

MSc thesis in Geomatics

# Spatial Height Prediction of ICESat-2 Satellite Data using Random Forest Regression

Leo Kan

January 2024



**MSc thesis in Geomatics**

**Spatial Height Prediction of ICESat-2  
Data using Random Forest Regression**

Ling Wo Leo Kan

January 2024

A thesis submitted to the Delft University of Technology in  
partial fulfillment of the requirements for the degree of Master  
of Science in Geomatics

Ling Wo Leo Kan: *Spatial Height Prediction of ICESat-2 Data using Random Forest Regression* (2024)

© This work is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

The work in this thesis was carried out in the:



3D geoinformation group  
Delft University of Technology

Supervisors: Hugo Ledoux  
Maarten Pronk  
Co-reader: Azarakhsh Rafiee

# Abstract

The Earth's surface is a complex landscape that is essential for a wide range of applications, from urban planning to environmental monitoring. Digital models of the Earth's surface are generated through mathematical calculations using elevation data collected from various sources and the Digital Terrain Model (DTM) which captures the bare earth's surface topography in 2.5D. The creation of DTM is an approximation of terrain in unsample locations, by using x-y coordinates and one z value. Traditionally, terrain interpolation uses deterministic or geo-statistical methods to calculate elevation. This research, based on [Hengl et al. \[2018\]](#), would use random forest regression as an alternative method and to compare the results against traditional interpolation. Comparing different locations against traditional interpolation yields similar results overall. Feature importance, within the the points that are closest to the sampled ICESat-2 data point are more significant than other features used in Random Forest model. The correlation between these datasets and the spatial relationship established would impact on the results of the elevation. The improvement overall of using traditional interpolation compared to random forest regression is limited depending on the location and using model trained with local datasets. For model trained on other geographical locations, which shows similar differences.



# Acknowledgements

I am deeply indebted to my family for their unwavering love and support throughout my studies. Their encouragement and belief in me have been a constant source of strength. I am especially grateful to my parents, for their sacrifices and unwavering support. My sister, who came to visit me during this time, is really wonderful. All of their support has been a driving force throughout my academic journey.

I would like to also express my sincere gratitude to my supervisors, Hugo Ledoux and Maarten Pronk, for their invaluable guidance and support throughout my thesis journey. Their expertise, patience, and encouragement were instrumental in shaping my research and helping me overcome challenges along the way.

I am particularly grateful to Hugo for his unwavering belief in my abilities and for always pushing me to reach my full potential. His dead-pan humour is of course confirmation of how intelligent he truly is. Maarten's insights and expertise were invaluable in refining my research methodology and ensuring the diligence of my analyses.

Finally, I would like to thank all my friends in the Netherlands and Hong Kong for their good spirits and support. Thanks to the hang out spot 'Jazzcafé Bebop' too. I have had great evenings and great conversations with friends. Their encouragement and willingness to offer a hand have been invaluable throughout my studies. Thank you.



# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Research Motivation . . . . .	1
1.2. Traditional interpolation Techniques . . . . .	2
1.3. Research Question . . . . .	3
1.4. Research Scope and constraint . . . . .	4
1.5. Study area . . . . .	4
1.6. Thesis outline . . . . .	5
<b>2. Background and Related Work</b>	<b>7</b>
2.1. The ICESat-2 Satellite Mission . . . . .	7
2.2. Technical Background . . . . .	8
2.3. Existing Spatial Interpolation Techniques . . . . .	9
2.3.1. Interpolation with Voronoi Diagram . . . . .	10
2.4. Random Forest Regression . . . . .	12
2.5. Random forest for spatial predictions framework (RFsp) . . . . .	13
<b>3. Methodology</b>	<b>15</b>
3.1. Download and Filtering Data . . . . .	16
3.2. Gathering Features Data for Training . . . . .	17
3.2.1. Geometric Features using Dataset Relationships . . . . .	17
3.2.2. Remote Sensing features via Google Earth Engine . . . . .	18
3.3. Random Forest Regression . . . . .	20
3.3.1. Data Pre-processing . . . . .	20
3.4. Testing geometric features . . . . .	21
3.4.1. Implementation of Random Forest Regressor . . . . .	22
3.4.2. Exporting and Combining Random Forest (RF) models . . . . .	24
3.4.3. Testing and Evaluation . . . . .	25
3.4.4. Comparison with Traditional Interpolation Methods . . . . .	25
3.4.5. Presenting final results and accuracy assessment . . . . .	26
<b>4. Implementation</b>	<b>29</b>
4.1. Access to Elevation Data . . . . .	29
4.2. Correlation Matrix Among Geometric Features . . . . .	29
4.3. Details on the Study Area . . . . .	31
4.4. Code Access and Data Downloading . . . . .	31

## *Contents*

<b>5. Results and Analysis</b>	<b>33</b>
5.1. Individually Training of Geometric Features . . . . .	33
5.1.1. Test 1: Distance to Nearest ICESat-2 Point . . . . .	34
5.1.2. Test 2: Nearest Neighbour Height . . . . .	35
5.1.3. Test 3: Gradient to Neighbourhood Points . . . . .	36
5.1.4. Test 4: Relative Height To Neighbours . . . . .	37
5.1.5. Test 5: Remote Sensing Features . . . . .	38
5.2. Training RF model by Combining Features . . . . .	40
5.3. Application on Geographic Locations . . . . .	41
5.3.1. Grand Canyon, USA . . . . .	42
5.3.2. South Limburg, Netherlands . . . . .	43
5.3.3. Mount Taranaki, New Zealand . . . . .	43
5.4. Use of Pre-trained RF model on Tasmania, Australia . . . . .	44
<b>6. Discussion and Conclusion</b>	<b>55</b>
<b>A. Reproducibility self-assessment</b>	<b>59</b>
A.1. Marks for each of the criteria . . . . .	59
A.2. Reflection . . . . .	60
<b>B. Features Data</b>	<b>63</b>

# List of Figures

1.1. Differences between Digital Surface Model (DSM) (Earth's surface with objects like buildings and vegetation) and DTM (Bare Earth's surface without any objects) [Croneborg et al., 2020] . . . . .	1
1.2. ICESat-2 Track . . . . .	3
2.1. ICESat-2 ATL08 Product Classification from Neuenschwander et al. [2021] . . . . .	7
2.2. ICESat-2 mission beam pattern [Smith et al., 2019] . . . . .	8
2.3. Adjusted IDW adapted from Li et al. [2018] . . . . .	10
2.4. Voronoi Diagram . . . . .	11
2.5. Inverse Distance Weighting (IDW) Interpolation with pre-defined searching circle . . . . .	12
2.6. Partial Structure of Decision Tree . . . . .	13
3.1. Procedure to predict terrain elevation . . . . .	15
3.2. Calculation of geometric features from ICESat-2 points . . . . .	18
3.3. One-hot encoding on categorical data . . . . .	20
3.4. Interpolation using Tasmania ICESat-2 Data using methods: (a) IDW (b) TIN (c) Laplace . . . . .	27
4.2. Location maps of the study area. (a) Tasmania, Australia (b) Grand Canyon, USA, (c) Limburg, Netherlands, and (d) Mount Taranaki, New Zealand . . . . .	32
5.1. (a) RF - Nearest Distance to ICESat-2 Point (b) Scatter Plot . . . . .	34
5.2. (a) RF - Nearest Height to ICESat-2 Point (b) Scatter Plot . . . . .	36
5.3. (a) RF - Gradient to ICESat-2 Point (b) Scatter Plot . . . . .	37
5.4. (a) RF - Relative Height to ICESat-2 Point (b) Scatter Plot . . . . .	38
5.5. Feature Importance from Remote Sensing Features Only . . . . .	39
5.6. Subsequently add remote sensing features to the random forest features and results from adding features: (a) Water Mask (b) Geomorphon (c) Land Use and Land Cover (d) Normalised Difference Vegetation Index (NDVI) . . . . .	45
5.7. RF model using All Features on ICESat-2 Data . . . . .	46
5.8. Random Forest Regression (RFR) using All Features without Nearest Neighbour Height on ICESat-2 Data . . . . .	47
5.9. <b>Grand Canyon, USA Comparing Different Methods</b> . . . . .	48

*List of Figures*

5.10. (a) <b>Grand Canyon</b> –Combined Geometric Features (b) Scatter Residuals Plot (c) Features Importance . . . . .	49
5.11. <b>South Limburg, Netherlands</b> Comparing Different Methods . . . . .	50
5.12. (a) <b>South Limburg</b> –Combined Geometric Features (b) Scatter Residuals Plot (c) Features Importance . . . . .	51
5.13. <b>Mount Taranaki, New Zealand</b> Comparing Different Methods . . . . .	52
5.14. (a) <b>Mount Taranaki</b> – Combined Geometric Features (b) Scatter Residuals Plot (c) Features Importance . . . . .	53
5.15. RF result trained from three geographic locations . . . . .	54
6.1. Figure showing prediction v actual scatter graph . . . . .	56
A.1. Reproducibility criteria to be assessed. . . . .	59

# List of Tables

3.1. Comparison between Traditional Interpolation Methods . . . . .	26
4.1. Coordinates of Study Area in EPSG:4326 . . . . .	31
5.1. Summary of Geometric Features Tests . . . . .	33
5.2. Distance to Nearest ICESat-2 Point . . . . .	34
5.3. Nearest Neighbour Height of ICESat-2 Point . . . . .	35
5.4. Nearest Neighbour Slope of ICESat-2 Point . . . . .	36
5.5. Relative Height to Nearest ICESat-2 Point . . . . .	37
5.6. Additional Features (a) Water Mask (b) Geomorphon (c) Land Use and Land Cover (d) NDVI . . . . .	39
5.7. Geometric Features Tests . . . . .	40
5.8. Tasmania Data Interpolation and RF Prediction . . . . .	41
5.9. Combined Features among Study Areas . . . . .	42
5.10. Prediction with Geometric Features . . . . .	42
5.11. Prediction with Geometric Features . . . . .	43
5.12. Prediction with Geometric Features . . . . .	43
5.13. Error Metrics from Combined Models and All Features Test from Tasmania, Australia . . . . .	44
A.1. Self Evaluation . . . . .	59



# Acronyms

<b>DEM</b>	Digital Elevation Model . . . . .	1
<b>DSM</b>	Digital Surface Model . . . . .	xi
<b>DTM</b>	Digital Terrain Model . . . . .	1
<b>LiDAR</b>	Light Detection and Ranging . . . . .	1
<b>TIN</b>	Triangular Irregular Network . . . . .	1
<b>NNI</b>	Natural Neighbour Interpolation . . . . .	9
<b>IDW</b>	Inverse Distance Weighting . . . . .	xi
<b>AIDW</b>	Adjusted Inverse Distance Weighting . . . . .	10
<b>RFR</b>	Random Forest Regression . . . . .	xi
<b>RF</b>	Random Forest . . . . .	ix
<b>ML</b>	Machine Learning . . . . .	21
<b>RMSE</b>	Root Mean Square Error . . . . .	8
<b>MSE</b>	Mean Squared Error . . . . .	24
<b>MDA</b>	Mean Decrease in Accuracy . . . . .	24
<b>GEDI</b>	Global Ecosystem Dynamics Investigation . . . . .	2
<b>MDI</b>	Mean Decrease in Impurity . . . . .	24
<b>RFsp</b>	Random forest for spatial predictions framework . . . . .	ix
<b>OK</b>	Ordinary Kriging . . . . .	13
<b>MAE</b>	Mean Average Error . . . . .	26
<b>gDEM</b>	Global Digital Elevation Model . . . . .	4
<b>InSAR</b>	Interferometric Synthetic Aperture Radar . . . . .	2
<b>NDVI</b>	Normalised Difference Vegetation Index . . . . .	xi
<b>MDA</b>	Mean Decrease in Accuracy . . . . .	24



# 1. Introduction

Earth's surface is filled with complex landscapes. Digital models of the Earth's surface are essential tools for a wide range of applications: from urban planning to environmental monitoring [Arun, 2013]. These models are generated through mathematical calculations using elevation data collected from various sources, including manual surveying, satellite imagery, and aerial photography. Figure 1.1 shows that the term Digital Elevation Model (DEM) encompasses two things: the DSM, which represents all surface objects like trees and buildings, and the Digital Terrain Model (DTM), which captures the bare earth's surface topography.

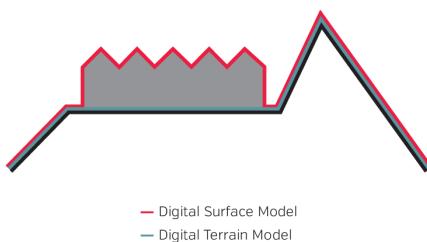


Figure 1.1.: Differences between DSM (Earth's surface with objects like buildings and vegetation) and DTM (Bare Earth's surface without any objects) [Croneborg et al., 2020]

To construct a digital representation of terrain, data is collected using diverse acquisition techniques such as aerial Light Detection and Ranging (LiDAR) scanning, land surveying, and photogrammetry. These datasets are subsequently processed, filtered, and interpolated to generate a DEM, commonly visualised through data structures such as rasters, Triangular Irregular Network (TIN), and contour lines.

## 1.1. Research Motivation

Achieving global coverage for Earth's surface elevation is challenging. It is physically impractical and time consuming to acquire such dense data when scaled up to a global scale. Using aerial LiDAR point cloud capture, it is able to achieve large land areas covered with density of 6-8 points per square meter in smaller countries such as the Netherlands [AHN, 2023]), it is, however, still a patchwork of varying levels of coverage in larger countries such as the United States, as can be seen on Open Topography OpenTopography [2023].

## 1. Introduction

Tracking changes to the earth's surfaces can be done with **InSAR!** (**InSAR!**), which uses radar images to measure ground deformation. Contrary to optical imaging satellites, it can overcome cloud cover and capture at night. Data from Interferometric Synthetic Aperture Radar (**InSAR**), such as Sentinel-1, can reveal topographic changes when the satellite returns to the same point as it orbits [Bürgmann et al. \[2000\]](#). The downside of **InSAR** is that it can be susceptible to different atmospheric conditions because radar back-scatter is sensitive to variations in water content in the troposphere [Michaelides et al. \[2021\]](#). ICESat-2<sup>1</sup>, in this case, is able to be employed to monitor and model Earth's terrain at higher temporal resolution and measure thin ice sheets that are not possible for **InSAR**. When modelling a **DTM** [Li \[1994\]](#) with sea, and ice changes, each method has its own advantages, and this project focuses on ICESat-2 satellite mission.

The ICESat-2 satellite data often come with its own limitations, with sparse data points acquired at standard intervals. Therefore, machine learning techniques explored by [Hengl et al. \[2007\]](#) is able to produce an accurate **DTM**. He explored the geometric relationships between the elevation points as features in "a similar way as Kriging" [Hengl et al. \[2007\]](#). With the addition of remote sensing features, this would contribute additional information to model's learning. Based on this context, this thesis aims to assess the accuracy of using ICESat-2 elevation data using random forest machine learning method data that is gathered from this satellite mission, and testing whether Geometric and Remote sensing features is able to produce a **DTM** that performs better than traditional interpolation.

### 1.2. Traditional interpolation Techniques

There are a range of techniques in terms of traditional interpolation techniques, such as **IDW** Interpolation [[Shepard, 1968](#)], Laplace Interpolation [[Burrough, 1986](#)], and other linear spatial interpolation techniques. Since it is common to have a limited number of elevation measurements in the area of interest, these intermediate points would need to rely on known measurement points to estimate its elevation. Some terms are used interchangeably to represent the bare-earth model that measures from the vertical datum. For this research, **DEM** will be used to mean the bare earth model. Therefore, the act of spatial interpolation methods are essential to estimate values within the data gaps where measurements are not available. Comparing **RF** and interpolation techniques would expand the understanding of machine learning, particularly on sparse space-borne **LiDAR** datasets.

Space-borne satellite missions like ICESat-2 and Global Ecosystem Dynamics Investigation (**GEDI**) are launched to detect changes and measure the Earth's land, ice, and vegetation surfaces with high precision. The datasets from these missions allow scientists to monitor changes in the Earth throughout its many orbits around Earth. The

---

<sup>1</sup>Ice, Cloud and land Elevation Satellite

### 1.3. Research Question

satellite orbits at an angle at certain intervals, and a ground track pattern is designed to cover the earth as much as possible. The ground track pattern of strong beams and weak beams of ICESat-2 is illustrated in [Figure 1.2](#).

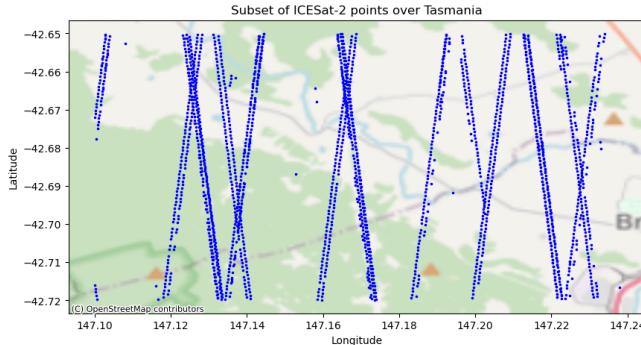


Figure 1.2.: ICESat-2 Track

## 1.3. Research Question

Elevation dataset using coordinates from the ICESat-2 mission will be used in this thesis to produce a [DTM](#) of several areas of interests, namely **Australia**, **New Zealand**, **the Netherlands** and **the United States**. This thesis aims to find the significance of Geometric and Remote Sensing Features to perform elevation prediction on the sparse measurements of the ICESat-2 mission, and would compare the results against Traditional Interpolation Techniques against ground truth. The main research question is as follows:

**To what extent would Random Forest (RF) elevation prediction improve on traditional interpolation methods for creating a DTM?**

To further investigate this research question, and the main question can be separated into several sub-questions:

- Which features in the [RF](#) model have the most significant impact on the accuracy?
- How does the accuracy of [RF](#) regression vary across different geographical locations within the study areas?
- Compared to traditional interpolation, how would the [RF](#) model created in this research compared against the method proposed in [RFsp](#) by [Hengl et al. \[2018\]](#)?

## *1. Introduction*

### **1.4. Research Scope and constraint**

To facilitate the research and the resulting digital terrain, the main study area for this research will focus on interpolation of the sparse nature of the ICESat-2 ATL08 dataset. Though the ICESat-2 mission covers near-globally, this study will concentrate on three distinct geographical regions, which will be further elaborated in [section 1.5](#).

In addition to traditional interpolation methods, machine learning will also be used to predict the expected height of the areas in between known data points. [RF](#) regression and the tools within the scikit-learn library will be extensively used in the scope of this research. The scikit-learn library offers flexibility and comprehensive functions that allows efficient testing and easy implementation.

The ICESat-2 ATL08 dataset is known for its sparsity, and machine learning can adapt to varying levels of density to make predictions possible in regions in between satellite transect. External datasets contribute as features in the algorithm, with their weights and biases being integrated into the decision trees to enhance its robustness. By using [RF](#) regression, this research can expect to generate a more comprehensive [DTM](#).

To fully implement the machine learning algorithm, a significant amount of data is necessary to effectively train the model. Consequently, a diverse collection of data will be employed as features. These features encompass a range of data types, including geometric data, land use and land cover data, settlement data, and remote sensing data. The full list of data sources can be found in [Appendix B](#).

The constraint for this study is that the use of Global Digital Elevation Model ([gDEM](#)) and other derivative products such as viewshed, accumulation flow, slope and aspect as feature data will not be used. It is because this would provide influence to the model and is prone to over-fitting. However, the relationships between the ICESat-2 points will be used based on the geometrical relationship between the ICESat-2 points (Geometric Features) and external remote sensing datasets (Remote Sensing Features) obtained from earth observation agencies that are found on Google Earth Engine.

### **1.5. Study area**

To demonstrate the suitability of [RF](#) as the predictive model on ICESat-2 points, a comparison of interpolation techniques with ICESat-2 points will be compared by using an area in **Tasmania, Australia**. **Tasmania, Australia** provides good range of terrain features, and is suitable to be a first step in understanding its performance before comparing with other study areas and [RF](#) regression method.

In addition to Tasmania, three more areas are chosen as study areas for comparison

## 1.6. Thesis outline

for this research. They are chosen to show the range of terrain features the random forest algorithm is expected to handle, namely hill, saddle, valley, ridge, and depression. These four areas chosen are as follows:

- Tasmania, Australia
- Grand Canyon, United States
- South Limburg, Netherlands
- Mount Taranaki, New Zealand

**Tasmania, Australia** offers a range of terrain features that are present with rivers and valleys. **Mount Taranaki, New Zealand** is a cone-shaped volcano situated in North Island of the country. It demonstrates one large hill—in this case a volcano—with a height of over 2500 metres. By contrast, **South Limburg, Netherlands** has more terrain features albeit the highest peak stands at around 300 metres, and finally, the **Grand Canyon, United States** can show a variety of terrain features within the area.

ICESat-2 data will come from the NASA Earthdata ([www.earthdata.nasa.gov](http://www.earthdata.nasa.gov)), and the ground truth data from the mapping agency for each country. They are Land Information New Zealand (LINZ) ([www.linz.govt.nz](http://www.linz.govt.nz)), Kadaster ([www.kadaster.nl](http://www.kadaster.nl), and USGS ([www.usgs.gov](http://www.usgs.gov)) respectively.

## 1.6. Thesis outline

In this research there are six chapters and broken down to subsections. A brief summary of each chapter are outlined as follows:

- **chapter 1:** Introduction provides the background, motivation, and objectives of this study. It also introduces the study area, dataset, and research question.
- **chapter 2:** Theoretical Background and Related Work elaborates on the technical aspects, including the concepts of different terrain analysis techniques, and various **DEM** analysis approaches. This chapter also details the machine learning methodologies that are available and discusses the techniques in which different research has explored.
- **chapter 4:** This chapter goes into detail about the algorithms for downloading different datasets, the interpolation techniques used.
- **chapter 3:** Methodology details the downloading and pre-processing of ICESat-2 data, feature data to be used as training data of the **RF** algorithm, the enabling of **RF** regression for elevation prediction, and the error metrics used to measure the suitability of the algorithm.

## *1. Introduction*

- **chapter 5:** This chapter presents and interprets the results obtained from the random forest machine learning algorithm, with a focus on comparing them to traditional interpolation techniques and identifying key similarities and differences across various geographical locations. Further analysis on the error metrics and scoring of each feature importance.
- **chapter 6:** This section represents the thesis's conclusion, encompassing the discussion of results, addressing the research questions posed in [chapter 1](#), and exploring potential opportunities for future research.

## 2. Background and Related Work

The characteristics of the ICESat-2 mission and its specific orbital ground track and measurements often results in significant spatial gaps in its data. Present research have addressed this challenge by using deterministic and geo-statistical interpolation techniques. Furthermore, there has been a notable shift towards application of machine learning algorithms to improve elevation accuracy. This background section aims to provide an understanding of the existing strategies for interpolating and, in addition, focusing on random forest regression as a technique to assess the suitability of using this method on sparse data points.

### 2.1. The ICESat-2 Satellite Mission

The satellite's orbit covers most of the Earth's surface due to its orbital inclination of  $92^\circ$ , with global coverage from  $88^\circ$ South to  $88^\circ$ North latitudes. [Neumann et al., 2019] Its ground-track orbits around the polar region. The satellite uses laser beams to measure Earth's surface, and it uses three beam pairs (one weak and one string beam in each pair) that are directed to the Earth's surface to measure the elevation of the Earth's surface. This thesis will focus on data from ATL08 product<sup>1</sup> which has data on the along-track heights above the WGS84 ellipsoid on the ground and canopy surfaces.

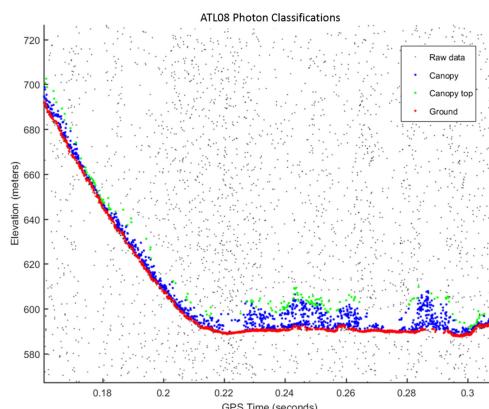


Figure 2.1.: ICESat-2 ATL08 Product Classification from Neuenschwander et al. [2021]

<sup>1</sup>Land and Vegetation Height Product. <https://nsidc.org/data/atl08/versions/5>

## 2. Background and Related Work

Out of all the products in the ICESat-2 mission, the ATL08 product offers data that has been classified as canopy and ground for land and vegetation. The data used is the height of the terrain that is measured as the photon is reflected to the satellite as seen in [Figure 2.1](#). The study conducted by [Neuenschwander et al. \[2021\]](#) demonstrated that following testing across various terrains, including Sonoma County characterised by its "complex topography and vegetation," [[Neuenschwander et al., 2021](#)] the Root Mean Square Error (RMSE) of ATL08 data show that terrain photon levels ranged from 0.5m to 2m. Although [Neuenschwander et al. \[2021\]](#) added that the performance and capabilities still needed more rigorous testing, the accuracy of the satellite still provides good quality for the extent of this research.

## 2.2. Technical Background

The ICESat-2 operates at a pulse of 10 kHz, which means that the laser fires 10,000 times per second. This high repetition rate enables dense sampling of the Earth's surface and allows for accurate measurements of surface elevation changes. As seen from [Figure 2.2](#), the gap between each track between each strong-weak beams is 3.3 kilometres and around 90 metres along each track. The ground track follows a near-polar orbit that completes the orbit around the Earth in 90 minutes. The satellite repeats itself every 91 days that covers the same ground track, and the data from repeated ground tracks enables temporal data for monitoring of changes of land, sea and ice on Earth [[Neumann et al., 2019](#)].

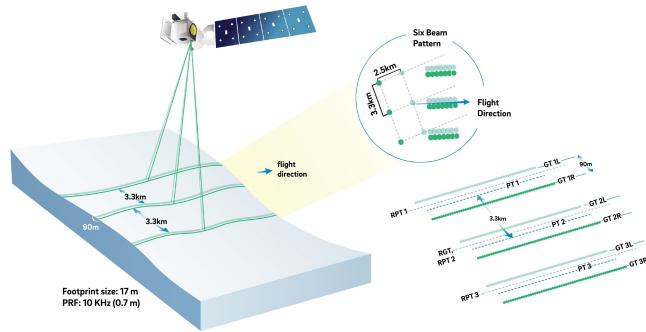


Figure 2.2.: ICESat-2 mission beam pattern [[Smith et al., 2019](#)]

Furthermore, the density of data points varies depending on the latitude of the orbit. In high-latitude polar regions, for instance, ground tracks are closely spaced, resulting in a higher density of measurement points, whereas at mid-latitude and at the equator results are wider in spacing and thus lower the density of measurement points.

For the extent of this research, all points from the ATL08 data product will be used from the beginning of the satellite mission until May 2023 within a predefined bounding box. In terms of the density of points, three areas will be chosen with similar sizes

### 2.3. Existing Spatial Interpolation Techniques

and density of points in order to find the interpolation method for such density of ICESat-2 data points.

In terms of the density and sparseness of the measurement points, this thesis aims to utilise ICESat-2 data and apply Random Forest Regression to fill in the missing areas represented as a digital elevation model raster. Specifically, this thesis will use a range of features to train the model and then evaluate its performance in predicting the missing areas, with bounding boxes having a similar density of measurements.

## 2.3. Existing Spatial Interpolation Techniques

To interpolate spatial data on a 2.5D surface, several techniques are available for filling gaps in the unsampled locations. 2.5D is a representation of terrain that encodes one  $x$ - $y$  (horizontal) coordinate with one  $z$  (vertical) coordinate. In a DTM, there is one elevation data point for every point on the 2-dimensional field [Kraus and Otepka \[2005\]](#).

Traditional techniques such as 'Inverse Distance Weighting ([IDW](#))' [[Shepard, 1968](#)], 'Natural Neighbour Interpolation ([NNI](#))' [[Sibson, 1981](#)], 'Triangular Irregular Network ([TIN](#))' [[Philip and Watson, 1982](#)], and 'Laplace interpolation (Laplace)' [[Burrough, 1986](#)] focuses on the geographical relationship between neighbouring data points, and have been a robust technique for any distribution of data points. All of these techniques aim to find elevation data at unsampled locations using the known elevation values in the surrounding neighbourhood. Each of these techniques uses proximity and assigning higher weighting to nearby points. Their simple operation make them convenient to use on many spatial data. Furthermore, [IDW](#), Laplace, and [NNI](#) methods are deterministic approaches, which are fully dependent on the parameters and input values. In the case of [IDW](#), for instance, the parameters including but not limited to searching circle distance, weighted power, number of neighbours etc. are decided for each terrain characteristics. In order to choose which technique works best, one should consider the complexity of terrain, distribution of data points, and resulting accuracy.

### Inverse Distance Weighting ([IDW](#)) Interpolation

In a spatial dataset with sampled elevation data, [IDW](#) is a common interpolation technique to model the terrain. It assigns weights to its neighbours, by a searching circle or ellipses based on the distance of nearby points within the search area, as illustrated in [Figure 2.5](#). The inverse of the distance of each of the known data points are used as weights. This means the closer the point, the more influence on the weight than further points. These weighted values are then averaged to obtain the interpolated value in the unsampled location.

## 2. Background and Related Work

[Equation 2.1](#) explains it in a formula where  $z_n$  represents the point to be interpolated at location  $n$ ,  $z_i$  is the height of the nearby known points within the searching circle,  $d_i$  is the distance between points between the interpolated point and the nearby points.  $p$  is the power, of which the weighting  $w_i$  are inverse to the distance power. The downside of [IDW](#) is the drawing of searching circle. The size of the circle will influence the results, especially on anisotropic datasets such as ICESat-2 elevation data. It is quite sensitive to areas with large gaps between measurement points.

$$z_n = \frac{\sum_{i=1}^n z_i w_i}{\sum_{i=1}^n w_i} \quad w_i = \frac{1}{d_i^p} \quad (2.1)$$

### Adjusted Inverse Distance Weighting ([AIDW](#)) Interpolation

[AIDW](#) produced by [Li et al., 2018] is an adjusted and improved [IDW](#) to adjust for clustering of data points, as illustrated in [Figure 2.3](#). The adjustment of this technique uses the angle at the interpolated point in relation to the neighbouring two points ( $\angle\alpha$ ), their intersection angle ( $\angle\beta$ ), and their respective distances are included in its weighting formula. Such a technique have indeed improved [IDW](#) in the study location of Zhejiang Province, China. However, it is worth noting that this enhancement comes at the cost of increased computation time, which can pose challenges when scaling up the method.

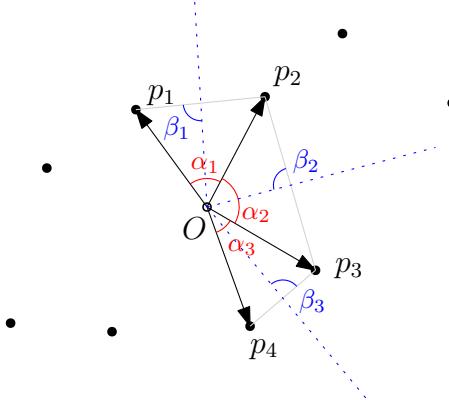


Figure 2.3.: Adjusted IDW adapted from Li et al. [2018]

### 2.3.1. Interpolation with Voronoi Diagram

Using Voronoi Diagram is a way of dividing space to a number of cells that are defined by the closest sampled point on each elevation data point. This method provides another way of estimating values for unsampled points. The continuous surface of a set of discrete points can be used to interpolate the elevation at the spatial

### 2.3. Existing Spatial Interpolation Techniques

extent within the study area. In simple terms, the partitioning of the points creates these cells, and each cell corresponds to the region that represent the areas where the data point is the closest as illustrated in [Figure 2.4](#) with  $O$  being one of the data points.

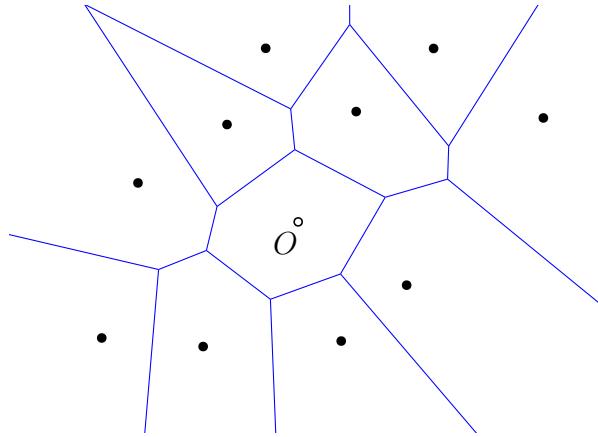


Figure 2.4.: Voronoi Diagram

Based on this diagram, interpolation such as Laplace interpolation and [NNI](#) interpolation are based upon the relationship of the calculation of weighting based on the distance or area respectively. The advantage of using Voronoi Diagram instead of [IDW](#) is that it reduces the number parameters involved, and calculation of is simpler based on the characteristic of the Voronoi Diagram. Compared to [IDW](#), however, these interpolation techniques only work within the convex hull that is constructed by outermost points that encloses all the Voronoi cells in a given dataset. To address the areas that are out of convex hull, cropping of the resulting [DTM](#) may be necessary.

## Kriging Interpolation

Kriging interpolation, on the other hand, is a geo-statistical method used for spatial interpolation that was developed in the 1950s by Krige [[Krige, 1951](#)]. It is a method that estimates the value of a point at an unsampled location using the spatial correlation and the variability of the data from its surrounding neighbourhood. It assumes that there is correlation among the distribution of data, and that data are stationary and normally distributed. Kriging works when the closer points are given higher weighting than those that are further away. The accuracy of the model suffers when the data points are limited in its spread and the number of sampled data is small. Kriging, also, can be slow due to the complexity of mathematical calculations, particularly with large datasets.

## 2. Background and Related Work

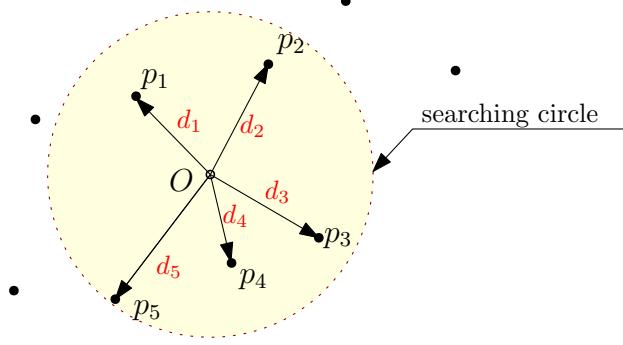


Figure 2.5.: IDW Interpolation with pre-defined searching circle

### Machine Learning as Interpolation Technique

In Machine Learning, filling unsampled data gaps is referred to as data imputation. For spatial interpolation, researchers are finding new interpolation techniques to improve accuracy from traditional interpolation as well as geo-statistical approach such as kriging. Incremental improvements have experimented and applied to different elevation data, including, notably 'Random Forest Spatial Interpolation' [Sekulić et al., 2020] and 'AIDW' [Li et al., 2018].

There has been a growing interest in the use of "more computationally intensive and data-driven algorithms" [Hengl et al., 2018] over the past decade. This interest coincided with the increased availability of high-performance computers, which leads to more experimentation in machine learning and deep learning algorithms. Hengl et al. [2018] introduced to machine learning in random forest and gradient boosting. The paper illustrates that the RFsp framework is able to apply on spatial and spatio-temporal prediction on many geo-spatial fields. It shows that using different kinds of features can improve the prediction of random forest.

[Sekulić et al., 2020] used an approach from random forest to predict values at unsampled locations with information such as precipitation data and their distances from the prediction location as features. Precipitation and daily mean temperature data were used to estimate the precipitation as data are sourced from land surface stations in Spain. With spatial resolution of 10km, The study found that RFSI technique overall outperformed simple deterministic interpolation techniques and had similar performance as IDW.

## 2.4. Random Forest Regression

Random forest is driven by a set of decision trees that improves on prediction accuracy by using a ensemble of these trees (hence the name 'random forest'). The forest of decision trees takes random data points within the features and training data, then

## 2.5. Random forest for spatial predictions framework ([RFsp](#))

taking the average in bootstrap aggregation [Breiman, 2001, Liaw and Wiener, 2002]. One of the downside of using [RF](#) is that coordinates for features in random forest ignores the fundamental relationship between the coordinates itself. This means that it breaks the spatial relationship as data enters each node in the decision tree. This results in overestimating or underestimating predictions. In order to solve this, Hengl et al. [2018] have introduced spatial relationship into random forest regression as a variable. This approach engages spatial relationship such as relative distances, distances to specific coordinates etc. These geometric relationships would form part of this study to be used as Geometric Features. With reference to [RFsp](#), da Silva Júnior et al. [2019] also commented that this method results are closer to [IDW](#) and Ordinary Kriging ([OK](#)) algorithms due to its "lower estimating error" [da Silva Júnior et al., 2019]. With machine learning modelling, some of the previous research is able reduce errors in terms of [RMSE](#) by using machine learning compared to traditional interpolation.

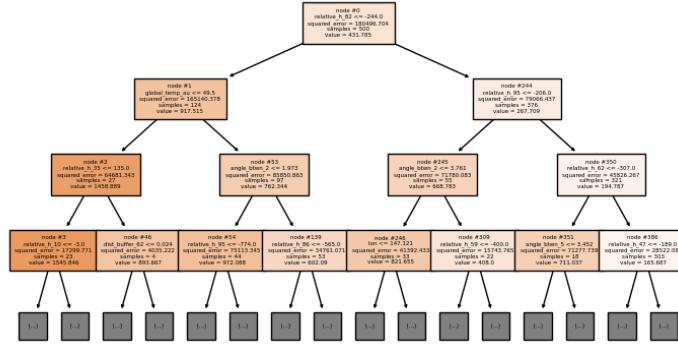


Figure 2.6.: Partial Structure of Decision Tree

## 2.5. Random forest for spatial predictions framework ([RFsp](#))

The upside of [RFsp](#) is that it uses the geographical relationship between the points and the observation points as buffer distances. The introduction of this framework also means features (also called "*covariates*" in Hengl et al. [2018] paper) pointed out having a set of features in random forest as Hengl et al. [2018] have pointed out, [Equation 2.5](#)

$$Y(s) = f(X_G, X_R, X_P) \quad (2.2)$$

## 2. Background and Related Work

The example in [Figure 2.6](#), shows that the decision of each node splitting in both direction based on the criteria set in the hyper-parameters, and with the help of features as exemplified by  $X_G$  for geographical relationships,  $X_R$  are reflectance data from remote sensing images, and  $X_P$  are process-based covariates such as soil reflectance index etc. The combination of these features in [RF](#) regression would have similar outcomes to Ordinary Kriging. The only difference is that variogram is not needed, also a search distance.

This thesis is going to model an area of Earth using spatial interpolation, and the model is a representation of a 2-dimensional surface in a 3-dimensional space. Since the Earth is round, the terrain model would represent poorly on a large scale. [[de Berg et al., 2008](#)] Hence, a smaller-scale representation will be implemented. There are many spatial interpolation methods, and each has their characteristic. This section will highlight existing works on spatial interpolation and their results and accuracy metrics.

There are increasing number of publications in recent years on using machine learning algorithms as a method for spatial interpolation [[Hengl et al., 2018](#), [Sekulić et al., 2020](#)], and is now becoming a useful method for [DEM](#) and [DEM](#) elevation interpolation and experimentation. [RFsp](#), as highlighted by [[Hengl et al., 2018](#)], shows the benefits of using [RF](#) as an approach to perform interpolation with the use of data points, 'covariates' (meaning machine learning features in his paper), and random forest algorithm to perform prediction of terrain elevation. This approach can decrease time for generating a variogram, a computationally intensive operation, and an accuracy that matches [OK](#). The contribution of [Hengl et al. \[2018\]](#) therefore provides a basis for this thesis.

### 3. Methodology

The method of using machine learning algorithms to model terrain is an important aspect regarding the analysis of this research. Using accurate terrain model leads to better understanding the dynamics of Earth's surface. This research aims to use machine learning to perform elevation prediction from ICESat-2 dataset, and [RFR](#) is a central part in this research. This section guides through the methodology of achieving this prediction.

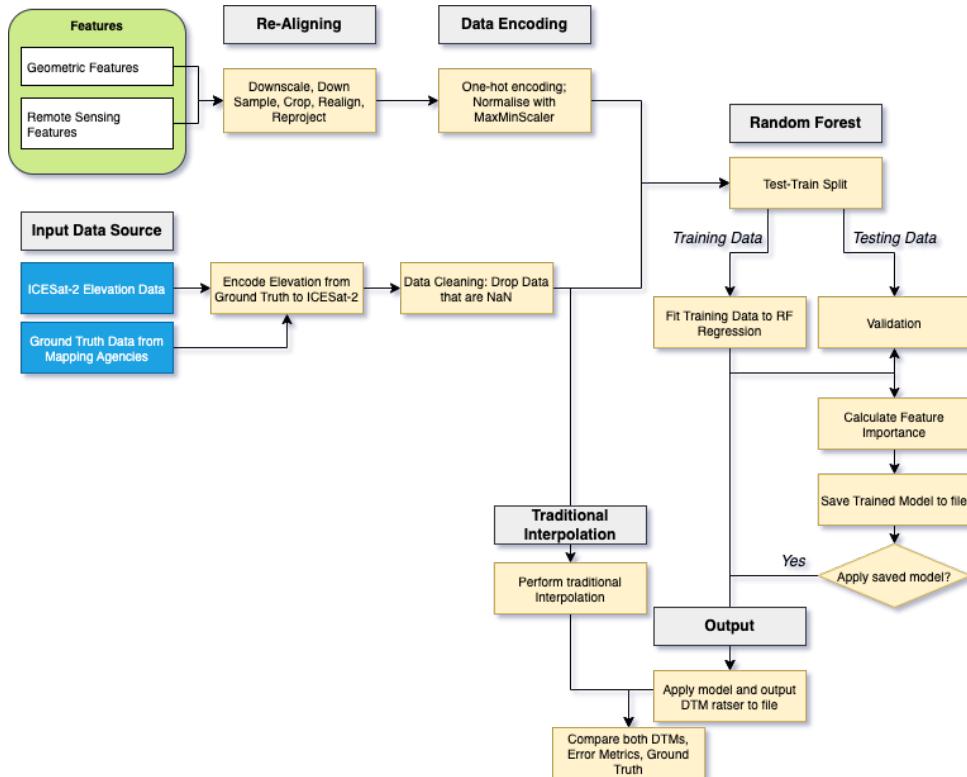


Figure 3.1.: Procedure to predict terrain elevation

Since 2018, the ICESat-2 data provided by the Advanced Topographic Laser Altimeter System (ATLAS) presents a "unique opportunity to quantify the Earth's surface processes" [[Rehman et al., 2023](#)]. Hence the focus will be on the ICESat-2 product,

### 3. Methodology

which captures ground and canopy height in 100-metre intervals. This interval is the distance that each [LiDAR](#) light beam captures the ground surface of Earth during its orbit. This dataset captures the latitude and longitude and the elevation will be provided by each of the official mapping agencies in different geographical locations. In the end, combining trained [RF](#) models from these locations will be tested on a new location. The methodology will follow the steps of gathering ICESat-2 data, followed by feature selection, feature ranking, and training and applying model. The diagram presented in [Figure 3.1](#) summarises this process.

The geographical location chosen for

## 3.1. Download and Filtering Data

The first step in this research involves the collection of satellite data from NASA's Earth Data repository. The objective of the ICESat-2 mission from NASA is to measure the elevation of Earth's ice sheets, sea ice and land surfaces using laser. It provides data for monitoring the polar ice and land surfaces to analyse the impact of climate change. To access the ICESat-2 data, the National Snow and Ice Data Centre (NSIDC) have made a python library *icepyx* that is free to download and open source. This simplifies the process of downloading ICESat-2 data. Since the research extensively investigates Earth's terrain, hence ATL08 is the product needed.

The ATL08 dataset focuses on the photons' heights captured during laser shots to capture the surface elevation and canopy height above Earth's surface [[Markus et al., 2017](#)]. The ATL08 dataset from ICESat-2 contains photon-level data, including the photon heights and times of laser shots, which are useful for deriving surface elevation and canopy height over land.

The use of the *icepyx* Python library makes the download of data easier and more efficient. The ease of data download can be defined by the bounding box. The resulting file is then packaged in [HDF5<sup>1</sup>](#) and available to download. All the data within the bounding box would be contained inside the HDF5 file.

All the necessary data are present and ready to be extracted including latitude and longitude of the ICESat-2 points as well as other features such as height, ground classification etc. Since this study aims to use height as information, a fair comparison would be to use the height that is extracted from the ground truth. The ground truth are elevation data of earth's surface from [DTM](#) that are sourced from each of the mapping agencies. Therefore, only the coordinates of the ICESat-2 point coordinates are necessary. Thus, the ICESat-2 track pattern is downloaded from pre-determined bounding boxes from each location that are presented in [Table 4.1](#).

---

<sup>1</sup>Hierarchical Data Format version 5

## 3.2. Gathering Features Data for Training

To understand the predictions and the role of weights and biases within the context of a random forest, a set of features is needed. These features are integral to the regression process, and it facilitates the decision-making within each node of the decision tree. Each decision split in each node is based on a criteria. The decision criteria will be further elaborated in [section 2.5](#).

The features used in this research can be classified in two broad categories: Geometric Features and Remote Sensing Features. Geometric Features focuses on the geometric relationship within the measurement points and the interpolated points. Whereas Remote Sensing Features are external datasets captured from other remote sensing satellites from reputable sources. These features data will be combined in a data frame that forms the dataset component of the random forest. This table combines all the information of latitude, longitude, elevation, Geometric Features, and Remote Sensing Features.

### 3.2.1. Geometric Features using Dataset Relationships

Geometric features picks on the relationship between the ICESat-2 points and its relationship with the surrounding points. These relationship utilises K-D tree as the primary data structure for data point organisation. The K-D tree enables the identification of N-nearest neighbours in proximity to each ICESat-2 point, which allows the encoding of various relationships among the points. The relationships are namely the relative height differences, slope, the angular relationship between the nearest neighbours, and the buffer distance between the interpolation point and the ICESat-2 points.

- **Height from nearest neighbour:** This feature is derived by considering the n-number of neighbouring ICESat-2 points, from which height information is extracted and encoded within the features. Each encoding represents the height from the n-number of closest neighbour. In [Figure 3.2a](#), each ICESat-2 points  $O$ , are encoded with three neighbouring points— $p_1$ ,  $p_2$ , and  $p_3$ —that has heights of 20, 33, and 29 units respectively.
- **Gradient to nearest neighbour:** Features in the nearest neighbour gradient represents the interpolated point with respect to n-number of neighbouring ICESat-2 points. In general, the gradient is derived from subtracting between two points  $(x_1, y_1, z_1)$  and  $(x_2, y_2, z_2)$ . This will give a vector. Dividing next by its magnitude will obtain the gradient. With n-number of points, the gradient is calculated by finite difference. The calculation of gradient requires data from the nearest neighbour height, and this data will be gathered from ten nearest height data point from the neighbourhood of ICESat-2 point  $O$ .
- **Buffer Distances between ICESat-2 data:** The distance between the ICESat-

### 3. Methodology

2 point ( $O$ ) and points  $p_{1-3}$ , and the interpolated points as illustrated in [Figure 3.2a](#). This means selecting closest points within the neighbourhood and calculate the distance between the sampled point and the ICESat-2 point.

- **Relative Heights between ICESat-2 data:** Features the closest neighbour height, and compare this height with the next height of the nearest neighbour. To establish the relative height differences between the ICESat-2 points. This height would mean that all heights in this feature would be relative to each other.

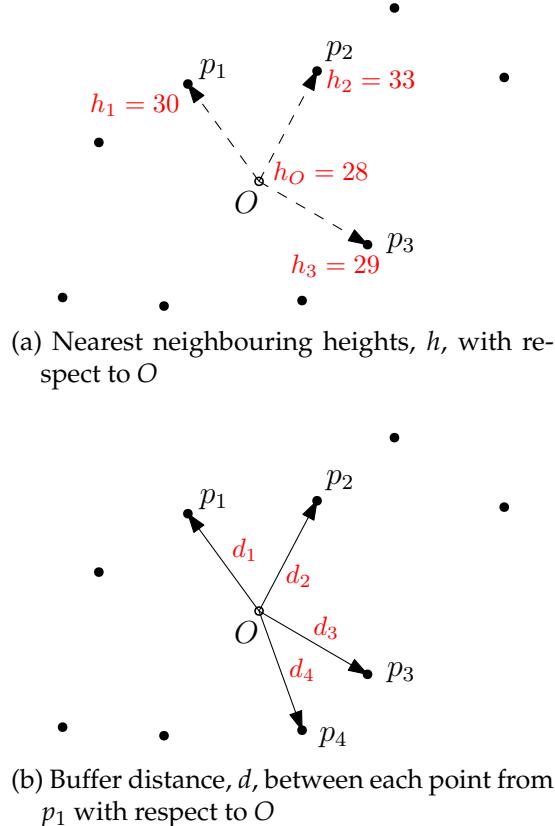


Figure 3.2.: Calculation of geometric features from ICESat-2 points

#### 3.2.2. Remote Sensing features via Google Earth Engine

Remote Sensing Features are data features that are sourced from third-party datasets. These are useful for understanding the relationships that extend beyond the scope of ICESat-2 points, and provides crucial geographical context in relation to the study area. More features data enables the random forest training to be more robust, which leads to improved predictive capabilities. These datasets are available to use freely via Google Earth Engine data catalogue and the Awesome GEE Community Catalogue [[Roy et al., 2023](#)].

### 3.2. Gathering Features Data for Training

- **World Settlement Footprint (WSF) 2019:** [Marconcini et al., 2020] This is a binary mask of which population settlements are represented in '1' and everything else '0'. The data sources used in this dataset stem from multi-temporal imagery captured by Sentinel-1 and Sentinel-2 satellites.
- **Geomorpho90m Geomorphometric Layers** [Amatulli et al., 2020]: Geomorphons are classification method used in geomorphology to categorise landforms based on the characteristics such as slope, aspect, and curvature. These are categorical data that describes the landscape. The 90m resolution data are derived from the MERIT-Digital Elevation Model. The 3 arc-second resolution dataset is used as feature.
- **ESRI 2020 Global Land Use Land Cover from Sentinel-2** [Karra et al., 2021]: This dataset captures the land use and land cover across the globe that is derived from the imagery captured from the Sentinel-2 satellites. This dataset includes information about land cover classes such as forests, urban areas, water bodies, crop lands and so on.
- **GFSAD Landsat-Derived Global Rainfed and Irrigated-Cropland Product (LGRIP)** [Teluguntla et al., 2023]: Croplands derived from the Landsat satellite that categorises from irrigated and rain-fed croplands and also the extent of these croplands. The dataset uses Landsat-8 temporal data at 30 metres resolution. The layers area classified as rain-fed cropland (dependent on precipitation only), irrigated cropland (with at least one irrigation during the crop growth season), non-cropland, and water bodies.
- **ASTER Global Water Bodies Database (ASTWBD)** [NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team, 2019]: This dataset covers global water that are larger than 0.2 square kilometres. The water bodies are further categorised into three categories: ocean, river, or lake. For oceans, there is a uniform elevation, and elevation information is included for river and lakes.
- **Bare Earth's Surface Spectra**[Demattè et al., 2020]: This data shows how often a show is detected as bare earth surface, and also indicating whether the bare earth is increasing or decreasing.
- **Global Forest Cover Change (GFCC) Tree Cover Multi-Year Global 30** [Townshend, 2016]: This dataset captures the temporal analysis of Landsat images over the changes in forest extent and forest changes over a series of years.
- **JRC Global Surface Water Mapping Layers** [Pekel et al., 2016]: This dataset shows open water locations on Earth's surface that spans for three decades using Landsat satellite captured images. It captures natural features like rivers, lakes, coastal waters and artificial water bodies etc. The global surface water maps are presented in 30m resolution.
- **Daily near-surface air temperature dataset** [Zhang and Zhou, 2022]: This is a

### 3. Methodology

global dataset of near-surface maximum and minimum temperatures. It uses a series of ground-station measurements and satellite observations to gather the near-surface air temperature.

## 3.3. Random Forest Regression

As mentioned earlier in [section 2.5](#) about [RFR](#), the decision tree splits the node in two directions based on the criteria in the data features. In the context for terrain analysis, [RF](#) is a collection of decision trees that are trained on a subset of data using different features sourced from external datasets. The final prediction is then obtained by aggregating the prediction and perform averaging for regression. This method helps to minimise over-fitting due to a single decision tree. In order to fully implement the [RF](#) regression, data pre-processing of features dataset will need to be performed, which will discuss further in [subsection 3.3.1](#).

### 3.3.1. Data Pre-processing

Data pre-processing is a crucial step to align features data with the ICESat-2 height dataset and the remote sensing features. Due to the amount of remote sensing data to be processed to be used in the [RF](#) regressor. Each features has different types of data. While continuous data fits nicely with machine learning models, not all datasets use continuous data. Categorical data, for instance, requires further processing in order to fit it into the [RF](#) regressor. One-hot encoding proves better data clarity as it represents each category in a column with their respective data. Having this process ensures that the [RF](#) regressor can perform the best with more detailed information to make its predictions.

#### One-hot encoding



<b>id</b>	<b>LULC</b>
0	Trees
1	Built Area
2	Built Area
3	Crops
4	Built Area
5	Trees

<b>id</b>	<b>LULC_Trees</b>	<b>LULC_Built_Area</b>	<b>LULC_Crops</b>
0	1	0	0
1	0	1	0
2	0	1	0
3	0	0	1
4	0	1	0
5	1	0	0

Figure 3.3.: One-hot encoding on categorical data

Features that use categorical variables are represented with different values, each having its own classification and representation. There are no ordinal relationship between each data point. In the context of this research, '**Geomorpho90m Geomorphic Layer**' and '**ESRI 2020 Global Land Use Land Cover**' are examples datasets

### 3.4. Testing geometric features

that only categorical data is used. These categorical data often use varying naming conventions across datasets. For instance, 'Copernicus Global Land Cover' uses '20' to represent shrubs, '30' for herbaceous vegetation, '40' for cultivated and managed vegetation/agriculture, and so on. To represent non-ordinal relationships, *one-hot encoding* is a technique used to deal with Machine Learning ([ML](#)) algorithms that require numerical input. In machine learning, it provides a necessary representation to convey information about the categories.

*One-hot encoding* is a technique commonly used in machine learning and data analysis to convert categorical variables into a binary matrix representation [[Brownlee, 2017](#)]. Since some of the features in the selection are categorical, such as land use/land cover data, the application of one-hot encoding becomes essential. Figure [Figure 3.3](#) illustrates how one-hot encoding transforms categorical data into a binary matrix. Each column in the encoded matrix corresponds to a unique category encountered in the data. Therefore, if there are  $n$  unique categories, the resulting binary matrix will have  $n$  columns.

#### Data Normalisation

$$h_{norm} = \frac{h - \min(h)}{\max(h) - \min(h)} \quad (3.1)$$

Data normalisation aims to standardise the elevation values. It is to ensure to retain the consistent scaling by constraining the elevation values of the elevation data to a certain scale or range. As the areas of study have different height ranges, normalisation allows the ICESat-2 elevation data, `h_te_interp`, to scale and perform within the range between 0 and 1. The resulting heights,  $h_{norm}$  in [Equation 3.1](#), takes into account the maximum and minimum values, thus each geographic location can be normalised to the same scale. Within this research, a Min-Max scaling is used such that the normalised dataset retains its original distribution of data. For  $h_{norm}$ , 0 corresponds to the minimum value in the original ICESat-2 data, and 1 corresponds to the maximum value.

## 3.4. Testing geometric features

Before inserting data into [RF](#) regression, the testing of each geometric features will be performed. Geometric features used in this thesis is solely based on the relationship between the ICESat-2 points. The testing of each of the geometric feature will be able to identify the most suitable parameters in relation to each geometric points. For example, the test consist of testing the number of neighbours and its distance from its ICESat-2 point based on the accuracy of [RMSE](#). A number of test on each location will find the suitable number of neighbour that yields a more accurate result. The random forest regression will be based on the these geometry relationships that are

### *3. Methodology*

encoded in the geometric features to perform its prediction, therefore having a more accurate result by testing the number of neighbours will impact on the accuracy of the results.

f

#### **3.4.1. Implementation of Random Forest Regressor**

Previous research, namely [Hengl et al. \[2018\]](#) and [Sekulić et al. \[2020\]](#) has demonstrated that the use of random forests can result more accurate interpolation from a sample of elevation points compared to traditional interpolation methods. Their approach relies on the programming language, R, with a specific focus on the 'ranger' package for random forest implementation. In short, the 'ranger' package also implement random forest in fast and efficient manner in classification and regression tasks. This study, however, uses sci-kit learn library as the main implementation. Benchmarking between these two implementations of random forest is out of scope for this study.

Random Forest, as its name suggests, randomises both the features and target data during the training process. Instead of constructing a single decision tree from the entire dataset, it creates an ensemble of decision trees. This means that different trees in the forest see different subset of the whole dataset. After training on multiple decision trees, the model performs bootstrap aggregation (bootstrapping). This combines the predictions from all the trees and takes the mean of all the trees. This reduces the variance and improves overall accuracy.

#### **Test-train Split**

Test-train split is an intermediate step for evaluating the performance of random forest or any machine learning predictions in general. It divides a given dataset into two subsets: the training set and the testing set. In this particular research, the test-train split for the random forest regressor is 80% training set and 20% testing set. The subset for each testing and training sets are randomly selected from the ICESat-2 dataset, so it is selected uniformly throughout the whole dataset.

The training set is used to train the [RF](#) model. As previously discussed in [section 2.4](#), training takes the training data and performs the splitting at each node in the tree and arrive at an overall prediction after bootstrapping. The testing set, on the other hand, is used to assess the model's performance and evaluate how well it performs. Knowing the [RF](#) performance metrics can provide insight to the effectiveness of the training. Validation plays a key role in this study as the testing set can provide a baseline for evaluation of the model as well as for feature importance ranking.

### 3.4. Testing geometric features

#### Training the RF Model

Having all the features dataset combined, and splitting training and testing dataset. The training of the RF model involves inserting in the training dataset into the random forest. The training of each RF model uses the same location as the input data. The training dataset is a large data frame that consists of multiple columns representing each features except the target variable: elevation. Since the prediction is elevation, this would belong to the target variable, and the goal of the random forest is to predict the elevation based on the training dataset. For instance, the training of Tasmania, Australia would use the Geometric Features and Remote Sensing Features from Tasmania. This ensures the spatial relationship in the model relies its own feature dataset information for the results. The training of the RF model also repeats on different geographical locations: **Mount Taranaki, New Zealand, Grand Canyon, USA, South Limburg, Netherlands.**

The role of the testing dataset is to ensure there is a control dataset to compare against. Usually the ratio between training and testing dataset varies between 80%-20% or 75%-25% depending on the size of the training dataset and the number of data points. The ratio represents the percentage used as training and as testing datasets. In this study, 80% of the dataset will be randomly selected for training, and the remaining 20% for testing. The use of testing dataset helps to prevent the model from overfitting and to monitor the performance of the training by assessing and is used a comparing against training dataset. It is used as a validation tool measure to monitor the performance of training.

The exception to this training is the last test, where the model will be used by combining training dataset from three locations. This means the training of the model is built upon its local dataset but not on this new location. This unseen data will test the model's ability to adapt and application to non-local data, so this tests whether the spatial relationship built upon the combined dataset from non-local datasets is able to be applied on a new location.

#### Feature Importance and Ranking

Predicting the outcome using Random Forest in this projects requires a basket of features. These features can influence the weights and biases in the decision trees in the random forest algorithm. In order to interpret the outcome of the results, it is convenient to know the feature importance and it enables us to gain insight into which of the features influences the results the most. Also, in order to answer the question from [section 1.3](#), this research is going to quantify the contribution of each feature that has the most significance on the RF model's prediction.

Feature importance plays a significant factor in this study. In random forest, feature importance measures the features with the most influence on the prediction of the model [Breiman, 2001]. This is typically calculated based on the how much of

### *3. Methodology*

the feature contributes to reducing the impurity of the nodes of the trees. There are in general three calculations of feature importance, each has its advantages. The choice of the calculation methods depends on classification or regression problem. In general there are two methods suitable for regression: Permutation importance and Mean Decrease in Accuracy Mean Decrease in Accuracy ([MDA](#)), and Mean Decrease in Impurity Mean Decrease in Impurity ([MDI](#)) suitable for classification problem. In this study, the regression calculation will be addressed.

Permutation importance involves shuffling the values randomly of a features and measuring the impact on the model's performance compared to the validation dataset as baseline. Each of columns in the training dataset are then shuffled independently, meaning only the values of the feature columns are shuffled, and all target values remains unchanged. Each shuffling of the column values would produce a out-of-bag score, and is compared to the baseline score. This shuffling will be repeated exhaustively on all values on all feature columns and the average difference will be calculated as the feature importance. If shuffling a particular feature has a significant difference compared to baseline, it suggests that the feature is more important. The advantage is that breaks the relationship between the features variable sand the target variable, such that bias is reduced. It is, however, computationally expensive and repeating on multiple tests would require significant process time.

Mean Decrease in Accuracy ([MDA](#)) is used for splitting nodes in the decision tree. This information aids in feature selection, understand the model's behaviour to find the most influential features [[Breiman, 2001](#), [Scikit-Learn, 2023](#)]. [MDA](#) quantifies the average reduction in accuracy using Mean Squared Error ([MSE](#)) as metric, and assess each split based on this metric. The values of each feature are shuffled, and its metric is re-assessed and compared to the baseline validation dataset. It evaluates how much a feature's contribution by assessing the decrease in disorder when used for splitting used. Features with the higher [MDA](#) values means they are considered more important, which indicates a more impact on the decision tree and an overall contribution to making accurate predictions.

#### **3.4.2. Exporting and Combining [RF](#) models**

In order to compare the performance with different geographical locations from the study areas, the [RF](#) model are exported after training in each location. Each model are assessed individually and be combined to form a single model. This model are applied to an untrained location, out of the three selected locations. This tests the model's ability to adapt to different terrains, and subsequently assess its accuracy.

### 3.4.3. Testing and Evaluation

In order to assess the quality of the resulting [DEM](#), it is important to compare the prediction against the ground truth. This entails comparing the elevations from the predicted surface. This serves as a validation of performance for the [RF](#) regression results. In order to holistically evaluate the prediction of elevation in the [DEM](#), both the [RF](#) regressor and the results will also be evaluated.

Testing is done by overlaying the resulting [DTM](#) on ground truth. The resulting [DTM](#) is then subtracted from the ground truth to establish the areas where there are higher or lower than the ground truth. This diagram is a measurement to find areas that have deviation away from the ground truth. To find more detail of this variation, a scatter plot represents the discrepancies between prediction values from [RF](#) and the expected values from testing dataset. Each data point on the plot corresponds to a pair of values, that lies on the y-axis representing the predicted data and the x-axis represent the testing or validation data, the diagonal line shows the closeness to the ground truth each test presents. The distance of each point from a defined reference '0' level illustrates the extent of deviation. This graphical representation enables outliers and data points that deviates to understand the relationships of the [RF](#) model.

### The Loss Function

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.2)$$

It is important to assess the machine learning algorithm and assess their performance during training, fitting, and testing of data. The loss function assess whether the training and testing data to determine the outcome. Since each decision tree node splits two children each time, we utilise a selection criterion to make the decision to split. The objective is to minimise the loss function in order to result in better prediction of elevation from training data.

The loss function is also a crucial evaluation of the random forest performance. It is characterised by [MSE](#). [MSE](#) indicates how close the prediction are closer to the actual target values. The goal of random forest regression is to minimise the [MSE](#) as much as possible. Using the ensemble of trees in the random forest, the [RFR](#) algorithm would aggregate the prediction from multiple trees to minimise the over-fitting ensuring the effectiveness of the Random Forest Regression algorithm.

### 3.4.4. Comparison with Traditional Interpolation Methods

The test on the results on the random forest elevation prediction against traditional interpolation, the traditional interpolation are tested against the results from [RF](#), and is tested with the same error metrics. The purpose is to find whether the random

### 3. Methodology

forest interpolation is a better interpolation technique compared to traditional interpolation technique.

For [IDW](#), the parameters are set to: the nearest nearest 10000m and power weighting of 2. This would mean that the searching circle would encompass the surrounding 10km and look for the nearest ICESat-2 points surrounding the circle. For Voronoi diagram based interpolation, [TIN](#) and Laplace Interpolation are used. These three interpolation techniques are compared to to find the most suitable interpolation technique that subsequently be used to compare with [RF](#) regression. The reason to choose one from traditional technique is find one that can represent to compare between between [RF](#) technique and traditional interpolation. [Table 3.1](#) compares different interpolation using metrics such as minimum, maximum Error, Mean Average Error ([MAE](#)) and Root Mean Square Error ([RMSE](#)). The error metrics are explained in more details in [subsection 3.4.5](#).

Interp'n Method	Min Error	Max Error	MAE	RMSE
<a href="#">IDW</a>	-967.729	-138.021	205.893	287.405
<a href="#">TIN</a>	-942.682	-137.694	206.4669	287.478
Laplace	-927.121	-139.236	206.567	285.489

Table 3.1.: Comparison between Traditional Interpolation Methods

This preliminary test uses ICESat-2 dataset for interpolation using three different methods: [IDW](#), [TIN](#), and Laplace. These three traditional interpolation techniques have different interpolation approaches, but largely similar results in terms of [MAE](#) and [RMSE](#) as [Figure 3.4](#) have shown. The comparison with Laplace interpolation shows that in terms of [RMSE](#), it is marginally better than other interpolation techniques at 285.49m. The comparison shows that Laplace would be used as represent to compare with results from [RF](#) regression.

#### 3.4.5. Presenting final results and accuracy assessment

"Vertical accuracy is the main criterion in the specification of the quality of elevation data." [Uuemaa et al. \[2020\]](#) One of the error metrics used in this context is the [RMSE](#). [RMSE](#) (as shown in [Equation 3.3](#)) calculates the square root of the mean of the squared differences between predicted height and ground truth values. This offers insights into both the magnitude and direction of errors. The height difference between the predicted height and the ground truth height is also more generally called vertical accuracy.

Additionally, [MAE](#) (as shown in [Equation 3.4](#)) is also used to measure the average absolute difference between the predicted values and the actual observations, providing a clear indication of how close the predictions are to the true values. These error metrics are also used in previous research [[Hu and Ji, 2022](#), [Hawker et al., 2022](#)]. In

### 3.4. Testing geometric features

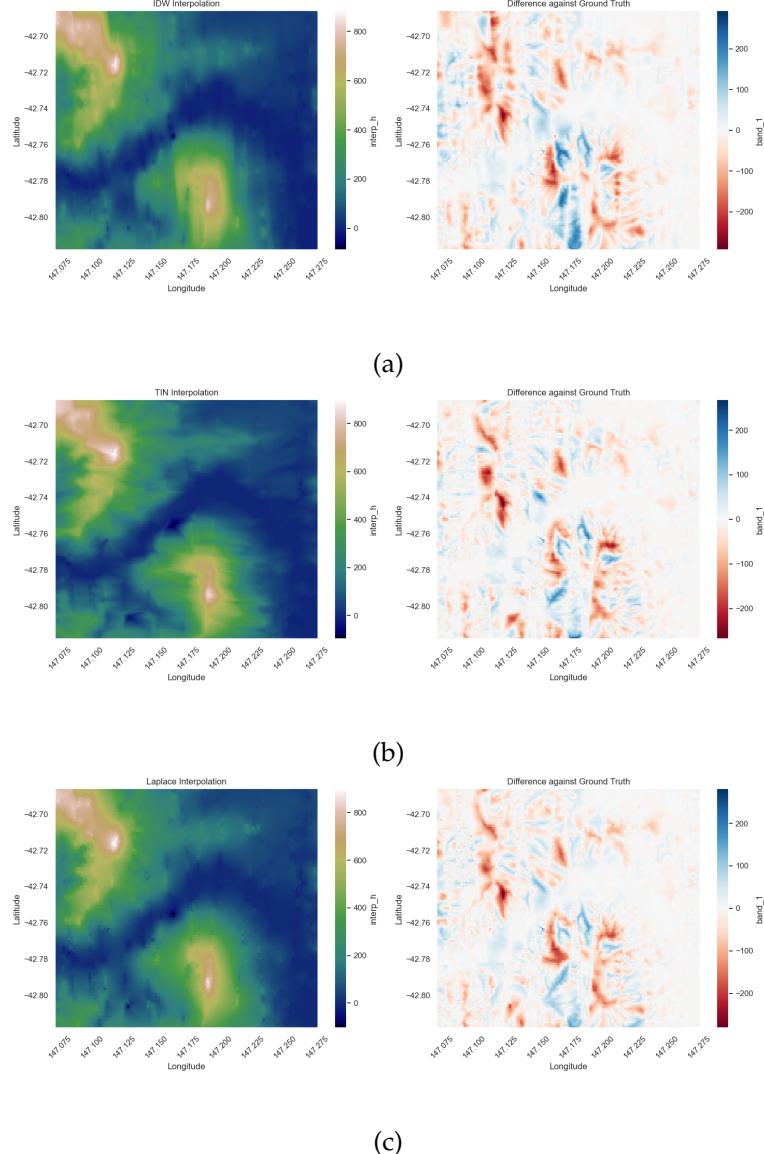


Figure 3.4.: Interpolation using Tasmania ICESat-2 Data using methods: (a) IDW (b) TIN (c) Laplace

addition to these error metrics, minimum error, maximum error, and standard deviation of error will also be used to determine the accuracy of the resulting [DEM](#).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{z}_i - z_i)^2} \quad (3.3)$$

### *3. Methodology*

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{z}_i - z_i| \quad (3.4)$$

To further assess the performance of the **RF** regression model, this would mean considering metrics such as R-squared coefficient ( $R^2$ ). The  $R^2$  is the proportion of variance in the testing data that is predicting the elevation from the **RF** regression. This is also one of the assessment of fit of the model in relation to the features and the prediction of the height variable. By incorporating a combination of these error metrics, a comprehensive evaluation of the Random Forest regression outcomes from ICESat-2 data can be conducted, ensuring the robustness and credibility of the results presented in this thesis.

# 4. Implementation

To elaborate on the datasets used, this chapter will delve into the two datasets used for training: the ICESat-2 Dataset and the Features Dataset that includes Geometric and Remote Sensing Features. The details on data access from NASA Earth-Data and Google Earth Engine are discussed and the implementation of RF regression, and the explanation of feature importance will be provided.

## 4.1. Access to Elevation Data

Since elevation dataset is the core of this study, the elevation information are accessed from the [DEM](#) from each of the local mapping agencies. Although the ICESat-2 mission has its elevation data acquired from [LiDAR](#), the comparison of results against ground truth needs to be comparable, and thus rules out the use of raw data from the ICESat-2 mission itself.

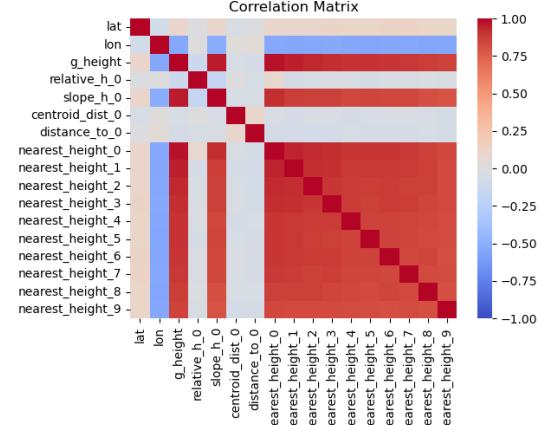
When seeking an appropriate [DTM](#) as ground truth, several criteria come into play. Firstly, the resolution must be less than 100 meters, ensuring that it remains equal to or smaller than the resolution presented in the results. Secondly, consider utilising a mapping agency directly sourced from the government. Lastly, prioritise data that is freely accessible, usable, shareable, and open for further development.

## 4.2. Correlation Matrix Among Geometric Features

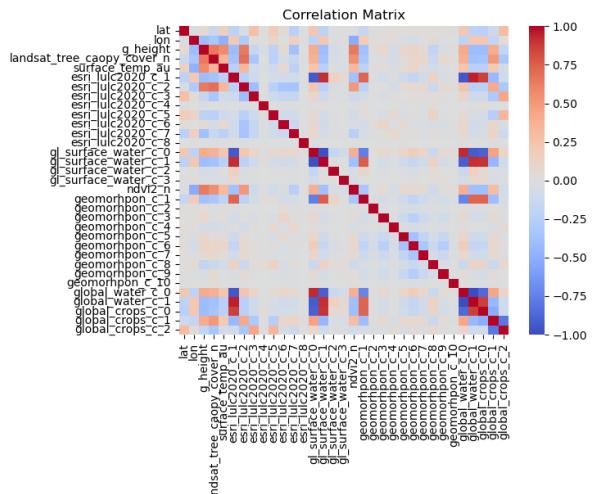
When assessing the Features Data, the correlation matrix provides a convenient way to visualise the relationships between different features used, and it is important to understand the relationships among the features. Having or not having correlation will have impact on the feature importance in the random forest, and this provides an insight to the relationship among the geometric features.

As "Tobler's First Law of Geography" [Miller, 2004] has claimed, the nearer points are more related than the points that are further away. [Figure 4.1a](#) shows the use of the correlation coefficient as a measure of strength and direction of relationship between two features. It ranges from 1 (indicating perfect positive correlation) and -1 (indicating perfect negative correlation), and 0 indicates no linear relationships.

#### 4. Implementation



(a) Correlation Matrix



(b) Correlation Matrix

Among the relationships in the correlation matrix, the nearest height (i.e. 1-nearest-neighbour) that are closer to the interpolated points shows a darker shade of red than ones that are further away (i.e. 9-nearest-neighbour). As the assumption indicates, more related points in darker shades are have distances that are close to the interpolated point, and as distance between the interpolated point and the ICESat-2 point mover further apart, the correlation decreases. As expected, the closer the points, the more related the data point is.

### 4.3. Details on the Study Area

In addition to Geometric Features, Remote Sensing Features shown in [Figure 4.1b](#) also has shown that there are correlations among the remote sensing features. The water-mask features, for example, shows opposite correlation with each other—areas water pixels has negative correlation with non-water pixels. Overall, however, there are limited amount of correlated features within the matrix save for water mask and land use and land cover dataset. Most of these are correlations to each group of features, and not across different features. This shows that the features does not have a strong leaning towards the other features in [RF](#) model.

## 4.3. Details on the Study Area

Four study areas is selected as the study for this thesis, and this will test the suitability of the machine learning algorithm on multiple terrains. Using **Tasmania, Australia** as a test case, this will be used to compare distributed dataset against ICESat-2 dataset on the performance of [RF](#) regression. And the other three locations represent different landscapes on Earth will be trained and tested, each with its own characteristics. [Figure 4.2](#) shows the maps for each of the respective study locations.

Study Area	Bounding Box ( $\min_{\text{lon}}, \min_{\text{lat}}$ ), ( $\max_{\text{lon}}, \max_{\text{lat}}$ )
Tasmania, Australia	(147.0611, -42.82818479), (147.2856, -42.6760)
South Limburg, Netherlands	(5.5826, 50.7417), (6.1637, 51.0827)
Grand Canyon, USA	(-112.2195, 35.9746), (-111.7676, 36.3165)
Mount Taranaki, New Zealand	(173.8336, -39.4686), (174.3158, -39.1469)

Table 4.1.: Coordinates of Study Area in EPSG:4326

## 4.4. Code Access and Data Downloading

In the spirit of *Open Science*, the datasets and code are publicly available on a GitHub repository: [https://github.com/leowlk/RF\\_Terrain](https://github.com/leowlk/RF_Terrain). The code includes algorithms such as filtering, data pre-processing, [RF](#) training and testing, and evaluation. The repository will also include a sample [DEM](#), provided as a sample of the expected results.

The *icespyx* python library provides all the necessary tools to download and extract the ICESat-2 dataset. This method requires an Earth Data account from [NASA \[2023\]](#). Alternatively, [OpenAltimetry \[2023\]](#) provides a user interface for downloading datasets within a predefined bounding box. This is useful for quick testing and deployment.

As for the Remote Sensing Features Dataset, individual datasets are found within their respective data organisations. However, it is recommended to utilise Google

#### 4. Implementation

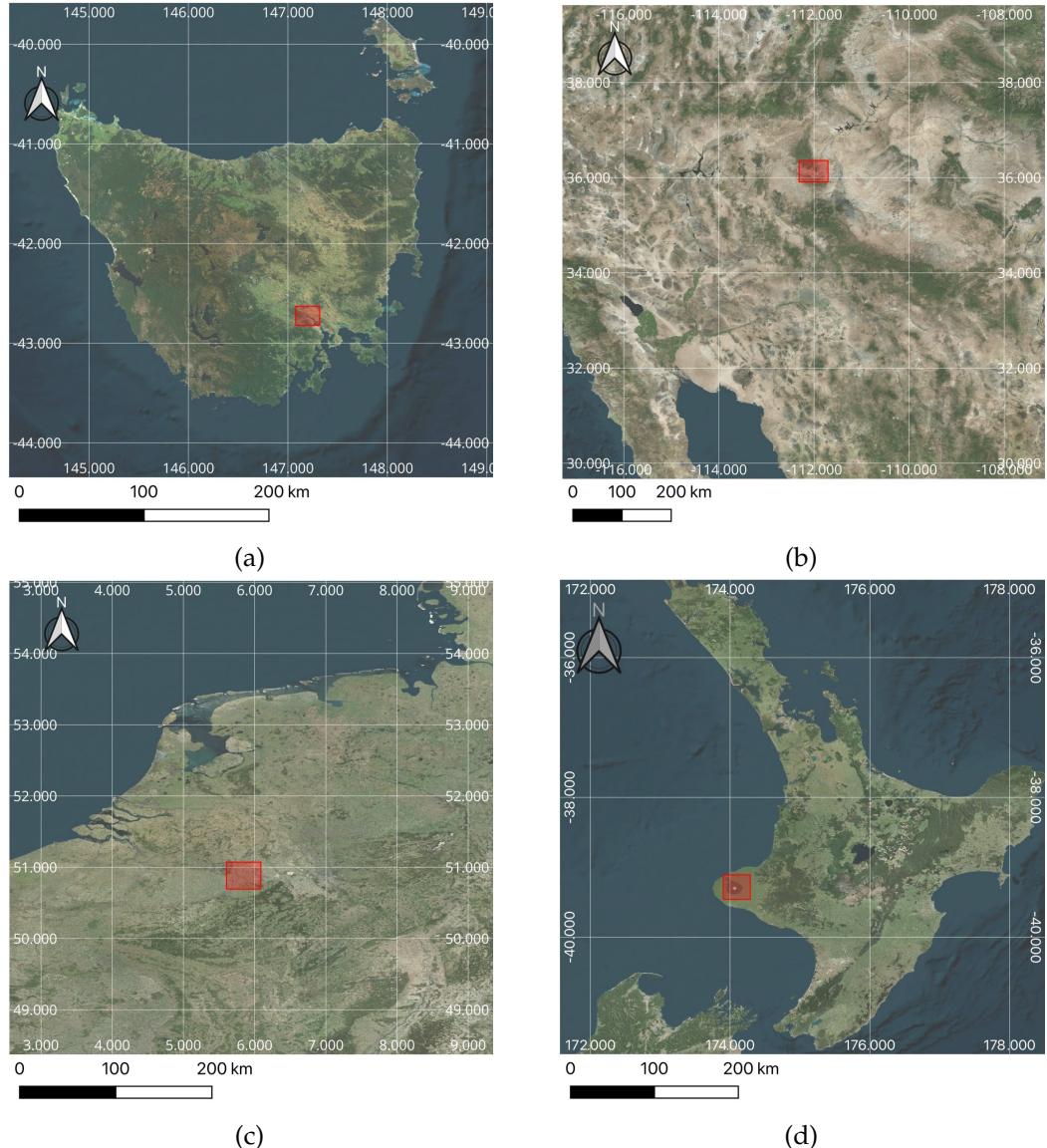


Figure 4.2.: Location maps of the study area. (a) Tasmania, Australia (b) Grand Canyon, USA, (c) Limburg, Netherlands, and (d) Mount Taranaki, New Zealand

Earth Engine as a centralised hub for accessing and downloading these data for machine learning features. Downloading datasets from Google Earth Engine requires a Google Account, and large files can be downloaded to Google Drive or other compatible cloud services.

# 5. Results and Analysis

Most of the ICESat-2 dataset consists of substantial gaps in non-polar regions due to the orbiting nature of the satellite. This chapter presents the results and provide analyses in order to answer the research questions on the suitability of using RF regression as an interpolation technique, and present the quality of the resulting predictions and its feature importance using this technique. It is used to show the data distribution and geographical differences among the study areas.

Since Geometric Features and Remote Sensing Features are used in this analysis, both are tested individually before combining the features to form the final results. These tests are important to assess the number of neighbours and remote sensing data in order to discuss their contribution to the RF model. These features are tested on the three different geographic locations, and the model of three locations are trained and applied to Tasmania, Australia. In addition, the use of RF model are compared against traditional interpolation techniques. To answer the research question, error metric, feature importance ranking, and the difference against ground truth are used in the results and analysis. Further discussion are made in this section.

## 5.1. Individually Training of Geometric Features

Each Geometric Feature are first tested with respect to the ICESat-2 points and measure its accuracy from its implementation. This step is significant as it helps to establish the optimal relationship of the features used in the RF model. These tests include the distance to neighbouring points of ICESat-2, nearest neighbour height, gradient between the elevations, and relative height difference. In summary, these tests will use the relationship of ICESat-2 points, and is summarised in [Table 5.1](#).

Tests		Number of Neighbours				
1	Distance to nearest ICESat-2	1	2	5	10	100
2	Nearest neighbour height	1	5	10	20	100
3	Gradient between ICESat-2 points	1	2	5	-	-
4	Relative height difference	1	2	5	10	-

Table 5.1.: Summary of Geometric Features Tests

In these tests, these features are used: Longitude, Latitude, and the tested Geometric Feature. The feature importance would be expected show similar or equal dominance

## 5. Results and Analysis

on latitude and longitude, but with only three features, the significance of feature importance are small in these preliminary tests and thus are not shown.

### 5.1.1. Test 1: Distance to Nearest ICESat-2 Point

While testing the features, n-number of neighbours are tested such that each subsequent neighbours reflects the relationship from the nearby elevation points. These relationships are important to analyse the sampled height that are part of the RF regression. From [Table 5.2](#), using low number of neighbours overall results in errors in the range of -372 to 571 metres, and [RMSE](#) of rangers from around 31 to 44 metres. The most accurate results from [Figure 5.2](#) shows that in terms of [RMSE](#), one neighbour displays the least error. Low number of neighbours—1 to 10—yields similar results, and as the number of neighbour increases, the [RMSE](#) increases.

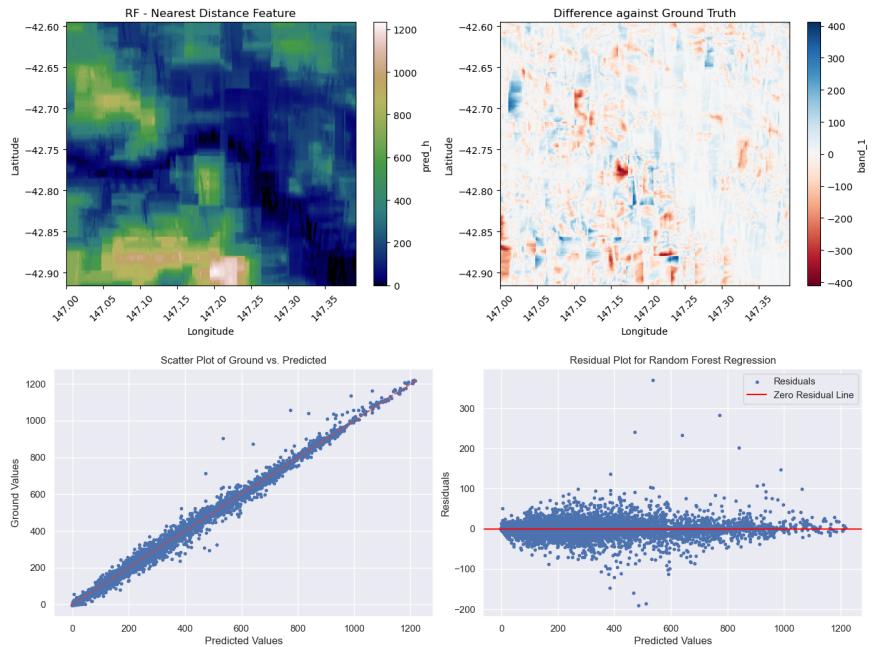


Figure 5.1.: (a) RF - Nearest Distance to ICESat-2 Point (b) Scatter Plot

N-neighbours	Min Diff	Max Diff	RMSE	MAE
1	-379.611	375.658	49.916	31.051
2	-372.352	352.439	50.075	31.276
5	-390.528	424.799	54.068	34.91
10	-398.333	371.783	56.325	36.784
100	-395.529	571.263	68.555	44.677

Table 5.2.: Distance to Nearest ICESat-2 Point

The reason for this rise in error is that the more data points in the features table

## 5.1. Individually Training of Geometric Features

mean that there are more data points that are not directly correlated to the nearby points. Owing to the fact that closer points are more correlated than neighbours that are further away. This is the reason smaller number of neighbours—from 1 and 10 neighbours—produces more accurate results. This results in the most accurate would be 1-closest neighbour to be used the feature used in the combined Geometric Features [RF](#) model. The reason for using distance to 1-closest-neighbour is the consideration of computational cost. Calculating larger number neighbours costs more computational power, especially with the magnitude of ICESat-2 data in this study of around 20 thousand to 30 thousand points in each study area. Even though having more than one neighbour would be more representative of the neighbourhood of the ICESat-2 points, the results show limited improvement for the computation resource required during computation.

### 5.1.2. Test 2: Nearest Neighbour Height

In this experiment, the relationship between the interpolated point and the height of the nearest ICESat-2 point as seen from [Table 5.3](#). This test tests the nearest neighbours of the interpolated point and using KD-tree to find the nearest n-neighbours from the interpolated point. Each number of neighbours will be aggregated in subsequent tests and to be used as features in the training and testing set.

It is expected that the relationship of the 1-nearest neighbour would follow a similar if not the same pattern as the nearest neighbour interpolation. This means the resulting [DTM](#) would be arranged in a Voronoi Diagram, with each Voronoi Cell of each ICESat-2 point would have a uniform height. As the number of nearest neighbour increases, the number of features in the features set would also increase.

N-neighbours	Min Diff	Max Diff	RMSE	MAE
1	-344.246	385.154	45.688	27.079
2	-341.152	387.564	44.025	25.242
10	-341.805	376.818	43.63	24.923
100	-337.345	377.037	43.737	25.06
200	-335.816	374.136	43.753	25.078

Table 5.3.: Nearest Neighbour Height of ICESat-2 Point

The errors of the nearest neighbour height yields similar results. As the number of neighbours increase, the [RMSE](#) gradually decreases to around 43 metres. The difference in maximum and minimum difference hovers from around -340 and around 380. The distribution of errors in the right hand chart shows that that both sides As the number of nearest neighbour increases from 20 to 200, the general errors have stabilised to around a similar range from -337m to 374m. In [Figure 5.2](#) shows the implementation of the most accurate solution as predicted form the [RF](#) regression at 10 neighbours.

## 5. Results and Analysis

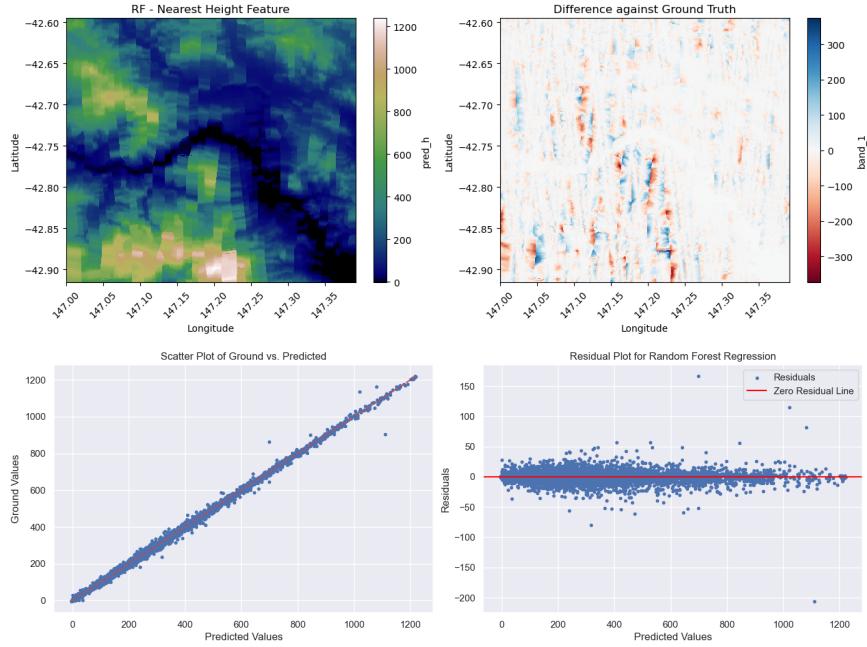


Figure 5.2.: (a) RF - Nearest Height to ICESat-2 Point (b) Scatter Plot

### 5.1.3. Test 3: Gradient to Neighbourhood Points

In [Table 5.4](#), this shows the test is going to test the relationship of the gradient slope and their neighbours. This relationship between the interpolated point and the height of the nearest ICESat-2 point will be tested using 1, 2, 5, and 10 neighbours. It will use KD-tree to find the nearest n-neighbours from the interpolated point. Each number neighbour are aggregated in subsequent tests and to be used as features in the training and testing set.

N-neighbours	Min Diff	Max Diff	RMSE	MAE
1	-391.77	394.306	47.677	29.076
2	-411.091	427.007	47.602	29.072
5	-537.632	412.667	52.541	32.416
10	-530.209	333.739	53.881	33.666

Table 5.4.: Nearest Neighbour Slope of ICESat-2 Point

As seen from [Table 5.4](#), the number of neighbour with the most accuracy is 2 with regard to the RMSE, and this is very close. In order to ensure the number of neighbours are included, having 2 neighbours can ensure that the gradient are from more than one source to prevent siding one direct line of gradient. A variety of data sources surrounding the neighbourhood will be used to calculate the gradient.

### 5.1. Individually Training of Geometric Features

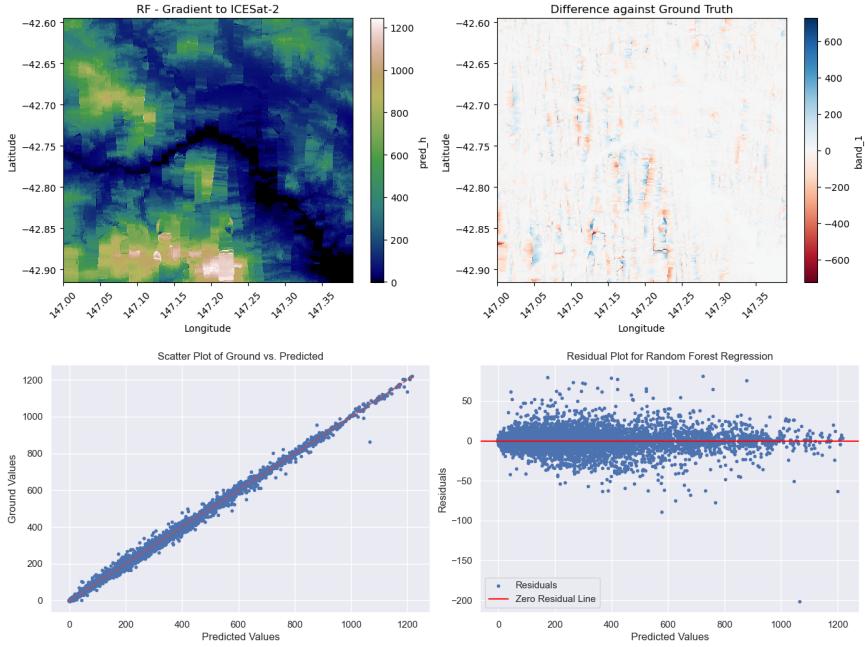


Figure 5.3.: (a) RF - Gradient to ICESat-2 Point (b) Scatter Plot

#### 5.1.4. Test 4: Relative Height To Neighbours

This test shown in [Table 5.5](#) is going to focus on the relative heights between the nearest ICESat-2 point and the heights of its surrounding neighbourhood. Similar to nearest height, this feature will, however, focus on the difference on the relative heights among the ICESat-2 points. For each relative distances for n-number of neighbours, this would contribute to n-number of columns in this feature.

N-neighbours	Min Diff	Max Diff	RMSE	MAE
1	-353.554	340.476	47.573	29.034
2	-390.509	334.424	48.724	30.545
10	-787.84	564.375	58.619	37.006
50	-773.813	550.006	69.181	42.806
100	-885.601	494.715	76.984	45.242

Table 5.5.: Relative Height to Nearest ICESat-2 Point

Similar to previous test, the most accurate from the [RF](#) prediction from using relative height is the nearest one neighbour. [Figure 5.4](#) shows the optimal from the nearest one neighbour from the prediction. In terms of errors, lower numbers of neighbours yield a better result than larger number of neighbours, where the [RMSE](#) of relative height between closest to 1-neighbour is 47.6 metres. As the number of neighbour increases, the [RMSE](#) and [MAE](#) increases accordingly. Smaller neighbour numbers result in better accuracy, and using just one neighbour saves computational power while providing

## 5. Results and Analysis

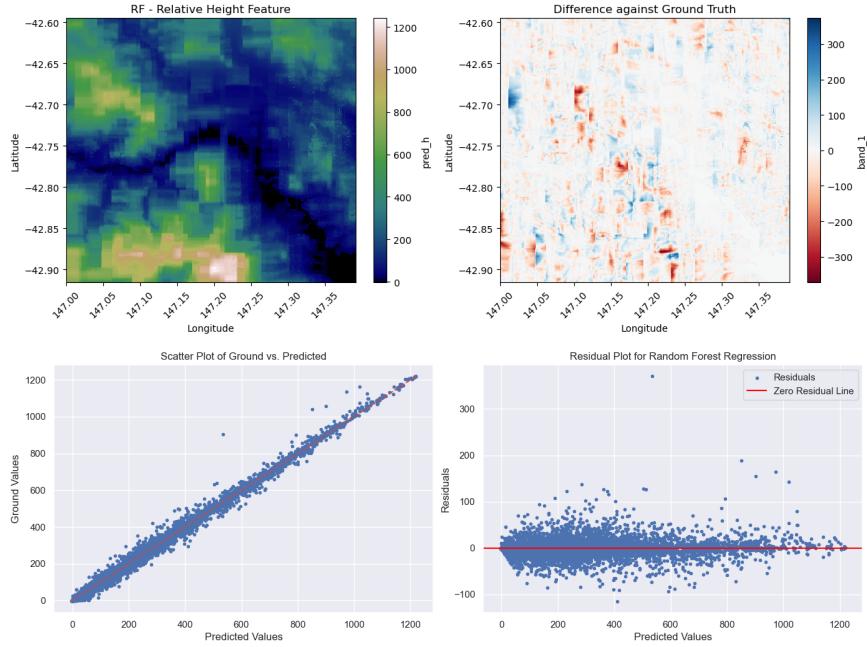


Figure 5.4.: (a) RF - Relative Height to ICESat-2 Point (b) Scatter Plot

the most accurate results in the tests.

### 5.1.5. Test 5: Remote Sensing Features

Remote Sensing Features adds a second layer of features in the [RF](#) regression, and would involve more information that may influence in the elevation prediction in the results. These remote sensing features are highlighted in [subsection 3.2.2](#). To test the Remote Sensing Features, these features data are added separately for each feature to analyse the effect in the [RF](#). This will then be reflected in the results.

This provides an insight into the random forest and its ability to incorporate itself into the results. The test is going to be based on the Tasmania Area. To find clearer comparison of the Remote Sensing Features, a baseline feature is used to provide the effect of new added features in the [RF](#) model. The baseline feature chosen is the nearest height, and this feature is added without Latitude and Longitude features. The addition Latitude and Longitude would create established underlying relationship and would be hard to test the effects of Remote Sensing Features. Therefore, a baseline feature using the distance to ICESat-2 geometric feature is used.

[Figure 5.6a](#) shows that by adding the water mask as a feature, the flat surface at the middle of the responds to the random forest. Water mask is a binary feature that indicates '1' where water is present and '0' when not present. This can be illustrated by the shape of the river present in the results. The error diagram on the

### 5.1. Individually Training of Geometric Features

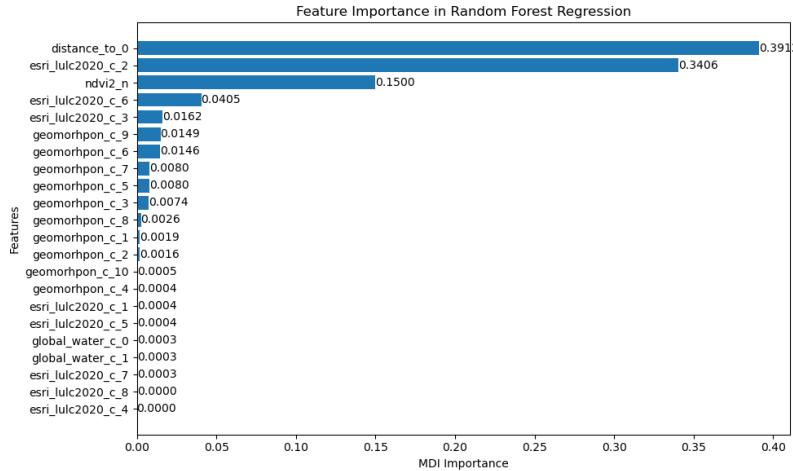


Figure 5.5.: Feature Importance from Remote Sensing Features Only

right in [Figure 5.6a](#) reflects that the area within the water mask feature is close to zero. This means adding water mask as feature would improve the error metric in the areas with water. Areas containing rivers, streams, sea, would therefore be a flat surface, which would also yield better results compared to ground truth.

Features	Min Diff	Max Diff	RMSE	MAE
RS Features (a)	-1236.442	1147.563	328.218	243.487
RS Features (a), (b)	-1238.844	1103.648	322.218	242.487
RS Features (a), (b), (c)	-1109.608	1098.945	237.764	160.393
RS Features (a), (b), (c), (d)	-1109.35	943.361	211.063	140.779

Table 5.6.: Additional Features (a) Water Mask (b) Geomorphon (c) Land Use and Land Cover (d) [NDVI](#)

When the features of (a) Water Mask and (b) Geomorphon were added to the regression, the flat surfaces responded as expected in the results. In flat surfaces as indicated by the water surfaces and the difference against ground truth are be minimised such that water surfaces are flat, and yielded close to the ground truth. The contribution of (b) Geomorphon in [Figure 5.6b](#) has also shows that the peaks and the valleys and flat surfaces of the terrain. This, however, did not contribute to the overall improvement in error metrics as was expected. It has only improved the [RMSE](#) very marginally from 328.218 metres to 322.218 metres, and [MAE](#) from 243.487 metres to 242.487 metres.

In addition to the last two features, a third feature was then added (c) Land User/-Land Cover in [Figure 5.6c](#). This indicates the land use and land cover classification. The response of the added feature here further minimise some areas closer to the ground truth. There is still presence of a lot of artefacts, but the effect of the feature

## 5. Results and Analysis

added to the random forest has improved the compared to using features (a) Water Mask and (b) Geomorphon.

Finally adding (d) NDVI to the random forest as seen in [Figure 5.6d](#) shows further sees improvement of the error metrics and also the difference against ground truth. This shows the random forest responds well to the remote sensing features as they are added one by one, and the improvement from one feature to four features is drastic. RMSE and MAE has improved from 328.218 to 211.063, and 243.487 to 140.779 respectively. Overall, this shows that with Remote Sensing Features alone is able to contribute to the prediction of elevation, albeit the predicted elevation does is not directly correlated from the elevation of ICESat-2 points. Their contribution to RF is nonetheless an improvement to the method.

[Figure 5.5](#) shows the individual feature contribution to the results with four features. It shows that the each of the features contributes roughly from 0.34 to 0.003. Land Use and Land Cover is overall the most influential among the features followed by NDVI. It is worth noting The Geometric Features of Distance to ICESat-2 point is retained in the features list in order to maintain the spatial relationship among the pixels in the prediction.

## 5.2. Training RF model by Combining Features

After testing Geometric Features and Remote Sensing Features, the training of the RF model are based on the previous tests, and Geometric Features uses the results of the number of neighbours listed in [Table 5.7](#). These are part of the training data set used to train the RF model. The training of the model involves organising the features columns into a data frame. This can ensure the tests from previous sections are honoured. Combining the Geometric Features and Remote Sensing Features allows further analysis on the behaviour of each feature together.

Tests	No. of Neighbours
1 Distance to Nearest ICESat-2	1
2 Nearest Neighbour height	10
3 Gradient between ICESat-2 points	1
4 Relative Height Difference	1

Table 5.7.: Geometric Features Tests

With reference to [subsection 3.4.4](#), Laplace interpolation has the best error metrics from traditional interpolation, therefore it is used as a comparison guide to measure the performance of RF models in subsequent tests.

[Figure 5.7](#) shows that using all the features yield results that shows the most dominant geometric feature is the 'nearest neighbour height'. At almost 1.0, the 'nearest neighbour height' feature almost dominates the feature importance in RF. In terms of

### 5.3. Application on Geographic Locations

Interp'n Method	Min Diff	Max Diff	RMSE	MAE
Laplace	-281.367	169.54	38.543	24.108
RF(All Features)	-345.907	357.817	43.342	24.384
RF(All Features, except Nearest Neighbour Height)	-597.919	620.033	55.711	34.986

Table 5.8.: Tasmania Data Interpolation and RF Prediction

error metrics, [Table 5.8](#) shows that in terms of [RMSE](#), RF(All Features) performs worse than Laplace Interpolation, but [MAE](#) shows similar average errors. This is unexpected given the fact that the use of both Geometric and Remote Sensing Features have given a range of dataset for [RF](#) model.

The [RF](#) regression has no significant improvement on the Tasmania's data. In fact, in terms of [RMSE](#) and [MAE](#), the [RMSE](#) have slightly worsen from to 278.836m to 295.312m but [MAE](#) improved from 207.965m to 205.955m. Comparing to the results from traditional interpolation, using dataset from randomly distributed data has slight worsen the overall prediction of elevation.

Due to its dominance of the feature 'nearest neighbour height' it is removed and tested again. This would let the remote sensing features contribute to more the prediction of the elevation, and it can be seen from [Figure 5.8c](#) that the distribution of features importance are more equal, where tree canopy remote sensing data and surface temperature forms came first and second at 0.36 and 0.24, as shown in [Figure 5.8](#). This shows that random forest can perform interpolation without the need of very closely correlated features, albeit produced less accurate results overall. [Figure 5.8](#) also shows the top five most important features except 'nearest neighbour height' (Geometric Feature) in descending order, are Landsat Tree Canopy Cover, Surface Temperature, Latitude, Longitude, and Land Use Land Cover. Water mask, even though its effectiveness established previously is able to bring water surfaces in the [DTM](#), ranks close to the bottom of the list of feature importance.

## 5.3. Application on Geographic Locations

Having tested on Geometric Features and Remote Sensing Features, combining these features on different geographical locations is the next logical step. This section focuses on implementation on other locations in this study, namely, **Grand Canyon, USA**, **South Limburg, Netherlands**, and **Mount Taranaki, New Zealand**.

All three locations have produced results that reflects local terrain, and would explore further the comparison with Laplace interpolation. In terms of the error metrics, all of them produces artefacts in the results that deviate from the ground truth. Individual performance from the scatter diagram (in [Figure 5.10a](#), [Figure 5.12a](#) and [Figure 5.14a](#)) shows that all these study locations produced results that are close to the

## 5. Results and Analysis

diagonal line of predicted and actual elevation. Most of the residuals shown in the scatter graphs in all locations concentrates around the zero residual line and fits well, and with small number of individual points that lies farther out of the zero residual line.

Location	Min Diff	Max Diff	RMSE	MAE
Tasmania, Australia	-342.941	362.697	43.343	24.378
Grand Canyon, USA	-658.465	656.926	76.173	39.438
South Limburg, Netherlands	-120.219	83.171	7.861	3.862
Taranaki, New Zealand	-758.779	546.25	28.45	11.107

Table 5.9.: Combined Features among Study Areas

### 5.3.1. Grand Canyon, USA

The location **Grand Canyon, USA** shows an area that represents more variations of the landscape. In the study area, the river valley and the peak ranges from around 600 to over 2000 meters in elevation. This is in stark contrast to the previous study location of the Netherlands. The differences in terrain and the distribution of ICESat-2 data should reveal the robustness and the accuracy of the [RF](#) model in terms of accuracy to the ground truth as well as to traditional interpolation, Laplace interpolation.

Interp'n Method	Min Diff	Max Diff	RMSE	MAE
Laplace	-792.423	931.947	109.017	69.526
RF	-653.547	661.584	76.177	39.448

Table 5.10.: Prediction with Geometric Features

The results of Grand Canyon shows that by using Geometric Features alone is able to produce a [DTM](#) that is close to the ground truth. Although the main differences, positive and negative differences, compared to the ground truth follows closely to to the ICESat-2 tracks. This is expected due to the fact that the nearest height being a dominant feature in the [RF](#) regression.

The Grand Canyon shows that more complex terrain areas produces more areas of errors, especially around the river gorges areas. Other terrain that area away from the river gorge, by contrast, are relatively less error prone, which signifies the difference between simpler and complex terrains. The improvement from Laplace interpolation shows that there are areas in Grand Canyon that performs better using [RF](#), these are due to the fact that the terrain have less variation. The results is that the influence from Remote Sensing Features and Geometric Features is able to establish relationship more accurately across the ICESat-2 track due to the simpler terrain.

### 5.3. Application on Geographic Locations

#### 5.3.2. South Limburg, Netherlands

The South of Limburg, Netherlands has more terrain features comparatively to the coastal region in Western Netherlands. The combined Geometric features shown in [Figure 5.12](#) shows that a simpler terrain produces comparatively less error compared to ground truth. The minimum differences shown in [Table 5.11](#) sees that there is improvement of the minimum difference from -206.447m to -120.75, and [RMSE](#) has overall from 9.544m to 7.867m. The simpler terrain features in the Netherlands was help due to the fact of the simpler terrain, and easier for the model to establish relationship across the ICESat-2 track similar to the Grand Canyon.

Interp'n Method	Min Diff	Max Diff	RMSE	MAE
Laplace	-206.447	65.012	9.544	4.235
RF	-120.75	86.09	7.867	3.867

Table 5.11.: Prediction with Geometric Features

The residual scatter graph shows that the predicted and actual points concentrates around the zero-residual line, and this also shows well in the [RMSE](#) and [MAE](#) being 7.861 and 3.862 respectively. This means using Geometric Features alone is able to produce results that are already quite close to the ground truth. The fact that the influence of nearest neighbour height is 0.90 shows correlation, but not as dominating as Tasmania of 0.99.

#### 5.3.3. Mount Taranaki, New Zealand

[Figure 5.14](#) shows that the cone shape of the volcano is a relatively simple terrain, and the peak is visible. Although some artefacts can be seen, which follows the ICESat-2 tracks, this overall performance of using random forest produces results that fits will in the residual plot.

Interp'n Method	Min Diff	Max Diff	RMSE	MAE
Laplace	-467.64	234.029	26.520	22.881
RF	-314.416	249.204	26.499	22.880

Table 5.12.: Prediction with Geometric Features

The overall difference between the ground truth shows that the errors concentrates around the slope close to the peak, this is due to the fact that the changes in terrain lies in between the tracks, and it can be difficult to register the details of the changing terrain between the tracks without additional information. This is an anticipated outcome due to the lack of data point between the tracks. This type of terrain example also confirms that simpler terrain features suits well because of the lack of variation in height differences between the tracks. Similar to the flatter surfaces Grand Canyon, [RF](#) is performs better in these conditions.

## 5. Results and Analysis

### 5.4. Use of Pre-trained RF model on Tasmania, Australia

After having seen differences of error metrics that are presented in previously in **Tasmania, Australia**, **Grand Canyon, USA**, **Mount Taranaki, New Zealand** and **South Limbrug, Netherlands**. The trained models from the latter three study areas are then applied to **Tasmania, Australia** to find whether these models are robust in modelling an unknown location using **RF** regression.

[Figure 5.15](#) shows that the resulting **DTM** performs quite similar to the ones tested from Tasmania, Australia. The fact that it used data trained from non-local data sources shows that spatial relationship are not entirely established from the error metrics presented in [Table 5.13](#)

Interp'n Method	Min Diff	Max Diff	RMSE	MAE
Laplace	-281.367	169.54	38.543	24.108
RF(Geometric Features)	-342.562	358.449	43.356	24.398
RF(All Features)	-345.907	357.817	43.342	24.384
RF(Combined Models)	-341.878	392.526	43.759	24.606

Table 5.13.: Error Metrics from Combined Models and All Features Test from Tasmania, Australia

[Table 5.13](#) shows that the in terms of **RMSE** and **MAE** figures, RF(Geometric Features) and RF(All Features) displayed similar results from previous models, the combined models, however shows little improvement overall, in fact, slightly less accurate compared to the models trained from local datasets. The **RMSE** slightly worsened from 43.356 to 43.759, though the significance of the difference is limited. It shows that using combined model does not always yield better results than using local datasets as feature data.

With all things considered, the significance of this change is not enough to conclude that using combined models trained from non-local models perform worse, but it shows no significant change can be established between using non-local models and combined features from local features. Compared to Laplace Interpolation, the performance from the combined model is worse in terms of **RMSE** at 43.759m compared to 38.543m; and **MAE** at 24.606 compared with 24.108. From [Figure 5.15](#), one can establish that overall, the combined model is able to perform elevation prediction in the **DTM**, and elevation difference against ground truth, like previous, follow closely along the ICESat-2 tracks. For other areas between the tracks can vary widely based on the terrain complexity.

#### 5.4. Use of Pre-trained RF model on Tasmania, Australia

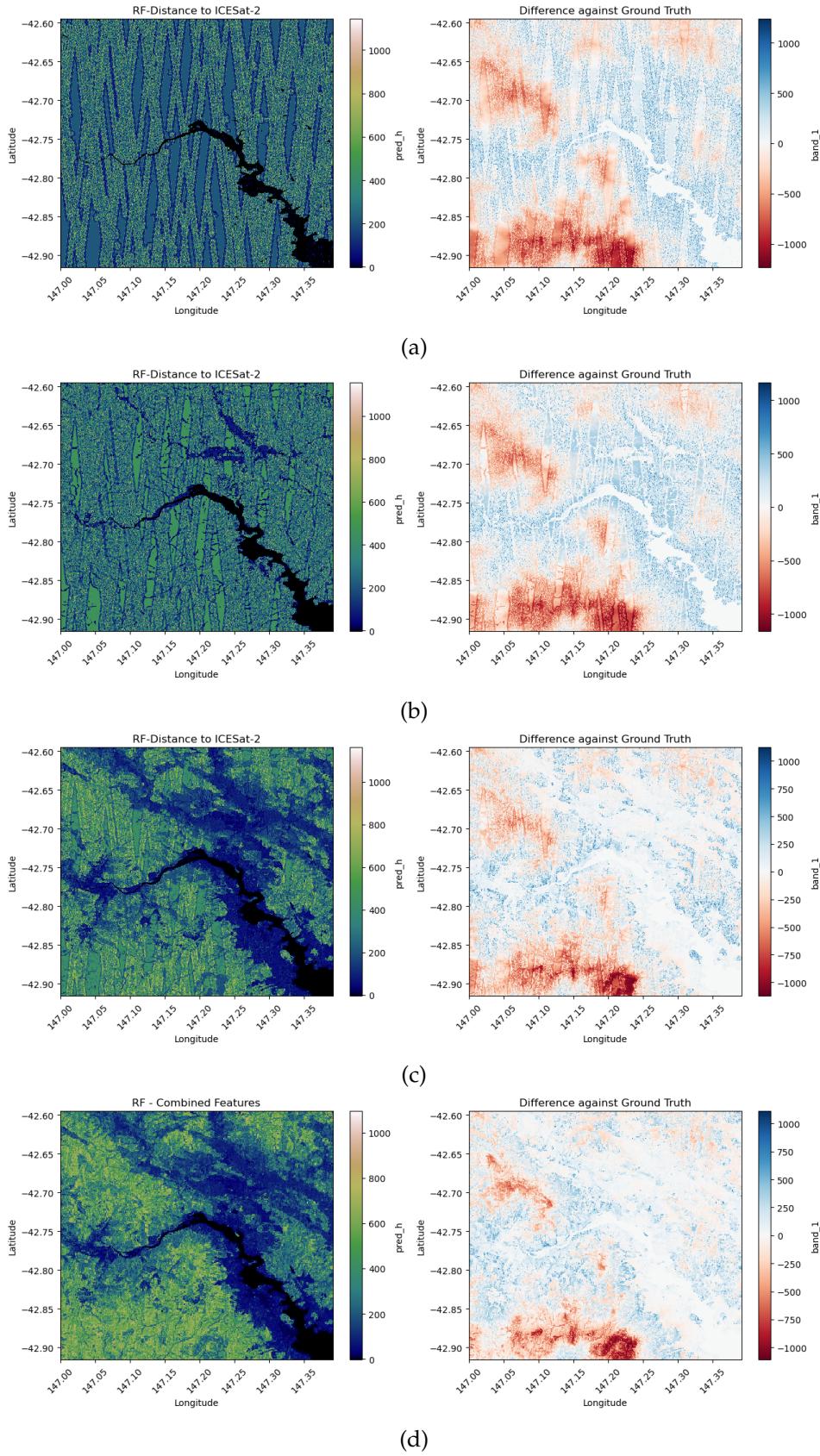


Figure 5.6.: Subsequently add remote sensing features to the random forest features and results from adding features: (a) Water Mask (b) Geomorphon (c) Land Use and Land Cover (d) **NDVI**

## 5. Results and Analysis

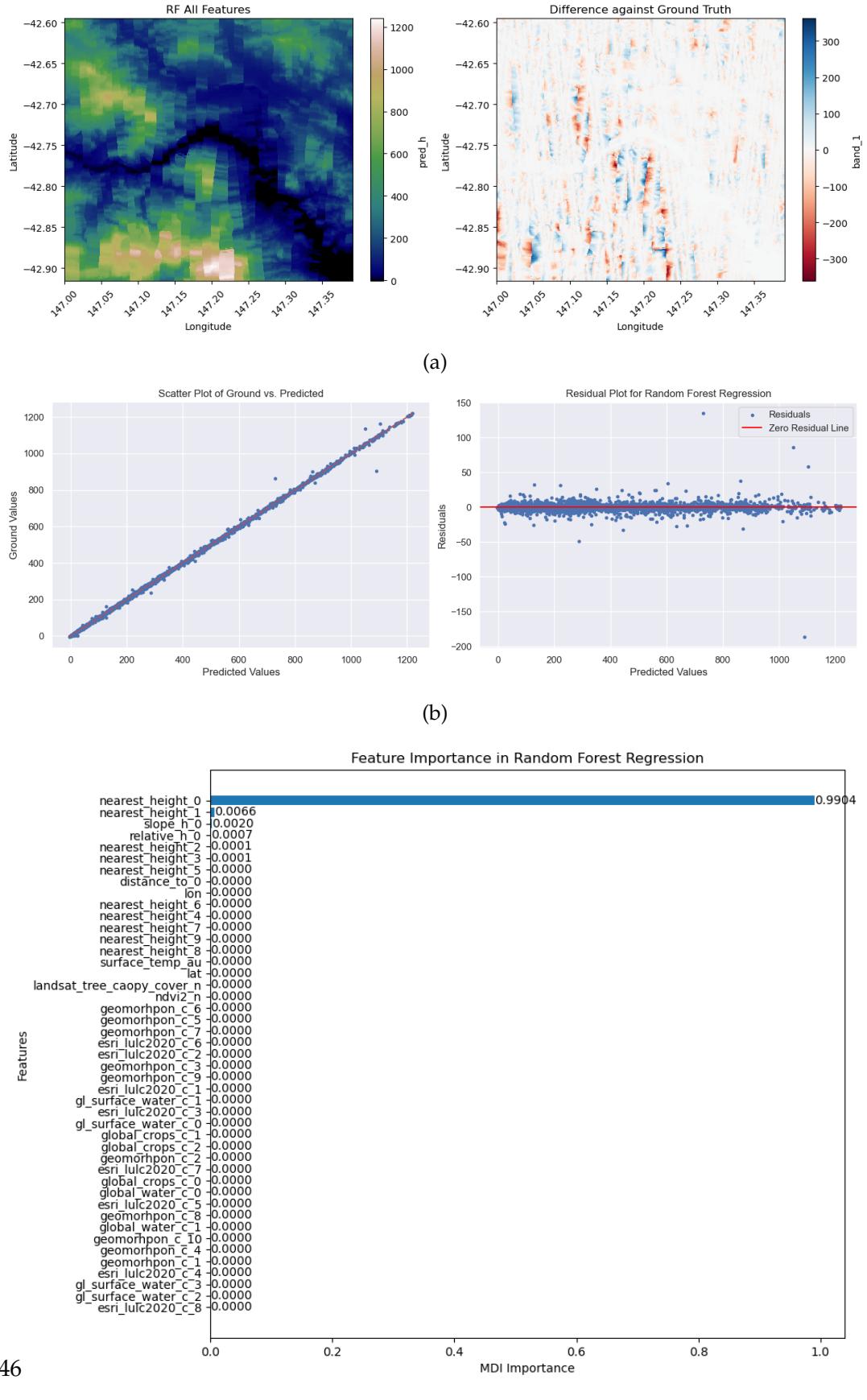


Figure 5.7.: RF model using All Features on ICESat-2 Data

#### 5.4. Use of Pre-trained RF model on Tasmania, Australia

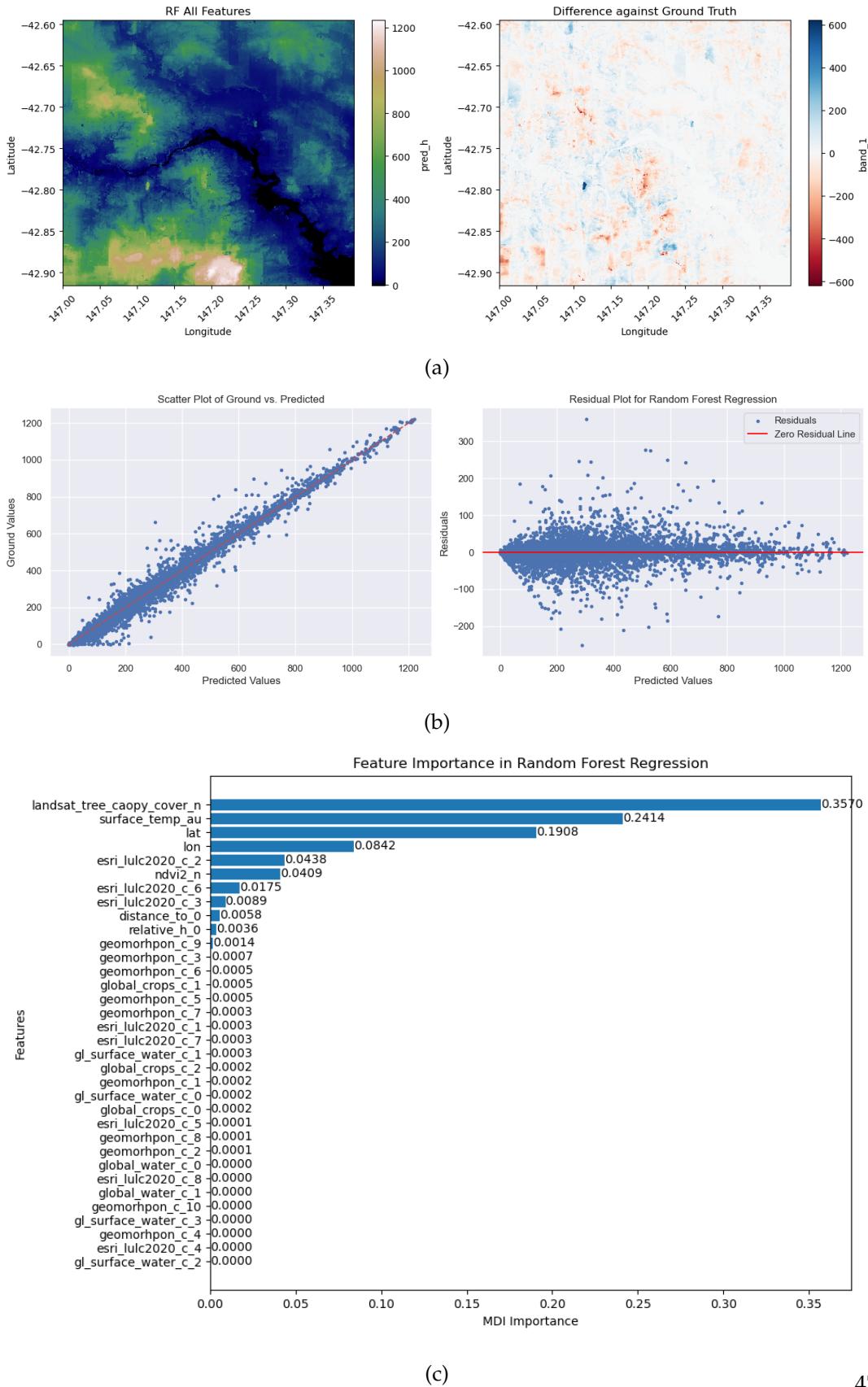


Figure 5.8.: RFR using All Features without Nearest Neighbour Height on ICESat-2 Data

## 5. Results and Analysis

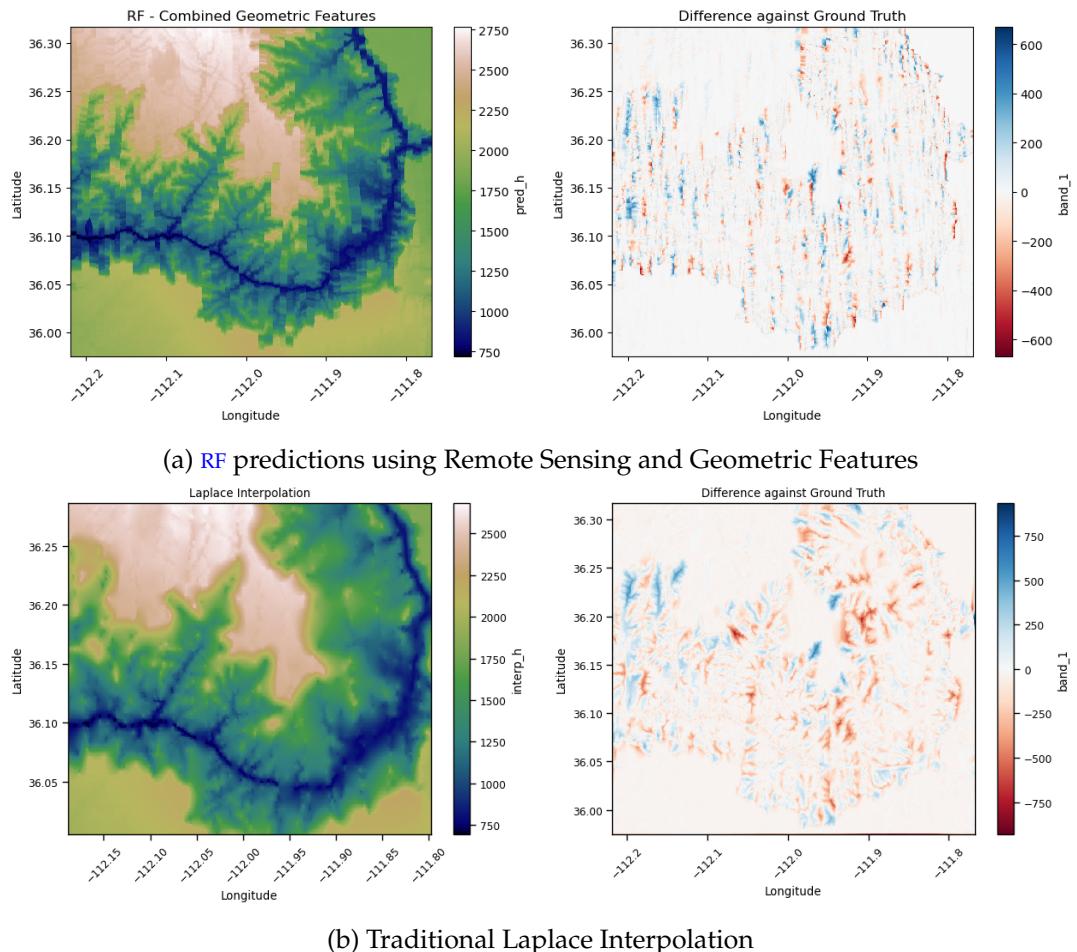


Figure 5.9.: **Grand Canyon, USA** Comparing Different Methods

#### 5.4. Use of Pre-trained RF model on Tasmania, Australia

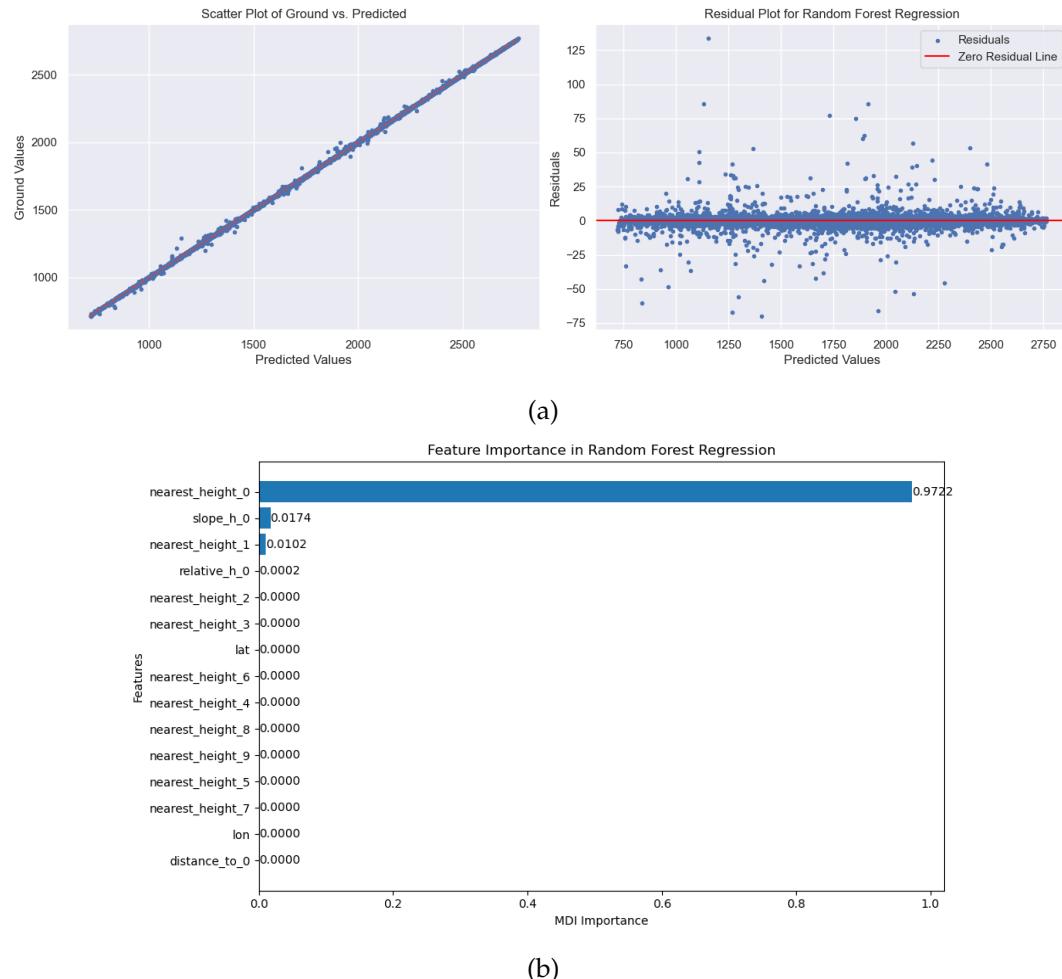


Figure 5.10.: (a) **Grand Canyon**–Combined Geometric Features (b) Scatter Residuals Plot (c) Features Importance

## 5. Results and Analysis

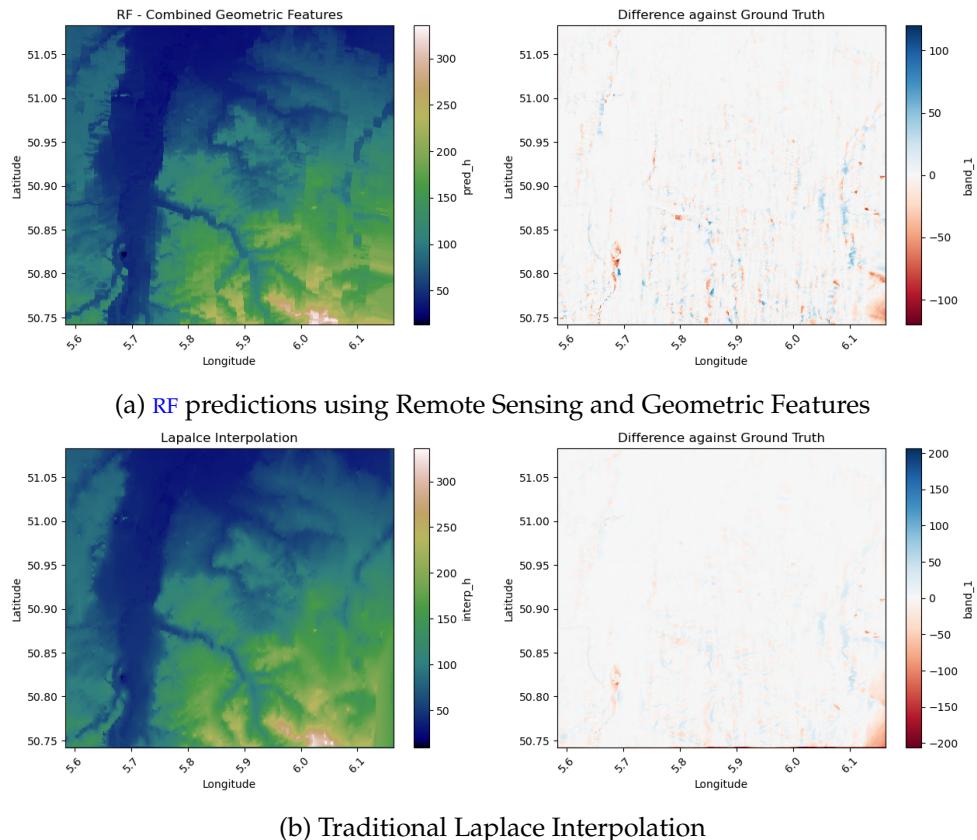


Figure 5.11.: **South Limburg, Netherlands** Comparing Different Methods

#### 5.4. Use of Pre-trained RF model on Tasmania, Australia

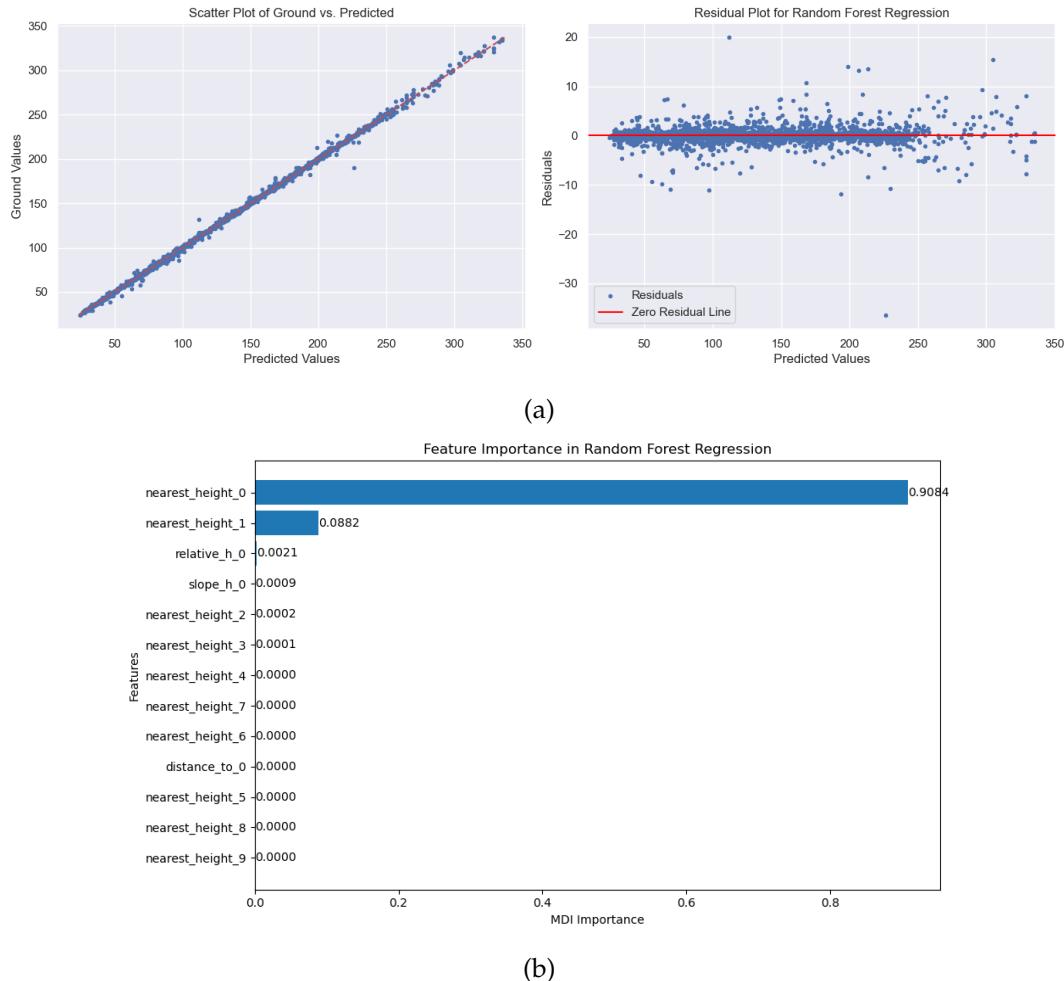


Figure 5.12.: (a) **South Limburg**—Combined Geometric Features (b) Scatter Residuals Plot (c) Features Importance

## 5. Results and Analysis

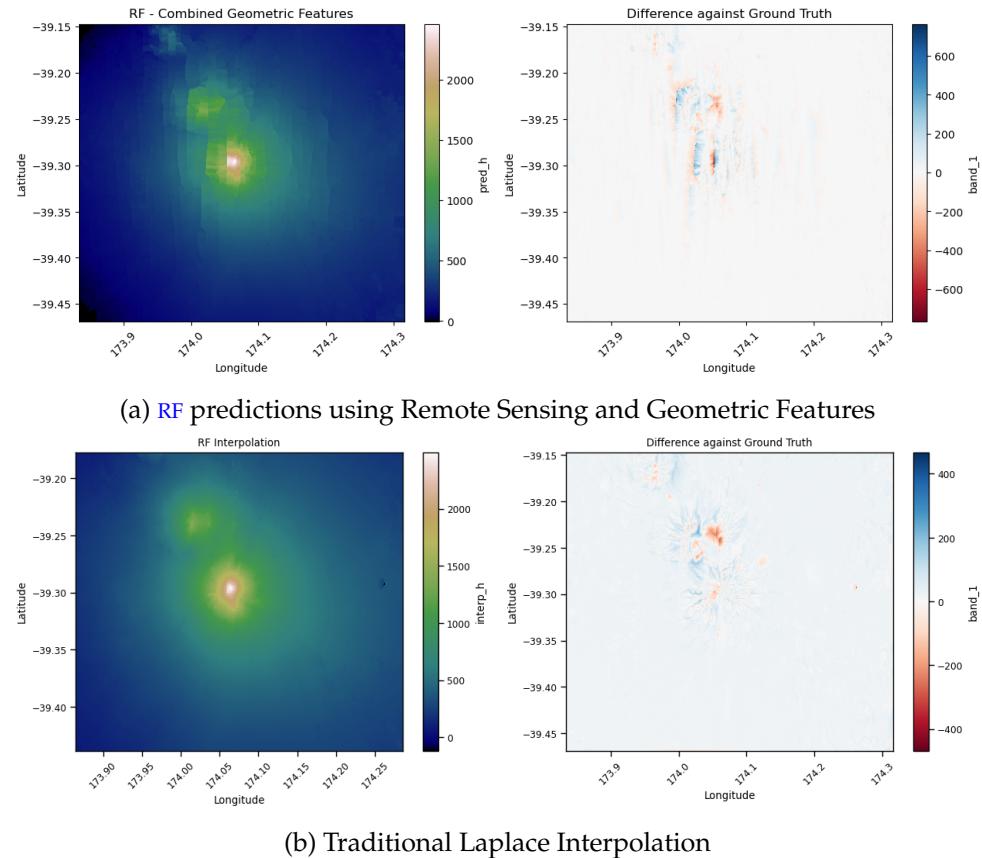


Figure 5.13.: Mount Taranaki, New Zealand Comparing Different Methods

#### 5.4. Use of Pre-trained RF model on Tasmania, Australia

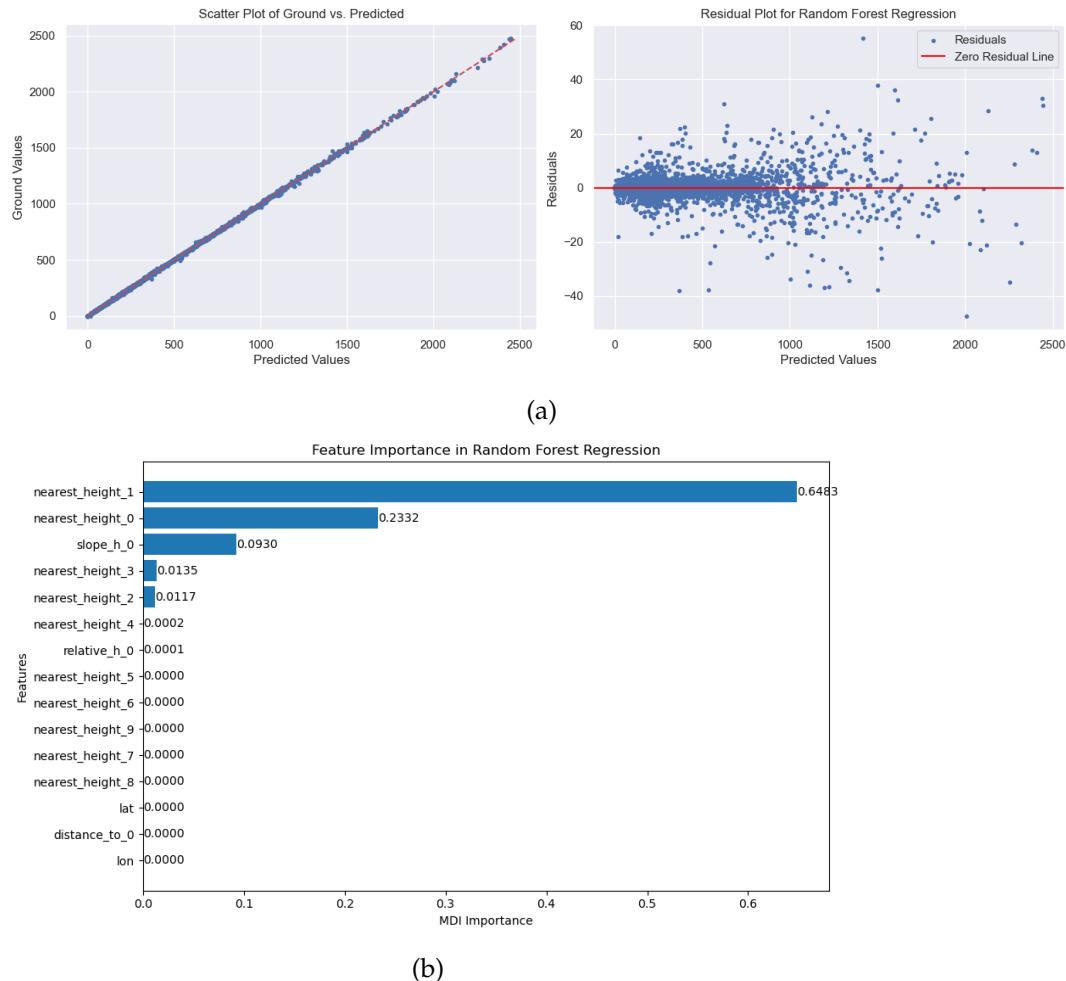


Figure 5.14.: (a) **Mount Taranaki**– Combined Geometric Features (b) Scatter Residuals Plot (c) Features Importance

## 5. Results and Analysis

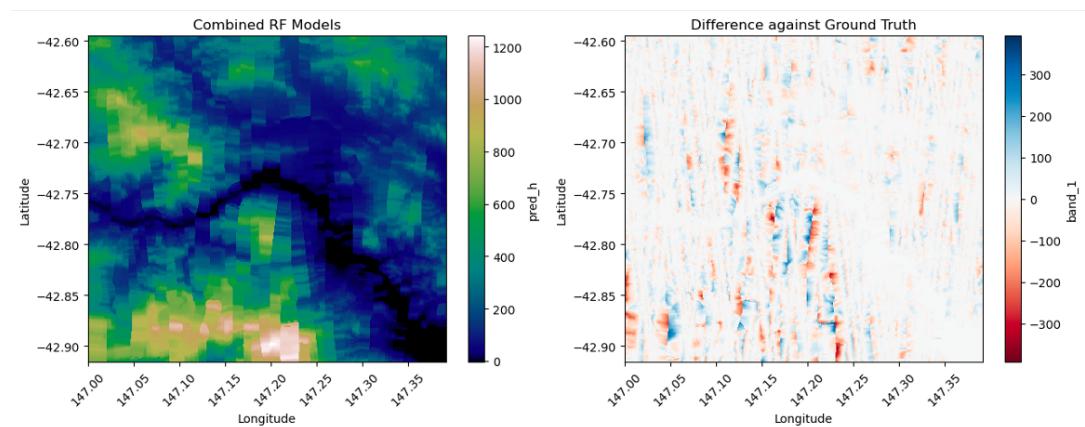


Figure 5.15.: RF result trained from three geographic locations

## 6. Discussion and Conclusion

To conclude this study, this section is going to sum up this study on random forest as a means of performing spatial interpolation in sparse datasets such as ICESat-2. The dataset used are elevation data sourced from local [DEM](#) and used the pattern of the ICESat-2 mission to perform spatial interpolation. The term interpolation in the context of this study means the prediction of elevation across the study area to find if random forest is suitable as a technique. Also, the use of Geometric Features and Remote Sensing Features allows the algorithm to build relationships between all of the features, and finally presenting the elevations and use the results to compare with the original [DTM](#) values.

Considering the fact that the resolution is set at 100m, the unsampled locations are tested on the accuracy of the machine learning algorithm along with the features used in this study. In general, Geometric Features have the most impact and present the most importance in the accuracy of the [RF](#) regression model. Closely correlated features such as **nearest neighbour height** is very dominant and consistently stands at 0.9 or in importance in all geographical locations. When the most dominant feature is eliminated, however, Remote Sensing Feature of Landsat Tree Canopy Cover stands at 0.36, which is significantly less influential. This shows that in most cases, Geometric Features are the most dominant features in elevation prediction, and the correlation plays a significant role in the results. The more correlated the feature is, the more significance in terms of feature importance to the [RF](#) model.

The accuracy of the results varies across the board depending on the complexity of the terrain. For terrains that such as the cone-shaped volcano, then [RF](#) performs comparatively worse compared to Laplace interpolation. More complex terrains such as the Grand Canyon, the [RF](#) algorithm is able to construct a terrain that focus more prominent terrain features due to the use of auxiliary data. Compared to Laplace interpolation, the peaks and valleys are less visible as a result. Overall, different geographical locations perform differently in terms of accuracy. Flatter terrain such as Netherlands tend to perform better than more complex terrains.

With reference to the background literature, [RF](#) does not inherently have an "explicit (geo)statistical model" [[Hengl et al., 2018](#)], and having more data, more covariates, more data crunching algorithms as training dataset does not necessary lead to drastic improvement in accuracy. Although there are sample size and resolution differences, it shows that the [RF](#) Loss function Mean Square Error (MSE) tend to be quite high at 48124.16, and R-squared: 0.6428 for 500 sample points at Lake Tahoe, California,

## 6. Discussion and Conclusion

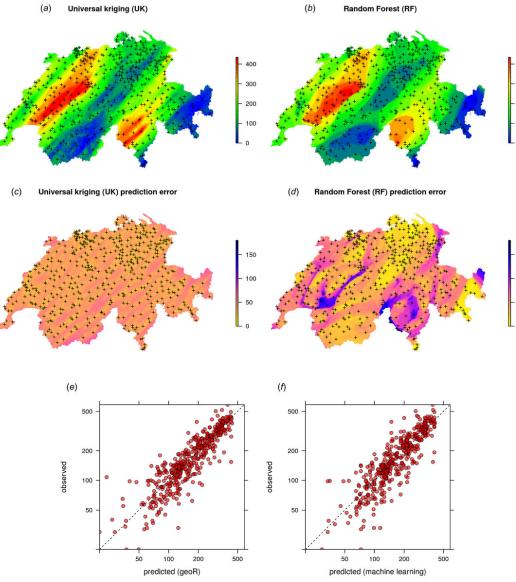


Figure 6.1.: Figure showing prediction v actual scatter graph  
[Hengl et al. \[2018\]](#)

USA; Loss function MSE of 0.0052395 and R squared of 0.8511794 at 157870 sample points at Illinois and Indiana, USA [Hengl et al. \[2018\]](#). Comparing this to the method used in this study, the [RF](#) metrics for Tasmania stands at Loss function MSE at 860.610 and R-squared at 0.985 using 49796 sampled points.

The comparison in Tasmania have shown that using [RF](#) regression is not always yield better results than traditional interpolation such as Laplace interpolation. In fact, there even some test where the [RMSE](#) and [MAE](#) has worsened after introduction of geometric features to the model. However, using the method to other geographic locations, the comparison between Laplace interpolation and [RF](#) regression have improved the error metrics with nearest neighbour height being the most dominant at over 0.9. This shows that using [RF](#) as a prediction of elevation is not always a accurate option compared to traditional interpolation. The cost of computation power and time to generate a model might cost more time instead.

After eliminating the most dominant Geometric Features (namely **nearest neighbour height**), it can be concluded that features without close correlation between each other and ground elevation can be used in the random forest to get results that are close to the ground truth. The advantage of using this approach is that the remote sensing features such as *water-mask*, *land-use* and *land-cover* features are able to provide more information to the random forest and get better estimation without using information from [gDEM](#).

Since the main constraint of this research is to perform [RF](#) regression on a terrain without any global DEM and its derivative of analytical features such as water shed

analysis, flow accumulation, surface roughness etc. This is the primary difference between this study and the study done by Hengl et al. [2018]. Despite the results was not as expected, more research can be further investigated. As the chapter 5 have established, different types of terrain complexity would produce different error measure, and it has been tested that simpler terrain produces better accuracy than traditional interpolation. Further investigation can focus on different complexity of terrains or focusing on adjacent terrains. Using a new measurement of Geometric and Remote Sensing Features that caters to terrain complexity may be an area of research. The final DTM, though without DEM derived features, RF regression can still be able to achieve a small improvement on elevation prediction against traditional interpolation techniques on selected terrains.



# A. Reproducibility self-assessment

## A.1. Marks for each of the criteria

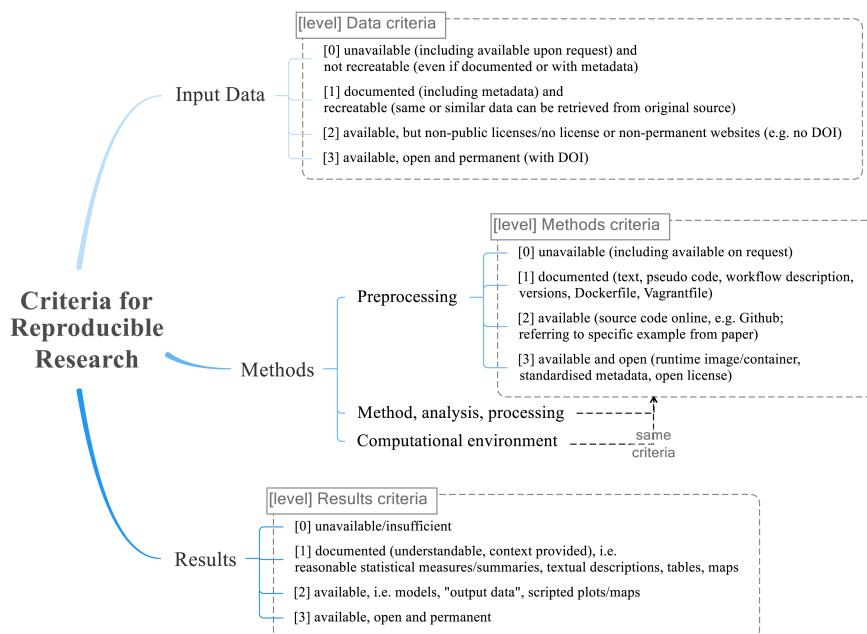


Figure A.1.: Reproducibility criteria to be assessed.

Grade/evaluate yourself for the 5 criteria (giving 0/1/2/3 for each):

Criteria	Grade
Input data	3
Preprocessing	2
Methods	2
Computational environment	2
Results	1

Table A.1.: Self Evaluation

## A. Reproducibility self-assessment

### A.2. Reflection

To reproduce the results from this thesis, there are several tools that needed accounts, namely NASA EarthData account and Google account for Google Earth Engine. The data used in this research can be accessed freely on the internet from data sources such as NASA, Open Altimetry [OpenAltimetry \[2023\]](#), and Open Topography [Open-Topography \[2023\]](#). However, obtaining the data requires more than clicking download on the browser. The data gathering took quite time for me, since the *icepyx* python package needed an account and the example code did not explain too much on the specific ATL08 data that I needed. The trial and error at the initial phase took some time, but I managed to download the required data after reading the docs and forums. The ICESat-2 elevation can be quite easily accessed once familiar with the *icepyx* package.

Moreover, Google Earth Engine is another data source that I used to gather Remote Sensing data. These data are freely available on Earth Engine itself, and also data from the GEE community catalogue. GEE community catalogue is a supplement data source that has more data featuring Population, Socioeconomic, Soil, Biological, Physical data etc. Earth Engine's data are normally sourced from government agencies whereas GEE community catalogue are sourced from other researchers that are willing to make data available openly for free.

To reproduce the results, analysis tools are crucial to begin analysis, and most of the packages are available in Python Anaconda, such as scikit-learn, geopandas, and xarray. These tools are also freely available. To use these tools, it took time to browse the documentation and online forums for its implementation is available for help. Overall, to reproduce the results from the thesis, the experimentation of different algorithms such as different types of Random Forest algorithms published form scikit-learn, Ranger, XGBoost since each of these algorithms can have different results, each of these tools are tested to find the most accurate.

The biggest hurdle in my opinion is the coordinate system change. Since the ground truth from different location uses different coordinate system, aligning of these coordinates system needs some work especially when calculating the euclidean distance between the ICESat-2 points and the unsampled location. The solution is to calculate it in their respective coordinate system and then transforming these into to EPSG:4326 for alignment between all the locations. For exporting results, I find that xarray is a better tool than geopandas for geospatial data because it allows me to export in the desired coordinate system in one line of code. I appreciate the use of xarray, and I find this tool largely hassle-free to use.

Although reproducing the results from data sources are not always straightforward, but using different python tools and reading documentation is able to find solution along the way. There are different consideration when aligning with different coordinate systems, and I find that there are always more elegant solution, but for the purpose of this research, the functionality of transforming coordinates systems is good

## *A.2. Reflection*

enough for producing the results for this research.



## B. Features Data

Name of Feature	Resolution	Derived
World Settlement Footprint (WSF) 2019	10m	Landsat imagery used for generating the WSF evolution <a href="#">Marconcini et al. [2020]</a>
Geomorpho90m Geomorphometric Layers	90m	Derived from MERIT-Digital Elevation Model <a href="#">Amatulli et al. [2020]</a>
ESRI 2020 Global Land Use Land Cover from Sentinel-2	10m	Sentinel-2 imagery at 10m resolution <a href="#">Karra et al. [2021]</a>
GFSAD Landsat-Derived Global Rain-fed and Irrigated-Cropland Product (LGRIP)	30m	Landsat 8 time-series satellite sensor data for the 2014-2017 <a href="#">Teluguntla et al. [2023]</a>
ASTER Global Water Bodies Database (ASTWBD) Version 1	30m	Created ASTER Global Digital Elevation Model (ASTER GDEM) Version 3 data product by the Sensor Information Laboratory Corporation (SILC) in Tokyo <a href="#">NASA/METI/AIST/Japan Spaceystems And U.S./Japan ASTER Science Team [2019]</a>
Bare Earth's Surface Spectra 1980-2019	250m	identify global bare surface areas and their dynamics based on multitemporal remote sensing images to aid the spatiotemporal evaluation of anthropic and natural phenomena <a href="#">Dematté et al. [2020]</a>
JRC Global Surface Water Mapping Layers	30m	Classification of Water <a href="#">Pekel et al. [2016]</a>



# Bibliography

- AHN. Home. Mar. 2023. URL <https://www.ahn.nl/>. Publisher: AHN.
- G. Amatulli, D. McInerney, T. Sethi, P. Strobl, and S. Domisch. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. *Scientific Data*, 7(1):162, May 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0479-6. URL <https://www.nature.com/articles/s41597-020-0479-6>. Number: 1 Publisher: Nature Publishing Group.
- P. V. Arun. A comparative analysis of different DEM interpolation methods. *The Egyptian Journal of Remote Sensing and Space Science*, 16(2):133–139, Dec. 2013. ISSN 1110-9823. doi: 10.1016/j.ejrs.2013.09.001. URL <https://www.sciencedirect.com/science/article/pii/S1110982313000276>.
- L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- J. Brownlee. Why One-Hot Encode Data in Machine Learning?, July 2017. URL <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>.
- P. Burrough. Principles of geographical information systems for land resources assessment. *Geocarto International*, 1(3):54–54, Jan. 1986. ISSN 1010-6049, 1752-0762. doi: 10.1080/10106048609354060. URL <http://www.tandfonline.com/doi/abs/10.1080/10106048609354060>.
- R. Bürgmann, P. A. Rosen, and E. J. Fielding. Synthetic Aperture Radar Interferometry to Measure Earth's Surface Topography and Its Deformation. *Annual Review of Earth and Planetary Sciences*, 28(1):169–209, 2000. doi: 10.1146/annurev.earth.28.1.169. URL <https://doi.org/10.1146/annurev.earth.28.1.169>. eprint: <https://doi.org/10.1146/annurev.earth.28.1.169>.
- L. Croneborg, K. Saito, M. Matera, D. McKeown, and J. van Aardt. *Digital Elevation Models*. Sept. 2020. doi: 10.1596/34445.
- J. C. da Silva Júnior, V. Medeiros, C. Garrozi, A. Montenegro, and G. E. Gonçalves. Random forest techniques for spatial interpolation of evapotranspiration data from Brazilian's Northeast. *Computers and Electronics in Agriculture*, 166:105017, Nov.

## Bibliography

2019. ISSN 0168-1699. doi: 10.1016/j.compag.2019.105017. URL <https://www.sciencedirect.com/science/article/pii/S0168169919302315>.
- M. de Berg, O. Cheong, M. Van Kreveld, and M. Overmars. Delaunay Triangulations: Height Interpolation. *Computational Geometry*, pages 191–218, 2008. doi: 10.1007/978-3-540-77974-2\_9. URL [http://link.springer.com/10.1007/978-3-540-77974-2\\_9](http://link.springer.com/10.1007/978-3-540-77974-2_9).
- J. A. M. Demattê, J. L. Safanelli, R. R. Poppiel, R. Rizzo, N. E. Q. Silvero, W. d. S. Mendes, B. R. Bonfatti, A. C. Dotto, D. F. U. Salazar, F. A. d. O. Mello, A. F. d. S. Paiva, A. B. Souza, N. V. d. Santos, C. Maria Nascimento, D. C. d. Mello, H. Bellinaso, L. Gonzaga Neto, M. T. A. Amorim, M. E. B. d. Resende, J. d. S. Vieira, L. G. d. Queiroz, B. C. Gallo, V. M. Sayão, and C. J. d. S. Lisboa. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. *Scientific Reports*, 10(1):4461, Mar. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61408-1. URL <https://www.nature.com/articles/s41598-020-61408-1>. Number: 1 Publisher: Nature Publishing Group.
- L. Hawker, P. Uhe, L. Paulo, J. Sosa, J. Savage, C. Sampson, and J. Neal. A 30 m global map of elevation with forests and buildings removed. *Environmental Research Letters*, 17(2):024016, Feb. 2022. ISSN 1748-9326. doi: 10.1088/1748-9326/ac4d4f. URL <https://dx.doi.org/10.1088/1748-9326/ac4d4f>. Publisher: IOP Publishing.
- T. Hengl, G. B. M. Heuvelink, and D. G. Rossiter. About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301–1315, Oct. 2007. ISSN 0098-3004. doi: 10.1016/j.cageo.2007.05.001. URL <https://www.sciencedirect.com/science/article/pii/S0098300407001008>.
- T. Hengl, M. Nussbaum, M. N. Wright, G. B. Heuvelink, and B. Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, Aug. 2018. ISSN 2167-8359. doi: 10.7717/peerj.5518. URL <https://peerj.com/articles/5518>.
- M. Hu and S. Ji. Accuracy evaluation and improvement of common DEM in Hubei Region based on ICESat/GLAS data. *Earth Science Informatics*, 15(1):221–231, Mar. 2022. ISSN 1865-0481. doi: 10.1007/s12145-021-00721-3. URL <https://doi.org/10.1007/s12145-021-00721-3>.
- K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby. Global land use / land cover with Sentinel 2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4704–4707, July 2021. doi: 10.1109/IGARSS47720.2021.9553499. ISSN: 2153-7003.
- K. Kraus and J. Otepka. DTM Modelling and Visualization – The SCOP Approach. Jan. 2005.

## Bibliography

- D. G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6): 119–139, Dec. 1951. doi: 10.10520/AJA0038223X\_4792. URL [https://journals.co.za/doi/10.10520/AJA0038223X\\_4792](https://journals.co.za/doi/10.10520/AJA0038223X_4792). Publisher: Southern African Institute of Mining and Metallurgy.
- Z. Li. A comparative study of the accuracy of digital terrain models (DTMs) based on various data models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 49(1): 2–11, Feb. 1994. ISSN 0924-2716. doi: 10.1016/0924-2716(94)90051-5. URL <https://www.sciencedirect.com/science/article/pii/0924271694900515>.
- Z. Li, K. Wang, H. Ma, and Y. Wu. An Adjusted Inverse Distance Weighted Spatial Interpolation Method. pages 128–132. Atlantis Press, Nov. 2018. ISBN 978-94-6252-620-4. doi: 10.2991/cimns-18.2018.29. URL <https://www.atlantis-press.com/proceedings/cimns-18/25907186>. ISSN: 2352-538X.
- A. Liaw and M. Wiener. Classification and Regression by randomForest. 2, 2002.
- M. Marconcini, A. Metz-Marconcini, S. Üreyen, D. Palacios-Lopez, W. Hanke, F. Bachofer, J. Zeidler, T. Esch, and E. Strano. World Settlement Footprint (WSF) 2015, 2020. URL [https://springernature.figshare.com/articles/World\\_Settlement\\_Footprint\\_WSF\\_2015/10048412/1](https://springernature.figshare.com/articles/World_Settlement_Footprint_WSF_2015/10048412/1). Artwork Size: 2688996507 Bytes Pages: 2688996507 Bytes.
- T. Markus, T. Neumann, A. Martino, W. Abdalati, K. Brunt, B. Csatho, S. Farrell, H. Fricker, A. Gardner, D. Harding, M. Jasinski, R. Kwok, L. Magruder, D. Lubin, S. Luthcke, J. Morison, R. Nelson, A. Neuenschwander, S. Palm, S. Popescu, C. Shum, B. E. Schutz, B. Smith, Y. Yang, and J. Zwally. The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment*, 190:260–273, Mar. 2017. ISSN 00344257. doi: 10.1016/j.rse.2016.12.029. URL <https://linkinghub.elsevier.com/retrieve/pii/S0034425716305089>.
- R. J. Michaelides, M. B. Bryant, M. R. Siegfried, and A. A. Borsa. Quantifying Surface-Height Change Over a Periglacial Environment With ICESat-2 Laser Altimetry. *Earth and Space Science*, 8(8):e2020EA001538, Aug. 2021. ISSN 2333-5084, 2333-5084. doi: 10.1029/2020EA001538. URL <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020EA001538>.
- H. J. Miller. Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, 94(2):284–289, June 2004. ISSN 0004-5608, 1467-8306. doi: 10.1111/j.1467-8306.2004.09402005.x. URL <http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.2004.09402005.x>.
- NASA. ICESat-2, Jan. 2023. URL <https://icesat-2.gsfc.nasa.gov/>.

## Bibliography

- NASA/METI/AIST/Japan Spacesystems And U.S./Japan ASTER Science Team. ASTER Global Digital Elevation Model V003, 2019. URL <https://lpdaac.usgs.gov/products/astgtmv003/>.
- A. L. Neuenschwander, K. L. Pitts, B. P. Jelley, J. Robbins, B. Klotz, S. C. Popescu, R. F. Nelson, D. Harding, D. Pederson, and R. Sheridan. ATLAS/ICESat-2 L3A Land and Vegetation Height, version 5, 2021. URL <http://nsidc.org/data/at108/versions/5>.
- T. A. Neumann, A. J. Martino, T. Markus, S. Bae, M. R. Bock, A. C. Brenner, K. M. Brunt, J. Cavanaugh, S. T. Fernandes, D. W. Hancock, K. Harbeck, J. Lee, N. T. Kurtz, P. J. Luers, S. B. Luthcke, L. Magruder, T. A. Pennington, L. Ramos-Izquierdo, T. Rebold, J. Skoog, and T. C. Thomas. The Ice, Cloud, and Land Elevation Satellite – 2 Mission: A Global Geolocated Photon Product Derived From the Advanced Topographic Laser Altimeter System. *Remote sensing of environment*, 233:10.1016/j.rse.2019.111325, Nov. 2019. ISSN 0034-4257. doi: 10.1016/j.rse.2019.111325. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6839705/>.
- OpenAltimetry. OpenAltimetry, Jan. 2023. URL <https://openaltimetry.org/>.
- OpenTopography. OpenTopography, Jan. 2023. URL <https://opentopography.org/>.
- J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, Dec. 2016. ISSN 1476-4687. doi: 10.1038/nature20584. URL <https://www.nature.com/articles/nature20584>. Number: 7633 Publisher: Nature Publishing Group.
- G. M. Philip and D. F. Watson. A PRECISE METHOD FOR DETERMINING CONTOURED SURFACES. *The APPEA Journal*, 22(1):205, 1982. ISSN 1326-4966. doi: 10.1071/AJ81016. URL <http://www.publish.csiro.au/?paper=AJ81016>.
- K. Rehman, N. Fareed, and H.-J. Chu. NASA ICESat-2: Space-Borne LiDAR for Geological Education and Field Mapping of Aeolian Sand Dune Environments. *Remote Sensing*, 15(11):2882, June 2023. ISSN 2072-4292. doi: 10.3390/rs15112882. URL <https://www.mdpi.com/2072-4292/15/11/2882>.
- S. Roy, V. Pasquarella, E. Trochim, and T. Swetnam. samapriya/awesome-gee-community-datasets: Community Catalog, Aug. 2023. URL <https://zenodo.org/record/8223455>.
- Scikit-Learn. Permutation Importance vs Random Forest Feature Importance (MDI), 2023. URL [https://scikit-learn/stable/auto\\_examples/inspection/plot\\_permutation\\_importance.html](https://scikit-learn/stable/auto_examples/inspection/plot_permutation_importance.html).
- A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat. Random Forest Spatial Interpolation. *Remote Sensing*, 12(10):1687, Jan. 2020. ISSN 2072-4292.

## Bibliography

- doi: 10.3390/rs12101687. URL <https://www.mdpi.com/2072-4292/12/10/1687>. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
- D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference on -*, pages 517–524, Not Known, 1968. ACM Press. doi: 10.1145/800186.810616. URL <http://portal.acm.org/citation.cfm?doid=800186.810616>.
- R. Sibson. A brief description of natural neighbour interpolation. *Interpreting Multivariate Data*, pages 21–36, 1981. URL <https://cir.nii.ac.jp/crid/1572543024689215360>. Publisher: John Wiley & Sons.
- B. Smith, H. A. Fricker, N. Holschuh, A. S. Gardner, S. Adusumilli, K. M. Brunt, B. Csatho, K. Harbeck, A. Huth, T. Neumann, J. Nilsson, and M. R. Siegfried. Land ice height-retrieval algorithm for NASA’s ICESat-2 photon-counting laser altimeter. *Remote Sensing of Environment*, 233:111352, Nov. 2019. ISSN 0034-4257. doi: 10.1016/j.rse.2019.111352. URL <https://www.sciencedirect.com/science/article/pii/S0034425719303712>.
- P. Teluguntla, P. Thenkabail, A. Oliphant, M. Gumma, I. Aneece, D. Foley, and R. McCormick. Landsat-Derived Global Rainfed and Irrigated-Cropland Product 30 m V001, 2023. URL <https://lpdaac.usgs.gov/products/lgrip30v001/>.
- J. Townshend. Global Forest Cover Change (GFCC) Tree Cover Multi-Year Global 30 m V003, 2016. URL <https://lpdaac.usgs.gov/products/gfcc30tcv003/>.
- E. Uuemaa, S. Ahi, B. Montibeller, M. Muru, and A. Kmoch. Vertical Accuracy of Freely Available Global Digital Elevation Models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM). *Remote Sensing*, 12(21):3482, Jan. 2020. ISSN 2072-4292. doi: 10.3390/rs12213482. URL <https://www.mdpi.com/2072-4292/12/21/3482>. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.
- T. Zhang and Y. Zhou. A global 1 km resolution daily near-surface air temperature dataset (2003 – 2020). 2022. doi: 10.25380/IASTATE.C.6005185.V1. URL [https://iastate.figshare.com/collections/A\\_global\\_1\\_km\\_resolution\\_daily\\_near-surface\\_air\\_temperature\\_dataset\\_2003\\_2020\\_/6005185/1](https://iastate.figshare.com/collections/A_global_1_km_resolution_daily_near-surface_air_temperature_dataset_2003_2020_/6005185/1). Publisher: Iowa State University.

## **Colophon**

This document was typeset using L<sup>A</sup>T<sub>E</sub>X, using the KOMA-Script class `scrbook`. The main font is Palatino.

