RWTH Aachen University

Master of Science in Applied Geosciences

Master Thesis

Surface-based Geolgical Modeling Using an Integrated Explicit-Implicit Method

December 27, 2022

Author:
Mosaku Adeniyi, B.Sc.

Examiners:
Prof. Florian Wellmann, Ph.D.
Dr. Jan Niederau, Ph.D.

Supervisor:
Mohammad Moulaeifard, M.Sc.

submitted to the
Institute of Computational Geoscience and Reservoir Engineering

Faculty of Georesources and Material Engineering · RWTH Aachen University

Printed in Aachen, Germany

Dedication

This research work is dedicated to my late parents, Mr. and Mrs. Mosaku, of blessed and loving memory, for their immense support from the very beginning.

December 27, 2022

Abstract

Surface-based geomodeling is a very critical aspect to understanding the heterogeneity of the subsurface structure. The effective understanding of these subsurface structures is important in many aspects of geological reservoir (geothermal, petroleum, or hydrology) modeling, such as simulation, for example, in heat flow, or optimization in the aspect of hydrocarbon production. However, constructing these heterogeneous geological surfaces is a strenuous task mainly due to the uncertainties involved, perhaps as a result of collected data or the inflexibility of model manipulation. In this research work, probabilistic machine learning through the integration of implicit and explicit modeling techniques is presented. The implicit method is implemented by a Python-based open-source package called *GemPy* which generates geological realizations in the form of a scalar field. These scalar fields are then transformed into a layer surface using an explicit method. The explicit method is a parametric surface created using a Cubic Basis Spline. These surfaces are optimized to have the highest accuracy possible when compared to the scalar field without using every point generated from the scalar field, thus having fewer control points as much as possible to allow wider spatial manipulations. The model generated is further subjected to probabilistic machine learning in order to estimate uncertainties and possibly reduce them to the barest minimum using the Bayesian inference approach. This approach involves the inclusion of additional data, such as geophysical data, expressed as a likelihood function to constrain the model, thereby reducing the uncertainties involved in the model. The explicit approach provides surfaces which allow for easy manipulation and the generation of several realistic reservoir models with the help of the estimated control points. Volumes between these surfaces through surface terminations are created. An optimization algorithm such as the Adam optimizer is used to optimize the simulation. The three theoretical model used are evaluated based on their prior and posterior distribution. In addition, attributes such as probability and information entropy as well as the error estimate are used for the evaluation of the maximum a posteriori estimation.

Acknowledgements

I would like to express my gratitude by thanking my supervisor, Mohammad Moulaeifard, for his support during my research work. I also want to sincerely thank Dr. Jan, particularly for his direction and continuous productive input during my research work.

Professor Florian Wellmann has my heartfelt gratitude for taking the time to closely monitor the development of my research work, providing essential insight into improving my research work and scientific research capabilities despite his busy schedule. Working as a research assistant at the Institute of Computational Geosciences and Reservoir Engineering has been a great experience for me, and I want to again express my gratitude to Professor Wellmann for giving me the opportunity to be part of the research group at the institute and participate in research work that is both inspiring and exciting at the same time.

Lastly, but certainly not the least, I would like to express my sincere gratitude to my brothers and sisters for their unflinching support since our parents passed away to pursue my desires and aspirations. Thank you very much for your support.

RWTH Aachen University
December 27, 2022

Mosaku Adeniyi, B.Sc.

Table of Contents

List of Figures

Acronyms and symbols

RWM Random Walk Metropolis

MAP Maximum a posteriori

ML Machine Learning

MLE Maximum Likelihood Estimation

MCMC Markov Chain Monte Carlo

PDF Probability Density Function

MC Monte Carlo

HMC Hamiltonian Monte Carlo

SGD Stochastic Gradient Descent

TDMA Tri-Diagonal Matrix Algorithm

TFP Tensorflow Probability

θ	True parameter
$\hat{\theta}$	Estimates
$\mathbb{E}$	Expected value
ϕ	Porosity
U	X direction knot vector
V	Y direction knot vector
p	degree
t	knot parameter
d	data point
y	observed data
k	permeability
H_t	Information entropy
m	multiplicity
C	Continuity
P	Control points
x, y, z	Coordinate axes, z is vertical (depth)

Chapter 1

Introduction

Many geophysical and geological investigations are interested in the Earth's subsurface (Wellmann & Caumon 2018) to better understand the materials in the subsurface and its complexities. Some of the most valuable reasons are for mineral exploitation, energy production, and environmental engineering. Therefore, it is important to build a subsurface model that is realistic and close to exactly what we have in the subsurface. The approach used in building these subsurface models, created to encompass the heterogeneity of the subsurface, is an approach known as geomodeling. The geomodeling approach could be surface-based where reservoir modeling and simulation are carried out in such a way that reservoir heterogeneity, which impacts fluid flow, is represented explicitly by bounding surfaces (Caumon et al. 2009). Another method includes the use of geostatistical procedures such as kriging, an estimation technique where weighted averages are created for samples based on their spatial relationship to estimate an initially unknown point or entity (Matheron 1963). Other conventional methods include the corner-point or Cartesian grid approach, where each grid is discretised to contain a uniform value (Liu et al. 2020). The surface-based approach, however, is said to preserve detailed, complex input geometries throughout the modeling workflow, from geological representation to fluid flow simulation (Jacquemyn et al. 2019*a*).

With all these approaches in mind, there are certain issues that stand out when it comes to geomodeling. They are uncertainties and flexibilities. Every aspect of a process that involves some form of decision making has, no matter how little, some form of uncertainties. In geomodeling, this is extremely important as it involves modeling a process, geobody or scenario that we have no complete or certain information about. As George (2007) puts it:

"All models are wrong, but some are useful."

Therefore, we will constantly be faced with uncertainties when dealing with situations that involves decision making. These decisions can be very complex and costly if not properly managed. Certainly, uncertainty estimation is a critical aspect when making decision with regards to geomodeling in order to be able to make plausible and rational decisions. However, due to the scarcity of data and the resulting need for geological interpretation to build these geomodels, uncertainty is prevalent and has typically gone unmeasured (Schaaf et al. 2021).

According to Wellmann & Caumon (2018), classification of geomodeling uncertainties can be directed along the aspect of data gathering and processes of model building. When we go through the classical, conventional workflow of geological model construction, we actually go through quite a number of steps. Typically, at the beginning, we may look at a region and decide first of all what kind of geometric features we are going to expect, what kind of depositional setting we are looking at, what kind of fault or fault sequence we are expecting, and so on. All of these considerations lead to a decision on a certain type of modeling method that might be suitable to use.

Once we decide on some sort of modeling method to use, we then need to define the structure of the underlying mathematical model required. And finally, we need to look at the type of input data that is required for the construction of the models. All these decisions, including the data acquired, may have some level of uncertainty attached to them, and this leads to potential uncertainties in our final constructed geological model.

One of the approaches to tackle this problem, if we could, would be to acquire the data a sufficient number of times and get the most common, which is the mode, or the average value of all the possible realizations. This approach is what is known as the "frequentist" approach. However, in reality of course, we cannot simply go map and model the same target location for a large number of times, sufficient enough to make a decision from. That is even if we know the number of times that would take. This is where stochastic modeling methods come in. Using this method, we could assign some uncertainties to the parameters required to build the model, such as upper and lower limits or the mean, and then sample from the distribution created by this range of values. This way, we have variability in the models created, thereby forming the possible realizations we could obtain in reality. This is what is usually done when considering Monte Carlo MC simulation or a forward stochastic modeling approach.

It is clear that these limits or mean values will in turn inform the type of realizations we will get. Therefore, our model realizations hugely depend on our initial model, which could also be very biased.

Another question is: if we mimic the uncertainties in reality with our stochastic approaches, do we get close to what we would observe in nature? In other words, does it actually reduce the uncertainties?

This question is important because, in terms of mathematics, we could have close to an infinite number of realizations that could be completely different from any possible geological structure we want to model. This is because there are certain geological laws that must be obeyed.

Therefore, important judgments must be made at each stage of this modeling process, partially based on limited information and general geological knowledge as well as measurements and observations, all of which are susceptible to uncertainty (Wellmann & Caumon 2018).

One way we hope to tackle this problem is through this project. This involves the concept of including additional knowledge back into this geological modeling process. That is our stochastic modeling approach to obtaining models that are, at least on average, more realistic. This approach is also known as "probabilistic machine learning".

On the aspect of flexibility, modeling techniques that involve pre-defined corner point grids tend to be less flexible when compared to surface-based modeling techniques. The key disad-

vantages of this approach include the difficulty involved in changing the fundamental geological principles of a comprehensive model, making it difficult to investigate a range of reservoir prototypes or the geometry of the geological intricacies that can be captured (Zhang et al. 2018). However, for surface-based geological models, due to its flexibility, it can easily be updated accordingly as new data is introduced.

Due to these inherent uncertainties and needed flexibility, the integration of an implicit and explicit approach becomes important and necessary. The implicit approach was implemented using a Python-based open-source package called Gempy (de la Varga et al. 2018). Gempy makes use of a probabilistic geomodeling method constrained by a stochastic simulation which has geological information in the form of topological information embedded in it, thus serving as a constraint to the resulting probabilistic ensembles (Schaaf et al. 2021). The explicit modeling approach makes use of the cubic spline algorithm through the process of surface resampling. The scalar field in the implicit modeling approach can be converted to surface cartessian points which is then converted into a geological surface-based model.

This combination allows for the use of structures from both modeling approaches to create many realizations though the use of its control points. Therefore provide the possibilities of uncertainty estimation of our realizations.

The integration of these two approaches, however, poses some obstacles. Essentially, what we are doing is to approximate the scalar field we got from the implicit model approach. Therefore, we will need to be able to verify how good our final model is when compared to the scalar field from the implicit model. In this research work, an approach called the least squares method is used. An approximated surface is generated through the minimization of an error term, which is the distance between the cubic b-spline surface and the surface point extracted from the scalar field. In addition to generating an approximated surface, the control points are re-estimated based on the surface generated.

These surfaces are then subjected to uncertainty estimation through the use of the control points and additional data such as data gotten from the geophysical measurement of the subsurface in the area of target formation. And this is what we achieved in this project.

In general, in this project, we developed a framework around probabilistic machine learning where we can combine additional information to learn about the subsurface for improved decision-making with surface-based geological models through the integration of an explicit-implicit approach.

It is important to note that technical terms will be written in italics only when they are first mentioned in this research work to show that they have an inherent underlying definition. Italics will always be used for software packages and programming languages. This research work will present results using theoretical models. For the models, an arbitrary, dimensionless property is assumed for testing purposes.

Chapter 2

Methodology

To begin, we will start from the fundamentals and basics of building the parametric surface using an explicit method. The approach utilized in this research work is a cubic basis spline. We will begin from understanding the basis function, B-spline curve (spline) and then move to surfaces, which is also known as spline of splines. Then we create theoretic geologic model using an open source package *GemPy* (de la Varga et al. 2018), an open-source tool written in the programming environment *Python*. The theoretical model generated is then re-sampled using our explicit modeling approach (cubic bspline) to generate not only the surface-based geological model but also the control points that determine the geometry of surface-based model. Furthermore, this approach is integrated with an implicit method which forms the framework for the presented geological models to which the explicit method is built to optimize. We present an uncertainty estimation known as Bayesian inference. Bayesian data analysis is introduced to estimate the uncertainty involved in the geological models constrained with an additional geophysical data. Focus is laid on Bayesian inference, estimation of uncertain values, loss functions and uncertainty evaluation. We apply these methods in the context of structural geological modeling specifically in a probabilistic framework.

Regarding numerical implementation, *GemPy*, our cubic basis spline algorithm representing the explicit modeling approach and Tensorflow Python package are presented as central tools used in a Python environment. We outline our methods to evaluate consequent modeling results. It is important to note that different authors make use of different technical terms, therefore we try to provide an overview and stick to the one most common in the cited literature.

2-1 B-Spline Curves

2-1-1 Basis of B-Spline Curves

Basis functions of B-splines are defined in a knot coordinate system u_i and by their degree. It was established by Carl deboor in the 1970s along side M.G. Cox. Through the application of

Leibniz's theorem, Carl deboor was able to derive the B-spline basis functions as a recursive relation, known today as Cox-de Boor formula.

The B-splines are defined by (Rogers 2001):

$$N_{i,0}(t) = \begin{cases} 1, \text{ if } x_i \leq t < x_{i+1} \\ \\ 0, \text{ otherwise} \end{cases} \tag{2-1}$$

and;

$$N_{i,p}(t) = \left(\frac{t - x_i}{x_{i+p} - x_i}\right) N_{i,p-1}(x) + \left(\frac{x_{i+p+1} - t}{x_{i+p+1} - x_{i+1}}\right) N_{i+1,p-1}(t) \tag{2-2}$$

for the i_{th} normalized b-spline basis function of degree p (order p + 1)

Where:

$\mathrm{N}_{i,p}(t) = \mathrm{i}_{th}$ b-spline blending function of degree p and order (p + 1).

u_i = non-decreasing set of real numbers also called the knot sequence.

t = parameter variable.

The variable u_i represents the active area of the real number line that defines the B-spline basis. Since the basis functions are based on knot differences, the shape of the basis functions is only dependent on the knot spacing and not specific knot values.

Using the recurrence equation 2-2 the non-uniform B-splines from degree 1 up to degree 3 are given by:

Linear B-spline:

$$N_{i,1}(t) = \begin{cases} \frac{t - x_i}{x_{i+1} - x_i} & \text{if } x_i \leq t < x_{i+1} \\ \\ \frac{x_{i+2} - t}{x_{i+2} - x_{i+1}} & \text{if } x_{i+1} \leq t < x_{i+2} \\ \\ 0, & \text{otherwise} \end{cases} \tag{2-3}$$

The figure 2-1 below describes the linear spline basis function of degree $p = 1$, order $p + 1$

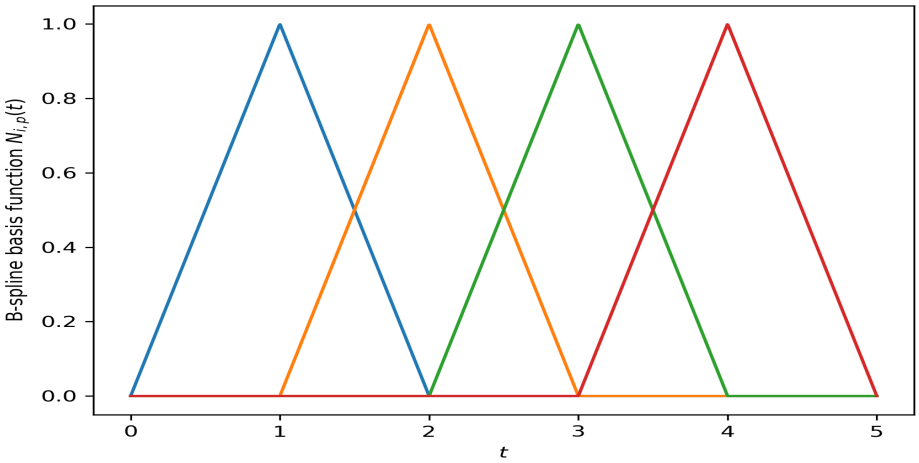

Figure 2-1: Linear Spline Basis

Quadratic B-spline:

$$N_{i,2}(t) = \begin{cases} \frac{(t-x_i)^2}{(x_{i+2}-x_i)(x_{i+1}-x_i)} & \text{if } x_i \leq t < x_{i+1} \\ \frac{(x_{i+2}-t)(x_{i+2}-t)}{x_{i+2}-x_{i+1}} + \frac{(x_{i+3}-t)(t-x_{i+1})}{(x_{i+3}-x_{i+1})(x_{i+2}-x_{i+1})} & \text{if } x_{i+1} \leq t < x_{i+2} \\ \frac{(x_{i+3}-t)^2}{(x_{i+3}-x_{i+1})(x_{i+3}-x_{i+2})} & \text{if } x_{i+2} \leq t < x_{i+3} \\ 0 & \text{otherwise} \end{cases} \tag{2-4}$$

The figure 2-2 below describes the quadratic spline basis function of degree $p = 2$, order $p+1$

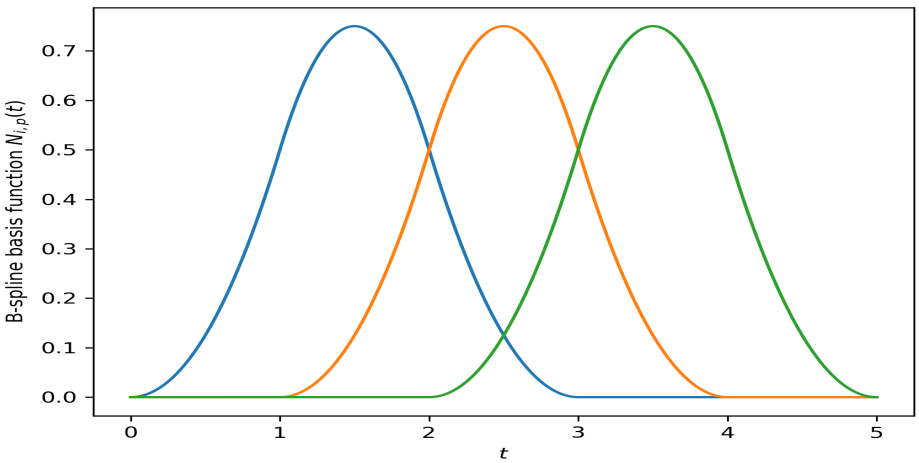

Figure 2-2: Quadratic Spline Basis

Cubic B-spline:

$$N_{i,3}(t) = \begin{cases} \frac{(t-x_i)^3}{(x_{i+3}-x_i)(x_{i+2}-x_i)(x_{i+1}-x_i)} & \text{if } x_i \le t < x_{i+1}, \\ \frac{(t-x_i)^2(x_{i+2}-t)}{(x_{i+3}-x_i)(x_{i+2}-x_i)(x_{i+2}-x_{i+1})} + \frac{(t-x_i)(x_{i+3}-t)(t-x_{i+1})}{(x_{i+3}-(x_i)(x_{i+3}-x_{i+1})(x_{i+2}-x_{i+1})} & \text{if } x_{i+1} \le t < x_{i+2} \\ + \frac{(x_{i+4}-t)(t-x_{i+1})^2}{(x_{i+4}-(x_{i+1})(x_{i+3}-x_{i+1})(x_{i+2}-x_{i+1})} & \\ \frac{(t-x_i)(x_{i+3}-t)^2}{(x_{i+3}-x_i)(x_{i+3}-x_{i+1})(x_{i+3}-x_{i+2})} + \frac{(x_{i+4}-t)(t-x_{i+3})(x_{i+3}-t)}{(x_{i+4}-x_{i+1})(x_{i+3}-x_{i+1})(x_{i+3}-x_{i+2})} & \text{if } x_{i+2} \le t < x_{i+3} \\ + \frac{(x_{i+4}-t)^2(t-x_{i+2})}{(x_{i+4}-x_{i+1})(x_{i+4}-x_{i+2})(x_{i+3}-x_{i+2})} & \\ \frac{(x_{i+4}-t)^3}{(x_{i+4}-x_{i+1})(x_{i+4}-x_{i+2})(x_{i+4}-x_{i+3})} & \text{if } x_i \le t < x_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{2-5}$$

$N_{i,p}(t) = 0$ if $x \notin [x_i, x_{i+p+1}]$, otherwise $N_{i,p}(t) > 0$ if $x \in [x_i, x_{i+p+1}]$. Therefore, for all B-splines, the basis function $N_{i,2}(x)$ is zero except on the interval $[i, x_i + p + 1]$ (Mungia & Bhatta 2015).

The figure 2-3 below describes the cubic spline basis function of degree $p = 3$, order $p + 1$

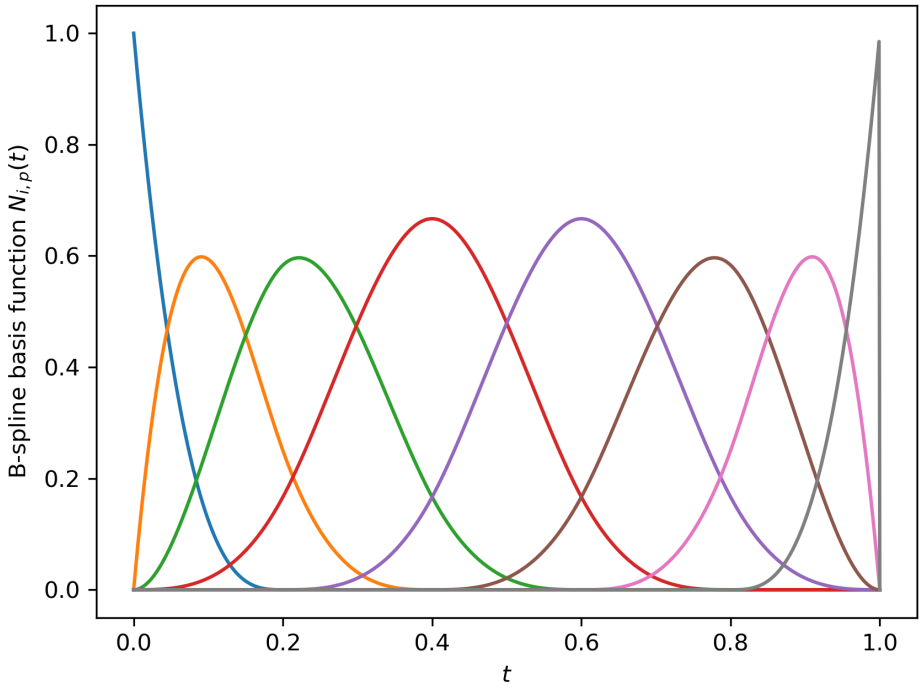

Figure 2-3: Cubic Spline Basis

We have to check for zero resultant value in the case of null denominator, which can occur in the case of knots with greater multiplicity of one. To tackle this, all that is needed to be done is to declare that whatever results into division by zero should be regarded as zero because we know that if $t_i = t_{i+1}$, we will get zero (Lyche & Mørken 2008).

2-1-2 Knot Vector

Different types of knot vectors include uniform knot vectors, non-uniform knot vectors, and open knot vectors. Uniform knot vectors have even spacing of knot values whereas non-uniform knot vectors do not, these spaces are known as segments. The point of connection between these segments are therefore called knots, and the collection of these knots are stored as a knot vector, which play an important role in the understanding of this kind of curve.

The knot vector is used to specify values in the evaluation interval where the curve changes segment. For an open basis function however, we need the first and the last data point to be interpolated. To achieve this, the first and last knots are repeated four (4) times for a curve of degree 3.

$$\vec{t}_{periodic} = \mathbf{[0,1,2,3,4,5]} \tag{2-6}$$

$$\vec{t}_{non-uniform} = \mathbf{[0,1.5,2,3,4,6]} \tag{2-7}$$

$$\vec{t}_{open} = \mathbf{[0,0,0,0,0.2,0.4,0.6,0.8,1,1,1,1]} \tag{2-8}$$

An example of a periodic knot vector is presented in equation 2-6 to 2-8 . An example of a non-uniform knot vector is shown in equation 2-7 and an example of a open uniform knot vector for a cubic B-spline of order $k = 4$, degree k - 1 is shown in equation 2-8.

As seen in the equation 2-8, open knot vectors have order $k - 1$ multiplicity of the first and last knot values. Repetition of a knot value has the effect of drawing the curve closer to a specific control point. If a knot value is repeated $p - 1$ times the B-spline curve will pass through the associated control point. Therefore in the case of an open B-spline (constructed from an open knot vector) the first and last control points will be forced to interpolate the end points by transforming the first and last segments to single points. What we have essentially done is to create a discontinuity at that particular point. The number of repeated knot values are often referred to as the multiplicity of the knot (Piegl & Tiller 1997).

The equation 2-9 below simplifies the relationship between continuity (C_s), order (k) and multiplicity (m) of knot values. For example a cubic open B-splines of degree 3 (order k = 4), multiplicity m (m = k -1) will have C^0 continuity at the first and last knot values in the knot vector as explained earlier. This relationship also indicate that continuity decreases with increasing multiplicity at the associated location of the b-spline curve.

$$C_s = k - m - 1 \tag{2-9}$$

2-1-3 Cubic B-Spline Curves

According to Piegl & Tiller (1997), the p_{th} degree B-spline curve is defined as:

$$C_s = \sum_{i=0}^{n} N_{i,p}(u) P_i \qquad a \leq u \leq b \tag{2-10}$$

$$U = \underbrace{a, ..., a}_{p+1}, u_{p+1}, ..., u_{m-p-1}, \underbrace{b, ..., b}_{p+1} \tag{2-11}$$

Where:

$\mathrm{N}_{i,p}(t) = \mathrm{p}_{th}$ degree B-spline basis function defined by non uniform knot vector.

P_i = control points.

U = non-decreasing set of real numbers knot sequence.

The B-spline $C(u)$ can thus be represented as a vector of Cartesian values which are a function of the parametric basis space as shown in equation 2-12 below, adapted to a cubic spline is defined as (Bindiganavle 2000):

$$C(u) = \left\{ \begin{array}{l} x(u) = \sum_{i=0}^{N+2} dx_i N_{i,p}(u) \\ y(u) = \sum_{i=0}^{N+2} dy_i N_{i,p}(u) \\ z(u) = \sum_{i=0}^{N+2} dz_i N_{i,p}(u) \end{array} \right\} \tag{2-12}$$

Now, because in our case, there is a desire to have a particular surface to be fitted by the b-spline curve and by extension, surface, the control points and basis functions therefore must be determined to produce a B-spline curve that represents the desired surface.

Piegl & Tiller (2000) developed an approach for B-splines curves that addresses this issue which he gave as an interpolation problem considering a set of points (Q) at knot vector (U).

The B-spline curve of set of points Q_r $r = 0, ..., n$ can therefore be written as:

$$C(t_r) = \sum_{i=0}^{n} N_{i,p}(u) P_i \qquad r = 0, ..., n \tag{2-13}$$

The B-spline curve will be produced by control points (P) and basis functions $N_{i,p}(u)$ of order k. It is important to note here that the difference between tr and u is the multiplicity condition we have added.

A system of linear equations could therefore be solved to determine the appropriate control point values p_i to produce the curve for the desired geometry (a_i).

$$a_i = j_i N_{i,k}(t_j) \tag{2-14}$$

To find the curve, the following information are required:

The parameters t_0, . . ., t_n

The knot vector U and

The control points P_0, . . ., P_n.

There are different methods of obtaining the parameters and knot vectors. These different methods will be discussed in section 2-1-5. In general, the number of control points $(n+1)$ must satisfy the following criteria: $k \leq n+1 \leq j$ where j is the number of points provided for the desired geometry. Using an arbitrary knot vector and order for the B-spline curve, the basis functions may be evaluated at the knot values t_j. For a set of data points given as $\mathrm{A} = (a_j)$ from the desired geometry, and the control points given as $\mathrm{P} = (p_i)$;

$$[\mathrm{N}][\mathrm{P}] = [A] \tag{2-15}$$

The equations can be written in matrix formulation which is useful as a visualization of the numerical realization in the provided code for implementation:

$$\begin{bmatrix} N_{1,p}(t_1) & \cdots & N_{n+1,p}(t_1) \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ N_{1,p}(t_{n+1}) & \cdots & N_{n+1,p}(t_{n+1}) \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_{n+1} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ \vdots \\ a_{n+1} \end{bmatrix} \tag{2-16}$$

And according to Rogers (2001), upon calculation of the basis functions, and given the desired geometry, the system of equations 2-16 may be solved for the necessary control point values using the equation 2-17 below.

$$[\mathrm{P}] = [N]^{-1}[A] \tag{2-17}$$

2-1-4 End Point Condition

For a smoother B-Spline curve, Piegl & Tiller (2000) use derivative up to $p-1$ at end points and setting the end derivatives with respective to the data points to zero.

$$\mathbf{D}_s^{(1)}, ..., \mathbf{D}_s^{(k)} \qquad \mathbf{D}_e^{(1)}, ..., \mathbf{D}_e^{(l)} \qquad k, l < p \tag{2-18}$$

where $D_s^{(i)}$ denotes the i_{th} derivative at the start point, and $D_s^{(j)}$ is the j_{th} derivative at the end. And in the bid to satisfying equation 2-13, the curve must fulfill the following additional conditions:

$$\mathbf{C}^{(i)}(t_0) = \mathbf{D}_s^{(i)} \qquad \mathbf{C}^{(j)}(t_n) = \mathbf{D}_s^{(j)} \qquad i = 1, ..., kj = 1, ..., l \tag{2-19}$$

The j_{th} derivative of equation 2-13 is given as (Piegl & Tiller 1997):

$$\mathbf{C}^{(j)}(t_n) = \sum_{i=0}^{n} N_{i,p}^{(j)}(u)\mathbf{P}_i \tag{2-20}$$

where;

$$N_{i,p}^{j}(u) = p\left(\frac{N_{i,p-1}^{j-1}(u)}{u_{i+p} - u_i} + \frac{N_{i+1,p-1}^{j-1}(u)}{u_{i+p+1} - u_{i+1}}\right) \tag{2-21}$$

Therefore,

$$C''(t_0) \ = \frac{d^2 N_0^3 t_0}{du^2}\mathbf{P}_0 + \frac{d^2 N_1^3 t_0}{du^2}\mathbf{P}_1 + \frac{d^2 N_0^3 t_2}{du^2}\mathbf{P}_2 = 0 \tag{2-22}$$

$$C''(t_n) \ = \frac{d^2 N_n^3 t_0}{du^2}\mathbf{P}_n + \frac{d^2 N_{n+1}^3 t_0}{du^2}\mathbf{P}_{n+1} + \frac{d^2 N_{n+2}^3 t_2}{du^2}\mathbf{P}_{n+2} = 0 \tag{2-23}$$

With all these in mind, the unknown control points P_0, . . .,P_m can be computed by the following system of equations:

$$\begin{bmatrix}
1 & & & & & & & & & & \\
N_1^0(t_0) & N_1^1(t_0)) & & & & & & & & & \\
\vdots & \ddots & & & & & & & & & \\
N_0^k(t_0) & \cdots & N_k^k(t_0) & & & & & & & & \\
 & & N_0(t_1) & \cdots & N_p(t_1) & & & & & & \\
 & & & \ddots & \ddots & & & & & & \\
 & & & & N_{m+p}(t_{n-1}) & \cdots & N_{m-p}(t_{n-1}) & & & & \\
 & & & & & N_{m-l}^l(t_n) & \cdots & N_m^l(t_n) & & & \\
 & & & & & & \ddots & \vdots & & & \\
 & & & & & & & N_{m-1}^1(t_n) & N_m^1(t_n) & & \\
 & & & & & & & & 1 & &
\end{bmatrix} \cdot \begin{bmatrix} P_0 \\ P_1 \\ \vdots \\ P_{m-1} \\ P_m \end{bmatrix} = \begin{bmatrix} Q_0 \\ D_s^{(1)} \\ \vdots \\ D_s^{(k)} \\ Q_1 \\ \vdots \\ Q_{n-1} \\ D_e^{(l)} \\ \vdots \\ D_e^{(l)} \\ Q_n \end{bmatrix} \tag{2-24}$$

We then solve the linear systems to obtain the x and y coordinates in 2 dimension or x, y, and z coordinates if 3 dimensional surface (or curve in 3D) is desired by making use of the control points derived from the solved solution.

2-1-5 Parameter Selection

Different strategies to estimate the parameters are used to calculate the distance between neighboring knots in a knot vector. These techniques are essential for modeling B-splines because, as was previously said, the spacing of the knot sequence affects the basis functions. It essentially amounts to specifying the size of each parametric interval, which, when transferred to modeling space, will specify each curve segment. According to Shene (1997), if the data points are d_0, ..., d_n, then $n+1$ parameters t_0, ..., t_n inside the domain of the curve must be found in order for data point d_k to corresponds to parameter t_k for k between 0 and n, which in turn means that for $C(t)$ to pass through all the data points, the d_k must be equal to $C(t_k)$ for all $0 \leq k \leq n$. There are three basic approaches that are frequently used to parameterize curve data: centripetal, chord length, and uniform. Below is a discussion of these strategies.

Uniform Method

When the knot spacing is set to be the same for each interval such that the knots are equally spaced, this is frequently viewed as the simplest technique to estimate parameters, however, it is known to result in erratic shapes or other unpleasant results (Shene 1997). This approach is too basic to handle the majority of real-world scenarios (Farin 2002). The uniform parameterization's general ineffectiveness can be attributed to the fact that the geometry of the data points is ignored. According to Amirfakhrian (2012), the equation is given as:

$$t_{i+1} = t_i + 1 \tag{2-25}$$

Chordal Method

Having the knot spacing proportionate to the data points' distances is one technique to avoid the potential for neglecting the geometry of the data points (Farin 2002). This can be accomplished by observing that if an interpolating curve follows the data polygon very closely, the length of the curve segment between two adjacent data points will be very close to the length of the chord of these two data points, and the length of the interpolating curve will also be very close to the total length of the data polygon (Shene 1997). The distance between the data points serves as the foundation for this parameter estimation technique. The knot spacing varies in direct proportion to the separation of the data points. It reflects the geometry of the data points and is said to be more accurate. The equation is given as (Amirfakhrian 2012):

$$\frac{t_{i+1} - t_i}{t_{i+2} - t_{i+1}} = \frac{\|d_{i+1} - d_i\|}{\|d_{i+2} - d_{i+1}\|} \tag{2-26}$$

Centripetal Method

Centripetal technique is another parameterization method proposed and named by Lee (1989). This is an extension of the chordal method in section 2-1-5. The equation is given as (Amirfakhrian 2012):

$$\frac{t_{i+1} - t_i}{t_{i+2} - t_{i+1}} = \left(\frac{\|d_{i+1} - d_i\|}{\|d_{i+2} - d_{i+1}\|} \right)^{1/2} \tag{2-27}$$

2-1-6 Knot Vector Generation

In order to construct a knot vector, for a parameter $t_0 = 0.0$ and $t_4 = 1.0$ which represent the first and last point of the parameter sequence, they have to be repeated four (4) times for a cubic b-spline curve of degree 3 and order 4 to interpolates first data point d_0 and the last data point d_n to generate the multiplicity of 0s and 1s. Therefore we can generate a knot vector from the parameter, for example, to be:

parameter t $= 0, 0.2, 0.4, 0.6, 0.8, 1$ and

knot vector $= 0, 0, 0, 0, 0.2, 0.4, 0.6, 0.8, 1, 1, 1, 1$

2-1-7 Solving Triadiagonal Linear System

As seen in equation 2-15, to solve for the control points, we will be dealing with a tridiagonal linear system or tridiagonal matrices. For a linear which could be put into a matrix-vector form can be written as:

$$[\mathrm{N}][\mathrm{P}] = [A] \tag{2-28}$$

Where N is an n x n coefficient matrix while P and A are the control point coordinates to be estimated and data point coordinates respectively. A Tridiagonal linear system is a type of linear system in which the main diagonal, the first diagonal above the main diagonal, and the first diagonal below the main diagonal are the only three diagonals in a Matrix with non-zero components (James 2013). In this section, we discussed the method used to solving this kind of equation. The Thomas algorithm was used which is also known as The Tri-Diagonal Matrix Algorithm (TDMA). Thomas Algorithm is a simplified form of Gaussian elimination which can be used to solving tridiagonal systems of equations with the advantage of requiring less calculations and thus less storage when compared with the Gaussian method (Lorena 2011). However the downside of this method is that it can be unstable in certain situations and requires certain condition which will be discussed below. The Thomas algorithm consists of two steps which are forward elimination and backward substitution. A tridiagonal linear system with n unknown variables can be written in a general form as (Lorena 2011):

$$a_i x_{i-1} + b_i x_{i+1} + c_i x_{i+1} = d_i, i = 1, 2, ..., n \tag{2-29}$$

and to a matrix form ($a_i = c_n = 0$)

$$\begin{bmatrix} b_1 & c_1 & & & 0 \\ a_2 & b_2 & c_2 & & \\ & a_3 & b_3 & c_3 & \\ & & \cdots & \cdots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & & & a_n & b_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix} \tag{2-30}$$

Considering the n_{th} equation in the matrix description above that is to be solved:

$$\begin{aligned} 0 + b_1 x_1 + c_1 x_2 &= d_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 &= d_2 \\ &\dots \\ a_i x_{i-1} + b_i x_{i+1} + c_i x_{i+1} &= d_i \\ &\dots \\ a_n x_{n-1} + b_n x_n + 0 &= d_n \end{aligned} \tag{2-31}$$

Starting from the forward elimination procedure, we can eliminate the x_1 from the system by using the first equation to eliminate x_1 from the second equation and likewise the second equation can be used to eliminate x_2 from the third equation (Thuan 2018). The effect is that x_1 has been eliminated from the second equation.

This will result into:

$$(b_2 b_1 - c_1 a_2) \cdot x_2 + c_2 b_1 \cdot x_3 = d_2 b_1 - d_1 a_2 \tag{2-32}$$

And as stated earlier, we can eliminate x_2, using the modified second equation and the third one, we can apply the same technique to the third and the forth equation, yielding:

$$[b_3(b_2 b_1 - c_1 a_2) - c_2 a_3] \cdot x_3 + c_3 \cdot (b_2 b_1 - c_1 a_2) \cdot x_4 = d_3 \cdot (b_2 b_1 - c_1 a_2) - (d_2 b_1 - d_1 a_2) \cdot a_3 \tag{2-33}$$

This technique can be repeated until the $(n-1)_{th}$ variable has been eliminated, we can obtain one equation which contains only one variable x_n since at this point, the only two unknowns are the x_n and x_{n-1}. Then the backward substitution is used to obtain the remaining $(n-1)$ variables through the back substitution into the n_{th} equation which provides the answer for the x_{n-1} and then the solution to the $(n-1)_{th}$ can be obtained from knowing x_{n+1} (Thuan 2018).

2-2 Basis Spline Surface

The cubic spline can be extended to surface by considering rows and columns spline with each spline having its control points and hence it is popularly called a bi-cubic basis spline. A bicubic surface is an extension of equation 2-13 applied on a two-dimensional grid of knots, where third order polynomials, with the same degree $d = 3$, are used in both directions to shape a bicubic parametric surface. The only assumption concerns the particular organization of the knots along x and y directions using the knot vector U and V respectively. These knot vector are obtained the same way however in different directions. For example, we simply apply the knot vector U to the x and z coordinates while the knot vector V is applied to the y and z coordinates.

The general formulation basis spline surface or bicubic surface can be written as (Piegl & Tiller 1997):

$$\mathbf{C}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(u)N_{j,q}(v)\mathbf{P}_{i,j} \tag{2-34}$$

with

$$U = \underbrace{a,...,a}_{p+1}, u_{p+1},...,u_{n-p-1}, \underbrace{b,...,b}_{p+1} \tag{2-35}$$

$$V = \underbrace{a,...,a}_{q+1}, u_{q+1},...,u_{m-q-1}, \underbrace{b,...,b}_{q+1} \tag{2-36}$$

Where:

$\mathrm{N}_{i,p}(u) = \mathrm{p}_{th}$ degree spline basis function defined by non uniform knot vector.

$\mathrm{N}_{i,q}(v) = \mathrm{q}_{th}$ degree spline basis function defined by non uniform knot vector.

$P_{i,j}$ = control points in U and V directions

U and V = non-decreasing set of real numbers knot sequence.

2-3 Volumetric Modeling

Surface-based reservoir modeling employs a boundary representation approach in which all heterogeneity of interest, whether structural, stratigraphic, sedimentological, or diagenetic, is modeled by its bounding surfaces, which are not grid-dependent (Jacquemyn et al. 2019*b*). A main topological requirement in volume modeling is that surfaces should only intersect along common borders (Mäntylä 1988). Surfaces can therefore be combined to produce a volumetric model whereby any heterogeneity to be modeled within such volumes is incorporated by adding surfaces. Surfaces must therefore be joined, terminated, and stacked in order to create a surface-bound volume. There are two criteria that must be met by surfaces that bind closed volumes, which are (Jacquemyn et al. 2019*b*):

1. They must have an acceptable geometry that depicts the shape of the geological heterogeneity to be modeled, and
2. To form closed volumes, they must have interactions with all neighboring surfaces.

According to Jacquemyn et al. (2019*b*), when portraying geologically realistic geometries as volumetric models, three forms of surface contact must be considered:

1. Joining of surfaces at their edges to form closed volumes.
2. Termination of surfaces arranged hierarchically.
3. Warping of surfaces and surface sets.

The required surfaces and curves can be generated using cubic basis splines as described in previous sections. These surfaces and curves have the potential to be manipulated and optimized in order to efficiently create a more realistic and accurate geometry capable of preserving detailed and complex geological structures in our geological models.

2-3-1 Surfaces Terminations

Because the generated volumes are utilized for calculations and simulations, the manner in which surfaces are truncated and joined to other surfaces is critical. The topography of the surface(s) utilized to produce the volumes, in turn, influences the results of these simulations. The design of a surface in which their respective edge points correspond is accomplished by ensuring that the number of points in the U and V directions, as well as their respective control points along the shared edge, are the same. Attaching surfaces with the same degree in the U and V directions of their shared edge is a simple procedure.

Below is the surface modeled to indicate how the linking of the two surfaces is best carried out. Two different surfaces will unlikely be the same distance at all points between one another and also unlikely to be in the same position in space with only a vertical difference. This means that in order to uniformly link all the points on the edges of surface A with their respective points on surface B, uniformly divided points must be used.

Hence, the distances between each point pair are extracted in three (3) dimensional space and then subdivided equally with distance, as seen below in figure 2-4 below. This allows for a symmetrical layering and truncation of the surface in three-dimensional space.

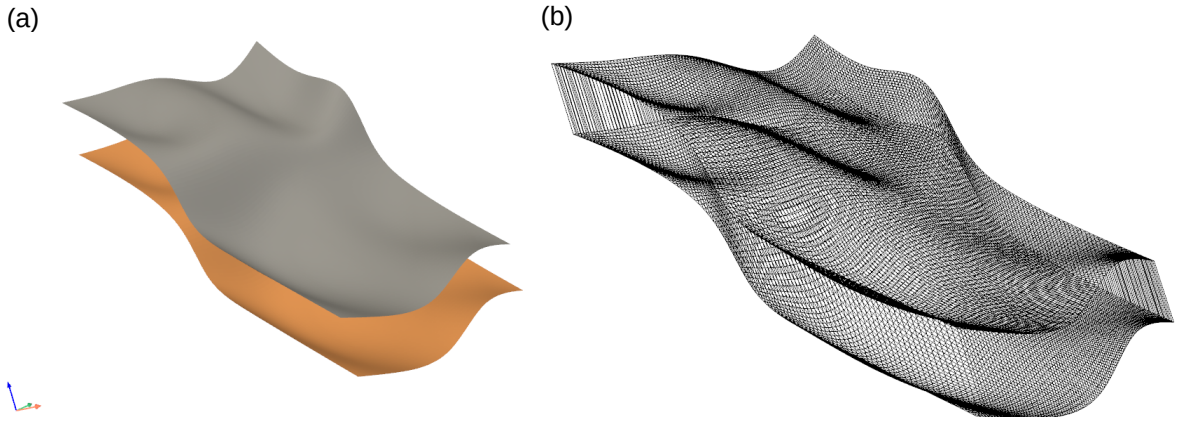

Figure 2-4: Conversion of surface-based model (a) to a volumetric model (b) through termination surface points

The geological formation obtained from the combination of this two surfaces are filled with grid points so that we can represent it as voxels and able to carry out further analysis such as uncertainty evaluation. Section 2-3-2 below shed more light on this important step.

2-3-2 Voxelizations

The term "voxelization," also known as "discretization," refers to the process of transforming data structures that store geometric information in a continuous domain into a discrete grid. Voxelization is an important processing step in this project because it not only provides the opportunity to utilize the Bayesian inference approach to estimate uncertainties but also provides the opportunity to evaluate the estimated uncertainties of our target formation.

The establishment of the occupancy of voxels located on a uniform 3D grid is the most general technique to express a voxel model (Patil & Ravi 2006). For each cell, either a binary value is stored indicating the presence or absence of lithology type or a numeric value is stored representing a physical quantity such as density, as we have in our case.

Voxel models can be represented using two major schemes which are the array and octree schemes (Patil & Ravi 2006).

At every step of partitioning, the actual implementation of octree employs pointers to express connections between parent and sibling nodes, although the array implementation approach is considerably simpler and straightforward (Patil & Ravi 2006). Unstructured data such as cartesian coordinates represented as points cloud can be turned into a voxel representation where points in the point cloud are designed to fall in certain voxels.

For genuine representation of very complex structure, the octree approaches its full size, and the array implementation may be more efficient in terms of storage than the octree (Patil & Ravi 2006). This is because we must keep pointers to the parent and eight siblings of each node in the octree.

In general, regardless of the voxel scheme adopted, the voxels with points in them are left alone, while all others that do not intersect any points are discarded. The number of voxel is given as resolution, hence the more the resolution, the less the number of points intersecting the same voxel and the more the chances of each point to intersect a unique voxel and get represented provided the points are in the zone of the voxel grid. Figure 2-5 below give the visual illustration in 2D how voxelization works, with respect to the two different schemes.

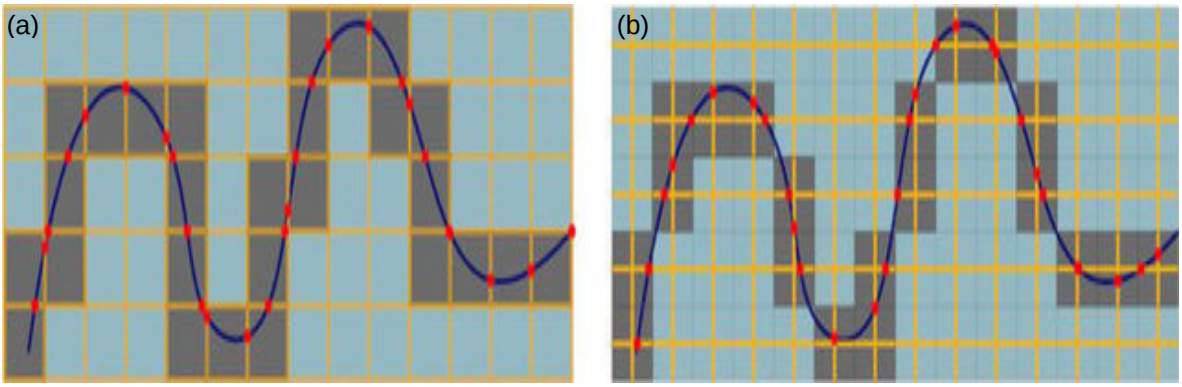

Figure 2-5: Different Voxelization methods using Array schemes (a) and Octree schemes (b) (Nourian et al. 2016)

In this project, lithology are modelled as surfaces and these lithologies are filled with cartesian coordinate grid points as explained earlier to enable a good volumetric representation of the 3D geological model as voxel grid. It is important to note that, to be able to get this done, the point cloud needs to be arranged as grids so that it can be fitted with the voxel grid as shown in figure below.

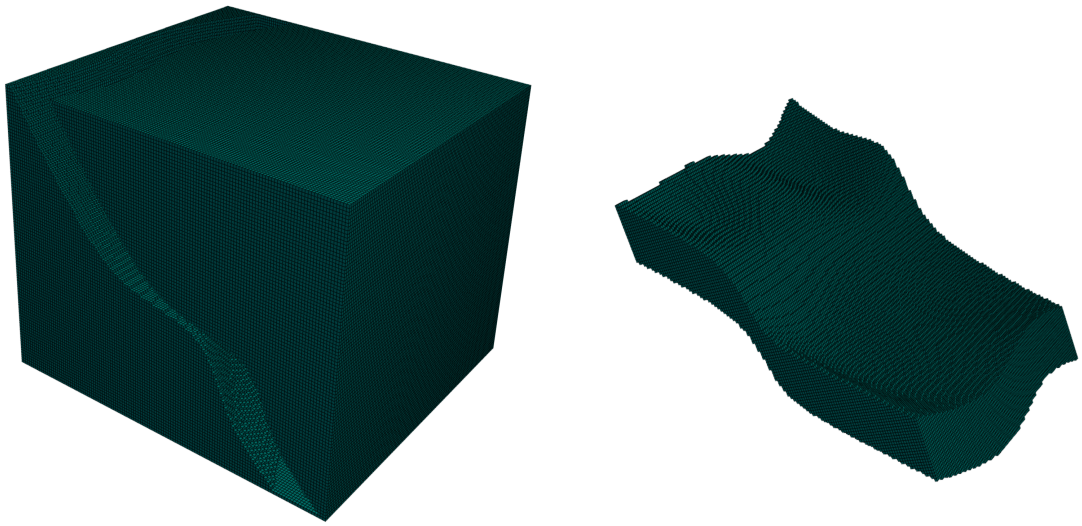

Figure 2-6: Complete voxelization of the grid point and extraction of the target formation voxels based on the target points enclosed by the surface layer

2-4 Uncertainty Estimation

As stated earlier in the introduction, to make a plausible decision when carrying out geomodeling, it is important to estimate the uncertainties involved in the process which can be very complex and costly if not properly managed. This is because in geological modeling, there are different sources of uncertainties which include initial geological observations, geological knowledge and from diverse source of input data (Wellmann & Caumon 2018).

Wellmann & Caumon (2018) developed an automated strategy for evaluating uncertainty in geological models that allows for the integration of the geological modeling phase into inversion approaches as well as integration with geophysical data by making use of geological models in joint inverse approaches.

This type of uncertainty estimation process makes use of a Bayesian Inference approach, which involves the use of Bayes theorem to calculate or update the probability of a scenario being true given certain constraint(s). Here, probability is treated as measure of belief or confidence that we have on a particular event occurring. This is what this section will dwell upon.

Many, if not most, statistical studies are carried out with the end purpose of decision making in mind (Davidson-Pilon 2015). However, more often than not, these decision making processes are affected by the uncertainties involved in the statistical studies which can appear with the information gathered or associated with the parameters used. Available statistical knowledge is employed in statistical decision analysis to gather more information on the nature of these uncertainties. Uncertain parameters can be thought of as numerical quantities. To identify the optimum solution to a particular problem, sample information may be combined with other factors such as the potential implications of decision making and the availability of earlier knowledge on our uncertainty. This is one of the reasons we use Bayesian inference in decision analysis to evaluate the conditional contribution of important parameters and outcomes based on past decision information (Davidson-Pilon 2015).

In order to investigate or estimate the possible uncertainties, we need to know how they are distributed and which pattern relative to the gathered data they follow. This is because the kind of probability distributions assigned to the input data and parameters have a direct effect on the subsequent step of uncertainty quantification (Wellmann & Caumon 2018). Some sources of geological uncertainty can reasonably be described with a normal distribution, for example the position of an interface in a wire-line well logs (Wellmann & Caumon 2018).

2-4-1 Bayesian Inference Analysis

Bayesian inference is generally regarded as a universal method used for summarizing uncertainty, making probabilistic estimates as well as making predictions using probabilistic statements which are constrained by an observed data and an assumed model (Gelman 2008). The Bayesian perspective is thus applicable to all aspects of statistical inference, while being open to the incorporation of information items resulting from earlier experiments and from expert opinions (Bois 2013). Therefore by regarding geomodeling as a Bayesian inference problem, additional information whether geological or geophysical information can be incorporated as likelihood functions linked to prior parameters in a probabilistic framework.

While trying to estimate uncertainty, Bayesian inference differs from more traditional statistical inference because it preserves uncertainty. In recent years, the use of Bayesian inference approach has increased, however prior to this period, the approach usually carried out as noted by Gelman et al. (2015) is what is known as the Frequentist approach in the form of Monte Carlo simulation where we have a long repeated experiment or sampling, after which the average of outcome distribution is made use of. A frequentist approach is not entirely bad, but it requires a lot of sampling. We know that if we sample from a known possible outcome in a *large number* of times, the resulting outcome tends towards the mean of the distribution. Now, the question would be, if we actually know the required number of sampling to attain the *large number* and more importantly if we can handle the computational difficulties of storage and execution on big data.

In the geomodeling field, there have been recent developments in the use of this probabilistic approach to address uncertainties in the field of geological modeling. de la Varga & Wellmann (2016) was able to develop structural geologic modeling as an inference problem, integrating additional information such as thickness as the constraint. By linking additional geological knowledge to prior model parameters in the form of likelihood functions and using MCMC sampling to explore the resulting probability spaces, de la Varga & Wellmann (2016) were able to reach posterior model distributions with reduced uncertainty.

It is important to note that Bayesian methodology returns distributions of the unknowns, known as the posterior distribution (Davidson-Pilon 2015). The wider the distribution, the less certain the posterior distribution is. However, the resulting posterior distribution can be used to acquire point estimates for the parameter θ true state. Therefore, used to describe the unknowns. The mode is a common and basic example of such an estimator which is known as the Maximum a Posteriori, MAP for short. However, Davidson-Pilon (2015) contends that while employing pure accuracy measurements is objective, it undermines the original aim of doing statistical inference in circumstances when choice payoffs are valued more than their accuracies. The introduction of loss and the application of loss functions provide a more appropriate approach (Davidson-Pilon 2015).

By making use of the traditional probability notation, the initial state of our unknown is denoted as parameter θ in form of distribution. We can then denote our belief or initial guess about event θ as $P(\theta)$ which is called the prior probability. In the presence of an additional information say y, we can update our belief on the unknown. This updated belief is known as posterior probability which can be denoted as $\mathbb{P}(\theta|y)$, interpreted as the probability of θ given the evidence y.

Therefore, the objective in Bayesian inference is to determine the posterior distribution $\mathbb{P}(\theta|Y = y)$ of a parameter set θ, given prior distributions $\mathbb{P}(\theta)$ and likelihood functions $\mathbb{P}(Y|\theta, \mathcal{M})$, which contain the additional information or the observation y related to the unknown parameter θ (de la Varga & Wellmann 2016).

Updating our belief is done via the following equation, known as Bayes' Theorem, named after its discoverer Thomas Bayes (Davidson-Pilon 2015):

$$\mathbb{P}(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{2-37}$$

The uncertainty space is proportional to priors $P(\theta|y)$ and likelihood $P(Y|\theta, \mathcal{M})$ as given below (de la Varga & Wellmann 2016):

$$\mathbb{P}(\theta|y) \propto P(y|\theta)P(\theta) \tag{2-38}$$

This way, we are introducing prior uncertainty about events, and by so doing, we are already admitting that any guess we make is potentially very wrong but after observing data, evidence, or additional information, we update our beliefs, and our wrongfulness of the initial guess decreases as the update continues.

Gelman et al. (2015) structured Bayesian inference in three (3) major steps:

1. Setting up a joint probability distribution for all observable and unobservable quantities in a problem.
2. Conditioning on observed data: calculating and interpreting the appropriate posterior distribution.
3. Evaluating the fit of the model and the implications of the resulting posterior distribution

These above steps is further discussed in section 2-4-4.

2-4-2 Bayesian Inference Methods

The choice of a suitable numerical method to obtain the posterior distribution depends on the requirements of the particular model (de la Varga & Wellmann 2016). This is because, when carrying out geomodeling, we can often be faced with multi-dimensional distributions spaces where unknown our priors reside in. Therefore, we might not be afforded a simple idealized case where a simple numerical method can be applied. In order to be able to search these multi-dimensional spaces, two methods were utilized. They are the MCMC and the MAP. MCMC provides the most likely distribution while MAP provides a single model outcome assumed to best match prior parameter distributions and the geologic likelihood functions (de la Varga & Wellmann 2016).

Markov Chain Monte Carlo, MCMC

Markov Chain Monte Carlo (MCMC) sampling has shown to be a broadly applicable and dependable approach for intelligently exploring multidimensional parameter spaces. MCMC is a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution (Gelman et al. 2015). In the ordinary Monte Carlo or frequentist approach, random independent samples are drawn from a target distribution in order to approximate its shape. However, the problem with this is that areas with long tails could complicate the exploration of the parametric space (de la Varga & Wellmann 2016).

The general principle of MCMC can be described as follows: Drawing representative samples from a target distribution of unknown shape is based on the conduction of a so-called random walk on the parameter distribution space i_{th} sampling steps are to be performed. The first sampling location is chosen at random. With each subsequent step, a new position is proposed.

The new sample value is then related to the previous step. According to a weight defined by the scaled-up candidate density of the value, the proposed step is then accepted or rejected. In the case of acceptance, the value is added to the sample trace and the process is continued from the current location.

According to Davidson-Pilon (2015), the MCMC procedure can be expressed at a high level as follow:

First we start by sampling from target distribution, using this as the current position. Then we propose moving to a new position through investigation the space and estimation of the likelihood of the sample around the space. We then accept or reject the new position based on the position's adherence to the defined tolerance and prior distributions. If we accept, we move to the new position and start the process all over again. If we reject, we move to a new position and start the process all over again. Then after a large number of iterations, we return all the accepted positions. In this manner, we advance in a general direction toward the regions with the posterior distributions, thereby collecting samples sparsely along the way. We can readily gather samples after we reach the posterior distribution since they almost certainly all belong to the posterior distribution. According to Davidson-Pilon (2015), the goal of this approach is to achieve sampling algorithm convergence towards areas of high probability.

There are different algorithms that have been developed and used for MCMC such as the Metropolis-Hastings Algorithms and the Gibbs algorithm. In this work, Hamiltonian MC sampling algorithm is utilized.

The original Markov Chain Monte Carlo algorithm still commonly in use today, utilizes a Gaussian distribution as its proposal mechanism which is generally referred to as the *Random Walk Metropolis* RWM (Betancourt 2018). Random Walk Metropolis is not only easy to build, but it also has an appealing intuition. However, the proposal distribution is skewed towards large volumes, thus the tails of the target distribution, whereas the Metropolis correction rejects proposals that jump into neighbourhoods with small density (Betancourt 2018), creating a form of bias.

The Hamiltonian MC sampling algorithm was initially proposed by Simon et al. (1987) and further developed by Neal (1995). It makes use of a phenomenon in Physics known as Momentum. Hamilton found a really nice way of representing such systems so that paths that explore the space efficiently naturally pop out from the representation. Hamilton showed that we can describe any mechanical system with just two variables, or two degrees of freedom, represented by position q and momentum p. The force acting on a particle can be calculated as gradient of a potential energy $E(\theta)$ of parameter θ at that point which is given as (Betancourt 2018):

$$\mathrm{E}(\theta) = -\log p(\theta) \tag{2-39}$$

We can then define the total energy of the particle as:

$$\mathrm{H}(\theta, y) = K(y) + E(\theta) \tag{2-40}$$

where $K(p, q)$ is called the kinetic energy, while the term corresponding to the density of the target distribution, $V(q)$ is known as the potential energy. HMC introduces the momentum

of the particle as an auxiliary variable and samples the joint distribution of the particle's position and momentum. This joint distribution of the particle's position and momentum is given by (Betancourt 2018):

$$\mathbb{P}(\theta, y) \propto \exp\{-H(\theta, y)\} \tag{2-41}$$

$$\mathbb{P}(\theta, y) = \exp\{-K(y)\} \times \exp\{-E(\theta)\} \tag{2-42}$$

$$\mathbb{P}(\theta, y) = \exp\left\{-\frac{|y|^2}{2}\right\} \times p(\theta). \tag{2-43}$$

Because the numerical simulation approach only in fact approximates, we have to account for errors, the acceptance probability is given by:

$$P_{\text{acc}}^{\text{HMC}} = \min\left\{1, \exp\{-[H(\theta^*, y^*) - H(\theta, y)]\}\right\}. \tag{2-44}$$

In general, as explained by Neal (2012) and Betancourt (2018), the Hamiltonian Monte Carlo technique begins with a predefined set of parameters θ. This value can be chosen by the user or produced at random, usually from a gaussian distribution, depending on the scenario or the event we want to model. The current value of the parameter is then updated using the leapfrog integrator with discretization time and number of steps according to Hamiltonian dynamics for a specific number of iterations i (Betancourt 2018). The Metropolis acceptance phase is then used, and a choice is taken as to whether to update to the new state or maintain the current state.

Maximum A Posteriori MAP

As we can see, in MCMC, we get the most likely solution in form of distribution. Density estimation, on the other hand, is the challenge of estimating the probability distribution for a sample of observations from a decision-making perspective. Typically, predicting the complete distribution is difficult, therefore having the predicted value of the distribution, such as the mean or mode, would be preferable. Maximum a Posteriori (MAP) is a Bayesian-based technique for estimating the distribution and model parameters that best describe an observed data set, in this case, the posterior distribution. With parameter θ, the maximum estimate is given by (de la Varga & Wellmann 2016):

$$\mathbb{P}(\theta|y) = \arg\max_{\theta}\left[\frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta}\right] \tag{2-45}$$

$$\mathbb{P}(\theta|y) = \arg\max_{\theta}\left[P(y|\theta)P(\theta)\right] \tag{2-46}$$

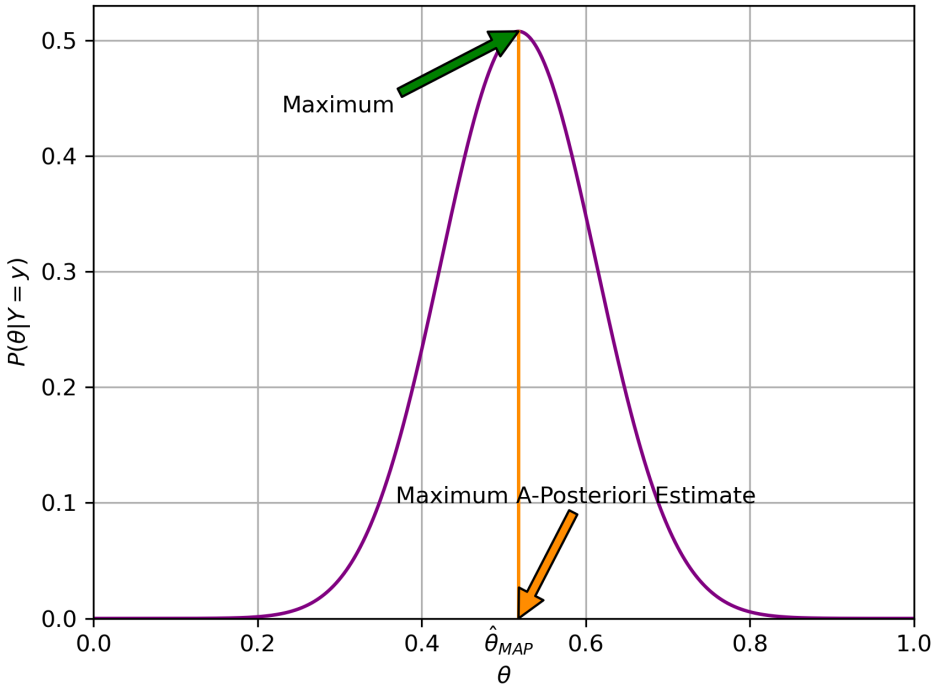

Figure 2-7: The probability of an event $\mathbb{P}(\theta)$ using the Bayesian Maximum A-Posteriori Estimate

Depending on the complexity of the statistical model, there are numerous ways for obtaining the MAP point estimate (de la Varga & Wellmann 2016). There are many algorithm that has been developed to optimize the MAP estimate, Adam, or adaptive moment estimate, stands out. This optimization technique is further discussed in section 2-4-3.

2-4-3 Loss Function

The MCMC and MAP has been discussed in the above section, but what was not explicitly discussed is how or what approach we want to make use of to get to the best estimate from the posterior distribution. There has to be a way to measure that we are moving in the right direction during the simulation, or that we have arrived at the best estimate after the simulation. As many statistician have suggested, the best way to carry out this check is to determine how wrong or right we are after each and every estimation. This is what brought the idea of a loss function which can be defined, as Davidson-Pilon (2015) rightly puts it, as the function of the true parameter, and an estimate of the same parameter. Loss is essentially a measure of how bad the current estimate is when compared to its true self after making a particular decision. The loss function is given as (Davidson-Pilon 2015):

$$\mathbf{L}(\theta, \hat{\theta}) = f(\theta, \hat{\theta}) \tag{2-47}$$

Error estimate of the error function can be calculated using different approach. The commonly

used error function are the absolute-error and the squared-error loss function. The squared-error loss function is given as:

$$\mathbf{L}(\theta, \hat{\theta}) = f(\theta - \hat{\theta})^2 \tag{2-48}$$

while the absolute error function is given as:

$$\mathbf{L}(\theta, \hat{\theta}) =\mid \theta - \hat{\theta} \mid \tag{2-49}$$

The difference between absolute error functions and squared-error function is that in absolute error function, losses increases linearly which means that all differences between relative errors are weighted equally. However, the loss of the squared-error function grows quadratically which means that large errors are weighted much stronger than small errors, thus puts disproportionate emphasis on large outliers (Davidson-Pilon 2015). This is why the absolute error is seen as more robust.

In as much as these examined error function might be sufficient in solving most cases, Davidson-Pilon (2015) propose designing a customized loss functions that specifically reflect an individual's objectives, preferences and outcomes. A probabilistic alternative to the actual loss is to consider each decision's expected loss and to make a decision that is optimal in relation to this expected loss (Davidson-Pilon 2015).

The idea is that, we need to distill our posterior distribution down to a single value, or vector when dealing with a multivariate case and faced with uncertainty. This is known as Bayesian point estimate.

Given a posterior distribution $P(\theta|y)$, the expected loss of choosing an estimate $\hat{\theta}$ over the true parameter θ (after evidence y has been observed) is defined by the function below (Davidson-Pilon 2015):

$$l(\hat{\theta}) = E_\theta \left[\mathbf{L}(\theta, \hat{\theta})\right] \tag{2-50}$$

By the Law of Large Numbers, the expected loss of can be approximated drawing a large sample size N from the posterior distribution, respectively applying a loss function L and averaging over the number of samples (Davidson-Pilon 2015):

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{L}(\theta, \hat{\theta}) \approx E_\theta \left[\mathbf{L}(\theta, \hat{\theta})\right] = l(\hat{\theta}) \tag{2-51}$$

If the value $\hat{\theta}$ is chosen intelligently, we can avoid the flaw of frequentist methodologies that mask the uncertainty and provide a more informative result. The value chosen here, if it is from a Bayesian posterior, is known as a Bayesian point estimate.

However, rather than having to deal with choosing the right $\hat{\theta}$ to estimate θ, we can make use an optimizer that can optimize the loss function for us. First we select the error function, then we optimize this error function by minimizing the estimated error using the optimizer. In this project, the error function approach is to make use of the log likelihood. The main

goal is to maximize the likelihood of observed data given our parameter θ. This function can be optimized to determine the MAP as stated earlier. We define the log likelihood by looking at the predicted probabilities assigned to the right labels (Mehta et al. 2019). This is given as:

$$\log \mathbb{P}(y|\theta) = \sum_{i=1}^{n} (y_i \log \hat{y}_{\theta,i} + (1 - y_i) \log(1 - \hat{y}_{\theta,i})) \tag{2-52}$$

This equation above represent the joint log probability. In order to minimize the log-likelihood, the log likelihood is set to negative. The minimizing negative log-likelihood will result in the optimal estimate (Mehta et al. 2019). The negative log-likelihood is therefore given as:

$$-\log \mathbb{P}(y|\theta) = -\sum_{i=1}^{n} (y_i \log \hat{y}_{\theta,i} + (1 - y_i) \log(1 - \hat{y}_{\theta,i})) \tag{2-53}$$

To optimize the MAP estimate, which essentially finds the local and global minimal, there are many methods that can be used. One of the method is the Adam optimizer as earlier stated.

Using Adam optimizer, individual adaptive learning rates for distinct parameters are calculated using estimations of the gradient's first and second moments. Adam's learning process is broken down into three steps (Kingma 2014):

1. It computes an exponentially weighted average of previous gradients, saves it in variables before bias correction and stores it in another variables after bias correction.

2. It also computes an exponentially weighted average of the squares of the previous gradients, saves it in variables before bias correction and stores it in another variables after bias correction.

3. Finally, it updates the parameters depending on the information extracted from step 1 and 2.

Adam has the following advantages: its parameter update magnitudes are invariant to gradient rescaling, its stepsizes are approximately bounded by the stepsize hyperparameter, it does not require a stationary objective, and works with sparse gradients (Kingma 2014). At this juncture, it is important not to confuse Maximum Likelihood Estimation MLE with MAP. This is because, even though both are point estimates, they differ. The result from MLE maximises the Likelihood $\mathbb{P}(y|\theta)$ while the result of MAP maximises the posterior probability $\mathbb{P}(\theta|y)$. Also the prior information is deterministic in MLE while MAP requires a random prior (De Luca & Termini 1972).

Because Bayesian statistics clearly permits prior belief about models to be included systematically, this flexible probabilistic framework may be utilized to offer a Bayesian foundation for various machine learning techniques.

It should be clearly emphasized that the outcome in Bayesian statistics is the entire posterior distribution across the model parameters, hence the MAP technique solution simply reflects one solution without revealing its associated probability (de la Varga & Wellmann 2016).

Depending on the form of the posteriors, this may or may not result in an adequate approximation of reality and may produce a deceptive result (de la Varga & Wellmann 2016).

No matter the algorithm used in MCMC and MAP, they all require enough iterations for sampling which is primarily dependent on the rate of convergence towards the true distribution. A large number of iterations i has to be chosen for a multi-dimensional space, so that a reliable and statistically significant exploration of the parameter space can be carried out from which different realizations of the 3D geological model can then be constructed based on the approximated posterior distributions. How large the dimensional space we are dealing with is, determines how large the iterations to be used should be (de la Varga & Wellmann 2016).

Other method involves the application of stochastic gradient descent on the negative log-likelihood (Mehta et al. 2019).

2-4-4 Bayesian Inference in Geological Modeling

The fundamental approach outlined by de la Varga & Wellmann (2016) on the application of Bayesian inference on geomodeling is closely related to this project. Therefore, we will be making use of the same approach in the application of Bayesian inference on surface-based geological modeling. However, first, we need to revisit the well researched approach and findings they have used and how it directly relates to this project. According to them, they opined that geomodeling can be regarded as statistical problem that can be solved using Bayesian inference approach, hence, the final goal is to quantify how certain our predefined statement or model is given a specific amount of data. With this perspective in mind, the elements of Bayesian inference can be specified in this context as follows:

1. Mathematical forward model ($\mathcal{M}$): There is a link or connections between parameters θ and observed data y which can be defined as a mathematical model (de la Varga & Wellmann 2016), therefore, the realization of a geological model M can be regarded as a direct function of a set of input parameters. In the context of this project, this would be b-spline function that requires the control points as its parameters θ. In fact, in this project, it is. To build a surface model M, the function as seen in equation 2-34 will be defined by the parameter θ.

2. Model parameters (θ): These model-defining parameters (θ) can be deterministic where a self defined value(s) is assigned to the parameters, or it could be stochastic, acting as an uncertain parameters where sample(s) from probability distribution is assigned to the parameters.

3. Observed data (y): This can be an additional information that can be related to the forward modeling results or to the parameters and their combinations. This comprises mostly secondary data that can not be incorporated directly as input parameters, such as geophysical measurements, for example gravity. They act as a form of constraint to guide the sampling algorithms, thereby reducing the uncertainty around the process.

4. Likelihood functions $P(y|\theta)$: Links between the previous parameters θ and the additional data y are established by these functions in a way that they reflect the likelihood of the parameter states given the observations. They are mathematically defined in the same way as probability functions, but they are a function of the data y, instead of the parameters θ (de la Varga & Wellmann 2016).

According to Gelman et al. (2015), these three procedure should be followed as the sequence of the Bayesian inference:

1. Setting up a full probability model: This can be seen as a multidimensional environment defined by every parameter θ that forms part of the model (de la Varga & Wellmann 2016). A joint probability space is to be generated, taking into account the probability distributions of every model parameter θ. Figure 2-8 below shows an example of a probability distribution of a bivariate parameter θ_x and θ_y. In a multi dimensional space, we could have more than two distributions. These parameters could be, in the case of geomodeling, properties that defines a geological formation such as porosity ϕ and permeability k.

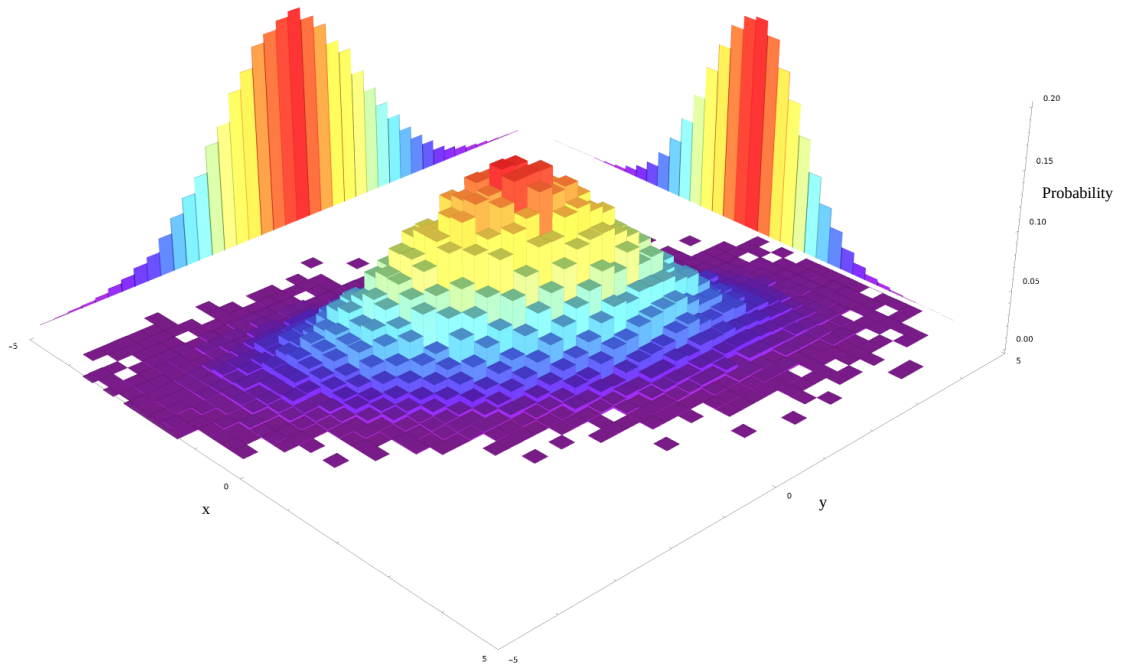

Figure 2-8: A bivariate joint probability distribution of two random parameters θ_x and θ_y. Created using Wolfram Research, Inc. (2010)

2. Conditioning on observed data: Afterwards, to estimate the posterior distribution $P(\theta|y)$, we have to condition the parameters θ prior distribution on the observed data

y (de la Varga & Wellmann 2016). This is the step of Bayesian updating of the belief about uncertain parameter having seen new information. In a chosen model M, this is achieved by linking parameters and data through deterministic operations to the likelihood functions. The result of this is evaluated against the likelihood function of the given observation. It is important to point out that not all the parameters necessarily must be related to all observations but any combination may be valid (de la Varga & Wellmann 2016).

3. Evaluation of the posterior model: Depending on the aim of the study, a post-processing analysis can be conducted accordingly. de la Varga & Wellmann (2016) focused on two aspect in the examination of the posterior distributions of the parameters θ which are the analysis of the posterior distribution of the parameter θ with a Gaussian kernel density estimation. The other has to be do with the stacking of generated models from the posterior distribution and then deriving the information entropy. Wellmann & Regenauer-Lieb (2012) gave a very good approach in visualizing the measures of information entropy which was first developed by Shannon (1948) using voxels. Hence, the need to voxelize our model. This approach is explicitly discussed in following section below 2-4-5.

2-4-5 Uncertainty Evaluation

As stated earlier, the kind of post-processing analysis we want to carry out after Bayesian inference will be dependent on the aim of the study. In this study, the aim is to demonstrate the integration of explicit-implicit approach in surface-based geomodeling. Therefore, the posterior distributions and surface-based model realizations must be assessed. There are quite numbers of ways to carry this out. When we carry out Bayesian inference, some of the output we get include the posterior distribution of the control points, which in fact defines the modal values that best explains our model. Other output include the sample traces. These posterior distribution can be used to construct 3D geological models to be evaluated. One of the best approach is to stack these 3D geological models together and estimate their probability and information entropy as described by de la Varga & Wellmann (2016). However, this can only be done through the voxelizations of each individual model where each and every model in the ensemble is made into a regular raster of equally sized cells(voxels) and measuring the accuracy for every such cell. The process of voxelizations is described in the section 2-3-2.

Wellmann & Regenauer-Lieb (2012) described the utilization of probability and information entropy for the visualization of uncertainties in 3D geological models. They adopted the shannon entropy developed by Shannon (1948) which is also known as information entropy. Shannon entropy is used to predict model correctness at each position in the model space by visualizing its uncertainty using the ensembles models and thereby delivering a measure of model quality. The idea here in geological modeling context is predicated on the knowledge of how frequently a specific geological characteristic occurs in a voxel (Wellmann & Regenauer-Lieb 2012). The use of these entropy metrics enables the computation and visualization of uncertainties in each voxel, the evaluation of uncertainties in whole geological units, and the quantification of total model uncertainty as a single number (Wellmann & Regenauer-Lieb 2012).

As stated by Wellmann & Regenauer-Lieb (2012), it was De Luca & Termini (1972)'s idea

to apply information entropy as a measure of fuzziness. For a fuzzy set, where $f \in [0, 1]$ is a measure of fuzziness of each part of the set, the most important properties are:

1. The measure should be 0, if, and only if f is 0 or 1 in all cell.
2. The measure has its maximal given $f = 0 : 5$ in all cell.

Using the conditions above, which are met by the Shannon (1948) entropy function, if we denote f as a P_m, a probability of an outcome $m \in M$, the fuzziness can be quantified as the entropy H_m normalized by the sum of cells N:

$$\mathbf{H}_m(t) = -\frac{1}{N}\sum_{x=1}^{N}\left[(P_m(x,t)\log P_m(x,t)) + (1 - P_m(x,t))(1 - \log P_m(x,t))\right] \tag{2-54}$$

Extending further, the total information entropy H_t can be calculated for a the whole model space, given as (Wellmann & Regenauer-Lieb 2012):

$$\mathbf{H}_T(t) = -\frac{1}{N}\sum_{x=1}^{N} H(x,t) \tag{2-55}$$

$$\mathbf{H}_T(t) = -\frac{1}{N}\sum_{x=1}^{N}\sum_{m=1}^{M} P_m(x,t)\log P_m(x,t) \tag{2-56}$$

The total entropy H_T is equal 0 when no uncertainty is found in any cell, which indicate that the cell belong to a unique formation. The opposite of this condition would mean the cell contains sub-parts of the model and in all the cell, the probability of all the member in the model ensemble is equal to $1/M$. The figure 2-9 give a pictorial explanation on how the shannon entropy is applied to visualize uncertainties in geological modeling procedures.

2-5 Surface-based Geological Model

In this section, we implemented the theoretical concept that we have described in the previous section by integrating explicit geological modeling approach with implicit geologic modeling approach.

2-5-1 Numerical implementation

The programming language utilized in this project is Python programming language. Python, being an open source and one of the most widely used in the world today for data science programming. It is easy to use and has been around for awhile. Python is an object-oriented language with very diverse and rich packages suitable to be used for scientific research. The fact that it is easy to learn and has a very large community makes a very good option for scientists who are beginners to carryout process like numerical computing. This large

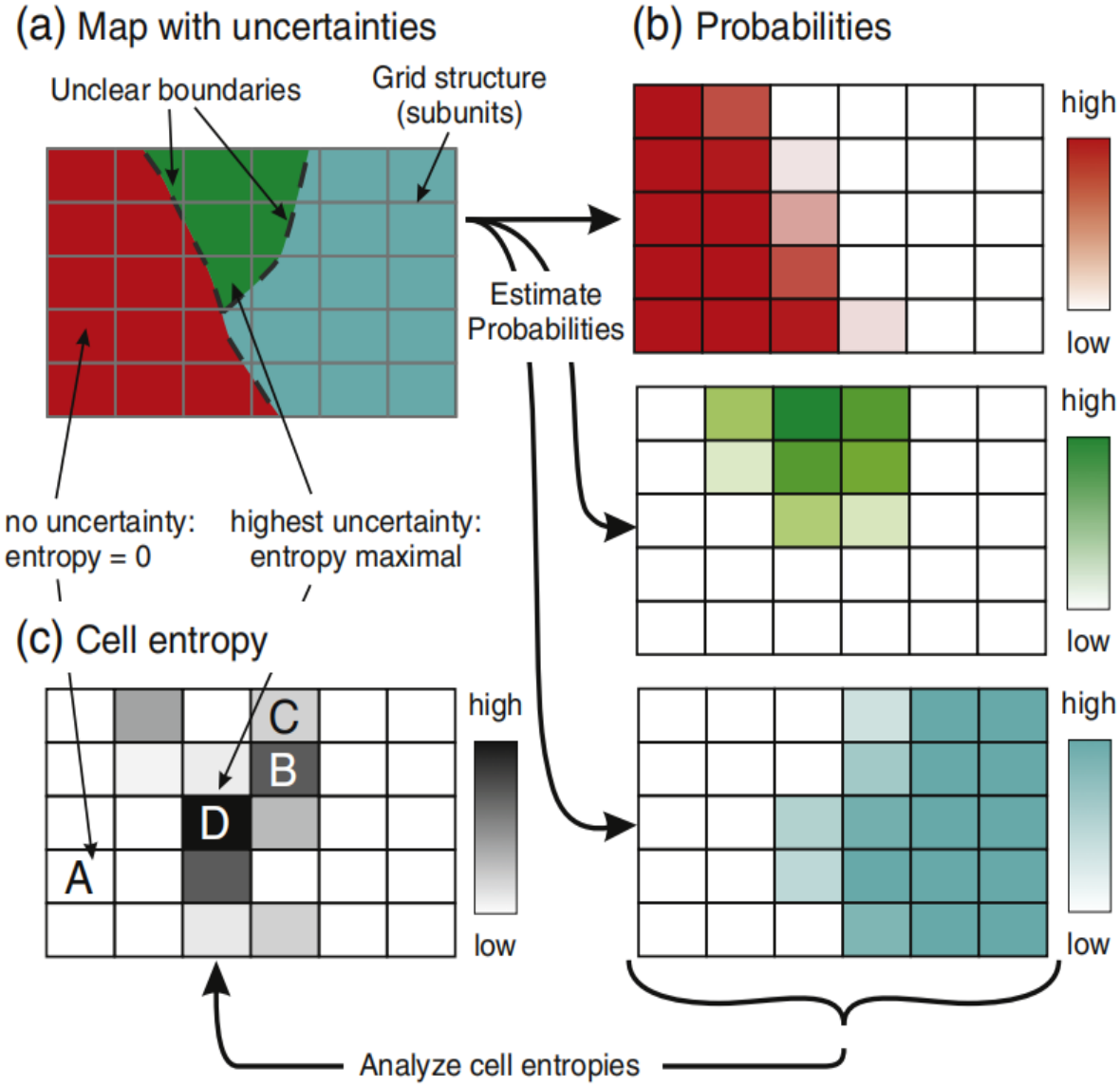

Figure 2-9: Uncertainty Visualization using Entropy: (a) subdivision of an uncertain member map into a regular grid; (b) Probabilities of various outcomes for each cell member; (c) Entropy is 0 when no uncertainty exists and the entropy is highest when all of the members are equally likely (Wellmann & Regenauer-Lieb 2012).

community has been providing well documented solutions both online and offline to people who might be interested in using the programming language for computations. As at the writing of this project, Python 3.9 has been released and basically utilised for this work. There are proper documentation that gave a vivid description of what python programming language is and how it can be integrated with other programming languages for numerical modeling.

The computational process used in this project can be divided into four major categories or steps.

The 3D geological modeling step in this work is implemented using GemPy, which is an open-source 3D geomodeling package. It is a Python-based package that is capable of generating and visualizing complex 3D geological models based on the potential field interpolation method (de la Varga et al. 2019). This package is used to create our theoretical geological model, from which surface points were extracted. This is further explained in section 2-5-3.

The second step involves the introduction of our explicit modeling approach. This involves the use of our cubic spline algorithm to create surface-based geological models. It is integrated with the Gempy package through the surface re-sampling process. The numerical computation in this step utilize Python and some of its libraries. Libraries such as Numba is utilized to speed up the computational process. Numba is an open-source python library that helps to compile python and Numpy code using the LLVM compiler infrastructure (Lam et al. 2015). Other library include SciPy etc.

The third step involves the embedding of the surface-based geological modeling step in a probabilistic context by conducting the Bayesian analysis on the resulting explicit geological models using TensorFlow Probability (TFP). TFP is an open source Python library built on TensorFlow to perform probabilistic programming (Dillon et al. 2017). It provides numerous sampling methods and other analytical and estimation tools such as optimizer that can be used as optimizer for MAP estimation. Such optimizer include Adam, SGD etc. TFP allows the use of two major types of variables to form statistical models, which are deterministic and stochastic numeric variables. Deterministic numeric variables are variables with fixed values. Stochastic variables, on the other hand, are random numerical variables used to either describe uncertain parameters θ or likelihood functions $\mathbb{P}(y|\theta)$. TFP is optimized for data scientists, statisticians, and other practitioners who want to encode domain knowledge to understand data and make predictions (Dillon et al. 2017). It also provides wide selection of probability distributions which can be used to define our stochastic numeric variables based on what we want to model. TFP is fully object-oriented, and can be extended with its own object definitions inherited from the deterministic and stochastic class descriptions. In the case of structural geologic modeling as considered in this work, input parameters θ to the geologic modeling are described as stochastic variables, while the explicit modeling method is implemented as a deterministic function, that requires deterministic numeric variables as denoted in equation 2-34. These two variable types allow us to create different realizations of our surface-based geological models. Since we are dealing with Bayesian inference problem, constraints are used. Python package known as SimPEG is used to generate Geophysical data for our target lithology based on the density. SimPEG is an open source framework for simulation of geophysical data and gradient based parameter estimation in geophysical applications (Cockett et al. 2015). The process of implementation can be divided into two broad categories which are the forward simulation, followed by the inversion. The result will

serve as our constraint.

The last step is to evaluate our models both as a distribution or as an overall estimation. In order to do this, we made use of the entropy and probability estimation approach given by Shannon (1948) and further expatiated by Wellmann & Regenauer-Lieb (2012). This process requires our target volume to be voxelized.

2-5-2 Theoretical Case Study

In order to demonstrate the complete idea in this project, we are making use of a theoretical geological model. These models are created using Gempy, an open-source python package. Gempy makes use of implicit interpolation for geological modeling, a method developed and elaborated by Lajaunie et al. (1997) and Calcagno et al. (2008). It enables the creation of 3D implicit geologic models based on surface contact points between layers and orientation measurements by making use of the potential field method. This idea assumes that points on the same interface are equivalent to a single scalar field value. The scalar field is then interpolated to form the model.

GemPy uses universal cokriging as the interpolation method. For the potential field method, it is possible to incorporate orientation measurements into the interpolation of the scalar field and obtain a gradient field as a result, consisting of vectors orthogonal to the iso-surfaces of the scalar field (Lajaunie et al. 1997). During the interpolation process, because of the potential field utilized, the layer interfaces (scalar field value) relate to one another. This is very fundamental to geological principles, for example, law of original horizontality. This global interpolation approach makes it possible to model geological events, that conform to this law. In order to account for depositional discontinuities like unconformities, the domain can be subdivided into different partitions called series, and each one of them will feature an independent scalar field.

GemPy generates a geologic model, which is typically made up of a series of stratigraphic units or layers. For the resampling process in this project, we will focus on one of those layers. With a scalar field value for each layer, the top and bottom layers will form a single geological lithology. The scalar field can be turned in a three (3) dimensional surface points that is utilized in the explicit modeling process, and afterwards, further subjected to stochastic modeling process to estimate the uncertainties necessary for decision making. Below are the brief geological description of the theoretical models that is used.

Theoretical Models

In this subsection, the structural features of each theoretical model is discussed along with their respective economic importance and geologic settings. These models are subjected to geological and geophysical approach. The resulting property distribution can be considered to be a ground truth during the uncertainty estimation process. All the models are unit neutral because they are conceptual model. Any unit of length can be added as replacement.

1. **Anticline:** An anticline is a geological structure formed by the folding of rock strata into an arch-like shape. The rock layers in an anticline were originally laid down horizontally

and then earth movement due to compressional forces caused it to fold into an arch known as an anticline. In an anticline, the structural contours form closed loops which could be symmetrical or asymmetrical. This type of structure can be made visible on a seismic section as surface reflection or exposed to the surface as an exposed outcrop as seen in figure 2-10.

An anticline has many economic benefit for example when exploring for hydrocarbons. Hydrocarbons that find their way into a reservoir rock that has been bent into an arch will flow to the crest of the arch, and get stuck provided we have a four-way closure and there is a seal rock overlying the arch to keep the entrapped hydrocarbons in place as hydrocarbon tends to flow upward due to buoyancy effect. As a result, this allows the anticline to act as a structural trap. Figure 2-10 (c) is a cross section of the earth showing typical anticline trap. The extent of the crest in terms of the thickness and width of the anticline determines the amount of hydrocarbon it can hold. If the contour below is not enough to accommodate all the hydrocarbon migrating into it, it will move into the next structure by spilling at the point known as the spill point. Knowing the geometry is therefore important and ability to estimate the uncertainties involve is key in making the right decision, for example in quantifying the amount of oil or gas in place, as well as to know where best to drill for these hydrocarbons.

To demonstrate this, a symmetrically deformed anticlinal structure is constructed using the GemPy package.

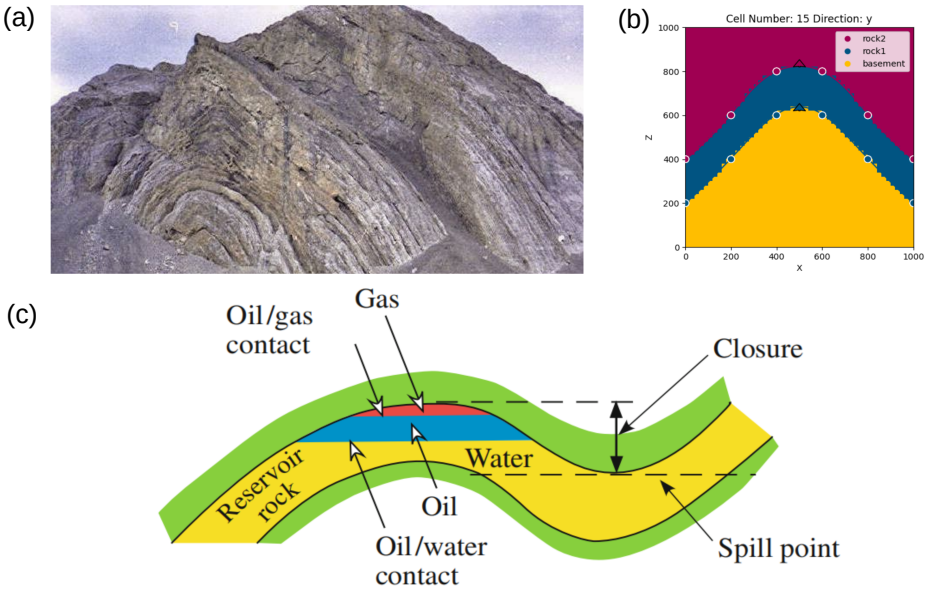

Figure 2-10: Outcrops showing an anticline fold Alberta, Canada (Geology In 2015) (a). Implicit anticline fold geologic model created with GemPy (b). Graphical representation of model economic importance (c) Bjørlykke (2015).

2. **Recumbent Fold:** A recumbent fold is also a geological structure formed by the folding of rock strata into an arch-like shape due to compressional forces. However, a recumbent fold is so much overturned that its axial plane is horizontal or nearly horizontal. Recumbent fold is believed to have formed from nappes which is formed

during continental plate collisions, thereby resulting in large-scale recumbent fold and as a result of this phenomenon, recumbent folds are structures typical of orogenic belts, therefore they usually develop in a compressional tectonic settings (Bastida et al. 2014).

Unlike an anticline, the development of a recumbent fold can tell us about a unique tectonic setting in which they are formed. Bastida et al. (2014) reviews how recumbent fold can be a key structural element in an orogenic belt. Bastida et al. (2014) stated that the physical conditions of these folds' development, the strain present in the folded layers, their formation kinematic mechanisms, the role of force of gravity, the tectonic context of their development and the structures associated with them are the determining factors for their structural geometry. Therefore, subjecting this to an uncertainty estimation process can help us better understand the orogenic process of the earth's crustal layer.

To demonstrate this, a recumbent fold is constructed using the GemPy package.

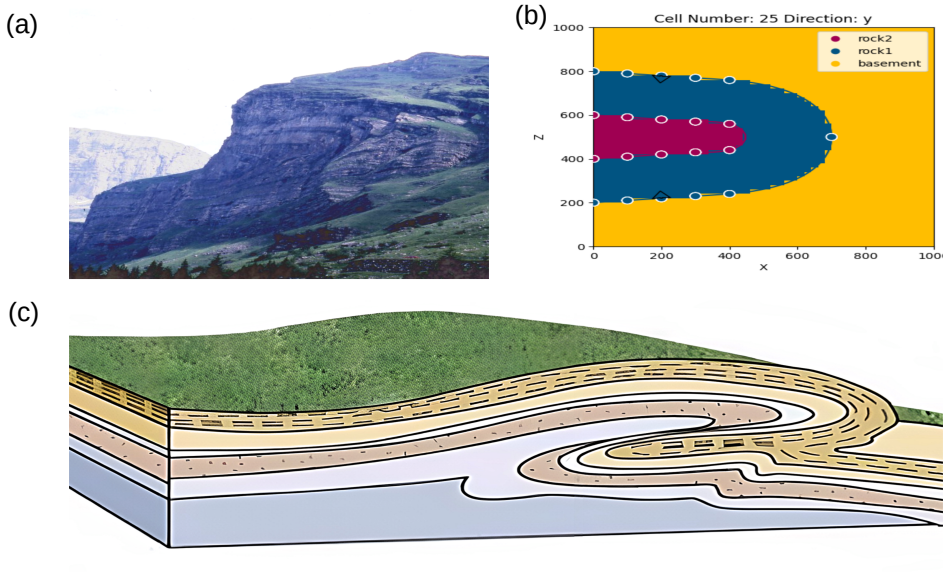

Figure 2-11: Outcrops showing a recumbent fold in Swiss Alps, Switzerland (Eastern Connecticut State University 2010) (a). Implicit recumbent fold geologic model created with GemPy (b). Graphical representation of model economic importance (c) (Eastern Connecticut State University 2010).

3. **Pinch Out Formation:** A pinch out is formed usually due to changes in facies, thereby forming a stratigraphic traps or as a result of structural deformation, for example pinching out of sand beds as a result of salt tectonics, thereby forming a structural trap (Bjørlykke 2015). A very good example of stratigraphic trap is the isolation of fluvial channel sandstone by impermeable clay-rich sediments resulting in the pinching out sandstone in a clay-rich sediments.

   Just like the anticline, pinch-out formation has economic benefit for example during hydrocarbon exploitation. Hydrocarbons that find their way into a pinch-out reservoir rock that is surrounded or isolated by non-permeable rocks can get stuck and trapped in place.

To demonstrate this, a pinch-out is constructed using the GemPy package.

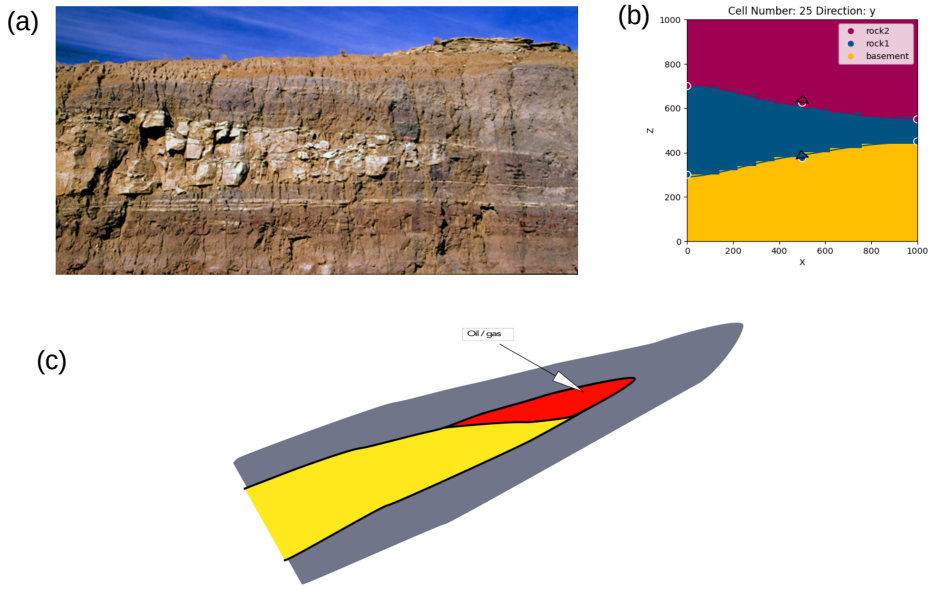

Figure 2-12: Outcrops showing a thinning out shale in the Jurassic Morrison Formation, Wyoming, United States (Sherman 2008) (a). Implicit pinch out geologic model created with GemPy (b). Graphical representation of model economic importance (c).

Uncertainty estimation is important and these above examples show how widely an uncertainty estimation process can be applied in geosciences which can range from the oil and gas exploration to tectonic activities investigation.

2-5-3 3D Explicit Surface Modeling

Like stated in section 2-5-2, the scalar field from GemPy models can be converted to surface points. The surface point is treated as a cloud points which is used in the resampling process. The resampling algorithm used is known as the cubic bspline algorithm. This algorithm is fully explained in section 2-1 and 2-2. The result of this modeling is a surface that represents our geological surface or interface and its control points. As stated earlier, the main reason we are integrating the implicit and explicit approach is to find a way to update our geological models during uncertainty estimation.

Control points in cubic spline determine the shape of a curve. Typically, each point of the curve is computed by taking a weighted sum of a number of control points. Each point's weight changes depending on the controlling parameter. For a curve of degree p, the weight of any control point is only nonzero in $p + 1$ intervals of the parameter space. Within those intervals, the weight changes according to a polynomial function (basis functions) of degree p. At the boundaries of the intervals, the basis functions go smoothly to zero, the smoothness being determined by the degree of the polynomial. It is important to note that the control points are equally weighted in this case, therefore they form uniform rational basis spline,

and then extended to a surface. Hence, for areas that requires more degree of polynomial, for example in a highly undulating surface at short distances, more control points are needed to accurately model such surface. This allows for more control over the shape of the curve without unduly raising the number of control points.

Adding more control points allows better approximation to a given curve or surface, however we want to limit their numbers as much as possible and get the more accurate surface as much as we can without having to deal with all the surface points of target of interest individually. This is why having a control point can be very useful. In the context of geological modeling, scale is also important. Therefore, the amount of control point required is a function of the numbers of geological structure or event we want to model. This, in the context of geomodeling and this project will be further expatiated in the following section 2-5-4.

The surfaces formed is transformed into a volume. Three (3) dimensional control points are quiet popular in 3D modeling, where they are often referred to as a point in a location in three (3) dimensional space.

2-5-4 Model Uncertainty Estimation

In the quest to estimate our uncertainties, there are priors needed. The control points generated above from the explicit modeling approach forms part of our prior. This is the basis of the integration of the implicit and explicit modeling approach for surface-based geological modeling.

To dissect this further, we know the control points controls our geological target which serves as our target geological formation. Therefore, making the control points the prior serves a lot of purpose. From previous work done by de la Varga & Wellmann (2016), they made use of priors that determines the z-component of surface depth values. As you can imagine, if all the surface points are utilized, it would be computationally too expensive and it gets more expensive when dealing with complex model. As rightly stated by de la Varga & Wellmann (2016), due to computational issues that result from multi-dimensional scenarios, it can be beneficial to reduce the number of priors. Therefore, we can make use of the control points without requiring too much computational powers relatively, and at the same time have the ability to access all the layer interface points in our geological model. It is also important to note that these control points affects the target surface points in all the three (3) dimensional space when dealing with three (3) dimensional interface and in all the two (2) dimensional space when dealing with a two(2) dimensional interface even though we vary the control points itself using only its z-component. Changing of the control point in that manner that it affects the different areas at different combinations in all the three (3) dimensional component will enable the optimally capturing of the salient properties attributed to the target formation, in this case geophysical properties in the target formation needed for uncertainty estimation. Figure 2-13 show the update of control points and its effect on the surface target.

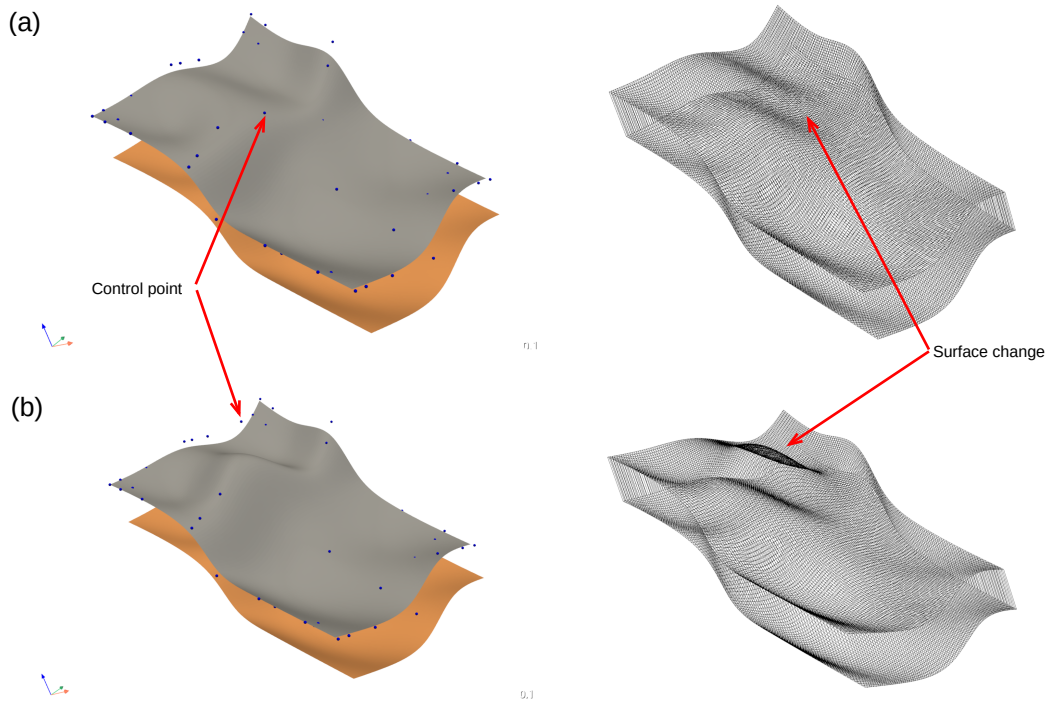

Figure 2-13: Change in layer surface due to an update of control point before (a) and after the control point has been moved (b).

Few priors, and by extension few control points thus make the uncertainty estimation computationally easy to handle. This is even more important when dealing with a complex geological surfaces that requires lots of surface points to model. This is where the geological aim and scale come in. According to Wellmann & Caumon (2018), one key aspect of geologic realism, is the question of how complex the geological settings is, and especially, how much of this complexity should be reflected in a subsequent geological model. These complexities can vary from those that results from the geological law that must be adhered to such as law of superposition to complexities arising from effect of tectonic activities on geologic layers after deposition such as faults and folds. Making use of control points allows us to model quiet complex geology just by relatively using a minute number of points when compared to using the surface points.

The control point relate with one another from the aspect of target surface. Depending on the scale, we might want to determine how many control points are required to model the geological interfaces for example, of the well tops and how will each control point affects the interfaces relative to the effect of other control points. This is because different part of geological layer reacts to tectonic activities differently depending on the distance between them. Therefore when estimating the uncertainties involves using a stochastic (implicit) approach that in turn involves making use of additional data such as geophysical data as a constraint, the number of control points will determine how well the salient features of the geologic settings can be captured.

It is worth knowing that few control points might be bad and too much control points might as well be bad. An example is modeling of geologic layer, depending on the scale or extent

of the geological layer, not all part of the formation will behave the same when subjected to tectonic activities. To model these changes in the presence of additional data expressed as likelihood function becomes important. In diagram 2-13, we see how modeling the interface of the two surface changes with change in the control points.

In as much as few control points is good, enough control points are equally important to allow enough flexibility during uncertainty estimation in the presence of additional information.

Starting from the frequentist perspective, we know if a known possible outcome is sampled from a normal distribution, the result tends towards the expected value $\mathbb{E}$ of the distribution (Davidson-Pilon 2015). The approach is utilised to help visualize possible realization of our model by making use of the Monte Carlo simulation. The point z-component is sampled from a normal distribution where the initial value is regarded as the ground truth. Several realizations was acquired and evaluated using the evaluation technique discussed in section 2-4-5.

The use of Monte Carlo simulation indicate that there are multiple geological realization that can be obtained and this gets even larger as the complexities increases.

In order to estimate the uncertainties, data driven modelling approach is utilized which allows for the consideration of various kinds of uncertainties and their implementation in several different ways. The uncertainties are included in the model by assigning uncertainties to the z-component of the control points which in turn helps to determine the geological layer interfaces in the 3D space. This is achieved by sampling from a probability distributions for each control points and then assigned to the z-component of these control points. As earlier stated, the type of distribution used has direct effect on the uncertainty quantification. In our case, a normal distribution is where the values are sampled from, and because we are dealing with multi variables, we simply sample from a multivariate normal distribution.

2-5-5 Setting the Likelihoods in 3D Models

To estimate the uncertainties involved using Bayesian inference, constraint(s) is required. These are expressed as likelihoods functions. These constrain serves as a probabilistic input into the model. In geomodelling, this additional information might have been derived from, for example seismic observations or from wire-line log data. Constraint(s) could also be derived from knowledge of the geological history of the formation or physical properties of the target formation such as thickness. Assuming there are some form of relationship between the target formation and these additional information, the uncertainty can be reduced by incorporating these additional information through the assessment of their likelihood in the target formation (Wellmann & Caumon 2018).

In the case of this project, the likelihoods are expressed as the z-components of the control points for the top and bottom layer, and gravity inversion of the target formation constructed using these control points. This is based on the fact that, using the density of the target formation, we assume a possible correlation between the output of a gravity inversion in the area occupied by the target formation, in the subsurface and the z-component of the control points used to construct the target formation.

These constraints are then implemented based on probability distributions, basically as a multivariate normal distribution from which the priors are sampled from. The multivariate

normal covariance matrix is symmetric positive semi-definite which makes Cholesky factorization one of the best and efficient way to parameterize the multivariate normal distribution. Hermitian, positive-definite matrix can be decomposed into the product of a lower triangular matrix and its conjugate transpose using the Cholesky decomposition (Muschinski et al. 2022). This makes the Cholesky decomposition very efficient, not only to easily create random covariance matrices, but also makes the parameterization of the multivariate normal PDF more efficient. In essence, the cholesky helps us to parameterize the covariance matrix such that the multivariate sampling from the distribution has some form of correlation by ensuring a positive definite and well defined PDF (Muschinski et al. 2022).

2-5-6 Determining Formation Volume

One of the aim of this project is to evaluate the integration of explicit and implicit geological modeling approach in constructing surface based models. To carry this task out, certain process are designed such as obtaining the posterior distribution of the control point and evaluate this distribution relative to one another. Another aspect is to actually see if a random value sampled from a normal distribution of wider standard deviation can be assigned to these control points, optimizing them using the same Bayesian inference approach and made to converge. The term converge simply means, the control points with their respective unique distribution can find themselves back to positions that produce the original geological formation, through introduction of additional data, in this case, gravity measurement, as constraints. We basically want to prove that our model can be learned and the original layer can be obtained.

The algorithm is allowed to learn the prior distribution of the control points of the two bounding surfaces to produce a posterior distribution given the constraint. Hamiltonian Monte Carlo, a Markov Chain Monte Carlo MCMC algorithm is used to obtain the posterior distribution while the Maximum a Posteriori MAP estimation is done with the same sampling process and optimized using an Adam optimizer. The error is estimated to further analyze the uncertainty when compared with the ground truth. The idea is to see how integration of the implicit and explicit can help predict the ground truth.

To achieve the above process, voxelization process is needed for the evaluation. Section 2-3-2 explains how the voxelization of target formation is carried out. A grid of voxels is constructed with preferred resolution. The points bounded by the target formation, for example an anticlinal formation, for every iteration is assigned voxels of the same value, indicating points of the same formation. Therefore, as the control points change, due to sampling, the position of the voxels, using their indices, also changes.

The voxels is also used to house the additional data acting as constraint. This is set as likelihoods as explained in section 2-5-6.

Therefore, for every iteration of the Bayesian inference process, error are estimated and posterior distribution are retrieved for further evaluation.

Model evaluation below in section 2-5-7 makes use of the resulting voxelized grid for assessment by calculating the probability and entropy of the combined voxelized grid.

2-5-7 Model Uncertainty Evaluation

After performing Bayesian inference, the posterior distributions and model realizations must be assessed. We can do this by looking at sample posterior distributions of single parameters or combinations of parameters. The full set of approved 3D geological models must be evaluated in terms of uncertainty quantification and visualization. Furthermore, a key component of this effort is assessing how successfully the explicit-implicit strategy was integrated. All of these can be established by making use of the posterior distribution, the probability and the information entropy.

For the visualization of the result, section 2-4-5 fully explains the process.

In general, the diagram 2-14 below, in the context of this project summarizes the Bayesian inference steps.

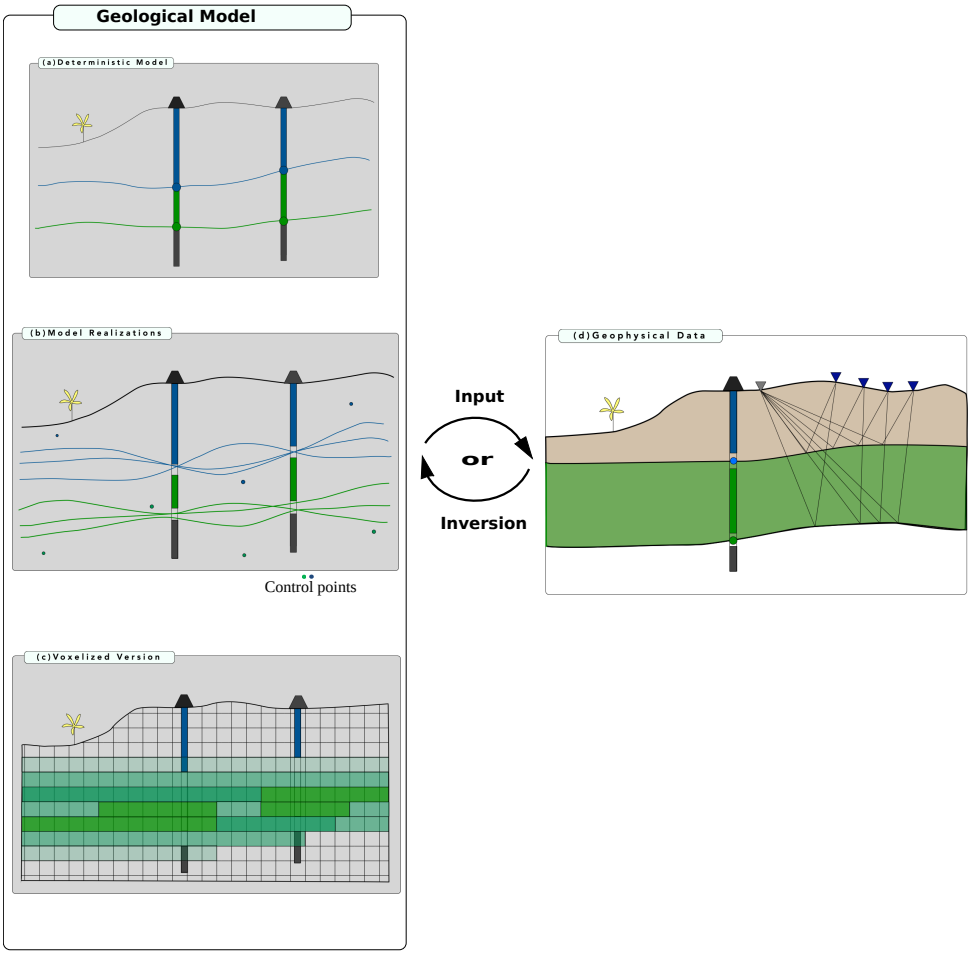

Figure 2-14: The fundamental approach behind most of methods for uncertainty quantification in geological models is to start from one deterministic representation (a) to multiple and plausible geological realizations (b). These realizations are eventually voxelized(c). Then, by subjecting these geological models to supplemental geophysical probability criteria (d), uncertainty is decreased (Wellmann & Caumon 2018)

Chapter 3

Results

In this chapter the results produced with the implemented algorithm to the theoretical 3D geological model will be shown. The results for each and every theoretical model can be divided to three (3) major parts. They are:

1. Prior Model
2. The Posterior Model
3. The Maximum a Posteriori Model

For each theoretical model, the prior involves MC simulation of 50 samples and the posterior MCMC sampling simulation involving $25,000$ sampling steps with a burn-in phase of 1000 iterations were carried out. For the optimization of the posterior distribution, the Adam algorithm was used with maximum iteration set at $10,000$ while the loss tolerance was set between 0 and 10^{-13}. The probability and information entropy were used for the visualization of result. The plot of both the prior and posterior distribution, the probability and information entropy plot were used as the assessment criteria for the HMC while the convergence of the optimization was assessed based on the loss plot. Specific criteria for each model are declared in their respective sections below.

3-1 3D Model I: An Anticline

To establish how efficient our combined approaches are, a prior distribution, a posterior distribution and maximum a posteriori estimate were created for the anticline fold model.

3-1-1 Prior Model

First, for the prior distribution, the original control points were used to generate our likelihood which in turn, serves as the constrain to our model. These original respective positions of

our control points were then subjected to a Monte Carlo simulation to generate a normal distribution from which a random sample was made. This sample of control points serves as our prior from which the prior normal distribution was obtained through another Monte Carlo simulation with no inclusion of likelihood function.

To set up the prior distribution, the uncertainties of the parameter used is given in the table 3-1 below:

Uncertainty: Control Point Z-position (normal distribution)		
	μ [m]	σ [m]
Layer Top	$\equiv$	5
Layer Bottom	$\equiv$	5

Table 3-1: Defining the prior parameter uncertainties by normal distributions with their standard deviations σ and their respective values as their means μ equivalent, thus represented as $\equiv$

The prior distribution is given by the figure 3-1 below:

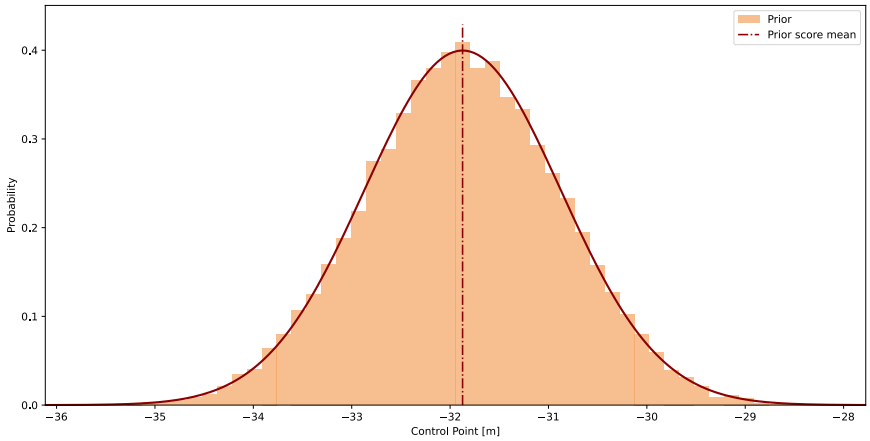

Figure 3-1: Prior distribution of the control point probability for the 3D model

To visualize 50 samples from this prior distribution, probability and information entropy in figure 3-2 below are used to describe the uncertainty involved in the model.

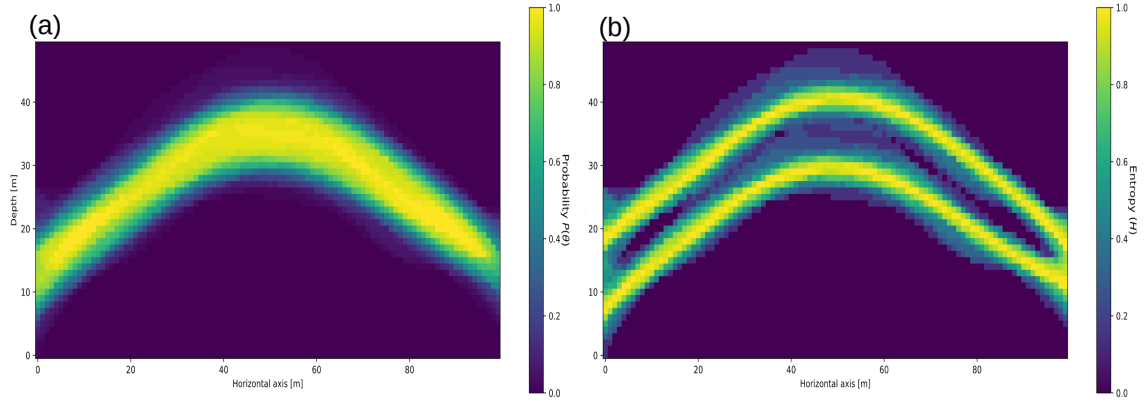

Figure 3-2: Visualization of Probability (a) and Information Entropy (b) in Prior model.

In figure 3-2 above, the high uncertainties involved in the prior distribution is located at the boundary of the anticlinal fold while it can clearly be seen that large portion of the distribution is found in a range that include the volumetric space enclosed by the top and bottom layer of the anticlinal fold, depicting zone of high probability.

3-1-2 Posterior Model

For the derivation of this posterior model, we included Gravity likelihood for the anticlinal fold probabilistic model. This was defined in a way that reinforced the probabilities of the relevant lithology.

The derived gravity value from every samples from prior distribution is then cross examined by using the likelihood through the calculation of the log probability.

Likelihoods: Gravity (normal distributions)		
	μ [m]	σ [m]
Gravity	$\equiv$	0.001

Table 3-2: Bayesian inference based on Gravity as likelihoods defined by normal distributions with their standard deviations σ and their respective values as their means μ equivalent, thus represented as $\equiv$

For the posterior distribution, at every iteration, we need to confront our model with the likelihoods. Our anticlinal fold gravity data expressed as likelihood function is given by figure 3-3.

The combined prior and posterior distribution is then compared by observing the means of their distribution as seen in figure 3-4.

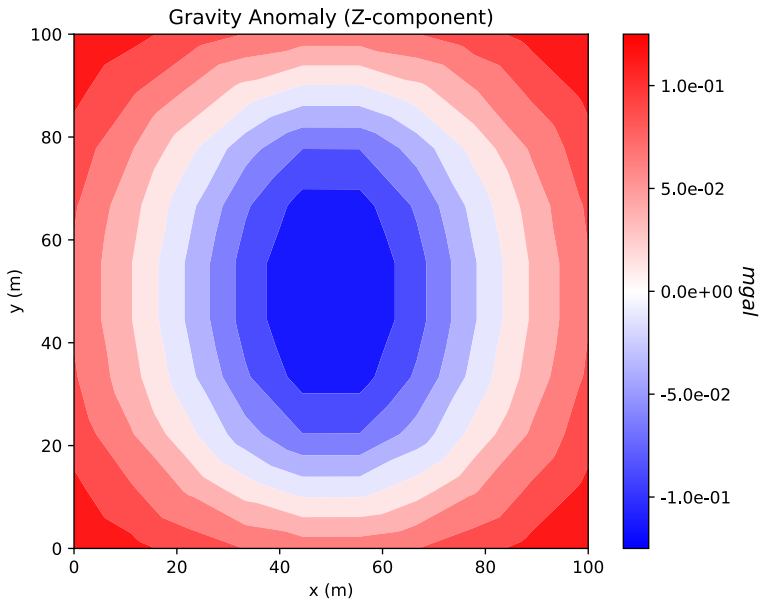

Figure 3-3: Anticline fold gravity simulation result

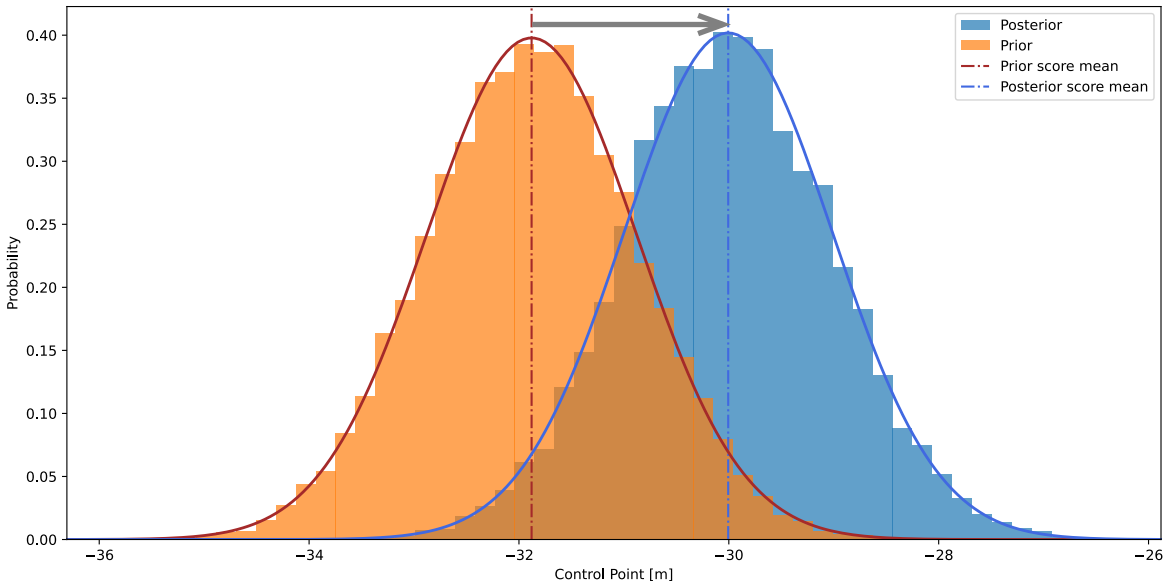

Figure 3-4: Prior and posterior control point probability distributions

We can see the change in the mean location from the prior distribution to posterior distribution indicated by the arrow.

3-1-3 Maximum a Posteriori Estimation

The maximum a posteriori was designed to get the point estimate. This way we can determine how the model converged and how well it converged around the tolerance range.

Figure 3-5 below shows the loss plot. The loss converged around 6000 iterations.

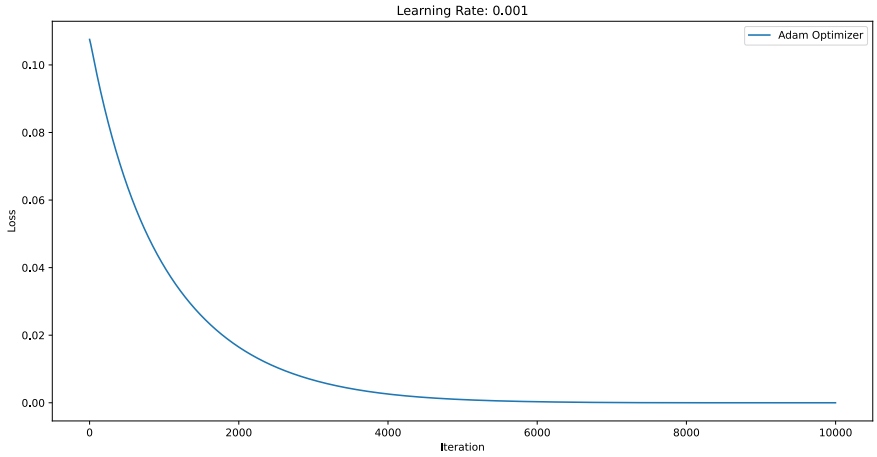

Figure 3-5: Optimization Loss Plot

To visualize 50 samples from this posterior distribution, probability and information entropy below describes the uncertainty involved in the model. The result shows our approaches worked really well as the entropy basically goes to zero.

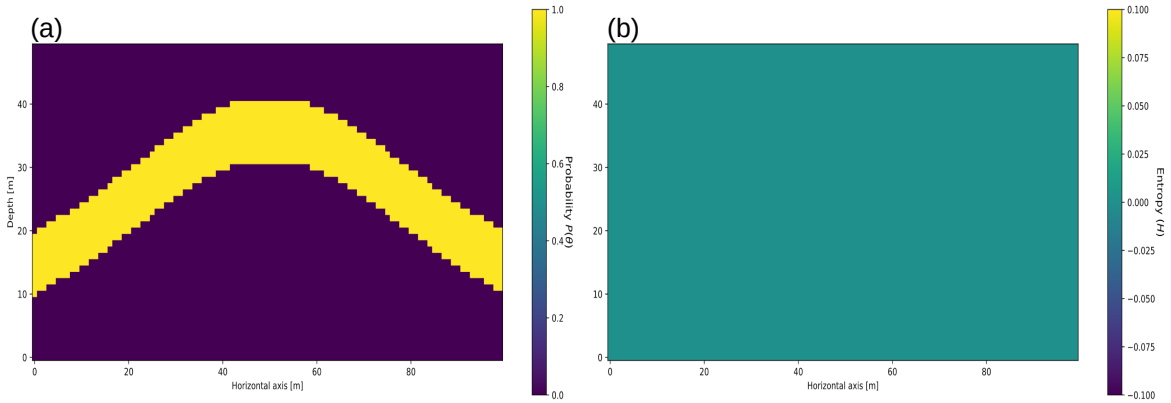

Figure 3-6: Visualization of Probability (a) and Information Entropy (b) in Posterior model.

3-2 3D Model II: A Recumbent Fold

To establish how efficient our combined approaches are, a prior distribution, a posterior distribution and maximum a posteriori estimate were created for the recumbent fold model.

3-2-1 Prior Model

First, for the prior distribution, the original control points were used to generate our likelihood which in turn, serves as the constrain to our model. These original respective positions of our control points were then subjected to a Monte Carlo simulation to generate a normal distribution from which a random sample was made. This sample of control points serves as our prior from which the prior normal distribution was obtained through another Monte Carlo simulation with no inclusion of likelihood function.

To set up the prior distribution, the uncertainties of the parameter used is given in the table 3-3 below:

Uncertainty: Control Point Z-position (normal distribution)		
	μ [m]	σ [m]
Layer Top	$\equiv$	5
Layer Bottom	$\equiv$	5

Table 3-3: Defining the prior parameter uncertainties by normal distributions with their standard deviations σ and their respective values as their means μ equivalent, thus represented as $\equiv$

The prior distribution is given by the figure 3-7 below:

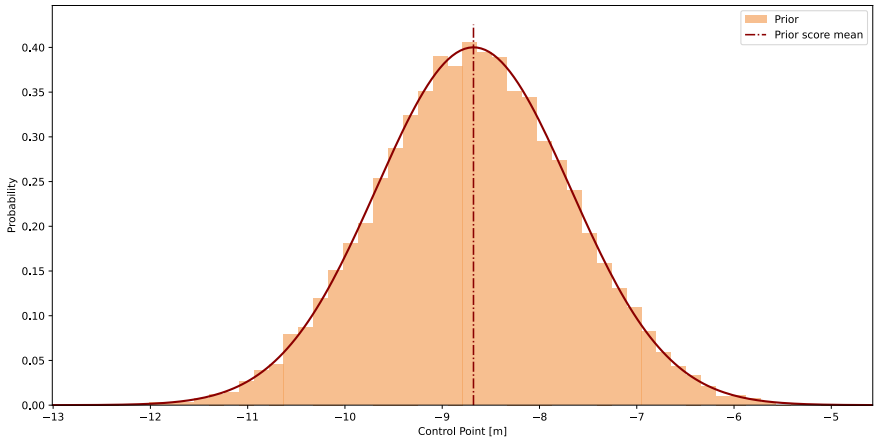

Figure 3-7: Prior distribution of the control point probability for the 3D model

To visualize 50 samples from this prior distribution, probability and information entropy in figure 3-8 below are used to describe the uncertainty involved in the model.

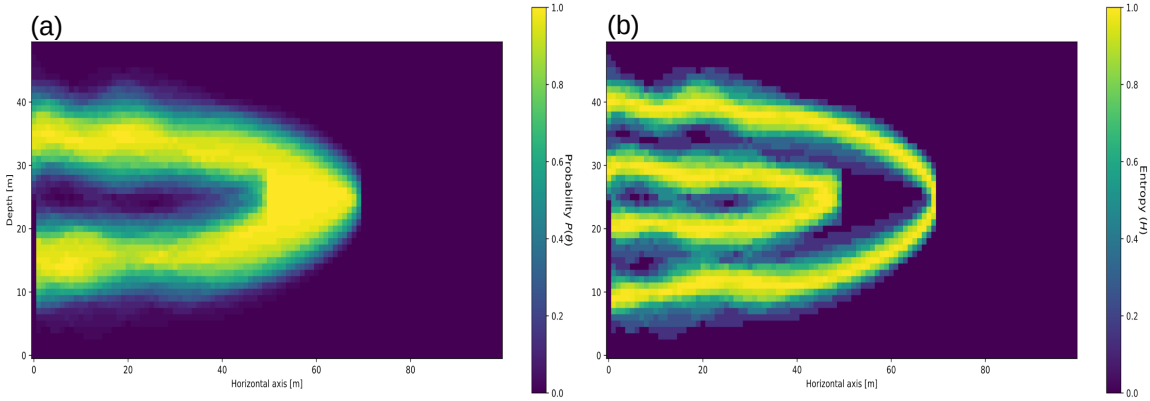

Figure 3-8: Visualization of Probability (a) and Information Entropy (b) in Prior model.

In figure 3-8 above, the high uncertainties involved in the prior distribution is located at the boundary of the recumbent fold while it can clearly be seen that large portion of the distribution is found in a range that include the volumetric space enclosed by the top and bottom layer of the recumbent fold model, depicting zone of high probability.

3-2-2 Posterior Model

For the derivation of this posterior model, we included Gravity likelihood for the recumbent fold probabilistic model. This was defined in a way that reinforced the probabilities of the relevant lithology.

The derived gravity value from every samples from prior distribution is then cross examined by using the likelihood through the calculation of the log probability.

Likelihoods: Gravity (normal distributions)		
	μ [m]	σ [m]
Gravity	$\equiv$	0.001

Table 3-4: Bayesian inference based on Gravity as likelihoods defined by normal distributions with their standard deviations σ and their respective values as their means μ equivalent, thus represented as $\equiv$

For the posterior distribution, at every iteration, we need to confront our model with the likelihoods. Our recumbent fold gravity data expressed as likelihood function is given by figure 3-9 below:

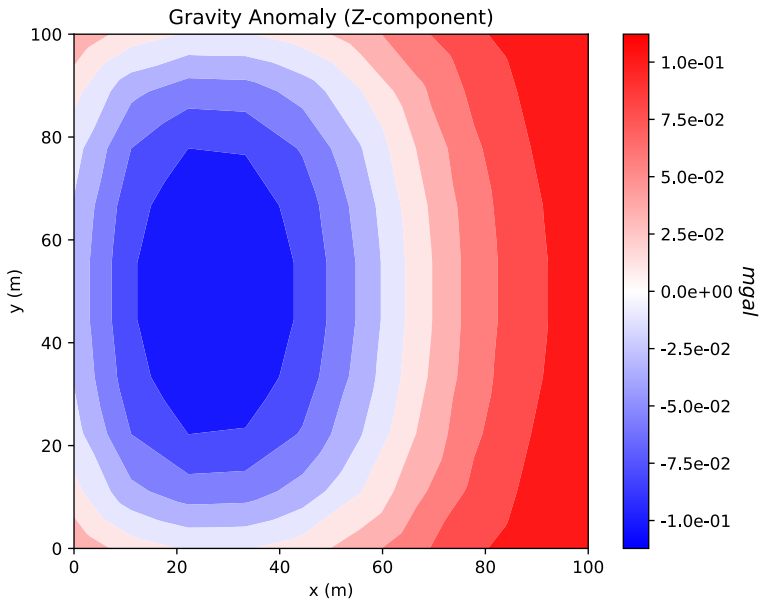

Figure 3-9: Recumbent fold gravity simulation result

The combined prior and posterior distribution is then compared by observing the means of their distribution as seen in figure 3-10 below:

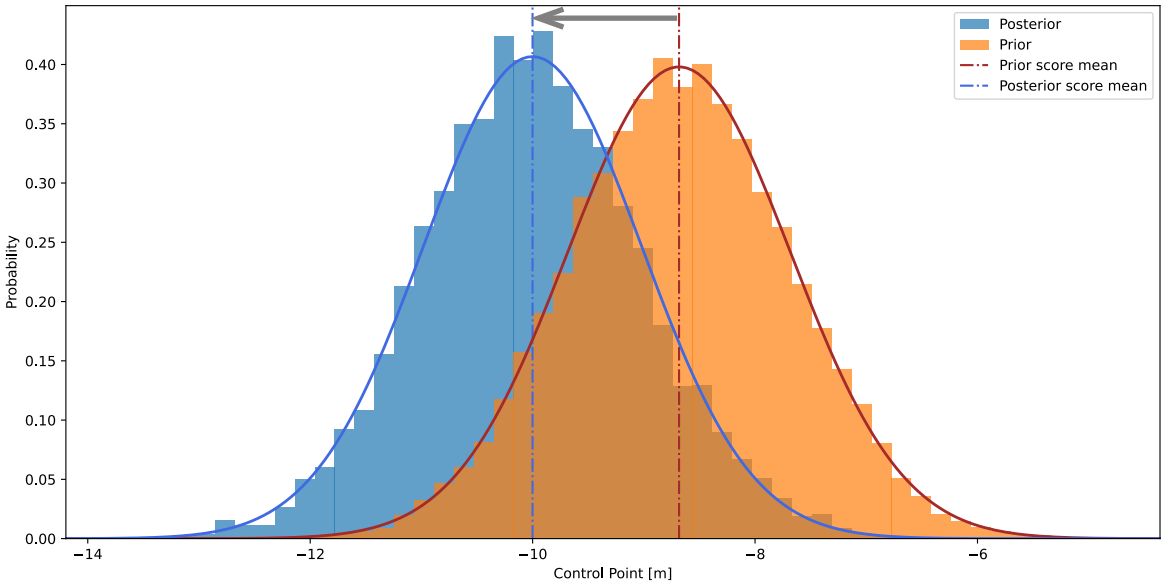

Figure 3-10: Prior and posterior control point probability distributions

We can see the change in the mean location from the prior distribution to posterior distribution indicated by the arrow.

3-2-3 Maximum a Posteriori Estimation

The maximum a posteriori was designed to get the point estimate. This way we can determine how the model converged and how well it converged around the tolerance range.

Figure 3-11 below shows the loss plot. The loss converged around 6200 iterations.

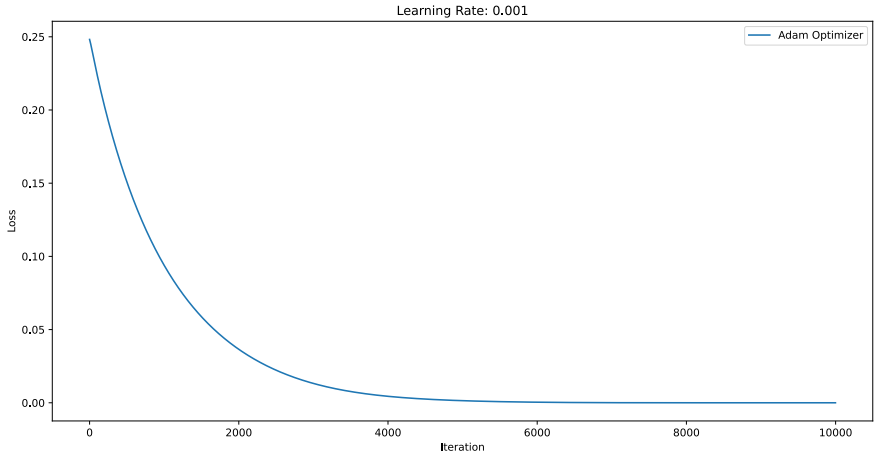

Figure 3-11: Optimization Loss plot

To visualize 50 samples from this posterior distribution, probability and information entropy below describes the uncertainty involved in the model. The result shows our approaches worked really well as the entropy basically goes to zero.

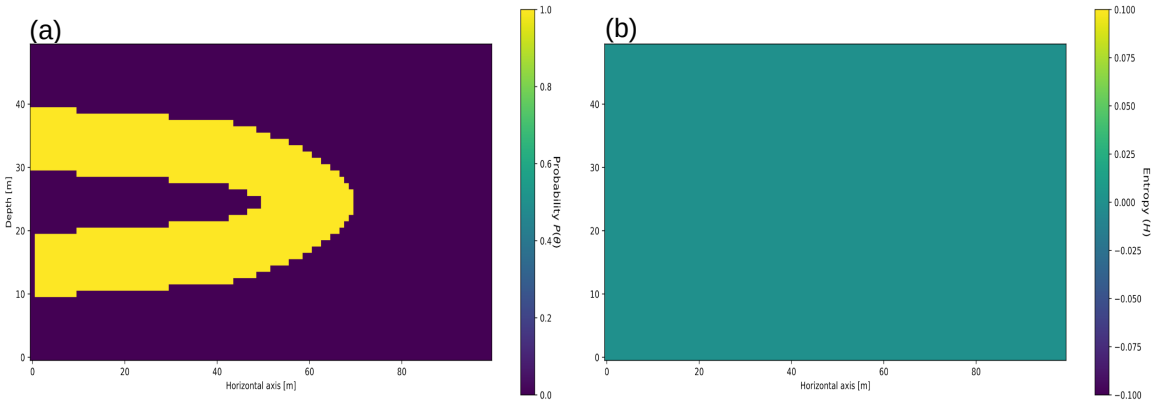

Figure 3-12: Visualization of Probability (a) and Information Entropy (b) in Posterior model.

3-3 3D Model III: A Pinch Out Formation

To establish how efficient our combined approaches are, a prior distribution, a posterior distribution and maximum a posteriori estimate were created for the pinch out model.

3-3-1 Prior Model

First, for the prior distribution, the original control points were used to generate our likelihood which in turn, serves as the constrain to our model. These original respective positions of our control points were then subjected to a Monte Carlo simulation to generate a normal distribution from which a random sample was made. This sample of control points serves as our prior from which the prior normal distribution was obtained through another Monte Carlo simulation with no inclusion of likelihood function.

To set up the prior distribution, the uncertainties of the parameter used is given in the table 3-5 below:

Uncertainty: Control Point Z-position (normal distribution)		
	μ [m]	σ [m]
Layer Top	$\equiv$	5
Layer Bottom	$\equiv$	5

Table 3-5: Defining the prior parameter uncertainties by normal distributions with their standard deviations σ and their respective values as their means μ equivalent, thus represented as $\equiv$

The prior distribution is given by the figure 3-13 below:

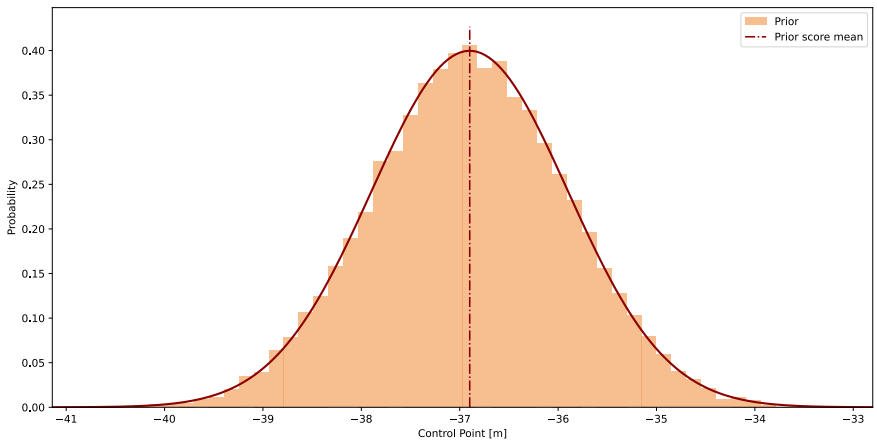

Figure 3-13: Prior distribution of the control point probability for the 3D model

To visualize 50 samples from this prior distribution, probability and information entropy in figure 3-14 below are used to describe the uncertainty involved in the model.

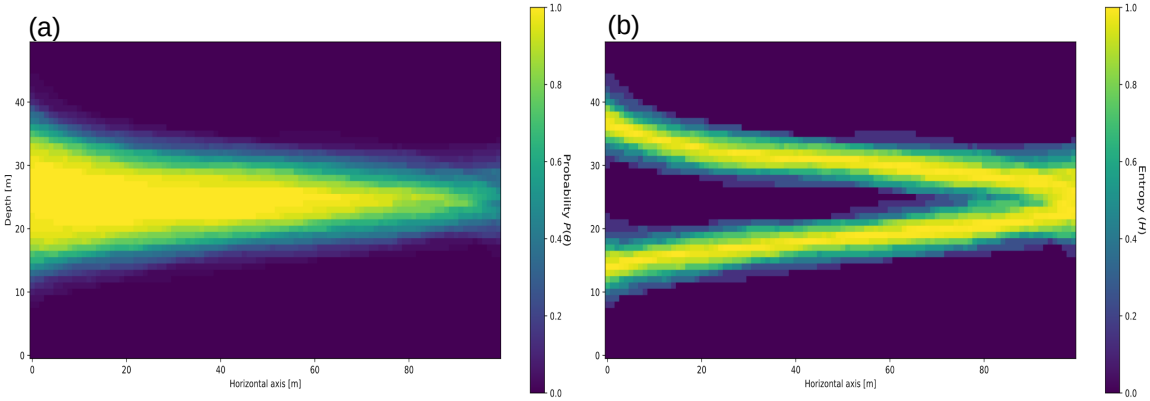

Figure 3-14: Visualization of Probability (a) and Information Entropy (b) in Prior model.

In figure 3-14 above, the high uncertainties involved in the prior distribution is located at the boundary of the pinch-out formation while it can clearly be seen that large portion of the distribution is found in a range that include the volumetric space enclosed by the top and bottom layer of the pinch-out model, depicting zone of high probability.

3-3-2 Posterior Model

For the derivation of this posterior model, we included Gravity likelihood for the pinch-out probabilistic model. This was defined in a way that reinforced the probabilities of the relevant lithology.

The derived gravity value from every samples from prior distribution is then cross examined by using the likelihood through the calculation of the log probability.

Likelihoods: Gravity (normal distributions)		
	μ [m]	σ [m]
Gravity	$\equiv$	0.001

Table 3-6: Bayesian inference based on Gravity as likelihoods defined by normal distributions with their standard deviations σ and their respective values as their means μ equivalent, thus represented as $\equiv$

For the posterior distribution, at every iteration, we need to confront our model with the likelihoods. Our pinch-out gravity data expressed as likelihood function is given by figure 3-15 below:

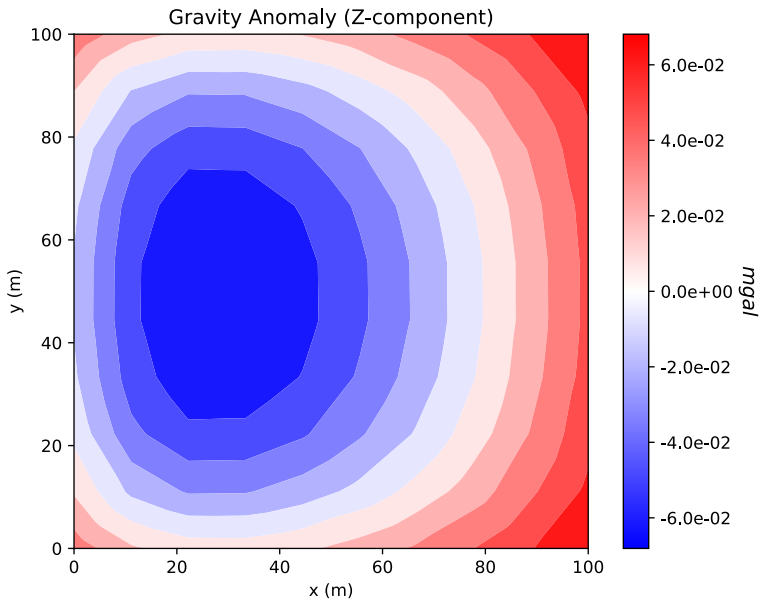

Figure 3-15: Pinch out gravity simulation result

The combined prior and posterior distribution is then compared by observing the means of their distribution as seen in figure 3-16 below:

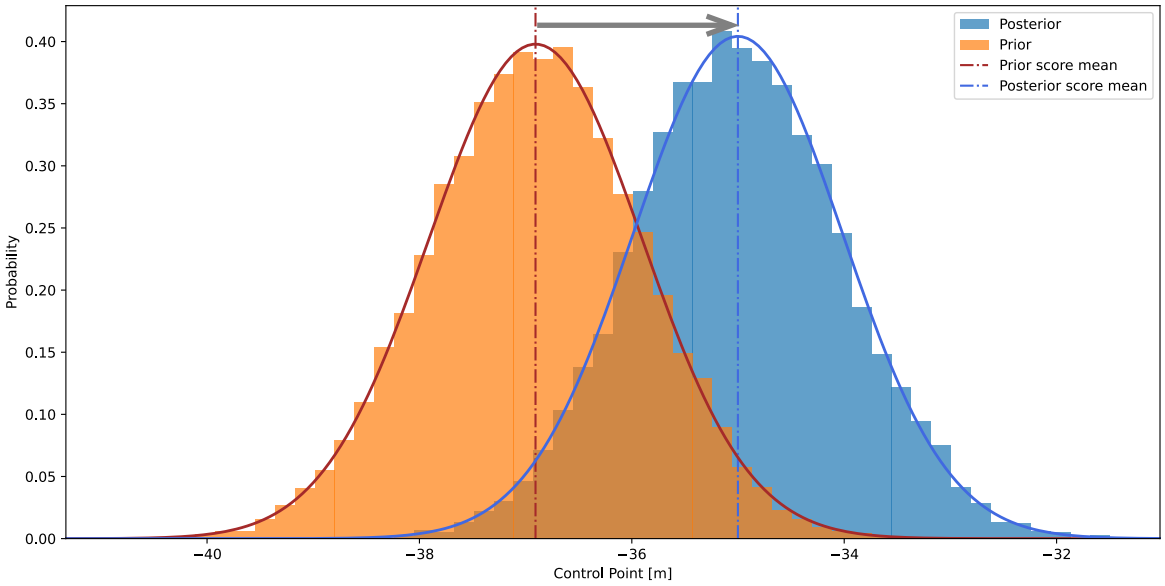

Figure 3-16: Prior and posterior control point probability distributions

We can see the change in the mean location from the prior distribution to posterior distribution indicated by the arrow.

3-3-3 Maximum a Posteriori Estimation

The maximum a posteriori was designed to get the point estimate. This way we can determine how the model converged and how well it converged around the tolerance range.

Figure 3-17 below shows the loss plot. The loss converged around 5900 iterations.

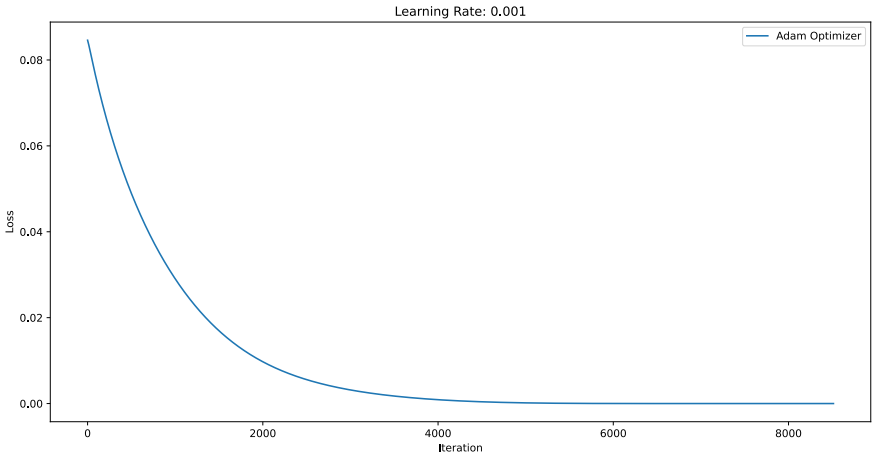

Figure 3-17: Optimization Loss plot

To visualize 50 samples from this posterior distribution, probability and information entropy below describes the uncertainty involved in the model. The result shows our approaches worked really well as the entropy basically goes to zero.

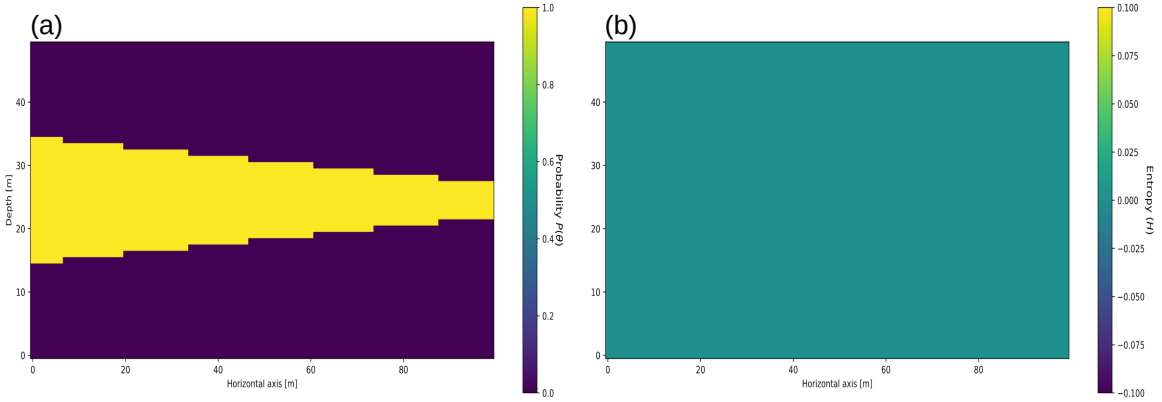

Figure 3-18: Visualization of Probability (a) and Information Entropy (b) in Posterior model.

Chapter 4

Discussion

The goal of this research work is to make use of an integrated implicit-explicit approach to construct a surface-based geological model through a re-sampling process, then set up this integration as a Bayesian inference problem in order to be able to make a more realistic and accurate economic decision through the process known as probabilistic machine learning by trying to estimate and evaluate the possible uncertainties involved in the geological model. Our hope is to see how additional information expressed as likelihood functions through the use of Bayes theorem can be added to our model simulation in order to influence our final model positively.

We understand that different industries involved in geological modeling, for example in the hydrocarbon industry, face high risks and uncertainties depending on how complex the model is. Because this process has direct economic implications, our goal is to extend our integrated approach to a point where it could be used to reduce the uncertainties involved and potentially reduce risks and give high rewards in terms of economic benefit. This goal was achieved, although there is always room for improvement. All these will be discussed in this section.

In general, two major tasks were achieved in this project:

1. The integration of an implicit-explicit approach to create a flexible surface-based geological model
2. The estimation and evaluation of uncertainties in surface-based geological models

In this section, we will discuss certain issues that make our approach very useful, the limitations, and conclude by providing an overview of options to properly test and verify the presented results, as well as suggestions for improvements.

To implement this integrated idea, theoretical geological models were used. They include an anticline fold, a recumbent fold, and a pinch-out formation. The idea here is to show that through probabilistic machine learning, using the explicit algorithm as the mathematical function and adding additional information, we can reduce the uncertainties attached to

these subsurface geological models, which in turn improves the decision-making process. A central part of our work was comprised of the development of algorithms capable of resampling the surface points of our geological models created by an implicit modeling approach. This re-sampling process comes with a set of control points that define every layer of interest. Different prior and posterior probability distributions for control points were generated to form the basis for estimating the uncertainties related to the layer of interest.

4-0-1 Surface-based modeling

There are different approaches that can be used for geomodeling, which could be surface-based or a point grid approach as described by Caumon et al. (2009) and Liu et al. (2020). The advantages of the surface-based modeling approach include its stronger linkage with geological concepts and interpretations, improved geometrical accuracy, and greater computational efficiency (Jacquemyn et al. 2019*a*).

The use of cubic spline surfaces in surface-based modeling preserves the direct conceptual link between the geological interpretation of heterogeneity and the final reservoir model. Our theoretical models as seen in figure 2-10, 2-11 and 2-12 show how surface-based modeling can be used to portray variety of geological model in a more geometrically realistic way. With the help of the created control points, a surface representation preserves geometrical precision to whatever level of detail that is required (Jacquemyn et al. 2019*a*). Explicitly describing heterogeneities several orders of magnitude smaller than the complete model volume only adds to the model's complexity locally, where surfaces that capture the small-scale heterogeneities are introduced, but the rest of the model remains unaffected. By utilizing surface terminations and the adaptability of control points, cubic spline surfaces can be used to obtain a greater level of geometric accuracy that captures the intended connections between one surface and the other. This approach is way better considering typical yet complex structure configurations that are difficult or impossible to portray using conventional pillar grid approaches. This is perhaps the best example of the benefit of employing parametric surfaces, such as cubic spline surfaces, to accurately express geological geometries (Qu et al. 2015).

The illustration in figure 2-4 demonstrates how little data is required to build a parametric surface with a comparatively complex geometry. This demonstrates its high construction efficiency and ease of manipulation. When producing many model realizations that are sampled from the same input data distributions, this computational efficiency is a considerable benefit over traditional reservoir modeling methodologies (Jacquemyn et al. 2019*a*). Because surfaces don't need to be discretized or gridded, they may be created rapidly and inexpensively. When complicated geometries are modeled with a cubic spline representation for curves and surfaces, the detail of one surface does not affect the detail of other surfaces.

The hierarchical arrangement of surfaces governs the interaction between surfaces and allows us to switch on and off specific heterogeneities or length scales because no modification of an underlying grid is needed (Jacquemyn et al. 2019*a*). Consequently, the computational efficiency of cubic spline representation makes it possible to quickly generate multiple realizations of surface-based models and to test the impact of different geological scenarios. Because cubic spline surfaces are computationally efficient, generating multiple realizations is not time-consuming. For example, the top and bottom surfaces of figure 2-4, which involve 160,000 surface points each, took 2.2 seconds to construct on a standard workstation. In

contrast, constructing and exporting cornerpoint-gridded or triangulated versions of surfaces takes a longer period of time, depending on the resolution of discretization being used.

In this project, we also made use of a gridded system, but that was mainly for the evaluation of uncertainties after the surface-based model had been created. We have the flexibility to determine the level of resolution of the grid for our uncertainty evaluation. Therefore, it is better to create a surface-based geological model, capture all the uncertainties, complexities, and heterogeneity attached to the model, and use the flexibility of the control point to update the model. Then we can convert these captured heterogeneities into a grid system, mainly for the uncertainty evaluation procedure. By doing that, our original model has not been in any way changed or affected. Using a grid system to create a geological model from the start, on the other hand, will prevent the capture of the heterogeneities involved. Another aspect that makes this approach important, particularly to us, is the ability to update the model through the use of its control points when estimating the uncertainty involved. This aspect is further discussed in the section below.

4-0-2 Uncertainty Estimation

As discussed in the above section, the presence of control points allows us to manipulate the surface of the geological model. This has really proven very efficient in different ways.

To estimate the uncertainties related to our target formation, several realizations that cover all possible scenarios in the subsurface of this target are needed. These scenarios have spatial implications, especially when dealing with formations in three-dimensional space. Looking at the computational cost it requires to create these realizations with way fewer control points compared to using the surface points of the model, we can manipulate and update the surface of our model more easily and more rapidly. This manipulation and updating of the geological surface can be done in all three dimensions using these control points. Furthermore, when we consider the alternatives, for example, the use of a relatively few surface points only in z-direction, it becomes evident that not all the intricacies of our geological model will be captured. This is because our subsurface target has a three-dimensional spatial interaction with other subsurface materials. Also, as we will see in the choice of likelihoods, depending on what properties we want to use, the use of modeling methods that will allow the capture of the details of our model in three dimensions has proven to be beneficial.

Having outlined the importance of using control points, it is also important to illustrate the importance of optimizing the number of control points we make use of. As our models become complex, the number of control points to replicate the complexity becomes higher. Hence, the termination and optimization of control points become more important.

4-0-3 Advantages of Bayesian Inference Method

The Bayesian inference approach has been around for a very long time, and it has proven to be very useful in tackling problems that revolve around decision making.

The Bayesian inference approach provides a natural and principled way of combining prior information with data within a solid decision-theoretic framework. You can create a prior distribution for future analyses by using past information about a parameter. When new

estimates that are better than the previous ones are made available, estimates that were previously regarded as posterior can be used as prior estimates. Therefore, the Bayesian inference approach allows us to quantify evidence and track its progression as new data are made available.

A wide range of models can also be constructed in a flexible and easy environment using the Bayesian inference approach. Input data can be regarded as deterministic or stochastic. It allows us to make use of different types of distribution that may be peculiar to the problem. Even though we often make use of a normal distribution, there are scenarios where we might need to use other forms of distribution. For example, estimating the rate of gas production can make use of an exponential distribution.

Expressing our uncertainty problem as a Bayesian inference problem is not complete without the presence of additional information to serve as constraint(s). Constraints are important because they are defined as likelihoods. This is the principle implemented in MCMC. On the other hand, MC simulation, for example, in reservoir estimation and risk assessment, has become common and is often used in combination with decision trees (Stamm et al. 2019). Davidson-Pilon (2015) refers to this approach as a "frequentist approach," in which probabilistic modeling is mostly used to obtain the best estimates in the form of means (Stamm et al. 2019). The idea is that by taking enough samples of a known distribution due to expert knowledge, the most likely event will occur at the center of the distribution. This practice does not harness the full potential of such a probabilistic distribution, and much of the inherent information potential is discarded. This is because we know that expert knowledge could also be biased. Furthermore, the difficulty of analysis using frequentist approaches usually increases depending on the complexity of the model because it requires lots of sampling to be made, enough to make a decision, thus becoming very expensive in terms of computational power and time required. The Bayesian inference approach, on the other hand, takes into account the full probability distribution and enables the inclusion of various conditions in the process of finding an optimal estimate. With the inclusion of additional information in a MC simulation, the ability to automatically find an optimal decision evidently increases, and the computational power and time requirement reduces, mainly because we can monitor its success as the simulation progresses.

Because we could add many constraints that are related to the data and, by extension, to themselves, the Bayesian inference approach provides inferences that are conditional on the data and are exact, without reliance on approximation. Therefore, inference from a small sample proceeds in the same manner as if one had a large sample. This is because the likelihood principle is well followed and its successes can be monitored. Therefore, if two distinct sampling designs yield proportional likelihood functions for parameter θ, then all inferences about θ should be identical from these two designs.

Applying Bayesian inference approaches to several aspects of geomodeling, such as structural geological modeling, is intended to support decision-making in the earliest stages of a subsurface investigation before the intrusive method is applied. This makes it very useful economically, as it can help prevent economic risks.

4-0-4 Choice of Likelihoods

As earlier stated in the previous chapter, we cannot overemphasize the importance of likelihood. However, the choice of likelihood to use is equally as important and must be taken into consideration. Constraints to be set up as likelihood function could be scalar quantities, for example, volume or thickness. These quantities have no spatial relationship with the target formation when dealing with geological models. They are values that collectively represent the whole formation. Therefore, we could have several geometrical realizations of our model that are wrong but conform to our constraints due to possibility of having different changes in different part of our target formation. This implies that we could have a scenario where the inclusion of additional data, depending on its type, could reduce uncertainty in one part of the spatial model but not all part of our spatial geological model.

Research carried out by Stamm et al. (2019) indicates that the impact of additional information on decision uncertainty, induced by Bayesian inference, is not necessarily strictly aligned with the change in uncertainty regarding model parameters and their combinations. They used thickness and volume as constraints. While there is improved certainty about their reservoir thickness, it had little to no impact on their decision-making (Stamm et al. 2019).

This same principle applies to using borehole data as likelihoods, for example, porosity and permeability. These properties can be ascertained in the well bore, perhaps through core sampling, and, to some extent, in the vicinity of the well bore. However, the uncertainty of these petrophysical properties varies greatly beyond the wellbore horizontally. On one part, this could be due to the effect caused by the process of obtaining these core samples, on the other part, might be due to heterogeneous nature of the subsurface itself. Therefore, using them as constraints becomes problematic, especially in modeling subsurface formation. They further identified two major aspects that are important to look at depending on the type of likelihood being used. They are:

1. where the model uncertainties are reduced
2. which outcome is enhanced in terms of probability

All these problems are inherent if our model has spatial properties and the likelihood used does not have direct spatial properties. Making use of spatial constraints, for example, seismic data or gravity, a geophysical data that has spatial properties and covers the area occupied by the target formation, is therefore important and better. This is evident in our model result. The algorithm was able to learn the model and determine where the uncertainties are and finds a way on how to reduce them.

4-0-5 Scalability and limitations of this methodology

So far, we have seen that some of our approaches to geomodeling and uncertainty evaluation and reduction have worked. This is pleasing to see, but we know that there can be room for improvement. To start with, we need to emphasize that the models constructed in this work are theoretical models. Although, they represent features that can be found in the subsurface, they tend to be less heterogeneous when compared to what is attainable in the subsurface.

It is therefore important to prepare for what is more likely in terms of heterogeneity, so that we can be sure our approach can work at scale.

To start with, the area of control point optimization would be a very good place to start. Control points can be made more robust so that when dealing with more complex models, control points concentrate only where needed. Combining this possibility with the ability to terminate layers will create an even more flexible approach to carrying out geomodeling. Further, it is important to carry out more extensive scientific testing using this approach.

Even though the Adam optimizer performed very well in our case, attempting other optimization techniques with more complex models would demonstrate how well the approach performed relatively. There are other optimization technique such as Nesterov-accelerated Adaptive Moment Estimation (NADAM), Stochastic Gradient Descent (SGD) etc.

Nevertheless, it is our belief that, for the purpose of this study, we were able to achieve a very good result.

Chapter 5

Conclusion

In this thesis work, we developed a framework around probabilistic machine learning where we can combine additional information to learn about the subsurface for improved decision-making with surface-based geological models through the integration of an explicit-implicit approach. We started with the integration of an implicit-explicit approach through a process called re-sampling, where we approximate the scalar field we got from the implicit model approach through the use of cubic splines, an explicit approach to form a surface-based model.

The implicit modeling step in this work is implemented using *GemPy*, which is an open-source 3D geomodeling package. It is a Python-based package that is capable of generating and visualizing complex 3D geological models based on the potential field interpolation method. This package is used to create our theoretical geological model, from which surface points were extracted. The extracted surface points were subjected to a resampling process using our explicit modeling approach. This involves the use of a cubic basis spline algorithm to create surface-based geological models. This surface-based model comes with control points that make it very flexible to manipulate and obtain several realizations that can be used for estimating the uncertainties involved in our geological model.

These surface-based models are then subjected to uncertainty estimation through the use of control points and additional data, such as gravity data gotten from the geophysical simulation of the subsurface in the area of the target formation. The additional data is set up as a likelihood function, acting as a constraint on our model. We used this integration to learn, estimate, and evaluate the uncertainties in our models.

This integration approach produced the desired result on the theoretical model, as we can see in the result section. However, additional research and optimization were proposed in order to develop a more robust method capable of handling more complex, heterogeneous subsurface scenarios.

Bibliography

Amirfakhrian, M. (2012), 'Approximation of 3d-parametric functions by bicubic b-spline functions', *International Journal of Mathematical Modelling Computations* **02**, 211–220.

Bastida, F., Aller, J., Fernández, F. J., Lisle, R. J., Bobillo-Ares, N. C. & Menéndez, O. (2014), 'Recumbent folds: Key structural elements in orogenic belts', *Earth-Science Reviews* **135**, 162–183.

Betancourt, M. (2018), 'A conceptual introduction to hamiltonian monte carlo', **2**(3).

Bindiganavle, K. (2000), 'An optimal approach to geometric trimming of b-spline surfaces'.

Bjørlykke, K. (2015), *Petroleum Migration; Petroleum Geosciences: From Sedimentary Environments to Rock Physics Book*, Springer, Berlin/Heidelberg, Germany.

Bois, F. (2013), 'Bayesian inference. methods in molecular biology'.

Calcagno, P., Chilès, J.-P., Courrioux, G. & Guillen, A. (2008), 'Geological modelling from field data and geological knowledge, Part I – Modelling method coupling 3D potential-field interpolation and geological rules', *Physics of the Earth and Planetary Interiors* p. 38.

Caumon, G., Collon, P., Le Carlier de Veslud, C., Viseur, S. & Sausse, J. (2009), 'Surface-based 3d modeling of geological structures', *Mathematical geosciences* **41**, 927–945.

Cockett, R., Kang, S., Heagy, L. J., Pidlisecky, A. & Oldenburg, D. W. (2015), 'Simpeg: An open source framework for simulation and gradient based parameter estimation in geophysical applications', *Computers & Geosciences* .

Davidson-Pilon, C. (2015), *Probabilistic programming and bayesian inference for hackers.*, 1st edn, Addison-Wesley Data Analytics.

de la Varga, M., Schaaf, A. & Wellmann, F. (2018), 'GemPy 1.0: Open-source stochastic geological modeling and inversion', *Geoscientific Model Development Discussions* pp. 1–50.

de la Varga, M., Schaaf, A. & Wellmann, F. (2019), 'GemPy 1.0: Open-source stochastic geological modeling and inversion', *Geoscientific Model Development* **12**(1), 1–32.

de la Varga, M. & Wellmann, J. F. (2016), 'Structural geologic modeling as an inference problem: A Bayesian perspective', *Interpretation* **4**(3), SM1–SM16.

De Luca, A. & Termini, S. (1972), 'A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory', *Information and Control* **20**(4), 301–312.

Dillon, J., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M. & Saurous, R. (2017), 'Tensorflow distributions'.

Eastern Connecticut State University (2010), 'Recumbent fold in swiss alps, Switzerland.', https://www.easternct.edu/cunninghamw/files/nice-big-swiss-fold.jpg.

Farin, G. (2002), 8 - b-spline curves, *in* G. Farin, ed., 'Curves and Surfaces for CAGD (Fifth Edition)', fifth edition edn, The Morgan Kaufmann Series in Computer Graphics, Morgan Kaufmann, San Francisco, pp. 119–146.

Gelman, A. (2008), 'Objections to bayesian statistics. bayesian analysis', *Interpretation* **3**(3), 445–450.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D.B., V. A. & Rubin, D. B. (2015), *Bayesian Data Analysis*, 3rd edn, Taylor Francis Group, Boca Raton, FL.

Geology In (2015), 'Anticline fold in Alberta, Canada.', https://www.geologyin.com/2015/02/types-of-folds-with-photos.html.

George, B. (2007), 'The george quotes', https://www.azquotes.com/quote/534227.

Jacquemyn, C., Jackson, M. D. & Hampson, G. J. (2019*a*), 'Surface-based geological reservoir modelling using grid-free nurbs curves and surfaces', *Mathematical Geosciences* **51**(1), 1–28.

Jacquemyn, C., Jackson, M. D. & Hampson, G. J. (2019*b*), 'Surface-based geological reservoir modelling using grid-free nurbs curves and surfaces', *Mathematical Geosciences* **51**(1), 1–28.

James, F. E. (2013), *An Introduction to Numerical Methods and Analysis*, 2nd edn, John Wiley Sons, New Jersey.

Kingma, Diederik Ba, J. (2014), 'Adam: A method for stochastic optimization.', **2**.

Lajaunie, C., Courrioux, G. & Manuel, L. (1997), 'Foliation fields and 3D cartography in geology: Principles of a method based on potential interpolation', *Mathematical Geology* **29**(4), 571–584.

Lam, S. K., Pitrou, A. & Seibert, S. (2015), Numba: A llvm-based python jit compiler, *in* 'Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC', pp. 1–6.

Lee, E. (1989), 'Choosing nodes in parametric curve interpolation', *Computer-Aided Design* **21**(6), 363–370.

Liu, Y., Shiqi, L. & Mao, P. (2020), 'Finite element simulation of oil and gas reservoir in situ stress based on a 3d corner-point grid model', *Mathematical Problems in Engineering* **2020**, 1–14.

Lorena, B. (2011), 'Tri-diagonal matrix algorithm'.
URL: *http://www.thevisualroom.com/01barbatheory/tridiagonalmatrix.html*

Lyche, T. & Mørken, K. (2008), *Spline Methods*, Vol. 10.

Mäntylä, M. (1988), 'An introduction to solid modeling.', *Rockville Md: Computer Science Press.* **26**(1), 45–60.

Matheron, G. (1963), 'Principles of geostatistics', *Economic Geology* **58**(8), 1246–1266.

Mehta, P., Bukov, M., Wang, C.-H., Day, A. G., Richardson, C., Fisher, C. K. & Schwab, D. J. (2019), 'A high-bias, low-variance introduction to machine learning for physicists', *Physics Reports* **810**, 1–124. A high-bias, low-variance introduction to Machine Learning for physicists.

Mungia, M. & Bhatta, D. (2015), 'Use of cubic b-spline in approximating solutions of boundary value problems', *Applications and Applied Mathematics: An International Journal* **10**, 750–771.

Muschinski, T., Mayr, G. J., Simon, T., Umlauf, N. & Zeileis, A. (2022), 'Cholesky-based multivariate gaussian regression', *Econometrics and Statistics* .

Neal, M. R. (1995), 'Bayesian learning for neural networks.', *Lecture Notes in Statistics* **118**(3), 445–450.

Neal, M. R. (2012), 'Mcmc using hamiltonian dynamics', *Handbook of Markov Chain Monte Carlo* **118**(3), 445–450.

Nourian, P., Goncalves, R., Zlatanova, S., Ohori, K. & Vo, A.-V. (2016), 'Voxelization algorithms for geospatial applications', *MethodsX* **3**.

Patil, S. & Ravi, B. (2006), 'Voxel-based representation, display and thickness analysis of intricate shapes', *Proceedings - Ninth International Conference on Computer Aided Design and Computer Graphics, CAD/CG 2005* **2005**, 6 pp.–.

Piegl, L. A. & Tiller, W. (2000), 'Curve interpolation with arbitrary end derivatives', *Engineering with Computers* **16**(1), 73–79.

Piegl, L. & Tiller, W. (1997), *The NURBS Book*, Vol. 10, 2 edn, Springer, Berlin.

Qu, D., Røe, P. & Tveranger, J. (2015), 'A method for generating volumetric fault zone grids for pillar gridded reservoir models', *Computers Geosciences* **81**, 28–37.

Rogers, D. F. (2001), *An Introduction to NURBS: With Historical Perspective*, 1st edn, Morgan Kaufmann, New York.

Schaaf, A., de la Varga, M., Wellmann, F. & Bond, C. E. (2021), 'Constraining stochastic 3-d structural geological models with topology information using approximate bayesian computation in gempy 2.1', *Geoscientific Model Development* **14**(6), 3899–3913.

Shannon, E. C. (1948), 'A mathematical theory of communication: Bell system technical journal.', **379–423**(27).

Shene, C. K. (1997), 'Introduction to computing with geometry notes', https://pages.mtu.edu/ shene/COURSES/cs3621/NOTES/.

Sherman, D. (2008), 'Sandstone pinch out in shales in the jurassic morrison formation, green river, Wyoming, United States.', https://www.pbase.com/image/93925514.

Simon, D., Anthony, K., Brian, P. & Duncan, R. (1987), 'Lattice quantum chromodynamics', https://en.wikipedia.org/wiki/Lattice$_QCD$.

Stamm, F. A., de la Varga, M. & Wellmann, F. (2019), 'Actors, actions, and uncertainties: optimizing decision-making based on 3-d structural geological models', *Solid Earth* **10**(6), 2015–2043.

Thuan, K. N. (2018), 'The thomas algorithm for tridiagonal matrix equations', https://www.cpp.edu/ tknguyen/egr511/Notes/.

Wellmann, F. & Caumon, G. (2018), Chapter one - 3-d structural geological models: Concepts, methods, and uncertainties, Vol. 59 of *Advances in Geophysics*, Elsevier, pp. 1–121.

Wellmann, F. & Regenauer-Lieb, K. (2012), 'Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models.', *Tectonophysics* **526-529**(3), 207–216.

Wolfram Research, Inc. (2010), 'Mathematica 8.0'.
URL: *https://www.wolfram.com*

Zhang, Z., Yin, Z. & Yan, X. (2018), 'A workflow for building surface-based reservoir models using nurbs curves, coons patches, unstructured tetrahedral meshes and open-source libraries', *Computers Geosciences* **121**, 12–22.

Appendix A

Appendix

A-1 Code Availability

GemPy and TFP are open-source Python packages. Gempy can be obtained from here https://github.com/cgre-aachen/gempy while TFP can be obtained from here https://github.com/tensorflow/probability. This project is also written in Python programming language, and the code developed during this research, including the algorithms, theoretical models, and functions, can be accessed from here https://gitfront.io/r/user-8622696/HGzgLWdvGNLS/PyC-Bspline-Opt/. Other files, such as the JupyterLab notebooks and input data for the provided theoretical models, are also provided.