovic*, 1998; *Hassan et al.*, 2001, 2002]. These differences occur when evaluating the magnitude and arrival time of the peak of the solute flux moments for the case of small contaminant source sizes. The relative dispersion results approaches and absolute dispersion framework as contaminant source size increases [*Andricevic and Cvetkovic*, 1998].

Although pore-scale dispersion was not included in the present analysis we recognize that it may play a role in risk assessment (in fact, this will be shown in the next chapter). Analytical expressions obtained in *Fiori et al.* [2002] for instantaneously injected plumes could be easily

adapted to the present framework. As shown in *Fiori et al.* [2002], pore-scale dispersion leads to the reduction in the peak of the mass flux due to the mass transfer between Darcy-scale stream tubes leading to smaller risk values. However, the impact of pore-scale dispersion is dependent on the scale of the sampling area that collects the contaminated water. If the size of the sampling area is of large dimensions, characterization efforts of dispersivities may not be as important since the mixing effect induced by sampling becomes more important than pore-scale dispersion.

It is worth mentioning that most of the results obtained in the present chapter are limited to simple scenarios (small values of σ_Y^2) and extending a similar analysis to more complicated problems may lead to many interesting conclusions when it comes to the uncertainty trade-offs between hydrogeology and physiology. However, the present methodology framework allows one to obtain *a priori* and important information through the use of the analytical expressions derived.

Notation

Symbols and their respective units:

$\mathbf{a}$: Initial location of the source [L]

$\mathbf{a}_o$: Point source location [L]

AT: Average time [t]

ADD_M: Average daily dose [M/(M t)]

ADD_G: Average daily dose for groundwater intake [M/(M t)]

ADD_H: Average daily dose for inhalation [M/(M t)]

ADD_D: Average daily dose for dermal sorption [M/(M t)]

A, B: Travel time variables [t]

C_f: Flux-averaged concentration [M/L^3]

CPF_M: Cancer potency factor [(M t)/M]

C_Y: Spatial covariance of the log-conductivity

CV_R, CV_R^o: Coefficient of variation for risk [-]

ED: Exposure duration [t]

EF: Exposure frequency [t/t]

$E_P, E_{P,O}$: Joint entropy for physiological parameters

$E_H, E_{H,O}$: Entropy for hydrogeological parameters [-]

f: Generic function

$f_C(c_f)$: The PDF for C_f

$f_R(r)$, $f_R^o(r)$: Risk PDF [-]

$f_P(\boldsymbol{\theta}_P)$: PDF for risk related parameter

$f_H(\boldsymbol{\theta}_H)$: PDF for hydrogeological parameters

f_{mo}: Metabolized fraction of contaminant ingested [-]

f_{mr}: Metabolized fraction of contaminant from dermal contact [-]

$F_R(r)$: Risk CDF [-]

$F_R^c(r)$: Risk CDF conditioned on measurements [-]

$g_1(\tau)$: Travel time PDF [1/t]

$g_2(\tau, \tau')$: Travel time joint PDF [(1/t)^2]

$G_\tau(\tau)$: Travel time CDF [-]

$H(\cdot)$: Heaviside function

h: Generic function [1/t]

I: Number of individuals in the exposed population [-]

I_Y: Integral scale of the aquifer [L]

IR/BW: Ingestion rate per body weight [L^3/(t M)]

K_i: Hydraulic conductivity at a specific location $\mathbf{x}_i$ [L/t]

$\mathbf{K}(\mathbf{x})$: Hydraulic conductivity at a generic location $\mathbf{x}$ [L/t]

L: Euclidean distance between the contaminant source and the control plane [L]

$\{m\}$: Set of measurements

$\{m_1\}$ Data unconditioned on $\mathbf{K}(\mathbf{x})$ measurements

$\{m_2\}$: Data conditioned on a sparse grid of $\mathbf{K}(\mathbf{x})$ measurements

$\{m_3\}$: Data conditioned on a dense grid of $\mathbf{K}(\mathbf{x})$ measurements

M_o: Mass injected [M]

$\dot{m}$: Mass release function [M $t^{-1}L^{-3}$]

m_Y mean of the log conductivity [-]

N: Number of locations sampled [-]

$\mathbf{P}_i$: Behavioral and exposure parameters for the i^{th} individual

$Q(\mathbf{x}, t)$: Total solute mass flux [M/t]

$\langle Q \rangle$: Expected value of solute flux [M/t]

$\langle Q^2 \rangle$: Second moment of solute flux [(M/t)^2]

$Q_w(\mathbf{x})$: Water flux at the control plane [L^3/t]

R_f: Retardation coefficient [-]

r: Increased cancer risk or simply risk [-]

$\langle R \rangle$: Expected value of increased cancer risk [-]

$\langle R^2 \rangle$: Second moment of increased cancer risk [-]

t_o: Time source injection begins [t]

T_o: Period of injection of contaminant [t]

T_i: Time exposure begins [t]

U, $\langle V_1 \rangle$: Mean velocity in the x_1 direction [L/t]

u': Velocity fluctuation in the x_1 direction [L/t]

$\mathbf{V}$: Velocity vector [L/t]

V_1, V_2, V_3: Components of the velocity vector [L/t]

Y: Logarithm of the hydraulic conductivity

$X_{11}(t)$: Particle displacement covariance [L^2]

$\mathbf{x}$: Cartesian coordinate system [L]

α: Ratio between entropies [-]

β: Risk related parameter [L^3/M]

δ: Dirac delta

ΔCV_R: Relative difference for the risk coefficient of variation [-]

ΔE_H, ΔE_P: Entropy difference [-]

γ: Reaction release function [t^{-1}]

μ_R^*: Mean of the random variable's logarithm [-]

η: Ratio between L and I_Y [-]

Ω: Finite source volume [L^3]

ϕ: Porosity [-]

σ_{fmo}: Standard deviation of the f_{mo} [-]

σ_{IRBW}: Standard deviation of the IR/BW [L^3/(t M)]

σ_Q^2: Variance of solute flux [(M/t)^2]

σ_Y^o: Non-perturbed log-conductivity standard deviation

σ_Y^2: variance of the log conductivity [-]

σ_R^2: Variance of increased cancer risk [-]

σ_R^*: Standard deviation of the random variable's logarithm [-]

τ: Travel time [t]

$\boldsymbol{\theta}_H$: Hydrogeological parameter vector

$\boldsymbol{\theta}_P$: Risk related parameter vector

χ: Number of dimensions of the contaminant source [-]

Chapter 3

The Concept of Comparative Information Yield Curves and Their Application to Risk-Based Site Characterization

3.1 Introduction

As described in the previous chapter, obtaining accurate predictions of potential human health risks from groundwater contamination is a challenge. The main difficulty lies in the fact that many of the factors that constitute risk are uncertain. Amongst these, we highlight two classes of parameters: (i) hydrogeological and (ii) physiological. Hydrogeological parameters are necessary

[2]This chapter is based on a published article in Water Resources Research, 2009. (*In Press*, doi:10.1029/2008WR007324)

to estimate fate and transport of pollutants in the subsurface as well as the level of contamination to which humans potentially will be exposed. Because of aquifer heterogeneity [*Dagan*, 1984, 1987; *Rubin and Dagan*, 1992; *Rubin*, 2003], the input values for hydrogeological parameters between measurement locations can influence the flow field and, consequently, the concentration values calculated by the model. Since we lack the full knowledge of the subsurface structure, we must account its uncertainty to fill the spatial gap not covered by measurements [*Beckie*, 1996; *Rubin*, 2003].

Physiological parameters are needed in order to link contaminant concentration to human health risk. Uncertainty within this component comes from dose-response studies [*McKone and Bogen*, 1991; *Chiu et al.*, 2007]. The dose-response relationship is often obtained by performing laboratory experiments on animals and later extrapolating the results to humans. Thus, because of this extrapolation, at low doses these dose-response models are uncertain. Besides the physiological component, human behavioral characteristics, such as ingestion rate of tap water, also add uncertainty and variability in the risk related parameters [*Burmaster and Wilson*, 1996; *Maxwell et al.*, 1998; *Daniels et al.*, 2000].

Understanding the impact from each of these factors in human health risk provides a rational guidance towards answering questions such as: What is the expected risk uncertainty reduction if additional measurements of hydraulic conductivity are sampled? Given the uncertainty present in physiology, when is a detailed site characterization campaign necessary?

Several studies have investigated risk due to groundwater contamination in a probabilistic framework. For example, risk-cost-benefit analysis can be found in *Massman and Freeze* [1987], *Freeze et al.* [1990] and *James and Gorelick* [1994]. They studied the trade-offs between financial costs and risk. Uncertainty is accounted within the hydrogeological parameters within a Bayesian

framework. In these previous studies, costs are associated with probability of failure (contamination above some regulatory threshold value occurring) and the worth of data was addressed. A few other articles investigated the dependence of risk to hydrogeological and physiological parameters under different types of contaminants (i.e. radionuclide or organic) [*Bogen and Spear*, 1987; *Andricevic et al.*, 1994; *Andricevic and Cvetkovic*, 1996; *Maxwell et al.*, 1998; *Maxwell and Kastenberg*, 1999]. *Maxwell et al.* [1999] addressed how increased sampling of hydraulic conductivities affects reduction of uncertainty in human health risk. *Benekos et al.* [2007] extended the studies performed by *Maxwell and Kastenberg* [1999] for multi-species transport.

To investigate the relative impact of uncertainty reduction in the hydrogeological component and in the physiological component on the final risk estimate, *de Barros and Rubin* [2008] developed a metric, based on the concept of information entropy that allows one to quantify the relative impact of information gathered on human health risk (see details in Chapter 2). This metric is used within a graphical tool that compares alternative strategies for risk uncertainty reduction.

However, the role of flow and transport scales in determining characterization needs in a risk-driven approach has not received much attention. There is still need for further investigation when counter-balancing the effects of hydrogeological site characterization with physiological uncertainty as a function of flow and transport scales. Hydraulic properties can vary on different scales and the value of hydrogeological information is dependent on these physical scales. Physical scales include the characteristic lengths that characterize subsurface heterogeneity, flow and transport. Such scales, as shown in Figure 3.1, are source size relative to the correlation length of aquifer heterogeneity, size and configuration of the exposure endpoint (screened well or control plane), pore-scale and capture zones induced by the action of pumping.

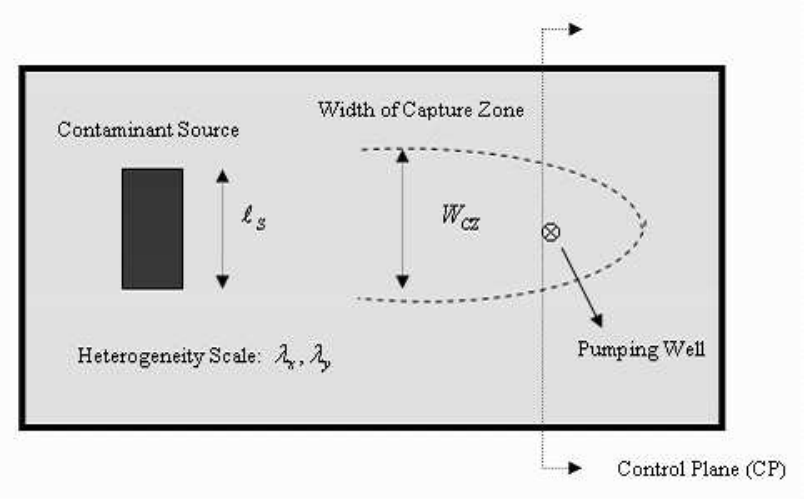


Figure 3.1: Illustration of the various length scales that define the flow, transport and consequently risk. Width of the contaminant source (ℓ_S), capture zone width (W_{cz}) and the representative geostatistical correlation lengths (λ_x,λ_y).

Furthermore, little attention has been given to the implication of different risk models in defining characterization needs and this issue will also be addressed. In this chapter, we employ the concepts presented in *de Barros and Rubin* [2008] to investigate the significance of various length scales that define the risk problem and their impact on hydrogeological site characterization. We extend the ideas from *de Barros and Rubin* [2008] (also presented in Chapter 2) to introduce the concept of comparative information yield curves in order to quantify the relative impact of uncertainty reduction of flow and health parameters in risk. The theoretical aspects of this concept are presented along with its implications on site characterization applications.

This chapter addresses the following fundamental question: Are there flow and transport characteristics in which uncertainty reduction in human health risk will benefit more from uncertainty reduction from human physiology or hydrogeology? We wish to investigate the role of these physical scales (e.g. plume-size, correlation lengths, etc) in determining characterization needs. The above question is relevant since assessing the value of data acquisition is an issue of concern in

real life applications. Questions concerning the expansion of existing, and sometimes even substantial, measurement networks, or issues regarding selecting between alternative targets for additional investment are of primary concern. Such efforts may not always be justified, because they can potentially yield only marginal improvement in the predictive capability. Due to an ever increasing demand on site characterization, many sampling techniques are available that vary in resolutions and offer direct or indirect information on the parameters relevant for modeling. Thus having a rational guide to manage all these alternatives becomes relevant since we live in a resource constrained world.

3.2 Mathematical Statement of the Problem

Given the uncertainty present in all components of human health risk assessment, it is rational to use a probabilistic framework to quantify risk due to groundwater contamination. Our objective is to obtain the ensemble distributions of human health risk for the exposed population. As in the previous chapter, we will consider r to represent the increased lifetime cancer risk. This is not a limitation and non-cancer risks can also be used within the framework. Here, $F_R(r)$ denotes the corresponding risk cumulative distribution function (CDF). $F_R(r)$ is evaluated for a given vector of hydrogeological parameters, $\boldsymbol{\theta}_H$, field site measurements, $\{m\}$, and for a given matrix containing the population's health-related parameters $\boldsymbol{\Theta}_P$.

The vector $\boldsymbol{\theta}_H$ contains the parameters that characterize the *Space Random Function* (SRF) of the hydrogeological variables [*Dagan*, 1984, 1987; *Rubin and Dagan*, 1992; *Rubin*, 2003] such as mean value and variance of the logconductivity, integral scales as well as other flow and transport related parameters such as porosity, source concentration, pumping rates and dispersion

coefficients. These parameters have a physical and chemical nature and can be deterministic or stochastic.

The exposed population with I individuals is characterized by the matrix $\boldsymbol{\Theta}_P = \{\boldsymbol{\theta}_{P,1}, \boldsymbol{\theta}_{P,2}, \boldsymbol{\theta}_{P,3}, ..., \boldsymbol{\theta}_{P,I}\}$ where $\boldsymbol{\theta}_{P,i}$ is the vector of behavioral and physiological characteristics of the i^{th} individual. Each $\boldsymbol{\theta}_{P,i}$, where i = 1,..,I, varies from individual to individual. The typical parameters present in $\boldsymbol{\theta}_{P,i}$ are, for example, the ingestion rate per body weight, exposure duration and cancer potency factors as well as their statistical moments if uncertainty exists. Statistical distributions for these parameters for different pathways are found in the literature [*Maxwell et al.*, 1998; *Binkowitz and Wartenberg*, 2001; *Portier et al.*, 2007]. The conditional risk CDF for the i^{th} individual of the exposed population is given as follows:

$$F_R(r|\boldsymbol{\theta}_H, \boldsymbol{\theta}_{P,i}, \{m\}) = \text{Prob}[R < r] \tag{3.1}$$

With equation (3.1), risk estimates can be obtained given an appropriate risk model. Most important, for a given regulatory acceptable risk value, for example $r = 10^{-6}$, equation (3.1) provides the probability of risk reliability or exceedance. This may be accomplished by calculating the complementary cumulative distribution function, CCDF, Prob$[R > r]$=1 $- F_R$.

3.3 The Use of Entropy to Quantify the Impact of Information on Risk

In the recent work of *de Barros and Rubin* [2008], information entropy was used to develop a metric that relates the amount of information in hydrogeology to the amount of information in physiology. This metric, denoted in the present work by α , was used to investigate uncertainty trade-offs between hydrogeological parameters (such as hydraulic conductivity K) and physiolog-

ical parameters (cancer potency factor). Before explaining the form of and how it functions, some definitions are required. We first introduce the concept of information yield curves. Afterwards, we extend the theoretical aspects towards applications to site characterization.

3.3.1 The Concept of Information Yield Curves

Following the work of *de Barros and Rubin* [2008], let us define E_H as the information entropy for hydrogeological parameters (including transport variables such as chemical parameters) and E_P as the entropy for physiological and behavioral parameters. The entropies are defined as [*Christakos*, 1992]:

$$\begin{aligned} E_H &= -\int_{-\infty}^{\infty} f_H(\boldsymbol{\theta}_H|I_H,\{m_a\})\ln[f_H(\boldsymbol{\theta}_H|I_H,\{m_a\})]d\boldsymbol{\theta}_H \\ E_P &= -\int_{-\infty}^{\infty} f_P(\boldsymbol{\theta}_P|I_P,\{s_a\})\ln[f_P(\boldsymbol{\theta}_P|I_P,\{s_a\})]d\boldsymbol{\theta}_P \end{aligned} \tag{3.2}$$

where f_H and f_P are the continuous probability density functions (PDF) for the vector of hydrogeological parameters $\boldsymbol{\theta}_H$ and for the health-related parameters $\boldsymbol{\theta}_P$ respectively. The integration in equation (3.2) is performed over the entire parameter space. For the sake of notation, we have omitted the subscript i from $\boldsymbol{\theta}_P$ as defined in the previous section. Equation (3.2) represents the total amount of information from each component at an initial stage of knowledge. These entropies can be evaluated with hydrogeological prior knowledge I_H, with a small amount of available hydraulic data $\{m_a\}$, physiological prior information I_P and finally, available health-related sample data $\{s_a\}$. From the distributions necessary to estimate E_H and E_P we can evaluate a corresponding F_R, defined by equation (3.1), and consequently its statistical moments. As more information becomes available, either from flow or health physics, E_H and E_P would decrease since the uncer-

tainty in both f_H and f_P is reducible with additional data collection.

Being able to estimate the values of E_H and E_P with no *a priori* information allows one to investigate relative value of information in human health risk. This is necessary since decision-makers need to decide where to invest resources towards risk uncertainty reduction. At this early stage of the risk analysis, only a small amount of information is available through prior knowledge or initial data. In order to decide whether or not more data is needed, one must evaluate its impact in the human health risk distribution. We now denote the unknown (*to be sampled*) hydrogeological measurements by $\{m_{na}\}$ and the unknown health-related by $\{s_{na}\}$. The following equations are the entropies averaged over all possible measurement values:

$$\begin{aligned} E_{H,O} &= \left\langle -\int_{-\infty}^{\infty} \hat{f}_H(\boldsymbol{\theta}_H|I_H,\{m_a\},\{m_{na}\}) \ln[\hat{f}_H(\boldsymbol{\theta}_H|I_H,\{m_a\},\{m_{na}\})] d\boldsymbol{\theta}_H \right\rangle; \\ E_{P,O} &= \left\langle -\int_{-\infty}^{\infty} \hat{f}_P(\boldsymbol{\theta}_P|I_P,\{s_a\},\{s_{na}\}) \ln[\hat{f}_P(\boldsymbol{\theta}_P|I_P,\{s_a\},\{s_{na}\})] d\boldsymbol{\theta}_P \right\rangle, \end{aligned} \tag{3.3}$$

with $E_{H,O}$ and $E_{P,O}$ being the expected entropy values over all possible measurements values that $\{m_{na}\}$ and $\{s_{na}\}$ can take. They are evaluated with the inferred PDF $\hat{f}_H$ and $\hat{f}_P$ such that $E_{H,O} \leq E_H$ and $E_{P,O} \leq E_P$, see equation (3.2). A general numerical procedure that can be used to obtain the entropies in equation (3.3) is as follows:

1. Generate a possible realization of N_o measurements for $\{m_{na}\}$ and S_o measurements for $\{s_{na}\}$ from prior knowledge. This requires the assumption that the models from which the measurements are generated are known. For example, a Gaussian or Exponential geostatistical model and a dose-response model;

2. Using the data drawn from this realization, the parameter's PDF, $\hat{f}_H$ and $\hat{f}_P$, are inferred. These parameters can be the mean or variance of the logconductivity data and integral scales

or cancer potency factor. As the number of measurements increases, these PDF become more informative. Details concerning the PDF estimation procedure is given in Appendix B;

3. With $\hat{f}_H$ and $\hat{f}_P$, the conditional entropies $E_{H,1}$ and $E_{P,1}$ can be calculated. Here, the subscript 1 corresponds to the first realization of the data sets $\{m_{na}\}$ and $\{s_{na}\}$. These entropies are conditional on the generated data and a known model;

4. Repeat steps 1-3 for several realizations of the data $\{m_{na}\}$ and $\{s_{na}\}$ such that two vectors with elements $E_{H,j}$ and $E_{P,j}$ are obtained. Here the subscript j=1, ...,J_{MAX} corresponds to the realizations. J_{MAX} is the maximum number to realizations;

5. With the entropies $E_{H,j}$ and $E_{P,j}$, given j=1, ...,J_{MAX}, the values for the ensembles averages, $E_{H,O}$ and $E_{P,O}$, are obtained;

6. Repeat steps 1-5 to evaluate the impact of an additional amount of data ($N \geq N_o$) in $\{m_{na}\}$ and $\{s_{na}\}$ ($S \geq S_o$).

The assumption in this outlined procedure is that some information about the site needs to be known. This includes the prior parameter PDF and the use of expert opinions or information borrowed from geologically similar formations (see Appendix B). If model uncertainty exists (for example: the geostatistical model of the underlying geological formation or the shape of dose response model), the current framework can incorporate Bayesian Model Averaging [*Hoeting et al.*, 1999; *Neuman*, 2003]. This is done by assigning different weights to each entropy ensemble evaluated for a given model and then averaging them. Mathematically this is equivalent to $E_{H,O} = \sum \omega_i \times (E_{H,O}|M_i)$, where ω_i is the i^{th} weight for the corresponding i^{th} model denoted by M_i). The term $(E_{H,O}|M_i)$ is the ensemble averaged entropy given a geostatistical model. In many situations,

the conceptual model for flow and transport could change as more data is collected. For instance, extra hydrogeological data may give evidence to the presence of a leaking aquitard or strong vertical pressure gradients, which would cause revisions of the initial conceptual model. Even if there is still uncertainty within the conceptual model, this additional data helps in updating the weights, ω_i, in the Bayesian Model Averaging procedure. The current framework allows for this model updating process as more data is collected and obtain new predictions.

For increasing number, N, of measurements in both $\{m_{na}\}$ and $\{s_{na}\}$, the average entropy estimates decrease as shown in Figure 3.2. The vertical axis represents the difference between the entropy evaluated with increasing N measurements and the initial entropy calculated with N_o measurements (with $N \geq N_o$).

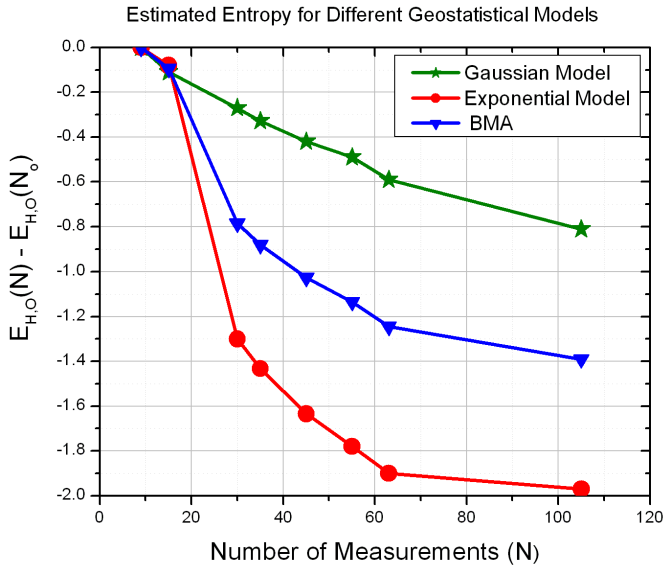


Figure 3.2: Entropy averaged over all possible measurement values generated by a geostatistical model.$E_{H,O}(N)$ and $E_{H,O}(N_o)$ are the entropies evaluated with N and N_o measurements respectively with $N \geq N_o$. If the model of the underlying formation is unknown, the methodology can account for the *Bayesian Model Averaging* (BMA). Here M_1 and M_2 corresponds the Gaussian and Exponential model respectively with $\omega_1 = \omega_2 = 0.5$.

This plot was obtained for the hydrogeological parameters by making use of steps given above together with Appendix B. A similar plot can be done with $E_{P,O}$ by averaging over all possible, non-available, physiological and behavioral data. Figure 3.2 also shows how different geostatistical models, Exponential and Gaussian, can lead to different entropy estimates. It also illustrates the Bayesian Model Averaging result if the geostatistical model is uncertain. Equal weights were assigned to each model for this demonstration. Now that we have presented the necessary definitions, we can write the following entropy differences for both hydrogeological and physiological parameters:

$$\Delta E_H = E_H - E_{H,O}$$
$$\Delta E_P = E_H - E_{P,O}. \tag{3.4}$$

Equation (3.4) define the differences between the expected entropies, $E_{H,O}$ and $E_{P,O}$, given in equation (3.3) and the current entropy stages denoted by E_H and E_P. By reducing uncertainty from both physiology and hydrogeology, ΔE_H and ΔE_P tend to values closer to zero. The metric that relates uncertainties from each risk component is given below (see Chapter 2):

$$\alpha = \frac{10^{\Delta E_H}}{10^{\Delta E_P}}, \tag{3.5}$$

where ΔE_H and ΔE_P are defined in equation (3.4). As explained in *de Barros and Rubin* [2008], loss of information in $\boldsymbol{\theta}_H$ means α increasing to values greater than one. This is obtained by increasing ΔE_H while keeping ΔE_P equal to zero. If uncertainty increases in $\boldsymbol{\theta}_P$, then α values are bounded between zero and one (keeping ΔE_H fixed and equal to zero). When α equals to one, we have ΔE_H = ΔE_P = 0. The point α =1 is considered the base case from which the relative contribution of information will be quantified. Figure 3.3 illustrates the α concept.

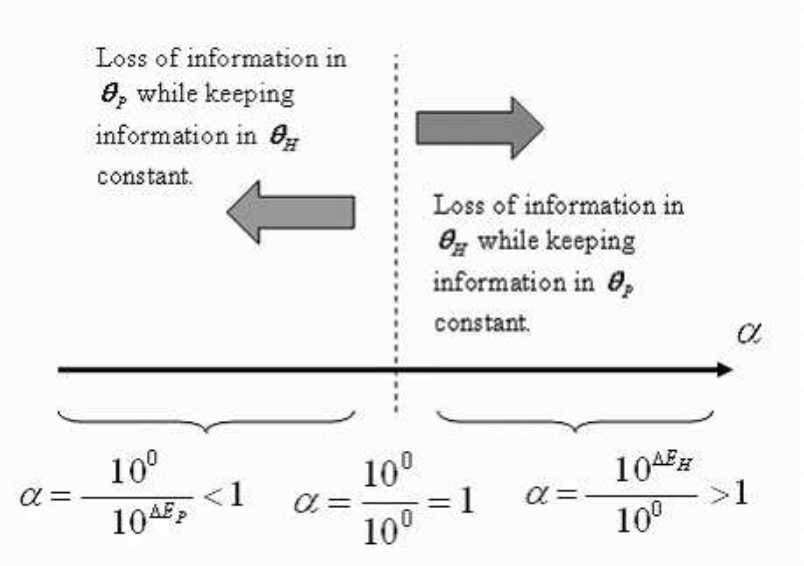


Figure 3.3: Graphical explanation of the α concept. At α equal to one, we have reached entropies $E_{P,O}$ and $E_{H,O}$. For each value of α, a corresponding risk variance or coefficient of variation is obtained. The plot of α versus the risk coefficient of variation is denoted here as the Comparative Information Yield Curves.

With equation (3.5) we can obtain a series of values to the right and left of α =1 and evaluate, through simulations, their corresponding uncertainty levels in human health risk. These corresponding uncertainty levels can be represented by risk variance, 95^{th} confidence intervals, or the risk coefficient of variation ($CV_R = \sigma_R/\mu_R$). For the current work, we will adopt the change of the coefficient of variation in the following way:

$$\Delta CV_R = \frac{CV_R - CV_R^o}{CV_R}, \tag{3.6}$$

where CV_R^o corresponds to risk evaluated with the entropies $E_{P,O}$ and $E_{H,O}$. We will obtain a series of these α versus ΔCV_R curves for several different physical scenarios to investigate uncertainty trade-offs. These graphs are denoted here as the *Comparative Information Yield Curves*. Summarizing, the value of α denotes a change in entropy values. It is a metric for comparing two stages of information. A financial cost value can be obtained by relating α to a given sampling

strategy. For this particular α value, a corresponding uncertainty reduction will occur. We represented this uncertainty reduction by ΔCV_R however, other representative measures besides ΔCV_R can also be evaluated from the Monte Carlo simulations (such as 95^{th} confidence intervals).

3.3.2 Application

Since $E_{P,O}$ and $E_{H,O}$ are speculative projections, in the sense that it needs to be defined in equation (3.4) and (3.5), one may want additional formulations of the approach described previously. An alternative application of the entropy concept for investigating uncertainty trade-offs in human health risk is obtained by changing the definition of $E_{P,O}$ and $E_{H,O}$, given in equation (3.3) and (3.4), such that we have $E_{P,O} \geq E_P$ and $E_{H,O} \geq E_H$. This means that the values for $E_{H,O}$ and $E_{P,O}$ correspond to the current state of information and are denoted as *base case* entropies. Note that this new inequality differs from the definition given in equation (3.4). This would bypass the need to calculate the entropy ensemble averages, $E_{P,O}$ and $E_{H,O}$, as given in equation (3.3) and (3.4). Based on a set of initial data or prior information, an estimate of $E_{P,O}$ and $E_{H,O}$ is obtained such that $\Delta E_P = \Delta E_H = 0$ corresponds to the initial uncertain case together with a corresponding coefficient of variation for risk. Now, with this alternative approach, $E_{H,O}$ and $E_{P,O}$ represents the available amount of information at the early stage of characterization. As more data are collected, new estimates of E_H and E_P are obtained and their values will be lower than the corresponding $E_{H,O}$ and $E_{P,O}$. As shown in the following paragraph and in Figure 3.4, a graphical approach could be used not only to investigate the value of information in uncertainty reduction in human health risk as more data is acquired but also to compare different sampling strategies by estimating their respective value of information using the 6-step procedure described earlier in Section 3.3.1.

Figure 3.4 shows the application of this alternative definition ($E_H \leq E_{H,O}$ and $E_P \leq$

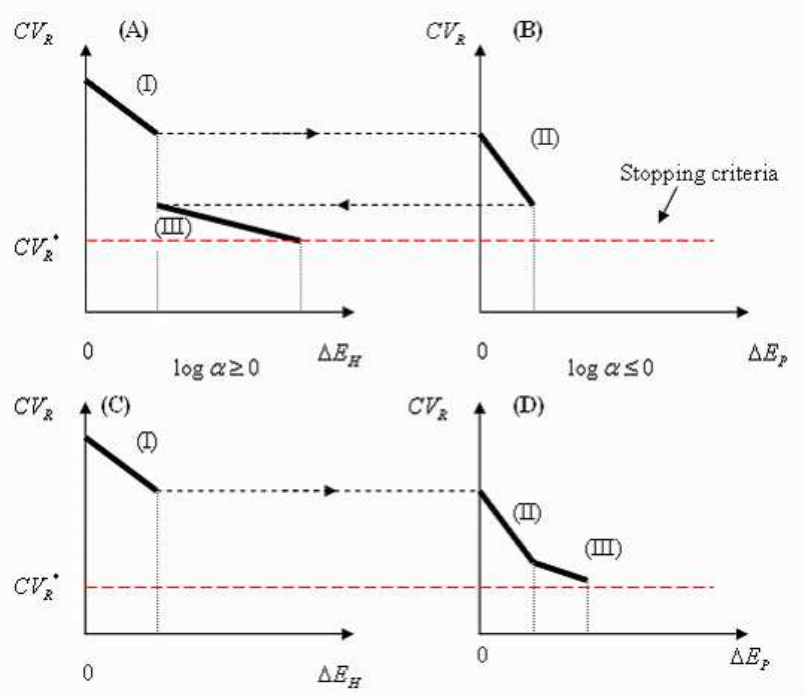


Figure 3.4: Illustration of the second alternative for the Comparative Information Yield Curves to investigate risk uncertainty reduction strategies between the hydrogeological and physiological component. Here, plots (A-B) show one sampling strategy while plots (C-D) illustrate another different sampling strategy. CV_R is the coefficient of variation of risk while CV_R^* is a stopping criteria associated with an environmental regulation.

$E_{P,O}$) and how one can reduce the uncertainty in risk by using different sampling strategies. Plots (A) and (B) in Figure 3.4 illustrates one sampling strategy: Reducing uncertainty from hydrogeology (segment I), then physiology (segment II) followed by hydrogeology again (segment III). A different strategy is shown in Figure 3.4, plots (C) and (D), by reducing uncertainty from hydrogeology (segment I), then physiology (segment II) and then physiology again (segment III). The sampling stops when a regulatory target is reached. Figure 3.4 gives a step-wise approach allowing one to direct efforts to obtain the best information yield.

As mentioned before, the advantage of this alternative is that it avoids the need to prespecify $E_{H,O}$ and $E_{P,O}$ values as defined in equations (3.3) and (3.4) thus allowing one to construct

the plots such as the one given in Figure 3.4. Also, the information yield curves based on this alternative definition offers a step-wise approach illustrated in Figure 4 that allows one to revise the conceptual model as more information becomes available. At each step, efforts can be allocated where the sampling strategy offers the best yield. The usefulness of this second approach will be illustrated in the end of this chapter. Both alternatives for the use of entropy in risk will be discussed. However, it is important to state that the second alternative allows one to select where to invest resources in a more practical manner.

3.4 Illustration Case

Consider a bounded 2D flow in an aquifer with spatially variable and isotropic hydraulic conductivity $K(\mathbf{x})$ and $Y = \ln K$. Due to incomplete information of the system, K is characterized by its *Space Random Function* (SRF) and is considered here as statistically stationary. Its covariance structure model is assumed to be exponential and isotropic with σ_Y^2 being the variance of Y and λ the correlation length of heterogeneity. A contaminant plume, considered here as a collection of particles, is released within a rectangular source domain with transversal length ℓ_S. Each particle represents a mass of contaminant and travels along a streamline of the flow field and are used to determine spatial contaminant distributions that may cause adverse health effects. We simulate the case of a hypothetical PCE contamination problem. The prescribed pressure head along the longitudinal direction are used as boundary conditions. Zero flux boundary conditions are assumed along the transversal direction. A drinking water well with pumping rate Q represents the environmentally sensitive location. The governing flow and transport equations are given in Appendix C.

Flow and transport is solved numerically using a Monte Carlo procedure. At each real-

ization, the flow and transport problem is solved for a specific image of the aquifer's properties, generated using the *Turning Bands Method* [*Tompson et al.*, 1989; *Rubin*, 2003]. Specific information concerning the numerical codes used in this work can be found in *Ashby and Falgout* [1996]; *Maxwell and Kastenberg* [1999]; *Jones and Woodward* [2001]; *Kollet and Maxwell* [2006]. Technical details concerning the numerical implementation are summarized in Appendix C. The framework presented can be used with analytical and numerical methods. Our choice for numerical implementation of flow and transport is for illustration and not to depend on simplifying assumptions. There exists a large amount of work published in the literature with analytical solutions that could be used to build the *Comparative Information Yield Curves* and many are summarized in *Rubin* [2003]. These same analytical solutions were also used to investigate human health risk [*Andricevic et al.*, 1994; *Andricevic and Cvetkovic*, 1996] and served as the basis of the work of *de Barros and Rubin* [2008] where the *Comparative Information Yield Curves* were used to evaluate uncertainty trade-offs. In the following, the exposure pathways considered in this work as well as the input data used in the simulations are described.

3.4.1 Exposure Pathways and Risk Formulation

We consider risk due to groundwater ingestion and inhalation for illustration of the methodology. These two pathways were shown to have a stronger impact in human health risk [*Maxwell et al.*, 1998]. Due to the nature and complexity of cancer mechanisms, cancer risk models are generally derived from dose-response curves. These curves are based on toxicological studies and are determined experimentally by observing adverse effects in animals for increasing applied doses (or concentrations) [*Fjeld et al.*, 2007]. The dose-response curve results are then extrapolated to humans. A common challenge is determining what shape the dose-response relationship in the ex-

trapolated zone (where uncertainty is highest), also known as the low-dose zone, should take. For low-doses, risk models can assume both linear and non-linear forms [*USEPA*, 2005; *Fjeld et al.*, 2007]. Figure 3.5 shows linear and non-linear models for the dose-response curves. In the present

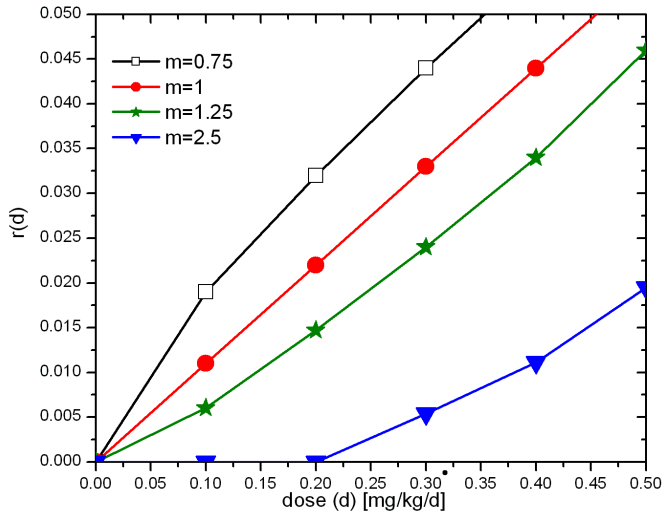


Figure 3.5: Example of dose-response relationship at the as a function of m shown in equation (7). The solid-dotted curve represents the linear model used by the [*USEPA*, 1989] with m=1

formulation we will treat the risk model in the following functional form:

$$r = CPF_G \times [LADD_G]^m + CPF_H \times [LADD_H]^m, \tag{3.7}$$

where CPF_G and CPF_H are the cancer potency factors for the ingestion and inhalation pathway respectively. These are also known as the cancer slope factors. The parameter m determines the non-linearity of the model. The value for m comes from fitting the model to toxicological data available from dose-response experiments. $LADD_G$ and $LADD_H$ are the average daily doses for tap water ingestion and inhalation during shower. The average daily dose is a function of the

concentration (C) and behavioral and exposure parameters.

$$\begin{aligned} LADD_G &= C \times \left(\frac{IR}{BW}\right) \times \frac{EF \times ED}{AT} \\ LADD_H &= AC_S \times ET_S \times \left(\frac{HR}{BW}\right) \times \frac{EF \times ED}{AT}. \end{aligned} \tag{3.8}$$

These behavioral and exposure parameters are the well-known EPA risk variables such as ingestion rate per body weight (IR/BW), exposure duration (ED), average lifetime (AT) and exposure frequency (EF), inhalation rate per body weight (HR/BW) and shower exposure time (ET_S) [*USEPA*, 1989, 2001]. The indoor air concentration is denoted by $AC_S = C(W_S \times TE_S)/\ VR_S$ with W_S being the tap water use rate, TE_S the transfer efficiency from tap water to air and VR_S is the air exchange rate [*USEPA*, 1989, 2001; *Maxwell et al.*, 1998]. For our work, we will use the peak concentrations to evaluate risk. Other works studied the effects of averaging concentration over the exposure duration [*Maxwell and Kastenberg*, 1999; *Hassan et al.*, 2001; *Maxwell et al.*, 2008] and the implications of using average versus peak concentration in hydrogeological site characterization [*de Barros and Rubin*, 2008]. To obtain the classic EPA linear low-dose model, we set m = 1 [*USEPA*, 1989, 2001].

3.4.2 Input Data used in the Case Study

Table 3.1 summarizes the deterministic data used for input in the simulations. The domain with longitudinal dimension L and width W (size: $50\lambda \times 32\lambda$) is discretized into a regular rectangular grid. Each grid block has dimensions Δx_1=Δx_2= $\lambda/5$ [*Rubin et al.*, 1999]. As mentioned previously, flow and transport are solved within the Monte Carlo approach and 300 realizations were performed.

To answer the research questions posed in the introduction, we select an aquifer with

Flow, transport and risk parameters			
Q	5 and 50 m^3/d	IR/BW	0.033 L/(d-kg)
$Pe = \lambda/\alpha_L$	100 and ∞	AT	70 y
λ	100 m	HR/BW	0.39 m^3/(d-kg)
n_e	0.3 [-]	W_S	480 1/h
R_f	1 [-]	EF	350 d/y
L	3000 m	ED	30 y
W	2500 m	TE_S	0.5 [-]
$\mathbf{x}_w$	(2500 m, 1000 m)	VR_S	12 mg/m^3
$\zeta = \ell_S/\lambda$	0.5 and 6 [-]	ET_S	0.13 h/d

Table 3.1: Data used in flow, transport and health risk models. Behavioral parameters are representative of the 50^{th} fractile of variability. Here, Q is the pumping rate, Pe is the Peclet number, λ is the heterogeneity correlation length, n_e is the effective porosity, R_f is the retardation factor, L is the longitudinal distance, W is the width, $\mathbf{x}_w$ is the location of the pumping well and ζ is the dimensionless source width. The other risk-related parameters are defined in Section 3.4.1.

parameters summarized in Table 3.1. This aquifer, denoted as the baseline, was selected from several realizations simulated with geometric mean K_G = 1 m/d and σ_Y^2 = 1 (see Table 2). From this baseline aquifer, we sampled values of hydraulic conductivity in fixed intervals of 8λ, 4λ and 2λ in a subdomain (18$\lambda \times$ 16λ) horizontally centered with the contaminant source and the environmentally sensitive target. We denote by $\{m_1\}$ the measurement density associated with the sampling interval 8λ, $\{m_2\}$ with 4λ and finally $\{m_3\}$ with 2λ.

For the present investigations, we assume that K_G and σ_Y^2 are uncertain parameters and its statistical distributions can be inferred as shown in Appendix B. Both K_G and σ_Y^2 vary between conditional simulations according to the three mentioned sampling densities shown in Table 3.2.

Hence, f_H in equation (3.2) corresponds to σ_Y^2 and K_G. If a distribution is assumed, say lognormal, the statistical moments of σ_Y^2 and K_G can be estimated by using the *Maximum Likelihood Function* [*Rubin*, 2003] (see Appendix D). Table 3.2 summarizes the estimated parameters from the sampled data set used in flow simulation.

Geostatistical Parameters Conditional on Hydraulic Data					
Sampling Strategy	$K_G[m/d]$	σ_Y^2	N^*	Var$[\sigma_Y^2]$	Var$[K_G]$
"Base Aquifer"	1.0	1.0	NA	NA	NA
$\{m_1\}$: 8λ	1.5	1.2	9	0.295	1.06
$\{m_2\}$: 4λ	0.9	0.6	25	0.035	0.16
$\{m_3\}$: 2λ	0.76	0.71	81	0.012	0.02

Table 3.2: Hydrogeological data used in the conditional simulations. Here N^* denotes the number of measurements sampled and NA means *Non-Applicable*.

From the physiological side, we assume that cancer potency factors are the uncertain parameter and uniformly distributed [*McKone and Bogen*, 1991]. Thus, f_P in equation (3.2), represents the uniform PDF for CPF_G and CPF_H. Table 3.3 summarizes the upper and lower bounds used in the following simulations. The coefficient of variation (CV) is also included in Table 3.3. We evaluate risk for different levels of parametric uncertainty in CPF_G and CPF_H. Due to the lack of data, we assume, without loss of generality, that the hydrogeological and physiological parameters are independent. This assumption is not a limitation in this work. If correlations between both components are known (for example, concentration data and the cancer potency factors), then joint entropies can be evaluated with the corresponding joint PDF between hydrogeological and phys-

iological parameters [*Christakos*, 1992]. For the current work, $E_{H,O}$ is evaluated with estimated uncertain parameters from denser measurement grid $\{m_3\}$ given in Table 3.2 and $E_{P,O}$ is evaluated with the statistical distributions in Case 4 from Table 3.3.

CPF_G				CPF_H			
CASE	Minimum	Maximum	CV	CASE	Minimum	Maximum	CV
1	0.001	0.025	0.53	1	0.0012	0.002	0.14
2	0.005	0.025	0.38	2	0.0015	0.002	0.08
3	0.01	0.025	0.24	3	0.0017	0.002	0.05
4	0.015	0.025	0.14	4	0.0017	0.0019	0.03

Table 3.3: Uniform distribution parameters for CPF_G and CPF_H along with the coefficient of variation (CV). Units of [(kg-d)/mg]^m , see equation (3.7).

3.5 Results and Discussion

In this section, results are presented based on the data set given previously. We first address the interplay between plume-scale, capture zones and pore-scale dispersion. The differences between using a screened well versus a control plane to evaluate the concentration in defining characterization needs within a risk driven approach is also addressed. Finally, we illustrate the how the value of information depends on the risk model used (i.e. linear versus non-linear model). Discussion and analysis are based on the concept of Comparative Information Yield Curves described in Section 3.3.

3.5.1 On Plume-Scale, Capture Zones and Pore-Scale

Here, we investigate the dependence of risk uncertainty reduction on plume-size by making use of the entropy concept defined in Section 3.3. Our goal is to evaluate the risk CDF conditioned on the contaminant source size and measurements, $F_R(r|\zeta, \{m_i\})$ where ζ is the ratio between the source width (ℓ_S) and the heterogeneity correlation length (λ) (see Figure 3.1). Results are shown for a small (ζ = 0.5) and large (ζ = 6.0) contaminant source given measurements densities $\{m_1\}$, $\{m_2\}$ and $\{m_3\}$. The *Comparative Information Yield Curves* in Figure 3.6 shows that the effect of conditioning in reduction of risk uncertainty is much more beneficial for small source when conditioning on hydrogeological data. However, gaining information on the physiology side has much more effect in risk uncertainty reduction when the source is large (ζ = 6.0).

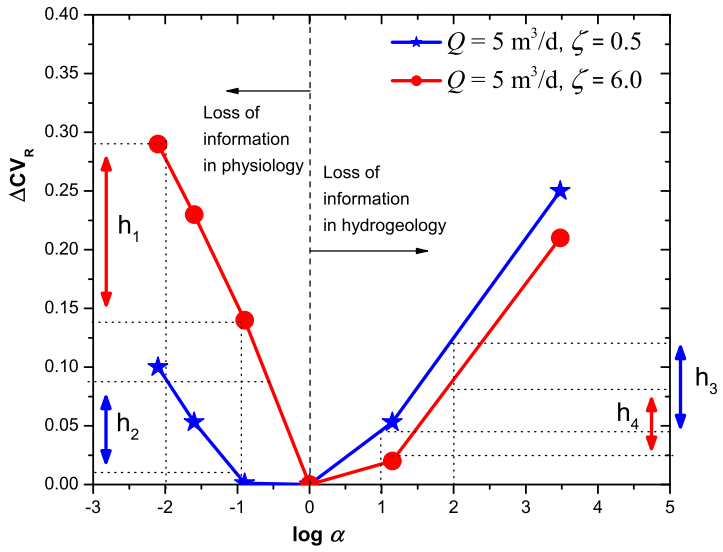


Figure 3.6: Illustration of the Comparative Information Yield Curves concept and the relative contribution of information for Q = 5 m^3/d, $Pe \to \infty$ given source sizes ζ = 0.5 and ζ = 6.0. Here $h_1 > h_2$ and $h_3 > h_4$ for a fixed change in $\log \alpha$. Risk evaluated with a linear model provided in equation (3.7) with m = 1.

For a given change in α at points $\alpha > 1$, we observe that the corresponding change in ΔCV_R is greater for the smaller plume case (ζ = 0.5) than for the larger plume (ζ = 6.0). This means that hydrogeological data acquisition has a stronger impact on risk uncertainty for smaller plumes by comparing h_3 and h_4 shown in Figure 3.6 ($h_3 > h_4$). One can also fix a change in ΔCV_R and compare the slopes of the curves and the corresponding changes in log to the left and right of α = 1. A similar effect was observed in *Maxwell et al.* [1999] by comparing conditional risk CDF for a 3D flow and transport test case. We will explore in more detail the physical mechanisms behind this result.

This effect can be explained as follows: As the scale of the solute body increases, the plume approaches the ergodic state. This means that the plume's centroid becomes less affected by small-scale fluctuations captured by hydraulic conductivity measurements [*Rubin et al.*, 1999; *Rubin*, 2003]. On the contrary, for small contaminant sources (ζ = 0.5), additional data contributes to reducing uncertainty about the location of the contaminant plume as well as the small-scale fluctuations of the streamlines. For example, a set of additional measurements may inform whether or not the contaminant plume will bypass the drinking water well.

The opposite effect is noted for $\alpha < 1$. Here we observe that for a given change in α, the larger ΔCV_R corresponds to the larger plume ($h_1 > h_2$). For larger plumes, uncertainty reduction from the physiological side causes a larger uncertainty reduction in the human health risk CDF when compared to the smaller plume case. This is quite intuitive since there is no or little uncertainty whether a larger plume will reach the environmental target. The only uncertainty is how severe the impact would be on the exposed population. This depends more on the population's physiological characteristics than on flow and transport processes.

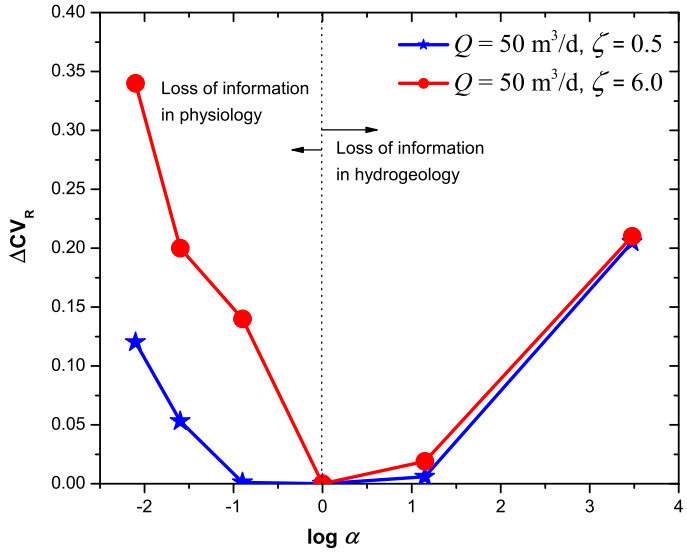


Figure 3.7: Illustration of the Comparative Information Yield Curves concept and the relative contribution of information for Q = 50 m^3/d, $Pe \rightarrow \infty$ given source sizes ζ = 0.5 and ζ = 6.0. Risk evaluated with a linear model provided in equation (3.7), m=1.

Now, we wish to extend this result to illustrate its dependence on the scale of the capture zone induced by aquifer pumping (see Figure 3.7). Juxtaposition of Figure 3.6 and Figure 3.7 shows that by increasing the pumping rate Q, the benefit of additional K sampling vanishes, regardless of source dimensions. The probability that the plume will reach the drinking well increases for larger Q, thus the additional data used to increase the accuracy of the plume's location has a smaller impact and becomes less relevant. On the other hand, improved physiological characterization is more beneficial for the bigger plume than the smaller one (similar to the conclusions drawn from Figure 3.6) because in the absence of uncertainty on whether the plume will be captured by the well, the only impact on risk uncertainty reduction is from the physiological side.

The effect of pore-scale dispersion on characterization needs is demonstrated in Figure

3.8. The Peclet number is defined as $Pe = \lambda/\alpha_L$ with α_L being the longitudinal dispersivity. It is varied in this case through different α_L values. Figure 3.8 shows that at small Peclet numbers, the benefits of K sampling diminish independently of the plume's dimension. Two cases for comparison are shown: An infinite Peclet scenario ($Pe \rightarrow \infty$) and a finite Peclet scenario (Pe = 100). Larger pore-scale dispersion smooths out details captured by hydrogeological site characterization for both large and small plumes. The role of a finite Peclet number in heterogeneous flows is to increase the rate of concentration variance destruction thus smoothing out the concentration field [*Fiorotto and Caroni*, 2002; *Rubin*, 2003; *Caroni and Fiorotto*, 2005]. This is observed in Figure 3.8 for $\alpha > 1$. By removing pore-scale dispersion, the effect of plume size starts to play a role in defining characterization efforts as shown in Figure 3.8 for points $\alpha > 1$. For large Pe, the plume centroid is influenced more by heterogeneity and hydraulic data contributes to risk uncertainty reduction. Furthermore, if the plume is small and transport is dominated by advective processes, pore-scale effects as well as macro-dispersion plays less of a role, thus increasing the importance of hydrogeological data acquisition. The information yield curve for this case (ζ = 0.5 and $Pe \rightarrow \infty$) is represented in Figure 3.8 for points $\alpha > 1$.

For $\alpha < 1$, we have the same results as shown in Figures 3.6 and 3.7. Note that, for finite Peclet, the curves corresponding to ζ = 0.5 and ζ = 6.0 are grouped closer compared to the previous figures and this is because dispersion tends to dilute the concentration field. On the contrary, by observing the slopes of the curves depicted in Figure 3.7, larger Peclet numbers imply larger variance in the concentration leading to higher probability for having larger concentration values. In the case of high Peclet, plume-scale makes a large difference in determining whether or not physiological uncertainty is important. For instance, Figure 3.8 shows that the physiological side becomes

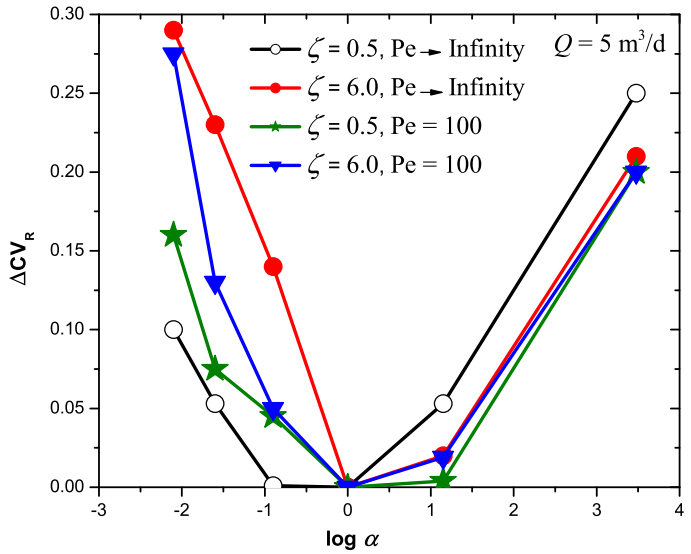


Figure 3.8: The influence of pore-scale dispersion in the analysis with $Pe = \lambda/\alpha_L$. Results obtained for Pe = 100 and given a fixed pumping rate Q = 5 m^3/d. The longitudinal dispersivity is α_L = 1 m^2 and the transversal dispersivity is α_T = 0.1 m^2. Risk evaluated with a linear model provided in equation (3.7) with m = 1.

more important to characterize for the ζ = 0.5 and $Pe \rightarrow \infty$ information yield curve. Summarizing, elements that reduce hydrogeological uncertainty about the environmental target being hit (larger plume, larger dispersivity, etc) increase the value of physiological characterization.

3.5.2 On the Significance of Concentration Averaging

Evaluation of human health risk may yield different results depending how the concentration is calculated or sampled. Some analysis makes use of the concentration at one or more fixed points in space represented by a drinking well [*Maxwell et al.*, 1998; *Maxwell and Kastenberg*, 1999; *Maxwell et al.*, 1999; *Benekos et al.*, 2007] whereas other studies have used the total solute mass flux (Q_s) over a control plane [*Andricevic et al.*, 1994; *Andricevic and Cvetkovic*, 1996,

1998]. Dividing the total solute discharge (Q_s) by the fluid volumetric discharge over the control plane (Q_f) yields the flux-averaged concentration [*Kreft and Zuber*, 1978]$C_f = Q_s/Q_f$. See also discussion in p.163 of *Rubin* [2003].

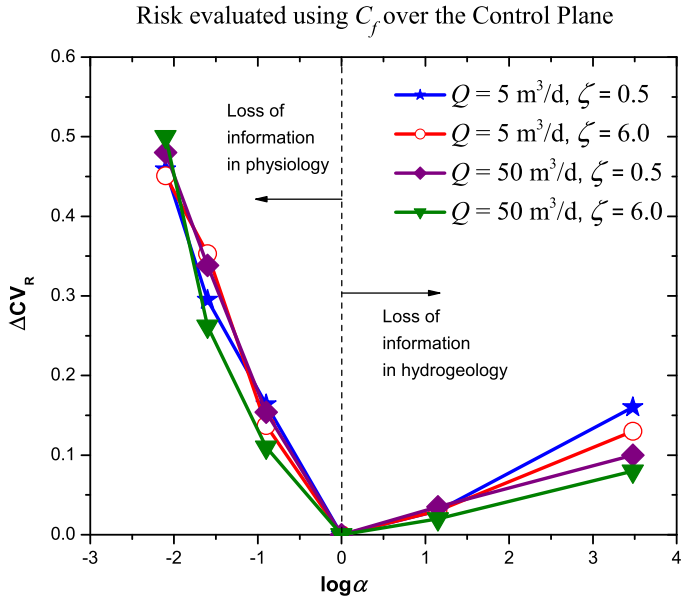


Figure 3.9: Measuring the relative contribution of information for pumping rates Q = 5 and 50 m^3/d and source sizes ζ = 0.5 and ζ = 6.0. Results evaluated using the flux-averaged concentration over the compliance plane. Risk evaluated with a linear model provided in equation (3.7) with m = 1.

The differences in the definition of the concentration are elucidated in the *Comparative Information Yield Curves* present in Figure 3.9. We obtained Figure 9 by making use of the flux-averaged concentration over the entire control plane (see Figure 3.1). When comparing with Figures 3.6 and 3.7, Figure 3.9 shows how the control plane approach dampens the effect of the differences in plume size and pumping. Note that for each pumping scenario, the curves for large and small plumes are closer together when compared to Figures 3.6 and 3.7. This is more evident to the curves to the left of α =1 where the relative gain of information in the physiological component of risk is

quantified. What this result illustrates is that parametric uncertainty reduction from the physiological component is less dependent of the both plume's dimension and the scale of the capture zone when human health risk is evaluated with the solute discharge over a control plane. One possible explanation is that when evaluating C_f over the control plane, the total mass of the solute present in that particular slice of the domain is being captured independently of its spatial distribution. Even if the peak of Q_s (say above a certain regulatory threshold value) occurs along a streamline that bypasses the well, the presence of the chemical (and it peak value) will still be lumped into C_f since the averaging process is over the entire control plane. This averaging procedure over the control plane also leads to smaller differences observed in the curves to the left of α = 1 when compared to Figures 3.6-3.8 since the breakthrough curves for C_f are smoother. From a regulator's point of view, an erroneous interpretation of the concentration term in risk may lead to unnecessary clean up costs. For example, the control plane approach may account for the contaminant mass along a streamline that bypasses the drinking water well leading to remediation costs. Still, from the information yield curves present in Figure 9, it is possible to observe that human health risk uncertainty reduction benefits more from physiological characterization.

As for the effects of additional measurements of K, the extra dilution added by averaging Q_s by Q_f smoothes out local details captured by characterization. In other words, the control plane approach adds an *enhanced diffusive mechanism* that removes some of the conditional effect gathered through site characterization and may mislead decision-makers. However, the control plane approach can be very helpful if regulations are based on travel times as shown in *Andricevic et al.* [1994]; *Andricevic and Cvetkovic* [1996].

3.5.3 The Effect of Alternative Risk Models

In the present subsection, the sensitivity of human health risk towards different dose-response models (see equation 3.7 and Figure 3.5) is illustrated. As explained previously, the main uncertainty in risk models is at the low-dose (or low concentrations) [*USEPA*, 2005; *Chiu et al.*, 2007; *Fjeld et al.*, 2007]. For instance, PCE is known to cause cell leukemia and kidney tumors in rodents however the shape of its dose-response in humans is uncertain *USEPA* [1998]. Here, we wish to point out how different dose-response models can lead to different characterization needs.

For illustration purposes, in the next results we will use a linear model (m = 1 in equation 3.7) and a non-linear model (m = 2.5 in equation 3.7). The linear model assumes zero risk only at zero concentration and is normally considered conservative [*USEPA*, 1989, 2001, 2005] . However, in recent years, the scientific community as well as environmental regulations acknowledges that the use of a non-linear model maybe more adequate depending on the amount of available data used to construct the dose-response model [*USEPA*, 2005; *Chiu et al.*, 2007; *Fjeld et al.*, 2007]. The applicability of these non-linear models may be expanded to both cancer and non-cancer risks [*USEPA*, 2005]. Now we illustrate how different risk models would possibly manifest in Information Yield Curves.

In Figure 3.10, we compare different risk models, their sensitivity to hydrogeological data acquisition and consequently parametric uncertainty reduction. Figure 3.10 shows how hydrogeological sampling has a stronger implication in risk uncertainty reduction for the non-linear model than for the linear model. This result indicates that when using a non-linear model to evaluate risk, the data worth of sampling hydraulic conductivity increases towards risk uncertainty reduction. Thus characterizing the behavior of the flow field becomes more important for this class of models.

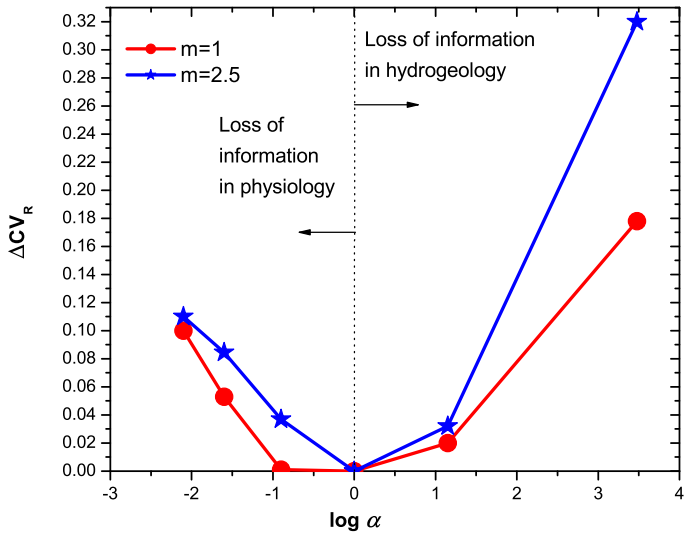


Figure 3.10: Sensitivity of human risk towards different dose-response models. Results evaluated with ζ = 0.5 and Q = 5 m^3/d.

Although this result is shown only for carcinogenic risk, it may also have implications for some non-carcinogenic compounds with threshold doses where an adverse effect is observed. In such cases, the worth of hydrogeological information might increase given that better understanding of the flow patterns lead to better estimation of the concentration (or dose) values since the knowledge of being above or below such threshold values becomes very important.

3.5.4 On the Definition of $E_{H,O}$ and $E_{P,O}$

As presented previously in Section 3.3, an alternative way to investigate uncertainty trade-offs is to change the definition of $E_{H,O}$ and $E_{P,O}$ such that we have $E_H \leq E_{H,O}$ and $E_H \leq E_{P,O}$. This implies that $E_{H,O}$ and $E_{P,O}$ corresponds to the most uncertain case (see Figure 3.4). These entropies are now evaluated with the most uncertain distributions, corresponding to the small

amount of information available a priori. In this new definition, $E_{H,O}$ and $E_{P,O}$ are now the starting points of uncertainty reduction. This avoids the need to pre-specify $E_{H,O}$ and $E_{P,O}$ values as defined in Section 3.3 thus allowing more flexibility.

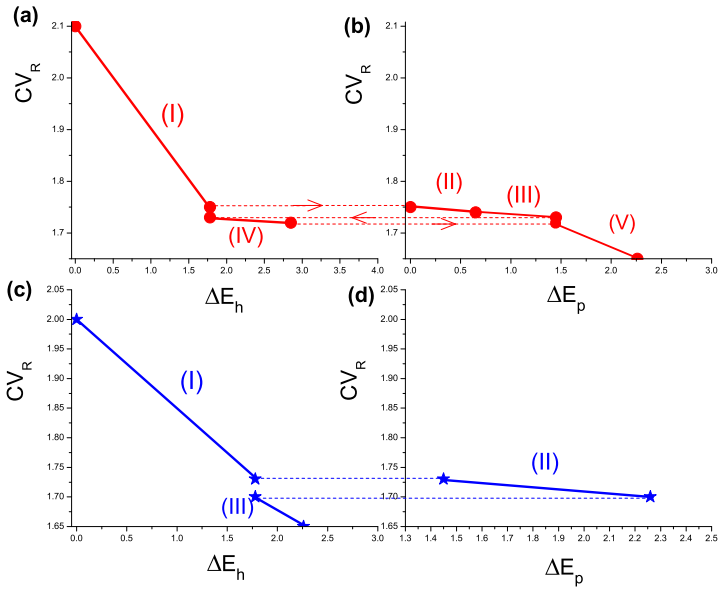


Figure 3.11: Sensitivity of human risk towards different dose-response models. Results evaluated with ζ = 0.5 and Q = 5 m^3/d.

Figure 3.11 depicts how decision makers could investigate risk uncertainty reduction strategies by plotting both the coefficient of variation of risk (CV_R) versus ΔE_H and ΔE_P (equivalent to $\log \alpha$). We have used the data given in Tables 3.1-3.3 to obtain plots similar to the diagram in Figure 3.4. As a starting point, risk is evaluated with $E_{H,O}$ and $E_{P,O}$ (initial information available, most uncertain case) as well as its corresponding $CV_R = CV_R^O$. By collecting additional data, one may perform conditional simulations from both the hydrogeological and physiological side (see Figure 3.11.a and 3.11.b) and evaluate a new CV_R. For instance, from the starting point CV_R^O, uncertainty reduction in risk can be done in a five-step procedure by collecting hydraulic

conductivity, K, data (Figure 3.11.a, segment I), then through physiology or behavioral parameters (Figure 3.11.b, segments II and III), then more K measurements (Figure 3.11.a segments IV) and finally health variables again (Figure 3.11.b, segment V). The stopping criterion is when CV_R meets the regulatory standards. Thus the necessity of additional sampling is risk-driven and can be decided upon based on an acceptable risk value (say the 95^{th} percentile confidence level or a coefficient of variation) that is in agreement with probabilistic risk assessment guidelines [*USEPA*, 1989]. However, this graphical approach can be useful to compare different characterization strategies by making use of the estimation procedure given in Section 3.3 to evaluate entropies for *a priori* unknown data.

Figures 3.11.a and 3.11.b showed a five-step procedure described in the previous paragraph. By summing up all the ΔE_H and ΔE_P needed to reduce CV_R (from 2.1 to 1.75) one may come with an estimate of the sampling efforts. Yet, a different strategy, three-step procedure, is given in Figure 3.11.c and 3.11.d which can yield a different summed entropy value when compared to the one given by the four-step procedure (see Figure 3.11.c and 3.11.d, segments I, II and III). By associating the risk uncertainty reduction with the corresponding total change in entropy (ΔE_H and ΔE_P) one may opt for the cheapest strategy to reach a compliance goal set up by environmental agencies. For example, the costs in hydrogeology could be associated with slug tests and sampling (laboratory experiments) while in physiology and other health-related parameters acquisition costs can be associated with number of animals used in toxicological studies or a more detailed survey of the behavioral characteristics of the exposed population.

Next, we show how sampling efforts can differ when using the same data acquisition strategy but different risk models. Figures 12.a and 12.b were evaluated using a linear risk model

while Figures 12.c and 12.d uses a non-linear model. One can see that the slopes of the curves in 12.a.b are different than 3.12.c.d for each corresponding segment. Also, Figure 3.12.a and 3.12.c depicts how the total change in ΔE_H , represented by summing δh_1 and δh_2 corresponds to two distinct changes in CV_R represented by Δ_H. The total change in ΔE_P (δp_1 + δp_2 + δp_3) and the associated total change in CV_R denoted by Δ_P, is highlighted in Figure 3.12.b and 3.12.d.

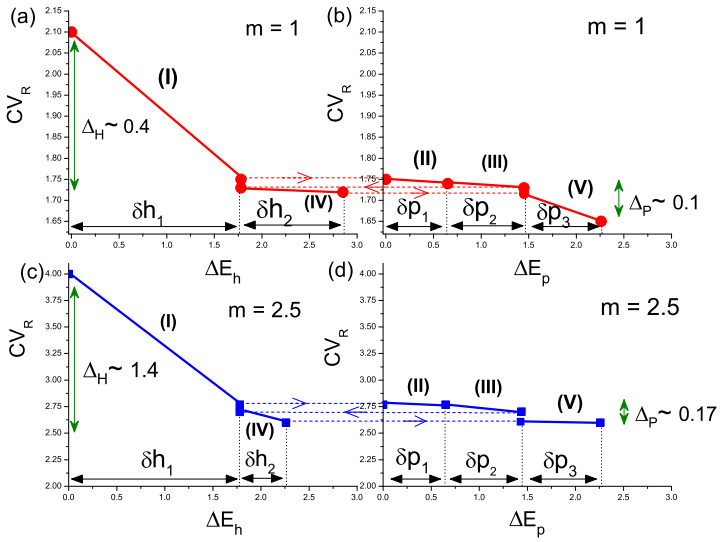


Figure 3.12: Sensitivity of human risk towards different dose-response models. Results evaluated with ζ = 0.5 and Q = 5 m^3/d.

In summary, the usefulness of the step-wise approach given in Figures 3.11 and 3.12 is that it allows one to see how different uncertainty reduction strategies could lead to different costs (associated with data) by avoiding the necessity of evaluating speculative values for both $E_{H,O}$ and $E_{P,O}$ required for equations 3.3 and 3.4 in Section 3.3.

3.6 Summary and Conclusions

In this chapter, we discussed the theoretical and practical aspects of the *Comparative Information Yield Curves* within a risk-driven context. The relevance of transport and flow scales in defining characterization needs in human health risk is addressed. Through numerical experimentation, conditions are identified where hydrogeological site characterization, through measurements of hydraulic conductivity, has a stronger impact in uncertainty reduction in risk as well as conditions in which physiological uncertainty reduction has a significant impact. In order to achieve this, we investigated the interplay between plume-dimension, capture zones induced by pumping action, Peclet number and sampling scales for different conditional simulations. We have quantified the relative gain of information through uncertainty reduction from both physiology and flow physics for a fixed, although not limited to, fractile of human variability. Results were analyzed for the low-dose risk curves. Based on the simulations results, physical configuration (2D groundwater flow and transport) and risk pathways, we highlight the following points:

1. The role of the plume's dimension proved important in defining characterization needs in the risk-driven context. Results show that uncertainty reduction in human health risk benefits more from hydrogeological site characterization if the contaminant source is small relative to the heterogeneity correlation length. The human health risk CDF is less sensitive to measurements of hydraulic conductivity if the contaminant source is large.

2. The value of information not only depends on plume's dimension but also on its interplay with the pumping rate related to the scale of capture zone. For high pumping rates, thus larger capture zones, the value of information from the hydrogeological component becomes less dependent of the plume's dimension. The opposite occurs as the pumping rate decreases

and the plume's dimension begins to gain a role in defining hydrogeological sampling needs.

3. Results indicate that uncertainty reduction in risk may benefit more from parametric uncertainty reduction from the physiological component as opposed to hydrogeological if the plume's dimension approaches ergodicity.

4. The significance of plume-dimension in defining hydrogeological characterization needs is also dependent on the phenomenon occurring at pore-scale. For high Peclet conditions, plume size relative to the heterogeneity scale has a role in defining characterization efforts. When pore-scale dispersion effects are increased (lower Peclet), the knowledge of whether the plume is large or small becomes less relevant in defining hydrogeological characterization strategies.

5. Similar conclusion was obtained when comparing concentration measured in a well versus the flux-averaged concentration at a control plane. The differences between concentration pumped by a well and concentration at a control plane is highlighted and can also lead to significant different characterization needs from both physiological and hydrogeological perspective.

6. We also showed how different risk models have different effects in risk uncertainty reduction and defining characterization needs.

For this work, we have made extensive use of information entropy to investigate uncertainty trade-offs in a graphical manner. We denote these entropy plots as *Information Yield Curves*. This is a useful concept since it allows one to easily view the relative contribution of information in risk from the physiological and the hydrogeological component. An important difference regarding

the use of these entropy plots as opposed to CDF is that one can assign changes in entropy for both physiology and hydrogeology for a fixed uncertainty reduction in human health risk. The challenge in this approach is to assign estimated financial values to these α values. This way, decision makers can verify which uncertainty reduction campaign is cheaper for a given uncertainty reduction in risk estimates. Translating values into financial terms allows one to cast the analysis in a cost-benefit framework as studied by *Massman and Freeze* [1987] and *Freeze et al.* [1990]. Section 3 provided a discussion of Information Yield Curves and how to make this concept practical in real site characterization problems. For our simulations, the mean value of risk did not vary significantly from each conditioning case and all were found to be within the same order of magnitude. However, if the mean value of risk varies significantly, other measures of uncertainty besides ΔCV_R might be more informative (for instance: relative entropy or 95th confidential interval). For our analysis, the uncertainty in the hydrogeological component was within SRF parameters. In this chapter we did not account for uncertainty in the chemical reactions, although this is not a limitation in the framework or the numerical implementation. These parameters can be accounted in $\boldsymbol{\theta}_H$. Another option would be to generalize the metric to other dimensions to account for specific subgroups of parameters. For example: a subgroup for chemical parameters, another for the SRF parameters and finally for source characteristics.

It is important to note that the framework and results presented here can be extended to different types of data used for conditioning. Also, other sources of uncertainty can be incorporated into the framework. We have used a two-dimensional model to answer the research questions addressed in the introduction of this chapter. Reproducing these numerical simulations and problem configuration in a three-dimensional physical model may enhance the results obtained. For an in-

stance, *Maxwell et al.* [1999] reported that finer sampling over the vertical direction is necessary and relevant to predict the expected plume path. Our approach can be extended to account for variability within the exposed population. Here we calculated the *Comparative Information Yield Curves* for a single fractile in population variability (see Section 3.2 for discussions on $\mathbf{\Theta}_P$ and $\boldsymbol{\theta}_{P,i}$), however one may obtain information yield curves for different fractiles. Given this, 3D surfaces of information yield could be evaluated. As for the pore-scale dispersion analysis, the slopes of the information yield curves are affected by both longitudinal and transversal Peclet numbers. For instance, increasing the transversal dispersion coefficient, more mass will be transferred between streamlines, thus smoothing the concentration field. This effect is reflected in the information yield curves. Nevertheless, one of the novelties of the present work is the illustration of the importance of considering flow and transport scales when defining characterization needs towards better resource allocation within a risk-driven approach. These results shows how any characterization effort should be task-oriented. Most importantly, the current chapter introduced the theoretical and practical aspects of the *Information Yield Curves* in human health risk assessment. This approach allows one to investigate uncertainty trade-offs from the health-related parameters and physical parameters.

Notation

Symbols and their respective units:

AC_s: Indoor air concentration [M/L^3]

AT: Average time [t]

b: Depth of the aquifer [L]

$LADD_G$: Average daily dose for groundwater intake [M/(M t)]

$LADD_H$: Average daily dose for inhalation [M/(M t)]

C, C_w: Resident concentration and well concentration [M/L^3]

C_f: Flux-averaged concentration [M/L^3]

CPF_G, CPf_H: Cancer potency factor $[(kg-d)/mg]^m$

C_Y, C_{YY}: Spatial covariance of the logconductivity

CV_R, CV_R^o: Coefficient of variation for risk [-]

$\mathbf{D}_d$: Dispersion tensor [L^2/t]

ED: Exposure duration [t]

EF: Exposure frequency [t/t]

ET_s: Shower exposure time [hr/d]

E_P, $E_{P,O}$: Joint entropy for physiological parameters

E_H, $E_{H,O}$: Entropy for hydrogeological parameters [-]

$f_R(r)$: Risk PDF [-]

$f_P(\boldsymbol{\theta}_P)$: PDF for risk-related parameter

$f_H(\boldsymbol{\theta}_H)$: PDF for hydrogeological parameters

$\tilde{f}$: Estimate of a PDF

f_{mo}: Metabolized fraction of contaminant ingested [-]

$F_R(r)$: Risk CDF [-]

$F_R^c(r)$: Risk CDF conditioned on measurements [-]

HR/BW: Inhalation rate per body weight [m^3/(d-kg)]

h: Hydraulic head [L]

h_1, h_2, h_3, h_4: Markings on Figure 3.6

I: Number of individuals in the exposed population [-]

IR/BW: Ingestion rate per body weight [L^3/(t M)] or [L/(d-kg)]

K_i: Hydraulic conductivity at a specific location $\mathbf{x}_i$ [L/t]

K_G: Geometric mean of the hydraulic conductivity [L/t]

$\mathbf{K}(\mathbf{x})$: Hydraulic conductivity at a generic location $\mathbf{x}$ [L/t]

ℓ_S: Dimension of the contaminant cloud in the x_2-direction [L]

m: Coefficient in the risk model.

$\{m_i\}$: Set of measurements

m_Y mean of the log conductivity [-]

N: Number of locations sampled and number of samples [-]

n_e: Porosity [-]

$\mathbf{P}_i$: Behavioral and exposure parameters for the i^{th} individual

$Q_s(\mathbf{x}, t)$: Total solute mass flux [M/t]

Q, Q_w: Pumping well [L^3/t] and volumetric fluid discharge [L^3/t]

R_f: Retardation coefficient [-]

r: Increased cancer risk or simply risk [-]

$\langle R \rangle$: Expected value of increased cancer risk [-]

$\langle R^2 \rangle$: Second moment of increased cancer risk [-]

U, $\langle V_1 \rangle$: Mean velocity in the x_1 direction [L/t]

TE_S: Transfer efficiency from tap water to air [-]

VR_S: Air exchange rate [mg/m^3]

$\mathbf{V}$: Velocity vector with components V_i for i = 1 and 2 [L/t]

Var$[\sigma_Y^2]$, Var$[K_G]$: Variance of σ_Y^2 and K_G

$\mathbf{x}$: Cartesian coordinate system [L]

$\mathbf{x}_w$: Location of the wth pumping well [L]

Y: Logarithm of the hydraulic conductivity

W_S: Tap water use rate [1/t] or [1/hr]

α: Ratio between entropies [-]

α_L, α_T: Dispersivities in the x_1 and x_2 direction [L^2]

δ: Dirac delta

δh_i, δp_i: Markings on Figure 3.12

ΔCV_R: Relative difference for the risk coefficient of variation [-]

ΔE_H, ΔE_P: Entropy difference [-]

λ: Correlation heterogeneity length of the aquifer [L]

σ_Y^2: variance of the logconductivity [-]

σ_R^2: Variance of increased cancer risk [-]

$\boldsymbol{\theta}_H$: Hydrogeological parameter vector

$\boldsymbol{\theta}_P$: Risk related parameter vector

ξ: Lag distance [-]

ζ: Dimensionless source dimension ℓ_s/λ

Chapter 4

Bayesian Geostatistical Design: Optimal Site Investigation When the Geostatistical Model is Uncertain

4.1 Introduction

Results from the previous chapters showed conditions when risk uncertainty reduction benefits more from hydrogeological data acquisition (for example, non-ergodic plumes). In this chapter we will focus on hydrogeological sampling. Scarcity of data and subsurface variability lead to the understanding of hydraulic conductivity as a *Space Random Function* (SRF) [*de Marsily*, 1986; *Dagan*, 1987; *Kitanidis*, 1997; *Rubin*, 2003]. This acknowledges the uncertainty in flow and transport models stemming from unresolved heterogeneity of soil parameters, patterns of flow, and their impact on contaminant transport [*Dagan*, 1984, 1987; *Rubin*, 2003]. Adopting the model-based

geostatistical approach [*Diggle and Ribeiro*, 2007], the SRF is defined by the global mean value, trend coefficients, and parameters for covariance models (e.g., the variance and integral scales of the logconductivity), often summarized under the term of structural parameters.

Incorporating hydrogeological flow and tracer data helps to reduce uncertainties, leading to smaller confidence bounds of model predictions, and supporting management decisions at a lower risk of liability. Two types of information are required: information on the spatial hydraulic conductivity field, and the spatial statistics of conductivity that allow to interpolate between unsampled positions. Given limited resources for hydrogeological characterization, this information need has to be satisfied in an efficient manner [*James and Gorelick*, 1994] via geostatistical optimal design. An extensive review is provided by *Herrera and Pinder* [2005], including the major contributions by Freeze and co-workers [*Freeze et al.*, 1990; *Massman and Freeze*, 1987]. The importance of setting priorities in allocating resources is also highlighted in the works of *Maxwell et al.* [1999]; *de Barros and Rubin* [2008] and *de Barros et al.* [2009]. These works showed the importance of task-oriented characterization.

Most of these studies presume prefect knowledge on the structural parameters. In realistic scenarios, however, information is too sparse to justify such strong *a priori* assumptions. Structural parameters tend to be poorly identifiable, especially from data sets limited in size and accuracy. Instead, there is a trend to perceive stochastic descriptions of structural parameters as much more adequate. *Pardo-Iguzquiza* [1999] illustrated the inadequacy of point estimates for the structural and trend parameters in synthetic case studies. Maximum Relative Entropy (MRE) techniques [*Woodbury and Ulrych*, 1993] allow transfer of general background knowledge from hydrogeological databases, from sites that are perceived to be geologically similar or from subjective expert opinion

[*Rubin*, 2003].

The principle of Bayesian geostatistics [*Kitanidis*, 1986] acknowledges that limited data do not support a unique geostatistical description. Instead, structural parameters are treated as yet another set of random variables on top of the hydraulic conductivity field. Further fundamental steps were provided by *Feinerman et al.* [1986] and *Rubin and Dagan* [1987, 1992]. MRE yields probabilistic distributions of structural parameters (no point estimates), and can be used as input to Bayesian geostatistics [*Woodbury and Ulrych*, 2000].

Uncertain structural parameters tend to increase the uncertainty of model predictions, such as contaminant levels or fluxes, because structural parameters have a substantial influence on macroscopic flow, plume dilution and dispersion [*Rubin*, 2003]. Conditional simulation (conditioned on direct local measurements of the space function) with uncertain structural parameters is provided by *Pardo-Iguzquiza and Chica-Olmo* [2008] and in *Rubin* [2003]. The application to geostatistical inverse modeling has been highlighted by *Woodbury and Ulrych* [2000] and by *Zhang and Rubin* [2009].

Because structural parameters are sufficiently known in only a few cases, inclusion of their uncertainly is all the more relevant in optimal design. Only small amount of data exists when setting out to plan site investigation via optimal design techniques. Although uncertainty of geostatistical models is largest prior to data collection, it is rarely recognized in optimal design studies.

The main focus of Chapter 4 is that uncertainty in the geostatistical model has to be accounted for in geostatistical optimal design. By optimal design, we mean an optimal sampling strategy that captures the geostatistical characteristics. Supported by the evidence and facts discussed above, it is reasonable that optimal design should fulfill the following four guidelines in the

context of uncertain geostatistical structure:

1. A common objective of optimal design is to minimize the uncertainty of predictions (such as contaminant levels or fluxes). Uncertainty of structural parameters contributes to the overall prediction uncertainty, so it must be assessed and accounted for.

2. Optimal design defines a rational way of data collection, and these data have a vast potential to better identify the geostatistical model and its structural parameters. This potential must be considered and utilized in defining and finding the optimal design.

3. Estimating structural parameters should be "*treated as a means to the primary end of spatial* [or hydrogeological] *prediction, rather than as an end in itself*", as suggested by *Diggle and Lophaven* [2006]. This asks for an optimal resource allocation between collecting spatial and structural information.

4. Optimal design patterns are sensitive to assumed values of structural parameters in geostatistical models. The resulting patterns, however, should be robust with respect to estimation error in structural parameters [e.g., *Christakos*, 1992, p. 438].

Until quite recently, no geostatistical optimal design study has addressed uncertainties resulting from unknown structural parameters, or investigated how their designs may aid in reducing that uncertainty. Only few optimal design studies in hydrogeology [*Criminisi et al.*, 1997] analyzed whether their resulting design patterns are robust when using inaccurate estimates of structural parameters.

The work on Bayesian design reported by *Müller* [2007, chapter 3] included uncertain trend parameters into geostatistical design, but left uncertain covariance functions untouched. Most

geostatistical design studies serve either the collection of spatial information or the identification of structural parameters. The former requires coverage of specific areas of the domain with samples, while the latter requires sampling certain lag distances. These objectives may seem contradictory to one another, but can be combined in multi-objective optimization [e.g., *Müller*, 2007, pp. 173].

Diggle and Lophaven [2006] introduced the concept of Bayesian Geostatistical Design, which accommodates for uncertainty in covariance parameters within the design procedure. These authors minimizes the spatial average of the kriging estimation variance and limited their study to direct measurements of the estimated quantity. The more recent work by *Marchant and Lark* [2007] may be seen as a direct extension, using a first-order approximation for the influence of structural parameters on the kriging variance for the sake of computational speed-up. Similar ideas, including the one by *Zimmerman* [2006], are summarized in *Müller* [2007, pp. 178].

This study may claim, given the information gathered in the literature review, to be the transfer and first-time application of Bayesian Geostatistical Design to geostatistical inverse problems. We extend the Bayesian Geostatistical Design framework to measurements of dependent state variables (such as hydraulic heads) and the prediction variance of yet other state variables (such as solute concentrations or arrival times at unobserved locations). The theoretical foundation of Bayesian Geostatistical Design is summarized and discussed in Section 4.2.

One step towards further generalization is to become more independent of arbitrarily chosen model shapes of covariance functions. Within hydrogeologic applications of geostatistical inverse modeling and optimal design, it is common practice to assume a prescribed parametric form of the covariance model. *Neuman* [2003] stressed that the geostatistical model choice will always be uncertain, and that this uncertainty should not be neglected. Especially when considering the initial

scarcity of data in realistic optimal design scenarios, we deem it highly inappropriate to assume a single fixed parametric form of the covariance model.

Bayesian Model Averaging is an attractive option to account for uncertainties in model selection. *Hoeting et al.* [1999] offer a very complete review of its principles and strengths. Application to hydrogeological problems is considered by *Neuman* [2003]. In this work, the Matérn family of covariance functions is used [*Matérn*, 1986] because it has an additional shape parameter. *Feyen et al.* [2003] mentioned briefly that this shape parameter could be used to represent uncertainty in the shape of covariances. Important details on controlled differentiability and smoothness, model averaging, and on the role in geostatistical inversion are summarized by *Zhang and Rubin* [2009]. Following their rationale, we utilize the parametric control on covariance shape to transform the model selection problem to a stochastic parameter inference problem. Treating the shape parameter as random variable resembles Bayesian Model Averaging over a continuous spectrum of models governed by the additional shape parameter. These matters are discussed in Section 4.3.

We illustrate the resulting framework in a synthetic case study. In Sections 4.5 and 4.6, we optimize sampling strategies for predicting (1) contaminant levels and (2) arrival times at an ecologically sensitive location, due to a plume that could evolve from a hypothetical future contamination. Considered measurements are hydraulic heads and logconductivity samples. In a series of scenarios, we vary between (a) known and (b) uncertain structural parameters in the geostatistical model. On this basis, we demonstrate and discuss how the resulting design patterns adapt to structural uncertainty, how well the design patterns allow to identify the geostatistical model, and how the conditional prediction variance is affected by structural uncertainty.

For the synthetic case study, we use a lean and computationally efficient implementation

based on the static Ensemble Kalman Filter [*Herrera and Pinder*, 2005] and the Kalman Ensemble Generator [*Nowak*, 2009b]. We extend these approaches by a first-order expansion in structural parameters, as shown in Section 4.4. It is important to stress that neither is Bayesian Optimal Design limited to the implementational choice used in this work, nor is it restricted to the exemplary choice of unknown parameters and data types.

In summary, the objective of this work is to develop an optimal design framework that more accurately characterizes, handles and accounts for the uncertainty in geostatistical model selection and in structural parameter values. The main benefits will be to reduce the arbitrariness of *a priori* choices and to shorten the list of assumptions that are hard to defend in the absence of sufficient data. The results presented in this chapter have direct implications related to human health risk. This is addressed in the conclusion section of this chapter.

4.2 Bayesian Geostatistical Design

4.2.1 Model-Based Bayesian Geostatistics

Model-based geostatistics refer to the choice of parametric models for the mean value, large-scale trends and the covariance function [*Diggle and Ribeiro*, 2007]. The Bayesian version adds uncertainty in the parameters of the geostatistical model, and forms the basis for Bayesian Geostatistical Design.

Consider $\mathbf{s}$ a $n_s \times 1$ random space vector $\mathbf{s} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathbf{s}$ (e.g., logconductivity discretized on a numerical grid). It is comprised of a trend model $E[\mathbf{s}] = \mathbf{X}\boldsymbol{\beta}$ plus zero-mean fluctuations $\boldsymbol{\varepsilon}_\mathbf{s}$. $\mathbf{X}$ is a $n_s \times p$ matrix containing p deterministic trend functions with p corresponding trend coefficients $\boldsymbol{\beta}$. $\boldsymbol{\theta}$ are the structural parameters , such as correlation scale and variance parameters of

a covariance function, so that $\boldsymbol{\varepsilon}_{\mathbf{s}}$ has a covariance matrix $\mathbf{C} = \mathbf{C}(\boldsymbol{\theta})$.

Conventional geostatistics consider known structural parameters, and the distribution of $\mathbf{s}$ is $p(\mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\theta})$. Bayesian geostatistics reflect the uncertainty of structural parameters by their joint distribution $p(\boldsymbol{\beta}, \boldsymbol{\theta})$. This is in contrast to classical variogram analysis [*Matheron*, 1971] and maximum likelihood estimation methods [*Kitanidis*, 1995], where structural parameters are represented by simple point estimates. The Bayesian distribution (marked by a tilde) is obtained by marginalization [*Kitanidis*, 1986]:

$$\tilde{p}(\mathbf{s}) \quad = \quad \int_{\boldsymbol{\beta}} \int_{\boldsymbol{\theta}} p(\mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\theta})\, p(\boldsymbol{\beta}, \boldsymbol{\theta})\; d\boldsymbol{\theta}\, d\boldsymbol{\beta}\,. \tag{4.1}$$

Now consider the $n_y \times 1$ vector $\mathbf{y}$ of measurements at locations $\mathbf{x}_m$ according to $\mathbf{y} = \mathbf{f}_y(\mathbf{s}) + \boldsymbol{\varepsilon}_{\mathbf{r}}$. Here, $\mathbf{f}_y(\mathbf{s})$ is a process model (e.g., the groundwater flow equation) that relates observable variables (e.g., hydraulic heads) to $\mathbf{s}$. $\boldsymbol{\varepsilon}_{\mathbf{r}}$ is a vector of random measurement errors with known distribution $p(\varepsilon_{\mathbf{r}})$. The distribution of $\mathbf{s}$ conditional on a given vector $\mathbf{y}_o$ of measurement data is according to Bayes theorem:

$$p(\mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y}_o) \quad \propto \quad p(\mathbf{y}_o|\mathbf{s})\, p(\mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\theta})\;, \tag{4.2}$$

The Bayesian distribution is obtained by marginalization:

$$\tilde{p}(\mathbf{s}|\mathbf{y}_o) \quad = \quad \int_{\boldsymbol{\beta}} \int_{\boldsymbol{\theta}} p(\mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y}_o)\, p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}_o)\; d\boldsymbol{\theta}\, d\boldsymbol{\beta}\,. \tag{4.3}$$

Note that the entire distribution $p(\mathbf{s}, \boldsymbol{\beta}, \boldsymbol{\theta})$ has been jointly conditioned on the field observations $\mathbf{y}_o$. For a detailed discussion, see *Kitanidis* [1986]; *Pardo-Iguzquiza* [1999]; *Woodbury and Ulrych* [2000]; *Diggle and Ribeiro* [2002].

The final purpose is the prediction of yet a different variable c (e.g., concentration or contaminant arrival time), related to $\mathbf{s}$ via $c = f_c(\mathbf{s})$ (e.g., the transport equation). The Bayesian

predictive distribution for c is obtained by integration over the distribution of $\mathbf{s}$:

$$\tilde{p}\left(c|\mathbf{y}_o\right) \propto \int_{\boldsymbol{\beta}}\int_{\boldsymbol{\theta}}\int_{\mathbf{s}} p\left(c|\mathbf{s}\right) p\left(\mathbf{s}|\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{y}_o\right) p\left(\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}_o\right)\, d\mathbf{s}\, d\boldsymbol{\theta}\, d\boldsymbol{\beta} \tag{4.4}$$

with Bayesian mean $\tilde{c}$ and increased variance $\tilde{\sigma}^2_{c|\mathbf{y}}$

$$\tilde{c}\left(\mathbf{y}_o\right) = E_{\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}_o}\left[E_{\mathbf{s}}\left[f_c\left(\mathbf{s}\right)|\mathbf{y}_o,\boldsymbol{\beta},\boldsymbol{\theta}\right]\right] \tag{4.5}$$

$$\begin{aligned}\tilde{\sigma}^2_{c|\mathbf{y}}\left(\mathbf{y}_o\right) &= E_{\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}_o}\left[V_{\mathbf{s}}\left[f_c\left(\mathbf{s}\right)|\mathbf{y}_o,\boldsymbol{\beta},\boldsymbol{\theta}\right]\right] \\ &\quad + V_{\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}_o}\left[E_{\mathbf{s}}\left[f_c\left(\mathbf{s}\right)|\mathbf{y}_o,\boldsymbol{\beta},\boldsymbol{\theta}\right]\right]\end{aligned} \tag{4.6}$$

where $E_a\left[\cdot\right]$ is the expected value operator over the distribution of a random variable a, and $V_a\left[\cdot\right]$ is the respective variance. Equation (4.5) follows from the double expectation theorem, and equation (4.6) reflects a variance increased by uncertainty in the conditional mean value [compare with *Kitanidis*, 1986].

4.2.2 Optimal Design

Outside geostatistics, optimal design theory has a long history in the traditional context of linear and non-linear regression [*Pukelsheim*, 2006] and its application to geostatistics is explained by *Müller* [2007]. A review and synthesis of specific geostatistical design criteria is provided by *Nowak* [2009a].

A design is a set of decision variables $\mathbf{d}$ that specify the number, location and types of measurements to be collected in the data vector $\mathbf{y}$. The objective is to minimize the uncertainty inherent in the predictive distributions $p\left(\mathbf{s}|\mathbf{y}_o\right)$ or $p\left(c|\mathbf{y}_o\right)$, before even knowing the data values $\mathbf{y}_o$. To this end, a task-specific measure of prediction uncertainty $\phi\left(\mathbf{d},p\right)$ is defined [*Müller*, 2007; *Nowak*, 2009a] and minimized. Characterization needs defined within a task driven approach are also given

in *Maxwell et al.* [1999]; *de Barros and Rubin* [2008]; *de Barros et al.* [2009]. For Bayesian Geostatistical Design, these distributions are simply replaced by their Bayesian counterparts $\tilde{p}\left(\mathbf{s}|\mathbf{y}_o\right)$ or $\tilde{p}\left(c|\mathbf{y}_o\right)$ (equations 4.3 and 4.4):

$$\phi\left(\mathbf{d},\tilde{p}\right) \;=\; E_{\mathbf{y}}\left[\phi\left(\mathbf{y}\left(\mathbf{d}\right),\tilde{p}\right)\right] = \int \phi\left(\mathbf{y}\left(\mathbf{d}\right),\tilde{p}\right)\tilde{p}\left(\mathbf{y}\right)\,d\mathbf{y}\,. \tag{4.7}$$

In allusion to the monetary context [*James and Gorelick*, 1994; *Feyen and Gorelick*, 2005], equations like equation (4.7) are sometimes called the expected data worth.

Equation (4.7) implicitly includes averaging over all possible values of the structural parameters, because

$$\tilde{p}\left(\mathbf{y}\right) = \int_{\boldsymbol{\beta}}\int_{\boldsymbol{\theta}} p\left(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\beta}\right)p\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)\,d\boldsymbol{\theta}\,d\boldsymbol{\beta}\,. \tag{4.8}$$

Diggle and Lophaven [2006] evaluated the integral in Eq. (4.8) only at the *truth* value $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ of the structural parameters. It is believed that it is inadequate to assume any value of $\boldsymbol{\theta}$ to be *true* within the Bayesian paradigm. To be precise, the credibility of any particular value of $\boldsymbol{\theta}$ to qualify as *truth* is quantified by its prior distribution, leading to integration over $p\left(\boldsymbol{\beta},\boldsymbol{\theta}\right)$.

In the illustrative test case (but not as a limitation of the general framework), we will choose to minimize the expected Bayesian prediction variance of c:

$$\begin{aligned} E_{\mathbf{y}}\left[\tilde{\sigma}^2_{c|\mathbf{y}}\right] \;=\; & E_{\mathbf{y}}\left\{E_{\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}}\left[V_{\mathbf{s}|\mathbf{y}(\mathbf{d}),\boldsymbol{\beta},\boldsymbol{\theta}}\left[f_c\left(\mathbf{s}\right)\right]\right]\right\} \\ & + E_{\mathbf{y}}\left\{V_{\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}}\left[E_{\mathbf{s}|\mathbf{y}(\mathbf{d}),\boldsymbol{\beta},\boldsymbol{\theta}}\left[f_c\left(\mathbf{s}\right)\right]\right]\right\}\,. \end{aligned} \tag{4.9}$$

It is important to highlight the two individual contributions to overall prediction uncertainty in the right-hand-side of the above equation: The first term resembles the prediction variance of concentration, averaged over all possible values of potential data and structural parameters. The second term

reflects how the estimate of concentration varies due to the uncertainty of structural parameters. The two terms result directly from Bayesian principles. They combine the objectives of interpolation and structural identification in a natural manner without requiring manual weighting.

The second term vanishes at the limit of known structural parameters, e.g., when the data $\mathbf{y}_o$ set allows strong inference of the structural parameters. If, in addition, $\mathbf{f}_y(\mathbf{s})$ and $f_c(\mathbf{s})$ are linear (i.e., the estimation problem is linear), the remaining prediction variance $V_{\mathbf{s}|\mathbf{y}(\mathbf{d}),\boldsymbol{\beta},\boldsymbol{\theta}}[\cdot]$ in equation (4.9) is independent of data values, and the operator $E_{\mathbf{y}}[\cdot]$ disappears. This property is inherited from the estimation variance of Kriging [e.g., *Journel and Huijbregts*, 1978; *Rubin*, 2003].

Note that this form of the minimum variance design criterion complies with all four guidelines stated in the introduction of this chapter:

1. It allows for uncertain structural parameters in the entire analysis.

2. It includes the identification of structural parameters in the objective function.

3. The importance of identifying the structural parameters is judged naturally via their contribution to the overall prediction uncertainty.

4. The optimal design is robust over the entire range of structural parameter specified by their prior distribution.

More details on the fulfillment of the four guidelines are provided in Sections 4.6 and 4.7.

4.3 Continuous Bayesian Model Averaging and the Matérn Family of Covariance Functions

So far we have formulated the Bayesian Geostatistical Design framework without specifying too much about its uncertain input components. In this section, we address the issue of geostatistical model uncertainty, and present an approach to incorporate the uncertainties of model selection into the framework of Bayesian Geostatistical Design. The fundamental principle of Bayesian Model Averaging [*Hoeting et al.*, 1999] is that each considered model alternative is assigned a prior probability to reflect its (subjective) credibility level. The modeling task is performed with all model alternatives. For each alternative, the mismatch between model predictions and available data is used to assign a likelihood. Posterior credibilities are then assigned as the product of prior credibility and likelihood. The final result is the ensemble of model outcomes, each one weighted by its posterior credibility. The overwhelming advantage is the increased robustness towards errors in individual conceptual models or in model selection.

The very same principle can be applied to the problem of geostatistical model selection [*Neuman*, 2003]. One could pick an arbitrary choice from the entire list of traditional parametric covariance models [*Rubin*, 2003, chapter 2], and then proceed with Bayesian Model Averaging. However, that the choice of model alternatives should not be restricted by traditional adherence to a small set of certain preferred covariance models.

Instead, we recommend a more elegant approach based on the Matérn family of covariance functions *Matérn* [1986]. *Zhang and Rubin* [2009] suggest to use the flexibility of the Matérn family in order to include uncertainty in covariance shape and smoothness into geostatistical inversion.

The Matérn function is given by:

$$\begin{aligned} C(\ell) &= \frac{\sigma_Y^2}{2^{\kappa-1}\Gamma(\kappa)} \left(2\sqrt{\kappa}\ell\right)^{\kappa} B_{\kappa}\left(2\sqrt{\kappa}\ell\right) \\ \ell &= \sqrt{\left(\frac{\Delta x_1}{\lambda_1}\right)^2 + \left(\frac{\Delta x_2}{\lambda_2}\right)^2 \ldots}, \end{aligned} \tag{4.10}$$

where σ_Y^2 is the variance of logconductivity, ℓ is the anisotropic effective separation distance, and $\kappa \geq 0$ is an additional shape parameter. $\Gamma(\cdot)$ is the Gamma function, and $B_k(\cdot)$ is the modified Bessel function of the third kind (Bessel's k) of order κ [*Abramowitz and Stegun*, 1972, section 10.2]. ℓ has λ_i as scale parameters for each spatial dimension. In the form provided here, ℓ is scaled by a factor $2\sqrt{\kappa}$ to make the integral scale roughly independent of κ [e.g., *Handcock and Stein*, 1993]. For the specific values of $\kappa = 0.5,\ 1,\ \infty$, the Matérn family simplifies to the exponential, Whittle and Gaussian covariance models, respectively (see Figure 4.1). More details on properties and specific additional advantages of the Matérn family are provided by *Zhang and Rubin* [2009] and by *Stein* [1999].

The novelty of this approach is the following: If one treats κ as a discrete random variable to resemble model selection, one arrives back at the principle of Bayesian Model Averaging. We suggest to keep κ a continuous parameter on the positive real line, introducing a continuous spectrum of model alternatives. We then simply include κ in the vector $\boldsymbol{\theta}$ and treat it no different than the other uncertain structural parameters. This way, we convert the problem of model selection to a problem of stochastic parameter inference, embedded in the Bayesian approach, with a long list of available methods to draw from. We refer to this approach as *Continuous Bayesian Model Averaging*.

The idea to treat Matérn's κ as uncertain structural parameter has already been used in *Handcock and Stein* [1993] and in *Diggle and Ribeiro* [2002] for the geostatistical description of a

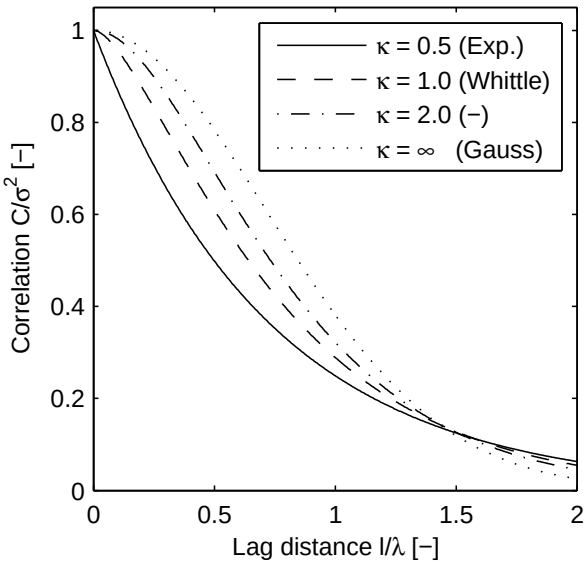


Figure 4.1: Examples from the Matérn family of covariance functions for different values of the shape parameter κ, including some special cases

digital elevation model. However, these authors did not discuss their choice in the light of the model selection problem, Bayesian Model Averaging, its extension to the continuous case, optimal design and the robustness of designs with respect to inadequate model selection.

4.4 Implementational Choices for the Illustrative Test Case

4.4.1 Computational Approach

In the upcoming test case, we will demonstrate the differences between optimal design patterns for known and uncertain structural parameters. For this purpose, we need a high spatial resolution of allowable sampling positions. In the present section, we provide a computationally efficient first-order approximation of structural uncertainty attached to an Ensemble Kalman Filter [*Evensen*, 1994; *Burgers et al.*, 1998; *Evensen*, 2003] and a sequential exchange optimization

algorithm.

However, Bayesian Geostatistical Design is not restricted to any of the choices and approximations taken in our study. In principle, any conditional simulation tool would suffice, given that it allows evaluation of conditional prediction variances. The advantage of the Ensemble Kalman Filter is its flexibility, the straightforward implementation, and its ability to evaluate prediction variances without conditioning the random fields.

4.4.2 Computational Efficiency

The computational costs for evaluating equation (4.9) should not be underestimated. This holds in particular if $\mathbf{f}_y(\mathbf{s})$ and $f_c(\mathbf{s})$ require to solve partial differential equations. To reduce computational costs, most studies restrict the design space to only a few allowable sampling locations, or by comparing a low number of design candidates: *Feyen and Gorelick* [2005] compared 25 different design candidates, and *Janssen et al.* [2008] considered 42 allowable sampling positions. *McKinney and Loucks* [1992] reached 200 allowable locations because they featured only direct measurements of logconductivity and linearized the prediction problem.

Linearized approaches are computationally very efficient and useful when understanding how to use them within their range of validity. *Cirpka et al.* [2004] used adjoint-state sensitivities in conjunction with FFT-based error propagation [*Nowak et al.*, 2003], reaching 90,000 allowable locations in a hydraulic design problem. The static Ensemble Kalman Filter allowed *Herrera and Pinder* [2005] to turn towards joint space-time optimization of sampling networks.

None of these studies considered structural uncertainty. Handling uncertain structural parameters complicates the evaluation of equation (4.9). *Diggle and Lophaven* [2006] restricted their study to the comparison of only two different design patterns, and featured only direct measure-

ments of the estimated parameter field. The implementation shown here allow us to consider a fine raster of 20,000 sampling position candidates on a contemporary desktop computer, while allowing for both structural uncertainty and complex relations between data, parameters and the prediction goal.

4.4.3 Multi-Gaussian First-Order Second-Moment Approximation

Casting Bayesian Geostatistical Design into a first-order second-moment framework paves the way to derive closed-form expressions for Bayesian design criteria [*Marchant and Lark*, 2007]. We model logconductivity as a multi-Gaussian vector $\mathbf{s}$ of discrete cell-wise values with $\mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathbf{N}\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{C}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right)\right)$, i.e., with mean vector $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{C}_{\mathbf{ss}}(\boldsymbol{\theta})$. In the generalized intrinsic case, uncertain $\boldsymbol{\beta}$ is absorbed in a generalized distribution of $\mathbf{s}$. We assume a Gaussian prior distribution $\boldsymbol{\beta} \sim \mathbf{N}\left(\boldsymbol{\beta}^{*}, \mathbf{C}_{\boldsymbol{\beta}\boldsymbol{\beta}}\right)$ with expected value $\boldsymbol{\beta}^{*}$ and covariance $\mathbf{C}_{\boldsymbol{\beta}\boldsymbol{\beta}}$. By assuming $\boldsymbol{\beta}$ multi-Gaussian and independent of $\boldsymbol{\theta}$, we can integrate over $p\left(\boldsymbol{\beta}\right)$ in equations (4.3) to (4.6) analytically [*Kitanidis*, 1986]: $\mathbf{s}|\boldsymbol{\theta} \sim \mathbf{N}\left(\mathbf{X}\boldsymbol{\beta}^{*}, \mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right)\right)$, where $\mathbf{G}_{\mathbf{ss}} = \mathbf{C}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right) + \mathbf{X}\mathbf{C}_{\boldsymbol{\beta}\boldsymbol{\beta}}\mathbf{X}^{T}$ is a generalized covariance matrix [*Kitanidis*, 1993]. This approach has already proven useful to generalize geostatistical inversion [*Nowak and Cirpka*, 2004].

The individual steps of linearizing $\mathbf{f}_{y}\left(\mathbf{s}\right)$ and $f_{c}\left(\mathbf{s}\right)$ in $\mathbf{s}$ are summarized in Appendix E, leading to:

$$E_{\mathbf{y}}\left[\tilde{\sigma}^{2}_{c|\mathbf{y}}\right] \;=\; E_{\boldsymbol{\theta}}\left[\sigma^{2}_{c|\mathbf{y}}\left(\boldsymbol{\theta}\right)\right] + E_{\mathbf{y}}\left\{V_{\boldsymbol{\theta}|\mathbf{y}}\left[\hat{c}\left(\mathbf{y}\left(\mathbf{d}\right), \boldsymbol{\theta}\right)\right]\right\}, \tag{4.11}$$

where $\sigma^{2}_{c|\mathbf{y}}\left(\boldsymbol{\theta}\right)$ is the conditional variance of concentration for given $\boldsymbol{\theta}$.

To further simplify equation (4.11), we expand in $\boldsymbol{\theta}$ about its prior mean value $\bar{\boldsymbol{\theta}}$, truncate

after first-order, and assume a prior covariance $\mathbf{C}_{\boldsymbol{\theta\theta}}$ to specify the structural uncertainty, similar to *Rubin and Dagan* [1987]. After executing $E_{\mathbf{y}}\{\cdot\}$, we obtain (see Appendix F for details):

$$\begin{aligned} E_{\mathbf{y}}\left[\tilde{\sigma}^2_{c|\mathbf{y}}(\mathbf{d})\right] &= \sigma^2_{c|\mathbf{y}}(\bar{\boldsymbol{\theta}}) + \sum_i \sum_j \left\langle \mathbf{C}_{\boldsymbol{\theta\theta}|\mathbf{y}} \right\rangle_{ij} \Bigg\{ \cdots \\ &\quad \cdots \left.\frac{\partial \bar{c}}{\partial \theta_i}\right|_{\bar{\theta}_i} \left.\frac{\partial \bar{c}}{\partial \theta_j}\right|_{\bar{\theta}_j} + \left(\left.\frac{\partial \boldsymbol{\kappa}}{\partial \theta_i}\right|_{\bar{\theta}_i} \right) \mathbf{G}_{\mathbf{yy}}(\bar{\boldsymbol{\theta}}) \left(\left.\frac{\partial \boldsymbol{\kappa}}{\partial \theta_j}\right|_{\bar{\theta}_j} \right)^T \Bigg\} \end{aligned} \tag{4.12}$$

where $\boldsymbol{\kappa} = \mathbf{H}_c \mathbf{G}_{\mathbf{ss}}(\boldsymbol{\theta}) \mathbf{H}_c^T \mathbf{G}_{\mathbf{yy}}^{-1}(\boldsymbol{\theta})$ is the Kalman gain of concentration in equation (F3), $\bar{c} = E_{\mathbf{s}}[c]$, and $\bar{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}[\boldsymbol{\theta}]$. $\left\langle \mathbf{C}_{\boldsymbol{\theta\theta}|\mathbf{y}} \right\rangle_{ij}$ is the i,j-the element in the conditional covariance of $\boldsymbol{\theta}$, here approximated by the inverse of the Fisher information $\mathbf{F}$. Details of the derivation are provided in Appendix F.

Once actual data values become available after the optimal design task, we can update the structural parameters with the technique by *Kitanidis and Lane* [1985] and *Kitanidis* [1995], later upgraded to the generalized intrinsic case by *Nowak and Cirpka* [2006]. The conditional covariance of $\boldsymbol{\theta}$ is again approximated by the inverse of $\mathbf{F}$, and the conditional mean $\widehat{\boldsymbol{\theta}}$ is approximated by

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &\approx \bar{\boldsymbol{\theta}} - \mathbf{F}^{-1}\mathbf{g} \\ \mathbf{C}_{\boldsymbol{\theta\theta}|\mathbf{y}} &\approx \mathbf{F}^{-1}, \end{aligned} \tag{4.13}$$

where $\mathbf{g}$ is the gradient and $\mathbf{F}$ is the Fisher information matrix as specified in Appendix F.

4.4.4 The Ensemble Kalman Filter and Kalman Ensemble Generator

Equation (4.12) and the equations in the appendices merely require auto- and cross-covariances between data and predicted variables, and their derivatives with respect to the structural parameters. We entrust this task to the static Ensemble Kalman Filter (sEnKF) by *Herrera*

[1998], and obtain the derivatives with respect to $\boldsymbol{\theta}$ from additional parallel sEnKF's with slightly different parameter values. *Nowak* [2009b] clarifies that Ensemble Kalman Filters are based on a certain type of optimal linearization that outmatches traditional first-order expansions in accuracy, adequately represent ensemble dispersion and dilution of solute transport, and hence avoid the non-trivial choice of dispersion coefficients when using estimated conductivity fields [e.g., *Rubin et al.*, 1999; *Nowak and Cirpka*, 2006].

Once the design is decided upon and the data become available, we condition the log-conductivity field by the Kalman Ensemble Generator [*Nowak*, 2009b]. The Kalman Ensemble Generator is an adaptation of the EnKF to geostatistical inversion, and has been upgraded by successive linearization, a Levenberg-Marquardt stabilization and an acceptance/rejection scheme.

4.4.5 Implementation

The quasi-linear Ensemble Kalman Generator and the static Ensemble Kalman Filter are implemented in MATLAB. A standard Galerkin Finite Element Method (FEM) for groundwater flow and the streamline upwind/Petrov-Galerkin FEM for solute transport are used [*Hughes*, 1987; *Fletcher*, 1996]. The resulting equations are solved using the UMFPACK solver [*Davis*, 2004]. We generate random hydraulic conductivity fields with the spectral method by *Dietrich and Newsam* [1993]. Each ensemble had a size of 4000 realizations, which is more than sufficient for Ensemble Kalman Filters in hydrogeological applications [*Chen and Zhang*, 2006].

Unlike many recent studies [*Reed et al.*, 2000; *Zhang et al.*, 2005; *Wu et al.*, 2006; *Janssen et al.*, 2008], we do not optimize our sampling locations with genetic algorithms [*Goldberg*, 1989]. Other studies [e.g., *McKinney and Loucks*, 1992; *Criminisi et al.*, 1997; *Cirpka et al.*, 2004] took the *greedy* search algorithm [e.g., *Christakos*, 1992, p. 411], which is computationally much faster.

We use the sequential exchange algorithm [e.g., *Christakos*, 1992, p. 411], a simple-to-implement upgrade to the *greedy* search algorithm with at least Pareto-optimal design solutions.

4.5 Synthetic Case Study

4.5.1 Scenario Definition and Relevance in Risk Assessment

We demonstrate the above methodology and the impact of parametric uncertainty on optimal design patterns in a synthetic case study. Consider a potential future groundwater contamination at an environmentally (or ecologically) sensitive location due to a hypothetical upstream groundwater contamination as part of a risk scenario. We will follow two different prediction objectives: to minimize the prediction variance of (1) contaminant concentration and (2) arrival time at the sensitive location. Objective (1) has been considered in the study by *McKinney and Loucks* [1992]. We extend their scenario to uncertainty in the geostatistical model and its structural parameters, and to measurements of dependent state variables such as hydraulic heads.

Two objectives are chosen because different objectives can yield fundamentally different design patterns. The actual choice of design objectives in site-specific applications of course depends on the modeling and management goals at the site under consideration. If necessary, multi-purpose design techniques [*Müller*, 2007] may be used to fuse different design objectives into one. The scope of the current case study, however, is not to compare different prediction objectives. The main point here is to illustrate the principle of Bayesian geostatistical design, i.e., how optimal sampling patterns change when allowing for structural uncertainty. For this purpose, our two exemplary objectives will suffice, and we will discuss the physical mechanisms that lead to the resulting patterns only in limited detail.

We will place 24 boreholes to obtain both core-scale measurements of transmissivity (e.g., from slug tests or disturbed-core grain-size analysis) and additional co-located measurements of hydraulic head (e.g., from minimum-cost groundwater level monitoring wells at the cored locations). To demonstrate the effect of structural uncertainty, we compare the results between (a) known and (b) uncertain structural parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in the geostatistical model. Combined with our two prediction objectives, this yields four different cases (1a, 1b, 2a and 2b; see Table 4.3).

This type of scenario is relevant, e.g., in the probabilistic assessment of human health risk [*de Barros and Rubin*, 2008; *de Barros et al.*, 2009], and its motivation shall be laid out in brief. Groundwater contamination in the proximity of drinking water wells or other environmentally sensitive locations may pose risks to human health risk. Many environmental regulations define concentration thresholds for such cases [*USEPA*, 1989, 1991, 2001]. This entails stochastic prediction of contaminant concentration at the sensitive location [e.g., *Rubin et al.*, 1994; *Andricevic and Cvetkovic*, 1996; *Maxwell et al.*, 1999]. Incorporating hydrogeological flow data helps to reduce the involved uncertainties, allows for a tighter prediction of contamination and health risk, and thus supports management decisions at a lower risk of liability [*Maxwell et al.*, 1999]. For the trade-off between uncertainties from site investigation and health related parameters, see *de Barros and Rubin* [2008]; *de Barros et al.* [2009] and Chapters 2-3 of this dissertation.

4.5.2 Flow and Transport Configuration

For simplicity, but not a limitation to the framework, we limit our solute transport problem to the late-time concentration and the arrival time down-gradient of a continuous line source in a depth-integrated 2D setting, and consider a point-like sensitive location. Depth-integrated steady-state groundwater flow is described by:

$$\nabla \cdot [T(\mathbf{x}) \nabla h] = 0\,, \tag{4.14}$$

where $T\left[L^2/t\right]$ is locally isotropic transmissivity and $h\left[L\right]$ is hydraulic head. The space coordinates are represented by $\mathbf{x} = (x_1, x_2)$. Boundary conditions are specified later. For the steady-state concentration, we use

$$\mathbf{v} \cdot \nabla c - \nabla \cdot (\mathbf{D}_d \nabla c) = 0\,, \tag{4.15}$$

where $c\left[M/L^3\right]$ is concentration, $\mathbf{v} = \mathbf{q}/n_e$ is velocity, $\mathbf{q}$ is the Darcy specific discharge, n_e is porosity, and $\mathbf{D}_d\left[L^2/t\right]$ is the pore-scale-dispersion tensor according to *Scheidegger* [1954]. We simulate the arrival time t_{50} using moment-generating equations [*Harvey and Gorelick*, 1995]:

$$\mathbf{v} \cdot \nabla m_k - \nabla \cdot (\mathbf{D}_d \nabla m_k) = k m_{k-1}\,, \tag{4.16}$$

with $t_{50} = m_1/m_0$, where m_0 and m_1 are the zeroth and first temporal moments of breakthrough for the related instantaneous release problem, respectively. *Cirpka and Kitanidis* [2000] discuss the physical meaning of temporal moments, and exemplary applications of the generating equations can be found in *Cirpka and Nowak* [2004]; *Nowak and Cirpka* [2006].

The physical configuration, domain size, and relevant parameter values are provided in Table 4.1 and Table 4.2. Boundary conditions are $\hat{h} = 1m$ and $\hat{h} = 0$ m at $x_1 = 0$ m and $x_1 = 600$ m, respectively. Uncontaminated groundwater enters at $x_1 = 0$ m, and the outflow boundary at $x_1 = 600$ m is unrestricted. The remaining two boundaries at $x_2 = 0$ m and $x_2 = 200$ m are no-flux boundaries for both flow and transport.

We consider a fixed-concentration source with unit concentration $c_0 = 1$ along a 50 m ($\approx$ 3 integral scales) wide line centered at $x_1 = 150$ m. A sensitive location is located at a

Numerical domain			
domain size	$[L_1,\, L_2]$	$[m]$	$[600,\, 200]$
grid spacing	$[\Delta_1,\, \Delta_2]$	$[m]$	$[2,\, 0.5]$
Transport parameters			
head difference	Δh	$[m]$	1
effective porosity	n_e	$[-]$	0.35
pore-scale dispersivities	$[\alpha_\ell,\, \alpha_t]$	$[m]$	$[2,\, 0.25]$
diffusion coefficient	D_m	$\left[m^2/s\right]$	10^{-9}
Transversal plume dimension	ℓ_S	$[m]$	50m
Geostatistical model parameters (prior mean values)			
global mean	$\beta_1 = \ln K_g$	$[-]$	$\ln\left(10^{-5}\right)$
trend x_1	β_2	$[-]$	0
trend x_2	β_3	$[-]$	0
variance	σ_Y^2	$[-]$	1.00
integral scales	$[\lambda_1,\, \lambda_2]$	$[m]$	$[15,\, 15]$
Matrn's kappa	κ	$[-]$	2.50
Measurement error standard deviations			
$Y \equiv \ln K$	$\sigma_{r,Y}$	$[-]$	1.00
head h	$\sigma_{r,h}$	$[m]$	0.01

Table 4.1: Parameter values used for the synthetic test cases.

Dimensionless numbers			
Longitudinal travel distance	$\xi = x_1/\lambda_1$	$[-]$	10.00
Transverse offset	$\eta = x_2/\lambda_2$	$[-]$	0.83
Contaminant source width	$\zeta = \ell_s/\lambda_2$	$[-]$	3.33
Longitudinal Peclet	$Pe_\ell = \lambda_1/\alpha_\ell$	$[-]$	7.50
Transverse Peclet	$Pe_t = \lambda_2/\alpha_t$	$[-]$	60.00

Table 4.2: Dimensionless representation of the relevant parameters used for the synthetic test cases.

longitudinal travel distance of 300 m down-gradient from the source, and transversely offset by 12.5 m ($\approx$ 1 integral scale) relative to the center of the line source. Dimensionless numbers (in terms of integral scales) for reference are also denoted in Table 4.2. The domain geometry, contaminant source and the sensitive location are illustrated in Figure (4.2).

4.5.3 Bayesian Geostatistical Setup and Test Cases

Predicting contaminant transport over some distance in heterogeneous formations requires assumptions on the structure of variability. In the scenario shown here, we assume that a single geostatistical model applies to the entire domain. For reasons of parsimony, this model is stationary in cases 1a and 2a, and intrinsic (due to a trend model with uncertain coefficients) in cases 1b and 2b, see Table 4.3. Cases 1b and 2b are less arbitrary and less subjective in their prior model assumptions: They do not claim to deterministically know the global mean, trend or the covariance function in absence of information, i.e., prior to design and data collection.

The remaining assumptions are that a single, domain-wide, intrinsic and multi-Gaussian geostatistical model can be justified, e.g., because geological maps indicate membership to a single

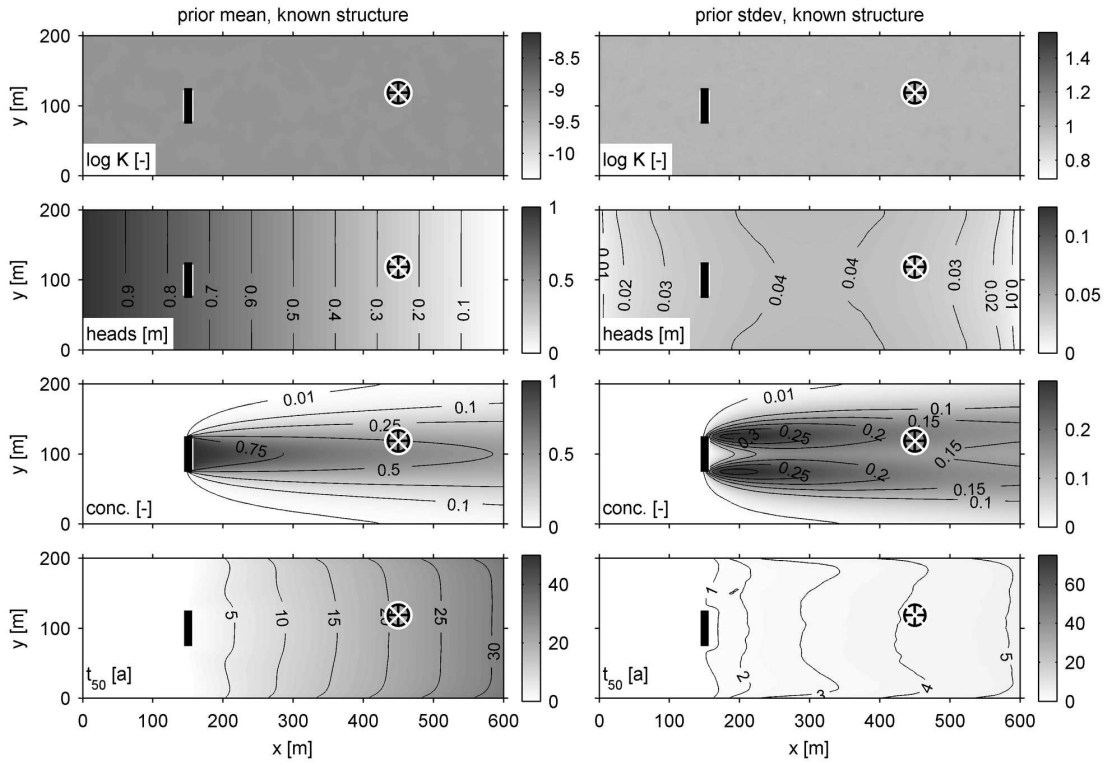


Figure 4.2: Illustration of the scenario for known structural parameters. Left: prior mean values of $Y = \ln K$, corresponding hydraulic heads h and hypothetical plume (late-time concentration c and arrival time t_{50}). Right: prior standard deviation. Crossed circle: sensitive location. Thick black line: hypothetical contaminant source. For parameter values, see Table 4.3 for cases 1a and 2a and Table 4.1. Grey-scale is identical to Figure 4.3 for direct comparison.

geological unit. Less parsimonic descriptions, e.g., different zones covered by different geostatistical models, can be adopted if necessary. This would increase the number of structural parameters to be identified and result in different sampling patterns.

We generate random log-transmissivity fields using the Matérn family of covariances plus a global mean and a linear trend. The two linear trend functions have a spatial mean of zero and cause a total variation of ± 0.5 over the respective length of the domain. Following the Bayesian rationale in geostatistics, we do not claim to know the parameter values or actual shape of the

Case Number	Objective	Assumptions	structural uncertainty
1a	σ_c^2	$\boldsymbol{\beta}, \boldsymbol{\theta}$ known	none
1b	σ_c^2	$\boldsymbol{\beta}, \boldsymbol{\theta}$ uncertain	$var\left[\beta_1, \beta_2, \beta_3, \sigma_Y^2, \lambda_1, \lambda_2, \kappa\right]$ $= [1, 1, 1, 0.5, 112.5, 112.5, 1.75]$
2a	σ_{t50}^2	$\boldsymbol{\beta}, \boldsymbol{\theta}$ known	none
2b	σ_{t50}^2	$\boldsymbol{\beta}, \boldsymbol{\theta}$ uncertain	$var\left[\beta_1, \beta_2, \beta_3, \sigma_Y^2, \lambda_1, \lambda_2, \kappa\right]$ $= [1, 1, 1, 0.5, 112.5, 112.5, 1.75]$

Table 4.3: Definition of test cases in our scenario. Objective: the quantity to be minimized by sampling (prediction variance of contaminant concentration or of arrival time at the sensitive location, respectively). Symbols: β_1 $[-]$: global mean of $\ln K$; β_2 and β_3 $[-]$: linear trend parameters; λ_1 and λ_2 $[m]$: scale parameters (spatial correlation); κ $[-]$: shape parameter of the Matrn function.

geostatistical model better than specified by a prior distribution.

Only for cases 1a and 2a, the structural parameters are considered known. For cases 1b and 2b, their values are uncertain, with squared coefficients of variation $CV^2 = 0.5$ each. Uncertain parameters are the global mean value β_1, the trend coefficients in x_1 and x_2 directions (β_2 and β_3, respectively), the variance σ_Y^2, the scale parameters in x_1 and x_2 directions (λ_1 and λ_2, respectively), and the Matérn shape parameter κ. Their prior mean values and variances are specified in Table 4.1. For simplicity, we assume prior stochastic independence among the structural parameters and use Gaussian measurement errors with $\sigma_r^2 = 1$ for measurements of $\ln T$ and $\sigma_r^2 = (0.01m)^2$ for hydraulic head.

4.5.4 Effect of Structural Uncertainty on Prediction Mean and Variance

Figures 4.2 and 4.3 compare prior mean values and standard deviations of Y, h, c and t_{50} for the case of known and uncertain structural parameters, respectively. They are obtained from Monte-Carlo analysis with 16000 realizations each, using the geostatistical settings described in Section 4.5.3.

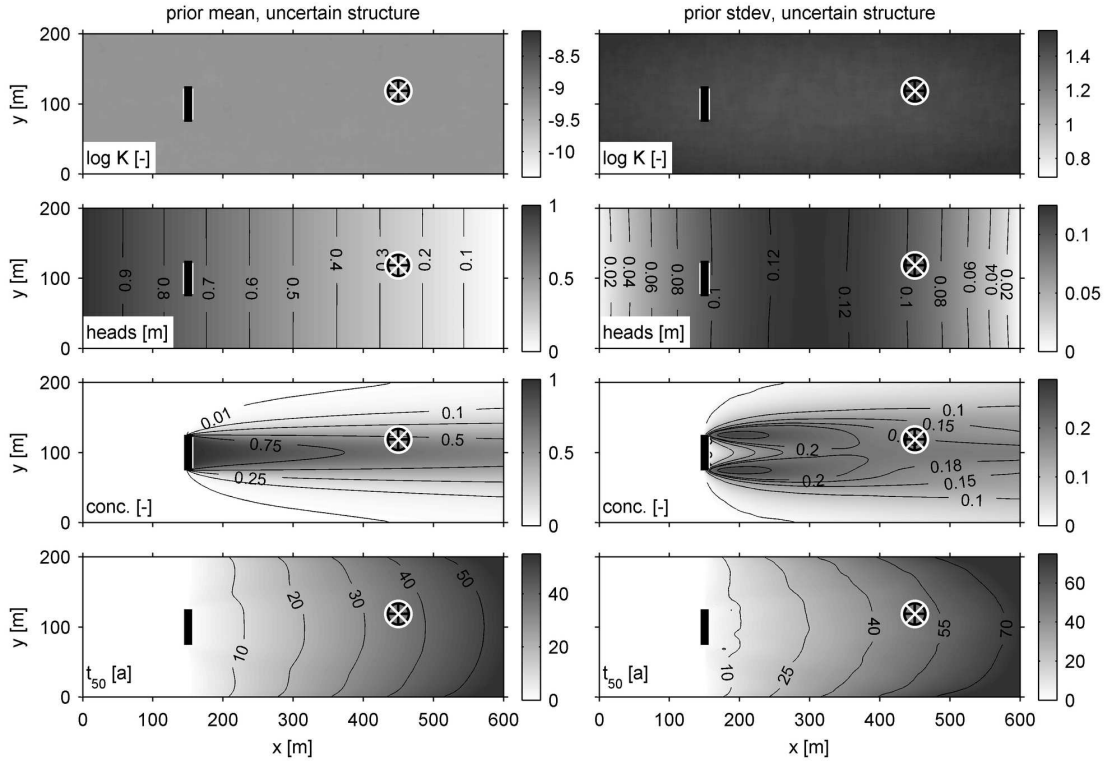


Figure 4.3: Illustration of the scenario for uncertain structural parameters. Left: prior mean values of $\ln K$, corresponding hydraulic heads h and hypothetical plume (late-time concentration c and arrival time t_{50}). Right: prior standard deviation. Crossed circle: sensitive location. Thick black line: hypothetical contaminant source. For parameter values, see Table 4.3 for cases 1b and 2b and Table 4.1. Grey-scale is identical to Figure 4.2 for direct comparison.

The impact of uncertain mean and trend manifests in the form of a prior standard deviation of logconductivity with values larger than $\sigma_Y = 1$ in the center of the domain, with increasing values

towards the domain boundaries (see Figure 4.3). The standard deviation of h for uncertain structure is dominated by the uncertain trend in x_1 direction, because the trend in x_1 controls whether the main energy loss appears in the first or in the second half of the domain.

Due to structural uncertainty, the standard deviation of concentration in Figure 4.3 also differs from the one with known structural parameters (Figure 4.2) or from the analytical expressions for line sources found in the literature [*Fiorotto and Caroni*, 2002; *Caroni and Fiorotto*, 2005; *Schwede et al.*, 2008]. The cited analytical solutions and our Monte-Carlo analysis for known structure display two distinct lines of high variance along the fringes of the plume. Structural uncertainty mainly increases the concentration variance along the center line of the plume, filling the space between those two lines. The explanation is that macro-dispersion and the approach rate to ergodicity become uncertain when the variance, integral scales and anisotropy are uncertain. The affected area extends from the source to far beyond the sensitive location. Results from different Monte-Carlo analyses (not shown here) indicate that the global trend functions have almost no impact on concentration variance.

With uncertain structure, the standard deviation for arrival time explodes by a factor of roughly ten; this can be traced back to the uncertain global mean of $Y = \ln K$, which dictates the average velocity. Variance, integral scales and anisotropy have an impact on large-scale effective hydraulic conductivity [*Zhang*, 2002; *Rubin*, 2003], so that uncertain covariance parameters increase the uncertainty of arrival time. The obtained arrival time statistics and their dependence on travel distance are in agreement with the findings of *Rubin and Dagan* [1992].

4.6 Results: Near-Optimal Sampling Patterns with Uncertain Structural Parameters

In this section, we present the sampling patterns resulting from Bayesian Geostatistical Design, considering the structural parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as uncertain. The main steps of analysis are:

1. Find a near-optimal design using the techniques described in Section 4.4;
2. Generate a respective synthetic data set for the suggested sampling pattern by unconditional random simulation of an aquifer.
3. With the sampled synthetic data from item 2 at locations determined by item 1, we compute the conditional ensemble statistics (e.g. mean and variance spatial maps for concentration and arrival times) for illustration using the Kalman Ensemble Generator by *Nowak* [2009b];
4. Steps 1-3 are repeated for both objectives defined in Section 4.5.1.

Results in the section are given only for cases with uncertain structural parameters (cases 1b and 2b in Table 4.3). Cases 1a and 2a and deeper discussion of the results will follow in the subsequent section, which analyze the relation between structural uncertainty and sampling design.

Figure 4.4 depicts spatial maps of logconductivity and the corresponding heads, concentrations and arrival times, obtained from unconditional random simulation with random structural parameters (see Table 4.4). We use these values to represent the *true* aquifer. We will read values of $\ln K$ and h at the near-optimal sampling locations and add random measurement error to obtain synthetic data. This way, we can compare the conditional results to fully known reference fields of $\ln K$, hydraulic heads, concentrations and arrival times, and to the random values of structural parameters used for generation.

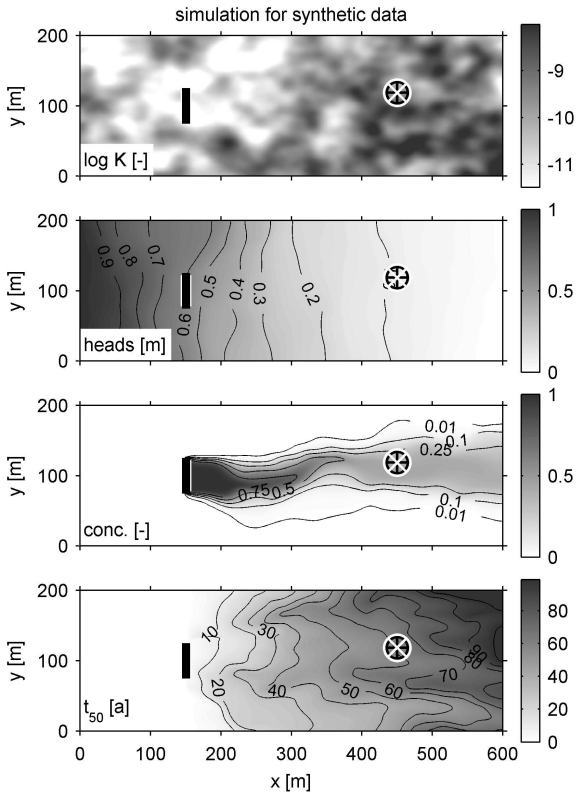


Figure 4.4: Random simulation used to obtain synthetic measurement values: a realization of $Y = \ln K$ and corresponding simulated hydraulic heads, late-time concentration and arrival time. Crossed circle: sensitive location. Thick black line: hypothetical contaminant source. For parameter values, see Tables 4.1, 4.3 and 4.4.

4.6.1 Sampling Patterns Optimized for Predicting concentration (Case 1b)

First, we present the results for case 1b (see Table 4.3): all structural parameters considered in our geostatistical model are uncertain, and we optimize the sampling pattern for optimal prediction of late-time concentration at the sensitive location. The resulting sampling pattern is shown in Figure 4.5. The figure also shows the conditional mean (left column) and standard deviations (right column) after applying the design and using the synthetic measurement values obtained from the random simulation shown in Figure 4.4.

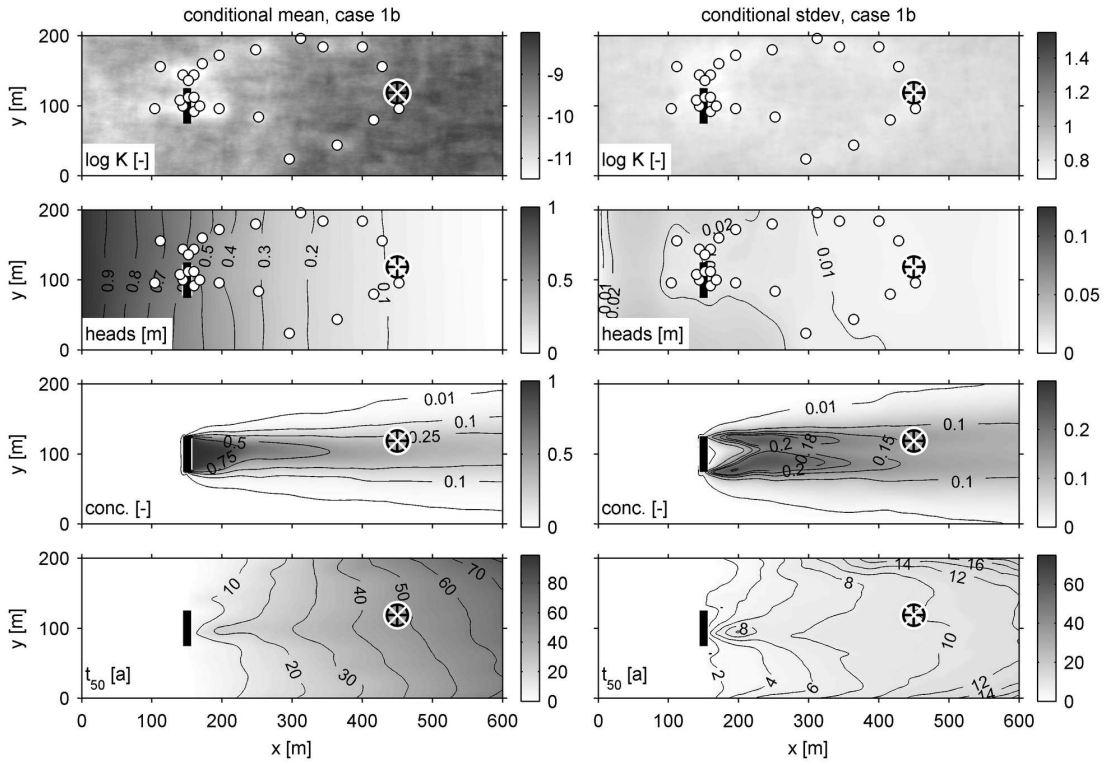


Figure 4.5: Results for case 1b. Left: conditional mean of $\ln K$, hydraulic heads h, and late-time concentration c and arrival time t_{50} of hypothetical plume. Right: corresponding conditional standard deviations. Crossed circle: sensitive location. Solid white circles: near-optimal sampling locations ($Y = \ln K$ and head measurements). Thick black line: hypothetical contaminant source. For parameter values, see Tables 4.1 and 4.3.

In principle, the original non-Bayesian prediction purpose leads to information needs in certain regions of the domain, where a certain function of the cross-covariance between measurable quantities and the prediction goal is highest [e.g., *Cirpka et al.*, 2004; *Herrera and Pinder*, 2005; *Zhang et al.*, 2005; *Nowak*, 2009a]. At the same time, the Bayesian approach requires a diversification of sampled lag distances in order to reduce structural uncertainty. Hence, the sampling pattern found for case 1b is in essence similar to the one found by *McKinney and Loucks* [1992], with small modifications due to structural uncertainty (see Section 4.7.2). Case 2b (see Section 4.6.2)

will provide an example where structural uncertainty leads to major modifications due to structural uncertainty. The current sampling pattern is asymmetrical because the environmental target is not aligned with the center of the contaminant source. All sampling locations fall into two groups, and each group provides a specific set of information:

1. Samples in and around the source.

2. Measurements flanking the average migration pattern of the hypothetical plume.

The contribution of the first group is to identify the release conditions. The most important factor is the actual volumetric flow through the source area. It substantially affects the total mass flux of contaminant leaving the source, and the resulting width and fate of the plume further down-gradient:

If the source is in a high volumetric flow region, larger mass flux will leave the source. Also, the plume will widen once if it re-enters medium or low volumetric flow regions and the streamlines diverge. If a situation like this happens, the chance of a wider plume to hit the sensitive location is higher than average. If the source is in a low volumetric flow region, a smaller mass flux leaves the source area, and the plume becomes thinner if it re-enters medium or high volumetric flow regions. In this case, the probability to hit the sensitive location is lower than average. Additionally, in such thin plumes (relative to the heterogeneity length scale), transverse hydrodynamic dispersion leads to a rapid dilution of peak concentrations at the plume's center line by mixing with background water, leading to a quick dissipation of peak concentrations. This is in agreement with theoretical derivations in the literature [*Fiorotto and Caroni*, 2002].

In summary, if the contaminant source is at a high volumetric flow zone, this will lead to larger levels of *expected* contaminant concentration further downstream, while low volumetric

flow zone, contaminant sources lead to much smaller *expected* levels of contamination. A parallel Monte-Carlo study (results not shown here) showed that up to 50% of concentration variance could be attributed to the volumetric flux through the source area. The K values within the source zone can be identified both from conductivity samples in the actual source zone, and from hydraulic heads measured around the zone. The latter help to detect focusing or divergence of flow caused by high or low volumetric flow areas, respectively. Thus, sampling locations even upstream or far beside the source are informative for downstream concentrations, leading to the sampling locations scattered around the source in Figure 4.5. In addition, it known that concentration variance is largest at early travel distances [*Fiorotto and Caroni*, 2002]. Due to the large concentration variance at early travel distances (and the fact that we are minimizing concentration variance), there is a tendency to place the samples in those locations near the source. With increasing travel times (or travel distances), dispersion starts to have a more significant role thus contributing towards concentration variance *destruction* and smoothing out the concentration variability (see works of *Fiorotto and Caroni* [2002] and *Caroni and Fiorotto* [2005]) .

The contribution of the second group of measurements is to identify the macroscopic transport direction through transverse gradients that deflect the plume from its expected mean trajectory. In other words, they convey information whether the plume is bypassing or directly hitting the sensitive location. Small-scale fluctuations are important only when they appear close to the sensitive location, while large-scale meandering is important even at a distance from the sensitive location. This explains why the two rows of samples flanking the expected plume trajectory draw closer to the plume's center in the vicinity of sensitive location. The integral impact of small-scale fluctuations further upstream is predicted sufficiently well in a stochastic sense by knowing

the structural parameters. In addition to this, measurements are located at the supposed fringe of the plume since uncertainty is highest at those locations. This intuition is in agreement with the theoretical derivation found in *Rubin* [1991]

For the current objective function, the area within the expected plume trajectory turns out to be least significant. Similar phenomena are common to studies that optimize sampling patterns to predict contaminant motion prior to release. For a scenario similar to ours but without structural uncertainty, *McKinney and Loucks* [1992] also found that optimal design patterns focus to a large degree on the direct vicinity of the source. A comparable situation occurred in the study of *Cirpka et al.* [2004], who investigated optimal placement of hydraulic measurements for the hydraulic design of funnel-and-gate systems. They found that the most useful sampling locations are at the corners of the funnel, because these locations inform best on the distribution of total mass fluxes in the funnel-and-gate control plane.

The availability of different data types is a second factor that influences sampling patterns, not less important than the choice of prediction objective. In our case study, the contamination has not yet occurred, so that concentration measurements are not available. This changes when monitoring contaminations that have already occurred, and concentration data are available as measurement types. In studies on optimal plume monitoring, for example, the resulting sampling pattern typically tries to determine the current outline of the plume, i.e., find its fringes and its current front [e.g., *Criminisi et al.*, 1997; *Herrera and Pinder*, 2005; *Wu et al.*, 2005; *Zhang et al.*, 2005].

Although the current study does not focus on the merits of conditional simulation in geostatistical inverse problems, it is worth while to compare the synthetic reality (see Figure 4.4) with the resulting conditional statistics (Figure 4.5). The global mean and trend of conductivity have

been captured well. In the synthetic example, the contaminant source happens to be in an area of slow flow, such that the bounding streamlines converge down-gradient of the source, leading to a very narrow plume with peak concentrations prevailing only over a short travel distance (see Figure 4.5, left). This effect has been captured by the dense cluster of samples around the source, so that the squeezing motion of the plume and the relatively short persistence of the $c = 0.5$ isoline is reflected in the conditional mean values. The large-scale features of the flow field have also been captured. In other randomly simulated reference fields (not shown here), the plume actually bypassed the sensitive location north or south of it. In the current example, the conditional ensemble mean plume is accurately hitting the sensitive location, with its center line passing only slightly south of the sensitive location.

The measurements convey sufficient information to reduce the uncertainty of hydraulic heads to almost zero between the source and the sensitive location. In combination, this leads to a significantly reduced prediction uncertainty of concentration up-gradient of and at the sensitive location. Also, the uncertainty of structural parameters is reduced by conditioning. At the transition towards known structural parameters, the concentration variance starts to exhibit the two distinct lines along the fringes of the plume. The actual reduction of concentration variance at the sensitive location is discussed in Section 4.7.1.

4.6.2 Sampling Patterns Optimized for Predicting Arrival Time (Case 2b)

In this section, we briefly repeat the above analysis for the design objective of minimal prediction variance of arrival time t_{50} at the sensitive location (case 2b) for later comparison to the previous results (case 1b). The main aspect of comparison will be the different impact of structural uncertainty onto designs from different objectives. The results for case 2b also illustrate that a

design which is optimal under one specific prediction objective (2b) will not necessarily perform well under a different objective (1b). If desired, multi-objective optimal design [e.g., *Müller*, 2007] may offer suitable situation-specific generalizations or compromises.

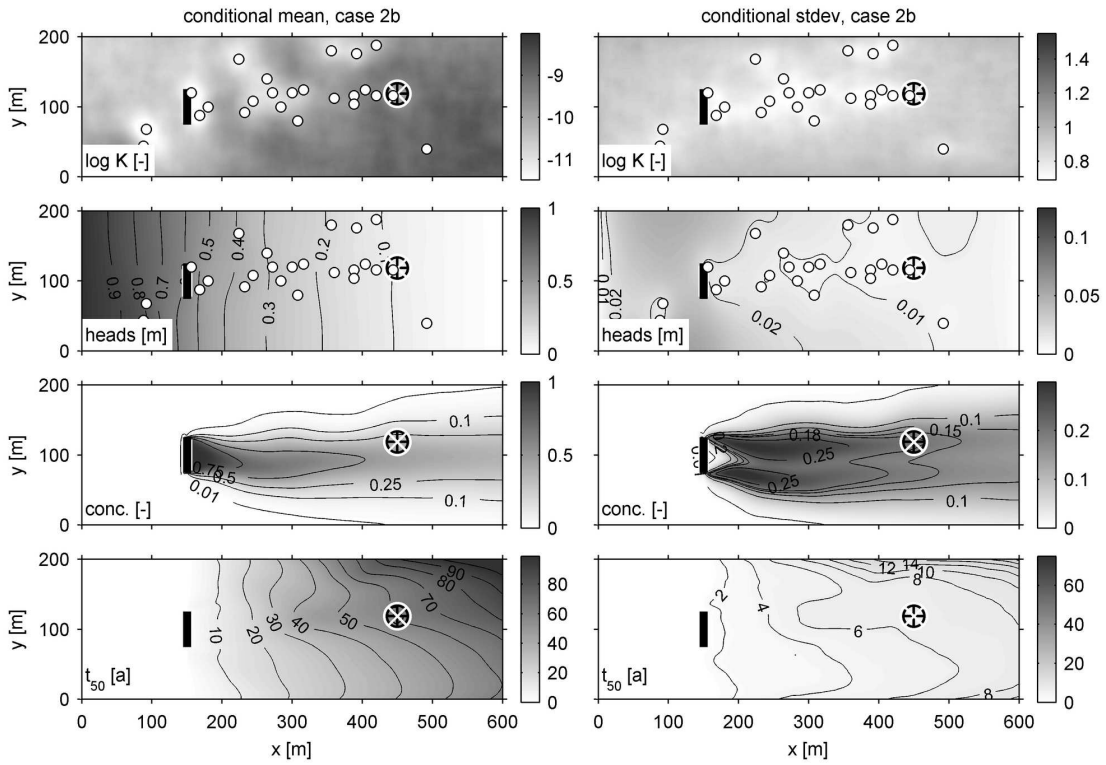


Figure 4.6: Results for case 2b. Left: conditional mean of $Y = \ln K$, hydraulic heads h, and late-time concentration c and arrival time t_{50} of hypothetical plume. Right: corresponding conditional standard deviations. Crossed circle: sensitive location. Solid white circles: near-optimal sampling locations ($\ln K$ and head measurements). Thick black line: hypothetical contaminant source. For parameter values, see Tables 4.1 and 4.3.

Figure 4.6 shows the results for case 2b, i.e. the near-optimal sampling pattern and the conditional mean (left column) and standard deviations (right column). The synthetic measurement values for conditioning are taken from the same random simulation as before (Figure 4.4). The main sampling effort goes to the area between the source and the sensitive location. This is because

arrival time is an integral outcome of the transport velocity along the entire distance. As shown later, the seemingly random scattering of measurements within and outside the area between source and sensitive location mainly addresses structural uncertainty. Some samples are scattered throughout the domain for better identification of the global mean and trend coefficients. Comparison of the conditional standard deviation between case 1b and 2b (Figures 4.5 and 4.6, respectively) shows that the pattern for case 1b is better in reducing the uncertainty of concentration, while pattern 2b performs better in reducing the uncertainty of arrival time.

4.7 Discussion

This section discusses the impact of added samples on the prediction variance, the reduction of structural uncertainty through sampling (model identification), and the impact of structural uncertainty on design patterns. Whenever applicable, the links between observed results and the four guidelines described in the introduction of this chapter will be given.

4.7.1 Effect of Sampling on Prediction Variance

How well did the near-optimal sampling patterns reduce the prediction variance of concentration? The design criterion, equation (4.12) promised (in the expected sense) that the near-optimal sampling patterns would reduce the prediction variances from $\sigma_c^2 = 0.0329$ to $\sigma_c^2 = 0.0215$ for late-time concentration, and from $\sigma_{t50}^2 = 2466.4$ to $\sigma_{t50}^2 = 1192.7$ for arrival time. σ_c^2 is dimensionless because we used $c_0 = 1[-]$ for generality.

An important caveat about expected prediction variances, such as equation (4.9), lies in their nature as expected value over yet unobserved data values. Therefore, actual prediction vari-

ances after collecting the data may of course differ from their expected values. *Feyen and Gorelick* [2005] discuss this issue within the context of expected monetary data worth. In addition, the Bayesian geostatistical framework averages over uncertain structural parameters that will later be updated with yet unobserved data. Using the synthetic data set, the near-optimal designs reduced the variances from $\sigma_c^2 = 0.0585$ to $\sigma_c^2 = 0.0204$ and from $\sigma_{t50}^2 = 2466.4$ to $\sigma_{t50}^2 = 41.121$ according to the conditional ensemble statistics. In our example, the reduction in variance is higher than expected mainly because the variance of $Y = \ln K$ smaller than its expected value.

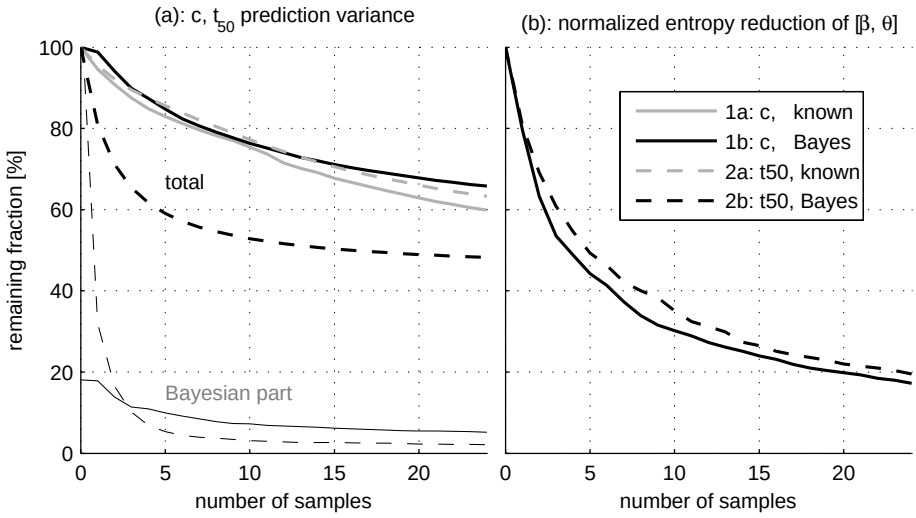


Figure 4.7: (a): reduction of prediction variance with increasing number of samples, normalized to the initial prediction variance. Upper curve set ("total", thick lines) is the expected prediction variance of c (solid) and t_{50} (dashed) according to Eq. (4.12). Lower set of curves ("Bayesian part", thin lines) is only the second term of Eq. (4.12). (b): relative entropy of structural parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ with increasing number of samples, similar to the *Information Yield Curves* according to *de Barros et al.* [2009].

Figure 4.7a (*total*) shows the expected prediction variance of concentration according to equation (4.12), recorded during the sequential placement of near-optimal sampling locations. Later modifications during to the exchange stage lead to rather minute changes in the patterns, and improved the expected prediction variance typically below one percent.

Across all cases, the planned sampling at 24 borehole locations reduce prediction uncertainties to between 50 and 70 percent of the initial uncertainty. The most effective samples are, of course, the first few ones that occupy the most informative locations. Samples placed later are displaced to less informative locations or suffer from redundancy of information if placed close by. The shape of the curves may suggest a residual asymptotic value of prediction variance that cannot be eliminated. This is indeed the case for imperfect measurements: even when exhaustively sampling the entire domain, the erroneous character of observation would still prevent a deterministic description of the system. Figure 4.7b will be addressed in section 4.7.3

4.7.2 Effect of Structural Uncertainty on Sampling Patterns

To illustrate the impact of structural uncertainty on near-optimal designs, we repeated the same analysis as above, but using known structural parameters this time. We then compare the resulting sampling patterns in Figure 4.8 (left column) and the respective sampled lag distances (Figure 4.8, right column).

The Bayesian approach to structural uncertainty honors the need for model identification. Geostatistical model identification generally leads to a diversification of sampled lag distances [*Diggle and Lophaven*, 2006; *Müller*, 2007]. The structural parameters $\boldsymbol{\theta} = \left[\sigma_Y^2,\, \lambda_1,\, \lambda_2,\, \kappa\right]$ require lag distances where they have the strongest impact on the covariance function (Figure 4.1). This information need appears in equation (4.12) as the derivatives of covariance functions with respect to structural parameters. For the global mean and variance, uncorrelated samples at great spacing are best, while covariance shape and scale additionally require a variety of low- to intermediate-range lags [*Bogaert and Russo*, 1999]. For the trend parameters, the most sensitive locations are close to the domain boundaries and corners, where the trend functions $\mathbf{X}$ have the largest impact on the

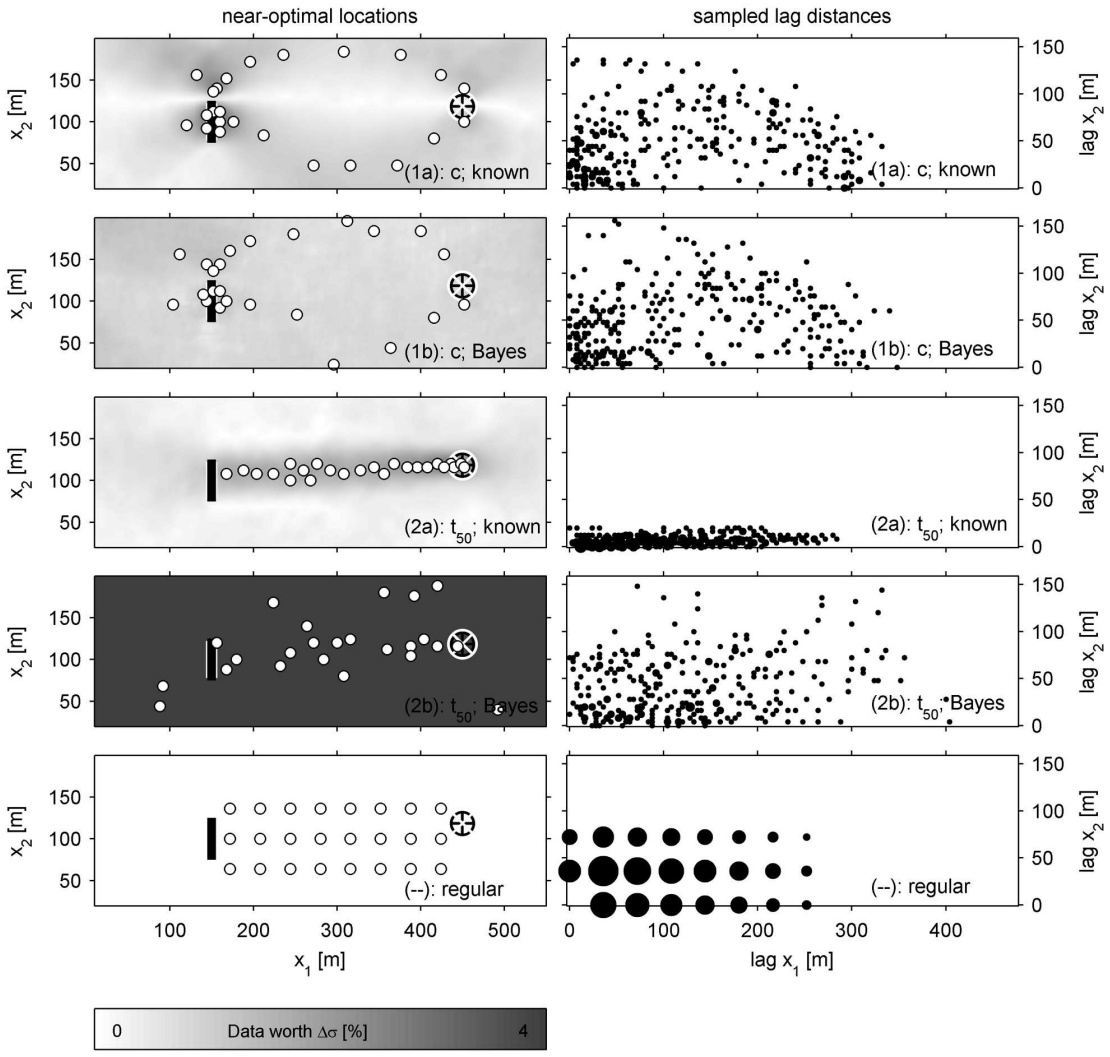


Figure 4.8: Left: Near-optimal design patterns for cases 1a-2b and a regular sampling grid. Right: respective sampled lag distances. Crossed circles (left): sensitive location. Solid white circles: 24 sampling locations; log-conductivity and hydraulic head measured jointly. Thick black line: hypothetical contaminant source. Grey-scale background: Maps of expected data worth (here: percent reduction of Bayesian predictive variance), evaluated before the first sample. Black dots (right): sampled lag distances. Dot area increases with multiple sampling of the same lag. Zero lag is not shown.

expected value of $Y = \ln K$.

We first compare cases 1a and 1b, i.e., at the patterns optimized for minimal concentration variance. The pattern for case 1a (with known structural parameters) does already offer a variety of lag distances, so that the patterns in case 1a and 1b do not differ much. Minor changes include a better coverage of long lag distances, which help to better identify the trend components.

This is drastically different between cases 2a and 2b. The pattern for case 2a is extremely narrow in the x_2-direction, and therefore does not support inference of the transverse trend or the transverse integral scale. Also, the samples are highly correlated due to their proximity along a single line, so that identification of the mean and variance are compromised. For these reasons, the pattern for case 2b is substantially different, offering a much wider range of lag distances for better identification of covariance parameters, and samples closer to the corners of the domain for better identification of the trend coefficients.

4.7.3 Effect of Sampling on Structural Uncertainty

The randomly generated structural parameters used to generate the synthetic reality, see Figure 4.4, are provided in Table 4.4. The table also provides prior mean values and posterior mean values after conditioning to the synthetic data from case 1b according to Eq. (4.13). Given the relatively small number of measurements and their level of measurement error, most structural parameters have been estimated very well.

The right half of Figure (4.7b) shows how structural uncertainty (measured by information entropy) decreases with increasing number of samples placed. We approximate the entropy

difference by [*Nowak*, 2009a]:

$$\Delta E\left(\boldsymbol{\beta},\boldsymbol{\theta}\right) \;=\; \det\left[\mathbf{C}_{\boldsymbol{\beta\theta}|\mathbf{y}}\mathbf{C}_{\boldsymbol{\beta\theta}}^{-1}\right]^{\frac{1}{d}} \tag{4.17}$$

where $\mathbf{C}_{\boldsymbol{\beta\theta}|\mathbf{y}}$ is the joint conditional covariance matrix of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, $\mathbf{C}_{\boldsymbol{\beta\theta}}$ is its prior version, and d is the total number of structural parameters. Apart from a sign flip, the same curves are a variation of the so-called *Information Yield Curves* by *de Barros et al.* [2009] (see details in Chapter 3 of the current dissertation). These curves illustrate how Bayesian Geostatistical Design considers and utilizes the potential of planned data to narrow down structural uncertainty and identify the geostatistical model (guideline 2).

The *Information Yield Curves* in Figure (4.7b) are less smooth than the prediction variance curves in Figure (4.7a), because model identification is not the primary or ultimate goal of Bayesian geostatistical design. The ultimate goal is confident prediction (here measured by prediction variance), and model identification is merely an implicit sub-goal to gain prediction confidence. Eq. (4.9) contains this sub-goal in a natural manner, and hence does not require a user-defined (and hence subjective) ranking between prediction and model identification (see guideline 3 in the introduction).

4.7.4 Cross-Case Validation, Robustness and Regular Sampling Grid

As final validation of design robustness, each near-optimal design pattern is applied to the conditions of all other test cases. We then scaled all performances (reduction of prediction variance) by the performance of the pattern that was designed for each specific case. This yields the performance indices summarized in Table 4.5. A pattern performing at 50%, for example, reduces the prediction variance only half as good as the pattern designed for the respective case.

Parameter			prior mean (and 95% CI)	synthetic values	posterior mean (and 95% CI)
global mean	β_1	$[-]$	-9.32 ($\pm$ 2)	-9.98	-9.50 ($\pm$ 0.14)
trend x_1	β_2	$[-]$	0 ($\pm$ 2)	+2.16	+2.24 ($\pm$ 0.23)
trend x_2	β_3	$[-]$	0 ($\pm$ 2)	-1.11	-0.39 ($\pm$ 0.93)
variance	σ_Y^2	$[-]$	1.00 ($\pm$ 1.41)	0.62	0.71 ($\pm$ 0.53)
integral scale λ_1	λ_1	$[m]$	15.00 ($\pm$21.21)	21.53	21.49 ($\pm$15.00)
integral scale λ_2	λ_2	$[m]$	15.00 ($\pm$21.21)	29.63	24.92 ($\pm$15.06)
shape parameter	κ	$[-]$	2.50 ($\pm$ 3.53)	3.89	2.12 ($\pm$ 3.36)

Table 4.4: Comparison of structural parameters: prior mean, synthetic reality and posterior mean values identified with synthetic data from case 1b. 95% confidence intervals are estimated from two times the posterior standard deviation, assuming a Gaussian distribution.

Of course, each sampling pattern performs best when applying it to the respective case it was designed for, surpassing all other patterns. In the cross-comparison, pattern 1b outperforms pattern 1a when applied to the respective other test case (see 4.5, first two rows). In other words, the under-achievements when designing for structural uncertainty are smaller than the under-achievements when falsely pretending a known structure during the design procedure. Quite contrarily, pattern 2a outperforms pattern 2b. The reason is that pattern 2b is adapted to the high impact of structural uncertainty onto the prediction objective, and is almost dominated by the requirements of model identification. If the geostatistical model is known (as in case 2a), most of the sampling effort of pattern 2b is spent uselessly on the diversification of lag distances. To obtain designs that are primarily robust under variation of geostatistical structure, a different objective function would have to be used, where averaging over structural parameters is performed, but all model identifica-

tion terms are omitted.

Section 4.5.4 indicated that some structural parameters do not contribute to one or both of the prediction variances discussed here. One may now ask why to consider a seemingly irrelevant geostatistical parameter as uncertain. A discussion on the role of trend parameters in the prediction of concentration is given as an example. The trends add to the variability of both log-conductivity and hydraulic head. Without properly de-trended data, the data values would falsely be interpreted towards a larger overall variance σ_Y^2, resulting in false interpretation of the data. In similar fashions, any unjustified assumption or mis-specification of geostatistical structure may introduce spurious error into data interpretation, and hence into either spatial interpolation or into the estimation of other structural parameters. In conclusion, even seemingly irrelevant structural parameters should be accounted for, thus providing robustness against mis-specified geostatistical models (guideline 4 listed in the introduction).

For additional illustration, reference and comparison to simplistic designs, we also tested a regular sampling grid with 24 sampling locations placed on a 8×3 grid with $40m \times 36m$ distance in x_1 and x_2 directions, centered between the source and target location, and aligned with the center of the source (Figure 4.8, bottom row). The regular grid is clearly defeated in all cases. Neither can it provide detailed information on the release conditions, nor does it cover the variety of lag distances to identify the structural parameters, nor does it focus on the process-specific most sensitive regions of the domain.

	case 1a	case 1b	case 2a	case 2b
pattern 1a	100%	60%		
pattern 1b	95%	100%		
pattern 2a			100%	98%
pattern 2b			68%	100%
regular grid	59%	75%	48%	96%

Table 4.5: Performance index of different patterns in different cases

4.8 Summary and Conclusions

This study transferred the concept of Bayesian Geostatistical Design to geostatistical inverse problems. Bayesian Geostatistical Design was introduced just recently by *Diggle and Lophaven* [2006]. Like other geostatistical design techniques, it optimizes site investigation or monitoring plans (called designs) for contaminated sites, while accounting for heterogeneous subsurface parameters as geostatistical random space functions. The optimal design is defined to achieve a minimal prediction uncertainty with respect to a given prediction objective.

In contrast to conventional techniques, Bayesian Geostatistical Design allows for uncertainties in the geostatistical model itself. Uncertainties in the geostatistical model may include uncertain mean values, uncertain trend coefficients, uncertain choices of covariance models, and uncertain parameters within the covariance model, all summarized under the term of structural uncertainty.

In realistic situations of site investigation, initial information on geostatistical model parameters such as the variance or integral scale of logconductivity is extremely scarce. This makes it

illegitimate to assume fixed values *a priori*, and forces to treat them as uncertain. Otherwise, overly optimistic small levels of uncertainty would be specified, and the design would be optimized under unjustified (and possible false) assumptions. It is argued that, under these premises, an adequate optimal design technique should fulfill four guidelines:

1. Uncertain structural parameters have a significant impact on prediction uncertainty. Their uncertainty must be assessed and accounted for accurately.

2. Sampling helps to reduce structural uncertainty. This potential has to be accounted for and utilized in finding a sampling design.

3. The objective of reducing structural uncertainty (i.e., geostatistical model identification) should be ranked versus the primary design objective in an optimal and natural manner.

4. Designs are sensitive to structural assumptions. Therefore, optimal designs should be robust with respect to estimation errors in structural parameters.

It is shown that Bayesian Geostatistical Design indeed reduces the number or *a priori* assumptions on geostatistical structure, and also fulfills the above four guidelines. The only remaining assumptions are that the variability of the site can be described by a reasonably parametric geostatistical model (regardless of its parameter values). However, several different parametric models may cover different parts of the domain, and there is little restriction to the complexity of the parametric models.

A key point is minimum arbitrariness when choosing a covariance model prior to sampling. To this end, we used the Matérn family of geostatistical covariance models. It offers an additional shape parameter, and includes the exponential, Whittle and Gaussian covariance function

as special cases. This way, as suggested by *Zhang and Rubin* [2009] and indicated earlier by [*Feyen et al.*, 2003], the problem of model selection becomes a problem of parameter estimation, with a wide range of methods available. We treat the shape parameter as yet another uncertain structural parameter, providing seamless integration of model uncertainty into the optimal design framework. This approach is called *Continuous Bayesian Model Averaging* because it is the limiting case of Bayesian Model Averaging over a continuous parametrized spectrum of models.

In a series of test cases, we demonstrated how structural uncertainty influences the optimal design. The test scenario featured the placement of 24 co-located hydraulic head and logconductivity measurements, optimized for minimal prediction variance of (1) contaminant concentration and (2) arrival time of contamination at an environmentally sensitive location. Structural uncertainty was represented by an uncertain global mean, uncertain coefficients of a linear trend model, and the Matérn covariance function with uncertain shape, variance and anisotropic integral scales. A variation of the test cases considered the structural parameters to be known for comparison.

Only a few samples placed optimally were sufficient to largely eliminate the additional uncertainty stemming from structural uncertainty. The list of uncertain structural parameters was shown to leave a distinct diversification in the fingerprint of the spatial pattern of the resulting optimal sampling layouts. The required diversification showed most clearly in the lag distances covered by the individual sampling patterns. The results of the test case positively confirmed that Bayesian Geostatistical Design fulfills our four guidelines listed above.

Within the risk assessment application context, Bayesian Geostatistical Design aligns well with the *Triad* principle of site investigation suggested by the USEPA [*Crumbling*, 2001]. *Triad* is an approach to decision-making for contaminated sites that offers a technically defensible methodology

for managing decision uncertainty by incorporating characterization tools and strategies. The *Triad* refers to three primary components: systematic planning, dynamic work strategies and real-time measurement systems. The *Triad* principle argues that information from ongoing site investigation should provide immediate feedback to adjust the sampling campaign in real-time, by continuously updating the site's conceptual model during the ongoing investigation effort. Bayesian Geostatistical Design extends the *Triad* principle from mere hydrogeological conceptualization to geostatistical conceptualization.

It is important to emphasize that the Bayesian Geostatistical Design framework is not in any way limited to the implementational choices taken in our study. The implementation used a first-order approximation for structural uncertainty, Ensemble Kalman Filters, and a sequential exchange optimization algorithm, yielding at least Pareto optimal designs. We are aware of the limited range of validity of first-order approximations and generality is not claimed within the results obtained in the illustrative test case. With willingness to accept substantially increased computational costs, our approximations can be replaced with brute-force Monte-Carlo or particle filter techniques, combined with genetic or simulated annealing optimization algorithms.

Notation

Symbols and their respective units:

b: Depth of the aquifer [L]

B_k: Modified Bessel function of the third kind of order k

c: Resident concentration [M/L^3]

$\tilde{c}$: Bayesian mean of the prediction goal (concentration or arrival time)

$\hat{c}$: Conditional prediction mean

C_Y, C_{YY}, $\mathbf{C}$: Spatial covariance of the log-conductivity

$\mathbf{C}_{ss}$, $\mathbf{C}_{\beta\beta}$: Covariance matrix

CV: Coefficient of variation[-]

$\mathbf{D}_d$: Dispersion tensor [L^2/t]

D_m: Molecular diffusion [L^2/t]

d: Number of structural parameters in the information entropy equation, see Eq. (4.17)

$E_a[\cdot]$: Expected value operator over the distribution of a random variable a

$\mathbf{F}$: Fisher information

$\mathbf{f}_y(s)$: Process model (such as groundwater flow equation)

$\mathbf{f}_c(s)$: Process model (such as transport equation)

$\mathbf{G}_{yy}$: Covariance matrix

h, $\hat{h}$: Hydraulic head [L]

$\mathbf{H}_c$: Sensitivity matrix

K_i: Hydraulic conductivity at a specific location $\mathbf{x}_i$ [L/t]

K_G: Geometric mean of the hydraulic conductivity [L/t]

$\mathbf{K}(\mathbf{x})$: Hydraulic conductivity at a generic location $\mathbf{x}$ [L/t]

ℓ: Lag distance [-]

ℓ_S: Dimension of the contaminant cloud in the x_2-direction [L]

L_1, L_2: Domain size [L]

m_Y mean of the log conductivity [-]

m_k: k^{th} temporal moment

N: Number of locations sampled and number of samples [-]

n_s: Vector length [-]

n_e: Porosity [-]

$p(\cdot)$: Probability density function (PDF)

$\tilde{p}(\cdot)$: Bayesian probability density function (PDF)

p: Number of trend functions

Pe_ℓ, Pe_t: Péclet number in longitudinal and transverse directions

$\mathbf{s}$: Vector of logconductivity

T: Transmissivity [L^2/t]

t_{50}: Arrival time derived from the moment generating equations [t]

$\mathbf{v}$: Velocity vector with components V_i for i = 1 and 2 [L/t]

$V_a[\cdot]$: Variance operator over the distribution of a random variable a

$\mathbf{X}$: $n_s \times p$ matrix

$\mathbf{x}$: Cartesian coordinate system [L]

$\mathbf{x}_m$: Measurement location coordinates [L]

$\mathbf{y}$, $\mathbf{y}_o$: Vector of measurements at locations $\mathbf{x}_m$

Y: Logarithm of the hydraulic conductivity ($\ln K$)

α: Ratio between entropies [-]

α_ℓ, α_t: Dispersivities in the x_1 and x_2 direction [L^2]

β: Trend coefficients

β^*: Expected value of the trend coefficients

δ: Dirac delta

Δ_i: Grid size in the i^{th} [L]

Δh: Head difference [-]

ΔE: Entropy difference [-]

$\boldsymbol{\varepsilon}_r$ Measurement error

$\boldsymbol{\varepsilon}_s$ Zero-mean fluctuations

$\Gamma(\cdot)$: Gamma function

λ_i: Correlation heterogeneity length of the aquifer in the i^{th} direction [L]

ϕ: Task-specific measure of prediction uncertainty

κ: Matérn shape parameter ($\kappa \geq 0$)

σ_Y^2: Variance of the logconductivity [-]

σ_c^2: Concentration variance [-]

σ_{t50}^2: Arrival time variance [-]

σ_r^2: Gaussian measurement error variance [-]

$\tilde{\sigma}_{c|\mathbf{y}}^2$: Increased variance conditional on $\mathbf{y}$ (Bayesian variance)

$\boldsymbol{\theta}$: Hydrogeological structural parameter vector

$\bar{\theta}$: Prior mean for structural parameter vector

$\hat{\theta}$: Conditional mean for structural parameter vector

ξ, η: Dimensionless longitudinal and transverse directions (x_i/λ_i)

ζ: Dimensionless source dimension (ℓ_s/λ)

Chapter 5

Summary

An approach to investigate the relative impact of uncertainty reduction from hydrogeological and physiological parameters in human health risk estimates is addressed. An important aspect of the approach used in this dissertation is that it unifies in a single framework all the major sources of uncertainties within a human health risk context (including parametric and model uncertainty). The results presented here illustrate the importance of considering uncertainty trade-offs in order to set priorities towards data acquisition efforts. In addition, it is highlighted how characterization needs vary within a task-oriented objective. One of the main messages of this work is to show how subsurface characterization efforts, as well as the design of sampling networks, are dependent on the prediction goal (say human health risk, concentration estimates or travel times).

In Chapter 2, a simple, yet general approach for addressing relative impacts of uncertainty reduction in human health risk is presented. The stochastic framework presented accounts for uncertainties and variabilities present in hydrogeology, human behavioral and physiological parameters. Lagrangian theory was applied to solve flow and transport analytically. Based on the

Lagrangian formulation, temporal moments of total solute mass flux were obtained. Consequently, these temporal moments were used to to derive a closed-form CDF for human health risk in terms of the relevant physical and health-related parameters. The impact of additional measurements of hydraulic conductivity on the increased cancer risk CDF was investigated. In addition to this, a single metric (α) that allows one to investigate trade-offs between sources of uncertainty was introduced. This metric is based on the concept of information entropy and was applied in a graphical approach to measure the relative impact of uncertainty reduction from flow physics and physiology.

The results in Chapter 2 show how the effect of uncertainty arising from human physiological parameters decreases as the distance between the contaminant source and the control plane increases. It was also observed that the impact of hydrogeological parametric uncertainty increases for larger distances between the control plane and contaminant source. Also, the interplay between contaminant exposure duration and hydrogeological site characterization was investigated. Results indicate that hydrogeological site characterization becomes dependent on the time the contaminant plume takes to cross the control plane if the concentration averaged over the exposure duration period is used to evaluate risk CDF, $F_R(r)$. Again, this result highlights how characterization needs should be task-oriented.

Chapter 3 focuses on the significance of flow and transport scales in defining characterization needs based on a task-oriented analysis. Again, the relative gain of information in human health risk was quantified through uncertainty reduction from both physiology and flow physics. The role of the plume's dimension proved important in defining characterization needs within the risk-driven context. Results in Chapter 3 show that uncertainty reduction in human health risk benefits more from hydrogeological site characterization if the contaminant source is small relative to the

heterogeneity correlation length. The human health risk CDF is less sensitive to measurements of hydraulic conductivity if the contaminant source increases relative to the heterogeneity correlation scale. In addition to this, it is highlighted how the value of information not only depends on the plume's dimension but also on its interplay with the scale of capture zone induced by the action of pumping wells. For higher pumping rates (thus larger capture zones), the value of hydrogeological characterization becomes less dependent of the plume's dimension.

Chapter 3 also emphasizes how uncertainty reduction in risk may benefit more from parametric uncertainty reduction from the health component as opposed to hydrogeological if the plume's dimension approaches ergodicity. The role of pore-scale dispersion is also addressed. For high Peclet conditions, plume-size relative to the heterogeneity scale is an important factor to be considered in defining characterization efforts. The contrary occurs for low Peclet conditions: information concerning the size of the plume relative to the heterogeneity scale becomes less relevant towards defining subsurface characterization strategies. Analogous observations were obtained when comparing concentration measured in a well versus the flux-averaged concentration at a control plane. The manner in which contaminated water is sampled has a strong influence in defining characterization needs within a risk-based approach as demonstrated in Chapter 3. It was also shown how different physiological dose-response models have different effects in risk uncertainty reduction and in defining characterization needs. One of the highlights of Chapter 3 is in extending the ideas presented in Chapter 2 to construct the concept of *comparative information yield curves*. Theoretical, methodological and practical aspects of the *comparative information yield curves* were given. For this work, these curves proved useful since it allows one to easily view the relative contribution of information in risk from the physiological and the hydrogeological component.

Chapter 4 explores in further detail the role of hydrogeological parametric uncertainty in reducing concentration and travel time variances at an environmentally sensitive target. The concept of Bayesian Geostatistical Design was transferred to geostatistical inverse problems. It was shown through a series of test cases how parametric uncertainty within the geostatistical model influences the optimal design. These results were compared to test cases with no parametric uncertainty. Measurements of head and logconductivity were optimized for minimal prediction variance of (i) contaminant concentration and (ii) contaminant travel time. In addition, it is highlighted how different objective functions lead to different sampling design. Furthermore, the Matérn family of geostatistical covariance models was used since it offers an additional shape parameter. For specific values of this shape parameter, the Matérn covariance model assumes the form of the classical geostatistical models (for example: exponential, Whittle and Gaussian covariance function) commonly used in the literature, see Chapter 2 of *Rubin* [2003]. This allowed conversion of the problem of model selection to a problem of parameter estimation. The shape parameter present in the Matérn covariance model was treated as yet another uncertain structural parameter. This approach of converting a model selection problem into a parameter selection problem is denoted as *Continuous Bayesian Model Averaging* since it is the limiting case of the Bayesian Model Averaging [*Hoeting et al.*, 1999; *Neuman*, 2003] over a continuous parameterized spectrum of models. The results in Chapter 4 show how important it is to account for parametric uncertainty in subsurface characterization and how the sampling patterns change drastically depending on the task to be minimized. Chapter 4 also explains the relevance of applying the Bayesian Geostatistical Design framework in monitoring contaminated sites and consequently evaluating potential human health risk.

References

Abramowitz, M., and I. A. Stegun (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*, Dover, New York.

Andricevic, R., and V. Cvetkovic (1996), Evaluation of Risk from Contaminants Migrating by Groundwater, *Water Resources Research*, *32*(3), 611–621.

Andricevic, R., and V. Cvetkovic (1998), Relative dispersion for solute flux in aquifers, *Journal of Fluid Mechanics*, *361*, 145–174.

Andricevic, R., J. Daniels, and R. Jacobson (1994), Radionuclide migration using travel time transport approach and its application in risk analysis, *Journal of Hydrology*, *163*, 125–145.

Ashby, S. F., and R. D. Falgout (1996), A parallel multigrid preconditioned conjugate gradient algorithm for groundwater flow simulations, . *Nuclear Science and Engineering*, *124*, 145–159.

Beckie, R. (1996), Measurement scale, network sampling scale, and groundwater model parameters, *Water Resources Research*, *32*(1), 65–76.

Bellin, A., A. Rinaldo, W. Bosma, S. E. van der Zee, and Y. Rubin (1993), Linear Equilibrium

Adsorbing Solute Transport in Physically and Chemically Heterogenous Porous Formations 1: Analytical Solutions, *Water Resources Research*, *29*(12), 4019–4030.

Benekos, I., C. A. Shoemaker, and J. R. Stedinger (2007), Probabilistic risk and uncertainty analysis for bioremediation of four chlorinated ethenes in groundwater, *Stochastic Environmental Research Risk Assessment*, *21*, 375–390.

Binkowitz, B., and D. Wartenberg (2001), Disparity in quantitative risk assessment: A review of input distributions, *Risk Analysis*, *21*(1).

Bogaert, P., and D. Russo (1999), Optimal spatial sampling design for the estimation of the variogram based on a least squares approach, *Water Resour. Res.*, *35*(4), 1275–1289.

Bogen, K. T., and R. C. Spear (1987), Integrating Uncertainty and Interindividual Variability in Environmental Risk Assessment, *Risk Analysis*, *7*(4), 427–436.

Burgers, G., P. J. V. Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble kalman filter, *Monthly Weather Review*, *126*, 1719–1724.

Burmaster, D., and A. Wilson (1996), An introduction to second-order random variables in human health risk assessments, *Human and Ecological Risk Assessment*, *2*(4), 892–919.

Caroni, E., and V. Fiorotto (2005), Analysis of concentration as sampled in natural aquifers, *Transport in Porous Media*, *59*, 19–45, doi:10.107/s11,242–004–1119–x.

Chen, Y., and D. Zhang (2006), Data assimilation for transient flow in geologic formations via ensemble kalman filter, *Adv. Water Resour.*, *29*, 1107–1122.

Chiu, W., C. Chen, K. Hogan, J. Lipscomb, C. Scott, and R. Subramaniam (2007), High-to-low dose extrapolation: Issues and approaches, *Human and Ecological Risk Assessment*, *13*, 46–51.

Christakos, G. (1992), *Random field models in earth sciences*, first ed., Academic Press.

Cirpka, O. A., and P. K. Kitanidis (2000), Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments, *Water Resour. Res.*, *36*(5), 1221–1136.

Cirpka, O. A., and W. Nowak (2004), First-order variance of travel time in non-stationary formations, *Water Resour. Res.*, *40*, doi:10.1029/2003WR002,851.

Cirpka, O. A., C. M. Bürger, W. Nowak, and M. Finkel (2004), Uncertainty and data worth analysis for the hydraulic design of funnel-and-gate systems in heterogeneous aquifers, *Water Resour. Res.*, *40*(W11502), doi:10.1029/2004WR003,352.

Criminisi, A., T. Tucciarelli, and G. P. Karatzas (1997), A methodology to determine optimal transmissivity measurement locations in groundwater quality management models with scarce field information, *Water Resour. Res.*, *33*(6), 1265–1274.

Crumbling, D. (2001), Using the triad approach to improve the cost-effectiveness of hazardous waste site cleanups, *Tech. Rep. EPA 542-R-01-016*.

Cushey, M., and Y. Rubin (1997), Field-scale transport of nonpolar organic solutes in 3-D heterogenous aquifers, *Environmental Science and Technology*, *31*(5), 1259–1268.

Cvetkovic, V., and G. Dagan (1994), Transport of kinetically sorbing solute by steady random velocity in heterogenous porous formations, *Journal of Fluid Mechanics*, *265*, 189–215.

Cvetkovic, V., A. Shapiro, and G. Dagan (1992), A solute flux approach to transport in heterogenous formations 2: Uncertainty Analysis, *Water Resources Research*, *28*(5), 1377–1388.

Cvetkovic, V., G. Dagan, and H. Cheng (1998), Contaminant Transport in Aquifers with Spatially Variable Hydraulic and Sorption Properties, *Proc. R. Soc. London A*, *454*, 2173–2207.

Dagan, G. (1984), Solute Transport in Heterogenous Porous Formations, *Journal of Fluid Mechanics*, *145*, 151–177.

Dagan, G. (1985), A note on higher-order corrections of the head covariances in steady aquifer flow, *Water Resour. Res.*, *21*(4), 573–578.

Dagan, G. (1987), Theory of Solute Transport by Groundwater, *Annual Review of Fluid Mechanics*, *19*, 183–215.

Dagan, G. (1989), *Flow and Transport in Porous Formations*, Springer Verlag, Berlin.

Dagan, G., and S. Neuman (1997), *Subsurface Flow and Transport: A Stochastic Approach*, first ed., Cambridge University Press.

Dagan, G., and V. Nguyen (1989), A Comparison of Travel Time and Concentration Approaches to Modeling Transport by Groundwater, *Journal of Contaminant Hydrology*, *4*, 79–91.

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

Daniels, J., K. Bogen, and L. Hall (2000), Analysis of uncertainty and variability in exposure to characterize risk: Case study involving trichloroethylene groundwater contamination at Beale Air Force Base in California, *Water, Air and Soil Pollution*, *123*, 273–298.

Davis, T. A. (2004), Algorithm 832: UMFPACK v.4.3 - an unsymmetric-pattern multifrontal method, *ACM Trans. Math. Software*, *30*(2), 196–199.

Dawoud, E., and S. Purucker (1996), Quantitative Uncertainty Analysis of Superfund Residential Risk Pathway Models for Soil and Groundwater: White Paper, *Tech. rep.*

de Barros, F. P. J., and Y. Rubin (2008), A Risk-Driven Approach for Subsurface Site Characterization, *Water Resources Research*, *44*, doi:10.1029/2007WR006,081.

de Barros, F. P. J., Y. Rubin, and R. Maxwell (2009), The concept of comparative information yield curves and their application to risk-based site characterization, *Water Resources Research*, *In Press*, doi:10.1029/2008WR007,324.

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

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

Diggle, P., and S. Lophaven (2006), Bayesian geostatistical design, *Scandinavian J. Statist.*, *33*, 53–64, doi:10.1111/j.1467–9469.2005.00,469.x.

Diggle, P., and P. J. Ribeiro (2002), Bayesian inference in Gaussian model-based geostatistics, *Geogr. and Environ. Mod.*, *6*(2), 129–146.

Diggle, P. J., and P. J. Ribeiro (2007), *Model-based geostatistics*, Springer series in statistics, Springer, New York.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Research*, *99*(C5), 10,143–10,162.

Evensen, G. (2003), The ensemble kalman filter: theoretical formulation and practical implementation, *Ocean Dynamics*, *53*, 343–367, doi:10.1007/s10,236–003–0036–9.

Feinerman, E., G. Dagan, and E. Bresler (1986), Statistical inference of spatial random functions, *Water Resour. Res.*, *22*(6), 953–942.

Feyen, L., and S. M. Gorelick (2005), Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas, *Water Resour. Res.*, *41*(W03019), doi:10.1029/2003WR002,901.

Feyen, L., J. J. Gómez-Hernández, P. J. R. Jr., K. J. Beven, and F. D. Smedt (2003), A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations, *Water Resour. Res.*, *39*(5), doi:10.1029/2002WR001,544.

Fiori, A., S. Berglund, V. Cvetkovic, and G. Dagan (2002), A first-order analysis of solute flux statistics in aquifers: The combined effect of pore-scale dispersion, sampling and linear sorption kinetics, *Water Resources Research*, *38*(8), 1–15.

Fiorotto, V., and E. Caroni (2002), Solute concentration statistics in heterogeneous aquifers for finite pclet values, *Transport in Porous Media*, *48*, 331–351.

Fjeld, R., N. Eisenberg, and K. Compton (2007), *Quantitative Environmental Risk Analysis for Human Health*, first ed., Wiley.

Fletcher, C. A. J. (1996), *Computational Techniques for Fluid Dynamics, Vol. 1: Fundamental and General Techniques*, Springer Series in Computational Physics, 2nd ed., Springer Verlag Telos, New York.

Freeze, A., J. Massmann, L. Smith, T. Sperling, and B. James (1990), Hydrogeological Decision Analysis: 1. A framework, *Ground Water*, *1*, 738–766.

Fritz, J., W. Nowak, and I. Neuweiler (2009, in press), Application of FFT-based algorithms for large-scale universal Kriging problems, *Math. Geosciences*, pp. 10.1007/s11,004–009–9220–x.

Gelhar, L. W. (1993), *Stochastic Subsurface Hydrology*, first ed., Prentice Hall.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Longman Publishing Co.., Inc., Boston, MA, USA.

Handcock, M. S., and M. L. Stein (1993), A Bayesian analysis of Kriging, *Technometrics*, *35*(4), 403–410.

Harvey, C. F., and S. M. Gorelick (1995), Temporal moment-generating equations: Modeling transport and mass-transfer in heterogeneous aquifers, *Water Resour. Res.*, *31*(8), 1895–1911.

Hassan, A., R. Andricevic, and V. Cvetkovic (2001), Computational issues in the determination of solute discharge moments and implications for comparison to analytical solutions, *Advances in Water Resources*, *24*, 607–619.

Hassan, A., R. Andricevic, and V. Cvetkovic (2002), Evaluation of analytical solute discharge moments using numerical modeling in absolute and relative dispersion frameworks, *Water Resources Research*, *38*(2), 1–8.

Herrera, G. S. (1998), Cost effective groundwater quality sampling network design, Ph.D. thesis, University of Vermont, Burlington.

Herrera, G. S., and G. F. Pinder (2005), Space-time optimiziation of groundwater quality sampling networks, *Water Resour. Res.*, *41*(W12407), doi:10.1029/2004WR003,626.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999), Bayesian Model Averaging: A Tutorial, *Statistical Science*, *14*(4), 382–417.

Hughes, J. R. (1987), *Finite Element Method*, Prentice Hall International, Inc.

James, B., and S. Gorelick (1994), When enough is enough: the worth of monitoring data in aquifer remediation design, *Water Resources Research*, *30*(12), 3499–3513.

Janssen, G. M. C. M., J. R. Valstar, and S. E. A. T. M. van der Zee (2008), Measurement network design including traveltime determinations to minimize model prediction uncertainty, *Water Resour. Res.*, *44*(W02405), doi:10.1029/2006WR005,462.

Jones, J., and C. Woodward (2001), Newton-Krylov-multigrid solvers for large-scale, highly heterogeneous, variably saturated flow problems, *Advances in Water Resources*, *24*, 763–774.

Journel, A. G., and C. J. Huijbregts (1978), *Mining Geostatistics*, Academic Press, New York.

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

Kitanidis, P. K. (1993), Generalized covariance functions in estimation, *Math. Geol.*, *25*(5), 525–540.

Kitanidis, P. K. (1995), Quasi-linear geostatistical theory for inversing, *Water Resour. Res.*, *31*(10), 2411–2419.

Kitanidis, P. K. (1997), *Introduction to Geostatistics*, Cambridge University Press, Cambridge.

Kitanidis, P. K., and R. W. Lane (1985), Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method, *J. Hydrol.*, *79*, 53–71.

Kollet, S., and R. Maxwell (2006), Integrated surface-groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model, *Advances in Water Resources*, *29*(7), 945–958.

Kreft, A., and A. Zuber (1978), On the physical meaning of the dispersion equation and its solution for different and initial boundary conditions, *Chemical Engineering Science*, *33*, 1471–1480.

Marchant, B. P., and R. M. Lark (2007), Optimized sample schemes for geostatistical surveys, *Math. Geol.*, *39*(1), doi: 10.1007/s11,004–006–9069–1.

Massman, J., and A. Freeze (1987), Groundwater Contamination From Waste Management Sites: The Interaction Between Risk-Based Engineering Design and Regulatory Policy, *Water Resources Research*, *23*(2), 351–367.

Matérn, B. (1986), *Spatial variation*, Springer, Berlin, Germany.

Matheron, G. (1971), *The Theory of Regionalized Variables and Its Applications*, Ecole de Mines, Fontainebleau, France.

Maxwell, R., and W. Kastenberg (1999), Stochastic Environmental Risk Analysis: An Integrated Methodology for Predicting Cancer Risk from Contaminated Groundwater, *Stochastic Environmental Research Risk Assessment*, *13*, 27–47.

Maxwell, R., S. Pelmulder, F. Tompson, and W. Kastenberg (1998), On the development of a new

methodology for groundwater driven health risk assessment, *Water Resources Research*, *34*(4), 833–847.

Maxwell, R., W. Kastenberg, and Y. Rubin (1999), A methodology to integrate site characterization information into groundwater-driven health risk assessment, *Water Resources Research*, *35*(9), 2841–2885.

Maxwell, R., C. Welty, and R. Harvey (2007), Revisiting the Cape Cod Bacteria Injection Experiment Using a Stochastic Modeling Approach, *Environmental Science and Technology*, *14*(15), 5548–5558.

Maxwell, R., S. Carle, and A. Tompson (2008), Contamination, Risk, and Heterogeneity: On the Effectiveness of Aquifer Remediation, *Environmental Geology*, *54*, 1771–1786.

McKinney, D. C., and D. P. Loucks (1992), Network design for predicting groundwater contamination, *Water Resour. Res.*, *28*(1), 133–147.

McKone, T., and T. Bogen (1991), Predicting the uncertainties in risk assessment, *Environ. Sci. Technol.*, *25*(10), 1674–1681.

Müller, W. G. (2007), *Collecting spatial data. Optimum design of experiments for random fields*, 3 ed., Springer, Berlin, Germany.

Neuman, S. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environmental Research Risk Assessment*, *17*, 291–305.

Nowak, W. (2009a), Measures of parameter uncertainty in geostatistical estimation and design, *Math. Geosciences*, (Under Review).

Nowak, W. (2009b), Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator, *Water Resour. Res.*, (doi:10.1029/2008WR007328).

Nowak, W., and O. A. Cirpka (2004), A modified Levenberg-Marquardt algorithm for quasi-linear geostatistical inversing, *Adv. Water Resour.*, *27*(7), 737–750.

Nowak, W., and O. A. Cirpka (2006), Geostatistical inference of conductivity and dispersion coefficients from hydraulic heads and tracer data, *Water Resour. Res.*, *42*(W08416), doi:10.1029/2005WR004,832.

Nowak, W., S. Tenkleve, and O. A. Cirpka (2003), Efficient computation of linearized cross-covariance and auto-covariance matrices of interdependent quantities, *Math. Geol.*, *35*(1), 53–66.

Nowak, W., R. L. Schwede, O. A. Cirpka, and I. Neuweiler (2008), Probability density functions of hydraulic head and velocity in three-dimensional heterogeneous porous media, *Water Resour. Res.*, *44*(W08452), doi:10.1029/2007WR006,383.

Pardo-Iguzquiza, E. (1999), Bayesian inference of spatial covariance parameters, *Math. Geol.*, *31*(1), 47–65.

Pardo-Iguzquiza, E., and M. Chica-Olmo (2008), Geostatistical simulation when the number of experimental data is small: an alternative paradigm, *Stoch. Environ. Res. Risk Assess*, *22*, 325–337.

Portier, L., K. Tolson, and S. Robert (2007), Body Weight Distributions for Risk Assessment, *Risk Analysis*, *27*(1).

Pukelsheim, F. (2006), *Optimal Design of Experiments*, Classics in Applied Mathematics, classic edition ed., SIAM, Philadelphia.

Reed, P., B. Minsker, and D. E. Goldberg (2000), Designing a competent simple genetic algorithm for search and optimization, *Water Resour. Res.*, *36*(12), 3757–3761.

Rubin, Y. (1991), Prediction of tracer plume migration in heterogeneous porous media by the method of conditional probabilities, *Water Resour. Res.*, *27*(6), 1291–1308.

Rubin, Y. (2003), *Applied Stochastic Hydrogeology*, first ed., Oxford Press.

Rubin, Y., and G. Dagan (1987), Stochastic identification of transmissivity and effective recharge in steady state groundwater flow. 1. Theory, *Water Resour. Res.*, *23*(7), 1185–1192.

Rubin, Y., and G. Dagan (1992), Conditional estimates of solute travel time in heterogenous formations: impact of transmissivity measurements, *Water Resources Research*, *28*(4), 1033–1040.

Rubin, Y., M. A. Cushey, and A. Bellin (1994), Modeling of transport in groundwater for environmental risk assessment, *Stochastic Hydrol. Hydraul.*, *8*(1), 57–77.

Rubin, Y., A. Sun, R. Maxwell, and A. Bellin (1999), The concept of block-effective macrodispersivity and a unified approach for grid-scale- and plume-scale-dependent transport, *Journal of Fluid Mechanics*, *395*, 161–180.

Scheidegger, A. E. (1954), Statistical hydrodynamics in porous media, *J. Appl. Phys.*, *25*, 994–1001.

Schwede, R. L., O. A. Cirpka, W. Nowak, and I. Neuweiler (2008), Impact of sampling volume on the probability density function of steady state concentration, *Water Resour. Res.*, *44*(W12433), doi:10.1029/2007WR006,668.

Shafer-Perini, A., and J. Wilson (1991), Efficient and accurate front tracking for two-dimensional groundwater flow models, *Water Resources Research*, *27*(7), 1471–1485.

Shapiro, A., and V. Cvetkovic (1988), Stochastic Analysis of Solute Arrival Time in Heteregenous Porous Media, *Water Resources Research*, *24*(10), 1711–1718.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, Berlin, Germany.

Taylor, S., L. Smith, R. Carr, A. Carson, and E. Darois (2003), Developing site specific derived concentration guideline levels for multiple media at the Connecticut Yankee Haddam Neck Plant, in *WM'03 Conference*, Tucson, Arizona, USA.

Tompson, A., R. Ababou, and L. Gelhar (1989), Implementation of the three-dimensional turning bands random field generator, *Water Resour. Res*, *25*(10), 2227–2243.

USEPA (1989), Risk Assessment Guidance for Superfund Volume 1: Human Health Manual (Part A), *Tech. Rep. Rep.EPA/540/1-89/002*.

USEPA (1991), Risk Assessment Guidance for Superfund Volume 1: Human Health Evaluation (Part B), *Tech. Rep. Rep.EPA/540/R-92/003*.

USEPA (1998), Appendix D: Dose Response Assessment in: Cleaner Technology Substitutes Assessment: Professional Fabricare Processes, *Tech. Rep. EPA 744-B-98-001*.

USEPA (2001), Risk Assessment Guidance for Superfund: Volume III - Part A, Process for Conducting Probabilistic Risk Assessment, *Tech. Rep. Rep.EPA 540/R-02/002*.

USEPA (2005), Guidelines for Carcinogen Risk Assessment, *Tech. Rep. EPA/630/P-03/001F*.

Woodbury, A. D., and T. J. Ulrych (1993), Minimum relative entropy: Forward probabilistic modeling, *Water Resour. Res.*, *29*(8), 2847–2860.

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

Wu, J., C. Zheng, and C. C. Chien (2005), Cost-effective sampling network design for contaminant plume monitoring under general hydrogeological conditions, *J. Contam. Hydrol.*, *77*(1), 41–65.

Wu, J., C. Zheng, C. C. Chien, and L. Zheng (2006), A comparative study of Monte Carlo simple genetic algorithm and noisy genetic algorithm for cost-effective sampling network design under uncertainty, *Advances Water Res.*, *29*(6), 899–911.

Zhang, D. (2002), *Stochastic Methods for Flow in Porous Media*, Academic Press, San Diego.

Zhang, Y., G. F. Pinder, and G. S. Herrera (2005), Least cost design of groundwater quality monitoring networks, *Water Resour. Res.*, *41*(W08412), doi:10.1029/2005WR003,936.

Zhang, Z., and Y. Rubin (2009), Inverse modeling of spatial random fields using anchors, *Water Resour. Res.*, *Under Review*.

Zimmerman, D. L. (2006), Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction, *Environmetrics*, *17*, 635–652.

Appendix A

Second Moment of the Solute Flux

A detailed derivation of the second moment of the solute flux, equation (2.29), is shown. Let us define a function $h(\tau)$ as:

$$h(\tau|R_f, T_o) = \{H[t - R_f\,\tau - t_o] - H[t - R_f\,\tau - t_o - T_o]\}. \tag{A1}$$

Assuming that the injected mass and the release duration are constant, the second moment is given by:

$$\langle Q^2(t, \tau|L, \mathbf{a}_o, T_o, t_o, M_o, R_f, \boldsymbol{\theta}_H, \{m\})\rangle = \frac{M_o^2}{T_o^2}\int_0^\infty\int_0^\infty h(\tau|R_f, T_o)h(\tau'|R_f, T_o)g_2(\tau, \tau'|L, \mathbf{a}_o, \mathbf{a}, t_o, \boldsymbol{\theta}_H, \{m\})d\tau d\tau' \tag{A2}$$

Inserting equation (2.28) into (A2) and with the aid of the properties of the Dirac delta, we obtain:

$$\begin{aligned}
&\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \\
&\quad \frac{M_o^2}{T_o^2}\int_0^\infty\int_0^\infty h(\tau|R_f,T_o)h(\tau'|R_f,T_o)g_1(\tau'|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})\delta(\tau-\tau')d\tau d\tau' \\
&\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \\
&\quad \frac{M_o^2}{T_o^2}\int_0^\infty h(\tau|R_f,T_o)h(\tau|R_f,T_o)g_1(\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})d\tau \\
&\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \\
&\quad \frac{M_o^2}{T_o^2}\int_0^\infty h^2(\tau|R_f,T_o)g_1(\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})d\tau. \qquad \text{(A3)}
\end{aligned}$$

Inserting equation (A1) into (A3):

$$\begin{aligned}
&\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \\
&\quad \frac{M_o^2}{T_o^2}\int_0^\infty \{H[t-R_f\,\tau-t_o]-H[t-R_f\,\tau-t_o-T_o]\}^2 g_1(\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})d\tau.
\end{aligned} \qquad \text{(A4)}$$

Now, we recall the properties of the Heaviside function. Note that since the values $H(\cdot)$ can take are either 1 or 0, the squared value of the Heaviside function can be neglected and equation (A4) can be further simplified:

$$\begin{aligned}
&\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \\
&\quad \frac{M_o^2}{T_o^2}\int_0^\infty \{H[t-R_f\,\tau-t_o]-H[t-R_f\,\tau-t_o-T_o]\} g_1(\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})d\tau,
\end{aligned} \qquad \text{(A5)}$$

where equation (A5) can be re-written in a more compact manner:

$$\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = \frac{M_o^2}{T_o^2}\int_A^B g_1(\tau|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\})\,d\tau',$$

with A and B defined in equation (2.25 and 2.26). Expression (A6) can be written in terms of the travel time cumulative distribution function as in equation (2.29):

$$\begin{aligned}\langle Q^2(t,\tau|L,\mathbf{a}_o,T_o,t_o,M_o,R_f,\boldsymbol{\theta}_H,\{m\})\rangle = &\frac{M_o^2}{T_o^2}\,G_\tau\left(B|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right)\\ &-\frac{M_o^2}{T_o^2}\,G_\tau\left(A|L,\mathbf{a}_o,t_o,\boldsymbol{\theta}_H,\{m\}\right). \qquad \text{(A6)}\end{aligned}$$

Appendix B

Estimating the Probability Density Function of $\boldsymbol{\theta}_H$

Given a set $\mathbf{Y} \equiv \{m\}$ consisting of $Y_i = \ln K_i$ measurements from a random field generator we are able to estimate a PDF for the uncertain parameter (with i=1, . . . , N, where N is the total number of measurements). From this sample, we obtain the SRF parameters, for example, $\boldsymbol{\theta}_H = \{m_Y, \sigma_Y^2, \lambda\}$ where m_Y and σ_Y^2 are the mean and variance of Y and λ is its correlation length. We need to infer the distribution of $\boldsymbol{\theta}_H$ given the measurements in $\mathbf{Y}$, $\hat{f}_H(\boldsymbol{\theta}_H|\mathbf{Y})$. The procedure is based on Bayes Theorem:

$$\hat{f}_H(\boldsymbol{\theta}_H|\mathbf{Y}) = \frac{f_{prior}(\boldsymbol{\theta}_H) f_Y(\mathbf{Y}|\boldsymbol{\theta}_H)}{f_Y(\mathbf{Y})}, \tag{B1}$$

where the assumption of a prior PDF, $f_{prior}(\boldsymbol{\theta}_H)$ is needed. Assuming that the PDF $f_Y(\mathbf{Y}|\boldsymbol{\theta}_H)$ is multivariate Gaussian, we have:

$$f_Y(\mathbf{Y}|\boldsymbol{\theta}_H) = \frac{1}{(2\pi)^{N/2}\|\mathbf{C_{YY}}\|} \exp\left[-\frac{1}{2}(\mathbf{Y}-\mathbf{m}_Y)^T \mathbf{C}_{\mathbf{YY}}^{-1}(\mathbf{Y}-\mathbf{m}_Y)\right], \tag{B2}$$

where $\mathbf{C_{YY}}$ is the geostatistical correlation model that depends on $\boldsymbol{\theta}_H = \{m_Y, \sigma_Y^2, \lambda\}$ and $\|\mathbf{C_{YY}}\| \equiv \det(\mathbf{C_{YY}})$. With equation (B2), the estimated PDF in equation (B1) can be obtained.

Appendix C

Flow and Transport Formulation and Numerical Implementation used in Chapter 3

A two-dimensional depth-averaged, saturated, steady-state flow is considered. The flow domain is considered bounded and defined by the aquifer's longitudinal length L and width W. The equation that governs flow is given as follows:

$$\nabla \cdot [\, b\mathbf{K}(\mathbf{x})\nabla h \,] = \sum_{w} Q_w \delta(\mathbf{x} - \mathbf{x}_w), \tag{C1}$$

where b is the average depth of the aquifer, h the hydraulic head, Q_w is the pumping rate of the w^{th} pump well at location $\mathbf{x}_w$. We consider no-flow boundary conditions on the transversal direction (x_2) and prescribed pressure head in the longitudinal direction (x_1). Flow occurs from the left to right. Assuming instantaneous linear chemical interactions with the soil particles and that the chemical, with initial concentration C_o, is instantaneously released within the aquifer along a line

source, we write:

$$R_f \frac{\partial C}{\partial t} + \mathbf{V}\nabla C - \nabla \cdot [\mathbf{D}_d(\mathbf{x})\nabla C] = \sum_w \frac{C_w Q_w}{n_e}\delta(\mathbf{x} - \mathbf{x}_w), \tag{C2}$$

with n_e being the effective porosity and $\mathbf{V}$ is the Eulerian velocity vector obtained through Darcy's Law, R_f is the retardation factor, $\mathbf{D}_d$ is the dispersion coefficient tensor, C is the concentration and finally C_w is the concentration at the pumping well. ParFlow was used to solve the flow field in the aquifer [*Ashby and Falgout*, 1996; *Jones and Woodward*, 2001; *Kollet and Maxwell*, 2006]. ParFlow is a watershed flow code that uses a multi-grid preconditioned conjugate gradient algorithm to efficiently solve the linear system resulting from the discretization of the flow equation. The contaminant transport is solved using a Lagrangian particle tracking algorithm with very minimal numerical dispersion and conservation of mass [*Maxwell and Kastenberg*, 1999; *Maxwell et al.*, 2007]. This code, called SLIM-FAST, simulates migration of dissolved, neutrally buoyant and reactive chemical in saturated porous media. To represent concentration and the spatial/temporal distribution of the contaminant, an explicit Lagrangian Random Walk Particle Method is implemented in the code. SLIM-FAST also benefits from the quasi-analytical formulation presented in *Shafer-Perini and Wilson* [1991].

Appendix D

Maximum Likelihood Estimator used in Chapter 3

In the case where N measurements of $Y = \ln K$ are available, the negative log-likelihood function for a multivariate normal PDF becomes [*Rubin*, 2003]:

$$-\ln \mathcal{L}(\boldsymbol{\theta}_H|Y_i) = \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln\|\mathbf{C_{YY}}\| + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(Y_i - \langle Y_i\rangle)(Y_j - \langle Y_j\rangle)}{C_Y(\mathbf{x}_i, \mathbf{x}_j)}, \qquad \text{(D1)}$$

where $\|\mathbf{C_{YY}}\|$ is the determinant of the variance-covariance matrix of order N by N. $C_Y(\mathbf{x}_i, \mathbf{x}_j)$ is the spatial covariance model. For our results, we used the case of an exponential isotropic $C_Y(\mathbf{x}_i, \mathbf{x}_j)$ such that $\boldsymbol{\theta}_H = \{m_Y, \sigma_Y^2, \lambda\}$. The Maximum Likelihood estimators are those that minimize equation D1.

Appendix E

Derivation for the Linearized $\mathrm{f}_y\left(\mathrm{s}\right)$ in Chapter 4.4

We define a linearized representation for $\mathbf{f}_y\left(\mathbf{s}\right)$ in the form of:

$$\mathbf{y} = \mathbf{f}_y\left(\mathbf{s}\right) \approx E\left[\mathbf{f}\left(\mathbf{s}\right)\right] + \mathbf{H}\left(\mathbf{s} - \bar{\mathbf{s}}\right) , \tag{E1}$$

where $\bar{\mathbf{y}} = E\left[\mathbf{f}_y\left(\mathbf{s}\right)\right]$ and $\bar{\mathbf{s}}$ are the mean values of $p\left(\mathbf{y}\right)$ and $p\left(\mathbf{s}\right)$, respectively. Within the linearized framework, the relevant mean values an covariances become:

$$\mathbf{G}_{\mathbf{yy}}\left(\boldsymbol{\theta}\right) = \mathbf{H}\mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right)\mathbf{H}^T + \mathbf{R} \tag{E2}$$

$$\hat{\mathbf{s}}\left(\mathbf{y}_o, \boldsymbol{\theta}\right) = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right)\mathbf{H}^T\mathbf{G}_{\mathbf{yy}}^{-1}\left(\boldsymbol{\theta}\right)\left(\mathbf{y}_o - \bar{\mathbf{y}}\right) \tag{E3}$$

$$\mathbf{G}_{\mathbf{ss}|\mathbf{y}}\left(\boldsymbol{\theta}\right) = \mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right) - \mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right)\mathbf{H}^T\mathbf{G}_{\mathbf{yy}}^{-1}\left(\boldsymbol{\theta}\right)\mathbf{H}\mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right) \tag{E4}$$

$$\mathbf{C}_{\boldsymbol{\beta}\boldsymbol{\beta}|\mathbf{y}} = \left(\mathbf{X}^T\mathbf{H}_{\mathbf{y}}^T\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{y}}\mathbf{X} + \mathbf{C}_{\boldsymbol{\beta}\boldsymbol{\beta}}\right)^{-1} , \tag{E5}$$

where $\mathbf{G_{yy}}$ is the generalized covariance of $\mathbf{y}$, $\hat{\mathbf{s}}$ and $\mathbf{G_{ss|y}}$ are the conditional mean and generalized covariance of $\mathbf{s}$, and $\mathbf{C}_{\boldsymbol{\beta\beta}|\mathbf{y}}$ is the conditional covariance of $\boldsymbol{\beta}$.

Employing a likewise linearized representation of $c = f_c(\mathbf{s})$ with coefficient matrix $\mathbf{H}_c$, the conditional predictive distribution for c becomes:

$$\begin{aligned} \hat{c}(\mathbf{y}_o, \boldsymbol{\theta}) &= \bar{c} + \mathbf{H}_c(\hat{\mathbf{s}}(\mathbf{y}_o, \boldsymbol{\theta}) - \mathbf{X}\boldsymbol{\beta}^*) \\ \sigma^2_{c|\mathbf{y}}(\boldsymbol{\theta}) &= \mathbf{H}_c \mathbf{G}_{\mathbf{ss}|\mathbf{y}}(\boldsymbol{\theta}) \mathbf{H}_c^T . \end{aligned} \tag{E6}$$

Due to linearization, the prediction variances for known $\boldsymbol{\theta}$ are independent of data values, and Eq. (4.9) simplifies to:

$$E_{\mathbf{y}}\left[\tilde{\sigma}^2_{c|\mathbf{y}}\right] = E_{\boldsymbol{\theta}}\left[\sigma^2_{c|\mathbf{y}}(\boldsymbol{\theta})\right] + E_{\mathbf{y}}\left\{V_{\boldsymbol{\theta}|\mathbf{y}}\left[\hat{c}(\mathbf{y}(\mathbf{d}), \boldsymbol{\theta})\right]\right\} . \tag{E7}$$

Linearization of $\mathbf{f}_y(\mathbf{s})$ is exact for direct measurements of $\log K$, and overwrites the responsible rows of $\mathbf{H}$ by a sampling matrix [e.g., *Fritz et al.*, 2009, in press]. *Dagan* [1985] showed analytically that linearized $\mathbf{f}_y(\mathbf{s})$ for hydraulic heads is highly accurate for variances of $\log K$ up to unity, and *Nowak et al.* [2008] demonstrated its reliability for up to $\sigma^2_Y = 5$ by Monte-Carlo analysis.

Appendix F

Derivation of $E_{\mathbf{y}}\left[\tilde{\sigma}^2_{c|\mathbf{y}}\left(\mathbf{d}\right)\right]$ given in Chapter 4.4

We now derive Eq. (4.12) from Eq. (4.11). For simplicity of notation, let $\omega\left(\boldsymbol{\theta}\right) \equiv \sigma^2_{c|\mathbf{y}}\left(\boldsymbol{\theta}\right)$. Expanding $\omega\left(\boldsymbol{\theta}\right)$ up to first-order in $\boldsymbol{\theta}$ yields:

$$\omega\left(\boldsymbol{\theta}\right) \approx \omega\left(\bar{\boldsymbol{\theta}}\right) + \nabla_{\boldsymbol{\theta}}\omega\boldsymbol{\theta}' \,, \tag{F1}$$

where $\bar{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}\right]$, $\boldsymbol{\theta}' = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$ and $\nabla_{\boldsymbol{\theta}}\omega$ is the row-vector Jacobian of ω evaluated at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. Due to $E\left[\boldsymbol{\theta}'\right] = 0$, the first term in Eq. (4.11) becomes:

$$E_{\boldsymbol{\theta}}\left[\sigma^2_{c|\mathbf{y}}\left(\boldsymbol{\theta}\right)\right] \;\approx\; \sigma^2_{c|\mathbf{y}}\left(\bar{\boldsymbol{\theta}}\right) = \mathbf{H}_c\mathbf{G}_{\mathbf{ss}|\mathbf{y}}\left(\bar{\boldsymbol{\theta}}\right)\mathbf{H}_c^T \,. \tag{F2}$$

The second term in Eq. (4.11) is obtained in a similar fashion by setting

$$\begin{aligned} \bar{c}\left(\boldsymbol{\theta}\right) + \boldsymbol{\kappa}\left(\boldsymbol{\theta}\right)\mathbf{y}' &\equiv \bar{c}\left(\boldsymbol{\theta}\right) + \mathbf{H}_c\mathbf{G}_{\mathbf{ss}}\left(\boldsymbol{\theta}\right)\mathbf{H}_c^T\mathbf{G}_{\mathbf{yy}}^{-1}\left(\boldsymbol{\theta}\right)\left(\mathbf{y} - \bar{\mathbf{y}}\right) \\ &= \hat{c}\left(\mathbf{y}\left(\mathbf{d}\right), \boldsymbol{\theta}\right) \,, \end{aligned} \tag{F3}$$

with $\mathbf{y}' = (\mathbf{y} - \bar{\mathbf{y}})$. $\boldsymbol{\kappa}$ can be interpreted as the Kalman gain of predicted concentration, and $\bar{c}(\boldsymbol{\theta})$ is the ensemble mean concentration given $\boldsymbol{\theta}$, prior to sampling. Now, we expand $\bar{c}(\boldsymbol{\theta})$ and $\boldsymbol{\kappa}(\boldsymbol{\theta})$ up to first order in $\boldsymbol{\theta}$:

$$\bar{c}(\boldsymbol{\theta}) \approx \bar{c}\left(\bar{\boldsymbol{\theta}}\right) + \nabla_{\boldsymbol{\theta}}\bar{c}\boldsymbol{\theta}' \tag{F4}$$

$$\boldsymbol{\kappa}(\boldsymbol{\theta}) \approx \boldsymbol{\kappa}\left(\bar{\boldsymbol{\theta}}\right) + \nabla_{\boldsymbol{\theta}}\boldsymbol{\kappa}\boldsymbol{\theta}'. \tag{F5}$$

The first-order perturbation of $\hat{c}(\mathbf{y}(\mathbf{d}), \boldsymbol{\theta})$ is

$$\hat{c}' = \nabla_{\boldsymbol{\theta}}\bar{c}\boldsymbol{\theta}' + \nabla_{\boldsymbol{\theta}}\boldsymbol{\kappa}\boldsymbol{\theta}'\mathbf{y}', \tag{F6}$$

and its variance over the distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is (accurate to first order in $\boldsymbol{\theta}$):

$$\begin{aligned} & V_{\boldsymbol{\theta}|\mathbf{y}}\left[\hat{c}(\mathbf{y}(\mathbf{d}), \boldsymbol{\theta})\right] \\ & \approx E_{\boldsymbol{\theta}|\mathbf{y}}\left[\sum_i \sum_j \theta_i'\theta_j' \left\{ \frac{\partial \boldsymbol{\gamma}}{\partial \theta_i}\left(\frac{\partial \boldsymbol{\gamma}}{\partial \theta_j}\right)^T + \frac{\partial \boldsymbol{\kappa}}{\partial \theta_i}\mathbf{y}'\mathbf{y}'^T\left(\frac{\partial \boldsymbol{\kappa}}{\partial \theta_j}\right)^T \right\}\right] \\ & = \sum_i \sum_j \left\langle \mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}|\mathbf{y}} \right\rangle_{ij} \left\{ \frac{\partial \boldsymbol{\gamma}}{\partial \theta_i}\left(\frac{\partial \boldsymbol{\gamma}}{\partial \theta_j}\right)^T + \frac{\partial \boldsymbol{\kappa}}{\partial \theta_i}\mathbf{y}'\mathbf{y}'^T\left(\frac{\partial \boldsymbol{\kappa}}{\partial \theta_j}\right)^T \right\} \end{aligned} \tag{F7}$$

where $\mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}|\mathbf{y}}$ is the conditional covariance of $\boldsymbol{\theta}$ and $\langle\cdot\rangle$ denotes the i,j-the element.

$\mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}|\mathbf{y}}$ is independent of actual data values when expressed via the inverse of the Fisher information $\mathbf{F}$ [e.g., *Kitanidis and Lane*, 1985]:

$$\mathbf{F} = E_{\mathbf{y}}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}}\ln p(\mathbf{y}|\boldsymbol{\theta})\right)^T\left(\frac{\partial}{\partial \boldsymbol{\theta}}\ln p(\mathbf{y}|\boldsymbol{\theta})\right)\right]. \tag{F8}$$

In the current context, we assume that the $\boldsymbol{\theta}$ has a prior covariance matrix $\mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}}$, so that the elements F_{ij} of $\mathbf{F}$ are given by [*Nowak and Cirpka*, 2006]:

$$F_{ij} = \frac{1}{2}Tr\left[\frac{\partial \mathbf{G}_{\mathbf{yy}}}{\partial \theta_i}\mathbf{G}_{\mathbf{yy}}^{-1}\frac{\partial \mathbf{G}_{\mathbf{yy}}}{\partial \theta_j}\mathbf{G}_{\mathbf{yy}}^{-1}\right] + \mathbf{e}_i^T\mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\mathbf{e}_j, \tag{F9}$$

where $\mathbf{e}_i$ is the i-th unit vector. $\mathbf{G_{yy}}$ and its derivatives are evaluated at $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$.

Now, we take the expected value over $p(\mathbf{y})$ to obtain the second term in Eq. (4.11):

$$\begin{aligned} & E_{\mathbf{y}}\left\{V_{\boldsymbol{\theta}|\mathbf{y}}\left[\hat{c}\left(\mathbf{y}\left(\mathbf{d}\right),\boldsymbol{\theta}\right)\right]\right\} \\ & \approx \sum_i \sum_j \left\langle \mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}|\mathbf{y}} \right\rangle_{ij} \left\{ \frac{\partial \boldsymbol{\gamma}}{\partial \theta_i}\left(\frac{\partial \boldsymbol{\gamma}}{\partial \theta_j}\right)^T + \frac{\partial \boldsymbol{\kappa}}{\partial \theta_i}\mathbf{G_{yy}}\left(\bar{\boldsymbol{\theta}}\right)\left(\frac{\partial \boldsymbol{\kappa}}{\partial \theta_j}\right)^T \right\} \end{aligned} \tag{F10}$$

Updating the structural parameters once data become available requires the gradient $\mathbf{g}$ [*Kitanidis and Lane*, 1985]. For the case of prior covariance $\mathbf{C}_{\boldsymbol{\theta}\boldsymbol{\theta}}$, its entries are [*Nowak and Cirpka*, 2006]:

$$g_i = \frac{1}{2} Tr\left[\frac{\partial \mathbf{G_{yy}}}{\partial \theta_i}\mathbf{G_{yy}^{-1}}\right] - \frac{1}{2}\left(\mathbf{y}_o - \bar{\mathbf{y}}\right)^T \mathbf{G_{yy}^{-1}}\frac{\partial \mathbf{G_{yy}}}{\partial \theta_j}\mathbf{G_{yy}^{-1}}\left(\mathbf{y}_o - \bar{\mathbf{y}}\right) + \mathbf{e}_i^T\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right) \tag{F11}$$