QUANTIFYING AND VISUALIZING UNCERTAINTY OF
3D GEOLOGICAL STRUCTURES WITH IMPLICIT METHODS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF GEOLOGICAL SCIENCES
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Liang Yang Candidate
June 2021
© 2021 by Liang Yang. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/zy244dn1573
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jef Caers, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Peter Kitanidis

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tapan Mukerji

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract

Geological structures significantly contribute the complex interaction of physical processes in subsurface systems. The evaluation of the spatial distribution of geological structures in the subsurface is crucial for various applications, so sophisticated methods are needed to model and visualize geological structures in 3D. However, uncertainties are unavoidable for these 3D models, due to sparsity and imprecision of data, as well as people’s lack of geological understanding. Both methodological and computational challenges exist in addressing uncertainties of 3D geological structures. This thesis addresses these challenges, by presenting new practical methods for quantifying and visualizing the uncertainty of geological structures with implicit methods.

To enhance people’s communication and perception about structural uncertainty, a new method based on the idea of stochastic motion is proposed in Chapter 2. Geological surfaces are represented as the addition of trend functions, initialized with signed distance functions, and residual functions, subject to constraints of data and geological age relationships. Two sampling approaches to create the stochastic motion are used: Monte Carlo and Markov chain Monte Carlo (McMC). The uncertainty is assessed by independent realizations drawn by Monte Carlo sampling. The uncertainty is visualized by a “smooth” movie of gradually evolving geological surfaces that have the same stationary distribution as Monte Carlo realizations, sampled by McMC. This idea is integrated into the level set equation, implemented with “velocity extension” from the level set method. The method is first illustrated with simple synthetic 3D examples, taking the constraints of data and geological age relationships into consideration. Finally, the method is illustrated using a synthetic data set from a copper deposit, where denser drillholes constrain an ore body with
seven different lithologies. The method provides a direct assessment and visualization of the uncertainty of 3D geological surfaces.

For handling more complex cases with even denser data and more geological rules, Chapter 3 introduces a framework to model large-scale geological structures. This chapter addresses how a single large-scale geological structural model that fits dense data and honor multiple geological rules is built. Implicit methods for modeling geological structures such as stratigraphy and faults have been developed for more than a decade. The implicit potential field method is such a method that is capable of incorporating multiple types of data, including contact points for geological boundaries and orientations. The implicit potential field method relies on the solution of a co-kriging system. However, applying the method to 3D modeling of large-scale geological structures constrained to dense data and strict geological rules remains challenging. Due to the non-stationary and complex nature of large-scale geological structures, and difficulty in estimating an adequate variogram model, performing global interpolation with all dense data together may create artifacts that are geologically unrealistic. Therefore, a framework that uses a divide-and-conquer strategy is proposed. The core idea is to create intermediate 3D geological models that match subsets of data and then recombine them into a single large 3D geological model, while maintaining data and geological rule constraints. The chapter proves that linear combinations of potential fields preserve properties of conditioning (i.e., matching data). The framework is successfully applied to model the stratigraphy model of a large banded iron formation (BIF) in Western Australia with dense boreholes, but scarce orientation measurements.

As an extension of the uncertainty visualization and assessment tool presented in Chapter 2, and the large-scale modeling framework presented in Chapter 3, Chapter 4 introduces an efficient Bayesian framework to quantify the uncertainty of implicit geological structures with geophysical data. Geophysical data provide critical information and constraints for validating subsurface models, so it is common to integrate geophysical data for solving inverse problems. Bayesian frameworks are often needed for quantifying uncertainty of 3D geological structures in inverse problems, but challenges exist, due to the high dimensional
nature of spatial models, inefficiency in building multiple geological models, and non-linearity of the forward model. Implicit representation of geological structures transforms discrete geological objects into a continuous variable, i.e., a scalar field; dimension reduction techniques such as principal component analysis (PCA) can be applied because of the implicit representation. Therefore, with the implicit representation of geological structures, PCA is applied for reducing the dimension of an ensemble of realizations of geological structures. After dimension reduction, the prior distribution is formulated in the dimension reduced space. Results show that computing time is saved when sampling new model realizations from the low dimensional space. Rejection sampling and Metropolis-Hastings sampling algorithms are designed to work in the case. The method is demonstrated with a mineral-hosting region in Western Australia with gravity data. Results show that both sampling algorithms work well in sampling posterior realizations of implicit geological structures.
Acknowledgements

Obtaining a PhD was probably the most challenging and rewarding thing I have ever accomplished in my life. I had a great opportunity to meet and work with some of the smartest and nicest people at Stanford University. Here, I would like to take the opportunity to express my most sincere gratitude towards all people who have helped me along the journey.

Firstly, I am deeply grateful to my advisor, professor Jef Caers, for guiding and supporting me through the entire PhD life. Jef has given me so many valuable advices in research work, including ways of posing good research questions, tackling challenges and communicating the results. Working with Jef has given me so much personal growth as well. I also thank professor Tapan Mukerji for his numerous and valuable suggestions given to my research work since my first year of PhD.

I am also grateful to so many talented and friendly people in the Stanford Center for Earth Resources Forecasting (SCERF). SCERFers, Lijing Wang, Noah Athens, Tyler Hall, Yizheng Wang, David Yin Zhen, Anshuman Pradhan, Marcelo Brasil Silka and many others, are my go-to people for suggestions in research and coursework. I would like to thank Peter Achtziger-Zupančič and Francky Fouedjio for sharing ideas and tackling great challenges together when we were collaborating on the same research project.

Last but not least, I am truly grateful to my family and friends for their continuous support and encouragement. They are always my last harbor. I thank my parents. I have been able to pursue all of my dreams, no matter big or small, with their love and support. I thank all my friends, Nana, Sam, Ruilin, Daniel and many others, for encouraging me when the PhD life becomes daunting. Besides, I would like to express my special thanks to my guitar for accompanying me all the time when completing the dissertation.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>vii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background and Scope</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Dissertation Outline</td>
<td>3</td>
</tr>
<tr>
<td>2 Assessing and Visualizing Uncertainty of Implicit Surfaces</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Methodology</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Level set method</td>
<td>9</td>
</tr>
<tr>
<td>2.2.2 Definition of the stochastic model</td>
<td>12</td>
</tr>
<tr>
<td>2.2.3 Level set implementation of the stochastic model</td>
<td>12</td>
</tr>
<tr>
<td>2.2.4 Modeling the mean function</td>
<td>15</td>
</tr>
<tr>
<td>2.2.5 Multiple level sets</td>
<td>16</td>
</tr>
<tr>
<td>2.2.6 Stochastic motion</td>
<td>18</td>
</tr>
<tr>
<td>2.3 Illustration cases</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1 Simulation results and uncertainty assessment</td>
<td>21</td>
</tr>
<tr>
<td>2.3.2 Output statistics: volume and surface roughness</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Case study: synthetic copper deposit</td>
<td>23</td>
</tr>
<tr>
<td>2.4.1 Data description</td>
<td>23</td>
</tr>
</tbody>
</table>
## 3 Implicit Modeling of Large-scale Geological Structures

3.1 Introduction ........................................................................... 30

3.2 Case study ........................................................................... 32

3.2.1 General background ......................................................... 32

3.2.2 Regional geology ............................................................. 33

3.2.3 Geology of the studied deposit: “Deposit 99” .................... 34

3.3 Methodology ...................................................................... 35

3.3.1 Potential field method ..................................................... 35

3.3.2 Issues with direct application of interpolation .................... 39

3.3.3 Overview of the entire approach ..................................... 40

3.3.4 Selection of subsets of contact points ................................. 42

3.3.5 Simulation of orientation data ........................................... 43

3.3.6 Validation ..................................................................... 44

3.3.7 Local linear combination of two potential fields ................. 47

3.3.8 Linear combination of multiple potential fields .................. 48

3.3.9 Walk-through of key steps with 2D synthetic example ........... 49

3.4 Full application to case study ............................................... 52

3.4.1 Data processing and analysis .......................................... 52

3.4.2 Input data generation ...................................................... 55

3.4.3 Intermediate 3D model generation and validation ............... 56

3.4.4 Linear combination of multiple implicit models .................. 60

3.4.5 Discussion on the runtime and RAM ............................... 60

3.5 Conclusions .................................................................... 62

## 4 Quantifying Uncertainty of Implicit Geological Structures

4.1 Introduction ...................................................................... 64
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2</td>
<td>Case study</td>
<td>67</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Geological background</td>
<td>67</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Gravity data and gravity forward modeling</td>
<td>68</td>
</tr>
<tr>
<td>4.3</td>
<td>Methodology</td>
<td>71</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Overview</td>
<td>71</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Prior distribution definition</td>
<td>73</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Prior distribution estimation</td>
<td>75</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Likelihood function definition</td>
<td>77</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Sampling</td>
<td>78</td>
</tr>
<tr>
<td>4.4</td>
<td>Application to case study</td>
<td>81</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Prior realizations and prior distribution</td>
<td>81</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Sampling results</td>
<td>82</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Runtime discussion</td>
<td>89</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussions</td>
<td>95</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusions</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>Conclusions and Future Work</td>
<td>98</td>
</tr>
<tr>
<td>5.1</td>
<td>Conclusions</td>
<td>98</td>
</tr>
<tr>
<td>5.2</td>
<td>Summary of contributions</td>
<td>100</td>
</tr>
<tr>
<td>5.3</td>
<td>Future research directions</td>
<td>100</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Bayesian inference of model parameters</td>
<td>100</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Integration of faults uncertainty</td>
<td>101</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Incorporating other types of data</td>
<td>101</td>
</tr>
<tr>
<td>A</td>
<td>Gradual deformation as a McMC sampler</td>
<td>102</td>
</tr>
<tr>
<td>B</td>
<td>Co-kriging system of the potential field method</td>
<td>106</td>
</tr>
<tr>
<td>C</td>
<td>Data fitting preservation with linear combination</td>
<td>109</td>
</tr>
</tbody>
</table>
### List of Tables

2.1 Runtimes for the 3D object example and the copper case .............. 29

3.1 Stratigraphic column of “Deposit 99” .................................. 37

3.2 Runtimes of time-consuming steps (for 5 intermediate models) ....... 62

4.1 Series names, lithology names and density values used for modeling .... 68

4.2 Parameters for the interpolator ........................................... 74

4.3 Reduction of standard deviation (SD) of lithology volumes (unit: number of voxels) .................................................. 89
List of Figures

2.1 Level sets represent surfaces implicitly as isocontours of higher-dimensional functions, in contrast to explicit surface methods. Here we compare the representations for the 2D circle \( x^2 + y^2 = 4 \). (Left) The zero contour of the level set paraboloid \( z = x^2 + y^2 - 4 \) is the desired circle (several contours in the \( z = 0 \) plane are drawn, with different colors indicating regions between contours). (Middle) An explicit (also called Lagrangian) representation of the circle is a collection of points (red) and edges (thick black). (Right) Computationally, discrete samples (red) of the level set are stored in a uniform grid data structure; interpolation of these values is used to identify a desired contour. .......................................................... 10

2.2 Different isocontours (blue) of a level set function (red) result in different interfaces (blue lines). The implicit nature of level sets makes it natural to perform topological operations like merging simply by shifting level sets or, with multiple level sets, via operations such as \( \min (\phi_1, \phi_2) \) (union) or \( \max (\phi_1, \phi_2) \) (intersection). (Source: (Gibou et al., 2018)) . . . . . . . . . . 11

2.3 Illustration of velocity extension. (a) shows a sphere (black circle) defined by its signed distance function \( \phi(\vec{x}) \). (b) shows a Gaussian realization for the entire grid. (c) shows the extended velocity field. Values (velocities) on the sphere are extracted and extended in the normal direction. For instance, the velocities at locations \( \vec{x}_1, \vec{x}_2 \) and \( \vec{x}_3 \) are equal to the velocities at \( \vec{x}'_1 = \vec{x}_1 - \phi(x_1)\vec{N}, \vec{x}'_2 = \vec{x}_2 - \phi(x_2)\vec{N} \) and \( \vec{x}'_3 = \vec{x}_3 - \phi(x_3)\vec{N} \). . . . . . . . . . 14
2.4 Comparison of the results of using velocity extension (middle) and directly adding a Gaussian random field (right) to the implicit function of a base surface (left). The same Gaussian realization is used for both methods in the test.

2.5 Illustrates the surface reconstruction process of (Zhao et al., 2001), which is used for initial model generation. (Left) 1,000 points from the surface of a 3D sphere of radius 10 centered at the center of the $100^3$ domain, a uniform Cartesian grid with resolution $\Delta x = 0.5$ are randomly sampled. The initial guess for a surface that fits these points is a box that fills the entire domain (i.e., the grid itself). (Middle) The fast tagging method efficiently refines the initial guess to a coarse approximation of a surface. (Right) After 30 steps of convection, the reconstructed surface appears quite smooth. The surface is not perfectly spherical due to the sparsity of the input points and the limited grid resolution (there is inherent $O(\Delta x)$ error in the final result).

2.6 A geological age relationship example. (a) Four level sets are used to represent four lithologies, which have an age relationship indicated in the figure; (b) Independent initialization of each level set without truncation.

2.7 Illustration of the gradual deformation of the residual function. The residual function is a normal extension of the Gaussian realization defined on the sphere in Figure 2.3.

2.8 Initial surfaces and velocity fields. (a) The initial three objects and three drill-holes are visualized together. Realizations of velocities are generated using three boreholes as conditioning data. Three realizations generated with (b) sill 5 and range 20, (c) sill 10 and range 20, and (d) sill 5 and range 60 are shown. The colors in the realizations represent the velocity magnitudes.
2.9 Snapshots of surfaces at time 60 (first column) and at time 90 (second column), sections of geological domains (third column) and the entropy maps (fourth column). The first three rows are the results of the conditional stochastic motion with sill 5 and range 20, sill 10 and range 20, and sill 5 and range 60, from top to bottom. The last row shows the results of the unconditional stochastic motion with sill 5 and range 20.

2.10 The change of volume proportion with three parameter settings. The horizontal axis shows time steps and the vertical axis shows the volume proportion of the domains separated by three surfaces.

2.11 The change of variance of mean curvature (vertical axis) w.r.t. time step (horizontal axis) with three parameter settings.

2.12 Synthetic copper data set. (a) shows 50 drillholes. (b) shows an interpreted model of the area. The interpreted model shows a complex shape. (c) a section view of the interpreted model. (d) shows the geological age relationship (old to young from the top to the bottom). Seven colors in (d) correspond to seven lithologies in the data.

2.13 Realizations of the synthetic copper case. (a) and (b) are two realizations from the stochastic motion (Lithology 1 is made transparent). (c) and (d) are sections of the models that show the geological age relationship. (e) and (f) are two realizations of Lithology 7. Note that all figures use the same color scheme as Figure 2.12 (d).

2.14 Probability and entropy maps. (a) is the probability that any location belongs to Lithology 4 (probabilities below 0.4 are made transparent). (b) is the entropy of all surfaces (entropies below 1.2 are made transparent).

3.1 All structural data in deposit “99”. All contact points (small pixels), orientations (large pixels) for modeling and the corresponding stratigraphic column (also see Table 1). Spatial extent: 5840 meters in EW, 3026 meters in NS, and 448 meters in depth.
3.2 Topographical map with certain, inferred and important faults. .............. 38

3.3 Issues of the model obtained by a direct application of the potential field method on the large BIF case. Shown in a 2D section, taken at Y = 247200 m. .................................................. 40

3.4 Framework for modeling large-scale geological structures. ........................ 42

3.5 Effect of orientation information on the interpolation result (only enriched BIF layers D1 - J6 are shown and they are all conformable layers). (a) Contact points (small pixels) and orientations (large pixels); (b) the interpolation result (Y = 26700 m, in 2D) with only observed orientations; (c) the interpolation result (Y = 26700 m, in 2D) with simulated orientations but without dividing the contact points. ......................... 45

3.6 Automatic geological rule check based on the topology of iso-contours. (a) a verified example based on the geological rule; (b) a rejected example based on the geological rule. ........................................ 46

3.7 Example of local linear combination. (a) $\mathbf{m}_1$ and (b) $\mathbf{m}_2$ are linearly combined with (c) $\mathbf{w}$ to get (e) the linearly combined potential field, and the linearly combined potential field is converted to (f) the linearly combined stratigraphy model using (d) spatially varying thresholding values (only the thresholding value for D1 is shown). ........................................ 48

3.8 Synthetic 2D example. (a) True model; (b) 500 data points; (c) Kriging result; (d) Kriging variance. .................................................. 50

3.9 One example of (a) sub datasets, (b) kriging results and (c) kriging variances. 51

3.10 Validation of 500 Kriging results. (a) First two dimension of PCA scores of 500 kriging results (blue dots) and left-out data (red cross); (b) Scree plot of PCA; (c) Mahalanobis distance between 500 kriging results (green to blue dots) and data (red diamond). ................................. 52

3.11 Comparison of linear combination results and kriging ($r = 5, L = 100$). (a-b) Two linear combination results; (c) Ensemble average of 100 linear combination results; (d) Kriging result with the whole dataset. .............. 53
3.12 Results with $r = 10$, $L = 50$. (a) Validation based on data; (b) Kriging result; (c) Linear combination result.

3.13 Observed gradients and their statistics. (a) Histograms of gradients; (b) Variograms for each component of the gradient. $G_x$, $G_y$, and $G_z$ from left to right. Variograms types: exponential, exponential, exponential; nugget effects: 0.08, 0.12, 0.02; sills: 0.12, 0.17, 0.025; ranges: 200, 200, 200 ...

3.14 Two subsets of contact points selected from all contact points.

3.15 Scatter plots of estimated dips vs observed dips with relative errors less than 0.5.

3.16 One cosgsim realization of three components of gradients. (a) $G_x$; (b) $G_y$; (c) $G_z$.

3.17 Data based validation for the BIF case. From left to right are plots of PC scores, cumulative explained variance and the Mahalanobis distance. Red cross and red diamonds indicate the observed data. (a) Validation using all stratigraphies’ data; (b) Validation using only data of TD3; (c) Validation using only data of D1.

3.18 Three intermediate models using subsets of data and simulated orientations. (a) One section of $m_1$; (b) One section of $m_2$; (c) Three sections of $m_3$ with data.

3.19 Probability map based the boundary of D4 from intermediate models.

3.20 Final models after linear combination. (a-b) two final models; (c) the fault block; (d) model in 2D with correct rules; (e) shows the consistency with data in 3D.

4.1 The studied sandstone greenstone belt and data (Wellmann et al., 2017). (a) Geological map of the studied region (adapted from Wellmann et al. (2017); (b) synthetic observed gravity anomaly; (c) contact points and orientations inferred from various sources of information, and grid points for gravity measurements (black dots).
4.2 Reference model generated using Gempy. (a) Reference model in 3D surfaces. (b) Top view of the reference model. .......................... 70

4.3 Gradients of observed gravity data. (a) Gradient in X direction; (b) Gradient in Y direction. ......................................................... 72

4.4 Features that are concatenated for PCA. (a) Potential field for BIF series; (b) potential field for Simple Mafic Series; (c) potential field for Early Granite Series; (d) Isocontours for representing lithology boundaries. ........ 76

4.5 Prior construction results. (a) Prior parameters (contact points and orientations); (b) Prior model realizations (lithology models); (c) Prior data realizations (forward gravity responses). ............................ 83

4.6 PCA of prior realizations and prior gravity responses. (a) PC scores of prior realizations in 2D; (b) PC scores of prior gravity responses and the observed gravity in 2D; (c) Cumulative explained variance of each PC component of prior realizations; (d) Cumulative explained variance of each PC component of prior gravity responses. ........................................ 84

4.7 Eigen-images of each potential field (Only first three PCs are shown). (a-c) Eigen-images of the BIF series’ potential fields; (d-f) Eigen-images of the Simple Mafic series’ potential fields; (g-i) Eigen-images of the Early Granite series’ potential fields. ............................................. 85

4.8 PCA loadings of each variable. (a) PCA loadings of the 1st and 2nd PCs; (b) PCA loadings of the 3rd and 4th PCs; (c) PCA loadings of the 5th and 6th PCs. ................................................................. 86

4.9 Global sensitivity analysis for PCs. Gravity responses are more sensitive to PCs that are indicated by red colors (darker red indicates higher sensitivity). 87

4.10 Kernel density estimation of sensitive PC scores. (a) PC scores in 3D (only 9th, 12th and 8th PC scores are shown); (b) PC scores colored by estimated densities (only 9th, 12th and 8th PC scores are shown). ................. 87
4.11 Metropolis-Hastings sampling of the prior PC scores. (a) Markov chain of
PC scores in 3D (with a burn-in period); (b) Independent MH samples in
first and second dimensions (burn-in samples are dropped); (c) Independent
MH samples in 50th and 51st dimensions (burn-in samples are dropped);
(d) Comparison of prior data realizations and MH samples of data real-
izations (burn-in samples are dropped); (e) Comparison of prior gravity
response profiles (green) with MH samples (orange) along X = 72500m.

4.12 Autocorrelation analysis for MH sampling. (a) Correlated MH samples
of 1st PC scores (prior) at each iteration that shows the convergence; (b)
correlated MH samples of 1st PC scores (posterior) at each iteration; (c)
autocorrelation analysis of MH samples (prior) that shows an autocorre-
lation length of about 1400; (d) autocorrelation analysis of MH samples
(posterior) that shows an autocorrelation length of about 800.

4.13 Comparison of prior and MH posterior realizations with PC scores and pro-
files along X = 72500m. (a) PC scores of model realizations; (b) PC sores
of data realizations; (c) Gravity profile; (d) Gradient in X; (e) Gradient in Y.

4.14 Comparison of prior and RS posterior realizations with PC scores and pro-
files along X = 72500m. (a) PC scores of model realizations; (b) PC sores
of data realizations; (c) Gravity profile; (d) Gradient in X; (e) Gradient in Y.

4.15 Posterior lithology models. (a-c) Posterior realizations by Metropolis-Hastings
algorithm; (d-f) Posterior realizations by rejection sampling.

4.16 Example of uncertainty reduction shown with probabilities of lithology
“Simple Mafic 1”. (a) Prior probability map of “Simple Mafic 1”; (b) Poste-
rior probability map of “Simple Mafic 1” (Metropolis-Hastings algorithm);
(c) Posterior probability map of “Simple Mafic 1” (rejection sampling).
Chapter 1

Introduction

1.1 Background and Scope

Subsurface resources are essential to maintain our livelihood. Human activities also have impacts on the subsurface through engineering, including resource extraction, waste disposal and underground construction. To aid these engineering activities, the subsurface is the subject of interest for many geological and geophysical investigations (Wellmann and Caumon, 2018). Geological structures (stratigraphy, lithology, fault, etc.) are particularly important because they are actively involved in physical processes of subsurface systems, such as mineralization, formation of reservoirs and water flow in aquifers (Pyrcz and Deutsch, 2014; Scheidt et al., 2018). The evaluation of the spatial distribution of geological structures and petrophysical properties in the subsurface are critical for a wide range of applications, such as reservoir forecasting (Thore et al., 2002), mining planning (Cowan et al., 2003), groundwater assessment (Hassen et al., 2016) and civil engineering (Culshaw, 2005; Hou et al., 2016).

Since the concept of “3D geological modeling”, or simply called “geomodeling” (Raper, 1989; Houlding, 1994; Mallet, 2002; Turner, 2006), proposed about three decades ago as an extension of traditional GIS methods, geologists have been putting efforts in visualizing subsurface geological structures using 3D models. From a computer graphics point of view, these geological structures can be either represented explicitly or implicitly. Explicit
methods rely on “explicit surfaces” to represent geological structures. For an explicit surface, one variable is explicitly expressed in terms of the other variables (e.g., \( z = f(x, y) \)). On the other hand, implicit methods represent geological surfaces using iso-contours of an implicit function (i.e., \( F(x, y, z) = c \)). Methods that rely on implicit functions, such as the signed distance function (Osher and Fedkiw, 2002) and the potential field (Lajaunie et al., 1997; Aug, 2004), have been developed and gained popularity over years due to its convenience in handling complex topologies in 3D geological models (Lajaunie et al., 1997; Cowan et al., 2003; Frank et al., 2007; Calcagno et al., 2008; Caumon, 2010; Yang et al., 2019). With developments in computing power and algorithms, techniques for 3D modeling of geological structures have been applied to more and more complex subsurface challenges (Mallet, 1997, 2014; Caumon et al., 2007, 2012; Cowan et al., 2003; Aydin and Caers, 2017; Grose et al., 2018; Thornton et al., 2018; Manchuk and Deutsch, 2019; Renaudeau, 2019; Renaudeau et al., 2019; Irakarama et al., 2020; Schaaf et al., 2020).

Despite recent advances in 3D modeling methods, due to the sparsity and imprecision of geological data, as well as a lack of geological understanding, uncertainty is an integral property of 3D geological structural models (Caumon, 2010; Mallet, 2014; Wellmann et al., 2010). Geophysical data provide information about rock properties in the subsurface, and also help to constrain subsurface models (Linde et al., 2015). Therefore, geophysical data are usually acquired to be integrated in inversion frameworks and reduce the uncertainty of geological structural models (Linde et al., 2015; Wellmann et al., 2017; Scheidt et al., 2018).

Methods and frameworks for quantifying and visualizing the uncertainty of 3D geological structures that incorporate various sources of information are needed to answer these fundamental research questions regarding 3D geological structures:

- How to integrate data and geological rules to model geological structures?
- What is the uncertainty of geological structures?
- How much uncertainty can be reduced with additional data?
The goal of the thesis is to present a suite of quantitative frameworks and methods to support the integration of geological observations, geophysical data and geological knowledge in modeling geological structures, and to tackle challenges in visualizing and quantifying uncertainty of geological structures. Studied cases in the thesis are mostly from contexts of mineral resources exploration, but methods presented here are for general purposes of addressing challenges regarding 3D geological structures. The thesis focuses on three specific challenges:

1. Challenge in the lack of a unified method for assessing and visualizing uncertainty of an existing 3D structural model constrained by input geological rules and data constraints.

2. Challenge in 3D modeling of large-scale geological structures constrained to dense data and strict geological rules. Reasons include the non-stationary and complex nature of large-scale geological structures, and also difficulty in estimating an adequate spatial continuity model.

3. Challenges in uncertainty quantification of geological structures with geophysical data. Such problems are challenging for three reasons. First of all, the dimension of spatial parameters is very high; no straightforward solution is available. Secondly, it is often time consuming to build multiple 3D geological models that are needed for uncertainty quantification. Last but not least, the forward model for generating geophysical data is usually non-linear and many methods based on linear assumptions or simple linear transformations may not work well.

1.2 Dissertation Outline

This thesis is composed of one introduction chapter, three research chapters, one conclusion chapter and two appendices.

Chapter 1 introduces the general background, challenges, research scope, and summarizes each chapter.

Chapter 2 proposes a method that assesses and visualizes uncertainty of implicit geological surfaces. Geological surfaces are represented as the addition of trend functions
(represented by signed distance functions) and residual functions subject to constraints of data and geological age relationships. Two sampling approaches to create the stochastic motion are used: Monte Carlo and Markov chain Monte Carlo (McMC). The uncertainty is assessed by independent realizations drawn by Monte Carlo sampling. The uncertainty is visualized by a “smooth” movie of gradually evolving geological surfaces that have the same stationary distribution as Monte Carlo realizations, sampled by Markov chain Monte Carlo (McMC). This idea is integrated into the level set equation. This new idea is first illustrated with simple synthetic 3D examples, taking the constraints of data and geological age relationships into consideration. Finally, the idea is illustrated using a synthetic data set from a copper deposit, where dense drillholes constrain an ore body with seven different lithologies. The method provides a direct assessment and visualization of the uncertainty of 3D geological surfaces. This chapter is published in Yang et al. (2019).

Chapter 3 introduces a framework to model large-scale implicit geological structures. This chapter addresses how a single large-scale geological structural model that fits dense data and honor multiple geological rules is built, which can be used as the initialization step in Chapter 2. The implicit potential field method is an interpolation algorithm that is capable of incorporating multiple types of data, including contact points and orientations, to build a complex structural model. However, directly applying the method to 3D modeling of large-scale geological structures constrained to dense data and strict geological rules remains challenging, due to the non-stationary and complex nature of large-scale geological structures, and difficulty in estimating an adequate variogram model. A framework that uses a divide-and-conquer strategy is proposed. The core idea is to create intermediate 3D geological models that match subsets of data and then recombine them into a single large 3D geological model, while maintaining data and geological rule constraints. The chapter proves that linear combinations of potential fields preserve properties of conditioning (i.e., matching data). The chapter presents an application of the framework in modeling the stratigraphy model of a large banded iron formation (BIF) in Western Australia with dense boreholes, but scarce orientation measurements. This chapter was accepted for publication in Natural Resources Research.
Chapter 4 presents a method to quantify the uncertainty of implicit geological structures with geophysical data. It is an extension of the modeling method in Chapter 3 and the uncertainty visualization and assessment method in Chapter 2, by incorporating geophysical data and focuses on uncertainty quantification. Geophysical data provide critical information and constraints for validating subsurface models, so it is common to integrate geophysical data for solving inverse problems. Bayesian frameworks are often used for quantifying uncertainty of 3D geological structures in inverse problems, but challenges still exist, due to the high dimensional nature of spatial models, inefficiency in building multiple geological models, and non-linearity of the forward model. Implicit representation of geological structures transforms discrete geological objects into a continuous variable, i.e., a scalar field; dimension reduction techniques such as principal component analysis can be applied because of the implicit representation. After dimension reduction, the prior distribution is defined in the dimension reduced space. Computing time is saved for sampling new model realizations from the low dimensional space. Therefore, the entire uncertainty quantification task becomes tractable. For sampling posterior realizations, rejection sampling and Metropolis-Hastings sampling algorithms are designed and compared in this chapter. The framework is demonstrated with a mineral-hosting region in Western Australia with gravity data.

Chapter 5 concludes the thesis and discusses future work.

Appendix A shows how the gradual deformation method is a McMC sampler and can be used for reflecting the actual uncertainty sampled by Monte Carlo, thereby used as a way to visualize the uncertainty. It provides the foundation for visualizing uncertainty of geological surfaces using animation, discussed in Chapter 2.

Appendix B reviews the co-kriging system of the implicit potential field method. It provides necessary mathematical details for understanding the implicit potential field method used in Chapter 3 and Chapter 4.

Appendix C proves that the conditioning to contact points and orientations is preserved with linear combination of two potential fields, which is the foundation for modeling large-scale implicit geological structures with linear combination in Chapter 3.
Chapter 2

Assessing and Visualizing Uncertainty of Implicit Surfaces

2.1 Introduction

Building 3D geological models is crucial in subsurface prediction and risk assessment. Model building is widely applied in reservoir forecasting (Thore et al., 2002), mining planning (Cowan et al., 2003), groundwater assessment (Hassen et al., 2016) and civil engineering (Hou et al., 2016). Many geological modeling exercises require building 3D surfaces of geological structures (e.g., ore bodies, reservoirs, faults). Due to the sparsity and imprecision of geological data, as well as a lack of geological understanding, uncertainty is an integral property of 3D surface models (Caumon, 2010; Mallet, 2014; Wellmann et al., 2010). Assessing and visualizing the uncertainty of these 3D surfaces have been acknowledged to be important and challenging tasks (Caers, 2011; Caumon, 2010; Lindsay et al., 2013; Scheidt et al., 2018).

Traditionally, uncertainty models are constructed with explicit methods, but the topology is fixed for all realizations (Lecour et al., 2001; Thore et al., 2002). This limitation is due to the fact that explicit methods require tedious manual effort to define the topology, such as the connectivity of a single surface and the spatial relationship among multiple
surfaces. Thus, with explicit methods, it is almost impossible to edit the topology for every single realization. Implicit methods, which use level sets to represent surfaces, were applied to model geological structures as early as Lajaunie et al. (1997) and have since gained considerable attention in geological modeling because little manual intervention is required in the modeling process (Calcagno et al., 2008; Caumon et al., 2012; Cowan et al., 2003; Gonçalves et al., 2017; Frank et al., 2007; Vollgger et al., 2015). Although implicit methods usually focus on interpolating a single deterministic model, they have the advantage of handling topological perturbations naturally, rendering future work about the joint quantification of geometric and topological uncertainty a feasible and automatic task (Aydin and Caers, 2017; Cherpeau and Caumon, 2015).

In general, two avenues have been explored for uncertainty assessment on 3D surfaces. One is base-case reasoning, which involves either estimating uncertainty based on the kriging variance (Chilès et al., 2004; Tacher et al., 2006), or perturbing a reference (base) model to generate multiple realizations (Caumon et al., 2007; Lecour et al., 2001; Mallet, 2014; Røe et al., 2014; Thore et al., 2002). Some base-case reasoning approaches specify an uncertainty boundary and perturb the base model within that boundary (Caumon et al., 2007; Lecour et al., 2001; Røe et al., 2014; Thore et al., 2002). Others consider the “coherency” among realizations while perturbing the base model (Mallet, 2014). The base-case reasoning approach has potential drawbacks; for instance, the base model may be biased, and the uncertainty of key prediction variables may be underestimated. However, for some applications where dense data sets (e.g., numerous drillholes) are available, such as mineral resources forecasting, the drawback may be alleviated, and it has computational advantages compared to the other avenue, fully stochastic. A fully stochastic approach, on the other hand, focuses on sampling multiple realizations, sometimes within a Bayesian setting (Aydin and Caers, 2017; Cherpeau et al., 2010, 2012; Cherpeau and Caumon, 2015; Holden et al., 2003; de la Varga and Wellmann, 2016; Wellmann et al., 2017; Grose et al., 2018). In order to address uncertainty caused by the inaccuracy of input data sets, some researchers propose sampling input data sets from user-defined distributions and using the implicit potential field method to directly simulate multiple realizations (Wellmann et al.,
This approach potentially ignores other sources of uncertainty, especially uncertainty between data locations (Aitken et al., 2013). In some fully stochastic approaches, usually geological structures (e.g., faults and horizons) are stochastically modeled in a consistent way enforced by specified rules (Holden et al., 2003; Cherpeau and Caumon, 2015). Recently, in order to account for prior information, uncertainty quantification of 3D geological structures using a Bayesian framework has emerged (Aydin and Caers, 2017; Cherpeau et al., 2012; de la Varga and Wellmann, 2016; Wellmann et al., 2017; Grose et al., 2018). The advantage of using a Bayesian framework is that the uncertainty can be reduced when new data or information becomes available, which is a logical way of reasoning (Scheidt et al., 2018). However, constraining the models with dense hard data within the Bayesian framework is still challenging.

In addition to the use of level sets in implicit modeling and uncertainty quantification in geosciences (Aydin and Caers, 2017; Cherpeau and Caumon, 2015; Mallet, 2014; Moreno and Aanonsen, 2011; Xie et al., 2011), sophisticated techniques have also been developed for generating and evolving implicit interfaces (curves and surfaces) in the fields of computer graphics and computer vision (Osher and Fedkiw, 2002). In particular, dynamic level set methods, first proposed in Osher and Sethian (1988), are a relatively simple yet powerful way to implicitly represent and dynamically evolve complex interfaces. Algorithms have been developed to reconstruct implicit interfaces based on optimization (Juan et al., 2006; Zhao et al., 2001), or to sample posterior interfaces (mostly curves) using Markov chain Monte Carlo (McMC) (Fan et al., 2007; Iglesias et al., 2016). However, few if any applications for assessing the uncertainty of complex surfaces with dynamic level set methods have been reported in the literature.

In this chapter, a new idea to incorporate stochastic motion with the level set methodology for assessing and visualizing the uncertainty of geological surfaces is proposed. Considering sampling the realizations of geological surfaces in a time sequence, a surface that is unknown, except where the hard data is located, can still “move” (assuming the measurement error in the data is negligible). In that sense, stochastic motion of the surfaces is
created, resulting in a “movie” of the surfaces. This idea is integrated with level set methods. With the level set methods, independent complex geological surfaces can be directly sampled by Monte Carlo and used for uncertainty assessment. Dependent realizations of geological surfaces are sampled by McMC and used to create a smooth movie to visualize the uncertainty, thereby potentially increasing people’s comprehension of where and how large the spatial uncertainty is (Ehlschlaeger et al., 1997; Hyde et al., 2018).

The chapter is structured as follows. First, the methodology is presented in Section 2.2. Then, in Section 2.3, several 3D examples are provided to illustrate the method with different parameters. Next, an application of multi-lithology simulation constrained by dense drill-hole data and a clear geological age relationship is shown in Section 2.4, along with the associated uncertainty analysis. Finally, Section 2.5 discusses the CPU runtime, and Section 2.6 concludes the chapter. Produced movies for all examples in Section 2.3 and Section 2.4 can be found at: https://github.com/SCRFpublic/LevelSetMovies.

2.2 Methodology

Our methodology for assessing and visualizing uncertainty involves applying stochastic velocity fields to level set representations of geological surfaces, and constraining this motion to data and geological rules. Therefore the level set method is reviewed first and then how this can be used to represent and initialize surfaces is shown. Finally, the notion of stochastic motion is explained.

2.2.1 Level set method

With the level set method, an $n$-dimensional interface $\Gamma$ (e.g., a curve in 2D, or a surface in 3D) is represented as the isocontour $\phi(\vec{x}) = c$ of an $(n + 1)$-dimensional function $\phi(\vec{x})$ called a level set function. This implicit approach fundamentally differs from explicit surface modeling (see Figure 2.1). Varying the chosen isocontour allows exploration of a family of geometries related by the geometry of the underlying level set. In particular, if the underlying level set is constrained to be a signed distance function (SDF), i.e., satisfying
Figure 2.1: Level sets represent surfaces implicitly as isocontours of higher-dimensional functions, in contrast to explicit surface methods. Here we compare the representations for the 2D circle $x^2 + y^2 = 4$. (Left) The zero contour of the level set paraboloid $z = x^2 + y^2 - 4$ is the desired circle (several contours in the $z = 0$ plane are drawn, with different colors indicating regions between contours). (Middle) An explicit (also called Lagrangian) representation of the circle is a collection of points (red) and edges (thick black). (Right) Computationally, discrete samples (red) of the level set are stored in a uniform grid data structure; interpolation of these values is used to identify a desired contour.

In this case, for an implicitly represented interface $\Gamma$, points satisfying $\phi(\vec{x}) < 0$ are inside $\Gamma$, points satisfying $\phi(\vec{x}) > 0$ are outside $\Gamma$, and points satisfying $\phi(\vec{x}) = 0$ lie exactly on $\Gamma$; moreover, for any point, $\phi(\vec{x})$ gives the signed distance to the nearest point on the surface, useful for many algorithms. Interfaces represented with level sets naturally support topological changes, such as merging and splitting, that are often challenging operations with explicit representations (see Figure 2.2).

Level set methods often dynamically evolve level sets by parameterizing them in time, $\phi(\vec{x}, t)$, and solving the “level set equation,”

$$\frac{\partial \phi(\vec{x}, t)}{\partial t} + \vec{v}(\vec{x}, t) \cdot \nabla \phi(\vec{x}, t) = 0.$$

(2.1)

The level set equation advects $\phi(\vec{x}, t)$ according to an $n$-dimensional velocity field $\vec{v}(\vec{x}, t)$, which in 3D is parameterized as $(v_x, v_y, v_z)$. Since only the component of velocity in the
normal direction has an effect on the level set equation, Equation 2.1 is equivalent to:

$$\frac{\partial \phi (\vec{x}, t)}{\partial t} + \vec{v}_n (\vec{x}, t) \cdot |\nabla \phi (\vec{x}, t)| = 0,$$

(2.2)

where $\vec{v}_n (\vec{x}, t)$ is the normal velocity. Temporally discretizing Equation 2.2 with the forward Euler method yields

$$\frac{\phi (\vec{x}, t + \Delta t) - \phi (\vec{x}, t)}{\Delta t} + \vec{v}_n (\vec{x}, t) \cdot |\nabla \phi (\vec{x}, t)| = 0.$$

(2.3)

Finally, note that using the level set method allows certain geometric quantities to be easily computed; for example, the normal $\vec{N}$ to and mean curvature $\kappa$ of a surface are given by

$$\vec{N} = \frac{\nabla \phi}{|\nabla \phi|}, \quad \kappa = \nabla \cdot \vec{N}.$$

(2.4)
2.2.2 Definition of the stochastic model

Let $T(\mathbf{x}, t)$ be a signed distance function that represents the mean function of an implicit geological surface, and $R(\mathbf{x}, t)$ denote a residual function that represents correlated variation made on that mean function, at time $t$. The mean function may be interpreted as an implicit conceptual model. A geological surface’s corresponding level set function at time $t$, $\phi(\mathbf{x}, t)$, is defined as the sum of $T(\mathbf{x}, t)$ and $R(\mathbf{x}, t)$, i.e.,

$$
\phi(\mathbf{x}, t) = T(\mathbf{x}, t) + R(\mathbf{x}, t).
$$

Equation 2.5 has similarities to the usual geostatistical assumption in random function theory of a mean plus a residual component (e.g. universal kriging) (Chilès and Delfiner, 2012; Kyriakidis and Journel, 1999). This formulation has also been applied in the uncertainty assessment of geological surfaces (Cherpeau et al., 2010; Cherpeau and Caumon, 2015). However, here the use of a random mean function is suggested, and not an unknown deterministic mean function. The stochastic mean is useful, for example, if some initial 3D surface model $\phi_0(\mathbf{x})$ has been built that is too smooth to have a geological meaning. Note that this approach has similarities with a Bayesian geostatistical approach (Omre, 1999).

To model a random mean function, the following definition is used:

$$
T(\mathbf{x}, t) = \phi_0(\mathbf{x}) + R_T(\mathbf{x}, t),
$$

where $R_T(\mathbf{x}, t)$ is another residual implicit random function, possibly with less variation than $R(\mathbf{x}, t)$. One may of course opt for the deterministic model, i.e., $T(\mathbf{x}, t) = \phi_0(\mathbf{x})$.

2.2.3 Level set implementation of the stochastic model

To implement the representation of Equation 2.5 in the level set formalism, $T(\mathbf{x}) = \phi(\mathbf{x}, t - \Delta t)$ and $R(\mathbf{x}, t) = |\mathbf{v}_n(\mathbf{x}, t - \Delta t)| \cdot \Delta t$ are set in Equation 2.5, which gives
the following level set equation:

$$\phi(x, t) = \phi(x, t - \Delta t) + |\bar{v}_n(x, t - \Delta t)| \cdot \Delta t,$$

(2.7)

where $\Delta t$ is always set to 1. To obtain $|\bar{v}_n(x, t - \Delta t)|$, at first, a Gaussian realization is defined on the surface $T(x, t) = 0$. This is achieved by creating a Gaussian realization for the entire grid and extracting the values on the mean surface, which themselves are also Gaussian. Then the values on the mean surface are extrapolated in the normal direction (Figure 2.3), called “velocity extension” (Adalsteinsson and Sethian, 1999), meaning velocities that satisfy:

$$\nabla |\bar{v}_n(x, t)| \cdot \nabla \phi(x, t) = 0.$$

(2.8)

are created. Velocity extension can be implemented by:

$$\bar{v}_n(x, t) = \bar{v}_n\left(x - \phi(x, t) \cdot \bar{N}, t\right).$$

(2.9)

$R_T(x, t)$ in Equation 2.6 can also be created by this velocity extension implementation. In contrast to directly adding a Gaussian random field to an initial level set function, such as in Cherpeau et al. (2010) and Cherpeau and Caumon (2015), velocity extension perturbs all isocontours consistently in order to avoid creating unexpected isolated objects distant from the base surface $T(x, t) = 0$ (see Figure 2.4 for an example). This does not mean that topological change is not possible. Velocity extension is useful since it ensures that $\phi(x, t)$ remains a SDF during evolution (Adalsteinsson and Sethian, 1999). The SDF properties are hence preserved, without ad-hoc modifications. Additionally, the velocity extension will be useful to create a random walk of the surface that has microscopic reversibility (Fan et al., 2007). The open-source library PhysBAM is used for representing and manipulating level sets.

---

Figure 2.3: Illustration of velocity extension. (a) shows a sphere (black circle) defined by its signed distance function $\phi(\vec{x})$. (b) shows a Gaussian realization for the entire grid. (c) shows the extended velocity field. Values (velocities) on the sphere are extracted and extended in the normal direction. For instance, the velocities at locations $\vec{x}_1$, $\vec{x}_2$ and $\vec{x}_3$ are equal to the velocities at $\vec{x}'_1 = \vec{x}_1 - \phi(x_1)\vec{N}$, $\vec{x}'_2 = \vec{x}_2 - \phi(x_2)\vec{N}$ and $\vec{x}'_3 = \vec{x}_3 - \phi(x_3)\vec{N}$. 
Figure 2.4: Comparison of the results of using velocity extension (middle) and directly adding a Gaussian random field (right) to the implicit function of a base surface (left). The same Gaussian realization is used for both methods in the test.

2.2.4 Modeling the mean function

To model the mean function, initialization of the signed distance function $\phi_0(\bar{x})$ in Equation 2.6 is needed. In some applications, analytic functions with uncertain parameters are enough to initialize simple shapes such as planes, ellipsoids and channels. In complex applications, more sophisticated approaches are required to initialize an implicit surface model. Data constraints are often placed on the generation of surface models. An efficient way to initialize the signed distance function is the surface reconstruction method proposed by Zhao et al. (2001). Zhao et al. (2001) reconstructs an interface $\Gamma$ from data points by first initializing $\Gamma$ to be a rough initial guess of the final surface (for instance, a bounding box such as the grid itself that encloses all the data points. The fast tagging method described in (Zhao et al., 2001) is then applied, which iteratively excludes from $\Gamma$ points that are outside a radius of any data points. The fast tagging method runs in $O(N \log N)$, where $N$ is the number of grid cells bounded by the initial guess for $\Gamma$. Finally, to obtain a smoother representation of $\Gamma$, the convection equation

$$\frac{d\Gamma}{dt} = -\nabla d(\bar{x})$$  \hspace{1cm} (2.10)
Figure 2.5: Illustrates the surface reconstruction process of (Zhao et al., 2001), which is used for initial model generation. (Left) 1,000 points from the surface of a 3D sphere of radius 10 centered at the center of the 100³ domain, a uniform Cartesian grid with resolution $\Delta x = 0.5$ are randomly sampled. The initial guess for a surface that fits these points is a box that fills the entire domain (i.e., the grid itself). (Middle) The fast tagging method efficiently refines the initial guess to a coarse approximation of a surface. (Right) After 30 steps of convection, the reconstructed surface appears quite smooth. The surface is not perfectly spherical due to the sparsity of the input points and the limited grid resolution (there is inherent $O(\Delta x)$ error in the final result).

is integrated for several (fictitious) time steps, where $d(\vec{x})$ is the distance from a point $\vec{x}$ to the nearest data point. Thus, $\Gamma$ is attracted towards the data set, creating a “shrink-wrapped” level set surface that matches the data. The initialization procedure is illustrated in Figure 2.5. After initialization, the conceptual models are constructed by Equation 2.6 and have larger variability than simply using $\phi_0$. If orientation information is also available, the potential field method (Section 3.3.1) may be preferable to be used to initialize a mean model, represented by potential fields, and potential fields can be converted into signed distance functions with a post-processing step, so that the proposed framework can be applied.

2.2.5 Multiple level sets

For representing multi-category geological surfaces, each category uses an independent level set function. Often, a geological age relationship constraint is necessary when a prior age relationship is known for multiple objects, such as stratigraphies, faults, and lithologies (Cherpeau et al. (2010). The geological age relationship can be expressed by the difference...
CHAPTER 2. ASSESSING AND VISUALIZING UNCERTAINTY

Oldest lithology
Second oldest, included in oldest
Third oldest, inside the second oldest but touching the boundary
Youngest, overprints all

\( f_1 = 10 \)
\( f_2 = 20 \)
\( f_3 = 30 \)
\( f_4 = 40 \)

if \( f_1 \) is intersected at \( f_2 = 0 \) then attachment is naturally handled

Figure 2.6: A geological age relationship example. (a) Four level sets are used to represent four lithologies, which have an age relationship indicated in the figure; (b) Independent initialization of each level set without truncation.

operation of level sets (Museth et al., 2003): consider a binary case with categories A and B, if category A is older than category B, then \( \phi(A) = \min(\phi(A), -\phi(B)) \), and \( \phi(B) = \phi(B) \); otherwise, \( \phi(A) = \phi(A) \), and \( \phi(B) = \min(\phi(B), -\phi(A)) \). In order to enforce the geological age relationship, multiple level sets are independently initialized for each object at first (see Figure 2.6). The initial model should reflect the initial geological states, meaning that the older level sets are not truncated by newer level sets yet. Then, the difference operation is applied to the level sets according to the geological age relationship.
2.2.6 Stochastic motion

For assessing and visualizing the uncertainty, stochastic motion is defined for geological surfaces, in the sense that both the mean function and the residual function change with time. Two sampling approaches, Monte Carlo and McMC, are proposed to draw independent and dependent realizations, respectively.

In the Monte Carlo sampling, a new realization of the mean function and a new realization of the residual function are drawn at every new time $t$ independently. Hence the time $t$ is simply a counter for realizations. For conditional simulation, the values of conditioning data points are set to zero and then conditional Gaussian realizations are drawn, so that the geological surfaces do not move away from the conditioning data. McMC sampling of the uncertainty of geological surfaces is also defined. McMC can be useful when needing to condition to more complex data that require the definition of a likelihood function. In this chapter, however, the research focuses on using McMC for drawing from the same distribution as with Monte Carlo, meaning drawing from a prior distribution. The gradual deformation method (GDM) (Hu, 2000), is used as an approximate McMC method for sampling surfaces. Note that GDM is not used as an optimizer in the classical sense as in Hu (2000). Instead, this approach uses GDM to generate a random walk in the prior space. For this random walk to converge to a stationary distribution, the Markov chain generated by GDM needs to have detailed balance (microscopic reversibility) (Green, 1995; Medhi, 1994; Mosegaard and Sambridge, 2002). In other words, the probabilities of going from $\phi(\vec{x}, t)$ to $\phi(\vec{x}, t + \Delta t)$ needs to be the same as the probability of going from $\phi(\vec{x}, t + \Delta t)$ to $\phi(\vec{x}, t)$. The time $t$ is now equivalent to the iteration in the chain. A proof that shows GDM provides a random walk in the prior space is given in the Appendix A. Additionally, defining the gradual deformation on the velocity extension, allows for the SDF properties to be preserved during the random walk. McMC can then be used to visualize the uncertainty as a continuously evolving “movie” of the geological surfaces it generates. Because of the properties shown in Appendix A, this movie has equivalent uncertainty as Monte Carlo sampling. Figure 2.7 shows the gradual change of a residual function $R(\vec{x}, t)$. 
CHAPTER 2. ASSESSING AND VISUALIZING UNCERTAINTY

\[ j = 1, k = \pi/5 \]
\[ j = 2, k = 2\pi/5 \]
\[ j = 3, k = 3\pi/5 \]
\[ j = 4, k = 4\pi/5 \]
\[ j = 5, k = \pi \]

Figure 2.7: Illustration of the gradual deformation of the residual function. The residual function is a normal extension of the Gaussian realization defined on the sphere in Figure 2.3.

2.3 Illustration cases

This section illustrates the method, studies the influence of parameters, demonstrates uncertainty assessment, and analyzes important output statistics, using a synthetic 3D setting with a grid size of $100 \times 100 \times 100$. In order to show how the method works for diverse surface types, the initial surfaces are three objects generated by three analytic functions: a plane, a sphere and a cosine-type wave (see Figure 2.8). The plane and the wave are open surfaces and the sphere is a closed body. To investigate the uncertainty affected by conditioning data, three drill-holes are used as conditioning data, and they pass through the plane and wave. A synthetic age relationship is assigned to three objects, so that the plane is truncated by the other two objects and the wave truncates the other two. The Gaussian realizations for defining the mean function and the residual function are created using sequential Gaussian simulation (SGS) with mean zero (Chilès and Delfiner, 2012; Remy et al., 2009). A Gaussian variogram model is used in all of our tests for creating smooth surfaces that may be more geologically plausible than using other variogram models. For comparing the influence of parameters, three series of conditional Gaussian realizations are generated, each series containing 100 realizations with sill 5 and range 20, sill 5 and range 60, and sill 10 and range 20, respectively (Figure 2.8). These parameters have the same unit as the grid size. One series of unconditional Gaussian realizations with sill 5 and range 20 is also generated as a comparison. McMC by GDM is used to create a smooth movie and each GDM cycle contains 10 deformation steps.
Figure 2.8: Initial surfaces and velocity fields. (a) The initial three objects and three drill-holes are visualized together. Realizations of velocities are generated using three boreholes as conditioning data. Three realizations generated with (b) sill 5 and range 20, (c) sill 10 and range 20, and (d) sill 5 and range 60 are shown. The colors in the realizations represent the velocity magnitudes.
2.3.1 Simulation results and uncertainty assessment

The simulation results are compared and the uncertainty is assessed in this section. Two realizations from the movie are selected and shown in the first and the second columns of Figure 2.9. Notice how the sill and range contribute to the variability of the surfaces. The surfaces separate the space into three domains (see the third column of Figure 2.9 for the sections of the domains with the truncation rule enforced). One quantitative way of assessing the uncertainty of 3D surfaces is using the information entropy (Shannon, 1948) and it has been applied in geoscience as a quality measurement of geological models (Wellmann and Regenauer-Lieb, 2012). As in Wellmann and Regenauer-Lieb (2012), every surface separates the grid into two sub-domains so the uncertainty of the surfaces is simply represented by the entropy of every grid cell. Only the independent geological surfaces sampled by Monte Carlo, (i.e., the first realization of every GDM cycle) are retained to calculate the entropy. The last column of Figure 2.9 shows entropy maps of the stochastic motion. In the entropy maps, the blue color indicates zero entropy (i.e., certainty), and other colors form a “color band” that mirrors possible locations of surfaces. The width of the “color band” reflects the variance of the Gaussian realization (i.e., the sill of the variogram). The larger the variance is, the wider the “color band” is. Conditioning is also shown by the entropy maps: the entropy values are zero on the drillholes and small near drillholes. The entropy map of the unconditional stochastic motion in the last row is shown for comparison. Most uncertainty lies in areas where multiple surfaces have a high possibility intersecting each other (see the orange and red regions in Figure 2.9).

2.3.2 Output statistics: volume and surface roughness

This section shows how the parameters affect some important output statistics of a geological structure model, such as the volume and surface roughness (Mallet, 2014; Thore et al., 2002). Volumes of the three domains separated by surfaces are calculated (see the third column in Figure 2.9 for the sections of the domains). Figure 2.10 shows the change of volume proportion during the conditional stochastic motion with three parameter settings.
Figure 2.9: Snapshots of surfaces at time 60 (first column) and at time 90 (second column), sections of geological domains (third column) and the entropy maps (fourth column). The first three rows are the results of the conditional stochastic motion with sill 5 and range 20, sill 10 and range 20, and sill 5 and range 60, from top to bottom. The last row shows the results of the unconditional stochastic motion with sill 5 and range 20.
It is shown that the initial volumes are maintained for all cases. The effect of maintaining the volume is in fact due to that the stationary isotropic Gaussian realizations are used, which are divergence-free (Rogers and Williams, 2000). A divergence-free velocity field is the condition for the conservation of mass (or volume if the density is constant) in continuum mechanics (Osher and Fedkiw, 2002). However, it should be noted that if the size of the mean surface is so small that the Gaussian realization is not stationary on the mean surface, the volume may no longer be maintained. Figure 2.9 suggests the surfaces are visually “roughe” with larger sills and smaller ranges. This observation coincides with the consequence of geostatistical simulation where larger sills and smaller ranges result in more variability in realizations. In order to quantitatively compare the surface roughness, the variance of the mean curvature: \( \text{Variance} (\kappa) \) is used as the global indicator of surface roughness. \( \text{Variance} (\kappa) \) of the three parameter settings in Section 2.3.1 are compared in Figure 2.11. The plots suggest that the surface roughness is related to the roughness of the mean surface, the sill and the range. Although some variations exist, there is a general mean of roughness decrease when increasing the range, or decreasing the sill.

### 2.4 Case study: synthetic copper deposit

#### 2.4.1 Data description

To evaluate the method’s performance in a more realistic setting where surfaces’ geometry and topology are complex, and constrained by dense hard data and a clear geological age relationship, the methodology is applied to a synthetic data set motivated by a real copper deposit. In this case, approximately 50 drillholes are used as hard constraints, indicating seven lithologies (Figure 2.12). An interpreted model was built, constrained by these drillholes and geologists’ understanding of the geological age relationship in this case (Figure 2.12) (this model was given as is, hence one can consider it as a smooth deterministic interpretation of the actual variability). The lithologies are named by their age hierarchy, so \( \text{Lithology 1} \supseteq \text{Lithology 2} \supseteq \text{Lithology 3} \supseteq \ldots \supseteq \text{Lithology 7} \). The older lithologies are
CHAPTER 2. ASSESSING AND VISUALIZING UNCERTAINTY

Figure 2.10: The change of volume proportion with three parameter settings. The horizontal axis shows time steps and the vertical axis shows the volume proportion of the domains separated by three surfaces.

Figure 2.11: The change of variance of mean curvature (vertical axis) w.r.t. time step (horizontal axis) with three parameter settings.
truncated by younger lithologies. Although the geological age relationship is known and fixed, the geometry and topology of each lithology are uncertain. Drillholes are dense in the center and at the top, but are sparse at the bottom and near the boundary of the grid.

2.4.2 Simulation

For sampling surfaces and maintaining the assumed geological age relationships, an interpreted model is used to initialize the level sets with the fast tagging method and convection equation (Equation 2.10) and get $\phi_0(\vec{x})$. To simulate, trends and residuals, i.e., $\phi_0 + R_T$ and $R$, are sampled by both Monte Carlo and McMC to generate geological surfaces. In order to maintain the age relationship during stochastic motion, the same truncation rule is used at all time steps. The simulation grid size is $120 \times 120 \times 100$. Since no additional information on surface roughness was made available to us for this virtual case, an arbitrary variogram was used for $R_T$ and $R$ with sill 10 and range 10 (same unit as the grid size). The output movies show a gradual change in the geometry and topology (see accompanied movies in https://github.com/SCRFpublic/LevelSetMovies). Figure 2.13 (a) and (b) show two copper deposit realizations drawn with Monte Carlo. One can observe their differences in the geometry and topology. The sections (Figure 2.13 (c) and (d)) indicate that the geological age relationship is maintained during the stochastic motion. Figure 2.13 (e) and (f) show the change of surface connectivity between two different realizations. It is clear that the two realizations have different connectivity compared to Figure 2.12(c). Because of the higher density of drill-holes, the surface variability is smaller in the center area and at the top, comparing to the bottom area. The probability that any spatial location belongs to Lithology 4 is shown in Figure 2.14 (a). It is shown that the probabilities are large near data points that indicate Lithology 4, and are small away from those points. The entropy of all surfaces is shown in Figure 2.14 (b). Notice that larger entropies are seen at the bottom, because of the sparsity of drillholes.
Figure 2.12: Synthetic copper data set. (a) shows 50 drillholes. (b) shows an interpreted model of the area. The interpreted model shows a complex shape. (c) a section view of the interpreted model. (d) shows the geological age relationship (old to young from the top to the bottom). Seven colors in (d) correspond to seven lithologies in the data.
Figure 2.13: Realizations of the synthetic copper case. (a) and (b) are two realizations from the stochastic motion (Lithology 1 is made transparent). (c) and (d) are sections of the models that show the geological age relationship. (e) and (f) are two realizations of Lithology 7. Note that all figures use the same color scheme as Figure 2.12 (d).
CHAPTER 2. ASSESSING AND VISUALIZING UNCERTAINTY

Figure 2.14: Probability and entropy maps. (a) is the probability that any location belongs to Lithology 4 (probabilities below 0.4 are made transparent). (b) is the entropy of all surfaces (entropies below 1.2 are made transparent).

2.5 CPU runtime

In terms of the method’s runtime, the runtime is discussed per time step of the stochastic motion, referred to as the “unit runtime,” and the runtime for a movie with 100 time steps, plus the initialization time, referred to as the “total runtime.” The number of level sets and the grid resolution are dominating factors for the unit runtime. Most of the computation time is spent on creating the Gaussian realizations and extending the velocity in the normal direction. Surfaces are initialized and all Gaussian realizations are created at the beginning. After that, the unit runtime is linear in the number of level sets since each level set is updated independently. Table 2.1 lists both the unit runtime and total runtime for the two cases in Section 3 and 4. All simulations are run on a laptop with a Intel Core i7-7700HQ CPU with a frequency of 2.80GHz. It is noted that on a single CPU core, the present method’s efficiency is promising; moreover, much of our algorithm could be easily parallelized in order to improve performance since each level set is treated independently.
Table 2.1: Runtimes for the 3D object example and the copper case

<table>
<thead>
<tr>
<th>Cases</th>
<th>Unit runtime</th>
<th>Total runtime (initialization + 100 steps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illustration case</td>
<td>3.9s</td>
<td>407.3s</td>
</tr>
<tr>
<td>Synthetic copper case</td>
<td>19.4s</td>
<td>1994.5s</td>
</tr>
</tbody>
</table>

2.6 Discussions and conclusions

In this chapter, a new methodology for assessing and visualizing the uncertainty of 3D geological surface models using level sets by the means of stochastic motion is proposed. The method takes advantage of the level set methods and directly simulates geological surfaces with complex geometries and topologies, subject to data and geological age relationship constraints. Monte Carlo sampling of independent realizations are retained for assessing the uncertainty. Markov chain Monte Carlo sampling is used for generating a continuously evolving movie that may enhance people’s perception of the uncertainty, as suggested by the growing body of work on building improved visualization systems for the geosciences as well as a recent survey showing the current difficulty of communicating uncertainty for mineral resource geologists (see Hyde et al. (2018) and the references therein). The uncertainty and output statistics such as the volume and surface roughness are related to the input parameters. The CPU runtime analysis shows the method is efficient enough for the presented cases.

What is not yet treated in this chapter is the specification of model parameters for both the mean and residual functions. Future work may include developing a Bayesian inference framework that infers those parameters based on data. With the same goal as other methods for stochastically simulating geological structures mentioned in the introduction, the method will complement geostatistical methods by creating the domain wherein the more classical approaches can proceed; often the geometry and topology of these domains are the zero-th order uncertainty in practical applications (Jones et al., 2013; Mallet, 2014; Thore et al., 2002).
Chapter 3

Implicit Modeling of Large-scale Geological Structures

3.1 Introduction

Geological three-dimensional (3D) modeling is important for visualizing and understanding the spatial distribution of subsurface geological structures. With developments in computing power and algorithms, techniques for 3D modeling of geological structures have been applied to more and more complex subsurface applications (Mallet, 1997, 2014; Caumon et al., 2007, 2012; Cowan et al., 2003; Aydin and Caers, 2017; Grose et al., 2018; Thornton et al., 2018; Manchuk and Deutsch, 2019; Renaudeau, 2019; Renaudeau et al., 2019; Irakarama et al., 2020; Schaaf et al., 2020). Despite possible lack of flexibility in model refining and expert knowledge incorporation (Caumon et al., 2004; Laurent, 2016; Renaudeau, 2019), methods that rely on implicit functions have been developed and gained popularity over years due to its convenience in handling complex topologies (Lajaunie et al., 1997; Cowan et al., 2003; Frank et al., 2007; Calcagno et al., 2008; Caumon, 2010; Yang et al., 2019). Implicit potential field method is one of them that is capable of incorporating multiple types of data including contact points and structural orientations by solving a co-kriging system (Lajaunie et al., 1997; Chilès et al., 2004; Calcagno et al., 2008;
Gonçalves et al., 2017). Although structural orientations are necessary inputs, which may not exist for some applications, the implicit potential field method has the strength of easily handling cases with both conformable and unconformable stratigraphies. The method has been successfully applied to various cases, from simple to complex ones (Aug., 2004; Guillen et al., 2008; Wellmann et al., 2017; Thornton et al., 2018).

However, challenges still exist when directly applying the implicit potential field method in real and complex cases. In applications such as mining, hydrology and urban geological survey, numerous boreholes (or wells), as many as several thousands, are often drilled (Vollgger et al., 2015; Jarna et al., 2015; Thornton et al., 2018; Zhong and Wang, 2020). Borehole data provide accurate locations of contact points between geological formations, as well as orientation measurements. Geological formations consist of conformable and unconformable sets coupled with known erosion and faulting processes, providing detailed rules on what final geologically consistent models should look like. Constraining large sequences to these thousands of boreholes with obeying specific geological rules at the same time poses a modeling challenge (Wellmann et al., 2014; Thornton et al., 2018). With more data, (co-)kriging is expected to be more accurate for prediction to some extent if the non-stationarity and spatial continuity are properly identified. However, an adequate variogram is difficult to be estimated for the potential field method (Chilès et al., 2004), especially for a large domain (Thornton et al., 2018), thus increasing the density and amount of data makes a (co-)kriging system more complex and potentially increases the likelihood of interpolation artifacts (Chilès and Delfiner, 2012; Thornton et al., 2018). As a result, geologically unrealistic models may be created due to fitting all input data at once.

Research on (co-)kriging with dense (and sometimes big) data has been conducted in the field of machine learning. Methods to make kriging more robust, accurate and more computationally tractable for prediction with dense or big data include sequential incorporation of old and new data (Vargas-Guzmán and Yeh, 1999; Kleijnen and van Beers, 2020), and simultaneous combination of multiple interpolation results using subsets of data (van Stein et al., 2015, 2020). The idea of aggregating multiple prediction results to improve prediction can be found in the famous bagging method (Breiman, 1996).
This chapter proposes a new framework for modeling large-scale geological structures with thousands of borehole data, using a divide-and-conquer strategy. Multiple subsets of contact points are selected and multiple sets of stochastic orientations are simulated to create many models each matching the subset of data. These models, in the form of implicit functions, are linearly combined to match all data. The method relies on a proof that any linear combination of potential fields is again a potential field that preserves conditioning to contact points and orientation data. To ascertain that the variability in models built with subset data is reasonable, a validation test with data and imposed geological rules is performed. This chapter will use a large banded iron formation (BIF) with 19 stratigraphical layers in Western Australia (Deposit “99”), with a spatial extent of about six kilometers in EW, three kilometers in NS, and 500 meters in depth, explored by 3100 boreholes, as an illustration case.

The chapter’s structure is as follows. First, a general geological background regarding BIF and the deposit in focus is introduced in Section 3.2. This sets the stage on the size of the problem and the challenge ahead. Then in section 3.3, the methodology is presented with simple illustrations. Finally, section 3.4 shows a full application of the method in modeling a large BIF and section 3.5 concludes the chapter.

3.2 Case study

3.2.1 General background

Iron is the world’s most used metal which is extracted from a variety of iron ores, such as magnetite, hematite or siderite ores. Around 1.6 Gt/y of iron ore were exported worldwide, with Australia being the most important country, shipping about 50%. Iron ore is found in magmatic settings mainly as hematite or in sedimentary settings as BIF or in detrital iron ore deposits. One of the largest iron ore districts is located in the Yilgarn and Pilbara Region in Western and North Western Australia.

Due to an incomplete understanding of the multiple processes forming the BIF’s, the
prediction of economical exploitable resources is difficult. However, the industry relies on very accurate predictions of the composition and spatial distribution of the ores and contaminants such as alumina, respectively. Thus, boreholes are comparatively dense already in the exploration phase of a project, reducing the uncertainties. Without other measurement, such as chemical characterization of drill cores, the stratigraphy is a first order proxy for economic accumulations of BIFs. Thus, structural models are required to match the observations exactly while obeying geologic rules. Building 3D geological models for the BIF is critical for estimating grades for the deposit due to their strong correlations in space (Lascelles, 2012).

3.2.2 Regional geology

The Pilbara Region consists of the Archean Pilbara Craton, the Precambrian Fortescue and Hamersley Basins. The Hamersley Basin is an about 100,000 km$^2$ large intra-cratic depression filled by 2.5 km of stratified volcano-sedimentary units of the Hamersley Group (2630 - 2445 Ma), including three major episodes of around 1000 m, laterally extensive BIFs. The Hamersley Group was probably formed as a sediment-starved basin deposited on a subsiding passive continental margin consisting of clastic sediments, dolostones, rhyolites and dolerites. Most formations within the Hamersley group exhibit partially interstratification of BIF’s, but only Marra Mamba and Brockman Iron Formations are economical with thicknesses of 230 m and 620 m (Trendall, 1983). The latter is further divided into the Dales Gorge and the Joffre Member by about 75 m of Whaleback Shale. BIF consists of fine grained, greyish interbeds of iron-poor layers like silicates, quartz (chert) or carbonates and iron rich minerals with little or no terrigenous sediment (Trendall, 1983) and low alumina content (Krapež et al., 2003; Lascelles, 2007). The bulk chemistry consists of more than 15% iron. The sediments are rhythmic laminated beds which are observed on a mm (micro-bands), cm (meso) and m-thick scale (macro) (Lascelles, 2007; Trendall, 2002). The sedimentation of the Weeli Wolli Formation (2.45 Ga) was accompanied by the intrusion of mafic, mainly doleritic dykes and sills, cross cutting all previous units (Harmsworth
et al., 1990). Presumably, the mineralization was upgraded by removal of silica and carbonate in Paleoproterozoic and Cenozoic, decreasing the thickness of high-grade beds locally by up to 40% (Trendall, 1983). Accumulations of the BIF scree and maghemite pisolite alluvials form up to 100 m thick, 150 km long and 5 km wide Tertiary iron channel deposits (CID) of economic importance, overlying the bedrock unconformably (Morris and Ramanaidou, 2007). They are a primary source of high-grade iron ores.

Although the Pilbara Craton ascended about 5 km since deposition of the Turee Group apart from very low-grade metamorphism, minor folding and uplift, the basin infill is essentially preserved as they were when deposited (Lascelles, 2007; Trendall, 1983, 2002). A “dome and basin” type fold pattern was imprinted in the S part of the Hamersley basin caused by minor deformations or deflections of open, gently N-S to NW-SE trending folds and en-echelon offset of anticlines and synclines (Lascelles, 2007; Trendall, 1983). The deformation intensity and metamorphic grade increases from north (prehnite-pumpellyte facies) to south (greenschist) (Lascelles, 2007; Trendall, 1983). The structural inventory is completed by two generations of steeply dipping SW and SE striking fault sets which are presumably post-folding (Trendall, 1983).

### 3.2.3 Geology of the studied deposit: “Deposit 99”

The “Deposit 99” is located in the wider Jimblebar region in the southern part of the Hamersley Basin where deposits are distributed along about 18 km in E-W extension at the far eastern end of the Ophthalmia Ridge. Topographically, the site is characterized by an E-W trending, S facing 100 m high ridge with deeply incised N-S to NE-SW trending valleys. The surface slopes smoothly to the North.

Six formations of the Hamersley Group have been explored (Table 3.1) by 3100 boreholes (Figure 3.1) penetrating up to 300 m of sediments between the Jeerinah Formation and the Brockman Formation, with the Brockman Formation BIF’s being the economic target. The stratigraphic column consists of BIF’s rhythmically interstratified with fine-clastic sediments and subordinate sequences of calcite and volcanic rocks (Harmsworth
et al., 1990; Trendall, 1983). Based on the relative amount of shale and BIF, Dales Gorge and Joffre Members of the Brockman Formation are subdivided into 10-50 m thick subdomains of alternating economic importance (Harmsworth et al., 1990). Doleritic dykes and sills penetrate the sediments unsystematically (Harmsworth et al., 1990). Only recently, the Marra Mamba Formation and the unconformable Tertiary Detrital deposits were recognized as potential economic targets. The iron ore distribution depends on a close match of stratigraphic and fault markers while maintaining geologic consistency, i.e. along unconformities. In this research, stratigraphies from O to Y are considered to be conformable, and JN, MU and TD2/3 are considered to be unconformable based on geological knowledge. Additional thickness variation originates from the unknown upgrading processes which cause the stratigraphies carrying the high-grade ore deposits to shrink.

Figure 3.2 shows the top of all boreholes projected onto the topographical map. Contact points shown on the map exhibit the first outcropping stratigraphy, discarding regolith and anthropogenic infills. The only fault altering the geometry of the stratigraphy noticeably at the entire deposit’s scale is indicated as a red line. This fault causes a vertical shift of around 5 meters for the West and East sides. It is not significant for the application but the fault is included into the modeling procedure for completeness of the approach.

3.3 Methodology

3.3.1 Potential field method

The potential field method is an interpolation algorithm (Lajaunie et al., 1997; Aug, 2004; Calcagno et al., 2008; Chilès and Delfiner, 2012; Gonçalves et al., 2017; de la Varga et al., 2019). It is a co-kriging system that takes contact points and orientations as inputs, and outputs an scalar field $Z(\vec{x})$ (or simply $Z$), called potential field, to represent geological structures. A geological interface corresponds to an iso-contour of the potential field - the set of points $\vec{x}$ that satisfy $Z(\vec{x}) = c$, a constant value. The iso-contour matches input
Figure 3.1: All structural data in deposit “99”. All contact points (small pixels), orientations (large pixels) for modeling and the corresponding stratigraphic column (also see Table 1). Spatial extent: 5840 meters in EW, 3026 meters in NS, and 448 meters in depth.
### Table 3.1: Stratigraphic column of “Deposit 99”

<table>
<thead>
<tr>
<th>Group</th>
<th>Age [Ga]</th>
<th>Formation</th>
<th>Member</th>
<th>Symbol</th>
<th>Max. Thickness [m]</th>
<th>Lithology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tertiary</td>
<td>0.01</td>
<td></td>
<td>TD2 - TD3</td>
<td>100</td>
<td></td>
<td>Fluvial sediments</td>
</tr>
<tr>
<td>Unconformity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamersley</td>
<td>2.45</td>
<td>Brockman</td>
<td>Y</td>
<td>40</td>
<td></td>
<td>chert, shale, dolerite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Joffre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Whaleback Shale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td>75</td>
<td></td>
<td>BIF with significant sheet silicate content</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dales Gorge</td>
<td>150</td>
<td></td>
<td>well-banded BIF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mt. McRae</td>
<td></td>
<td>Colonial Chert</td>
<td>D1</td>
<td>15</td>
<td></td>
<td>enriched in BIF</td>
</tr>
<tr>
<td>Shale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mt. McRae Shale</td>
<td>R</td>
<td>85</td>
<td></td>
<td>mudstone, siltstone, chert, IF, dolomite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mt. Sylvia</td>
<td></td>
<td></td>
<td>S/S7</td>
<td>30</td>
<td></td>
<td>mudstone, siltstone, chert, IF, dolomite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wittenoom</td>
<td></td>
<td></td>
<td>O</td>
<td>700</td>
<td></td>
<td>calcitic dolomite, chert, shale, volcaniclelastic sandstone</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unconformity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marra Mamba</td>
<td></td>
<td>Nammuldi</td>
<td>M/MU</td>
<td>100</td>
<td></td>
<td>Chert and cherty BIF beds with some shale bands</td>
</tr>
<tr>
<td>Unconformity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fortescue</td>
<td>2.68</td>
<td>Jeerinah</td>
<td>JN</td>
<td>1800</td>
<td></td>
<td>Interbedded volcano-sedimentary units</td>
</tr>
</tbody>
</table>

The co-kriging estimator (Chilès and Delfiner, 2012) has the form of:

$$Z^*(\bar{x}) - Z^*(\bar{x}_0) = \sum_{\alpha=2}^{m} \lambda_\alpha \left[Z(\bar{x}_\alpha) - Z(\bar{x}_{\alpha-1})\right] + \sum_{\beta=1}^{n} \sum_{i=1}^{3} u_{i\beta} \frac{\partial Z}{\partial u_i}(\bar{x}_\beta) \quad (3.1)$$
\( x_0 \) is an arbitrary reference point in space. \( \bar{x}_\alpha \) and \( \bar{x}_{\alpha-1} \) are contact points that belong to the same geological interface. \( \bar{x}_\beta \) denotes locations of orientation data (i.e., gradients). \( m \) and \( n \) denote numbers of observations for contact points and gradients, respectively. \( \lambda_\alpha \) and \( v_{i\beta} \) are weights for potential increments, \( Z(\bar{x}_\alpha) - Z(\bar{x}_{\alpha-1}) \), and weights for each gradient component, respectively. Note that \( Z(\bar{x}_\alpha) - Z(\bar{x}_{\alpha-1}) = 0 \), because \( \bar{x}_\alpha \) and \( \bar{x}_{\alpha-1} \) belong to the same geological interface. Though the potential increment for \( \bar{x}_\alpha \) and \( \bar{x}_{\alpha-1} \) is 0, its presence affects the weights on the gradients.

The estimator focuses on estimating the increment of potential field \( Z^*(\bar{x}) \) relative to the reference point \( x_0 \)’s potential field \( Z^*(\bar{x}_0) \). The potential field itself does not necessarily have any physical meaning, though it could be interpreted as “geological time” (Chilès and Delfiner, 2012). One special feature of the implicit potential field method is that its primary variable - potential field, cannot be observed in data. Therefore, usually a pre-determined variogram model is used and weights are estimated by a co-kriging system (Aug, 2004; de la Varga et al., 2019). \( Z(\cdot) \) is assumed to be a random function with a polynomial
drift and a stationary covariance (Chilès et al., 2004) (see Appendix B for mathematical details about the drift function and choice of covariance/variogram). The method assumes the mean of the scalar field changes in the same direction as the orientation data and it is estimated by a function of known basis functions (de la Varga et al., 2019). Faults are considered as offsets of the potential field defined by adding additional drift functions into the co-kriging system (Calcagno et al., 2008). Details about the co-kriging system for the implicit potential field method can be found in Appendix B.

In addition to represent geological structures using iso-surfaces $Z^*(\bar{x}) = c$, geological layers can also be visualized with a categorical model obtained by thresholding the potential field using $Z^*(\bar{x}) = c$. Geological layers that belong to the same conformable set are converted from the same potential field, and un-conformable ones are from different potential fields. A truncation rule is needed to define a geological relationship (Calcagno et al., 2008; de la Varga et al., 2019). In the rest of the chapter, $m$ is used to denote the implicit structural model for simplicity.

### 3.3.2 Issues with direct application of interpolation

An open source implementation of the potential field method called GemPy\(^1\) (de la Varga et al., 2019) is used for this research. Figure 3.3 shows a single vertical section computed from 3100 boreholes and 19 layers of the BIF only using the standard interpolation algorithm of GemPy. The most noticeable issue in Figure 3.3 is that geological rules such as the conformity (O-Y) and geological age constraints are not obeyed. Even with proper data analysis and processing (as shown in Section 3.4.1), a direct application is clearly fraught with difficulties. While ad-hoc changes could resolve some of these issues, a more fundamental solution is needed.

There are several reasons that may make direct application of the implicit potential field method challenging:

- The contact points are too dense to be handled by a single co-kriging system. A

\(^1\)https://www.gempy.org/
Figure 3.3: Issues of the model obtained by a direct application of the potential field method on the large BIF case. Shown in a 2D section, taken at $Y = 247200$ m.

local co-kriging is likely to generate artifacts in layering because of the use of a limited neighborhood. A global co-kriging, however, may also create artifacts if the data density is high, due to the difficulty in estimating an adequate spatial continuity model for a complex large-scale structure.

b. Insufficient orientation information exists to constrain the drift function of the scalar field. This may result in unrealistic folding of structures.

c. The co-kriging method itself simply interpolates data but does not guarantee that geological rules, as interpreted from data or based on prior geological knowledge, are followed.

These limitations motivate the proposal of the following approach.

3.3.3 Overview of the entire approach

The previous section shows that a single pass approach of the potential field method remains challenging to large cases. This section presents a new approach aiming to address
such cases. Before providing the details of each step, all steps are shortly outlined and
motivated. A framework is illustrated in Figure 3.4. Extensive results for the full case are
shown in Section 3.4.

Step 1: Data processing and analysis. We need to process and understand the data of a
specific case before moving forward. Faults that mostly affect the deposit are included and
topography constraints are added to build more realistic models. Histogram and variogram
analysis on the orientation information is conducted.

Step 2a: Selection of multiple subsets of contact points. It reduces the complexity of
the co-kriging system in this large-scale scenario. To aid the eventual conditioning of all
contact points, subset contact points are selected in a way such that all contact points are
included in a few (finite) subsets of contact points (i.e., each contact point is a part of at
least one subset).

Step 2b: Stochastic simulation of orientation data. Orientation data are not measured
at all boreholes, but orientation data are important constraints in the implicit potential field
method to generate geological realism. To include such data with spatial uncertainty in-
corporated, stochastic simulation of the missing orientations at contact point locations is
performed, using estimated orientations from contact points as soft data and measured ori-
entations as hard data.

Step 3: Creation of multiple 3D models for validation. The implicit potential field
method is used to create one 3D model by co-kriging one subset of contact points and one
set of conditionally simulated orientation data. Multiple 3D models are created by repeat-
ing this process with different contact point subsets and different simulated orientations.
Because the set of partly constrained geological models does not match all data, a vali-
dation test can be performed. The validation step includes both data-based validation and
geological rule-based validation.

Step 4: Generation of the final 3D model by linear combination of implicit functions.
The previous step only generated intermediate 3D geological models with partial condi-
tioning. In this step, local linear combination of potential fields is performed to create a 3D
model that match all data.
CHAPTER 3. LARGE-SCALE IMPLICIT MODELING

Figure 3.4: Framework for modeling large-scale geological structures.

The methodology is now detailed in Step 2a, 2b, 3 and 4. Details for step 1 are discussed in the real case application section.

3.3.4 Selection of subsets of contact points

When selecting subsets of contact points, the following constraint is considered: the finite sets of subset contact points constitute the set of all contact points. This constraint is needed for the final model to match all contact points after performing a linear combination (step 4). Instead of randomly selecting contact points, the approach selects them by boreholes, containing contact points from geological layers. In this way, the geological sequence could be better honored since boreholes already contain necessary stratigraphical information such as relative positions of layers, while random contact points could miss out
such information if poorly selected. Therefore, the procedure for selecting multiple subsets of contact points is as follows:

a. Define the number of sets of boreholes used for linear combination (denoted as $r$);  

b. Randomly divide boreholes into $r$ subsets.

The number of possible subsets that can be obtained in this way can be calculated. Suppose we have $n$ boreholes in total, and we divide $n$ boreholes into $r$ sets, then there are $n/r$ boreholes per set for co-kriging. The number of different sets of contact points is:

$$C\left(n, \frac{n}{r}\right) = \frac{n!}{(n-n/r)!(n/r)!}$$

(3.2)

For linear combination, $r$ sets of boreholes are needed to generate one fully conditioned model, and the total number of models that can be obtained by linear combination is: $C\left(n, \frac{n}{r}\right) / r$. Although the proposed workflow uses a stochastic approach for subset data selection, other approaches that are based on clustering methods (van Stein et al., 2020), or incorporate some statistical properties (Carmichael and Ailleres, 2016) may be helpful for other geological applications.

### 3.3.5 Simulation of orientation data

Due to the role of drift function in a scalar field, orientations take on an important role in maintaining geological realism in cases such as modeling folded sequences of BIFs. Figure 3.5 shows that without orientation data on the East side (right side of the dashed line), the model violates geological age constraints in this BIF case (e.g., D1 – J6 should all be conformable layers and no circular shapes are expected). Previous research has shown that additional orientation information could help to model folds (Laurent et al., 2016). Here, a statistically driven approach different from the structurally driven one in Laurent et al. (2016) is used. To guide the 3D modeling of the deposit and create geological realistic folding sequences, orientations are assigned to each contact point through co-simulation (Journel, 1999). First, orientation data from contact point information are estimated. To do it, a plane is fitted using a few nearest contact points of a contact point to estimate the
orientation at that contact point. Then, the estimated orientations are used as soft data and the measured orientations are used as hard data to create multiple realizations of orientation at contact points locations. The following bijective transformation is used to calculate gradients $G_x$, $G_y$, and $G_z$ from the angular data - azimuths ($A$) and dips ($D$) (Aug, 2004):

$$
\begin{align*}
G_x &= \sin(D) \times \sin(A) \times \text{polarity} \\
G_y &= \sin(D) \times \cos(A) \times \text{polarity} \\
G_z &= \cos(D) \times \text{polarity}
\end{align*}
$$

where polarity takes either 1 or -1, depending on the observation. Equation 3.3 ensures that $G_x^2 + G_y^2 + G_z^2 = 1$, which is used for normalization of gradients after co-simulation.

Figure 3.5 shows that by expanding the orientation information, without the previous subset data selection step, a result that is visually more plausible can be obtained. However, the result still contains topological errors, especially on the East side, so the subset data selection step is still needed.

### 3.3.6 Validation

Using multiple subsets of contact points and simulated orientations, the original potential field method is applied to create multiple (intermediate) models of the deposit. The question is, however, how many subsets should be created, i.e., what is $r$ in Equation 3.2? A large $r$ (i.e., less data in each set) will result in overly smooth and unrealistic models, and a small $r$ (i.e., more data in each set) will likely create unrealistic models just as using all data at once. To assess a suitable value for $r$, the approach relies on a validation test as follows. Suppose that $N$ 3D intermediate geological models have been generated, each matching a subset of data:

a. For each intermediate 3D geological model, extract the potential field value, at those locations where boreholes were not used as input data;

b. Using the potential field value, create a matrix of size $N \times L$, where $L$ is the number of boreholes used;
Figure 3.5: Effect of orientation information on the interpolation result (only enriched BIF layers D1 - J6 are shown and they are all conformable layers). (a) Contact points (small pixels) and orientations (large pixels); (b) the interpolation result ($Y = 26700$ m, in 2D) with only observed orientations; (c) the interpolation result ($Y = 26700$ m, in 2D) with simulated orientations but without dividing the contact points.
c. Perform principal component analysis (PCA) (Hotelling, 1933; Jackson, 1991) to reduce the dimensionality in this dataset, but keeping all components;

d. Perform outlier detection, through the Mahalanobis distance in principal component space, to assess the difference between data and models (Rousseeuw and Driessen, 1999; Yin et al., 2020).

In the application section (Section 3.4), a more detailed validation analysis with results will be presented. In addition to data-based validation, a geological rule-based validation is also needed to ensure the 3D models are geologically realistic. In the BIF context, models that have circular/closed objects should be rejected. This geological rule can be automatically checked based on the topology of iso-contours extracted from the 3D geological model. Figure 3.6 shows an example of accepting or rejecting models based on the existence of closed objects. A closed object can be detected by checking if the starting point of an iso-contour is the same as its ending point.

**Figure 3.6:** Automatic geological rule check based on the topology of iso-contours. (a) a verified example based on the geological rule; (b) a rejected example based on the geological rule.
3.3.7 Local linear combination of two potential fields

At this point multiple validated 3D geological models (and corresponding potential fields), each matching a subset of borehole data, have been generated. The question is on how to combine this into a single 3D model that matches all data. To do so, a local linear combination of the potential fields is performed, meaning that the weight \( w \) for linear combination is spatially varying. Let us first analyze the case on linearly combining two potential fields \( m_1 \) and \( m_2 \):

\[
m = w \cdot m_1 + (1 - w) \cdot m_2, \quad 0 \leq w \leq 1
\]  

(3.4)

\( m_1 \) and \( m_2 \), each matches a subset of the data (Figures 3.7 (a-b)). Then, at those locations where \( m_1 \) matches the data, \( w \) has to be equal to 1 (model \( m_1 \) gets all the weight), similarly for model \( m_2 \). For other spatial locations, the inverse distance weighting (IDW) method is used to interpolate a smooth map of \( w \), with weights always between 0 and 1, as in equation 3.4. This condition ensures that the method is not “extrapolating” from input potential fields but only creating an “in-between” version of them. Note that there are other ways to obtain \( w \), such as in van Stein et al. (2015, 2020) where the kriging variance is accessible. In this case, let \( v_i \) denote the kriging variance of the \( i^{th} \) interpolation result, then the weight given the \( i^{th} \) interpolation is:

\[
w_i = \frac{1}{(v_i + \epsilon) \sum_j^n \left( \frac{1}{v_j} + \epsilon \right)}
\]  

(3.5)

\( \epsilon \) is a small value (e.g., \( 10^{-9} \)) used to avoid a 0 divider.

With either IDW or kriging variances to calculate weights, the spatial variation of \( w \) is function of the data locations (Figure 3.7c). It should be noted that threshold values for converting a potential field into a categorical model are no longer constant values, but spatially varying in the same manner as \( w \). That means, the new isovalue to represent a geological interface should be \( cw \), if it were \( c \) before. Figure 3.7d shows one of such threshold maps (there are 11 of them in total for D1 - J6). A linear combination of two
potential fields is still a potential field, because the sum of two differentiable scalar fields is still differentiable. In addition, in Appendix C, it is shown that at data locations with either contact points or orientation data, conditioning to such data is preserved with linear combination. Therefore, this spatially varying \( w \) can be used to linearly combine two inputs and get a new model that matches both input data (Figure 3.7e and Figure 3.7f).

### 3.3.8 Linear combination of multiple potential fields

Previously, linear combination of two potential fields has been discussed. In our case, \( r \) intermediate potential fields are generated, hence the linear combination is generalized to
multiple potential fields. We use $m_1, \ldots, m_r$ to denote input potential fields, then the final potential field $m$ is a weighted linear combination of all input potential fields:

$$m = w_1 \cdot m_1 + w_2 \cdot m_2 + \ldots + w_r \cdot m_r$$  

(3.6)

$w_i, i = 1, 2, \ldots, r$ are weights, and they have to satisfy:

$$\sum_{i} m_i = 1, w_i \geq 0$$  

(3.7)

Similar to the binary case shown in Section 3.3.7, for each contact point, the interpolation result with the zero (or the smallest) error has a weight of 1, and other interpolation results have weights of 0. For locations other than contact points, IDW is used to produce a spatially smooth map of $w_i$. Similar to the binary case, the condition in Equation 3.7 and the smoothness of $w_i$ help to maintain the correct topology of validated input potential fields when they are linearly combined.

### 3.3.9 Walk-through of key steps with 2D synthetic example

In this section, a synthetic 2D example is used to show how linear combination of multiple scalar fields works in general. This example is simpler than a potential field case because it does not use orientation data as a secondary variable, nor geological rule constraints. The purpose of this example is to explain how some of the approach’s key steps work, including: subset data selection, data-based validation and linear combination for fitting. The “true model” is one Gaussian realization generated with a Gaussian type variogram with a range of 20 and a sill of 1 and $n = 500$ data points are sampled from the true model to mimic a case with dense data and smooth variation (Figure 3.8 (a-b)). To obtain a reference for comparison, global simple kriging is performed using all data with the same variogram as the “true model” (Figure 3.8 (c-d)).

Next, subsets of data are selected. If all data are divided into 5 sets (i.e., $r = 5$), then 100 points are used per kriging (i.e., $L = 100$), and all 5 sets are needed to fully condition
Figure 3.8: Synthetic 2D example. (a) True model; (b) 500 data points; (c) Kriging result; (d) Kriging variance.
the final linear combination. 500 different subsets of data points and 500 kriging results are generated. Figure 3.9 shows one possible division of data (5 sets) and corresponding kriging results and kriging variances.

Then, the validation step (Section 3.3.6) is conducted to check if those 500 kriging estimates can be validated by data points that were left out for kriging (Figure 3.10). To do this, PCA is applied to the vectors of kriged values observed at data locations. At the same time, the actual data is projected onto the same principal component vectors. The plot of the first two PC scores (Figure 3.10 (a)) shows that data are inside the cloud of kriging. Figure 3.10 (b) is the scree plot that shows the accumulative variance of PCA. To extend this check to all PC scores dimensions, the Mahalanobis distance between PC scores is calculated and a robust outlier detection method is used to detect if the data is an outlier (Rousseeuw and Driessen, 1999). Figure 3.10 (c) shows the data (red diamond) is below the outlier detection line, with 95% confidence (dashed line). These results indicate that $r = 5$ is suitable.

Finally, linear combination of 5 intermediate kriging estimates is conducted. In this
Figure 3.10: Validation of 500 Kriging results. (a) First two dimension of PCA scores of 500 kriging results (blue dots) and left-out data (red cross); (b) Scree plot of PCA; (c) Mahalanobis distance between 500 kriging results (green to blue dots) and data (red diamond).

In case, the kriging variance is used to determine weights $w$ (see Equation 3.5) because the variance is easy to obtain for the simple kriging method. 100 different linear combination results are obtained. Figure 3.11 (a–b) shows a comparison of two linear combination results. Notice that they are very similar. If comparing the mean of 100 linear combination results with kriging (Figure 11 (c–d)), one can see that the linear combination result is quite close to kriging.

Consider now a much lower number of data points ($r = 10$, $L = 50$) per subset. Figure 3.12 shows that the kriging result is overly smooth and the linear combination creates artifacts, as predicted by the validation test.

### 3.4 Full application to case study

#### 3.4.1 Data processing and analysis

Boreholes were drilled from the Earth’s surface, hence a separate implicit potential field for the topography, based on the position of borehole mouths, is modeled. The resulting topographic surface has been used to truncate the stratigraphic column. It is also important to analyze the existing orientation data because of its control on thickness variation.
Figure 3.11: Comparison of linear combination results and kriging \((r = 5, L = 100)\). (a-b) Two linear combination results; (c) Ensemble average of 100 linear combination results; (d) Kriging result with the whole dataset.

Figure 3.12: Results with \(r = 10, L = 50\). (a) Validation based on data; (b) Kriging result; (c) Linear combination result.
Figure 3.13: Observed gradients and their statistics. (a) Histograms of gradients; (b) Variograms for each component of the gradient. \( G_x, G_y \) and \( G_z \) from left to right. Variograms types: exponential, exponential, exponential; nugget effects: 0.08, 0.12, 0.02; sills: 0.12, 0.17, 0.025; ranges: 200, 200, 200

and thereby geological realism. Orientation data are lacking on the East side, hence the data measured at the Western part of the deposit is used to extract the spatial structure of orientation data and apply it to the East. This is a reasonable assumption because the regional geology does not indicate a significant change in folding geometries. To estimate the spatial structure of the Western part of the deposit, angles are converted into gradients using Equation 3.3 and then calculate variograms. Figure 3.13 (a) shows histograms of three components. It is noticeable that \( G_x \) and \( G_y \) are approximately Gaussian distributions, while \( G_z \) is a skewed distribution. Most \( G_z \) values are near one because BIF layers are mostly horizontal in the studied region. Figure 3.13 (b) shows a correlation structure with a range of 200 meters for all three components.
3.4.2 Input data generation

All boreholes are randomly sampled without replacement and distributed into 5 subsets. The number five is chosen according to the validation step (see results of the validation step in Section 3.4.3, where Figure 3.17 shows $r = 5$ is validated for the BIF case). Figure 3.14 shows two subsets of boreholes. The Markov Model II (MMII, Journel (1999)) is used for co-simulation using cosgsim for its simplicity (Remy et al., 2009). Only the correlation coefficient between hard and soft data is needed. Figure 3.15 shows the correlation of estimated orientations at locations where observed orientations exist. The correlation is high when the estimated orientations are accurate compared to the observed data. Figure 3.16 shows one realization of simulated gradients.
3.4.3 Intermediate 3D model generation and validation

Using simulated orientations and selected subsets of contact points, multiple 3D models are generated with co-kriging, each matching a subset of boreholes and all simulated gradients. As discussed in Section 3.3, we need to validate the parameter $r$ and select validated models. Figure 3.17 shows the data-based validation step of the entire deposit for all stratigraphies using all data, and using each stratigraphy’s data separately (only TD3 and D1 are shown as examples for unconformity layers and conformity layers, respectively). The distribution of Mahalanobis distances demonstrates that models are validated by the data. Rule-based validation is also performed as discussed in Section 3.3.6. Figure 3.18 (a-b) exemplifies two models, each of them fitting a subset of contact points and simulated orientations, validated by data and geological rules for a BIF. Sections from a third model are shown with contact points in Figure 3.18 (c).

To investigate the variability among the intermediate models, the frequency of any spatial location being on top of the boundary of D4 is calculated to get a probability map (Figure 3.19). One notices that more variations exist in locations with less data and less variations than in locations with denser data.
Figure 3.16: One cosgsim realization of three components of gradients. (a) $G_x$; (b) $G_y$; (c) $G_z$. 
Figure 3.17: Data based validation for the BIF case. From left to right are plots of PC scores, cumulative explained variance and the Mahalanobis distance. Red cross and red diamonds indicate the observed data. (a) Validation using all stratigraphies’ data; (b) Validation using only data of TD3; (c) Validation using only data of D1.
Figure 3.18: Three intermediate models using subsets of data and simulated orientations. (a) One section of $m_1$; (b) One section of $m_2$; (c) Three sections of $m_3$ with data.
3.4.4 Linear combination of multiple implicit models

Validated potential fields are used to perform the final linear combination step. Figure 3.20 (a-b) exemplifies two possible linearly combined models that fit all observed data (BIF-enriched D1 – J6 are shown in 2D) with very little small-scale variability. Figure 3.20 (c) shows the fault in 3D, along with stratigraphies D1 – J4. Figure 3.20 (d) shows all stratigraphies, including topography and basement, in one model resulted from averaging multiple linear combinations. One can see that the model complies with the geological age constraint, rules of conformity (O - Y) and unconformity (JN, MU, TD2/3). Figures 3.20 (a-d) show that the fault causes noticeable but not significant geometrical changes at the entire BIF’s scale. Figure 3.20 (e) shows contact points conditioning for all stratigraphies.

3.4.5 Discussion on the runtime and RAM

The approach is implemented based on GemPy. Modeling is conducted on 3.5 GHz Quad-Core Intel Core i5 CPU and 64GB RAM. Table 2 shows the runtime of time-consuming steps in the approach for the full application case, with all data as inputs and a section grid.
Figure 3.20: Final models after linear combination. (a-b) two final models; (c) the fault block; (d) model in 2D with correct rules; (e) shows the consistency with data in 3D.
with a dimension of 300 * 150 * 1 as output. Other steps such as data analysis, subsets selection of boreholes, and model validation take negligible time compared to these steps. As shown in Table 3.2, model generation (i.e., the interpolation procedure itself) is the most time-consuming. The linear combination step also takes some time because of the IDW interpolation step for computing weights. All steps in the proposed approach can be accomplished using less than 64GB RAM; however, a direct application of interpolation with all data would exceed the maximum RAM capacity due to the large amount of data in the co-kriging system, which would bog down the computation.

Table 3.2: Runtimes of time-consuming steps (for 5 intermediate models)

<table>
<thead>
<tr>
<th>Step</th>
<th>Gradients simulation</th>
<th>Intermediate model generation</th>
<th>Linear combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>7s</td>
<td>41 min 38 s</td>
<td>14 min 19s</td>
</tr>
</tbody>
</table>

3.5 Conclusions

This chapter presents a methodology to model the variability of large implicit geological structures in the context of numerous borehole data. The large contact point dataset is divided into multiple sets, and the scarce orientation dataset is expanded with co-simulation. An illustration case shows the that the division of the large contact point dataset is necessary, even with expanded orientation information. Each subset of contact points and stochastic gradient information are then used as inputs for co-kriging to create implicit geological models. To make sure generated models are consistent with unused data and geological rule, the method contains an automatic and effective validation procedure to check and select models. Data and geologically validated models obtained with sub-datasets are finally combined linearly to obtain final fully conditioned models. Theoretical proofs show the rationale of using linear combination of potential fields for modeling geological structures.

The framework has been applied to modeling a large-scale banded iron formation in
Western Australia. Results show that the method is able to generate one geologically realistic 3D structural model that fit all data. The result benefits both from the division of dense contact points and the addition of stochastic orientations. It can be seen that robustness is brought to co-kriging by linearly combining results obtained using sub-datasets. In that sense the idea of combining multiple kriging on smaller sets of data could be extended to big data type kriging problems such as filling gaps in remote sensing (Mariethoz and Caers, 2015). Other applications of large-scale implicit structural modeling with dense data may also benefit from the proposed method. This chapter did not discuss interpolation uncertainty of the interpolation method, which can be found in Chilès et al. (2004), and the interpolation uncertainty can be assessed and visualized using the idea from the previous chapter Chapter 2, or Yang et al. (2019).
Chapter 4

Quantifying Uncertainty of Implicit Geological Structures

4.1 Introduction

The spatial relationship between rock properties (e.g., porosity, permeability, density, and mineral grade) and corresponding geological structures is an important aspect in many subsurface systems (e.g., reservoirs, mineral deposits, aquifers, and urban areas) (Caumon, 2010; Caers, 2011). 3D geological structural models are tools to help convey this geological knowledge. However, the challenge is that uncertainties are common in these structural models (Mallet, 2014). Often, geophysical data are acquired to reduce the uncertainty of geological structural models, because they provide information on the physical properties of subsurface, as well as critical information and constraints for subsurface forecasting and management (Linde et al., 2015; Wellmann et al., 2017; Scheidt et al., 2018).

There are different ways to include geophysical data into modeling workflows, and one of the most common is via inversions (Sambridge and Mosegaard, 2002; Tarantola, 2005; Guillen et al., 2008; Jessell et al., 2014). Inversions can be done in either a probabilistic or an optimization framework (Tarantola, 2005; Caers, 2011). A lot of related work has been done in solving spatial inverse problems other than geological structures,
to name a few, Caers (2007); González et al. (2008); Saibaba and Kitanidis (2015); Grana et al. (2017); Pradhan and Mukerji (2020). For geophysical inversion with time-consuming forward models, many computational advances have been made to reduce the number of forward model computations within the classic Bayesian geostatistical framework to solve large-dimensional inverse problems (Liu and Kitanidis, 2011; Kitanidis and Lee, 2014; Zhao and Luo, 2020, 2021). Kitanidis and Lee (2014) utilizes a matrix free in terms of the Jacobian matrix Gauss-Newton method and improves the scalability of the geostatistical inverse problem. Zhao and Luo (2020) develops a reformulated framework of Bayesian inverse modeling for large-scale spatial fields on reduced dimensions comprised by principal components. In Zhao and Luo (2021), the inverse problem is transformed from direct estimation of the underlying parameter fields to estimating the coefficients or projections on the dominant principal components.

In recent years, an uncertainty quantification protocol called Bayesian evidential learning (BEL) has gained popularity (Hermans et al., 2018; Scheidt et al., 2018; Athens and Caers, 2019; Yin et al., 2020; Pradhan and Mukerji, 2020; Athens, 2021). BEL is a prescriptive and normative data-scientific protocol for designing uncertainty quantification within the context of decision making, which has been used in various applications and inverse problems (Athens, 2021). BEL relies on learning a statistical relationship between the prediction variable and data, or spatial models and data. Posterior samples are drawn from the learned statistical relationship. BEL also takes advantage of dimension reduction techniques to make uncertainty quantification a computational efficient task. However, one challenge of BEL lies in making sure the learned statistical relationship is reliable, especially when the forward model is non-linear.

Though uncertainty quantification of subsurface systems is an active research topic, research in uncertainty quantification of geological structures that entails using a Bayesian framework is not as extensive as other spatial problems (Dorn et al., 2013; de la Varga and Wellmann, 2016; Aydin and Caers, 2017; Wellmann et al., 2017; Grose et al., 2018). Dorn et al. (2013) uses rejection sampling to obtain a set of discrete fracture network models
that are consistent with both hydrogeological and geophysical data. de la Varga and Wellmann (2016) treats geological modeling as an inversion problem and combines uncertain prior information with geologically motivated likelihood functions in a Bayesian inference framework. Aydin and Caers (2017) uses a Markov Chain Monte Carlo sampler (specifically, a marked Strauss point process) to sample posterior fault network realizations that reflect tectonic knowledge and honor fault observations interpreted from seismic images. Wellmann et al. (2017) used an McMC framework with the adaptive Metropolis algorithm (Haario et al., 2001) for inversion of implicit geological structures with gravity data; however, the high dimensionality of the problem makes it challenging to accurately represent posterior distributions and difficult to evaluate the convergence of iterative samplers. Grose et al. (2018) frames structural modeling as an inverse problem and uses a Bayesian framework to reconcile structural parameters and structural data uncertainties.

In summary, uncertainty quantification of complex geological structures is still challenging. First of all, the dimension of spatial parameters is very high; no simple solution is available. Secondly, it is often time consuming to build multiple geological structural models. Last but not least, the forward model is often non-linear, so many methods based on linear assumptions (Grana et al., 2017) or simple linear transformations (Scheidt et al., 2018) may not work well.

To address these challenges, a framework for quantifying uncertainty of implicit geological structures with geophysical data (e.g., gravity data) is presented in this study. It adopts the dimension reduction idea from recent work in uncertainty quantification of large-scale spatial problems (Kitanidis and Lee, 2014; Scheidt et al., 2018; Athens and Caers, 2019; Zhao and Luo, 2021). Implicit representation of geological structures transforms discrete geological objects into a continuous variable, i.e., a scalar field; dimension reduction techniques such as principal component analysis (PCA) (Hotelling, 1933; Jackson, 1991) can be applied with the implicit representation. Therefore, in this research, we take advantage of the implicit representation to apply PCA for reducing the dimension of an ensemble of realizations of geological structures. After dimension reduction, the prior distribution is formulated in the dimension reduced space and save computing time for
sampling new model realizations from the low dimensional space. The non-linear forward model is kept and used for generating data realizations in the original space. Therefore, the entire uncertainty quantification task becomes tractable.

This research uses the same sandstone-greenstone belt case with gravity data from Wellmann et al. (2017) for illustration. First, a geological background regarding the studied region and the gravity data and the forward model are introduced in section 4.2. Then in section 4.3, the framework and details of each step are explained. Next, section 4.4 shows results of uncertainty quantification for the studied sandstone-greenstone belt. Finally, section 4.5 discusses limitations of the method and section 4.6 concludes the chapter.

4.2 Case study

4.2.1 Geological background

The studied region is introduced in Wellmann et al. (2017). It is a sandstone-greenstone belt in the Youanmi Terrane of the Yilgarn Craton in Western Australia (Figure 4.1). The studied sandstone-greenstone belt is located near the major north–south-trending Younami Shear Zone and is surrounded by younger Neoarchaean granites. The structure of the belt is interpreted as a refolded fold: the initial north–south compression resulted in an east–west syncline, which was refolded under east–west compression, resulting in a northwards closing fold (Wellmann et al., 2017). The greenstone consists of several Neoarchaean volcano–sedimentary rock sequences. The stratigraphy consists of a mafic sequence, a schist layer, basalts and banded iron formations (BIFs) with some ultramafics and metasediments at the top (Figure 4.1). The contrast of high density rocks (BIFs and mafics) and the surrounding low-density basalts can be observed in the gravity measurement (see Figure 4.1b). Greenstone belts draw people’s interests because they often host gold, iron and base metals in quantities that are economically significant for mineral exploration (Groves et al., 1998).

For the studied region, structural information including contact points and orientations of interfaces were inferred from geological maps, a few drill-holes, the seismic survey
and a preliminary gravity forward modelling result (Figure 4.1c). For this reason, the inferred contact points and orientations are considered uncertain. A reference model is built using the implicit potential field method (see Chapter 3) with the inferred contact points and orientations (Figure 4.2). For simplifying the system, the region is simplified to four lithologies, which are then grouped into three lithology series in total, as in Wellmann et al. (2017). Lithologies that belong to the same series are considered geologically conformable, including “Simple BIF” and “Simple Mafic 2” forming the conformable “BIF Series”. Each lithology’s rock density was obtained based on information of a similar greenstone area (Williams, 2009; Wellmann et al., 2017). Table 4.1 shows series names, lithology names and each lithology’s average rock density of the studied region.

<table>
<thead>
<tr>
<th>Series</th>
<th>Lithology</th>
<th>Density ($g \cdot cm^{-3}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIF Series</td>
<td>Simple Mafic 2</td>
<td>2.92</td>
</tr>
<tr>
<td>BIF Series</td>
<td>Simple BIF</td>
<td>3.1</td>
</tr>
<tr>
<td>Simple Mafic Series</td>
<td>Simple Mafic 1</td>
<td>2.91</td>
</tr>
<tr>
<td>Early Granite Series</td>
<td>Early Granite</td>
<td>2.61</td>
</tr>
</tbody>
</table>

**4.2.2 Gravity data and gravity forward modeling**

In recent years gravity measurements have been a valuable additional geophysical data source to support geological modeling (Phelps et al., 2018; Athens, 2021). For this research, a synthetic gravity data generated with the forward modeller detailed below, based on the reference 3D lithology model (Figure 4.2), is used. Using the method of Nagy et al. (2000) for rectangular prisms, the gravitational pull of every voxel of the lithology model is calculated, for a given density and distance in the vertical component, by differentiation with respect to $z$:

$$
F_z = G \rho \cdot u_z = G \rho \cdot ||| x \ln (y + r) + y \ln (x + r) - z \tanh \frac{xy}{zr} || x_1 y_1 || x_2 y_2 || x_1 y_1 || x_2 y_2 || x_1 y_2 \| z_1 \| z_2 \| z_1 $$

(4.1)
Figure 4.1: The studied sandstone greenstone belt and data (Wellmann et al., 2017). (a) Geological map of the studied region (adapted from Wellmann et al. (2017)); (b) synthetic observed gravity anomaly; (c) contact points and orientations inferred from various sources of information, and grid points for gravity measurements (black dots).
Figure 4.2: Reference model generated using Gempy. (a) Reference model in 3D surfaces. (b) Top view of the reference model.
where \(x, y\) and \(z\) are the spatial locations of the prism, using the measuring point as the origin, \(r\) is the Euclidean distance from the prism to the measurement point, \(G\) is the gravity constant, and \(\rho\) is the density.

The same forward model (Equation 4.1) is used throughout the study, for generating both the synthetic gravity observation (Figure 4.1b) and prior data realizations (see details in Section 4.3). Though synthetic gravity data is used for methodology development purposes, one may find the observed Bouguer gravity anomaly from the government of Western Australia\(^1\). The synthetic gravity data is shown in Figure 4.1b. The synthetic gravity data is gridded with about 2.5 km spacing in NS and 2 km spacing in EW over the study area. Physical devices (e.g., airborne, onshore, or offshore) that collect gravity measurements are usually not as evenly distributed as the grid, and an additional step (e.g., interpolation or simulation) is usually needed to fill in gaps between devices (Athens, 2021). This synthetic data mimics data after the additional gap-filling step. Gravity gradients are calculated on the same grid as they are proved to be useful for detection of geological interfaces (Figure 4.3).

### 4.3 Methodology

#### 4.3.1 Overview

An implicit representation of geological structures is denoted as \(\mathbf{m}\); it actually contains two components: scalar fields \(\mathbf{m}_{sf}\) and iso-contour values of geological interfaces \(\mathbf{c}\), so:

\[
\mathbf{m} = [\mathbf{m}_{sf}, \mathbf{c}]
\]  

\(^1\)The government of Western Australia provides public access to a grid of merged ground and airborne Bouguer gravity generated with a cell size of approximately 400 metres (Geological Survey of Western Australia, 2020 gravity merged 400 m grid of Western Australia (version 1), which can be downloaded from https://www.dmp.wa.gov.au/Geological-Survey/Regional-geophysical-survey-data-1392.aspx/. This grid combines onshore Complete Spherical Cap Bouguer gravity anomalies with offshore Free-air anomalies for Western Australia.
A Bayesian framework is used to quantify the uncertainty of implicit geological structures, with an aim to obtain posterior realizations given the geophysical data (Tarantola, 2005; Scheidt et al., 2018). Using \( p(m) \) to denote the prior distribution of implicit geological structures, \( d_{obs} \) to denote observed geophysical data and \( L(d_{obs}|m) \) to denote the likelihood function, the posterior distribution \( p(m|d_{obs}) \) of implicit geological structures is defined as:

\[
p(m|d_{obs}) \propto p(m) L(d_{obs}|m)
\]

Realizations of implicit geological structures are stored in 3D voxel models that contain \( N \) grid cells (\( N \) is usually very large), which mathematically, are high dimensional (\( N \)-dimensional) vectors. Due to the difficulty in directly obtaining and sampling from \( p(m|d_{obs}) \) because of its high dimensionality, the method instead aims to sample posterior realizations from a dimension reduced space. Afterwards, dimension reduced posterior realizations are transformed back to the original space. Using \( m^* \) to denote dimension
reduced realizations, the posterior distribution the method actually samples from is:

\[
p(m^*|d_{\text{obs}}) \propto p(m^*)L(d_{\text{obs}}|m^*)
\] (4.4)

Therefore, remaining questions to address in the research are: what is the definition of prior distribution \(p(m^*)\); how to estimate the prior distribution’s parameters; what is the likelihood function \(L(d_{\text{obs}}|m^*)\); how to sample from \(p(m^*|d_{\text{obs}})\)? These questions are answered in following subsections.

### 4.3.2 Prior distribution definition

The prior distribution \(p(m^*)\) is defined by the probability density function (pdf) of an ensemble of dimension reduced prior model realizations \(\{m^*(1), m^*(2), \ldots, m^*(L)\}\). They are dimension reduced results of an ensemble of prior model realizations \(\{m(1), m(2), \ldots, m(L)\}\) in the original space.

To sample \(\{m^*(1), m^*(2), \ldots, m^*(L)\}\), one needs to first sample \(\{m(1), m(2), \ldots, m(L)\}\) from the original space and then do dimension reduction. A model realization \(m(l)\) is the output of an interpolator \(f\) (e.g., the implicit potential field method) with input parameters \(\psi(l)\):

\[
m(l) = f(\psi(l))
\] (4.5)

In this research, with the implicit potential field method (implemented in Gempy (de la Varga et al., 2019)), uncertain parameters \(\psi\) consist of locations of contact points, locations of orientations, and orientations (dips and azimuths). These parameters are shown in Table 4.2. A Gaussian distribution is assumed for these parameters’ uncertainty. For each parameter listed in Table 4.2, each contact point or orientation’s mean is set to be the reference data inferred by experts, and the standard deviation (SD) is chosen based on geologists’ experience. The criterion for choosing the SD of prior parameters is that enough variations in \(\{m(1), m(2), \ldots, m(L)\}\) can be obtained but realizations that are geologically
unrealistic should be avoided. \( \psi \) is uncertain so Monte Carlo is used to sample input parameters \( \{ \psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(L)} \} \), and \( \{ \mathbf{m}^{(1)}, \mathbf{m}^{(2)}, \ldots, \mathbf{m}^{(L)} \} \) are outputs of the interpolator in Equation 4.5.

For this research, co-kriging parameters including range, sill and nugget effect are fixed for practical reasons (de la Varga et al., 2019). Regarding the range, since the research focuses on modeling smooth geological interfaces, it is chosen to be the maximum extent of the data in order to avoid artifacts created by the implicit potential field method, according to de la Varga et al. (2019). The absolute value of the sill is not very important because the potential field’s absolute value is meaningless (only relative values are needed to get isocontours to represent geological interfaces). A fairly small value is used for the nugget effect to give stability to the co-kriging system.

**Table 4.2:** Parameters for the interpolator

<table>
<thead>
<tr>
<th>Type of parameter</th>
<th>Parameter name (Unit)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>X of contact point (m)</td>
<td>reference</td>
<td>400</td>
</tr>
<tr>
<td>Location</td>
<td>Y of contact point (m)</td>
<td>reference</td>
<td>400</td>
</tr>
<tr>
<td>Location</td>
<td>Z of contact point (m)</td>
<td>reference</td>
<td>50</td>
</tr>
<tr>
<td>Location</td>
<td>X of orientation (m)</td>
<td>reference</td>
<td>200</td>
</tr>
<tr>
<td>Location</td>
<td>Y of orientation (m)</td>
<td>reference</td>
<td>200</td>
</tr>
<tr>
<td>Angle</td>
<td>dip (degree)</td>
<td>reference</td>
<td>10</td>
</tr>
<tr>
<td>Angle</td>
<td>azimuth (degree)</td>
<td>reference</td>
<td>6</td>
</tr>
</tbody>
</table>

Principal component analysis (PCA) is used to reduce the dimensionality of the ensemble of original high dimensional prior model realizations \( \{ \mathbf{m}^{(1)}, \mathbf{m}^{(2)}, \ldots, \mathbf{m}^{(L)} \} \). After PCA, each realization \( \mathbf{m}^{(l)} \) is a \( L \)-dimensional vector, so the dimension is reduced from \( N \) to \( L \) (\( N \gg L \)). There are two advantages of using PCA as the dimension reduction method in the proposed framework:

1. Realizations can be reconstructed using the linear combination of principal components (PCs) with their corresponding PC scores as weights. \( \mathbf{m}^* \) can be transformed to \( \mathbf{m} \) without loss of information if all PCs are used for reconstruction. To obtain a new realization, one only needs to sample PC scores from all dimensions and use them as weights to
linearly combine all PCs. This sampling is done in a much lower dimensional space, rather than the original high dimensional space.

2. PCA produces linear combinations of the original variables to generate the principal components (PCs). As discussed in the previous Chapter 3, linear combination of differentiable scalar fields results in differentiable scalar fields. Therefore, eigen-images (i.e., eigenvectors visualized as images) are essentially also potential fields (see Figure 4.7 for examples of eigen-images). When the reconstruction with PCs is performed, eigen-images are linearly combined to obtain new realizations, so new realizations are still potential fields (i.e., new realizations are not some random scalar fields that may be noisy or unrealistic to represent geological structures).

These two advantages of PCA provide the foundation for the proposed workflow. For implicit modeling, isocontour values that represent geological interfaces are also critical to obtain the final lithology model. Therefore, when doing PCA, iso-contour values that represent interfaces are concatenated with potential fields and used as features (Figure 4.4).

### 4.3.3 Prior distribution estimation

The prior distribution \( p(m^*) \) has been defined in the previous section. To apply the definition for sampling purposes, a pdf needs to be fitted to \( \{m^{*(1)}, m^{*(2)}, ..., m^{*(L)}\} \) (i.e., parameters of \( p(m^*) \) need to be estimated). Data used for fitting the pdf are PC scores obtained by PCA. Kernel methods such as the kernel density estimation (KDE) can be used to estimate the density of PC scores when the distribution of PC scores is continuous. The KDE idea has been successfully applied in geophysical applications (Scheidt et al., 2015; Phelps et al., 2018). However, a limitation of KDE is that it is only applicable for low dimensions (e.g., less than five), because the space tends to be empty and the estimated pdf becomes very close to zero, with more dimensions included. Therefore, one could select a few PCs and use KDE to estimate the pdf for those PCs, while assuming Gaussian distributions for other PCs. To determine which PCs should be more accurately estimated using KDE, an idea is to assess which PCs have the most impact on simulated gravity responses. The
CHAPTER 4. QUANTIFYING UNCERTAINTY

Figure 4.4: Features that are concatenated for PCA. (a) Potential field for BIF series; (b) potential field for Simple Mafic Series; (c) potential field for Early Granite Series; (d) Isocontours for representing lithology boundaries.
sensitivity is analyzed using a global sensitivity analysis method, called DGSA (Park et al., 2016). Eventually, the prior pdf is the multiplication of KDE estimated pdf and Gaussian distributions. In summary, our method for estimating the prior distribution in the dimension reduced space contains three steps:

1. Use kernel density estimation (KDE) for estimating the pdf of the most sensitive PCs found by DGSA.
2. Fit univariate Gaussian distribution independently to dimensions other than ones chosen in the first step.
3. The multiplication of all pdfs (both the KDE estimation and fitted Gaussian distributions) is used as the prior distribution.

4.3.4 Likelihood function definition

In geophysical applications, each model realization corresponds to a data realization (e.g., gravity), so prior model realizations \( \{m^{(1)}, m^{(2)}, \ldots, m^{(L)}\} \)’s corresponding prior data realizations are \( \{d^{(1)}, d^{(2)}, \ldots, d^{(L)}\} \). Data realizations are computed using a deterministic forward function \( g \):

\[
d = g(m)
\] (4.6)

This forward function \( g \) is Equation 4.1 for forward gravity data. The observed gravity data can be compared with data realizations using a mismatch function. The mismatch between the observed gravity data and a data realization \( d^{(l)} \) is defined as:

\[
J^{(l)} = \frac{1}{n} \|d^{(l)} - d_{obs}\|^2
\] (4.7)

Since this research is interested in uncertainty of geological structures, the gravity gradients \( d' \) and \( d'_{obs} \) are also incorporated in the mismatch function:

\[
J^{(l)} = \frac{1}{n} \|d' - d'_{obs}\|^2 + \frac{1}{n} \|d^{(l)} - d_{obs}\|^2
\] (4.8)
The likelihood function $\mathcal{L}$ is then defined as the exponential form of the mismatch function:

$$
\mathcal{L} = \exp \left( -\frac{(J^{(l)})^2}{2\sigma^2} \right)
$$  \hspace{1cm} (4.9)

where $\sigma$ is a parameter that should reflect the quality of data that may be affected by accuracy of devices and errors in data processing. The likelihood function increases when a data realization is more similar to the observed data. Different weights could be given to the gravity term and the gravity gradient term in the mismatch function to account for their influences. Equivalently, one could opt for using separate $\sigma$s in the likelihood function.

### 4.3.5 Sampling

The prior distribution and the likelihood function are both defined, so posterior realizations can be sampled. Two sampling schemes are used and compared: Monte Carlo and Markov chain Monte Carlo (McMC) to sample posterior realizations of implicit geological structures. The rejection sampling (RS) algorithm (Algorithm 1) (Von Neumann, 1963; Caers, 2011; Linde et al., 2015) and the Metropolis-Hastings (MH) algorithm (Algorithm 2) (Hastings, 1970; Caers, 2011; Linde et al., 2015) are methods that correspond to these two sampling schemes. For this research, both sampling methods use protocols described in Caers (2011) and Linde et al. (2015), but with details that fit into our context: instead of sampling in the original space, a new candidate is sampled from the low dimensional space; posterior realizations are obtained both from the original space and the low dimensional space.

Rejection sampling (Algorithm 1) is the only exact sampler. For each iteration, a new candidate is directly sampled from the low dimensional prior space in step 1. Then, the candidate is transformed to the original space for forward modelling. In step 4 of rejection sampling, the likelihood function and its ratio to a predefined supremum of likelihood functions $\mathcal{L}_{sup}$ are calculated. $\mathcal{L}_{sup}$ is set to be 1. Finally, a uniform random number is drawn, and the candidate is only accepted if the random number is smaller than the ratio
calculated in step 4.

For the MH algorithm (Algorithm 2), instead of sampling from the prior distribution, a new candidate is sampled from a proposal distribution. A normal distribution is used for the forward proposal, where the mean is the previous iteration’s result, and the variance $\sigma_{MH}$ is a pre-defined parameter. $\sigma_{MH}$ is chosen to make the average acceptance ratio of MH to be around 0.1 to 0.6 for a good algorithm efficiency (Rosenthal et al., 2011). When calculating the Hastings ratio (step 4) for the MH algorithm, the forward proposal and backward proposal are cancelled out because they are symmetrical (see the normal distribution in step 1). The posterior distribution in the Hastings ratio is a multiplication of prior distribution and the likelihood function in the dimension reduced space (Equation 4.4). The outputs of Metropolis-Hastings algorithm are correlated realizations and some of them are from the period before convergence (i.e., burn-in period (Hastings, 1970)). Samples from the burn-in period are discarded and an autocorrelation analysis is performed to obtain independent posterior realizations after MH converges.
Algorithm 1: Rejection sampling algorithm for sampling posterior realizations

**Input:** Number of posterior realizations ($L_{\text{post}}$)

**Output:** Posterior realizations $m_{\text{post}}^{s(1)}, m_{\text{post}}^{s(2)}, \ldots, m_{\text{post}}^{s(L_{\text{post}})}$ and $m_{\text{post}}^{(1)}, m_{\text{post}}^{(2)}, \ldots, m_{\text{post}}^{(L_{\text{post}})}$

$l = 1$, maximum likelihood function $L_{\text{sup}} = 1$

while $l \leq L_{\text{post}}$

1. Sample $m_{\text{cand}}^* \sim p(m^*)$; \(\triangleright\) Sample a candidate in the low dimensional space;
2. Reconstruct $m_{\text{cand}}$ using $m_{\text{cand}}^*$; \(\triangleright\) Transform the candidate to the original space;
3. $d_{\text{cand}} = g(m_{\text{cand}})$; \(\triangleright\) Forward modelling;
4. Calculate $\alpha = \frac{L_{\text{cand}}}{L_{\text{sup}}}$; \(\triangleright\) Calculate the ratio of likelihood functions;
5. $u \in [0, 1]$; \(\triangleright\) Draw a uniform random number;

if $u \leq \alpha$ then

- Accept $m_{\text{cand}}^*$ and $m_{\text{cand}}$;
  - $m_{\text{post}}^{s(l)} = m_{\text{cand}}^*$;
  - $m_{\text{post}}^{(l)} = m_{\text{cand}}$;
  - $l += 1$;

else

- reject $m_{\text{cand}}^*$ and $m_{\text{cand}}$;
Algorithm 2: Metropolis-Hastings algorithm for sampling posterior realizations

**Input:** Maximum number of iterations ($T$)

**Output:** Correlated realizations $m^*(1), m^*(2), \ldots, m^*(T)$ and $m^{(1)}, m^{(2)}, \ldots, m^{(T)}$

$t = 1$, initial sample $m^{*(0)}$;

while $t \leq T$ do

1. Sample $m_{cand}^* \sim N(m^{*(t-1)}, \sigma_{MH})$; $\triangleright$ Propose a candidate in low-dimensional space;

2. Reconstruct $m_{cand}$ using $m_{cand}^*$; $\triangleright$ Transform the candidate to the original space;

3. $d_{cand} = g(m_{cand})$; $\triangleright$ Forward modelling;

4. Calculate $\alpha = \min \left(1, \frac{p(m_{cand}^*|d_{obs})}{p(m^{*(t-1)}|d_{obs})} \right)$;

5. $u \in [0, 1]$; $\triangleright$ Draw a uniform random number;

if $u \leq \alpha$ then

   - Accept $m_{cand}^*$ and $m_{cand}$;
   - $m^{*(t)} = m_{cand}^*$;
   - $m^{(t)} = m_{cand}$;
   - $t++ = 1$;

else

   - reject $m_{cand}^*$ and $m_{cand}$;

4.4 Application to case study

4.4.1 Prior realizations and prior distribution

With uncertain parameters defined in Table 4.2, 300 sets of contact points and orientations are sampled by Monte Carlo. Then 300 prior model realizations and data realizations are obtained. Figure 4.5 shows a few realizations of uncertain parameters (contact points and orientations), corresponding prior lithology models and prior data realizations (forward gravity responses). One can notice variations in prior model realizations and how those
variations are reflected in the simulated gravity.

PCA is used to reduce the dimension of prior realizations of potential fields and isocountour values (see Equation 4.2) and prior data realizations. Figure 4.6 shows PCA scores and cumulative explained variances of PC components for both prior model realizations and prior data realizations. Figure 4.7 shows eigen-images of the first three PC components that correspond to three lithology series’ potential fields. Figure 4.8 shows several dimensions of PCA loadings for each series of potential fields. PCA loadings are the coefficients of the linear combination of the original variables from which the principal components (PCs) are constructed.

A sensitivity analysis using DGSA (Park et al., 2016) is conducted to determine which PCs have more impact on the forward gravity, so that a more accurate pdf needs to be estimated for those PCs (Figure 4.9). The result shows that simulated gravity is more sensitive to 9th, 12th, 8th, 24th and 15th PCs compared to other PCs. Based on the scree plot (Figure 4.6c), PCs from higher dimensions do not contribute a lot to the cumulative explained variance, so KDE is only used for estimating the probability density of 9th, 12th and 8th PC scores (Figure 4.10). Pdfs of remaining PCs (all PCs except 9th, 12th and 8th PCs) are approximated with independent univariate Gaussian distributions. The prior distribution is the multiplication of all PCs’ pdfs.

### 4.4.2 Sampling results

In order to evaluate if the estimated prior distribution is reasonable and if the MH sampler is an effective sampler for sampling from the estimated prior distribution, the Metropolis-Hastings sampling is tested for sampling from the prior space, without the use of the likelihood function defined in Equation 4.9 (or equivalently, $L = 1$). For prior sampling with MH, the average acceptance ratio is about 39%. Figure 4.11 (a) shows all samples from the MH sampler. It is shown that there is burn-in period before the MH sampler converged to the cloud of prior PC scores. The convergence is also shown by the first PC’s score at each
Figure 4.5: Prior construction results. (a) Prior parameters (contact points and orientations); (b) Prior model realizations (lithology models); (c) Prior data realizations (forward gravity responses).
Figure 4.6: PCA of prior realizations and prior gravity responses. (a) PC scores of prior realizations in 2D; (b) PC scores of prior gravity responses and the observed gravity in 2D; (c) Cumulative explained variance of each PC component of prior realizations; (d) Cumulative explained variance of each PC component of prior gravity responses.
Figure 4.7: Eigen-images of each potential field (Only first three PCs are shown). (a-c) Eigen-images of the BIF series’ potential fields; (d-f) Eigen-images of the Simple Mafic series’ potential fields; (g-i) Eigen-images of the Early Granite series’ potential fields.
Figure 4.8: PCA loadings of each variable. (a) PCA loadings of the 1st and 2nd PCs; (b) PCA loadings of the 3rd and 4th PCs; (c) PCA loadings of the 5th and 6th PCs.
Figure 4.9: Global sensitivity analysis for PCs. Gravity responses are more sensitive to PCs that are indicated by red colors (darker red indicates higher sensitivity).

Figure 4.10: Kernel density estimation of sensitive PC scores. (a) PC scores in 3D (only 9th, 12th and 8th PC scores are shown); (b) PC scores colored by estimated densities (only 9th, 12th and 8th PC scores are shown).
iteration in Figure 4.12 (a). Samples are highly correlated, as shown by the autocorrelation analysis of the first PC’s score at each iteration (Figure 4.12 (b)). Therefore, to obtain independent samples, samples are only retained every 1400 iterations based on the autocorrelation analysis. Samples from the burn-in period are also discarded because the sampler has not converged yet. Figures 4.11 (b-c) compare PC scores of prior model realizations and MH samples and Figure 4.11 (d) compares prior data realizations and simulated gravity anomalies using MH samples. Gravity profiles sampled along a section X = 72500m are also compared for the prior and MH sampled prior (Figure 4.11 (e)). The similarity in MH samples and prior realizations verifies that the MH sampler is working well within the framework.

The $\sigma$ in the likelihood function can be set based on the quality of data. If a device is very accurate, then $\sigma$ should be very close to 0 so that most realizations are rejected based on the likelihood function (i.e., only realizations that can generate gravity responses almost identical to the observed gravity will be accepted). By assuming that each device that collects gravity data has some degree of measurement error, $\sigma$ is set to be a heuristic value of $22.4 \mu gals$ (i.e., $\sigma^2 \approx 500 \mu gals^2$) for this research. The same value is used for MH and rejection sampling to make their results comparable.

For posterior sampling with MH, the average acceptance ratio is about 38%. Figure 4.12 (d) shows the autocorrelation length is about 800 for posterior samples of MH, so realizations every 800 iterations are retained for further analysis. Figure 4.13 shows the posterior sampling result with MH. Uncertainty reduction is noticeable when comparing posterior data realizations with prior data realizations (Figure 4.13 (b-e)). However, uncertainty reduction is not significant in model realizations when comparing the first two PCs of prior model realizations and posterior model realizations (Figure 4.13 (a)). In gravity profiles and gravity gradient profiles, more significant uncertainty reduction in a range from $Y = 6860000m$ to $Y = 6880000m$ is seen. It corresponds to the Southern part of lithology models where more variations are shown in prior realizations (Figure 4.5 (b)). Less uncertainty reduction is shown from $Y = 6880000m$ to $Y = 6920000m$, where the prior uncertainty is also relatively smaller. The assumption of data uncertainty in the observed data,
implemented by the specific choice of $\sigma = 22.4 \mu gal$ in the likelihood function, results in the uncertainty reduction not significant for $Y = 6880000m$ to $Y = 6920000m$.

For rejection sampling, the acceptance ratio is about 2.5%, which is much lower than Metropolis-Hastings algorithm. Figure 4.14 shows the posterior sampling result using rejection sampling. Similar to the MH result, again, one can see more uncertainty reduction in data realizations compared to the uncertainty reduction in model realizations shown by the first two PCs (Figure 4.14 (a)). Similarly, more significant uncertainty reduction is in a range from $Y = 6860000m$ to $Y = 6880000m$.

To investigate on the uncertainty reduction in the model realizations, Figure 4.15 shows a few posterior realizations obtained by MH and rejection sampling. Figure 4.16 compares the prior and posterior uncertainties of the boundary of lithology “Simple Mafic 1” using the probability map. Similar to what is seen Figures 4.13 and 4.14, uncertainty reduction is more noticeable in the South of the model where prior uncertainty is larger ($Y = 6860000m$ to $Y = 6880000m$). Table 4.3 compares standard deviations of volumes in prior lithology models and posterior lithology models. Uncertainty reduction results are similar for MH and rejection sampling.

Table 4.3: Reduction of standard deviation (SD) of lithology volumes (unit: number of voxels)

<table>
<thead>
<tr>
<th>Lithology</th>
<th>Prior (voxels)</th>
<th>MH posterior (voxels)</th>
<th>RS posterior (voxels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Mafic 2</td>
<td>364.45</td>
<td>252.61</td>
<td>267.27</td>
</tr>
<tr>
<td>Simple BIF</td>
<td>170.55</td>
<td>128.87</td>
<td>124.75</td>
</tr>
<tr>
<td>Simple Mafic 1</td>
<td>531.42</td>
<td>255.36</td>
<td>273.89</td>
</tr>
<tr>
<td>Early Granite</td>
<td>3808.61</td>
<td>2120.11</td>
<td>2332.74</td>
</tr>
</tbody>
</table>

4.4.3 Runtime discussion

About the runtime, it only takes around 1 second to sample and reconstruct a new lithology model with the use of PCs, but it takes about 9 seconds for generating one lithology model by doing interpolation with the implicit potential field method. The forward model itself
Figure 4.11: Metropolis-Hastings sampling of the prior PC scores. (a) Markov chain of PC scores in 3D (with a burn-in period); (b) Independent MH samples in first and second dimensions (burn-in samples are dropped); (c) Independent MH samples in 50th and 51st dimensions (burn-in samples are dropped); (d) Comparison of prior data realizations and MH samples of data realizations (burn-in samples are dropped); (e) Comparison of prior gravity response profiles (green) with MH samples (orange) along X = 72500m.
Figure 4.12: Autocorrelation analysis for MH sampling. (a) Correlated MH samples of 1st PC scores (prior) at each iteration that shows the convergence; (b) correlated MH samples of 1st PC scores (posterior) at each iteration; (c) autocorrelation analysis of MH samples (prior) that shows an autocorrelation length of about 1400; (d) autocorrelation analysis of MH samples (posterior) that shows an autocorrelation length of about 800.
Figure 4.13: Comparison of prior and MH posterior realizations with PC scores and profiles along X = 72500m. (a) PC scores of model realizations; (b) PC scores of data realizations; (c) Gravity profile; (d) Gradient in X; (e) Gradient in Y.
Figure 4.14: Comparison of prior and RS posterior realizations with PC scores and profiles along X = 72500m. (a) PC scores of model realizations; (b) PC scores of data realizations; (c) Gravity profile; (d) Gradient in X; (e) Gradient in Y.
Figure 4.15: Posterior lithology models. (a-c) Posterior realizations by Metropolis-Hastings algorithm; (d-f) Posterior realizations by rejection sampling.
Figure 4.16: Example of uncertainty reduction shown with probabilities of lithology “Simple Mafic 1”. (a) Prior probability map of “Simple Mafic 1”; (b) Posterior probability map of “Simple Mafic 1” (Metropolis-Hastings algorithm); (c) Posterior probability map of “Simple Mafic 1” (rejection sampling).

takes 2 seconds to run. This shows that with the dimension reduction technique, the runtime for sampling lithology models (either prior or posterior) can be significantly reduced because the interpolation is no longer needed once a prior distribution is defined in the dimension reduced space. In Wellmann et al. (2017), an McMC approach with the adaptive sampling algorithm (Haario et al., 2001) is used, and the interpolation algorithm (implicit potential field method) is needed for every iteration. Therefore, one can expect significant improvement in efficiency of sampling posterior realizations compared to Wellmann et al. (2017). Each iteration of rejection sampling and MH takes a total of 3-4 seconds. The experiment is conducted on a 3.5 GHz Quad-Core Intel Core i5 CPU.

4.5 Discussions

This research uses a synthetic gravity dataset for the purpose of methodology development. However, more care should be taken when using the actual observed data in real applications for two reasons: first, a forward model is only an approximation of the true physical process; second, gravity observations should be corrected based on the topographical effect (Ince et al., 2020). Uncertainty in the forward model and data processing errors may
result in different posterior distributions. Another possible limitation of the method is the use of a homogeneous density for each geologic unit (i.e., lithology block). This research focuses on the uncertainty of geological structures so the spatial variability of density is not considered. However, research shows that heterogeneous distribution of density within each geologic unit may contribute to the forward gravity anomaly (Phelps, 2016). In the future, geostatistical methods such as sequential Gaussian simulation (Goovaerts et al., 1997; Chilès and Delfiner, 2012) may be used to model spatial uncertainty of density within each geologic unit, so that uncertainties of geological structures and spatial properties can be jointly quantified (Yin et al., 2020).

4.6 Conclusions

In this chapter, a framework for quantifying uncertainty of implicit geological structures is presented. It is applied to a sandstone-greenstone belt in Western Australia. The main contribution is to take advantage of the implicit representation of geological structures with a continuous variable - the potential field, so that we can easily reduce the dimension of geological structures and thereby making uncertainty quantification a tractable problem. The prior distribution is estimated in the dimension reduced space. The use of principal component analysis (PCA) as the dimension reduction method makes sampling new realizations a fairly easy and efficient task. After dimension reduction, defining and fitting the prior distribution in the low dimensional space are possible. The runtime analysis shows it is more efficient to sample from the dimension reduced space compared to sampling in the original space, where each realization is the output of a time-consuming interpolator - the implicit potential field method.

Rejection sampling and Metropolis-Hastings algorithms are compared for sampling posterior realizations. Both methods work well in the study and are efficient enough to sample posterior realizations and reduce the uncertainty of the system. Rejection sampling is the only exact sampler, but the acceptance ratio is very low, so many samples are wasted. It may raise more difficulty with more complex geological structures and higher resolution
data (Linde et al., 2015). The acceptance ratio for MH is much higher. However, as shown in the autocorrelation analysis, MH samples are correlated, so the autocorrelation analysis is necessary for obtaining independent realizations.

By generating a suite of posterior lithology models, the proposed uncertainty quantification framework helps people to understand the subsurface variability and may improve mineral exploration with uncertain gravity data. It is shown that gravity data contribute mostly to reduce the uncertainty in regions where prior uncertainty is large. The framework may be extended and applied with different types of geophysical data such as magnetic data (Williams, 2009).
Chapter 5

Conclusions and Future Work

5.1 Conclusions

The thesis presents a suite of new methods that take advantage of implicit methods, including the level set method and the implicit potential field method, to tackle challenges in modeling large-scale geological structures and addressing structural uncertainties. Both real and synthetic cases are explored to demonstrate these methods.

To assess and visualize uncertainty of an implicit geological model, a new methodology using level sets by the means of stochastic motion is shown. The method uses the level set method to define a stochastic model for geological surfaces with complex geometries and topologies, subject to data and geological age relationship constraints. Two sampling strategies are used. Independent realizations sampled by Monte Carlo are retained for assessing the uncertainty of implicit geological surfaces. Results show that the uncertainty and output statistics such as the volume and surface roughness are related to the input parameters. A gradual deformation method (GDM) based McMC sampler is used for generating a continuously evolving movie that may enhance people’s perception of the uncertainty. GDM is proved to be McMC sampler with a few approximations. The runtime analysis shows the method’s efficiency for the illustrated cases.

In the context of numerous borehole data, a divide-and-conquer strategy is proposed to model large-scale implicit geological structures. The dense contact point
dataset is divided into multiple sets, and the scarce orientation dataset is expanded with co-simulation. Each subset of contact points and stochastic gradient information are then used as inputs for co-kriging to create implicit geological models. To make sure generated models are consistent with unused data and geological rules, the method contains an automatic and effective validation procedure to check and select models. Data and geologically validated models obtained with sub-datasets are linearly combined to obtain final models that fully match data. Proofs show that linear combination of potential fields preserves the conditioning to contact points and orientations. The framework has been applied to modeling a large-scale banded iron formation in Western Australia. Results show the success of generating a geologically realistic implicit 3D structural model that fit all data. It proves that robustness is brought to co-kriging by linearly combining results obtained using sub-datasets.

A Bayesian framework for quantifying uncertainty of implicit geological structures with geophysical data is presented. The framework is illustrated with a sandstone-greenstone belt case with gravity data in Western Australia. The main idea is to take advantage of the implicit representation of geological structures with a continuous variable - the potential field, so that the dimension of geological structures can be reduced and thereby making uncertainty quantification an efficient task. The prior distribution is estimated in the dimension reduced space. The use of principal component analysis (PCA) as the dimension reduction method makes sampling new realizations a fairly easy and efficient task. Algorithms based on protocols of Rejection sampling (RS) and Metropolis-Hastings (MH) are proposed and compared for sampling posterior realizations. The runtime analysis shows the efficiency of sampling from a dimension reduced space compared to sampling in the original space with the implicit potential field method. Results show that both sampling methods work well in the study and are efficient enough to sample posterior realizations and reduce the uncertainty of the system. The proposed uncertainty quantification framework may help people to understand the subsurface geological structures and guide mineral exploration with geophysical data that have uncertainties.
5.2 Summary of contributions

Here is a summary of main contributions of the thesis:

- Adapt “velocity extension” from the level set method to define a stochastic model for assessing and visualizing the uncertainty of an existing model.

- Animate a geological structure’s uncertainty using the level set method, to enhance people’s experience in perceiving uncertainty.

- Propose the divide-and-conquer strategy and use linear combination of implicit functions to model large-scale geological structures that match dense data and honor geological rules.

- Implement automatic validation of structural models based on both data and geological rules.

- Utilize geological structures’ implicit representation and perform dimension reduction for efficient uncertainty quantification.

- Design rejection sampling and Metropolis-Hastings algorithms for efficiently sampling posterior realizations of implicit geological structures.

5.3 Future research directions

5.3.1 Bayesian inference of model parameters

One of the questions left in Chapter 2 is the specification of model parameters for both the trend and residual functions. Without an inference procedure, the method presented in Chapter 2 is only an uncertainty assessment and visualization tool, but not an uncertainty quantification tool. Future work may focus on developing a Bayesian inference framework that quantifies uncertainty of model parameters based on data. An example of parameter inference for the trend and residual functions with a similar setup is shown in Yin et al.
In Yin et al. (2020), the uncertainty of reservoir thickness is modelled as the sum of trend and residual functions, and parameters of trend and residual functions are inferred within a Bayesian protocol.

### 5.3.2 Integration of faults uncertainty

For case studies in the thesis, faults are either deterministic or not incorporated. Integration of faults into uncertainty visualization and uncertainty quantification can be further investigated. First of all, the stochastic motion idea in Chapter 2 may be extended to visualize the uncertainty of faults, since level sets have been used to model faults, as indicated in the introduction of Chapter 2. For instance, Aydin and Caers (2017) shows one way to quantify uncertainty of implicit fault networks, using the marked Strauss point process. Though uncertainty quantification is mainly studied for lithologies in Chapter 4, but faults can readily be incorporated in the uncertainty quantification framework since faults are represented by additional potential fields and can be concatenated into Equation 4.2 for defining an implicit model. In that way, the uncertainty of geological layers such as stratigraphy or lithology and faults may be jointly quantified.

### 5.3.3 Incorporating other types of data

The thesis has discussed incorporating structural data and gravity data to model and quantify uncertainty of geological structures, but these methods may be extended to other types of data as well. A related research direction is regarding the incorporation of multiple sources of data into uncertainty quantification in addition to gravity data discussed in Chapter 4, such as magnetic data (Guillen et al., 2008) and flow data (Cherpeau et al., 2012). de la Varga and Wellmann (2016) presents a hierarchical Bayesian model to include structural data and “geological likelihood function” for geological modeling. Methods (dimension reduction, prior distribution estimation, sampling, etc.) presented in Chapter 4 can readily be integrated into a hierarchical Bayesian framework to incorporate other types of data for uncertainty quantification of geological structures.
Appendix A

Gradual deformation as a McMC sampler

In Hu (2000), gradual deformation was proposed to solve non-linear inverse problems with Gaussian prior distributions. The main idea is to write a Gaussian realization \( v(t) \) as a linear combination of two independent Gaussian realizations \( v_k \) and \( v_{k+1} \), i.e.,

\[
v(t) = \sin(t) \cdot v_k + \cos(t) \cdot v_{k+1}.
\]  

(A.1)

Then, a two-loop optimization on both \( k \) (outer loop) and \( t \) (inner loop) is used to optimize an objective function (difference between data and simulated data), constrained to the stated Gaussian prior. Multiple solutions are obtained by starting from different initial samples. As pointed out in Hansen et al. (2012), this form of sampling is not a Markov chain Monte Carlo since the method is dependent on the optimization algorithm and hence on the initial model. As a result, sampling from a posterior (or prior) is only approximate. Here, the GDM is used in a completely different setting. Instead of optimizing, GDM is used to generate a Markov chain; we generate a chain of realizations by changing (not optimizing) \( k \) discretely and \( t \) using a time step \( \Delta t \) in Equation A.1. In this appendix, a proof that GDM is a Markov chain Monte Carlo method regardless of the time steps \( \Delta t \) with reasonable approximations is provided. To do this, we need to show that gradual deformation has the
property of detailed balance (microscopic reversibility, see Green (1995), Medhi (1994) and Mosegaard and Sambridge (2002)). In other words, when

$$P(v(t) \rightarrow v(t + \Delta t)) = P(v(t + \Delta t) \rightarrow v(t)).$$  \hspace{1cm} (A.2)$$

If Equation A.2 does not hold, then in order to adjust the balance, we need to add an acceptance/rejection step with a probability given by the ratio

$$\alpha = \frac{P(v(t) \rightarrow v(t + \Delta t))}{P(v(t + \Delta t) \rightarrow v(t))}. \hspace{1cm} (A.3)$$

To simplify notation, call $t = x$, $t + \Delta t = y$, $v_k = v_1$ and $v_{k+1} = v_2$. Then

$$v(x) = \sin(x) \cdot v_1 + \cos(x) \cdot v_2 \hspace{1cm} (A.4)$$

Then we have:

$$v(x + y) = \sin(x + y) \cdot v_1 + \cos(x + y) \cdot v_2$$

$$= \sin(x) \cdot (\cos(y) \cdot v_1 - \sin(y) \cdot v_2)$$

$$+ \cos(x) \cdot (\sin(y) \cdot v_1 + \cos(y) \cdot v_2)$$

$$= \sin(x) \cdot v'_1 + \cos(x) \cdot v'_2 \hspace{1cm} (A.5)$$

Note that $v'_1$ and $v'_2$ are independent random variables. Similarly, going back from $x + y = z$ to $x$ yields:

$$v(z) = \sin(z) \cdot v_1 + \cos(z) \cdot v_2 \hspace{1cm} (A.6)$$

Similarly to Equation A.5, we have:

$$v(x) = \sin(z - y) \cdot v_1 + \cos(z - y) \cdot v_2$$

$$= \sin(z) \cdot (\cos(y) \cdot v_1 + \sin(y) \cdot v_2)$$

$$+ \cos(z) \cdot (- \sin(y) \cdot v_1 + \cos(y) \cdot v_2)$$

$$= \sin(z) \cdot v''_1 + \cos(z) \cdot v''_2 \hspace{1cm} (A.7)$$
Since distributions of all variables are Gaussian, detailed balance entails that both the conditional mean and variance are equal, i.e.,

\[
E[v(x + y)|v(x)] = E[v(x)|v(x + y)] \tag{A.8}
\]

and

\[
Var[v(x + y)|v(x)] = Var[v(x)|v(x + y)] \tag{A.9}
\]

In terms of conditional expectation this then entails that:

\[
E[v'_1|v_1] = E[v''_1|v_1] \tag{A.10}
\]

\[
E[v'_2|v_2] = E[v''_2|v_2]
\]

A simple calculation shows that

\[
E[v'_1|v_1] = E[v''_1|v_1] = \cos(y)v_1
\]

\[
E[v'_2|v_2] = E[v''_2|v_2] = \cos(y)v_2 \tag{A.11}
\]

Similarly, for the variance we get:

\[
Var[v'_1|v_1] = Var[v''_1|v_1] = \sin^2(y)
\]

\[
Var[v'_2|v_2] = Var[v''_2|v_2] = \sin^2(y) \tag{A.12}
\]

For small $\Delta t$, it is easy to show that using a Taylor expansion:

\[
v(t + \Delta t) = v(t) + \Delta t \cdot (\cos(x)v_1 - \sin(x)v_2)
\]

\[
= v(t) + \Delta t \cdot v \tag{A.13}
\]

In this case, detailed balance is trivial to show because of the addition of a random component with mean zero and standard deviation $\Delta t$. Caers (2007) showed with examples that gradual deformation leads to divergence when the size of the random vector is small relative to the number of perturbations made. This can now be understood as follows. Consider
$v$ to be a random vector with covariance matrix $C$. A simple way to generate samples of $v$ is using Cholesky decomposition, meaning:

$$v = m + L^T v_0$$  \hspace{1cm} (A.14)

with $v_0$ is vector of independent Gaussian deviates and $C = L^T L$. Because the vector $v_0$ is of finite size, so $v_0^T v_0$ is not necessarily exactly equal to zero (spurious correlation). This also means that there is no detailed balance since for another sample $v'$ we have $v_0^T v'_0$, hence

$$P(v \rightarrow v') \sim \exp(-(v' - m)^T C^{-1} (v' - m)^T) = \exp(-v_0^T v'_0)$$  \hspace{1cm} (A.15)$$

Likewise

$$P(v' \rightarrow v) \sim \exp(-(v - m)^T C^{-1} (v - m)^T) = \exp(-v_0^T v_0)$$  \hspace{1cm} (A.16)

Hence GDM requires an acceptance/rejection probability of

$$\alpha = \exp(-v_0^T v'_0 + v_0^T v_0)$$  \hspace{1cm} (A.17)

This probability is however close to 1 for high-dimensional problems, or problems where the sample size is orders of magnitude less than the dimension of the problem, which is the case for large spatial problems. The implications here for stochastic motion is that McMC is possible using GDM, and that the visualization, i.e. the “movie” of dependent realizations reflects the actual uncertainty sampled using Monte Carlo by independent realizations.
Appendix B

Co-kriging system of the potential field method

This appendix reviews a few details about the co-kriging system used for the implicit potential field method based on relevant literature (Lajaunie et al., 1997; Aug, 2004; Chilès et al., 2004; Calcagno et al., 2008; Chilès and Delfiner, 2012; de la Varga et al., 2019).

Using $Z(\vec{x})$ (or simply $Z$) to denote the scalar field, there is an algebraic relationship between the scalar field $Z(\vec{x})$ and its gradients:

$$\frac{\partial Z}{\partial u}(\vec{x}) = \lim_{\rho \to 0} \frac{Z(\vec{x} + \rho u) - Z(\vec{x})}{\rho}$$

The algebraic dependency between $Z$ and $\frac{\partial Z}{\partial u}$ must be considered in the computation of the drift function. It also determines the relationship of covariance functions for $Z$ and $\frac{\partial Z}{\partial u}$.

The drift of $Z$ is defined as a polynomial function:

$$\mu(\vec{x}) = \sum_{l=0}^{L} b_l f^l(\vec{x})$$

where $l$ is the grade of the polynomials, $f^l(\vec{x})$ are basis functions, and $b_l$ are weights.

Then the co-kriging system (without faults) has the form of:
APPENDIX B. CO-KRIGING SYSTEM OF THE POTENTIAL FIELD METHOD

\[
\begin{bmatrix}
C_{\partial Z/\partial u, \partial Z/\partial v} & C_{\partial Z/\partial u, Z} & F_{\partial Z/\partial u} \\
C_{Z, \partial Z/\partial v} & C_{Z, Z} & F_Z \\
F'_{\partial Z/\partial u} & F'_Z & 0
\end{bmatrix}
\begin{bmatrix}
\lambda_{\partial Z/\partial u, \partial Z/\partial v} & \lambda_{\partial Z/\partial u, Z} \\
\lambda_{Z, \partial Z/\partial v} & \lambda_{Z, Z} \\
\mu_{\partial u} & \mu_{u}
\end{bmatrix}
\begin{bmatrix}
C_{\partial Z/\partial u, \partial Z/\partial v} & C_{\partial Z/\partial u, Z} & F_{\partial Z/\partial u} \\
C_{Z, \partial Z/\partial v} & C_{Z, Z} & F_Z \\
F'_{\partial Z/\partial u} & F'_Z & 0
\end{bmatrix}^{-1}
\begin{bmatrix}
\partial Z \\
0 \\
0
\end{bmatrix}
\]

where \(C_{Z, Z}\) is the covariance matrix of potential increments; \(C_{\partial Z/\partial u}\) is the covariance matrix of gradients; \(C_{\partial Z/\partial u, Z}\) is the cross-covariance function; \(F_Z\) and \(F_{\partial Z/\partial u}\) are drift functions of potential increments and gradients, respectively; weights corresponding to each term are denoted as \(\lambda\); constants of drift functions are denoted as \(\mu\); terms on the right side correspond to covariances (e.g., \(c_{Z, Z}\)) and drift functions (e.g., \(f_{10}\)) at the spatial location of interest.

For efficient computation of the potential at any point \(\bar{x}\), the co-kriging system is preferred to be solved in its dual form:

\[
Z^* = \begin{bmatrix}
a'_{\partial Z/\partial u} & b'_{Z, Z} & c'
\end{bmatrix}
\begin{bmatrix}
c_{\partial Z/\partial u, \partial Z/\partial v} & c_{\partial Z/\partial u, Z} \\
c_{Z, \partial Z/\partial v} & c_{Z, Z} \\
f_{10} & f_{20}
\end{bmatrix}
\begin{bmatrix}
a_{\partial Z/\partial u} \\
b_{Z, Z} \\
c
\end{bmatrix}
\]

in which

\[
\begin{bmatrix}
a_{\partial Z/\partial u} \\
b_{Z, Z} \\
c
\end{bmatrix}
= \begin{bmatrix}
\partial Z \\
0 \\
0
\end{bmatrix}
\begin{bmatrix}
C_{\partial Z/\partial u, \partial Z/\partial v} & C_{\partial Z/\partial u, Z} & F_{\partial Z/\partial u} \\
C_{Z, \partial Z/\partial v} & C_{Z, Z} & F_Z \\
F'_{\partial Z/\partial u} & F'_Z & 0
\end{bmatrix}^{-1}
\begin{bmatrix}
C_{\partial Z/\partial u, \partial Z/\partial v} & C_{\partial Z/\partial u, Z} & F_{\partial Z/\partial u} \\
C_{Z, \partial Z/\partial v} & C_{Z, Z} & F_Z \\
F'_{\partial Z/\partial u} & F'_Z & 0
\end{bmatrix}
\begin{bmatrix}
c_{\partial Z/\partial u, \partial Z/\partial v} & c_{\partial Z/\partial u, Z} \\
c_{Z, \partial Z/\partial v} & c_{Z, Z} \\
f_{10} & f_{20}
\end{bmatrix}
\]

Gempy uses this dual form for implementation of the implicit potential field method (de la Varga et al., 2019).

Because of the algebraic dependency between \(Z\) and \(\partial Z/\partial u\), the covariance matrix for gradients \(\partial Z/\partial u\) can be derived from the covariance matrix of \(Z\). That means, the covariance function has to be twice differentiable at the origin. The cubic function is the most
commonly used in the potential field method to make sure the covariances are sufficiently differentiable. The cubic covariance function is:

\[
C(r) = \begin{cases} 
C_0 \left(1 - 7 \left(\frac{r}{a}\right)^2 + \frac{35}{4} \left(\frac{r}{a}\right)^3 - \frac{7}{2} \left(\frac{r}{a}\right)^5 + \frac{3}{4} \left(\frac{r}{a}\right)^7\right), & \text{for } 0 \leq r \leq a \\
0, & \text{for } r \geq a.
\end{cases}
\]

where \(r\) is the distance, \(a\) is the range and \(C_0\) is the variance (i.e. sill of the variogram).

Faults are incorporated into the co-kriging system with extra drift-related terms, including \(F_1\partial Z/\partial u\), \(F_1Z\), \(\mu\partial f_1\) and \(\mu f_1\):

\[
\begin{bmatrix}
\partial C Z/\partial u, \partial Z/\partial u & C Z/\partial u, Z & F Z & F 1 Z \\
C Z, \partial Z/\partial u & C Z, Z & F Z & F 1 Z \\
F Z/\partial u & F Z & 0 & 0 \\
F 1 Z/\partial u & F 1 Z & 0 & 0
\end{bmatrix}
=\begin{bmatrix}
\lambda \partial Z/\partial u, \partial Z/\partial v & \lambda Z/\partial u, Z \\
\lambda Z, \partial Z/\partial v & \lambda Z, Z \\
\mu \partial f & \mu f \\
\mu \partial f_1 & \mu f_1
\end{bmatrix}

\begin{bmatrix}
\partial C Z/\partial u, \partial Z/\partial v & \partial C Z/\partial u, Z \\
C Z, \partial Z/\partial v & C Z, Z \\
f_{10} & f_{20} \\
f_{10} & f_{20}
\end{bmatrix}
\]
Appendix C

Data fitting preservation with linear combination

In the appendix, we prove that the conditioning to contact points and orientations is preserved with linear combination of two potential fields. Let \( f(\vec{x}) \) and \( g(\vec{x}) \) denote two potential fields, \( h(\vec{x}) \) the linearly combined potential field and \( w(\vec{x}) \) the spatially varying weight.

**Statement 1.** The linear combination of two potential fields maintains fitting contact points that both potential fields already fit.

**Proof 1.** At contact points locations \( \vec{x}_c \), the linear combination of two potential fields \( f(\vec{x}_c) \) and \( g(\vec{x}_c) \) is:

\[
h(\vec{x}_c) = w(\vec{x}_c)f(\vec{x}_c) + (1 - w(\vec{x}_c))g(\vec{x}_c)
\]

Let \( \tilde{f} \) and \( \tilde{g} \) denote iso-values that fit \( \vec{x}_c \) for \( f(\vec{x}_c) \) and \( g(\vec{x}_c) \), respectively. Let \( \tilde{w} \) denote the weight at location \( \vec{x}_c \). Then we have:

\[
\tilde{f} = f(\vec{x}_c)
\]

\[
\tilde{g} = g(\vec{x}_c)
\]

\[
\tilde{w} = w(\vec{x}_c)
\]
The statement “\( \tilde{h} \) fits contact points”, entails \( \tilde{h} = h(\tilde{x}_c) \): \( \tilde{h} \) is the iso-value in \( h(\tilde{x}) \) that represents the boundary. We know that

\[
h(\tilde{x}_c) = w(\tilde{x}_c)f(\tilde{x}_c) + (1 - w(\tilde{x}_c))g(\tilde{x}_c)
\]

\[
= \tilde{w}\tilde{f} + (1 - \tilde{w})\tilde{g}
\]

In other words, the iso-value used to represent the boundary in \( h(\tilde{x}) \) should be the linear combination of iso-values in \( f(\tilde{x}) \) and \( g(\tilde{x}) \), with the same weight used for combining \( f(\tilde{x}) \) and \( g(\tilde{x}) \). To make sure new boundaries fit contact points, we not only need to do linear combination for two potential fields, but we also need to do a linear combination for the thresholding values and use it to create the categorical model.

**Statement 2.** For any contact point, if only one potential field fits it, then the linear combination of two potential fields maintains fitting that contact point if the fitted potential field has a weight of 1 and the unfitted potential field has a weight of 0 at the contact point location.

**Proof 2.** Let’s assume the potential field that fits \( \tilde{x}_c \) is \( f(\tilde{x}_c) \). As stated, \( w(\tilde{x}_c) = 1 \), so

\[
h(\tilde{x}_c) = f(\tilde{x}_c)
\]

and

\[
\tilde{h} = \tilde{f}
\]

Therefore, \( h(\tilde{x}_c) \) also fits \( \tilde{x}_c \) as \( f(\tilde{x}_c) \) does. This means that as long as we assign a weight of 1 to the fitted potential field at the contact point location, and a weight of 0 to the unfitted potential field, then our linearly combined potential field will fit to the contact point.

**Statement 3.** The linear combination of two potential fields maintains fitting orientations that both potential fields already fit.

**Proof 3.** At first, we know the existing linearity of gradient theorem, which states that, for two scalar fields \( f(\bar{x}) \) and \( g(\bar{x}) \) at any location \( \bar{x}_o \), we have
\[
\n\nabla (\alpha f + \beta g)(\vec{x}_o) = \alpha \nabla f(\vec{x}_o) + \beta \nabla g(\vec{x}_o)
\]

It’s easy to see that if we replace \( \alpha \) with \( w(\vec{x}_o) \), and \( \beta \) with \( 1 - w(\vec{x}_o) \), the linearity of gradient still holds.

Then, for orientation observation locations \( \vec{x}_o \) where \( f(\vec{x}_o) \) and \( g(\vec{x}_o) \) have the same gradients, i.e., \( \nabla f(\vec{x}_o) = \nabla g(\vec{x}_o) \), we can obtain the following using the linearity of gradient:

\[
\nabla h(\vec{x}_o)
\]

\[
= \nabla (w f + (1 - w)g)(\vec{x}_o)
\]

\[
= w(\vec{x}_o) \nabla f(\vec{x}_o) + (1 - w(\vec{x}_o)) \nabla g(\vec{x}_o)
\]

\[
= w(\vec{x}_o) \nabla f(\vec{x}_o) + (1 - w(\vec{x}_o)) \nabla f(\vec{x}_o)
\]

\[
= \nabla f(\vec{x}_o)
\]

In other words, the gradient remains the same after linear combination.

Therefore, as long as we use simulated orientations with the same observed “hard data” to generate \( f(\vec{x}) \) and \( g(\vec{x}) \) (i.e., \( \nabla f(\vec{x}_o) = \nabla g(\vec{x}_o) \) for all orientation observation locations), all orientation observations will always be matched when doing linear combinations.
Bibliography


