Uncertainties in Modeling Groundwater Contamination

Bruno Mendes
Abstract

The scientific problem of nuclear waste disposal suffers from strong uncertainties arising both from its object (the system being studied being unknown) and also from its ambitious objectives (prediction of future events and their consequences). This work focuses on identifying, structuring and quantifying the main uncertainties involved in this field. It also shows the way how to integrate all the identified uncertainties into any inference and prediction a risk analyst might want to make in this field.

An overall conceptual support is provided by using Bayesian statistics combined with hierarchical modeling techniques. Scenario and parametric uncertainty and their implications was studied in the context of a expert assessment project called GESAMAC, where Functional Data Analysis techniques were used to study the variability of the dose vs. time curves. Markov Chain Monte Carlo simulation was also used to make inference studies of parametric uncertainty using data from a field study.

Although this study is not exhaustive, it does show that the different sources of uncertainties do have an important impact on what we really know about the development system. Acknowledging scenario uncertainty, for instance, shows that we should expect uncertainties of dose responses on the order of 20 times what is expected to happen by default. It is shown that there is an asymmetry on the uncertainty of the dose curves; it seems there is more uncertainty on how the curve reaches its maximum than on how it relaxes afterwards. It is also found that the influence of the physical parameters on the dose curve is far from being linear and it seems to depend in a very important way on large scale physical inhomogeneities of the system.
List of Publications


VI B. Mendes and A. Pereira, *PMCD- Parallel Monte Carlo Drive, A software package for Monte Carlo simulations in parallel*, *Computer Physics Communications*, in press, July 2002


Author’s contributions to the papers

In [Papers I, II and III] I shared the responsibilities of all the numerical simulations. In [Papers IV and V] I was involved in the development of software for the parallel computer and in the performing all the simulations in that computer. In [Papers VI and VII] I performed everything that is contained in it in conjunction with the co-authors.
Contents

List of Publications 3

1 Introduction 11

2 Groundwater modeling 15
   2.1 Hydrodynamic and transport processes .............................. 15
   2.1.1 Analytical solution ............................................. 18
   2.1.2 Calculation of dose from contaminant flux .................... 20
   2.2 The Level E/G input specification ................................. 20
   2.3 Monte Carlo simulations ........................................... 21

3 Statistical methods 23
   3.1 Hierarchical modeling ............................................. 23
   3.1.1 Model uncertainty in the Level E/G specification .............. 24
   3.2 Markov Chain Monte Carlo ....................................... 26
   3.2.1 The prior ..................................................... 27
   3.2.2 The likelihood function ....................................... 28
   3.2.3 Maximum Likelihood Methods ................................ 29
   3.2.4 Markov Chains ............................................... 29
   3.2.5 The Metropolis algorithm .................................... 30
   3.3 Functional data analysis ......................................... 30
   3.3.1 Principal components analysis ................................ 31
   3.3.2 Linear modeling .............................................. 33

4 Markov Chain Monte Carlo. Results 37
   4.1 Experimental setup and experimental data .......................... 37
   4.2 Exploratory analysis .............................................. 37
   4.3 MCMC results ..................................................... 40

5 Conclusions 45

Appendix A 47

Acknowledgements 50

Paper I 53

Paper II 55

Paper III 57

Paper IV 59

Paper V 61

Paper VI 63
List of Figures

1.1 *Diagram of a possible high-level nuclear waste disposal site (taken from the French National Agency for Radioactive Waste Management’s web site).* ........................................... 12

2.1 *Surface plot of a solution to the simplest advection-dispersion differential equation.* ... 19

4.1 *Experiment setup for the SCN tracer experiment of [7].* ........................................ 37

4.2 *Breakthrough curve for the SCN tracer experiment of [7].* ..................................... 38

4.3 *Contour plot of the likelihood function with R held constant at 1. Broad plotting region for the parameter space.* ................................................................. 39

4.4 *Contour plot of the likelihood function with R held constant at 1. Parameter space region of maximum.* ................................................................. 40

4.5 *Contour plot of the likelihood function with R held constant at 1.* .............................. 40

4.6 *Surface plot of the likelihood function with R held constant at 1.* .............................. 41

4.7 *Markov Chain Monte Carlo output and diagnostics for water velocity (c.f. text).* .... 42

4.8 *Markov Chain Monte Carlo output and diagnostics for dispersion c.f. text.* ............ 42

4.9 *Markov Chain Monte Carlo output and diagnostics for σ.* ..................................... 43

4.10 *Plot of sampled values for the two parameters of interest, dispersion and water velocity.* .......... 44

4.11 *Curves resulting from every sampled values of water velocity and dispersion.* ............ 44
List of Tables

2.1 Chemical reactions and bands of uncertainty for their respective reactions rates simulated by GTMCHEM, for each scenario. The parameters correspond to maximum and minimum values for uniform distributions. .............................................. 21

2.2 Level E/G parameters for Iodine. The parameters correspond to maximum and minimum values for uniform distributions; the only exceptions are the water velocity and the stream parameters which have log-uniform distributions. Contaminant time is measured in years, leaching rate in years$^{-1}$, water velocity in meters/year, geosphere lengths in meters, flow rates in m$^3$/year and dose conversion factors in Sievert/moles. ................................. 22

4.1 Maximum Likelihood estimators for water velocity, dispersion and mean squared error of the fit. ................................................................. 41
Chapter 1

Introduction

Sweden and other nations with nuclear power production capabilities are facing an important challenge. With 30 to 40 years of operation of their nuclear power stations already behind them, finding a good answer to the following question becomes more pressing: what to do with the spent nuclear fuel? This work is intended as a contribution to one possible solution to this problem.

The problem in its broadest sense is known as disposal of nuclear waste. The proposed solution that has received the most attention internationally is to bury the waste in specially built deep underground facilities that look something like the example in Figure 1.1.

In order to study this approach, one typically breaks the system of underground facilities down into three main components. The first part is concerned with the immediate region surrounding the nuclear waste, which is usually referred to as the near field. The second component starts where the previous region finishes and goes all the way to the biosphere; this is usually called the far field or the geosphere. It deals mainly with the rock and groundwater around the whole site. The third component deals with the biosphere and it comprises such things as lakes, rivers, plants, animals, cities, and people.

There are two types of scientific work to be done on each of these components. The first type of work has to do with the characterization and setting up of the whole system, i.e., to learn as much as possible from it and to engineer it in a way that makes its conditions more suitable for its purpose. The preventive engineering that can be done is restricted to the near field component. In the far field and the biosphere we are reduced to monitoring, as far as preventive measures are concerned. The second type of work consists of trying to predict the system’s behavior during its long time of operation (which can range between ten thousand and a hundred thousand years) under as many scenarios as possible [19].

The site will typically be located about 500 meters under the surface and the fuel will be enclosed in several layers of different materials (such as stainless steel, lead, titanium, and copper) to render the canister as chemically stable as possible. The storage holes (where the canisters are supposed to rest) will be filled with different types of materials, including clays for example, to counter the effects of possible water infiltration and leakages from the canister. The rock immediately surrounding the storage holes will have a special treatment, including pumping concrete into the rock fractures, to improve its containing capabilities.

As for the third component of the system, the biosphere, studies are centered on knowing the paths that radionuclides might take to humans and the effects of possible radiologic doses on human health.

In this dissertation I deal mainly with the far field. My general goal is the prediction of the fate of radionuclides in the geosphere once there is a leakage serious enough to leave the near field. So the problem can be defined in this simple question:

**Assuming a serious leakage from the near field of a repository, how will the radionuclides disperse in the geosphere?**

Of course, it is implicit that one’s interest lies ultimately in how this will affect the human population. In this perspective, the far field will be the second of the three barriers that are
placed between the nuclear waste and humans.

When trying to answer the above question one comes up against two different types of uncertainty:

- uncertainty in the present characteristics of the geosphere, and
- uncertainty in the future development of those same characteristics.

The first type of uncertainty is usually tackled by means of experiments, but this approach is far from giving definite answers. Although countless efforts have been done in the past, this topic still lacks informative, reliable experimental data. Noninvasive imaging of underground disposal systems has proven ineffective; therefore, to observe such a system in action it is necessary to drill holes in the rock, and these holes dramatically change the behavior of the system under study. As for the second type of uncertainty, it is the main difficulty intrinsic to the subject of risk analysis, the science that tries to predict possible paths to disasters and their consequences. I believe the only way we have a chance of overcoming these problems is through mathematical, statistical, and numerical modeling of the physical system we are concerned with. These instruments are the only rational, consistent and verifiable way we have to extract the maximum information from experiments and use it to illuminate what lies ahead in terms of possible future scenarios.

The subject of uncertainties in the modeling of fate of contaminants is a central one in this dissertation. I believe that it is the single most important challenge facing all the scientists and professionals working in this field. No one working in this field would disagree with this statement, what to do about it has, on the other hand, brought about very different attitudes. These range from the very pragmatical but narrow scoped strategy of making the most reasonable assumptions and then carrying on as there was no uncertainty, to very thoroughly trying to elicit all sources of uncertainty and therefore build a universe of possible outcomes in leakage accidents, and using mathematical/statistical tools to try and extract useful predictive information [8, 5]. My approach is very much in line with this last position. As I will show in this document, we already have the conceptual tools that are sophisticated enough to be applied to this problem [Papers IV, V and VI] and, better still, extract very useful information of what to expect in an accident; what kind of uncertainties we have and what impact they will have in our predictions [Papers I, II, III, V and VII]. Furthermore, I obtained some results that indicate that a lot is still to be learned about the implications of all the uncertainties that plague this field. The main tools I use are: Bayesian statistics and hierarchical modeling to integrate all the sources of uncertainties in a mathematical valid framework, Markov Chain Monte Carlo simulation methods to implement all the calculations and finally I use simple Monte Carlo simulations and Functional Data Analysis to quantify and visualize the impact of scenario uncertainty in the prior distributions for dose curves produced by a numerical model.
In sum, my main goal in this work was to methodically quantify all the sources of uncertainty and also to show that there are strong indications these uncertainties have important impact on risk analysis work.

This document is divided into three main parts. In the first part I introduce the basic concepts necessary to understanding the deterministic part of the mathematical and numerical modeling: the differential equations used, the physical meaning of their parameters, known analytical solutions, and the description of the different future scenarios that we expect may develop. In the second part I present the theoretical background of my statistical modeling approach, by describing the main concepts in Functional Data Analysis as developed by Ramsay and Silverman in [17] and also the main concepts in Markov Chain Monte Carlo and incorporation of different sources of uncertainty in our statistical modeling as developed by David Draper in [2, 1]. Finally, in the third part I present the practical details of the theoretical tools described in the previous parts, the results obtained and some analysis and conclusions.
Chapter 2

Groundwater modeling

Here I introduce the main concepts involved in the mathematical modeling of groundwater contamination and contaminant transport.

2.1 Hydrodynamic and transport processes

This section outlines the mathematical models for the physical transport of contaminants. See [10, 18, 4] for more detailed discussions.

Mathematically one usually describes the movement of a contaminant in groundwater by two main processes: advection and dispersion. The first involves the bulk movement of the solute (contaminant) together with the water; the second is the spreading of the contaminant as it moves in a porous medium (and not necessarily in the same direction as the water). This last process can be divided in two underlying behaviors: mechanical dispersion and molecular diffusion. The mechanical dispersion occurs due to velocity variations in the pore channels and the tortuous nature of the flow in a porous medium. Molecular diffusion occurs due to concentration gradients. Its effect is to counter those gradients (a behavior that can be traced to basic thermodynamic concepts).

The idealized system I intend to model consists of a one-dimensional column of porous material through which the radionuclides will be transported by water. The principle of mass conservation is applicable and it will lead us to the following transport Equation [13]:

\[
\frac{n \partial C}{\partial t} + (1 - n) \rho \frac{\partial S}{\partial t} = Dn \frac{\partial^2 C}{\partial X^2} - Vn \frac{\partial C}{\partial X} - \lambda[nC + (1 - n)\rho S].
\] (2.1)

Here \(X [L]^{1}\) denotes space, \(t [T]\) time, \(n [-]\) is the porosity, \(C [M/L^{3}]\) is the concentration of radionuclide in the water, \(S [M]\) is the sorbed (absorbed or adsorbed) amount of radionuclide in the porous medium, \(\rho [ML^{-3}]\) is the dry bulk density of the porous medium, \(\lambda [T^{-1}]\) is the radioactive decay constant, and \(V [LT^{-1}]\) is the groundwater pore velocity. In (2.1) \(D [L^{2}T^{-1}]\), the hydrodynamic dispersion coefficient, is given by

\[
D = D^0 + \frac{\kappa \alpha \Delta h}{n \Delta l},
\] (2.2)

where \(D^0 [L^{2}T^{-1}]\) is the molecular diffusion coefficient, \(\kappa [LT^{-1}]\) is the hydraulic conductivity, \(\alpha [L]\) is the dispersion length, \(\Delta h [L]\) is the variation in height, and \(\Delta l [L]\) is the variation in distance.

As one can see, Equation (2.1) includes the radioactive decay of the substance being transported (\(C\)) and it also includes the substance that is sorbed in the porous medium during the transport (\(S\)). For this latter part of the modeling one usually assumes equilibrium between the dissolved substance and the sorbed one. In this case it can be shown [10] that

\(^{1}\text{In this chapter I indicate the dimensions of a quantity right after its definition.} \quad L \text{ stands for length,} \ T \text{ for time and} \ M \text{ for mass. This information is enclosed by square brackets.}\)
\( \frac{\partial C}{\partial t} \) and \( \frac{\partial C}{\partial x} \) are proportional to each other and the proportionality constant is given by:

\[
R = 1 + \frac{\rho_s (1 - n)}{n} k_d,
\]

where \( R \) is called the retention factor, \( \rho_s \) [ML\(^{-3}\)] represents the dry density of the porous medium and \( k_d \) [L\(^3\)M\(^{-1}\)] represents the distribution coefficient. This equation represents the so-called fast adsorption model.

In this manner one can write the transport Equation (2.1) in one variable only (the concentration of the substance in the water) as

\[
\frac{\partial C}{\partial t} + \frac{D}{R} \frac{\partial^2 C}{\partial x^2} - \frac{V}{R} \frac{\partial C}{\partial x} - \lambda C = 0.
\] (2.4)

The formulation is not complete until we determine the initial and boundary conditions:

\[
C(X, t) = 0 \text{ for all } X \text{ when } t \leq \tau
\]

and

\[
C(0, t) = f(t),
\]

where I use \( f(t) \) to denote a function of time that describes the leakage of contaminant from the repository; also

\[
\left. \frac{\partial C(X, t)}{\partial X} \right|_{X=L} = 0 \text{ for all } t.
\]

Here \( \tau \) denotes the time after which the leak starts and \( L \) denotes the geosphere path length of the radionuclides.

So far I have tried to keep the complexity of the model low, in order to better introduce the main processes that occur in this system, so I have only modelled one type of chemical behavior: the reaction between the contaminant in the water and the rock surface (or porous medium), and I have assumed this reaction to be in equilibrium. These two assumptions do not account for all possible conditions present in groundwater; for example, in order for equilibrium to happen the flow of the dissolved substance has to be sufficiently slow. When the flow does not satisfy this requirement the formulation of the chemical reaction model has to change to include kinetic reactions. We should also try to include other types of chemical reactions that the contaminant might participate in: reactions with other components present in the water or the porous medium.

A list of the main reactions one could expect to occur would include:

- Advection;
- Dispersion;
- Rapid adsorption by linear isotherm in the porous medium;
- Complexation in solution. This process involves substances dissolved in water. When a substance loses some of its electrons, its net positive charge makes it a target for groups of anions (negatively charged ions) or neutral molecules. The anions and neutral molecules are known as ligands or complexing agents;
- Biodegradation/filtration. Biodegradation denotes the transformation of a chemical substance by micro-organisms (such as bacteria and fungi). Filtration is an abbreviation for colloid filtration. Colloid is a general term for particles that are larger than molecules but not so large that their components of the mixture separate under the influence of gravity. They are defined as a function of their diameter, so a particle with dimensions between 10 and 2000 Å is considered to be a colloid. Colloids can be formed by atoms, ions, molecules, or even a giant molecule; and
Adversely, non-instantaneous kinetics. These models try to capture the situation when the reaction between the mobile (dissolved) contaminant and the rock is so slow that it does not have time to reach equilibrium. In this case the mathematical description cannot make use of a simple constant term to relate the quantity of mobile contaminant and the quantity sorbed in the rock; in other words, the reaction constant has to depend on both the mobile and the sorbed concentrations.

Each of these reactions is modelled differently. We have already seen that advection and dispersion are modelled by the first and second derivative of contaminant concentration with respect to the space variable, and rapid adsorption is modelled by a constant term (the retention). The complexation is not as straightforward since it is a non-linear process involving other chemical components in solution. This means that we first have to include an additional differential equation modeling advection, dispersion, adsorption and biodegradation/filtration for these additional species (chemical compounds), and then include in both differential equations a coupling term that represents the interaction (complexation) between the two species. Biodegradation and filtration are simpler to model since they can be assumed to act like a sink of contaminant, so this can be represented mathematically by a term negatively proportional to the concentration of contaminant. The nonequilibrium adsorption is modelled in a similar fashion to complexation; it usually involves an additional differential equation to track the amount of contaminant present in the porous medium. In the simplest case the adsorption rate increases proportionally with the concentration of mobile contaminant, but as more and more contaminant gets sorbed, the sorbed quantity should saturate; mathematically this situation can be described as

$$\frac{\partial C}{\partial t} = k_1 C - k_2 S,$$

where $k_1$ and $k_2$ are positive constants and $S$ is the amount of contaminant sorbed in the rock.

In the next display (2.6) I show the differential equation for the contaminant in water and all the additional terms due to the new chemical reactions being considered; below I introduce the remaining equations for the contaminant in the porous medium and a new set of coupled equations for the species that might complexate with the contaminant.

$$R \frac{\partial C^i}{\partial t} = \begin{cases} \\
\text{dispersive part} & \text{advective part} & \text{reaction in the solution} \\
D \frac{\partial^2 C^i}{\partial X^2} & -V \frac{\partial C}{\partial X} & -[S^{ij} C^j - S^{ji} C^i] \\
-\zeta^i C^i & -K^{i1} \frac{C_M - C^i}{C} C^i + K^{i2} C^i & \end{cases},$$

where $i$ denotes each species in solution, $S^{ij}$ and $S^{ji}$ are the forward and backward kinetic reaction rate constants for reactions involving species $i$ and $j$, $\zeta^i$ is the reaction rate for the biodegradation/filtration, $K^{i1}$ and $K^{i2}$ are, respectively the forward and backward reaction rates for the adsorption in the porous medium, $C^i$ is the amount of contaminant in the porous medium, and $C_M$ is the maximum amount of contaminant that can be adsorbed in the porous medium. When a chemical reaction is not in equilibrium the rate of production of products is different from the rate of production of reactants; one then has to make a distinction between the two ways in which the reaction can go, and this is why I refer to forward and backward reaction rate constants.

Notice that the adsorption of the contaminant is modelled in two different ways: using the retention factor (linear adsorption) or using the last two terms in the last equation (slow adsorption of the Langmuir type). This might at first seem contradictory, but it is included in the expert elaborated Level E/G specification (see Section 2.2 for more details), and it
might be regarded as different, parallel reaction paths to achieve sorption; some reactions will go quickly and the equilibrium can be assumed, and other reactions won’t be fast enough and can’t be assumed to be in equilibrium.

For the chemical species in solution the differential equations are

\[
R \frac{\partial C_i}{\partial t} = D \frac{\partial^2 C_i}{\partial X^2} - V \frac{\partial C_i}{\partial X} - [S_{ij} C^j - S_i C^i] - \zeta^i \dot{C}^i - K^{i1} (\bar{C}_M - \bar{C}^i) C^i + K^{i2} \bar{C}^i,
\]

\[
\frac{\partial \bar{C}^i}{\partial t} = K^{i1} (\bar{C}_M - \bar{C}^i) C^i - k^{i2} \bar{C}^i.
\]

(2.7)

We can include an additional feature in our mathematical model: the spatial inhomogeneities of some of the characteristics of the porous medium. This can be done in a rough but simple manner: one simply uses the model above for each stretch of the path that is homogeneous, with the output of one stretch as input for the next stretch. This procedure is an approximation to the geological concept of layers where each single layer would represent a zone of the geosphere with roughly homogeneous characteristics.

I used a program called GTMCHEM [13] to obtain numerical solutions to the general set of differential equations presented above. This program was developed and enhanced as part of the work of the GESAMAC project, which produced sensitivity and uncertainty analysis upon which this work builds.

### 2.1.1 Analytical solution

There is an analytical solution to the simpler Equation (2.4), which can be re-written as

\[
R \frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial X^2} - V \frac{\partial C}{\partial X} - \lambda R C,
\]

with initial and boundary conditions given by

\[
C(X, t) = 0 \text{ for } t = 0 \quad \text{and} \quad \frac{\partial C}{\partial X} = 0 \text{ for } X = L.
\]

The input concentration has the following profile:

\[
f(t) = \left\{ \begin{array}{ll}
C_0 e^{-\lambda t} & 0 < t < t_0 \\
0 & t > t_0
\end{array} \right.
\]

where the second condition means that there is no more leakage after time \( t_0 \). Its analytical solution [9] is as follows: for \( t < t_0 \)

\[
C(x, t) = C_0 e^{-\lambda t} \left[ \frac{V}{V + U} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX - Ut}{2\sqrt{DRt}} \right) + \frac{V}{V - U} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX + Ut}{2\sqrt{DRt}} \right) + \frac{V^2}{2DR(\lambda - \Lambda)} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX + V t}{2\sqrt{DRt}} \right) \right],
\]

and for \( t > t_0 \), after the leakage has stopped,

\[
C(x, t) = C_0 e^{-\lambda t} \left[ \frac{V}{V + U} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX - Ut}{2\sqrt{DR(t - t_0)}} \right) + \frac{V}{V - U} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX + Ut}{2\sqrt{DR(t - t_0)}} \right) + \frac{V^2}{2DR(\lambda - \Lambda)} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX + V(t - t_0)}{2\sqrt{DR(t - t_0)}} \right) \right. + \left. \frac{V}{V + U} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX - U(t - t_0)}{2\sqrt{DR(t - t_0)}} \right) + \frac{V}{V - U} e^{\frac{V}{2u} \lambda \tau} \text{erfc} \left( \frac{RX + U(t - t_0)}{2\sqrt{DR(t - t_0)}} \right) \right],
\]

(2.10)
Figure 2.1: *Surface plot of a solution to the simplest advection-dispersion differential equation.*

where $U = \sqrt{V^2 + 4DR(\lambda - \Lambda)}$ and \( \text{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz \) is the so called *complementary error function*. Here \( \Lambda \) is the leaching rate and is a constant, \( C_0 \) is the initial inventory of contaminant in the containers, and \( V \) is the water velocity.

This solution provides valuable information on what to expect from a typical run of GTMCHEM. Considering previous studies [12],[Papers IV and V], I expect the solution for equations (2.7) to be a smoothed version of the analytical solution above, especially for large values of \( X \). The reason is simple to grasp qualitatively: advection-dispersion equations (of which Equation (2.4) is a common example) typically deform the input function (i.e., the time behavior of the quantity of contaminant being inserted in the system) by way of smoothing the function out and at the same time transporting it, and this behavior becomes more apparent with the system of geological layers referred above; the input for a second layer consists of the output of the previous layer which results in an even smoother function at the end of the second layer. The coupled system of equations (2.7) can *potentially* yield very different type of behaviors, but in this situation I did not use its full modeling power: I did not consider reactions in the solution (see Section 2.6) which is why I have only a system of two coupled equations. Even the coupling of these two equations is expected to be weak. For example, in [Papers IV and V] I have hints that the chemical constants that create that coupling have next to no influence in the variability in the maximum radiologic dose. Although this *per se* does not assure us that one will have a solution similar to (2.4), it does hint that the perturbation to the simpler solution should not be that significant.

The conclusion to be drawn from these comments is that I expect the solutions provided by the computer program GTMCHEM to look like the analytical solution above, especially away from the source.

A typical graphical representation of this solution is given in Figure 2.1 (in this graph, concentration is expressed relative to \( C_0 \)).

2.1.2 Calculation of dose from contaminant flux

GTMCHEM assumes that once the radionuclides leave the geosphere they will reach a path that will connect directly to the human population, via a stream which contributes to the supply of drinking water, which means that the time for transport in the biosphere will be almost instantaneous when compared with the scale of times for the transport in the porous medium. The dose to which the human population will be exposed will be a function of the stream’s flow rate and the flux of the radionuclide reaching the stream. This flux is the quantity obtained by multiplying the concentration by the velocity of the water; since the latter is constant in our context, flux is just a scaled version of the concentration. The dose\(^2\) in \([Sv/year]\) is then given by [16]

\[
F_i(t) = \kappa_i \frac{w}{W} G_i^L(t),
\]

where \(\kappa_i [Sv/mol]\) is the dose conversion factor for radionuclide \(i\), \(w [m^3/year]\) is the annual individual drinking water requirement, \(W [m^3/year]\) is the stream flow rate and \(G_i^L [mol/year]\) is the flux of radionuclide \(i\) leaving the geosphere at time \(t\).

2.2 The Level E/G input specification

Once the numerical models have been defined one has to specify the context in which those numerical tools will be applied/tested, by which I mean defining the sets of parameters one is interested in studying, and the bands of uncertainty associated with each of those parameters. Here I use probabilities to quantify uncertainties. It was found useful in earlier work to define a number of different scenarios, each corresponding to different conditions of the geosphere. This specification is called Level E/G and is described in detail in Paper IV and [14].

In the following I describe the physical conditions that motivate each scenario and then (for each scenario) the sets of all model parameters which implement them in the numerical model.

- **Reference (REF) Scenario.** Assumes that present, known, conditions will be maintained in the future.
- **Fast Path (FP) Scenario.** The distance from the repository to the surface is reduced considerably, due to erosion or drilling.
- **Additional Geosphere (AG) Scenario.** Involves changes dealing with an additional layer in the geosphere: accumulation of debris left behind by a retreating glacier and geological dynamics that creates a longer rock pathway.
- **Glacial Advance (GA) Scenario.** Tries to predict the conditions when the next glacial age comes; represents conditions typical of an advancing glacier.
- **Environmentally Induced Changes (EIC) Scenario.** Changes induced by tectonic movement of the rock, or changes induced by human activities in the hydrologic characteristics of the area surrounding the disposal site.
- **Human Disposal Errors (HDE) Scenario.** This includes errors made during repository construction, the disposal of the wastes or any other activities connected with the repository operation.

The REF scenario embodies present day conditions and also assumes that all the barriers (human built and natural) perform as expected. In qualitative terms, the FP scenario is characterized by the following situations: geologically, a rock fault could develop that would

---

\(^2\)Due to the peculiarities of dose units, I show the units in square brackets here instead of the physical dimensions.
### Table 2.1: Chemical reactions and bands of uncertainty for their respective reaction rates simulated by GTMChem, for each scenario. The parameters correspond to maximum and minimum values for uniform distributions.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Reactions</th>
<th>Parameters [year⁻¹]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG</td>
<td>Slow, reversible adsorption (layer 2) \ Biodegradation/filtration (layer 3)</td>
<td>[10⁻²; 10⁻⁴], [10⁻⁴; 10⁻³]</td>
</tr>
<tr>
<td>EIC</td>
<td>Slow, reversible adsorption (layer 1) \ Biodegradation/filtration (layer 2)</td>
<td>[10⁻³; 10⁻⁷], [10⁻⁴; 10⁻³]</td>
</tr>
<tr>
<td>FP</td>
<td>Slow, reversible adsorption (layer 1) \ Biodegradation/filtration (layer 2)</td>
<td>[10⁻³; 10⁻⁷], [10⁻⁴; 10⁻³]</td>
</tr>
<tr>
<td>GA</td>
<td>Slow, reversible adsorption (layer 1) \ Slow, reversible adsorption (layer 2) \ Biodegradation/filtration (layer 2)</td>
<td>[10⁻³; 10⁻⁷], [10⁻³; 10⁻⁷], [10⁻⁴; 10⁻³]</td>
</tr>
<tr>
<td>HDE</td>
<td>Slow, reversible adsorption (layer 1) \ Slow, reversible adsorption (layer 2) \ Biodegradation/filtration (layer 2)</td>
<td>[10⁻³; 10⁻⁷], [10⁻³; 10⁻⁷], [10⁻⁴; 10⁻³]</td>
</tr>
</tbody>
</table>

pass through the repository connecting it directly to the biosphere; a direct connection between the repository and the biosphere could also be induced by the climate through severe erosion, and the geosphere barrier could have its containment role severely reduced. Another way this situation could be achieved is by direct human intervention, like mining or drilling of a pumping well.

The AG scenario includes several different situations. Geologically it could be arrived at by the development of a rock fault passing close to the repository, but with a correspondingly longer pathway than the existing faults. Climatically this scenario is connected with a glacial retreat and the additional geosphere due to the materials deposited by the retreating glacial layer. The GA scenario could happen due to a new ice age and corresponding hydrological changes.

The EIC scenario describes a situation which could correspond to a rock fault developing relatively far away from the repository site. Its influence on the site could only be felt indirectly. This scenario also includes situations where human intervention could alter the conditions of the system, although not in such a dramatic way as in the FP scenario. The HDE scenario deals with errors in the construction and operation of the repository, so its impact is felt directly in the near field part of the model.

It will be clear in Section 3.1 that there is a need to establish an \textit{a priori} set of probabilities for each scenario. These probabilities are meant to roughly denote beliefs about the likelihood of each scenario in the future. In subsequent sections I use the following probabilities: AG: 0.0125, EIC: 0.0325, FP: 0.0225, GA: 0.0125, HDE: 0.02, REF: 0.90. The details of how these probabilities were elicited (including the notion of micro-scenarios which I omitted here for brevity) are given in Paper IV and [14].

In Tables 2.1 and 2.2 I present the list of GTMChem parameters and their bands of variation for each scenario for a typical contaminant (Iodine-129), which is the only contaminant examined in this thesis. These bands were obtained via expert elicitation of likely conditions specific to each scenario.

### 2.3 Monte Carlo simulations

I choose the Monte Carlo method to explore the space of possible solutions of the differential equations that GTMChem solves.

When one is trying to make inferences or predictions about a given system, one of the main problems are the uncertainties in the parameters of the mathematical model that describes that system. One way to incorporate those uncertainties in the inferences or predictions is resorting to Monte Carlo simulations. The procedure is quite straightforward: one tries to describe the uncertainty for each parameter by means of a probability density function. One such distribution is constructed for each parameter of the model. Then one samples each distribution to obtain a single value for each parameter. Once that is accomplished for all the parameters, the solution to the differential equations is run with these values. This process is then repeated many times to obtain a distribution of solutions that reflect the uncertainty in the parameters. This approach is particularly useful when dealing with complex systems where the uncertainties in the parameters are difficult to quantify precisely.
<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>REF</th>
<th>FP</th>
<th>AG</th>
<th>GA</th>
<th>EIC</th>
<th>HDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Containment Time</td>
<td>[100;1000]</td>
<td>[100;1000]</td>
<td>[100;1000]</td>
<td>[100;1000]</td>
<td>[100;700]</td>
<td>[50;500]</td>
</tr>
<tr>
<td>Leaching Rate</td>
<td>[0.001;0.01]</td>
<td>[0.001;0.01]</td>
<td>[0.001;0.01]</td>
<td>[0.005;0.05]</td>
<td>[0.003;0.03]</td>
<td>[0.001;0.01]</td>
</tr>
<tr>
<td>Water Velocity Layer 1</td>
<td>[0.001;0.1]</td>
<td>[0.1;1]</td>
<td>[0.01;1]</td>
<td>[0.003;0.3]</td>
<td>[0.003;0.3]</td>
<td>[0.002;0.2]</td>
</tr>
<tr>
<td>Geosphere Path Length Layer 1</td>
<td>[100;500]</td>
<td>[200;500]</td>
<td>[100;500]</td>
<td>[200;600]</td>
<td>[200;500]</td>
<td>[100;500]</td>
</tr>
<tr>
<td>Retention Factor Layer 1</td>
<td>[1;5]</td>
<td>[1;2]</td>
<td>[1;5]</td>
<td>[1;5]</td>
<td>[1;3]</td>
<td>[1;5]</td>
</tr>
<tr>
<td>Water Velocity Layer 2</td>
<td>[0.01;0.1]</td>
<td>—</td>
<td>[0.1;1]</td>
<td>[0.01;0.1]</td>
<td>[0.03;0.3]</td>
<td>[0.02;0.2]</td>
</tr>
<tr>
<td>Geosphere Path Length Layer 2</td>
<td>[50;200]</td>
<td>—</td>
<td>[50;200]</td>
<td>[50;100]</td>
<td>[50;200]</td>
<td>[50;200]</td>
</tr>
<tr>
<td>Retention Factor Layer 2</td>
<td>[1;5]</td>
<td>—</td>
<td>[1;5]</td>
<td>[2;7]</td>
<td>[1;5]</td>
<td>[1;3]</td>
</tr>
<tr>
<td>Water Velocity Layer 3</td>
<td>—</td>
<td>—</td>
<td>[0.5;1]</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Geosphere Path Length Layer 3</td>
<td>—</td>
<td>—</td>
<td>[50;100]</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Retention Factor Layer 3</td>
<td>—</td>
<td>—</td>
<td>[1;10]</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Stream Flow Rate</td>
<td>$[10^5;10^6]$</td>
<td>$[10^4;10^6]$</td>
<td>$[10^5;10^6]$</td>
<td>$[10^5;10^7]$</td>
<td>$[10^5;10^7]$</td>
<td>$[10^5;10^7]$</td>
</tr>
<tr>
<td>Dose Conversion Factor</td>
<td>56</td>
<td>2800</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
</tbody>
</table>

Table 2.2: Level E/G parameters for Iodine. The parameters correspond to maximum and minimum values for uniform distributions; the only exceptions are the water velocity and the stream parameters which have log-uniform distributions. Contaminant time is measured in years, leaching rate in years$^{-1}$, water velocity in meters/year, geosphere lengths in meters, flow rates in m$^3$/year and dose conversion factors in Sievert/moles.

parameters, the numerical code implementing the mathematical model is run once resulting in a final value that will represent the model’s output for the sampled set of parameters.

I developed a code that will not only automatize this procedure, but also do it in a parallel computer. In the field of parallel computation, this problem belongs to the class of problems sometimes called ‘embarrassingly’ parallel. This program’s task consists solely in performing a domain decomposition together with some simple load balancing duties. We were able to build a code that is simple, very flexible and easy to use with basically no cost in performance. The driver can take full advantage of the potential of concurrent processing in Monte Carlo simulations. A detailed description of the algorithm, its expected and observed performance analysis are presented in [Paper VI].
Chapter 3

Statistical methods

3.1 Hierarchical modeling

Some systems are better understood if instead of using an all encompassing model, one uses some kind of structure of simpler models, each related to some or all of the other models. This breaking down of a model into a set of simpler ones is what hierarchical modeling is about and in this situation conditional probability will come handy: one can model observables conditional on certain parameters, which can themselves be modelled conditional on another set of parameters. The idea is that by introducing some additional mathematical structure which brings forward some symmetries or characteristics of the system being modelled we can, in one stroke, improve the modeling abilities of the statistical structure and at the same time improve the model’s tractability. This concept will become even more clear when it is applied to the problem at hand.

This section is based mostly on work done by David Draper in [1] and [Papers IV and V]. It describes an abstract framework for model building that will enable Bayesian statistics to intervene effectively and give us tools to measure how different sources of uncertainty, for a given set of data, will influence our knowledge of the workings of the system we are studying. We will be considering uncertainty arising from mathematical structures, parameters of these structures and also from scenarios. We start by defining precisely what is meant by model, structure, parameters and scenario.

Structure will denote a particular mathematical function that will describe the behavior of a variable of interest. Parameters will be particular to each different structure. One can describe a model $M$ in two parts: a set of its structure assumptions $S$, and (for each structure) a set of model parameters $\theta$, so the model is completely defined by those two entities $M(S, \theta)$.

Now we will imagine a situation where one has more than one structure that can reasonably describe the data. Up until recently the usual procedure in research was to make an a priori choice of the structure we think is most appropriate and then try to optimize the number and values of the parameters of the structure that best describe the data. Most of the time one tries to quantify how much uncertainty in the structure’s parameters will influence inferences about the data (as in linear regression analysis and weighted least squares), but seldom does one try to go one step back and do the same thing to uncertainty in the structure (and then the model) itself. For example: one starts by assuming (say) a linear relationship between a response variable and some explanatory variables (and implicitly discarding any other types of relationships such as quadratic, logarithmic), and then concentrates only on improving that fit by tweaking the model parameters. What I consider here in the first place is not to discard competing mathematical forms for the model, but rather to acknowledge my level of uncertainty in the models and include all of them in our inferences or predictions.

To see how this works we start by putting everything we know in one ‘bag’ and call it $\mathcal{D}$ (data and any additional information one has on the system under study) and include all the things we want to know in another ‘bag’ and call it $y$ (e.g., future values of some quantities or parameters of some data distributions that are common to all model structures
under study). We will be ultimately interested in getting something like \( p(y|D) \), i.e., our uncertainty about the values we are interested in knowing (the \( y \)s), conditioning on the data/knowledge one has about the system now (the \( D \)).

If there are several different models that can explain the data correctly and if one includes all those models in the set \( M \), then we can say that we are provisionally conditioning on \( M \), i.e., we might like to work with \( p(y|D) \) but the best we can do in all honesty is to work instead with \( p(y|D, M) \). Using basic rules of probability we then have

\[
p(y|D, M) = \int_M p(y|D, M) p(M|D) dM = \int \int p(y|D, \theta, S) p(\theta|S, D) d\theta dS,
\]

(3.1)

Said in other words, the left hand side means: “What is the uncertainty in the quantities we want to learn about if we condition on the information from the data and a set of possible models?” The right hand side of the expression is the answer and consists of a weighted average of the conditional inferential/predictive distributions \( p(y|D, M) \), with weights \( p(M|D) \) (the posterior model probabilities, i.e., the plausibility of each model in the light of the data). The last expression can be broken down into finer detail:

\[
p(M|D) = c \ p(M) \ p(D|M) = c \ p(S, \theta) \ p(D|\theta, S) = c \ p(S) \ p(\theta|S) \ p(D|\theta, S),
\]

(3.2)

where \( p(S) \) is the prior probability of structure \( S \), \( p(\theta|S) \) is the probability distribution for each parameter once the model has been chosen and \( p(D|\theta, S) \) represents the likelihood function for the data. Here it is explicit that the choice of parameters depends on the structure. I used the definition of conditional probability to get the third equality.

In the equations introduced so far I always assumed a continuum of all possible structures, but in practical situations the closest we can often come to this goal is to choose a discrete set of structures, upon which to condition, that is as exhaustive as possible; call this set \( S = \{S_1, \ldots, S_m \} \). Then, \( p(y|D, M) \) (Equation 3.1) is approximated by

\[
p(y|D, S) = \sum_{i=1}^{m} \int p(y|D, S_i, \theta_i) p(S_i, \theta_i|D) d\theta_i = \sum_{i=1}^{m} p(S_i|D) p(y|D, S_i),
\]

(3.3)

where the conditional distributions \( p(y|D, S_i) \) are given by

\[
p(y|D, S_i) = \int p(y|D, S_i, \theta_i) p(\theta_i|S_i, D) d\theta_i;
\]

(3.4)

and the posterior structural probabilities \( p(S_i|D) \) can be expressed in the following way,

\[
p(S_i|D) = c \ p(S_i) p(D|S_i),
\]

(3.5)

where

\[
p(D|S_i) = \int p(\theta_i|S_i) p(D|\theta_i, S_i) d\theta_i.
\]

(3.6)

3.1.1 Model uncertainty in the Level E/G specification

In the last section we introduced an abstract framework for integrating several sources of uncertainty in inferential statistics. Here follows a description of the implementation of the above model uncertainty framework to the specific case of the Level E/G implementation [Paper IV] described in section 2.2. I will start by identifying the main statistical quantities in this new context.

- **Past Data** (\( D \)) consists of data from previous experiments or accidents in conditions relevant to the context of the GESAMAC project.
• **Future observables** \( (y^*) \) correspond to dose or flux values in the biosphere in the different situations considered.

• **Scenarios** \( (X) \) represent the different situations (with corresponding geosphere parameters) considered as those that might develop in the future. See [14] for more details.

• **Structural possibilities** \( (S) \) are the different mathematical descriptions of the system.

• **Parametric uncertainty** (for parameters \( \theta \)): some of the parameters of the numerical model will not be known with certainty and I will model their uncertainty by means of probability densities over the values that are admissible (these values are given in Tables 2.1 and 2.2).

• **Predictive uncertainty** \( (p(y|D, M)) \), would be translated as differences between predicted quantities and results from the experiments.

The assumed known quantities are: past data \( (D) \), the set of possible scenarios \( (X) \) and structures \( (S) \). The future observable outcomes \( (y^*) \) are unknown and to be predicted.

For each scenario the goal is to calculate the *scenario-specific predictive distribution*, \( p(y^*|S, x, D) \), (where \( x \) denotes a particular choice of a scenario), namely

\[
p(y^*|S, x, D) = \int_S \int \theta_d p(y^*|\theta, S, x) p(\theta|S, x, D) p(S|x, D) d\theta_d dS, \tag{3.7}
\]

where \( p(y^*|\theta, S, x) \) denotes the conditional predictive distribution for \( y^* \) given specific choices for scenario, structure and parameters (which in this case is the output from the deterministic contaminant transport model described before, GTMCHM), \( p(\theta|S, x, D) \) is the posterior distribution for the parameters and \( p(S|x, D) \) the posterior for the structure. We can also calculate the *composite predictive distribution*, \( p(y^*|S, X, D) \) (where \( X \) denotes the set of all possible scenarios), which is computed by averaging over all the scenarios:

\[
p(y^*|S, X, D) = \int_X p(y^*|S, x, D) p(x|D) dx; \tag{3.8}
\]

here \( p(x|D) \) denotes the posterior distribution for the scenario given the past data.

For all the posterior distributions mentioned so far there are the corresponding prior distributions. For example, the Bayesian description for the structural uncertainty (given past data and a specific choice of a scenario) is

\[
p(S|x, D) = c \ p(S|x) \ p(D|S, x), \tag{3.9}
\]

where the first distribution on the right hand side is the prior and the second is the likelihood. The structural parametric posterior distribution in 3.7 can be computed as

\[
p(\theta|S, x, D) = c \ p(\theta|S, x) \ p(D|\theta, S, x) \tag{3.10}
\]

and the posterior scenario distribution is

\[
p(x|D) = c \ p(x)p(D|x). \tag{3.11}
\]

In practice, the integrals above are approximated via Monte Carlo methods, as follows:

• Draw a scenario \( x \) at random (according to \( p(x) \)); these scenario prior probabilities were given at the end of Section 2.2;

• Select a structure according to a second probability distribution \( p(S|x) \) (here, since structural uncertainty is not being assessed, this distribution has mass 1 on the model specified by the differential equations given in Chapter 2);

• Choose a parameter vector \( \theta \) according to \( p(\theta|S, x) \).
• Approximate the actual radionuclide transport relevant to this \( \theta \) with the numerical code GTMCHEM (or any other mathematical structure that would be an acceptable description of the data), thereby calculating the predicted dose from \( p(y^*|S, x, \theta) \), or—in the case of the work here—the entire predicted dose curve.

If one puts the above list in a loop and continues for a long time and then collects the output in histograms at each time point, this would correspond to a Monte-Carlo integration to calculate the relevant predictive distributions 3.7 and 3.8.

This finishes the description of how to implement a uncertainty audit of a typical groundwater contamination problem.

In practice I found many difficulties implementing it. For example, using experimental data. The ideal situation would have been to design and perform an experiment with Level E/G specifically in mind, but one can quickly realize the complexity of such task. The next best thing would have been to get hold of some data from the scores of field experiments already performed around the world in situations resembling very much the Reference scenario described above.

Considering these difficulties, the implementation of these ideas branched out into different projects, each being a partial application of the whole framework presented above.

In Papers IV and V, one can see how a partial application of this framework would work in the absence of structure uncertainty and experimental data. In that study one can see how important scenario uncertainty is for the maximum magnitude of dose observed in the biosphere, following a contaminated underground water source. In practice Monte Carlo sample sizes of approximately 1,000 for each scenario were used; in the GESAMAC project we then re-weighted the results using the scenario prior probabilities to produce conclusions that aggregated across scenarios (this is a form of stratified sampling which will produce more accurate results, since many of the scenario probabilities were small).

Following that study I worked on the whole “dose versus time” curve instead of just its maximum. The results I obtained and the fertile questions raised are shown in [Paper VII] and in the later chapters of this thesis. In the functional data analysis presented here I focused on scenario-specific results, with the exception of the functional ANOVA to be described below.

A second project dealt with the inclusion of data in our inferential work. Considering the constraints of the experiment where this data was collected, the context of these new calculations is not comparable to the whole Level E/G case. So it is as if I did not consider scenario uncertainty (and restricted myself to the Reference scenario situation). Due to time constraints, it was not possible to expand the number of mathematical structures that could explain the data, so structure uncertainty couldn’t be addressed either. But I was able to do a great leap forward in Bayesian terms, since the previous project had restricted itself to studying the priors, now I could update these priors and study posterior distributions for some of the quantities of interest. This project was fundamental in proving that the integrals developed in the last section can indeed be tackled and produce results. Maybe not in its full glory, but still an important result.

3.2 Markov Chain Monte Carlo

As mentioned in the last section, finding experimental data that would fit the Level E/G case study was not possible due not only to its inherent theoretical nature, but also to the peculiarities of the field of risk analysis in the nuclear waste community; experiments are technically difficult and more importantly very expensive to set up. So experimental results are very difficult to come by in general, and more so if one needs data with very stringent specifications.

However, my training as a physicist never let me forget the importance of experimental results in the process of scientific thought. In this chapter I will include the beginning of a work whose main objective is to show that the integration mentioned at the end of the last
chapter could in principle be performed with some kind of data. The data was taken from an old paper referring to a pumping experiment in the WIPP site [7]. Therefore the whole practical context of this section is different from the rest of the work.

My main goal was to produce posterior distributions for the most important parameters of the one dimensional transport equation: the water velocity, the dispersion and the retention. In statistical notation this means getting a probability distribution for the physical parameters $\theta$ given the data $\mathcal{D}$, $p(\theta|\mathcal{D})$. These distributions are an improvement over the usual curve fitting procedures (as least squares for example), because instead of producing a single number for each parameter of interest (the value of best fit) it gives us a whole probability distribution for the range of values of each parameter that can describe the data. Considering the high degree of uncertainties mentioned before I think this procedure to be a more sensible and honest way to produce estimates.

As it turns out, it is often not so trivial to develop the analytical form of this posterior distribution. Specially for high dimensions of the vector of parameters $\theta$, calculation of a subtle (but important) quantity like the normalizing constant $c$ of the posterior function 3.10, $p(\theta|\mathcal{D}) = c \ p(\theta) \ p(\mathcal{D}|\theta)$, can be a challenging task. This constant is needed to make sure the posterior is normalized to 1 (as any probability distribution should), and it is equal to $c^{-1} = p(\mathcal{D}) = \int p(\mathcal{D}, \theta) d\theta = \int p(\mathcal{D}|\theta) \ p(\theta) d\theta$. If $\theta$ is a vector of parameters, this last integral will be multidimensional, and therefore quite complicated to calculate analytically.

This is just the beginning of difficulties; to calculate a simple expected value for a specific $\theta_j$, for example,

$$E[\theta_j|\mathcal{D}] = \int \theta_j \ p(\theta_j|\mathcal{D}) \ d\theta_j$$

we would again come across a k-dimensional integral. ¹ One of the most important distributions for risk analysis, the predictive distribution for the next observation $y_{n+1}$, 

$$\int p(y_{n+1}|\theta) p(\theta|\mathcal{D}) \ d\theta,$$

would also be plagued with a k-dimensional integral. The problem of evaluating these integrals has been referred to as the central technical challenge of Bayesian statistical work for the past two and a half centuries by David Draper [2].

There are several techniques to calculate integrals like the ones mentioned above, and I elected to use Markov Chain Monte Carlo (MCMC) because of its power and flexibility.

Very succinctly, MCMC amounts to carefully building a stochastic process whose realizations form a Markov chain whose long-run, or stationary, or equilibrium, distribution is the posterior distribution (equivalent to simulating random (correlated) draws from this distribution). Under certain conditions this process will generate a sample from which one can extract all the information that is of use to us: means, standard deviations, correlations, histograms or kernel density estimates, telling us whatever is of interest about the posterior distribution.

I will start by constructing carefully the parts of the posterior distribution that we know about, beginning with the prior distribution for the parameters and then tackling the trickiest part, the likelihood.

### 3.2.1 The prior

The priors for parameters of interest are all chosen here to be flat, i.e. giving equal probability to all the possible values of the parameters in physically meaningful regions. This

¹The marginal distribution for $\theta_j$ is a $k-1$ dimensional integral of the form,

$$p(\theta_j|\mathcal{D}) = \int \ldots \int p(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k|\mathcal{D}) \ d\theta_1 \ldots d\theta_{j-1} \ d\theta_{j+1} \ldots d\theta_k.$$
is equivalent as saying that we know very little before obtaining the data. Of course, one of the strengths of the Bayesian approach is in fact being able to include all types of prior information, but in this case we will take a conservative approach and have non-committal priors and let the data lead the way. Flat priors sound very reasonable and sometimes even desirable but in practice we should use them with care. For example, if we naively choose a uniform distribution between the allowable values of each parameter, this might lead to divergent integrals for the posterior, specially if the information content of the likelihood function is low. There are ways to go around this like, for example, using a Gamma distribution with very small parameters, this leads to a very flat distribution in the support of values of interest, but it does not lead to divergences in the posteriors. Nevertheless, we will soon see that we do not expect such difficulties to appear in the case I studied.

### 3.2.2 The likelihood function

We first start with the relationship we expect to exist between the dose of contaminant, at the end of its dispersion-transport path, time and all the other physical parameters. We take the solution for the one-dimensional transport equation given in [9] and shown in this document before (Equation 2.10).

The data we had available was taken well after leaching stopped so it is safe to assume that the response will look very much like the second of the last two equations (Equation 2.10).

In statistical terms we will say that the observed data will be the result of whatever the deterministic model says, but with “noise” superimposed on it. In other words,

\[ y_i = f(D, R, V, t_i) + e_i \]  \hspace{1cm} (3.12)

where \( y_i \) will denote the natural logarithm of the observed dose, \( f(D, R, V, X_i) \) denotes natural logarithm of Equation 2.10, \( e_i \) will denote the random error, and we will assume that the error for one measurement will be independent but identically distributed to the errors in any other measurement and that they will be normally distributed with mean zero and standard deviation \( \sigma \) (these assumptions would not be reasonable on the raw dose scale, because symmetric errors around the predicted dose would go negative at early time periods, but may well be reasonable on the log scale).

The likelihood will tell us how likely are the values given each of the parameters of interest in the light of the data obtained. It is basically a joint distribution for all the data, \( P(Y_1 = y_1, \ldots, Y_n = y_n) \), but reinterpreted as a function of the parameters for fixed data values. In the model description above we have already made the assumption that each observed value would be like a sampling from a normal distribution centered at the expected value from 2.10 with a standard deviation of \( \sigma \), \( Y_i \sim N(f_i, \sigma^2) \), \( i = 1, \ldots, n \). We are also assuming that all data points are independent from each other, so the joint distribution is just the multiplication of each marginal distribution,

\[ P(Y_1 = y_1, \ldots, Y_n = y_n) = P(Y_1 = y_1) \cdots P(Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i) \]

Also assuming each data point is identically distributed we will obtain for the final form of the likelihood,

\[ l(D, R, V, \sigma | D) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(y_i - f(D, R, V, t_i))^2}{2\sigma^2} \right) = \sigma^{-n} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(D, R, V, t_i))^2 \right) \]  \hspace{1cm} (3.13)

where \( D \) represents the whole set \( y_i \). This function multiplied by the chosen prior will constitute the posterior distribution of interest. Remember that this is not yet a proper distribution because it is not normalized. It is this difficulty (among others) that will make us resort to MCMC methods to obtain the true characteristics of it.
3.2.3 Maximum Likelihood Methods

There are two ways to interpret function 3.13: before we ever observe any data this represents how the data will be likely to behave if you were to sample from the normal distribution, and then the likelihood can be thought of as being a function of $D$ for fixed parameters $D,R,V$ and $\sigma$ (as was mentioned in the last section). Once having observed some data we can say that $D$ is now fixed, and the likelihood could be looked upon as a function of the parameters. This interpretation opens the doors to Maximum Likelihood Methods (MLM) to perform inference work about the parameters. The practical outcome of these methods is the same for both interpretations; we will obtain a set of values for the function parameters that will best fit that function to the data.

I will use MLM only as part of a batch of exploratory techniques to help explore the posterior distribution. Since I am using flat priors for my posterior, I expect the latter to look very much like the likelihood, so any work that helps me to know the likelihood will certainly give at least good tips on how the posterior will look like. For example, we can use MLM to produce maximum likelihood estimators for the parameters $D,R,V$ and $\sigma$. These numbers are obtained by maximizing the likelihood (hence the name...) and will therefore be a good approximation to the maximum of the posterior itself. This way one will be more sure of starting the MCMC simulations in points of interest of the posterior and therefore avoid having to wait for the chain to look for them by itself.

Maximizing Equation 3.13 is the same as maximizing its logarithm, which I will gladly do, since it improves its numerical implementation,

$$\ln[l(D,R,V,\sigma|D)] = \frac{1}{2\sigma^2} \sum_{i=0}^{n} (y_i - f_i)^2 - n \ln(\sigma),$$

It is evident that MLM is equivalent to the least squares procedure when the measurement errors are independent and normally distributed (the second part in the last equation is just a constant).

3.2.4 Markov Chains

In general a Markov Chain (MaCh) is a stochastic process that unfolds in time in such a way that the past and future states of the process are independent, given the present state (in order to decide where to go next the chain does not need information where it has been, it just needs to know where it is now). More formally, the probability of the chain moving to a new state $\theta_{t+1}$ will depend only on the present state ($\theta_t$),

$$P(\theta_{t+1}|\theta_0, \ldots, \theta_t) = P(\theta_{t+1}|\theta_t).$$

Let’s emphasize that the decision whether to move to a new state or not is done using a stochastic rule, which we will call the jump distribution and denote by $J_t(\theta_{t+1}|\theta_t)$.

It is not just any Markov Chain that will be of use to us. Our interest will be centered in MaChs that are irreducible (no matter where the chain starts, the chain has a positive probability to reach any other state in a finite number of steps), aperiodic (there is no subset of states where the chain will indefinitely be stuck in) and not transient (for any state that it starts from, it has probability 1 of returning to it after a finite time). These three properties will warrant that any simulated sequence by this chain will converge to a unique stationary distribution. The first property holds if the jump distribution is able to eventually jump to all states with positive probability. The last two properties will hold for any stationary distribution that is proper.

There is one last condition that we must make sure holds so that we can say that not only are we sure the MaCh will settle on a unique distribution, but that distribution is the one that we are interested in (in our case, the posterior $p(\theta|D)$). This will depend specifically on the specific properties of the jump distribution, so now it is a good time to introduce the
simplest one. In the next subsection I will introduce one of the simplest implementations of MCMC called the **Metropolis algorithm**, and for this case I will show that its stationary distribution will indeed be our desired posterior distribution for the parameters \( \theta \).

### 3.2.5 The Metropolis algorithm

Given an approximation to the distribution we are interested in (in this case \( p(\theta|\mathcal{D}) \)) that we assume can be computed up to a normalizing constant, the following procedure will converge to it.

1. Choose a starting value \( \theta_0 \). There is a whole list of alternative ways to cleverly choosing good starting points (see [6]).

2. For \( t = 1, 2, \ldots \)
   
   (a) choose a candidate new \( \theta^* \) from what we will call a jump distribution, \( J_t(\theta^*|\theta^{t-1}) \). This distribution is not yet the transition distribution mentioned above. The only requirement for the jump distribution is that it is symmetric, \( J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a) \).
   
   (b) Calculate the ratio
   
   \[
   r = \frac{p(\theta^*|\mathcal{D})}{p(\theta^{t-1}|\mathcal{D})}
   \]

   (c) Set
   
   \[
   \theta_t = \begin{cases} 
   \theta^* & \text{with probability } \min(r,1) \\
   \theta_{t-1} & \text{otherwise}
   \end{cases}
   \]

   The actual transition distribution will be a mixture of the jump distribution and a point mass at \( \theta_t = \theta_{t-1} \).

   These rules basically implement a search algorithm where a new move will always be accepted if it increases the posterior density, and only sometimes be accepted even if it decreases it.

   Let’s show that with this specific implementation of a MaCh we will be sure that its stationary distribution will be our sought after posterior.

   If the jump distribution is symmetric \( (J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)) \), where the states \( \theta_a \) and \( \theta_b \) are such that \( p(\theta_b|\mathcal{D}) \geq p(\theta_a|\mathcal{D}) \) (notice here the intervention of our target distribution) then the joint probability \( p(\theta_{t-1} = \theta_a, \theta_t = \theta_b) \) will be symmetrical also. In this case the marginal distributions of \( \theta_t \) and \( \theta_{t-1} \) will be identical, so \( p(\theta|\mathcal{D}) \) will be the stationary distribution of this MaCh.

### 3.3 Functional data analysis

Statistical models are often applied to sets of data with a single outcome variable and we have indeed performed such studies in this same context before [Papers IV and V], where we studied the values of maximum dose. In fact the deterministic model (GTMCHM) produces much more informative output than that: among other things it produces a collection of values for radiological dose for different time points, and this for a fixed point in space, which I

\[
p(\theta_{t-1} = \theta_a, \theta_t = \theta_b) = p(\theta_a|\mathcal{D}) \ J_t(\theta_b|\theta_a)
\]

and

\[
p(\theta_t = \theta_a, \theta_{t-1} = \theta_b) = p(\theta_b|\mathcal{D}) \ J_t(\theta_a|\theta_b) \frac{p(\theta_a|\mathcal{D})}{p(\theta_t|\mathcal{D})} = p(\theta_a|\mathcal{D}) \ J_t(\theta_a|\theta_b)
\]
take to be the one at the biosphere. This collection of data points can be seen to approximate a continuous curve of dose versus time. In this section I describe statistical methods that are useful when the outcome of interest is an entire function rather than just a single numerical summary of the function. The goal is to introduce a set of statistical tools that will study the variability of radiologic dose with time. These tools will be the generalization of familiar concepts like Principal Components Analysis, ANOVA-like calculations and linear regression as applied to functions. The section provides the technical background for [Paper VII].

### 3.3.1 Principal components analysis

The general idea of Principal Components Analysis (PCA) (on scalars rather than functions) is a simple and attractive one: given a data set, identify the main modes of its variability, and in this way (hopefully) reduce the number of variables that can explain the data. For example, for multivariate data the central mathematical quantity in PCA is a linear combination of variable values,

$$f_i = \sum_{j=1}^{P} \eta_j y_{ij}, \quad i = 1, \ldots, N,$$

(3.14)

where \( y_{ij} \) denote the values of the variable \( j \), \( \eta_j \) are weights that have to be calculated, \( p \) is the number of variables, and \( N \) is the number of data values.

The calculations are conceptually simple: one starts by trying to find a set of weights \( \eta_j \) for which the mean of the squared values of \( f_i \) \( \left( \sum f_i^2 / N \right) \) is maximized (subject to the constraint \( \sum_j \eta_j^2 = ||\eta||^2 = 1 \), otherwise the sum of squares could be made arbitrarily large). In other words we are trying to find a new combination of the original variables which will give rise to maximum variability of the weighted sum.

We can also re-write 3.14 using inner-product notation, \( f_i = \langle \eta, y_i \rangle, i = 1, \ldots, N \) where \( \eta \) is the vector \((\eta_1, \ldots, \eta_p)\) and \( y_i \) is the vector \((y_{i1}, \ldots, y_{ip})\). This notation brings the perspective from linear algebra, i.e., that this expansion can be seen as a projection of the data values into a new set of basis variables, and we are simply trying to find which set of variables “exposes” the biggest variability. But the procedure does not stop here; we would be interested in finding other secondary modes of variability, and to achieve that we can use the same idea: in subsequent steps (say in step \( m \)) one computes new weight vectors \( \eta_m \) such that the values \( f_{im} = \langle \eta_m, y_i \rangle \) will be maximized, but this time in addition to the usual constraint that \( ||\eta_m||^2 = 1 \) we must have a second constraint,

$$\langle \eta_k, \eta_m \rangle = 0, \quad k < m,$$

(3.15)

which means that each new set of weights has to be perpendicular to all the other vectors of weights calculated previously. Each set of values of the vector projections \( f_{im} \) is called the principal component scores; these typically reflect how much of the character of each component of the new basis the data will have.

In practice, if one subtracts the mean from the data values, maximizing the mean square of the principal component scores corresponds to maximizing their sample variance. An alternative way to calculate principal component analysis involves calculating the eigenvalues and eigenvectors of the covariance matrix. Let the \( N \times p \) matrix \( X \) contain the values \( x_{ij} \) and let the vector \( \eta \) of length \( p \) contain the weights for a linear combination. Then the mean square criterion for finding the first principal component weight vector can be written as

$$\max_{\eta' = 1} ||f_i||^2 = \max_{\eta' = 1} N^{-1} \eta' Y' Y \eta,$$

(3.16)

where the equality comes from the fact that \( f_i = Y \eta \). Since the \( p \times p \) sample variance-covariance matrix \( V \) is equal to \( N^{-1} Y' Y \), the above criterion can be re-written as

$$\max \eta' V \eta,$$

(3.17)
and its solution is given by finding the largest eigenvalue $\rho$ of the eigenequation $V\eta = \rho\eta$. There will be a whole sequence of eigenvalue-eigenvector pairs $(\hat{\rho}_j, \eta_j)$ satisfying this last equation, with all eigenvectors naturally orthogonal to each other, and for each $j$ the corresponding eigenvector $\eta_j$ will also satisfy the original maximization condition 3.16. So, in conclusion, the problem of finding the principal components of a multivariate data set is equivalent to the problem of solving the eigenequation $V\eta = \rho\eta$.

**Functional PCA (FPCA)**

When the data consists of a set of curves, the procedure for PCA can be properly generalized and one can still obtain modes of variability for the curves. The main concepts are maintained from the last section but instead of variable values one has function values $y_i(s)$, so that instead of a discrete index $j$ for each variable, one has a continuous index $s$ (i.e., instead of $p$ variables one has the function values). The dot product is generalized to an integral form,

$$f_i = \int \eta(s)y_i(s)ds = \langle \eta, y_i \rangle,$$  \hspace{1cm} (3.18)

and $||\eta||^2$ will represent the squared norm $\int \eta^2(s)ds$ of the function $\eta$. So, as one can see, the integral makes sure that the whole curve is taken into account when calculating the scores.

The eigenequation that will let us calculate each principal component is generalized, in the functional data space, as

$$\int v(s,t)\eta(t)dt = \rho\eta(s),$$ \hspace{1cm} (3.19)

where the covariance function is

$$v(s,t) = N^{-1}\sum_{i=1}^{N}y_i(s)y_i(t),$$ \hspace{1cm} (3.20)

and $\eta(s)$ is the functional equivalent of the weights introduced in the previous section ($\eta$ is now an eigenfunction instead of an eigenvector).

The integral in Equation 3.19 can also be seen to represent an integral transform $V$ of the weight function $\eta$ with the kernel of the transform $v$ defined by

$$V\eta = \int v(\cdot,t)\eta(t)dt.$$ \hspace{1cm} (3.21)

This integral transform is called the covariance operator $V$ and it allows us to re-write the eigenequation in the following form

$$V\eta = \rho\eta.$$ \hspace{1cm} (3.22)

This is exactly the same general equation described above for the multivariate case, with the differences that each factor in the equation is suitably generalized.

So, as we will able to see clearly when I apply these concepts to data, by plotting the new basis functions $\eta_i$ one can gain some insight into what they really mean. For example, a higher absolute value of the function means higher variability. Also, if a data function has high, positive scores for a specific basis function, then that means that there is strong agreement between the basis function and the data function, since the scores are basically a dot product which involves the projection of one function on the other.
### 3.3.2 Linear modeling

**Univariate ANOVA**

Sometimes it is useful to look at the data as belonging to different groups and try to get some information from the comparison between them. Most of the time, we want to see if the fact that they belong to different groups is even relevant or not. The most popular technique to answer this question is called ANOVA (Analysis Of VAriance); here I introduce its common usage for multivariate datasets and the generalization for functional data.

The general situation is when we have $g$ random samples of size $n_g$, from each of $g$ populations:

$$
\begin{align*}
\text{population 1} & : \mathbf{Y}_{11}, \ldots, \mathbf{Y}_{1n_1} \\
\text{population 2} & : \mathbf{Y}_{21}, \ldots, \mathbf{Y}_{2n_2} \\
\vdots & \\
\text{population } g & : \mathbf{Y}_{g1}, \ldots, \mathbf{Y}_{gn_g}
\end{align*}
$$

where it is assumed that each sample $\mathbf{Y}_{ij}$, $i = 1, 2, \ldots, g$, is a random sample of size $n_i$ from a population with mean $\mu_i$, $i = 1, 2, \ldots, g$, and that samples from different populations are independent from each other. It is also assumed that all populations have the same covariance matrix $\Sigma$ and each population follows a multivariate normal distribution.

Here the data are univariate in the sense that each $Y_i$ is a scalar instead of a vector. So we will have random samples $Y_{ij}$ from $g$ different populations, each distributed as a $N(\mu_i, \sigma^2)$ with $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, n_i$. The ANOVA procedure starts by re-writing the mean of each population in such a way that one can split that mean in two parts: one that will reflect the overall behavior of the data, and another part which will be intrinsic to the population the data is assumed to originate from. Once that is accomplished, one has simply to find a meaningful way to compare those two parts and see if they are fundamentally different from each other.

We rewrite the mean of each population

$$\mu_i = \mu + (\mu_i - \mu), \quad (3.23)$$

where $\mu$ is the overall mean. We denote the quantity between brackets as just $\tau_i$ and call it the treatment effect. The response $Y_{ij}$ can then be re-written as

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where the $e_{ij}$ are independent and follow a $N(0, \sigma^2)$ distribution. By imposing the constraint

$$\sum_{i=1}^{g} n_i \tau_i = 0$$

we can be sure that the model parameters $\tau_i$ will be uniquely defined; otherwise our model will have more parameters than are needed, i.e., we will have an overdetermined system of equations.

We can obtain the same relationship for the estimated quantities and that will lead us in a natural way to the core of ANOVA:

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

where $\bar{y}$ is an estimate of $\mu$, $\hat{\tau}_i = \bar{y}_i - \bar{y}$ is an estimate of $\tau_i$ and $y_{ij} - \bar{y}_i$ is an estimate of the error $e_{ij}$. We can further subtract $\bar{y}$ from both sides and square it,

$$(y_{ij} - \bar{y})^2 = (\bar{y}_i - \bar{y})^2 + (y_{ij} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i), \quad (3.24)$$
summing over \( j \) (within each population),
\[
\sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = n_i (\bar{y}_i - \bar{y})^2 + \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,
\]  
(3.25)
where the third part on the right side of 3.24 disappears because
\[
\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0.
\]
If we further collect all the sum of squares (SS) of each population and sum them we get the following quantities,
\[
\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{g} n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.
\]
Total, corrected SS between samples SS Within samples SS
Finally, if we expand the square on the left side and leave only the sum of squared observations, we get
\[
\sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}^2 = \sum_{i=1}^{n_i} \bar{y}_i^2 + \sum_{j=1}^{g} n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.
\]
SS of observations SS of the mean SS between samples SS of residuals
The arrays representing the mean, treatment effects and residuals are orthogonal. Using these simple algebraic manipulations we can reduce the comparison between \( g \) quantities to the comparison between just two quantities: the SS of the overall mean and the SS between samples. This is usually done by testing a null hypothesis of equality of means
\[
H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0
\]
by comparing the ratio of sum of squares between and within groups with a suitable percentile of the \( F \)-distribution with appropriate degrees of freedom.

**Functional ANOVA (FANOVA)**

Once again the functional generalization will look very much like the usual ANOVA, with just some of the components having different meanings. We still decompose the overall mean in different parts, but every parameter will be dependent on time. In the level E/G context I am interested in bringing out the influence of each scenario in the overall mean dose, so if \( D_{kl}(t) \) denotes the \( k \)th curve of dose as a function of time, for scenario \( l \), we can split the overall mean in the same way as in the previous subsection:
\[
Dose_{kl}(t) = \mu(t) + \tau_l(t) + \epsilon_{kl}(t).
\]  
(3.26)
Here \( \mu \) is the grand mean as a function of time, indicating the average dose curve across all scenarios, \( \tau_l \) are the effects specific to each scenario \( l \), and \( \epsilon_{kl} \) will account for variability not accounted by the previous two quantities, i.e., variability from one curve to another, but within the same scenario. In order for the \( \tau_l \) to be uniquely determined we need to introduce the additional condition \( \sum_l \tau_l(t) = 0 \).

This subsection is under the heading of linear modeling because one can look at 3.26 as being a model for the data, and we can separate the data according to whether it belongs to a certain scenario or not by using some ingenuity in building the design matrix \( \mathbf{Z} \). This matrix is composed of dummy variables whose columns are constructed in the same way as the matrix \( \mathbf{Z} \) in the previous section, so that \( \mathbf{Z} \) has number of columns equal to the number of scenarios plus one, and has number of rows equal to the number of curves. If additionally we define a matrix \( \beta = (\mu, \tau_1, \tau_2, \ldots, \tau_l) \), then we can re-write model 3.26 as follows,
which basically just re-states the model in matrix terms. The vectors involved in these equations all contain entities that vary with time, i.e., they contain functions instead of numbers. This equation looks exactly like the one developed above for the univariate case.

The functional equivalent of the least squares criterion can be developed quite easily; for example the squared norm represented as

\[ \| \text{Dose} - Z\beta \|^2, \]  

(3.28)
is maintained as long as we can define an appropriate *functional* norm. Each component of \( Z\beta \) can be re-written as the inner product of the vectors \( z_{k,l} \) and \( \beta \), where the first vector is the transpose of each single row of the design matrix \( Z \), so \( \langle z_{k,l}, \beta \rangle \) (the inner product) will represent the role of predictor variables and fitting the linear model will amount to finding the \( \beta \)s that will minimize the quantity

\[ \sum_i \sum_k \int \| \text{Dose} - \langle z_{k,l}, \beta \rangle \|^2 dt, \]  

(3.29)
but subject to the constraint that all the scenario effects sum to zero, i.e., \( \sum_{i=0}^n \tau_i = 0 \).

**Regression analysis**

This method consists of predicting the value of one or more response (dependent) variables from a collection of predictor (independent) variable values and also assessing the influence of the latter on the former. This is put into practice by building a statistical model in the following way:

\[
Y = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_r x_r + \epsilon
\]

response = mean (depending on \( x_1, x_2, \ldots, x_r \)) + error.

Here \( Y \) is the response variable, the \( x_i \) are the predictor variables, \( \epsilon \) will model random errors, and the \( \gamma_i \) will be coefficients that will weigh how much of each variable will contribute to reproducing the behavior of the response variable. The latter are found by calculating which set of \( \gamma \)'s minimizes the quantity

\[
SS = \sum (Y - \gamma x)^2,
\]

which is the *sum of squares*; this measures the discrepancy between the actual observed values \( Y \) and the weighted values of the \( x \), with “candidate” weights \( \gamma \).

When the response variable is a whole function and also the predictor variables are functions, the linear model will be changed to

\[
y_i(t) = \alpha(t) + \int_{T_x} x_i(s)\gamma(s, t)ds,
\]

(3.30)
where the subscript \( i \) denotes each of \( n \) experimental curves, \( \alpha \) is a constant term (in the sense that it does not depend on any explanatory variable), \( x_i(s) \) denotes each of the explanatory functions, and \( \gamma \) are the weights to be determined as the optimum values that minimize the sum of squares.

When looking at Equation 3.30 one should interpret it cross-sectionally (in time) only for the response variable, i.e., the left side of 3.30 represents the variability of each time point \( t \) of \( y_i \). The right side of the equation will try to explain that variability but not cross-sectionally; it will be through a weighted average over all the time points of the explanatory variables. In summary, the linear model presented above will explain each time point of the response variable by a weighted average of all the time points of the explanatory variable. If we subtract out \( \alpha(t) \) from \( y_i(t) \) the sum of squares will be denoted by

\[
\| \text{Y}_{\text{data}} - \text{Y}_{\text{fitted}} \|^2 = \sum \int_{T_x} [y_i(t) - \alpha(t) - \int_{T_x} x_i(s)\gamma(s, t)ds]^2 dt,
\]


and our job is to calculate the $\gamma(s, t)$ that minimize this last quantity (see [17] for details on how this is done in practice).

The data in this thesis yield slightly simpler equations since my explanatory variables are all constant in time. This means that the linear model looks a little different:

$$y_i(t) = \alpha(t) + \sum_{j=1}^{m} x_{i,j} \gamma_j(t), \quad (3.31)$$

where $i$ spans through all the functional data, there are $m$ explanatory variables and the regression coefficients $\gamma$ are found by the usual least square method, i.e. by minimizing the sum of squares with the constraint that $\sum_j \gamma_j = 0$,

$$\|Y_{data} - Y_{fitted}\|^2 = \sum_i \int [y_i(t) - \sum_{j=1}^{m} x_{i,j} \gamma_j(t)]^2 dt \quad (3.32)$$

The usual multivariate sum of squares for non-functional response variables looks very similar,

$$\|Y_{data} - Y_{fitted}\|^2 = \sum_i [y_i - \sum_{j=1}^{m} x_{i,j} \gamma_j]^2$$

The only difference being the presence of an integral over time for the former expression (and that the response variable and regression coefficients are not functions). In the non-functional response situation the regression parameters are the ones that minimize the squared discrepancies for all variables and all data taken together. In this sense the functional generalization will introduce one extra entity to “sum” over; time.

Also, Equation 3.32 can be interpreted as (looking at it from the inside to the outside) first being a squared measure of discrepancy between the response function and a linear combination of functions, and the integral will just summarize that discrepancy for the whole length of the support of those functions (being effectively a distance between two functions). So the least square procedure is just a search for the set of functions that will reduce that discrepancy. We still have a global measure of discrepancy rather than cross-sectional regressions for each time point.
Chapter 4

Markov Chain Monte Carlo. Results

4.1 Experimental setup and experimental data

The experiment was set up as follows (see [7] for a more complete description). It took place in the Waste Isolation Pilot Plant (WIPP) near Carlsbad, New Mexico. It consisted in a two-well recirculation tracing test, which was performed using an extraction-injection well couplet as one can see in Figure 4.1.

![Diagram of the experiment setup](image)

Figure 4.1: Experiment setup for the SCN tracer experiment of [7].

This couplet was located in the Rustle formation of the site. The two wells were 23 meters away from each other, the injection well (IW) was drilled through to a depth of roughly 202 meters and the pumping well (PW) was drilled to 227 meters, leaving the culebra dolomite member of the formation as the liquid-bearing zone between the two of them. From other tests, the culebra average transmissivity was estimated to be $0.06 \text{ [m}^2/\text{d}]$ and the storage coefficient to be about $1.3 \times 10^{-5}$.

The pumping was allowed to reach a steady state before starting the tracer injection. The tracer used was sodium thiocyanate (SCN). The injection lasted for 28 days at an average rate of $1.9 \text{ ml/min}$. Pumping continued for 274 days. Below in Figure 4.2 you can see a plot of the breakthrough curve.

4.2 Exploratory analysis

In order to obtain the posterior distribution for the physical parameters of interest (water velocity and dispersion) for the experiment mentioned before, we decided to use Markov
Chain Monte Carlo simulation techniques.

It is good common sense in studies like these [6] to have good starting values for the “random walk” that the MCMC performs. One of the ways to obtain those is by looking at the likelihood as being a function of the parameters of interest (in this case; water velocity, dispersion and measurement error $\sigma$ as introduced in 3.2.2) and try to obtain the numerical values of these parameters that maximize the likelihood. Since we have flat priors, this will most likely start up the MCMC machinery very close to the mode of the posterior, and in this way start the random walk in the place of greatest interest for us.

I used several numerical tools for this search for the maximum of the likelihood function. I started by transforming the likelihood to the log scale and then used a combination of graphical plots and the steepest descent numerical method to look for the maximum of the function. It is well known [15] that neither assures us that the maximum will be a global one, but in my case I also had one additional aid: two of the parameters have very definite physical meaning, so the space of their possible values gets very restricted. For example, for the water velocity, it’s obviously restricted to positive numbers and very unlikely to be larger than 10 [m/day]. The same happens for the dispersion, with an even more restricted range of possible values.

Initially the log-likelihood was a function of 4 parameters, but the retention, the water velocity and the dispersion are linearly dependent. This can be seen by looking directly at the advection-dispersion differential Equation 2.4; dividing it by $R$, for instance, will not affect the shape of the solution, or conversely, multiplying these 3 variables by the same constant will not change the solution. The tracer used in the experiment was not expected to interact with the rock, so I assumed a value for retention to be equal to 1, and maintained that value constant during all the calculations \footnote{I also performed some maximization calculations using $R$ as a variable, and the best solution quoted the value 1.5 for $R$.}

One can try to maximize the log-likelihood analytically, by differentiating it with relation to each parameter ($V, D$, and $\sigma$) and finding which values of these variables make the derivatives equal to zero. The solution for $\sigma$ is the simplest and looks like

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - f_i(\hat{V}, \hat{D}))^2}{n},$$  \hspace{1cm} (4.1)$$

where $y_i$ is the concentration measured in the experiment, $f_i(\hat{V}, \hat{D})$ the concentration calculated from the analytical solution 2.10 with the maximum likelihood estimates of $V$ and $D$ plugged in, and $n$ the number of measurements done in the experiment.
We can plug this into the expression for the log-likelihood and in this way further reduce its number of parameters, and now the log-likelihood will be a function of only two parameters, making it easier to study its shape in the parameter space.

\[
ll(V, D) = \frac{n}{2} - \frac{n}{2} \ln \frac{\sum_{i=1}^{n}(y_i - f_i(V, D))^2}{n}
\]

(4.2)

**Figure 4.3: Contour plot of the likelihood function with R held constant at 1. Broad plotting region for the parameter space.**

I started by doing plots with very broad ranges for the parameters, and very quickly I found patches of considerable size in the parameters space that led to physically unacceptable solutions (concentrations smaller than zero) (see Figure 4.3). These patches are the product of numerical errors that are difficult to deal in the logarithmic scale. The truth is that these pathological patches are not in regions of physical interest, and the MCMC machinery can produce meaningful results in spite of this.

The broad view of the log-likelihood shows several maxima, which I explored carefully until I found what looks like a global one in the region \( \{ V \in (0.0, 0.02) \text{ [m/day]} \cup D \in (0.5, 0.9) \text{ [m}^2/\text{day}] \} \) (see Figure 4.4).

The main feature of Figure 4.4 is an obvious non-linear relationship between \( V \) and \( D \). This will in fact reduce the efficiency of the MCMC simulation unless one does one of two things: either we find a transformation of the variables such that the two new variables are roughly linearly dependent, or we change the MCMC search algorithm to take into account the non-linear relationship between the variables.

The first option is the easiest to implement (as long as the nonlinear relationship is not too complex). In fact transforming both variables \( (V, D) \) to the logarithmic scale produced almost linear relationship between them (see Figure 4.5). This figure shows that the log-likelihood, when re-interpreted in a Bayesian way as an approximation to the posterior distribution with flat priors, is very much like the log-likelihood of a bivariate normal distribution, which is relatively easy to sample from using MCMC.

Once all these preparations were done, the log-likelihood function looked like Figure 4.5. I then used several methods (a simple simulated annealing program, and steepest descent) to numerically look for the maximum. Both methods gave roughly the same numerical estimates for \( V \) and \( \sigma \): 0.01279 [m/day], 0.63778 [m²/day] and 0.33080 [mg/l], respectively. In Table 4.1 I show the maximum likelihood estimators (MLE) for all the quantities of interest. The correlation between \( V \) and \( D \) was very strong and was calculated to be equal to \(-0.892\).
4.3 MCMC results

Once in possession of good starting values for the MCMC simulation, I had only to specify the priors for the parameters. To keep things simple I assumed no prior information and used completely flat distributions for every parameter. In general this raises questions, since in situations where the likelihood does not contain much information, one can be confronted with non-integrable posteriors. In my situation, the number of data points allows the likelihood to dominate the posterior, and therefore to keep it proper, so I do not expect any problems in this respect.

I used a version of Metropolis sampling called random-walk Metropolis in which the jump distribution was a multivariate normal distribution centered at each iteration at the current value of the chain and with a covariance matrix that was a multiple of the covariance matrix for the maximum-likelihood estimates. This approach creates an optimal or nearly-optimal Metropolis sampler for posterior distributions whose surface plots look like Figure 4.6, provided that the multiplier is chosen to achieve an acceptance rate of around 30% when
there are three parameters to sample.

I used the usual 1000 iterations as burn-in before starting to monitor the statistical parameters of the simulation. I performed a total of 100 000 iterations in order to obtain 5 significant digits for the estimator of $\sigma^2$. The acceptance rate for the simulation was close to the optimal value for a 3-parameter run (nearly 30%). The whole run took 4918 seconds on a 400MHz Sun workstation.

In the next three figures (Figs. 4.7, 4.8, 4.9) I show the time series produced for each of the parameters of interest ($V$, $D$ and $\sigma$), their marginal posterior distributions, autocorrelation and partial autocorrelation functions. The time series seems reasonably well mixed, ie. it does not seem to be spending too long on the same region of the parameters space, and it is also not jumping too wildly from one value of the parameter to another.

The autocorrelation at lag $k$ of a time series measures the extent to which the series at time $(t + k)$ and at time $t$ are linearly related for $k = 1, 2, \ldots$. In order to understand the meaning of this quantity we can first imagine two lists containing the values produced by the simulation, then if we shift one of the lists by $k$ positions (and dropping the $2k$ that will have no corresponding values to be compared to in the other list), then this correlation will be the one calculated between these two lists. The sample estimate of this quantity is calculated by

$$ r_k = \frac{c_k}{c_0}, $$

where

$$ c_k = \frac{1}{m - k} \sum_{t=1}^{m-k} (\theta_t^* - \bar{\theta})(\theta_{t+k}^* - \bar{\theta}) $$

(4.4)

If we want the standard error for $\sigma$ to be equal to 0.00005 – assuring in this way 5 digit precision in its estimate – then

$$ SE(\sigma) = 0.00005 = \frac{\tilde{\sigma}}{\sqrt{m}} \sqrt{\frac{1 + \hat{\rho}}{1 - \hat{\rho}}}, $$

(4.3)

once the autocorrelation of lag 1 ($\hat{\rho}_1$) is known, we can solve the equation for $m$ (the number of runs to be performed).
Figure 4.7: Markov Chain Monte Carlo output and diagnostics for water velocity (c.f. text).

Figure 4.8: Markov Chain Monte Carlo output and diagnostics for dispersion c.f. text).
to relate the sampled values $\theta^*_i$ with its values at previous times $p$,

$$\theta^*_i = \alpha_1 \theta^*_{i-1} + \ldots + \alpha_p \theta^*_{i-p} + e_i$$  

(4.5)

if $e_i$ is white noise error with mean 0 and variance $\sigma^2$, then $\theta^*_i$ is called an autoregressive process of order $p$ ($AR_p$). If there is any correlation left unexplained by the linear model we can diagnose that by looking at the residuals of the previous linear model, and the partial correlation function measures it. So, if a series is $AR_p$, then the partial correlation function will be significantly different from zero at lags $1,\ldots, p$ and close to zero for lags larger than $p$. The lower right panels in these 3 plots all show $AR_1$ series for all parameters of interest. It’s good to know that the series for a given parameter is $AR_1$ because a simple formula is then available for calculating how long the monitoring run needs to be to achieve a target Monte Carlo accuracy for the posterior mean of that parameter (see the footnote on page 41).

We can also plot the sampled values for each parameter against each other, as I show in Figure 4.10. It is not surprising to see that it has a shape very similar to the log likelihood plot, since the data dominates the posterior distribution.

Finally I show a very interesting picture; it helps to visualize the implications of the uncertainty in parameters. We used all the sampled values for each parameter, used them to produce the corresponding curves and superimposed them on the experimental data (Figure 4.11). It is great visual evidence that the analytical solution is actually quite inflexible and will never give a perfect description of the data.

As a final comment, the uncertainty bands for $D$ and $V$ are quite reasonable even if the fit is not that perfect. This study is more important not in the specific results that produces, but in showing that we have a very powerful tool that works and should allow us to perform the useful integrals mentioned in Section 3.1.
Figure 4.10: Plot of sampled values for the two parameters of interest, dispersion and water velocity.

Figure 4.11: Curves resulting from every sampled values of water velocity and dispersion.
Chapter 5

Conclusions

The main, strongest conclusion at the end of this study is that parametric, model and scenario uncertainties can be brought together and expressed in a coherent, logical mathematical framework. Better still, within that framework we can integrate those uncertainties in inferential work.

The first three papers constitute a first attempt to explore the effect of different parametric uncertainties in the prediction of contamination accidents. In that preliminary work I did not have yet a complete framework on which to integrate different sources of uncertainty, but in each instance we can show indications of important consequences arising from different forms of parametric uncertainty.

For instance, in [Papers I and II] there are strong indications that uncertainty in the characteristics of the far field media (such as longitudinal dispersion, diffusivity, wet surface and different forms of adsorption) can have important consequences in the predicted dose for the biosphere. Comparison of simulations with and without uncertainty in those parameters, showed increased spread in the distribution of mean and the median doses expected at the end of the contaminant paths.

In [Paper III] the exploratory work was continued, but in this case the main interest was in seeing how different numerical implementations of the uncertainty in the contaminant path could have impact on the peak dose observed at the end of those paths. In that study, spatial variability of the media properties is implemented by comparing simulations where each single pathway has the same properties for its entire length to simulations where those properties are allowed to vary for that same length. It is then observed that the two sets of simulations lead to significant differences in the shape, the spread and the center of the peak dose distributions.

The conceptual context in which [Papers IV, V and VII] were done is different, but the goal is the same. They are all done in a more general framework both in terms of the geophysical system and its possible future scenarios (set up by the Level E/G specification in [Paper IV]) and also in mathematical terms (which is set up by the hierarchical model also specified in [Papers IV and V]).

It was shown that MCMC methods can be applied in this field, by using a simple analytical solution to a transport equation, and therefore produce uncertainty bands for physical parameters, based on that solution and experimental data. The simulation produced pleasingly acceptable values for the parameters of interest, showing that this elegant technique can work with relative ease. No major conceptual difficulty can be seen in including uncertainty in the mathematical models (by using different, competing models). This future work is one of the most exciting things to look forward to, and the next logical step. The way seems clear to producing the full integration of all sources of uncertainty in a case study.

FDA techniques were also applied successfully to dose versus time curves, producing a plethora of information which still has not been fully explored. Functional data analysis is indeed a very powerful method capable of producing results that require considerable interpretive effort. The ability to provide so much extra information, especially when compared with approaches based on single numerical summaries of entire functions, is so great, that
we have not yet come close to exhausting all of the possible substantive conclusions.

FDA is so powerful that it makes many results almost self evident (like the modes of variability of the curves), but the truth is that there was no idea they were there before performing this analysis. Many other results were not that clear, and it was the careful thinking induced by the plots produced by the FDA methods that brought to light many interesting aspects of the statistical characteristics of the dose functions, among which we can count: the nature of the asymmetry between the left and right tails of the dose curve, the tendency of certain harmonic types of variabilities to occur in conjunction with other types of harmonic 2 variabilities, the scenario effect and the clustering of the FP and AG scenario on one side and the REF, GA and HDE scenarios on another, and the inversion in the behavior of regression parameters over time associated with the same type of variables but for different geosphere layers. Also, the fact that by expanding the typical Reference scenario to 6 different ones has a big impact on the variability of the curves. The "opening up" of uncertainty in the scenarios can increase the curve variability by 40 times, as shown in my FANOVA calculations! Even more surprising, this increase in variability happens on the part of the curve to the left of the maximum, showing very explicitly something that had only been hinted in previous calculations I had done; the uncertainty in the dose curve on the way to the maximum seems to be qualitatively different from the way the dose curve relaxes back to zero. These conclusions appear innovative in the field of groundwater contamination.

Of course, there are lots of improvements I would like to be able to perform in this study: first of all, having access to laboratory or site data of real contamination experiments/accidents. This would allow me to fully study predictive uncertainty, as noted in Section 3.1. Secondly, I could benefit from further developments in FDA theory, especially in the ability to formulate a standardized version of the functional linear regression model (which is still under investigation by Prof. Ramsay), and I would certainly like to perform a functional regression analysis with explanatory variables that would vary in time; FDA theory is ready to tackle that kind of problem and its results would be an important step forward in groundwater contamination issues. It is evident that FDA techniques have the potential to shed new light on the field of environmental risk assessment.

Finally the most ambitious goal would be to produce a parametric, model and scenario uncertainty integration using the full dose vs. time functions combining in this way all the tools that were used in this study. Combining the power of these innovative tools should be the final goal, and one believed to yield important scientific returns.
Appendix A. ANOVA calculations using linear modeling

Here I show how one can use linear modeling to perform ANOVA calculations.

There is an alternative way of setting up the comparison of each population’s mean with the overall mean, and that involves using linear modeling (regression) with dummy variables. In the following I begin by expressing the problem in terms of a linear model and my aim is to show that the least squares procedure will lead us to the same result as the one just presented above. The advantage of this approach is that the ANOVA calculations will come out automatically from the least square procedures, which is easily implemented in a computer. This next part is based on the work of Draper and Smith in [3].

We will model the response variable explicitly with one explanatory variable $Z_i$ per treatment:

$$E(Y) = \mu Z_0 + \tau_1 Z_1 + \tau_2 Z_2 + \ldots \tau_g Z_g,$$

where we are still assuming there are $g$ different types of treatments/groupings of the whole data set, and $E(\cdot)$, denotes expected value of the quantity between brackets. This equation has to express the fact that when we take a data value $Y_{ij}$ from group $i$ it must have expectation $\mu + \tau_i$; in order for that to happen we have to construct each $Z_i$ in a particular way. This becomes clearer by using matrix notation; let $Y$ be a column matrix with the first $n_1$ entries filled with all the data values from group 1, the next $n_2$ with all the data values from group 2 and so on:

$$Y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{g1} \\ y_{g2} \\ \vdots \\ y_{gn_g} \end{pmatrix}$$

The matrix containing the values of the explanatory variable is constructed in a special way to ensure that the expected value of $y_{ij}$ has the right decomposition just mentioned. This can be done by filling the first column of $Z$ with only 1s; the second column has its first $n_1$ entries filled with 1s and the rest with 0s; the third column has 0s in its first $n_1$ entries, 1s in the following $n_2$ entries and 0s in all the rest of the column; the fourth column will have its first $n_1 + n_2$ entries filled with 0s, the next $n_3$ entries filled with 1s and the remaining filled with 0s; and so on, with one column per group. Then $Z$ will have $g + 1$ columns and $\sum_{i=1}^{g} n_i$ rows.
Here I use the dashes to make more visible the grouping of rows for each group of data. Lastly we can define a column vector \( \beta \) containing all the means,

\[
\beta = \begin{pmatrix}
\mu \\
\tau_1 \\
\tau_2 \\
\vdots \\
\tau_g
\end{pmatrix}
\]

Equation 5.1 can then be re-written in matrix notation,

\[
E(Y) = Z\beta. \tag{5.2}
\]

Our goal is to estimate the \( \tau_i \). Now that we have expressed the problem in the form of a linear model, it is straightforward to apply the usual methods in this area to find its solutions.

The normal equations are a well known result from least squares fitting of curves; their solution gives us the \( \hat{\beta}_i \) that will minimize \( Z'Z \hat{\beta} = Z'Y \). In my case \( Z'Z \) will look like

\[
Z'Z = \begin{pmatrix}
N & n_1 & n_2 & \cdots & n_g \\
n_1 & n_1 & 0 & \cdots & 0 \\
n_2 & 0 & n_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
n_g & 0 & 0 & \cdots & n_g
\end{pmatrix}
\]

and

\[
Z'Y = \begin{pmatrix}
N\bar{y} \\
n_1\bar{y}_1 \\
n_2\bar{y}_2 \\
\vdots \\
n_g\bar{y}_g
\end{pmatrix}
\]

If we call \( \hat{\beta}_0 \) the least squares estimate of \( \mu \), and \( \hat{\beta}_i \) the least squares estimates for each \( \tau_i \), and we gather them in a column matrix \( \hat{\beta} \), the normal equations \( (Z'Z)\hat{\beta} = Z'Y \) will produce a system of equations looking like

\[
N\hat{\beta}_0 + n_1\hat{\beta}_1 + n_2\hat{\beta}_2 + \cdots + n_g\hat{\beta}_g = N\bar{y} \\
n_1\hat{\beta}_0 + n_1\hat{\beta}_1 = n_1\bar{y}_1 \\
n_2\hat{\beta}_0 + n_2\hat{\beta}_2 = n_2\bar{y}_2 \\
\vdots \\
N\hat{\beta}_0 + \vdots = n_g\bar{y}_g
\]
Here we can see the result of the aforementioned over-determination of the original linear equation: the first equation is not linearly independent of all the other equations, in fact it is equal to their sum and this will lead to a singularity of the matrix $Z'Z$. It is at this point that we take the restriction

$$\sum_{i=1}^{g} n_i \tau_i = 0$$

I mentioned in the beginning of this section. This restriction will supply us with a needed extra equation and dropping (for instance) the first equation from $Z'Z$ will guarantee that our system will have a unique solution. It will look like

\[
\begin{align*}
\hat{\beta}_1 & + n_2 \hat{\beta}_2 + \cdots + n_g \hat{\beta}_g = 0 \\
n_1 \hat{\beta}_0 & + n_1 \hat{\beta}_1 = n_1 \bar{y}_1 \\
n_2 \hat{\beta}_0 & + n_2 \hat{\beta}_2 = n_2 \bar{y}_2 \\
\vdots & \quad \quad \vdots = \\
N \hat{\beta}_0 & + n_g \hat{\beta}_g = n_g \bar{y}_g
\end{align*}
\]

which, in matrix notation, is

\[
\begin{pmatrix}
0 & n_1 & n_2 & \cdots & n_n \\
n_1 & n_1 & 0 & \cdots & 0 \\
n_2 & 0 & n_2 & \cdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
n_g & 0 & 0 & \cdots & n_g
\end{pmatrix}
\begin{pmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\vdots \\
\hat{\beta}_g
\end{pmatrix} =
\begin{pmatrix}
0 \\
n_1 \bar{y}_1 \\
n_2 \bar{y}_2 \\
\vdots \\
n_g \bar{y}_g
\end{pmatrix}
\]

The solution is simple; from equations 5.3 we get,

$$\hat{\beta}_i = \bar{y}_i - \hat{\beta}_0,$$  \hspace{1cm} (5.4)

and using this in the first equation in the same set of equations,

$$\sum_{i=1}^{g} \bar{y}_i \hat{\beta}_i = 0 \quad \iff \quad \sum_{i=1}^{g} n_i (\bar{y}_i - \hat{\beta}_0) = 0 \quad \iff \quad \sum_{i=1}^{g} n_i \bar{y}_i + \hat{\beta}_0 \sum_{i=1}^{g} n_i = 0 \quad \iff \quad \bar{y} - N \hat{\beta}_0 = 0 \quad \iff \quad \hat{\beta}_0 = \bar{y}$$

so finally (going back to 5.4),

$$\hat{\beta}_i = \bar{y}_i - \bar{y},$$  \hspace{1cm} (5.5)

which turns out exactly as it should be: each $\hat{\beta}_i$ (which represents the estimated treatment effect $\tau_i$) should just be the difference between the estimated group mean and the estimated overall mean, and $\hat{\beta}_0$ equals the estimated overall mean.

The ANOVA procedure involves the use of the sum of squares of the previous quantities, and the sum of squares can actually be obtained from the normal equations, even if $ZZ'$ is singular. The SS is $\hat{\beta}'Z'Y = NY^2 + \sum n_i (\bar{y}_i - \bar{y})^2$, in which the sums of squares for the overall mean and SS for “between groups”, respectively, are clear. The sum of squares for “within groups” is obtained by the difference $Y'Y - \hat{\beta}'Z'Y$. 
Acknowledgements

Work done partly within the GESAMAC project (European Commission- Contract nr. FI4W-CT95-0017) and partly with a scholarship from the Portuguese Fundação para a Ciência e Tecnologia under the program Praxis XXI (reference PRAXIS XXI/BD/16031/98). I would like to thank Prof. Christoph Bargheer, Dr. Antonio Pereira and Prof. David Draper (University of California at Santa Cruz), the persons I learn most from during my work. In particular Prof. Draper, whom on top of being the scientist I most admire and being fun to work with, I consider to be a dear friend. I would also like to acknowledge the important support given by Prof. Jim Ramsay from McGill University.
Bibliography


Paper I
Paper II
Paper III
Paper IV
Paper V
Paper VI
Paper VII