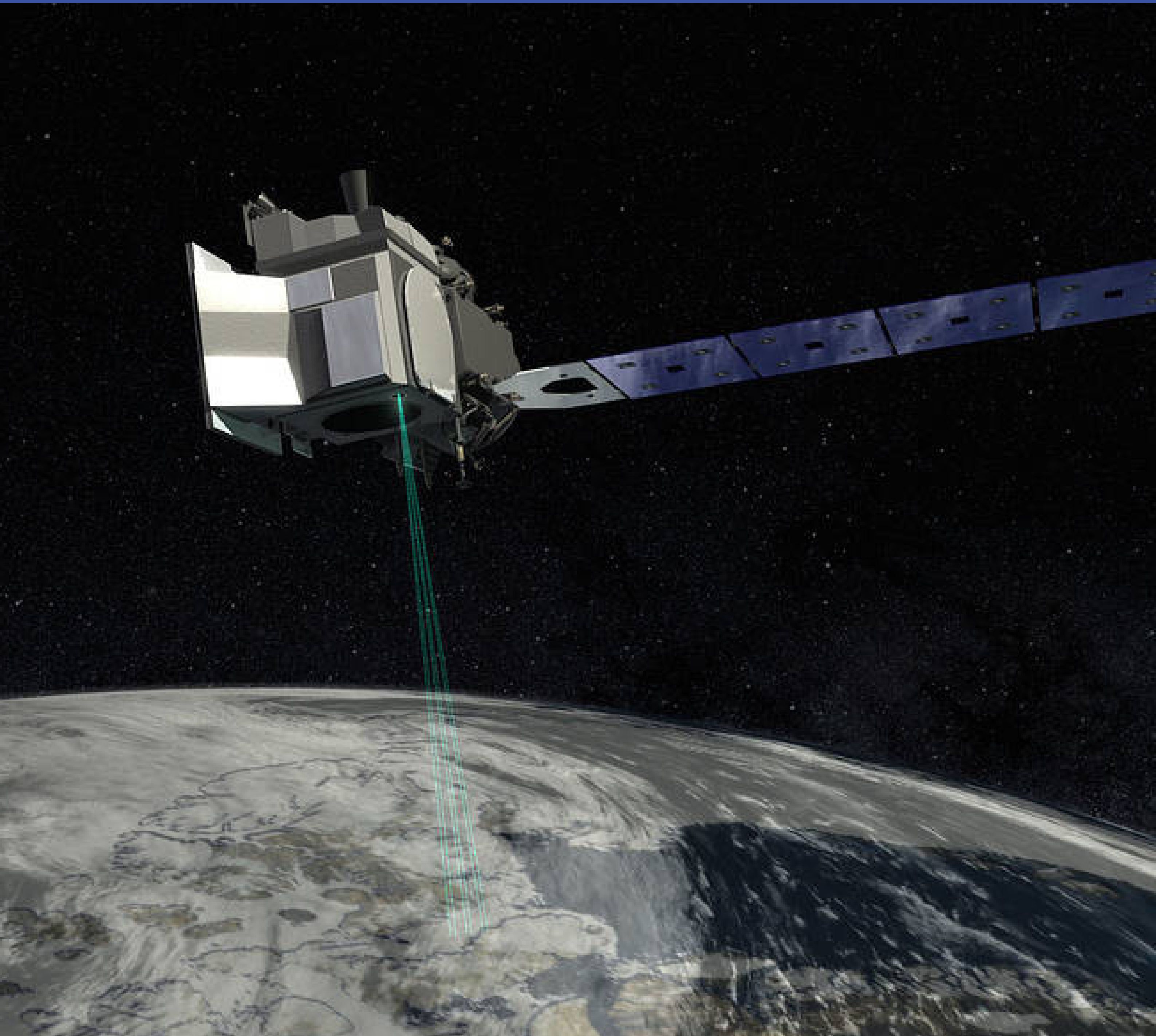
DTU Space
Department of Space

Satellite Derived Bathymetry from ICESat-2 using Machine Learning

Master Thesis



Satellite Derived Bathymetry from ICESat-2 using Machine Learning

Master Thesis
June, 2022

By
Freja Rose Borre

Supervisors
Ole Baltazar Andersen and Heidi Ranndal

Approval

Freja Rose Borre - s162599

....

Signature

....

Date

Abstract

There is increasing interest in getting bathymetry information, particularly for studying climate changes, assessing the marine environment, nautical navigation, and fishing, among others. The possibilities of determining bathymetry using machine learning are explored in this project.

An unsupervised density-based clustering model, called DBSCAN, has been applied to ICESat-2 data in the Heron reef area, part of the Great Barrier Reef. The model's performance has been compared with an empirical model based on a statistical interpolation approach. The performances of the two models are evaluated using a high-accuracy satellite-derived model from EOMAP.

Results show that both models perform well in determining bathymetry from ICESat-2 data. The DBSCAN model has challenges as it includes noise around the sea surface and leaves out bathymetry with low point density resulting in a slightly higher RMSE value than the empirical model. However, the empirical model results show a bias of 21 cm when compared with the EOMAP heights.

Based on the results from the DBSCAN model and the comparative study with the empirical model, it is evaluated that machine learning has great potential as a tool for determining bathymetry.

For future work, suggestions for improving the DBSCAN model are given. Furthermore, recommendations for other possible machine learning models are provided.

Acknowledgements

Heidi Ranndal, *Researcher*, DTU Space
Thank you, Heidi, for answering many questions, looking through scripts with me, helping me build a structured approach for the project, and giving constructive and detailed feedback on my results and interpretations.

Ole Baltazar Andersen, *Professor*, DTU Space
Thank you, Ole, for proposing this project, ideas for developments throughout the project, and giving constructive feedback on my results and the report.

Mads Ehrhorn, *PhD Student*, DTU Space
Thank you, Mads, for guidance related to Python and the machine learning aspect of the project. Thank you for your constructive and detailed feedback on the report.

Mikkel Lydholm Rasmussen, *Remote Sensing Specialist*, DHI Gras
Thank you, Mikkel, for the data and insight into using machine learning for determining bathymetry.

Jonathan Friis Schouenborg, *Msc. Student*, DTU Physics
Thank you, Jonathan, for discussing ideas concerning the thesis, inputs on contents, and for posing challenging questions concerning the topic and instrumentation related to the thesis.

Acronyms

ATLAS Advanced Topographic Laser Altimeter System.

DBSCAN Density-based spatial clustering of applications with noise.

EOMAP EOMAP GmbH & Co KG.

GNSS Global Navigation Satellite System.

ICESat-2 Ice, Cloud, and Land Elevation Satellite-2.

KDE Kernel Density Estimation.

kNN k Nearest Neighbor.

LiDAR Light Detection and Ranging.

LO Local Oscillator.

LSTM Long Short Term Memory.

MSL Mean Sea Level.

NASA National Aeronautics and Space Administration.

NSIDC National Snow and Ice Data Center.

RMS Root Mean Square.

RMSE Root Mean Squared Error.

SDB Satellite Derived Bathymetry.

SDG Sustainable Development Goal.

SNR Signal-to-Noise Ratio.

SVR Support Vector Regression.

Contents

Preface ii
Abstract iii
Acknowledgements iv

1 Introduction **2**
1.1 Motivation 2
1.2 Aims and Objectives 3
1.3 Structure of the Report 3

2 Theory **5**
2.1 Bathymetry 5
2.2 LiDAR 8
2.3 Machine Learning 10

3 Data **14**
3.1 Study Area 14
3.2 ICESat-2 15
3.3 EOMAP 17

4 Method **19**
4.1 Preprocessing 20
4.2 DBSCAN 21
4.3 Empirical Model 23
4.4 Performance Evaluation 24

5 Results **25**
5.1 DBSCAN 26
5.2 Empirical Model 32
5.3 Comparison 37

6 Discussion **40**
6.1 DBSCAN 40
6.2 Empirical Model 41
6.3 Comparative Analysis 42

7 Conclusion **44**

8 Future Work **46**

Bibliography **47**

A Data **49**

A.1 List of Data Files 49

B Results **50**

B.1 Figures 50

Chapter 1

Introduction

1.1 Motivation

Bathymetry information is one of the essential parameters which plays a significant role in planning near-shore structure activities such as engineering work, port management, fishing, and aquaculture, among others [1].

The topic is related to the Sustainable Development Goals (SDGs) 13 on climate action and 14 on life below water. Bathymetry can be used for climate change monitoring by studying sea level anomalies and disaster management, e.g., beach erosion and tsunamis. Furthermore, bathymetry can be used to monitor marine and coastal resources and aquaculture. Concerning SDG 14, life below water, an initiative called Seabed 2030 has been made to produce a complete, high-resolution bathymetric map of the world's seabed from the coasts to the deepest trenches by the year 2030 [2].

Bathymetry is traditionally determined locally by acoustic echo soundings or airborne LiDAR. Satellite data can be used instead to create bathymetric maps of a larger scale and reduce cost. Satellite Derived Bathymetry (SDB) is a technique that allows for global coverage, and is a cheap alternative to the local measurements, as the data can be accessed for free, e.g., via. from the European Space Agency or NASA. However, using satellite data comes with the challenge of obtaining high accuracy like the data acquired from techniques closer to the ground.

Several kinds of satellite data can be used to determine bathymetry; LiDAR data from ICESat-2 data is used in this project. The use of machine learning to determine bathymetry is a rising topic as machine learning is generally good for pattern recognition and prediction. However, machine learning often requires an extensive amount of data and computational power. In this report, a machine learning algorithm called DBSCAN is used, which has been specifically designed for large spatial datasets [3].

1.2 Aims and Objectives

The project's original objectives were to determine bathymetry in a particular area, compare the performance of two different machine learning models, and evaluate machine learning as a tool to acquire bathymetry by comparing it with an empirical model. However, the scope of the objectives turned out to be over-ambitious.

Instead, the objectives were slightly modified to the following,

- Determine bathymetry in a particular area.
- Create a machine learning model that can determine bathymetry
- Compare the machine learning model with an empirical model and evaluate machine learning as a tool for determining bathymetry.

As it appears from the new objectives, it was decided to focus on making a single machine learning model and perfect it rather than making two and comparing them.

The first objective is to determine bathymetry in a particular area. The bathymetry is determined using machine learning and an empirical model using an existing script, which Heidi Ranndal has provided. Initially, an area in the Bahamas was chosen for this project after conversing with DHI GRAS. However, it was decided to change the study area due to unavailable in situ data in the Bahamas area. The area was changed to the Heron reef, which is part of the great barrier reef in Australia due to its clear waters and the existence of a high accuracy Satellite Derived Bathymetry (SDB) map from EOMAP.

The second objective involves creating a machine learning model to determine bathymetry from ICESat-2 data. Originally, a model called Long Short Term Memory (LSTM) was chosen. However, due to unavailable in situ data in the area and a limited number of datasets, it was decided to change the model to the Density-based spatial clustering of applications with noise (DBSCAN) model. The DBSCAN model is an unsupervised model designed for large spatial datasets, where there is no need for pre-labeled ground truth data.

Finally, the third objective will conduct a comparative study between the machine learning and statistical interpolation models. It will be evaluated using the high accuracy SDB data from EOMAP to evaluate whether machine learning is optimal for determining bathymetry.

1.3 Structure of the Report

The report has a classical structure, in Chp. 2 relevant knowledge concerning bathymetry and refraction correction, LiDAR, which is the instrument used to collect the data from ICESat-2 is explained. Furthermore, there is a section explaining machine learning and, more generally, the DBSCAN algorithm, including how it works and determining parameters for the algorithm.

The data and the area of interest is presented in Chp. 3, along with details about the ICESat-2 mission, the ATLAS instrument, the particular dataset from ICESat-2 that is used, and the SDB EOMAP dataset.

In Chp. 4, the general data processing workflow is presented. The chapter includes the specific preprocessing steps used for both the ICESat-2 data and the EOMAP data to ensure an equivalent base for the comparison. Furthermore, the steps in applying the DBSCAN algorithm are explained in detail, and the reflections necessary to determine the parameters for the algorithm are described. The workflow of the empirical model is described according to [4], and the statistical parameters used to evaluate the models' performances are presented.

In Chp. 5, results from both the DBSCAN model and empirical model, which are necessary for the analysis, are commented on and presented.

In Chp. 6, the findings in Chp. 5 are discussed and combined with a priori knowledge. The performance evaluations of the models are compared with the performances of models in similar studies in the literature.

Finally, a conclusion is presented in Chp. 7 and suggestions for future work are given in Chp. 8.

Chapter 2

Theory

2.1 Bathymetry

Bathymetry is the study of underwater topography and is used in most human activity in the marine environment [1]. The bathymetric data can be used to investigate the environmental quality, sustainable fishery, aquaculture, infrastructure, boundaries definition, history, and composition of rocks and sediments.

The most common ways of determining bathymetry is using,

- Sonar
- LiDAR
- Radar

Local bathymetry maps are mostly made from sonar data acquired from the surface by ships or from the near bottom using remote vehicles. To map the bathymetry in larger areas, data can be obtained either from the air or space using LiDAR or radar data. In this thesis, the data being processed are acquired from a LiDAR; hence, this will be the primary focus.

2.1.1 Refraction Correction For Bathymetry

One of the essential things to account for when using LiDAR data to determine bathymetry is the refraction error that occurs due to a change in the speed of light at the air-water interface.

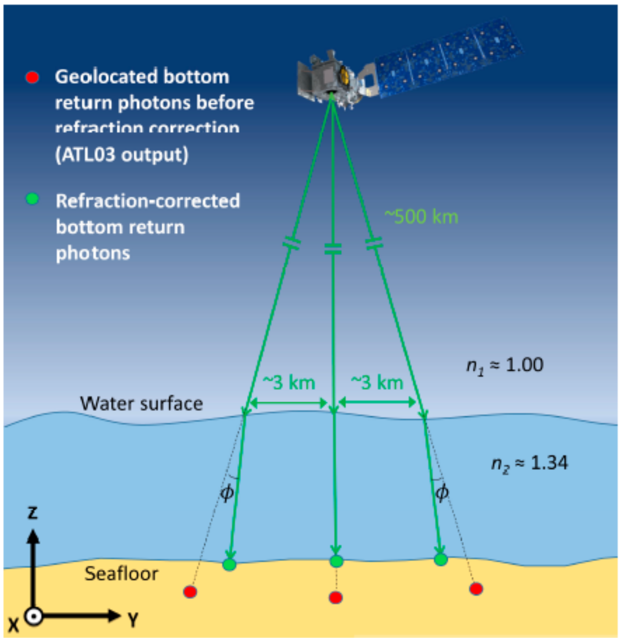

● Geolocated bottom return photons before refraction correction (ATL03 output)
● Refraction-corrected bottom return photons
~500 km
~3 km
~3 km
$n_1 \approx 1.00$
Water surface
$n_2 \approx 1.34$
ϕ
ϕ
Z
Seafloor
X
Y

Figure 2.1: Illustration of the need for refraction correction [5].

If not accounted for, errors in both horizontal and vertical directions will occur, resulting in locations that are deeper and further from nadir than the true measurement [5]. This effect is displayed in Figure 2.1.

The refraction correction can be considered a rotation and scaling as illustrated in Figure 2.2.

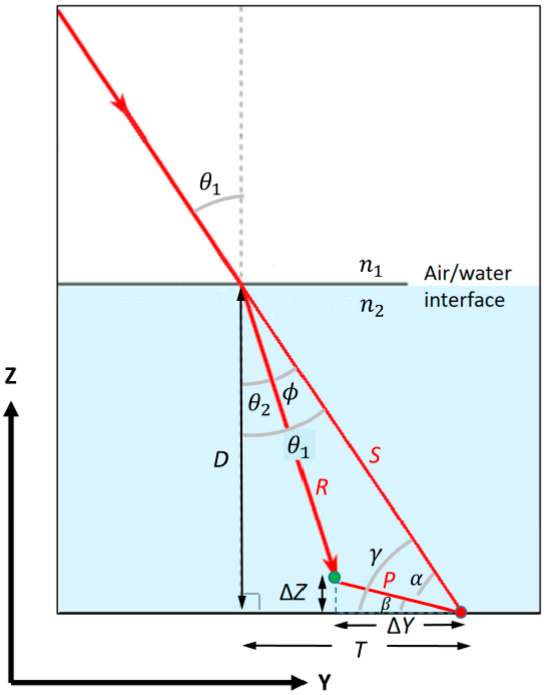


Figure 2.2: Geometry of refraction correction [5].

From Snell's law, the angle of refraction is given by,

$$\theta_2 = \sin^{-1}\left(\sin\theta_1 \frac{n1}{n2}\right) \tag{2.1}$$

where θ_1 is the angle of incidence and $n1, n2$ are the refractive indices for air and water, respectively.

Due to the change in speed of light, the corrected slant range can be found via. the relationship,

$$R = S\frac{n1}{n2} \tag{2.2}$$

where the slant range, S, can be found using,

$$S = \frac{D}{\cos\theta_1}, \tag{2.3}$$

where D is the uncorrected depth. Then, the angle, γ, is found,

$$\gamma = \frac{\pi}{2} - \theta_1 \tag{2.4}$$

Applying the law of sines to triangle RPS, cf. Figure 2.2,

$$\alpha = \sin\left(\frac{R\sin\phi}{P}\right) \tag{2.5}$$

where $\phi = \theta_1 - \theta_2$. Applying the law of cosines,

$$P = \sqrt{R^2 + S^2 - 2RS\cos\theta_1 - \theta_2} \tag{2.6}$$

Now the corrections in the Y and Z directions can be found by the following expressions,

$$\delta Y = P\cos\beta \tag{2.7}$$

$$\delta Z = P\sin\beta \tag{2.8}$$

where,

$$\beta = y - \alpha = \frac{\pi}{2} - \theta_1 - \sin^{-1}\left(\frac{R\sin\phi}{P}\right) \tag{2.9}$$

and finally, projecting δY onto (E, N) axes using azimuth, κ,

$$\delta E = \delta Y \sin \kappa \tag{2.10}$$

$$\delta N = \delta Y \cos \kappa \tag{2.11}$$

Hence, the new heights and geo-coordinates can be found by adding $\delta Z, \delta E$, and δN, respectively.

2.2 LiDAR

Light Detection and Ranging (LiDAR) or laser altimetry is a remote sensing method that measures range by measuring the time between transmitting a light signal and receiving it. Hence, the distance can be determined by,

$$\text{distance} = \frac{c\tau}{2} \tag{2.12}$$

where c is the speed of light, and τ is the measured time interval between the transmitted and received photons. The range is divided by 2 to obtain the one-way distance. From the distance, the elevation can be obtained by subtracting the distance from the satellite altitude,

$$\text{elevation} = \text{altitude} - \text{distance}. \tag{2.13}$$

Traditionally, two types of LiDAR's are used [6],

- Topographic
- Bathymetric

where the topographic LiDAR transmits light at near-infrared wavelength, and the bathymetric LiDAR transmits the photons at a wavelength corresponding to the green light such that it is possible to penetrate the water column. Fog has a wavelength at ≈ 1-100 μm in diameter, and rain drops ≈ 0.5-5 mm in diameter. Hence, a LiDAR signal will be highly scattered by clouds, fog, or rain [7].

There are three main components in a LiDAR,

- The laser source
- The receiver
- Optical system for pointing the LiDAR

Most LiDARs use either a diode laser or a diode-pumped solid-state laser. The diode lasers can be very efficient and inexpensive. However, they cannot store energy and tend to have a broad laser line width and a broad beam, whereas the solid-state laser has a narrower line width [7].

The LiDAR receivers can either be a single detector or and a array of detectors. To increase the Signal-to-Noise Ratio (SNR), two approaches can be used: Increase of intensity in the Local Oscillator (LO), or by use of gain [7].

For a LiDAR, a single aperture can be used for both the transmitter and receiver or by using separate apertures. Pointing the LiDAR can be done either mechanically, e.g., by a tilted mirror, or nonmechanically by tilting without moving a mechanical device [7].

The working principle of a LiDAR is that photons are transmitted from the laser source via multiple beams. Then, the photons are scattered, absorbed, or reflected when they hit a surface. In the case of a bathymetric LiDAR, the photons will be reflected at both the sea surface and the sea bottom as illustrated in Figure 2.3. The reflected photons are received in the detector, and by using the measured time interval between transmitting and receiving, the elevation can be obtained from Eq. 2.13. Combined with data from a Global Navigation Satellite System (GNSS), the elevation can be stored along with the corresponding latitude and longitude.

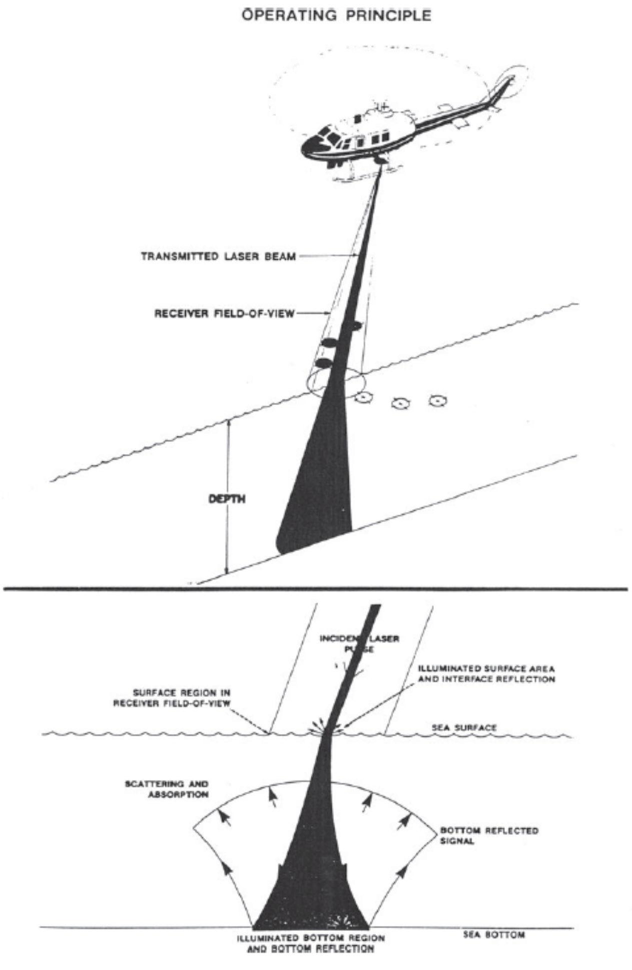

OPERATING PRINCIPLE
TRANSMITTED LASER BEAM
RECEIVER FIELD-OF-VIEW
DEPTH
INCIDENT LASER PULSE
ILLUMINATED SURFACE AREA AND INTERFACE REFLECTION
SURFACE REGION IN RECEIVER FIELD-OF-VIEW
SEA SURFACE
SCATTERING AND ABSORPTION
BOTTOM REFLECTED SIGNAL
ILLUMINATED BOTTOM REGION AND BOTTOM REFLECTION
SEA BOTTOM

Figure 2.3: LiDAR working principle [8].

2.3 Machine Learning

Machine learning is a way to train a computer to automate a process and is typically divided into three types,

- Supervised learning.
- Unsupervised learning.
- Reinforcement learning,

where supervised learning is applied for data sets that are composed of a data matrix and a set of target values, unsupervised is used for data sets that are only composed of the data matrix and where the goal is to infer structure to the data, and reinforcement learning is learning by trial and a reward system.

In this project, the focus is on unsupervised machine learning and, more specifically, clustering.

2.3.1 Clustering Analysis

Clustering analysis is a common way to analyze a dataset with a multipeak distribution of the observations, i.e. when there is more than one peak in the density distribution of at least one variable, cf. Figure 2.4 [9].

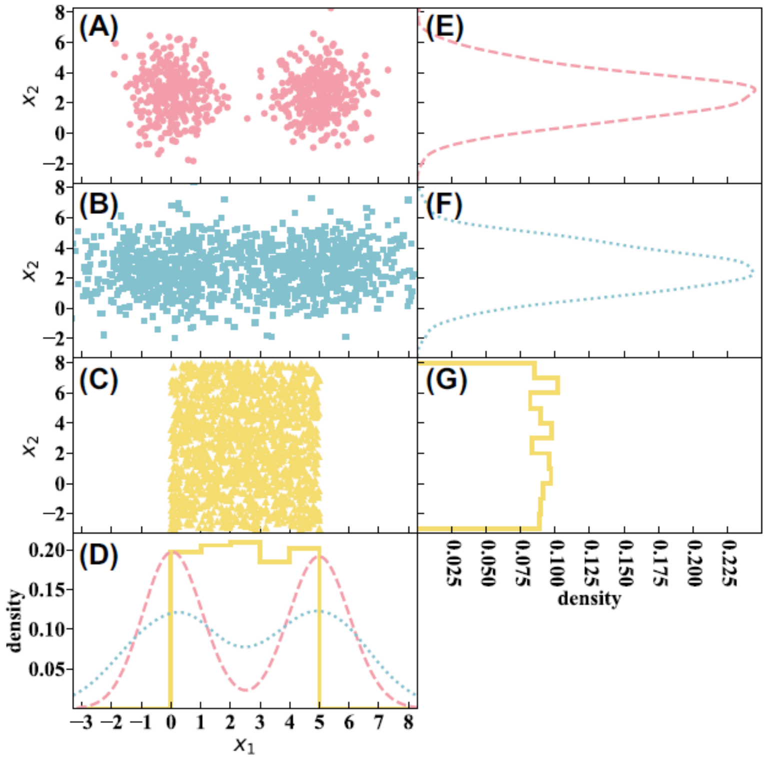


Figure 2.4: Examples of multipeak and unform distributions **(A)** multipeak distribution with clear separation; **(B)** multipeak distribution with less clear separation; **(C)** Uniform distribution; **(D)** Densities across X_1; **(E)**-**(G)** densities across X_2 [9].

Most cluster algorithms involve the following main tasks,

- feature selection,
- choice of a similarity metric,
- application of the grouping criterion,
- and cluster validation,

where the grouping criterion is vital as it defines how the observations are assigned to each cluster [9].

The different clustering algorithms can be classified based on whether it is; Partitional or hierarchical, hard or soft, centroid-based or density-based. Partitional or hierarchical refers to whether it can be divided into simple groups or groups and subgroups; hard or soft refers to whether it assigns each observation to a single class or receives a probability of belonging to each class. Finally, centroid-based or density-based refers to whether the algorithms assign each observation wrt. their distance to the center of the cluster or assigns based on the local density around the observation [9].

2.3.2 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm where observations are separated into high-density and low-density regions.

The algorithm is defined by two hyperparameters,

- Epsilon, ε
- MinPts,

where ε is the radius of the circle defining how close the observations must be from one another to be classified into the same cluster. MinPts is the minimum number of points that should be within ε of a point to be considered a core point [10].

In the DBSCAN algorithm, there are three main classifications of a point: Core point, border point, and noise point. A core point specifies a dense area based on the description mentioned above. The border point has fewer than MinPts with ε but is close to a core point. A noise point is neither a core point nor a border point and is considered an outlier [10]. The classification of the points is illustrated in Figure 2.5.

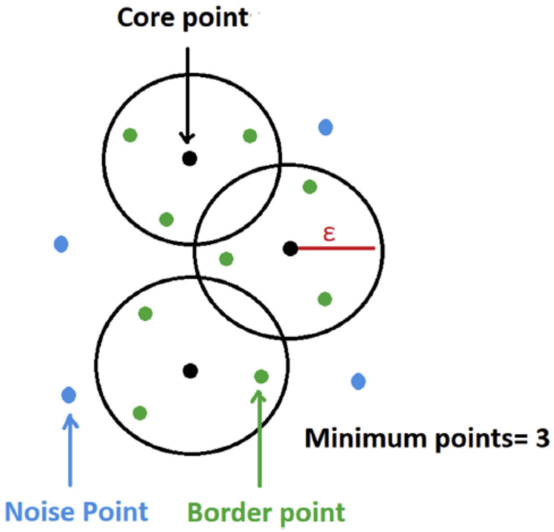


Figure 2.5: Illustration of core points, border points, and noise points [10].

The DBSCAN algorithm can be limited to the following steps [3],

```
DBSCAN Algorithm
1.  Start with arbitrary point, $p$
2.  Retrieve all points density-reachable from $p$ wrt. $\varepsilon$ and MinPts.
3.  if   p is a core point, this procedure yields a cluster wrt. $\varepsilon$ and MinPts
    if   $p$ is a border point, no points are density-reachable from p and DBSCAN
         visits the next point of the database.
```

A cluster, C, is defined wrt. ε and MinPts as a nonempty subset of a database, D, if it satisfy the following [3],

1. $\forall\ p, q$: if $p \in C$ and q is density-reachable from p wrt. ε and MinPts, then $q \in C$.

2. $\forall\ p, q \in C$: p is density-connected to q wrt. ε and MinPts,

where the terms density-reachable and density-connected are illustrated in Figure 2.6. The noise points are defined as not belonging to any cluster [3],

$$\text{noise} = \{p \in D \mid \forall\ i : p \notin C_i\}\,. \tag{2.14}$$

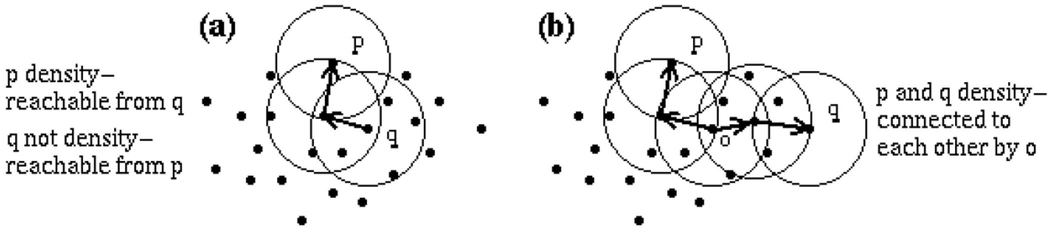

(a)
p density–reachable from q
q not density–reachable from p
p
q
(b)
p
q
o
p and q density–connected to each other by o

Figure 2.6: Density-reachability and density-connectivity [3].

One of the biggest challenges in applying DBSCAN is estimating the parameters, ε, and MinPts. Choosing the proper parameters requires a priori knowledge of the data. However, a general rule of thumb is to use a minimum number of points equivalent to $2 \times D$, where D is the dimension of the dataset.

After selecting the MinPts value, ε can be determined. One technique to automatically determine the optimal ε value is calculating the average distance between each point and its k nearest neighbors (kNN), where k = MinPts. The average k-distances are then plotted with the sorted distance along the x-axis and the kNN distance along the y-axis. All points with an equal or smaller distance will be the core points [3]. The optimal ε value will be at the ”elbow” of the curve, i.e., the point of maximum curvature, as displayed in Figure 2.7 as all points with a higher distance value are considered to be noise, and all other points are assigned to a cluster [3].

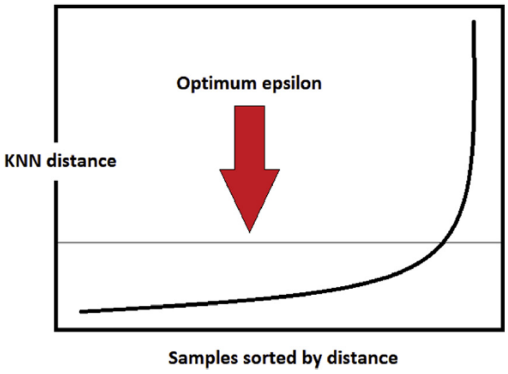


Figure 2.7: Points sorted by distance to the kNN [10].

Chapter 3

Data

This chapter presents the study area and the data used for the later analysis in this project. The data consists of raw data from the National Aeronautics and Space Administration (NASA) ICESat-2 mission and SDB data from EOMAP.

3.1 Study Area

The study area in this project is Heron Reef, which is part of the coral sea on the east coast of Australia. The reef is part of the great barrier reef and can be seen in Figure 3.1.

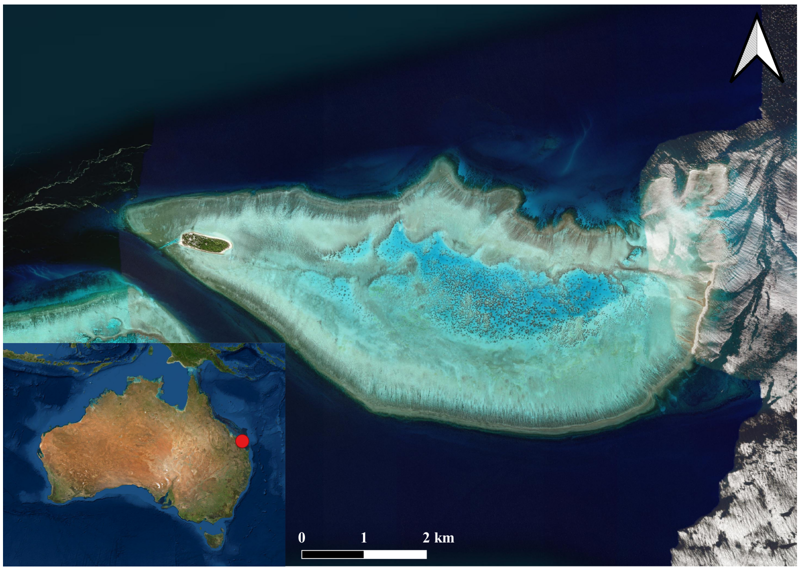


Figure 3.1: Heron reef, the red dot indicates the position of Heron reef in Australia.

This particular area has been chosen because it was possible to acquire the high-quality SDB data from EOMAP and because of its clear waters.

3.2 ICESat-2

The Ice, Cloud, and Land Elevation Satellite-2 (ICESat-2) mission is a NASA follow-up mission to ICESat. It was launched on 15. September 2018 to measure the elevation of ice sheets, glaciers, and sea ice in the cryosphere. The satellite also measures heights across Earth's temperate and tropical regions and thus, the Heron reef, which is the considered area in this project.

The orbital specifications of the ICESat-2 mission are stated in Table 3.1 below.

Table 3.1: Specifications of ICESat-2.

Orbital altitude	≈ 500 km
Inclination	92°
Repeat cycle	91 days

The primary payload on the ICESat-2 mission is the Advanced Topographic Laser Altimeter System (ATLAS).

3.2.1 ATLAS

ATLAS is an altimeter that carries a primary and secondary laser. The primary laser sends out six beams, cf. Figure 3.2 and the secondary is used as a backup. It transmits at a wavelength of 532 nm corresponding to green light with a rate of 10,000 pulses per second.

The ground tracks from the six different beams are typically about 14 m wide, and each ground track is numbered according to the laser spot number that generates it. The beams are in pairs; GT1L and GT1R, GT2L and GT2R, and GT3L and GT3R, with one being a strong beam and one being a weak beam with an energy ratio of approximately 1:4. The paired tracks are about 90 m apart in the across-track direction and 2.5 km in the along-track direction, as displayed in Figure 3.2. The distance between the beam pairs is approximately 3 km in the across-track direction [11]. ATLAS has an across-track resolution of 70 cm [4], meaning that the laser footprint moves at a 70 cm increment across the surface [12].

The technical specifications of ATLAS can be found in Table 3.2.

Table 3.2: Specifications of the ATLAS Transmitter.

Wavelength	532 nm
Pulse repetition frequency	10 kHz
Footprint diameter	< 17.4 m
Optical throughput efficiency	73%
Number of beams	6
Beam energy ratio (strong:weak)	4 : 1

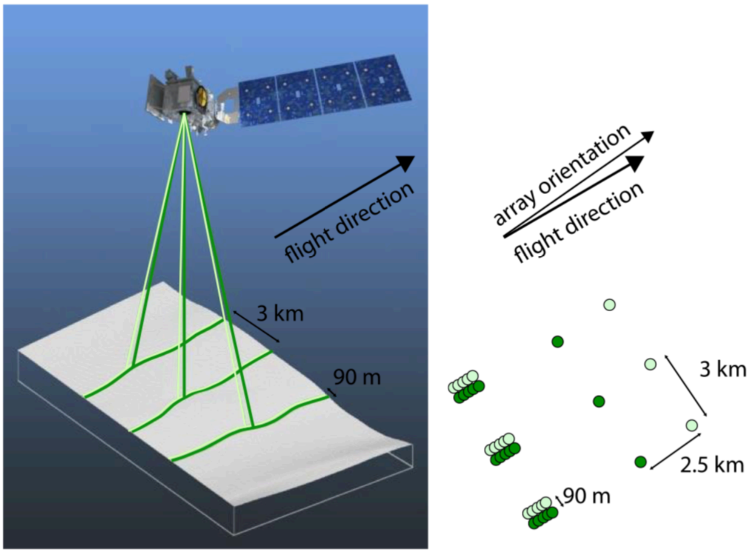


Figure 3.2: ATLAS acquisition principle [11].

There are three major tasks, that ATLAS is carrying out,

1. Send pulses of laser light to the ground,
2. collect the returning photons in a telescope,
3. record the photon travel time.

The returning photons are focused on six fiber optic cables corresponding with where the six laser beams return. From the fibers, the photons are passed through filters, letting only green light pass to prevent sunlight from swamping the detectors [13].

Then, from the recorded photon travel time, the photon height is found using Eq. 2.13.

3.2.2 Data

The dataset from the ICESat-2 satellite used for the project is the level 2 product, ATL03. The data provides the ellipsoidal height (above WGS84 ellipsoid), time, geodetic latitude, and longitude for all photons received in ATLAS.

Furthermore, the photons are classified based on the surface type from 0-4, where 0 corresponds to background photons, and 1-3 denotes a signal photon with low, medium, and high confidence, respectively. In this project, all the photons are included as the machine learning algorithm has the purpose of classifying them. The raw data is displayed in Figure 3.3.

The photon heights are already corrected according to solid earth tides, ocean, solid earth pole tides, ocean tidal loading, and range corrections for tropospheric delays [11]. Furthermore, the ATL03 file contains values for referencing, including the geoid height, which will be used in this project.

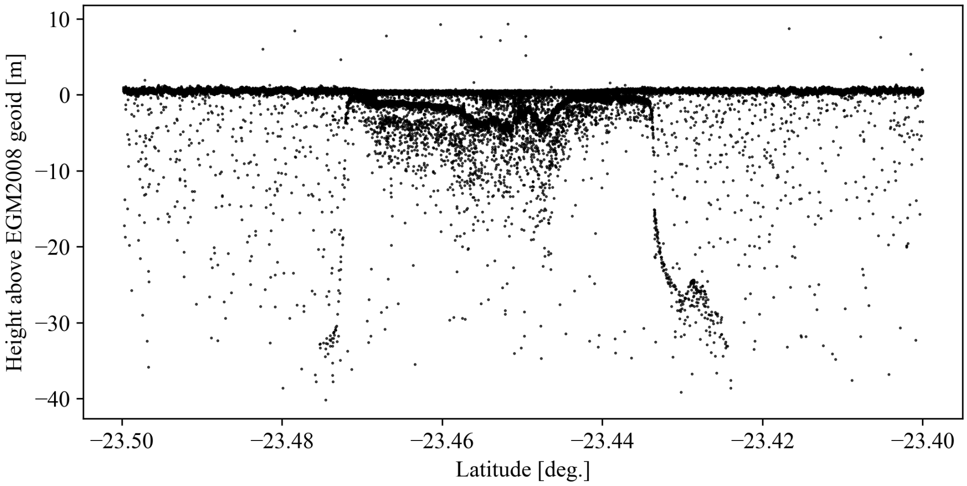


Figure 3.3: ICESat-2 photon height as a function of latitude, acquired on April 8, 2019, beam gt1l.

The data was acquired from The National Snow and Ice Data Center (NSIDC) website and were collected from the reference ground tracks 0154 and 1213. The temporal baseline ranges from March 2019 to November 2021. The acquired tracks are visualized in Figure 3.4.

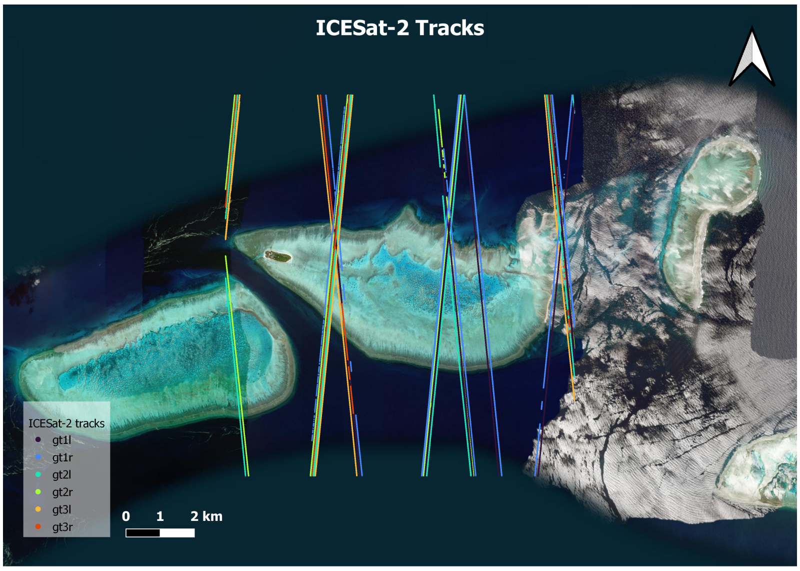


Figure 3.4: Map of tracks used in this project.

3.3 EOMAP

To evaluate the results from the machine learning model, SDB data from EOMAP is used. The SDB data is obtained from optical satellite images from the ESA Worldview-2 mission, launched on the 8. October 2009 to monitor the environment.

EOMAP applies a physics-based inversion method, known as Modular Inversion Program (MIP), which is independent of in-situ data and can be applied globally. Using the method and reflected sunlight energy in different wavelengths and various corrections, the bathymetry can be determined [14].

The product is displayed in Figure 3.5a, where the bathymetry is relative to the Mean Sea Level (MSL). The used MSL is 1.44 m, which is acquired from a local station [14]. The EOMAP product covers 42.6 km^2 down to a maximum depth of 25 m. The data has a spatial resolution of 2 m, and the horizontal accuracies are 5 m CE 90. The vertical accuracy is 50 cm absolute and 10% depth cf. Figure 3.5b.

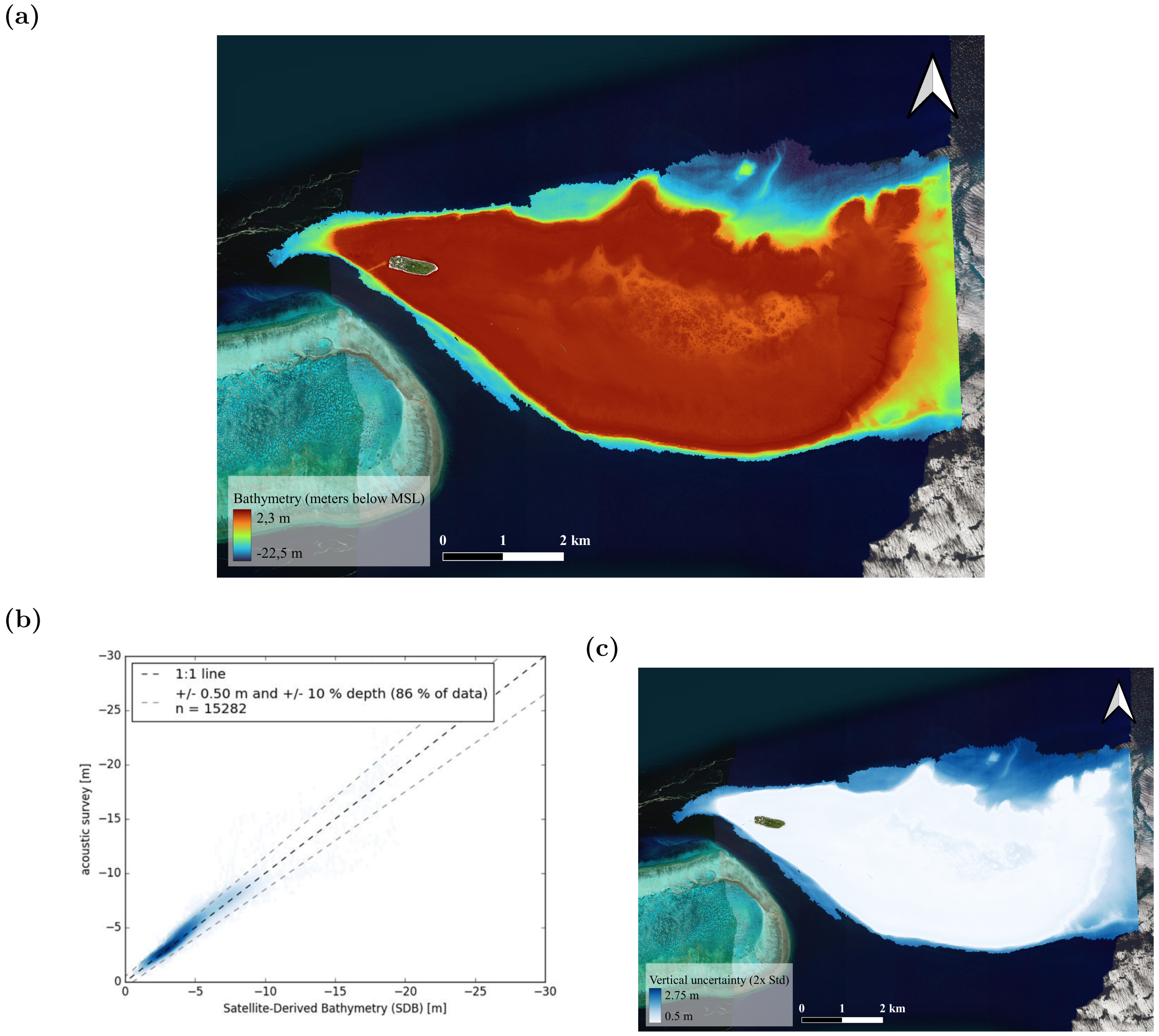

(a)
Bathymetry (meters below MSL)
2,3 m
-22,5 m
0
1
2 km
(b)
1:1 line
+/- 0.50 m and +/- 10 % depth (86 % of data)
n = 15282
acoustic survey [m]
−30
−25
−20
−15
−10
−5
0
Satellite-Derived Bathymetry (SDB) [m]
0
−5
−10
−15
−20
−25
−30
(c)
Vertical uncertainty (2x Std)
2.75 m
0.5 m
0
1
2 km

Figure 3.5: **(a)** EOMAP bathymetry model; **(b)** validation plot of the SDB data versus acoustic survey from single beam lines; **(c)** vertial uncertainties of the SDB data from EOMAP.

Chapter 4

Method

In this section, the data processing workflow is explained in detail, along with the necessary considerations for the preprocessing and the model parameters. The workflow is illustrated in the flowchart in Figure 4.1.

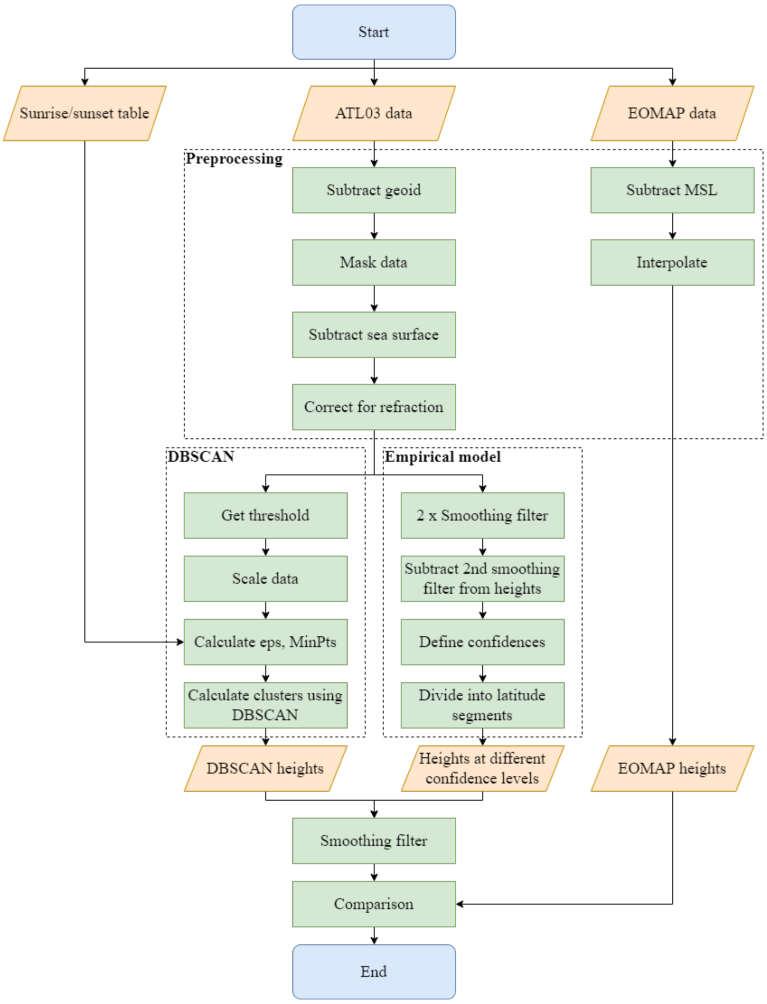


Figure 4.1: Flowchart of the data processing illustrating the start and end of the workflow (blue), the inputs/outputs (orange), and the processing step (green).

4.1 Preprocessing

4.1.1 ATL03 Data

The preprocessing steps of the ICESat-2 data are based on those in [4] and consist of the following,

- Remove geoid.
- Mask area and signal confidence.
- Remove sea surface.
- Correct for refraction.

First, the data is corrected for the geoid using the geoid data from the ATL03 file. The geoid heights are interpolated and subtracted from the photon heights.

Then, the data are masked by area and signal confidence as the file contains information for an entire track. The area masking is done by creating a bounding box made from specific latitude and longitude limits. Heron reef is placed approximately within the coordinates,

$$\text{latitude} : [-23.5, -23.4], \quad \text{longitude} : [151, 152],$$

which are used to limit the data in the ATL03 file. Then, the data signal confidence masking is done such that photons with confidence 0-4 are included, cf. the flags in Sec. 3.2.2.

After the masking, the sea surface is found using the median of the photon heights and a buffer of 0.5 m as in [4],

$$\text{sea surface} = \text{Med}_{(\text{height})} - 0.5\,\text{m}. \tag{4.1}$$

The sea surface is then removed from the heights. Finally, the observations are corrected for the refraction, cf. Sec. 2.1.1 assuming a salinity of 33 psu and a temperature of 29 °C.

4.1.2 EOMAP Data

The EOMAP data used to verify the DBSCAN model also needs to be corrected such that they have the same reference as the ICESat-2 data. It requires the following preprocessing steps,

- Correct for MSL.
- Interpolation.

As the EOMAP heights are relative to the MSL, the MSL needs to be subtracted from the heights. Hence, the heights will have the same reference, as the data has already accounted for the geoid. Furthermore, the data are interpolated so that the locations correspond to those in the ICESat-2 data.

The used interpolation model is constructed by triangulating the input data and making a piecewise cubic interpolating Bezier polynomial on each triangle, using a Clough-Tocher scheme [15][16]. The gradients of the interpolant are chosen so that the curvature of the interpolating surface is approximately minimized.

4.2 DBSCAN

In the application of the DBSCAN algorithm, the following steps are performed,

- Get threshold.
- Scale the data.
- Calculate ε and MinPts.
- Compute clusters using DBSCAN.

Before applying the DBSCAN algorithm, the preprocessed data is standardized to reduce computational time using,

$$z = \frac{x - \mu}{\sigma}, \tag{4.2}$$

where z is the standardized observation, x is the observation, μ is the mean, and σ is the standard deviation.

To calculate the clusters, the DBSCAN algorithm has been applied as mentioned in Chp. 2. The major keystone to the application of DBSCAN is to determine the parameters: ε and MinPts.

First, MinPts is determined based on a priori knowledge about the data. Given that the receiver of the ATLAS instrument is prone to pick up noise from the sun, the sun elevation is considered.

The sun elevations for that particular time for each ATL03 file are found using the National Oceanic and Atmospheric Administration (NOAA) solar calculator [17]. The inputs for the solar calculator are the latitude, longitude, time zone, date, and local time. The latitude and longitude limits used for masking the ATL03 file in Sec. 4.1 are given as input, and the date and local time are achieved from the ATL03 filename that is converted from UTC to the Australia/Brisbane time zone corresponding to UTC+10h. From the output, a table consisting of the date, time, apparent sunrise, apparent sunset, azimuth, and elevation is made. The solar calculator gives the azimuth and elevation output the value ”dark” when it is after the astronomical twilight. This value is treated as `NaN` throughout the data processing.

The parameter, MinPts, is determined by,

$$\text{MinPts} = \left\lfloor \frac{40}{\text{SNR}} \right\rfloor, \tag{4.3}$$

where the SNR denotes a signal-to-noise ratio estimated from a cumulative histogram of the preprocessed ATL03 heights, the number 40 is the expected number of points in a cluster, which is found by iteratively increasing the number from a beginning point of $2 \times D$, where D denotes the dimension of the data which is 3, as suggested in Sec. 2.3.2. Ultimately, 40 yielded the best result. The relation between MinPts and the SNR is presented in Figure 4.2.

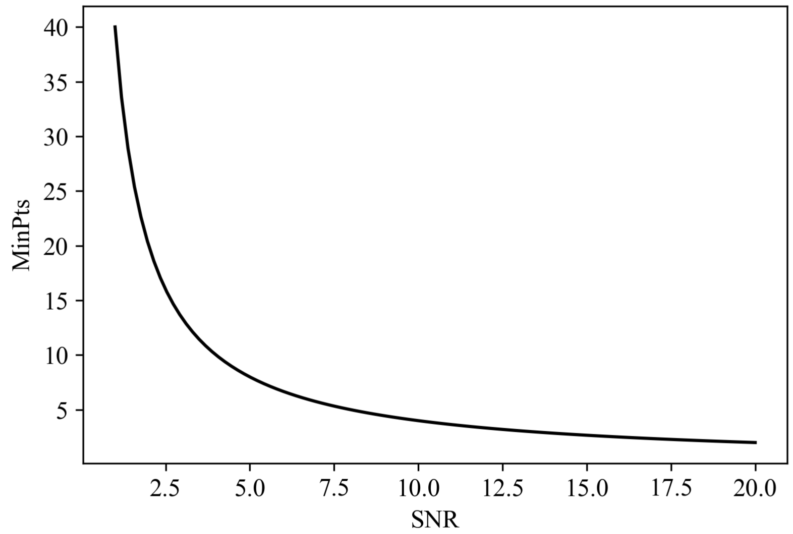


Figure 4.2: MinPts as a function of SNR.

SNR is found by the equation, inspired by [18],

$$\mathrm{SNR} = \frac{p_{\mathrm{signal}} + p_{\mathrm{noise}} \cdot 4\sigma}{p_{\mathrm{noise}} \cdot 4\sigma}, \tag{4.4}$$

where σ is the standard deviation of the heights, $p_{\mathrm{signal}} = 1 - p_{\mathrm{noise}}$ and,

$$p_{\mathrm{noise}} = \begin{cases} \rho_{\mathrm{noise}} + \rho_{\mathrm{noise}} \sin\theta_{\mathrm{sun}} & \text{if } \theta_{\mathrm{sun}} \neq \texttt{NaN} \\ \rho_{\mathrm{noise}} & \text{otherwise,} \end{cases} \tag{4.5}$$

where θ_{sun} is the sun elevation ρ_{noise} is an estimated noise density found from a threshold, which is defined by,

$$\mathrm{threshold} = \mathrm{Med} - 2\sigma, \tag{4.6}$$

where Med is the median of the heights. The threshold creates a clear distinction between estimated signal and noise such that all the photon heights below the threshold are considered noise.

A lower limit to MinPts has been implemented such that the minimum value for MinPts is $2 \times D = 6$.

The ε parameter is automatically determined from the technique mentioned in Sec. 2.3.2, using the kNN distance with k = MinPts. The optimal ε is found at the maximum curvature of the distance plot.

Several suggestions from the literature have been tested in the process of finding the final procedure for determining the parameters, including the generic approach, where MinPts = $2 \times D$ and ε are found from a kNN distance plot. However, it was found that a single value for MinPts was not optimal for this particular problem. Another approach from [19] has been tested, where two fixed values for ε were used for day and night, respectively, whereas ε is used to determine MinPts. This approach was also not found fit, as the MinPts values were generally too low; hence, the results contained a lot of noise.

Similar to the approach in [19], a previous article [20] uses a fixed value for ε, and determines MinPts afterward using a vertical segmentation. Once again, the output included too much noise. Finally, the method in [21] was tested. In this case, there are two fixed values for MinPts depending on whether the beam is weak or strong, and then the dataset is divided into ten vertical segments, where a unique ε parameter is found for each segment. Once again, the output was not satisfying.

As the noise depends significantly on the amount of green light from the sun, the idea of including the sun elevation occurred. The method for integrating the sun elevation into the noise component came from inspiration from [18], where it is included in the computation of SNR.

4.3 Empirical Model

The empirical model used for comparison is the one presented in [4]. To compute the results from the empirical model, a `MATLAB` code has been provided by Heidi Ranndal. The preprocessing of the data is the same for both DBSCAN and the empirical model, and after the preprocessing, the bathymetry is extracted via the following steps [4],

1. A moving median with a window of 50 observations is computed, and heights above 3m from the median are removed.

2. Another moving median with a window of 30 observations, H_{smooth}, is calculated along with a moving standard deviation of the difference between the observations and H_{smooth}.

3. The heights are divided into high, medium, and low confidence bathymetry.

4. The heights are divided into latitude segments, each with a length of 0.001 degrees $\approx 100\,\text{m}$. If there are less than 10 points in a segment, they are not considered bathymetry data.

The division of the data into high, medium, and low confidence bathymetry is done by excluding outliers using different thresholds. Hence, the indices for points within a distance, k_{diff} of the moving median are found using,

$$ix = |H_{\text{smooth}} - H| < k_{\text{diff}}, \tag{4.7}$$

and the data points within k_{diff} and with a standard deviation lower than k_{std} are kept,

$$H_{\text{keep}} = H[ix \ \& \ \text{mov}_{\text{std}} < k_{\text{std}}], \tag{4.8}$$

where the thresholds k_{diff} and k_{std} for the different bathymetry confidences are provided in Table 4.1.

Table 4.1: Thresholds for bathymetry confidence [4].

Threshold	**Low**	**Medium**	**High**
k_{diff}	0.75 m	1 m	2 m
k_{std}	1.5 m	2 m	4 m

4.4 Performance Evaluation

To evaluate the performance of the models, the resulting bathymetry is compared with the high resolution, high accuracy SDB data from EOMAP. The comparison is done by examining the residuals between the SDB data and the model bathymetry,

$$\text{Residual} = \text{Heights}_{\text{EOMAP}} - \text{Heights}_{\text{model}}, \tag{4.9}$$

and include statistical analysis. The statistical analysis includes a mean squared error (MSE), a coefficient of determination from linear regression made from the model bathymetry as a function of the SDB EOMAP data. The mean squared error and the coefficient of determination are computed using the following equations,

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2 \tag{4.10}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4.11}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$.

In addition to the linear regression and the RMSE, the distribution of the residuals is investigated by a histogram, and a statistics summary is also used to evaluate the performance.

Chapter 5

Results

This chapter presents and comments on the results of the data processing.

First, the distribution of all the ICESat-2 photon heights for Heron Reef is investigated. In Figure 5.1, the distribution of the preprocessed heights is illustrated.

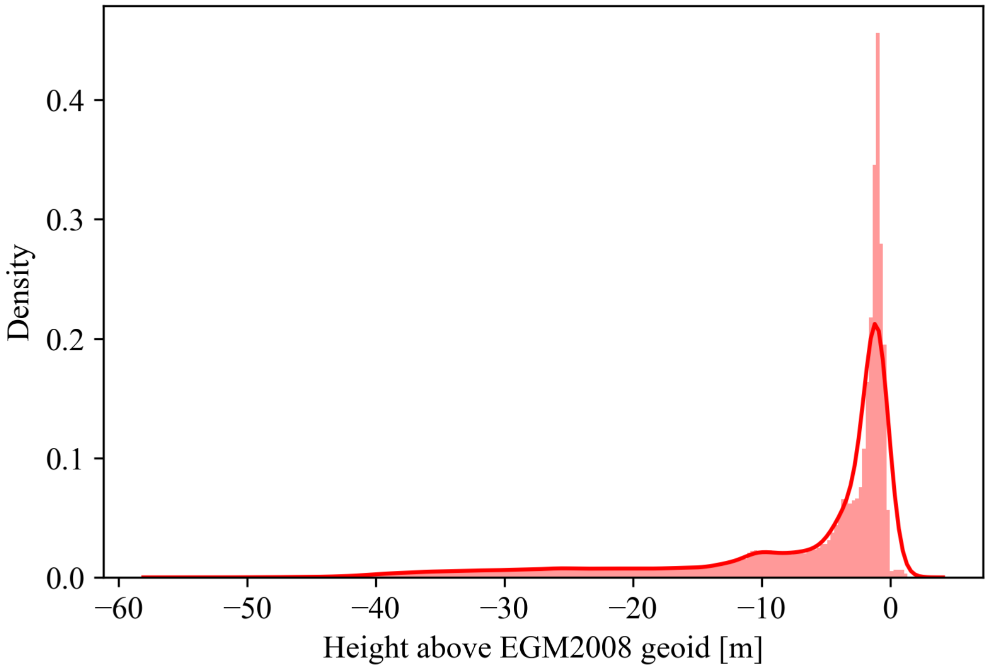


Figure 5.1: Histogram illustrating the distribution of all the preprocessed photon heights below the sea surface including a smoothed kernel density estimator (KDE) curve.

From Figure 5.1, the distribution appears to have more than one peak, at $\approx -0.5\,\mathrm{m}$ and $\approx -10\,\mathrm{m}$ and thus, meets the criteria of a multipeak distribution as described in Sec. 2.3.1.

5.1 DBSCAN

When using the DBSCAN algorithm, it is crucial to determine the parameters, MinPts and ε, cf. Sec. 2.3.2. The MinPts parameter is determined as described in Sec. 4.2 and used to calculate the kNN, where k = MinPts. The resulting distance curve is presented in Figure 5.2 along with the computed optimal ε value.

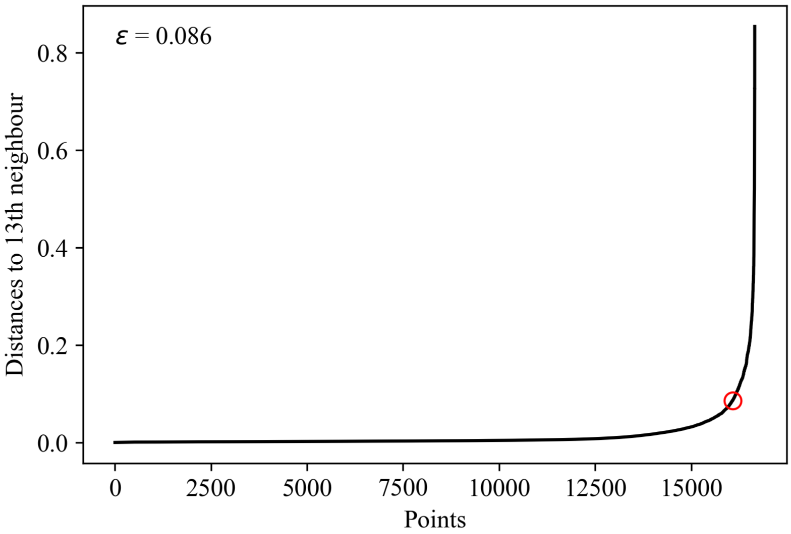


Figure 5.2: kNN distance plot for sample, where ICESat-2 passed on April 8, 2019, beam gt1l. The red circle indicates the optimal ε value.

From Figure 5.2, the optimal ε parameter is found as described in Sec. 4.2 at y value of the maximum curvature. For this particular file, MinPts = 13 and $\varepsilon = 0.086$.

After determining the parameters, the clusters can be found using the DBSCAN algorithm, which can be seen in Figure 5.3.

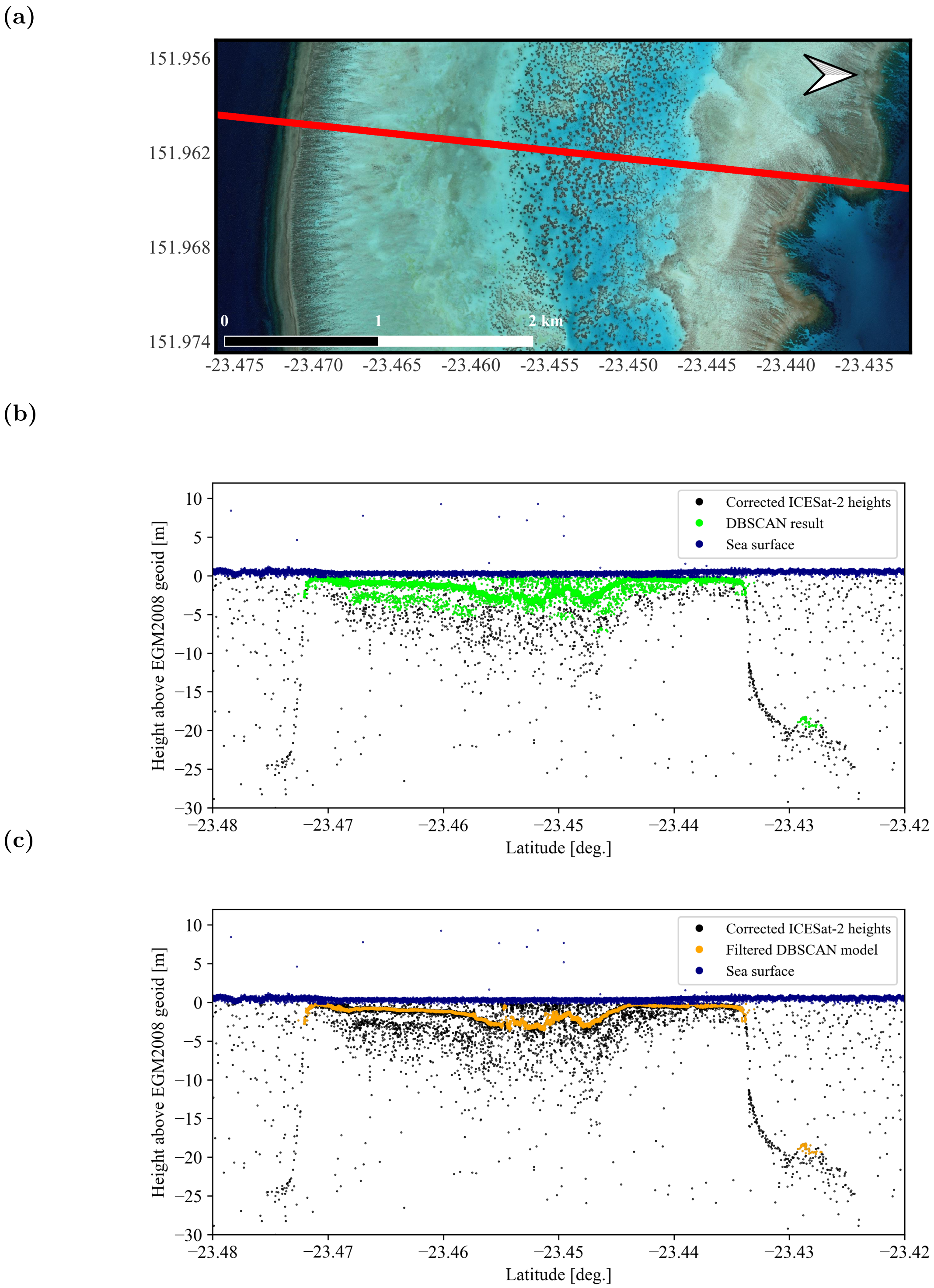


Figure 5.3: Sample, where ICESat-2 passed on April 8, 2019, beam gt1l **(a)** Map showing the ground track; **(b)** Corrected ICESat-2 photon heights with DBSCAN results and sea surface; **(c)** Corrected ICESat-2 photon heights with the filtered DBSCAN result and the sea surface.

From Figure 5.3, the results show that the model includes some noise near the sea surface. However, most noise disappears when applying the smoothing filter except for a few photons at $\approx 20\,\text{cm}$ height at ≈ -23.455 deg. latitude. Furthermore, it appears that there is some bathymetry signal that does not appear in the result from the model at latitudes $\approx [-23.48, -23.47]$ deg. and $\approx [-23.44, -23.42]$ deg.

The performance of the model is evaluated by creating a bias plot; the EOMAP heights as a function of the DBSCAN heights, whereas a linear regression is fitted to the data and visualized along with the statistical parameters, RMSE and a coefficient of determination, R^2. The bias plots can be seen in Figure 5.4.

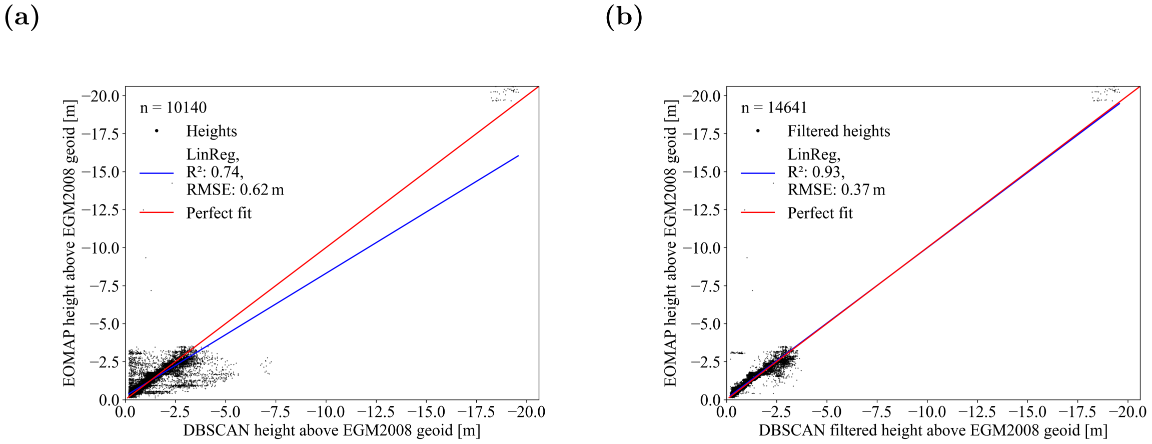


Figure 5.4: Bias plot of sample, where ICESat-2 passed on April 8, 2019, beam gt1l **(a)** EOMAP as function of DBSCAN; **(b)** EOMAP as a function of filtered DBSCAN.

From Figure 5.4, it is evident that there is a significant improvement in the application of the smoothing filter with a difference of $\delta R^2 = R^2_{\text{unfiltered}} - R^2_{\text{filtered}} = -0.19$ and $\delta\text{RMSE} = \text{RMSE}_{\text{unfiltered}} - \text{RMSE}_{\text{filtered}} = 0.25\,\text{m}$.

5.1.1 Overall Results

The results are computed based on all 12 files, which are listed in Table A.1 in App. A.1 in the following subsection.

The model's overall performance is evaluated in the bias plot in the following Figure 5.5.

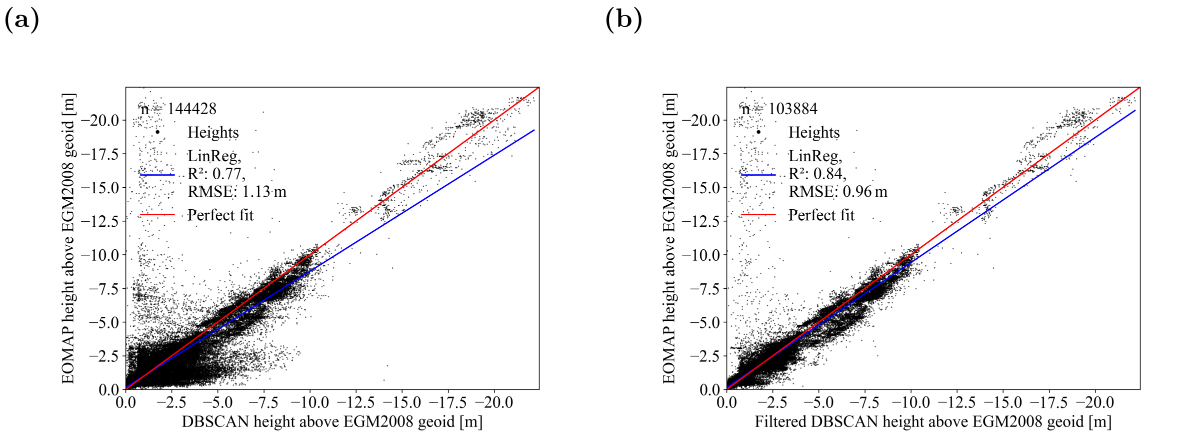


Figure 5.5: Bias plot of the **(a)** unfiltered heights from the DBSCAN model; **(b)** filtered heights from the DBSCAN model.

From Figure 5.5, the overall performance appear to be improved significantly with the application of the smoothing filter with a difference of $\delta R^2 = R^2_{\text{unfiltered}} - R^2_{\text{filtered}} = -0.07$ and $\delta\text{RMSE} = \text{RMSE}_{\text{unfiltered}} - \text{RMSE}_{\text{filtered}} = 0.17\,\text{m}$.

The residuals between the EOMAP heights and the DBSCAN heights and filtered DBSCAN heights respectively are displayed as a function of photon height in the following Figure 5.6.

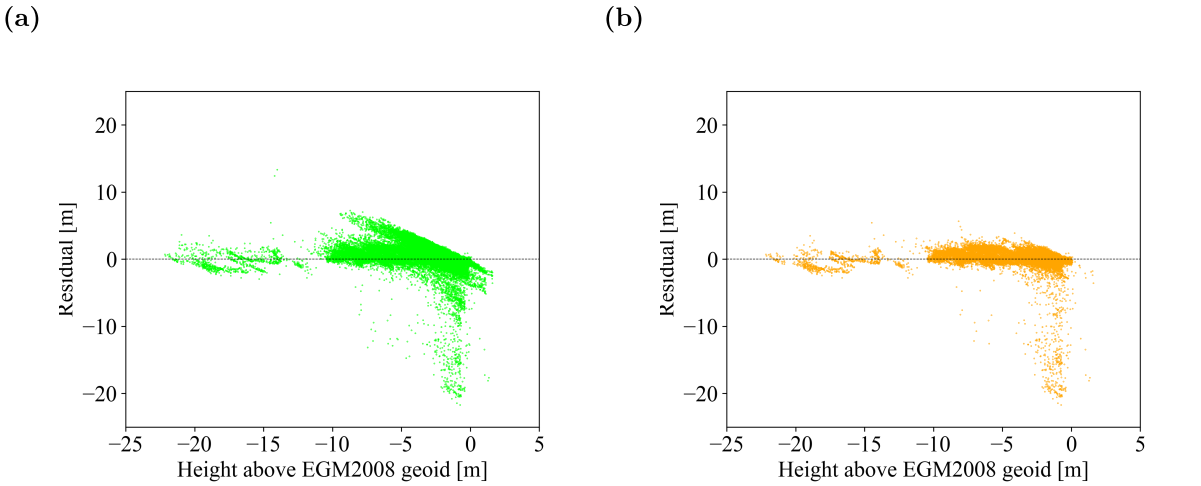


Figure 5.6: Residuals as a function of heights for **(a)** DBSCAN; **(b)** Filtered DBSCAN; Dashed line indicating residual = 0.

From Figure 5.6, a small bias appear on heights $[-10, 0]$ m. However, this bias disappears for the filtered DBSCAN residuals. Furthermore, there are several negative residuals near the sea surface.

The distribution of the residuals is displayed in a histogram in Figure 5.7.

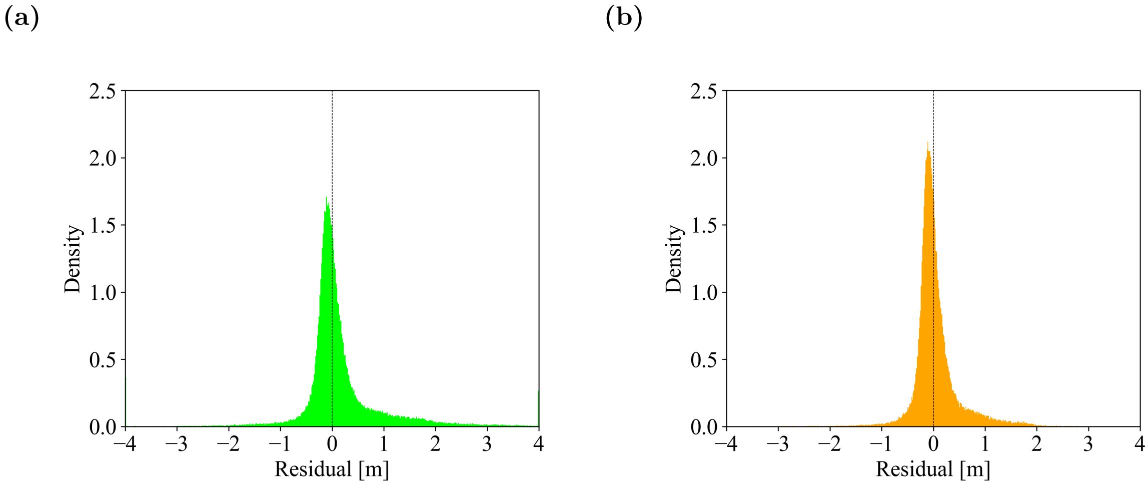


Figure 5.7: Histogram of **(a)** DBSCAN residuals; **(b)** filtered DBSCAN residuals. The histogram is narrowed to show the residuals between [-4,4] m, where the outliers are gathered in the outermost bins.

From Figure 5.7, the residuals appear to be centered around 0 as desired. However, the distribution appears slightly right-skewed for both the DBSCAN residuals and the filtered DBSCAN residuals indicating a small bias.

A statistics summary of the distribution of the residuals is displayed in Table 5.1.

Table 5.1: Statistics summary of residuals between EOMAP and DBSCAN and filtered DBSCAN respectively.

Residual	**DBSCAN**	**Filtered DBSCAN**
Count	144428	103884
Mean [m]	−0.11	−0.001
Std [m]	1.12	0.96
Min [m]	−21.70	−21.70
Max [m]	13.33	5.67
25% [m]	−0.17	−0.16
50% [m]	−0.01	−0.04
75% [m]	0.24	0.15

Table 5.1 reveal that there are outliers down to −21.70 m and up to 13.33 m. However, these are insignificant when considering a standard deviation of 1.12 m and 0.96 m, respectively. Furthermore, the median of −0.01 m and −0.04 m, indicate a small bias of 1 cm and 4 cm. The bias appears to increase after the smoothing filter is applied.

A map visualizing the resulting filtered photon heights from the DBSCAN model can be seen in Figure 5.8a along with the residuals between those and the EOMAP heights.

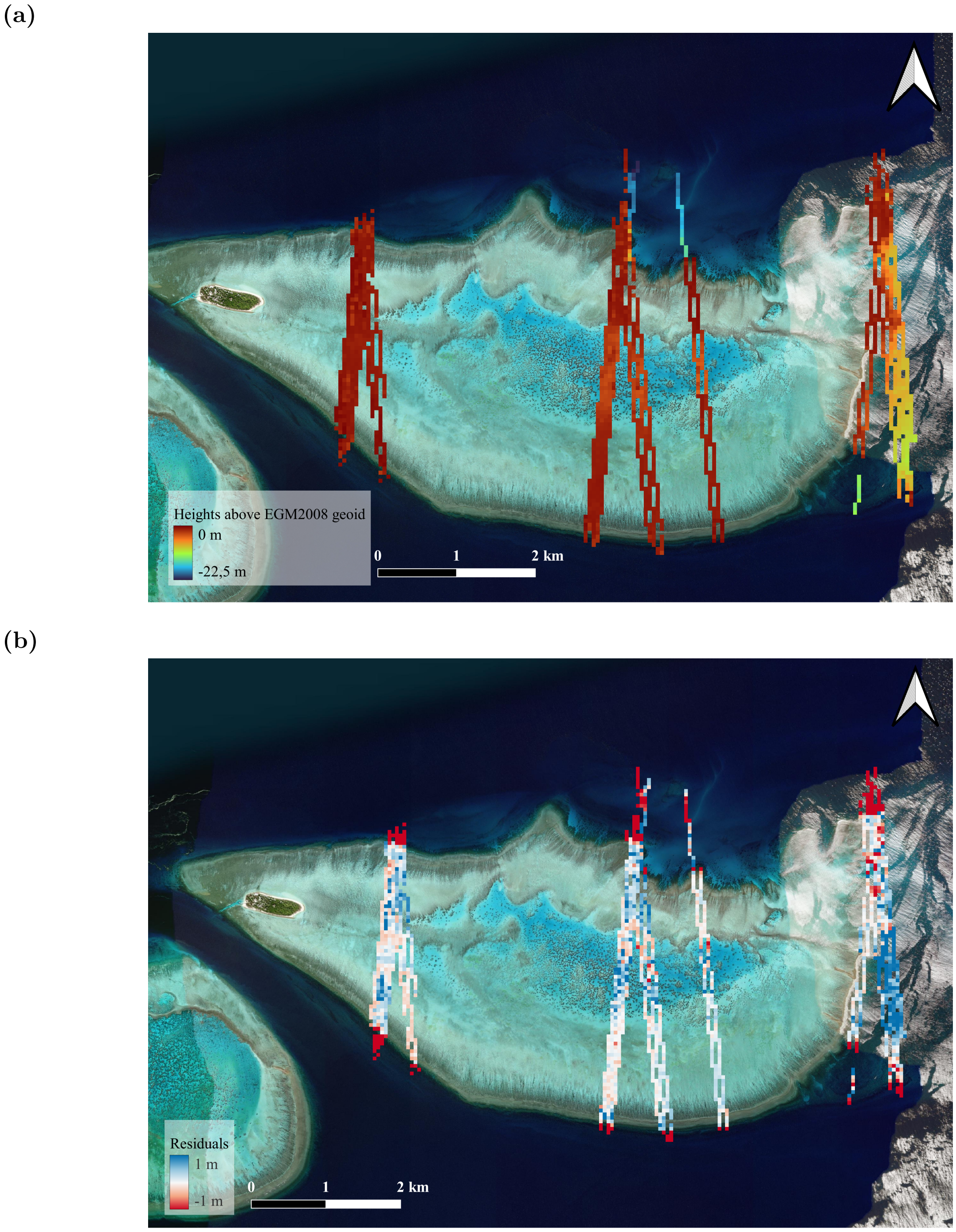


Figure 5.8: Map of **(a)** the resulting filtered heights from the DBSCAN model; **(b)** the residuals between the resulting filtered heights from the DBSCAN model and the EOMAP heights limited to residuals between [-1,1] m. The pixelsize in the maps has been exaggerated to make the heights visible.

From Figure 5.8a, the model appears to have caught some of the low heights at the top of the reef but is missing several in the high slope areas, see Figure B.1 in App. B.1. The minimum height in the overall results is -22.16 m, and the maximum height is 1.61 m. Considering Figure 5.8b, the biggest residuals appear in high slope area. It should be noted that the residual map is limited to be between [-1,1] m to reveal a detailed overview, and thus, there are residuals ≤ -1 m and ≥ 1 m, cf. Table 5.1.

5.2 Empirical Model

In this section, the results from the empirical model are presented.

In Figure 5.9, the computed high confidence bathymetry is displayed along with the filtered version.

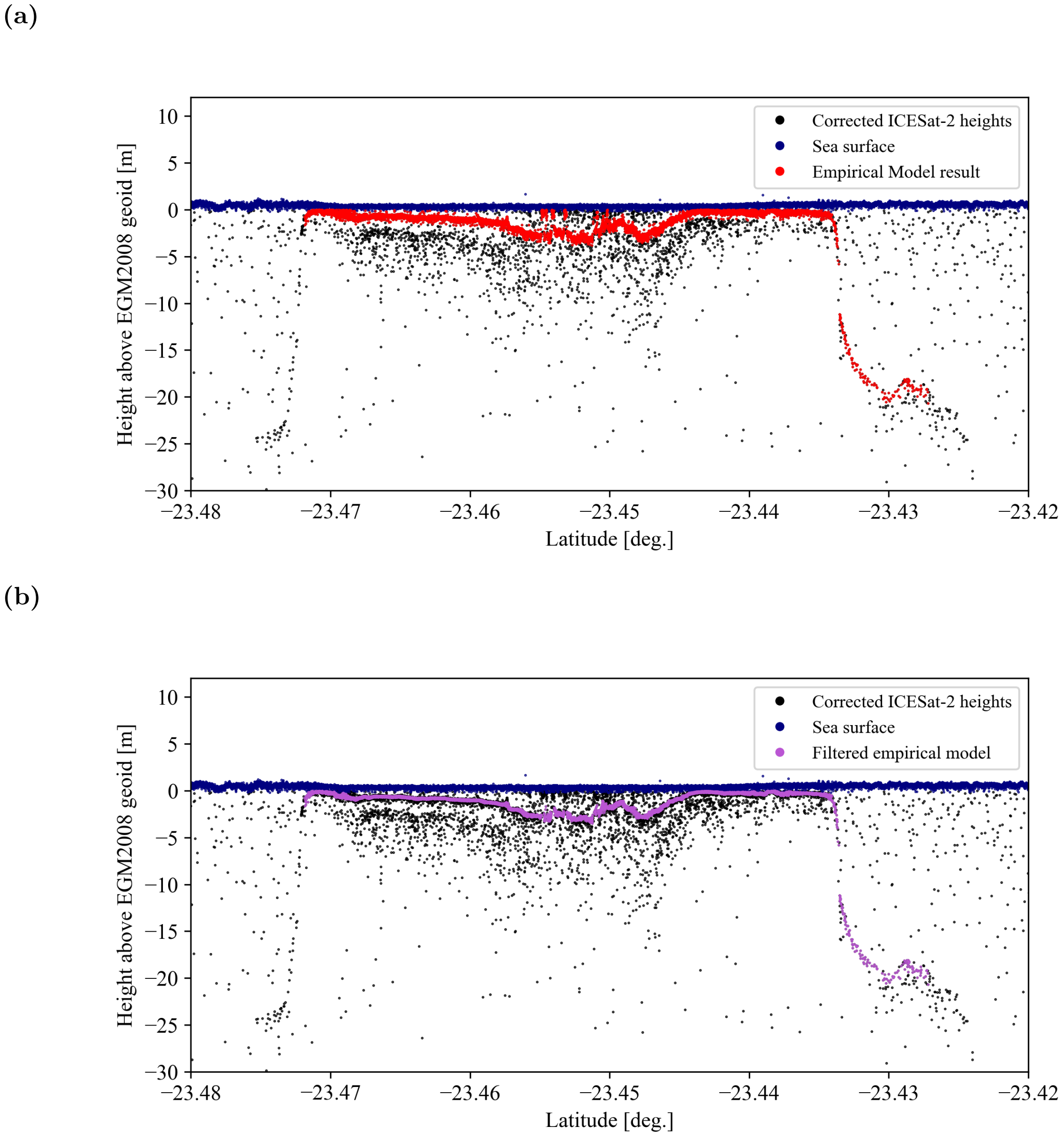


Figure 5.9: Sample, where ICESat-2 passed on April 8, 2019, beam gt11 **(a)** Corrected ICESat-2 photon heights with empirical model results and sea surface; **(b)** Corrected ICESat-2 photon heights with the filtered empirical model result and the sea surface.

From Figure 5.9, the results show that model includes some noise near the sea surface in height [-0.5,0] m between latitude [-23.46,-23.44] deg. However, the noise disappears when applying the smoothing filter. Furthermore, it appears that some bathymetry signal at latitude $\approx [-23.48, 23.47]$ deg is not included.

The performance is evaluated in a bias plot with linear regression, RMSE, and coefficient of determination, R^2, similar to that in Sec. 5.1. The bias plot is displayed in Figure 5.10.

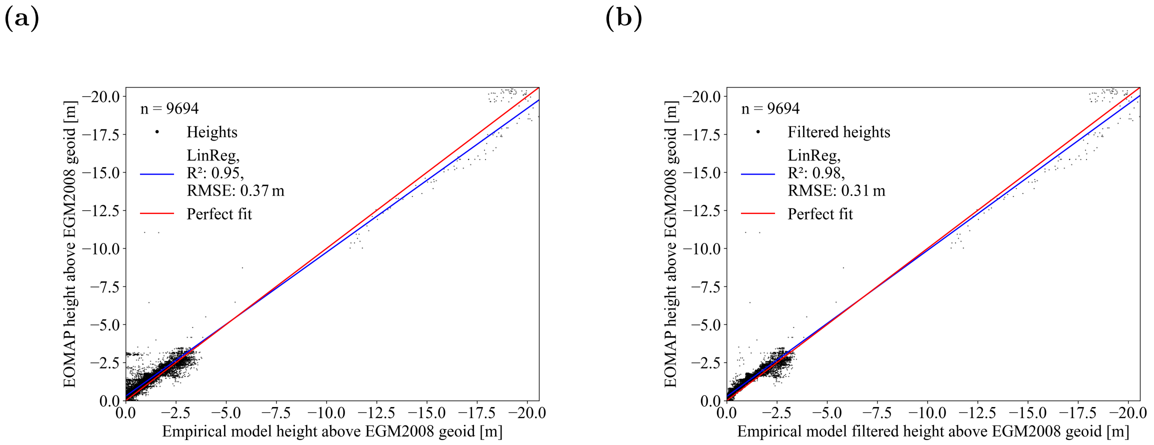


Figure 5.10: Bias plot of **(a)** EOMAP as function of empirical model heights; **(b)** EOMAP as a function of filtered empirical model heights.

From Figure 5.10, the results seem to have improved slightly with difference of $\delta R^2 = R^2_{\text{unfiltered}} - R^2_{\text{filtered}} = -0.03$ and $\delta\text{RMSE} = \text{RMSE}_{\text{unfiltered}} - \text{RMSE}_{\text{filtered}} = 0.06\,\text{m}$.

5.2.1 Overall Results

In this subsection, the results based on all 12 files are reviewed.

The model's overall performance is evaluated in the bias plot in the following Figure 5.11.

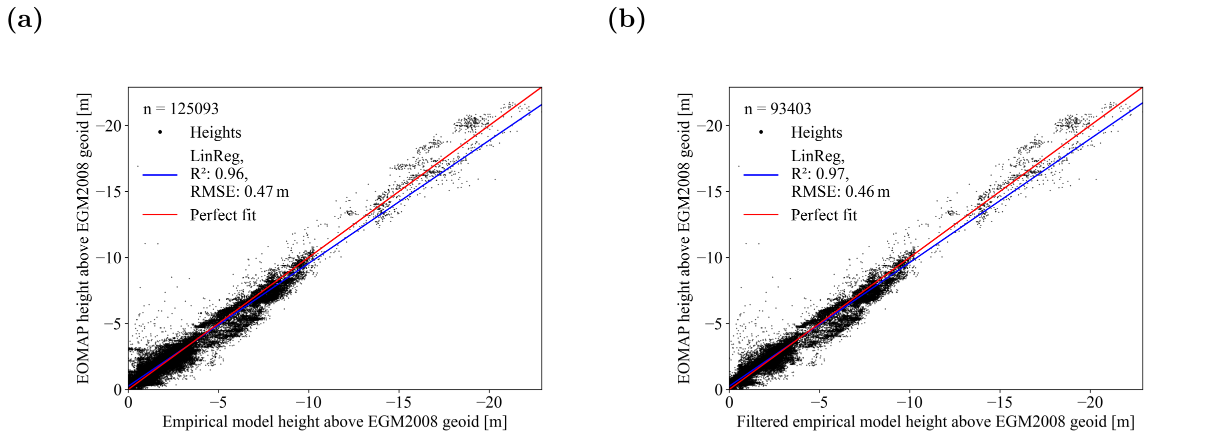


Figure 5.11: Bias plot of the **(a)** unfiltered heights from the empirical model; **(b)** filtered heights from the empirical model.

From Figure 5.11, the performance of the model seem to have been improved slightly with a difference of $\delta R^2 = R^2_{\text{unfiltered}} - R^2_{\text{filtered}} = -0.01$ and $\delta\text{RMSE} = \text{RMSE}_{\text{unfiltered}} - \text{RMSE}_{\text{filtered}} = 0.01\,\text{m}$.

The residuals between the EOMAP heights and the empirical model heights and filtered empirical model heights, respectively, are displayed as a function of photon height in Figure 5.12.

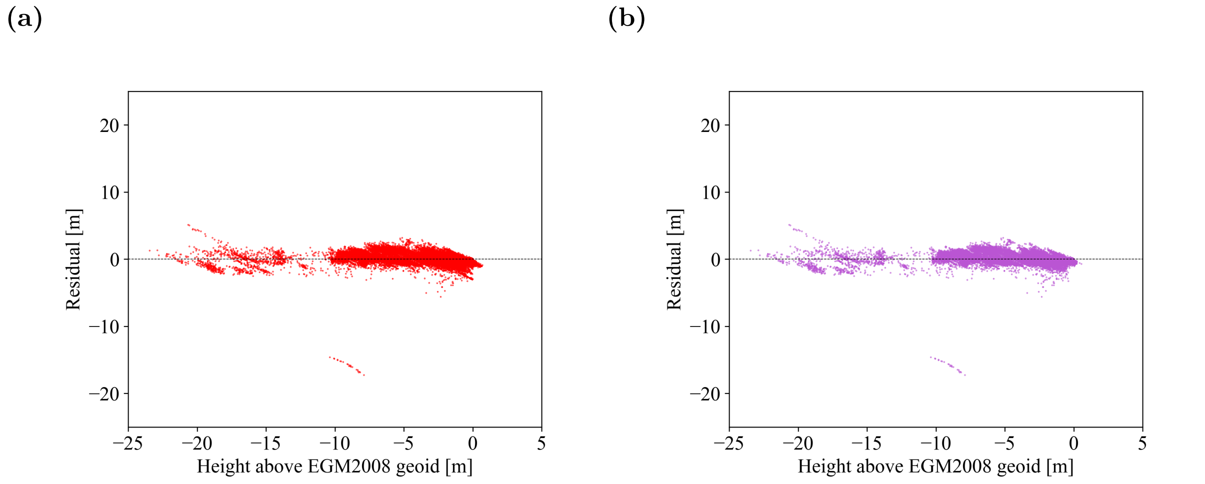


Figure 5.12: residuals as a function of height of **(a)** Empirical model; **(b)** Filtered empirical model.

From Figure 5.12, the biggest residuals appear to be at heights [-20,-15] m. The smoothing filter does not appear to impact the residuals significantly.

The distribution of the residuals is displayed in a histogram in Figure 5.13.

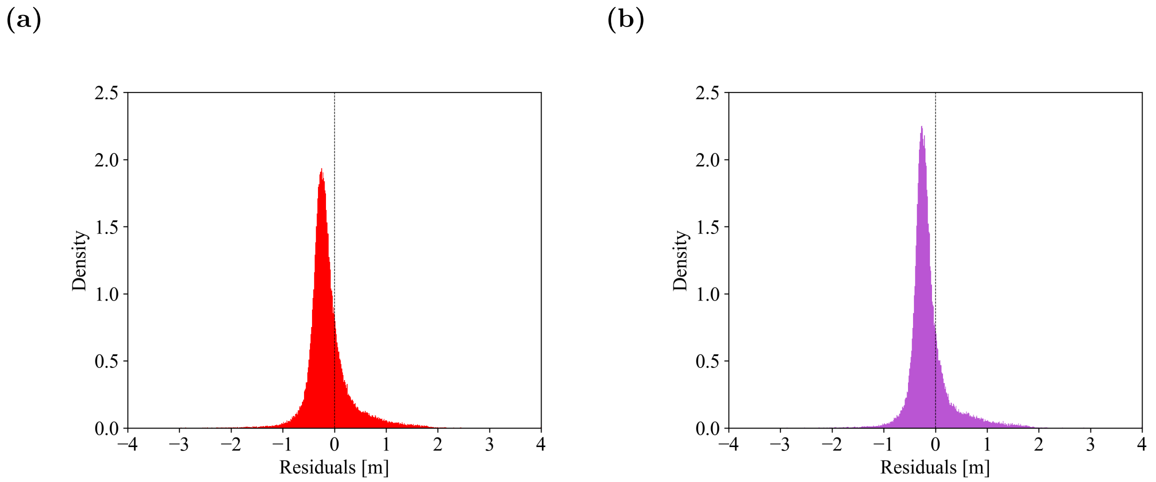


Figure 5.13: Histogram of residuals from **(a)** Empirical model; **(b)** Filtered empirical model. The histogram is narrowed to show the residuals between [-4,4] m, where the outliers are gathered in the outermost bins.

From Figure 5.13, there is a clear bias towards the negative residuals meaning that the empirical model results are closer to the sea surface than the heights from EOMAP. From the histogram, it is easier to see the improvement of the results with the smoothing filter, as the peak of the histogram is higher than that for the unfiltered results.

A statistics summary of the distribution of the residuals is displayed in Table 5.2.

Table 5.2: Statistics summary of residuals between EOMAP and the empirical model and the filtered empirical model respectively.

Residual	**Empirical Model**	**Filtered Empirical Model**
Count	125347	93657
Mean [m]	-0.13	-0.13
Std [m]	0.48	0.48
Min [m]	-17.23	-17.23
Max [m]	5.13	5.13
25% [m]	-0.33	-0.33
50% [m]	-0.20	-0.21
75% [m]	-0.01	-0.04

From Table 5.2, outliers down to -17.23 m and up to 5.13 m appear. Furthermore, the medians are -0.20 m, and -0.21 m, indicating that the resulting heights are generally 20 cm and 21 cm above those from the EOMAP results, which supports the fact there is a bias. The smoothing filter does not appear to impact the results from the empirical model greatly.

A map visualizing the resulting filtered photon heights from the empirical model can be seen in Figure 5.14a along with the residuals between those and the EOMAP heights.

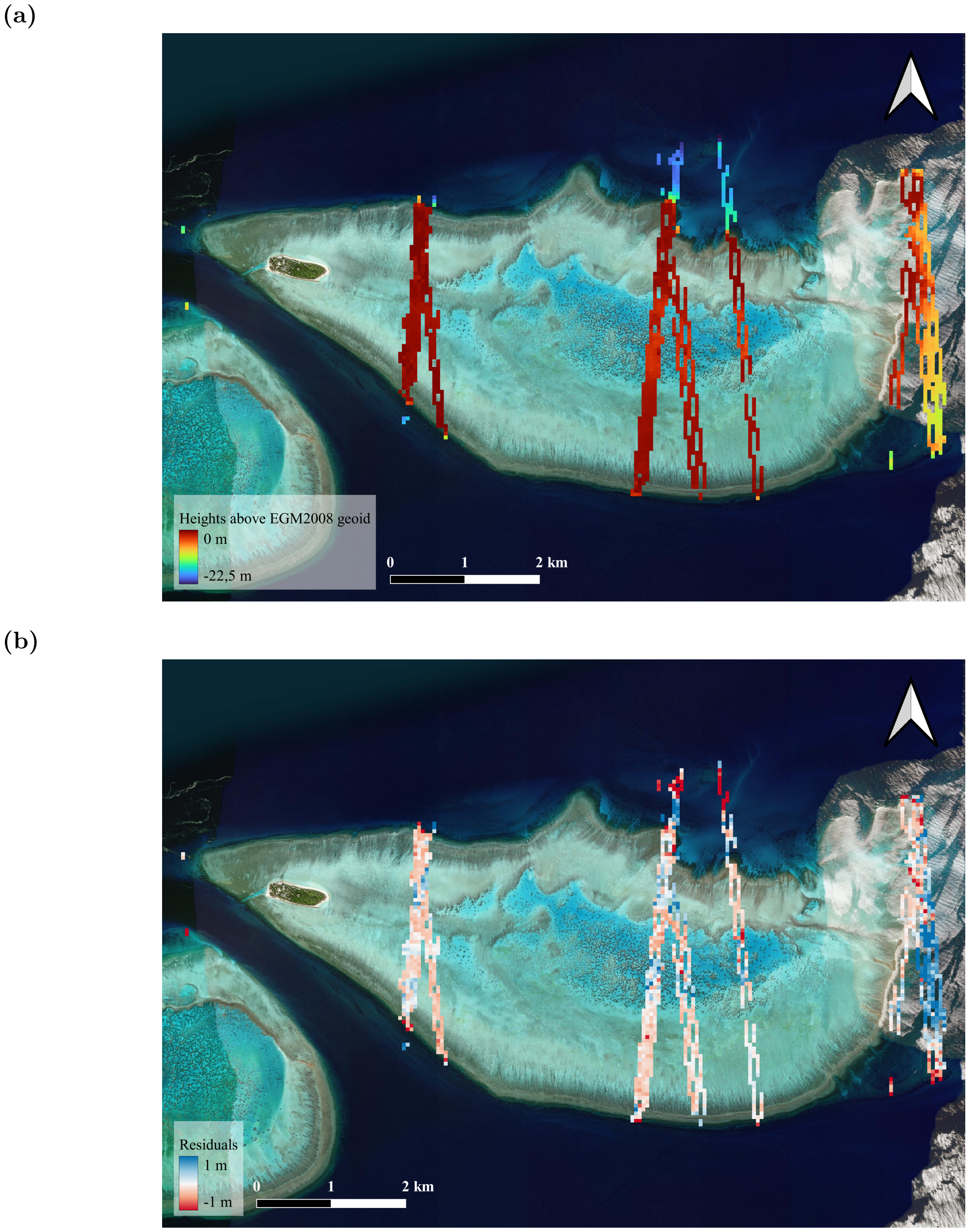


Figure 5.14: Map of **(a)** The resulting filtered heights from the empirical model; **(b)** The residuals between the resulting filtered heights from the empirical model and the EOMAP heights limited to residuals between [-1,1] m. The pixelsize in the maps has been exaggerated to make the heights visible.

From Figure 5.14a, the empirical model appears to have included several of the lower height areas. However, the empirical model seems to have difficulty catching the high slope areas, cf. Figure B.1 in App. B.1. The minimum height in the overall results is −23.45 m, and the maximum height is 0.55 m.

In Figure 5.14b, the biggest residuals appear to be located in the deep areas. In the areas closer to the surface, there appears to be an overload of red areas indicating that the results from the empirical model estimate the bathymetry heights to be closer to the sea surface compared to the EOMAP heights. The residual map is limited to show the residuals between [-1,1] m, there are residuals ≤ -1 m and ≥ 1 m, cf. Table 5.2.

5.3 Comparison

This section compares the filtered results from the DBSCAN model and the empirical model directly with the EOMAP heights.

The heights from the EOMAP data and the filtered heights from the DBSCAN model and the empirical model are displayed in the following Figure 5.15.

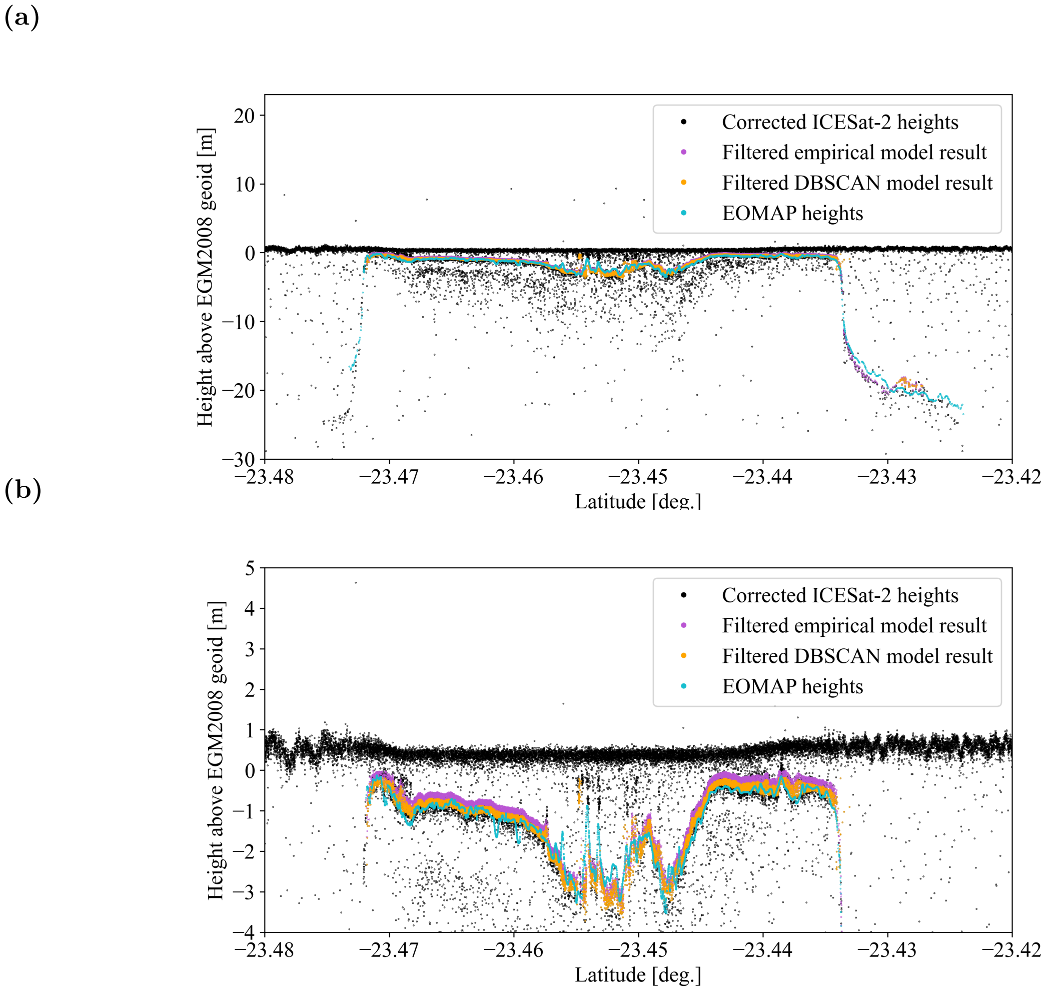


Figure 5.15: Sample, where ICESat-2 passed on April 8, 2019, beam gt1l with heights from the EOMAP data, and the filtered heights from the DBSCAN model and the empirical model **(a)** Full size; **(b)** zoomed in on heights $[-4, 0]$ m.

In Figure 5.15, there are peaks in the EOMAP heights that do not appear for either the filtered heights from the DBSCAN model or the filtered heights from the empirical model. Furthermore, the resulting heights from the DBSCAN model also include some noise at latitude ≈ -23.455 deg. and the resulting heights from the empirical model appear to be slightly above the other heights. The lower EOMAP heights seem to be significantly different from those from the DBSCAN model and the empirical model at latitude ≈ $[-23.435, -23.425]$ deg.

The residuals between the EOMAP heights and the filtered heights from the DBSCAN model and the empirical model, respectively are displayed in Figure 5.16.

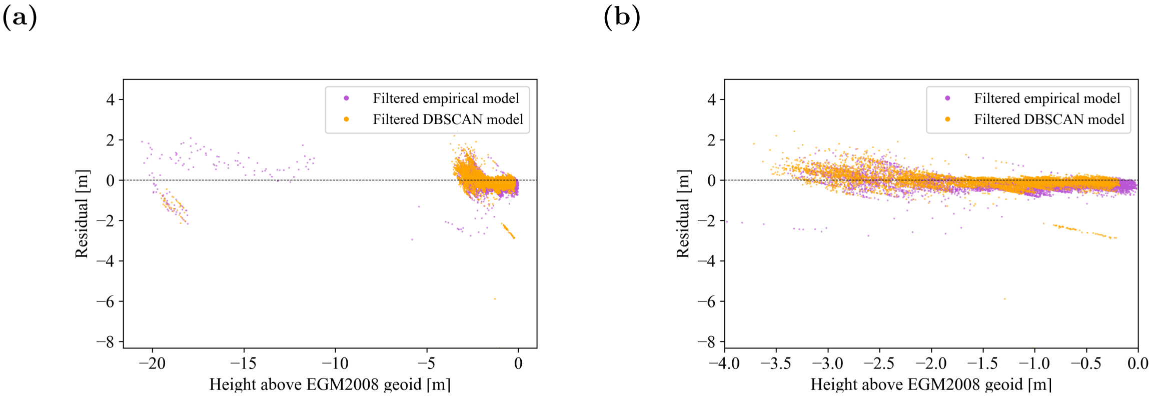


Figure 5.16: Residual as a function of height **(a)** Full size; **(b)** zoomed in on heights $[-4, 0]$ m.

The distribution of the residuals is displayed in Figure 5.17.

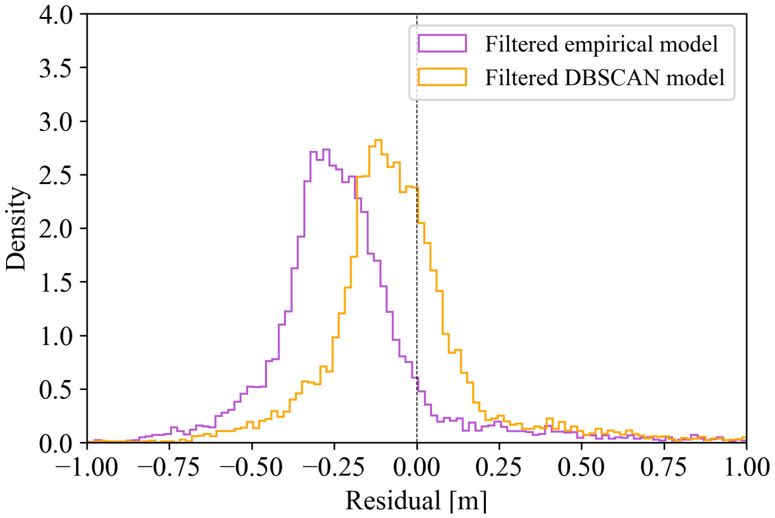


Figure 5.17: Histogram of residuals between EOMAP and the empirical model, and between EOMAP and the filtered DBSCAN for a sample, where ICESat-2 passed on April 8, 2019, beam gt1l.

From Figure 5.17, a bias appears for the residuals for both the DBSCAN model and the empirical model. However, the peak for each of the distributions seems to be at approximately the same density.

A statistics summary of the residuals is displayed in Table 5.3.

Table 5.3: Statistics summary of residuals between EOMAP data and the filtered empirical model results and the filtered DBSCAN results respectively.

Residual	**Filtered Empirical model**	**Filtered DBSCAN**
Count	9694	10088
Mean [m]	−0.21	−0.07
Std [m]	0.29	0.33
Min [m]	−2.94	−8.33
Max [m]	2.08	2.43
25% [m]	−0.34	−0.17
50% [m]	−0.24	−0.08
75% [m]	−0.13	0.03

From Table 5.3, the medians of −0.24 m and −0.08 m indicate a bias for both models of 24 cm for the empirical model results and of 8 cm for the DBSCAN model. However, the standard deviation is smaller for the empirical model results of 0.29 m, compared to that for the DBSCAN model, which is 0.33 m.

Chapter 6

Discussion

In this chapter, the results from Chp. 5 are evaluated and discussed. The chapter is divided into sections, where the model performance is discussed individually. Finally, a comparative study is made, where the performance is compared with the performance of models from other studies.

6.1 DBSCAN

The report distinguishes between the resulting unfiltered heights from the DBSCAN model and the heights with an applied smoothing filter. The results were compared with the SDB EOMAP data, where residuals between the heights were computed. The residuals were analyzed as a function of height, and the statistics of the residuals were calculated.

Generally, the DBSCAN model succeeds in finding the bathymetry from the noisy ICESat-2 files. However, the unfiltered results from the DBSCAN model reveal that the model fails to exclude much noise near the sea surface and detect the low-density bathymetry.

The noise around the sea surface is not excluded because the noise's point density is very close to the actual bathymetry. Hence, the DBSCAN models treat these points as bathymetry; however, most noise is removed when the smoothing filter is applied. Concerning the low-density bathymetry, the DBSCAN model searches for points with the same density based on the parameters ε and MinPts; thus, if the bathymetry has different densities, some of the bathymetry is excluded.

When comparing the results to the EOMAP data in Figure 5.6, there are a lot of significant residuals around the sea surface, which can occur because the DBSCAN model includes sea surface photons, where there is underlying bathymetry. The significant residuals are very likely a result of the included sea surface noise and excluded low-density bathymetry. The statistics in Table 5.1 also reveal that with a median of the residuals, of $\text{Med}_{\text{DBSCAN}} = -4\,\text{cm}$, the results from the DBSCAN model tend to be estimated 4 cm closer to the sea surface than the EOMAP heights, which is within the uncertainty of the EOMAP data, cf. 3.5b.

6.2 Empirical Model

The smoothing filter that was applied to the DBSCAN model was also applied to the empirical model results to ensure that the DBSCAN model and the empirical model are compared on the same basis. The empirical model results are compared to the EOMAP heights by computing the residuals and analyzing them as a function of height and by regarding the statistics as in Sec. 6.1.

The empirical model captures the bathymetry very well as the results include very little noise. However, it fails to include some of the low-density bathymetry, cf. Figure 5.3. When comparing the results to the EOMAP data, the residuals reveal that there are generally no significant deviations. However, statistics in Table 5.2 indicate that the results from the empirical model tend to be 21 cm closer to the sea surface than the heights EOMAP heights.

Concerning the fact that the bias, visible in Figure 5.17, in the empirical model results is more significant than that in the DBSCAN model results, hours of investigation by comparing the preprocessing scripts have been spent. The investigation shows that the odd outliers are removed in the same way, the same approach for correcting for the refraction is used, and the exact data for subtracting the geoid is used; thus, the origin of the bias was not found.

An issue occurred during the interpolation process between the results from the empirical model and the EOMAP data. As there are more decimals in the coordinates of the EOMAP data than there are in the coordinates of the empirical model results, 23 points outside the EOMAP area are present in Figures 5.12 and 5.14b, cf. the red circle in Figure 6.1.

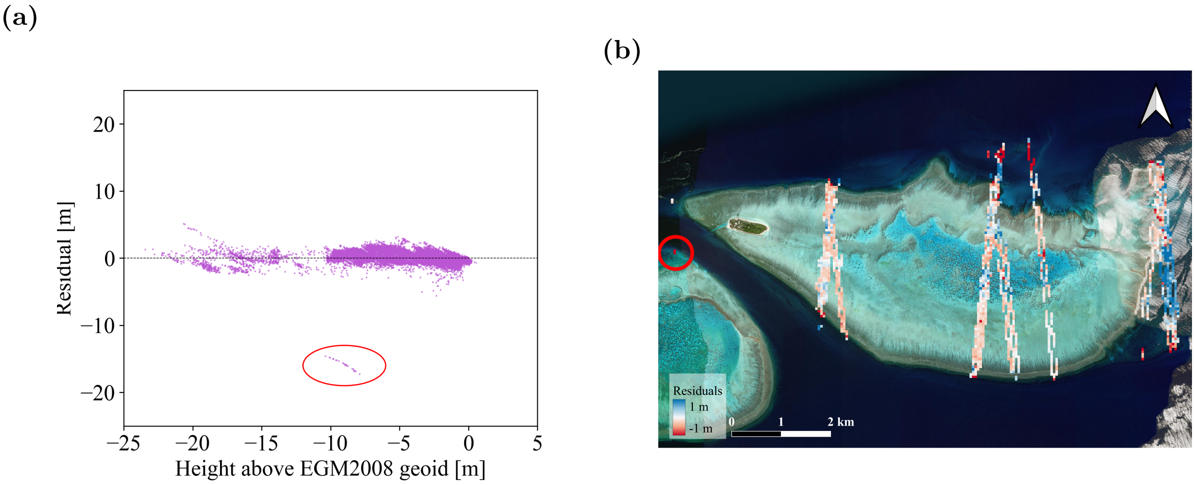


Figure 6.1: Figures 5.12 and 5.14b with a red circle indicating the 23 points.

These 23 points make 0.02% and do not impact the overall performance measures of the empirical model, except for the minimum value in Table 5.2. Furthermore, the points do not change any conclusions based on the performance measures. Hence, it was decided not to recompute the results with higher precision.

6.3 Comparative Analysis

In [18], the performance was evaluated by creating a confusion matrix of labeled signal and noise photons, where the precision, recall, and F-score are computed. The truth values were manually labeled based on the land type signal confidence of the ATL03 dataset. A confusion matrix was not chosen for this comparison as the output of the empirical model does not label the photons per se. If a confusion matrix was desired, the labels would need to be generated using the high confidence heights and the raw ICESat-2 data. Furthermore, a confusion matrix will not provide information where the model fails to predict the bathymetry.

In [19] and [22], they both used RMSE to evaluate the performance of the results from ICESat-2 data. In [22], the resulting depths are compared with in situ depths, and the RMSE is computed to be 0.68 m in the St. Thomas area in the Caribbean. In [19], an indirect way of estimating the accuracy of the bathymetry derived from the ICESat-2 data is used as there was no available in situ data, and the RMSE was estimated to be lower than 0.5 m in Yongle Atoll, in the South China Sea.

As there are no available in situ data for Heron reef, it was decided to estimate the accuracy by comparing with the high accuracy SDB EOMAP data. The accuracy of the heights from the DBSCAN model and the empirical model is evaluated from the coefficient of determination, R^2, and the RMSE. The difference in the performance parameters, RMSE and R^2 are calculated to be,

$$\begin{aligned} R^2_{\text{EM}} - R^2_{\text{DBSCAN}} &= 0.97 - 0.84 = 0.13 && (6.1) \\ \text{RMSE}_{\text{EM}} - \text{RMSE}_{\text{DBSCAN}} &= 0.46\,\text{m} - 0.96\,\text{m} = -0.50\,\text{m}, && (6.2) \end{aligned}$$

which indicates that the results from the empirical model make a better linear fit with the EOMAP data as the R^2 parameter is closer to 1 and the RMSE parameter is lower than that for the DBSCAN result. The performance of the DBSCAN model is poorer because the model results contain much of the sea surface, cf. Figure 5.5.

The sea surface photons might be included because of uncertainties tied to the sea surface's computation. In this project, the sea surface is calculated from the median of all heights with a buffer of 0.5 m as in [4]. Thus, the calculated sea surface is a rough estimate. The sea surface might be slightly above or below, affecting the refraction correction as it depends on the depth computed from the sea surface. With a more precise computation of the sea surface, the DBSCAN model would likely have included less of the sea surface and thus, have improved performance. In [19], the sea surface is computed by including the fluctuations in the sea surface, where the local MSL and the Root Mean Square (RMS) wave height were calculated by the mean and standard deviation from the detected photons on the sea surface.

Comparing the RMSE from the two models with the ones in [19] and [22], the heights from the empirical model have a lower RMSE compared to both [19] and [22]. The RMSE for the DBSCAN model is significantly higher than that in [19] and slightly higher than the one in [22].

The differences in the medians for the residuals of the overall results of the two models and the EOMAP data is,

$$\text{Med}_{\text{EM}} - \text{Med}_{\text{DBSCAN}} = -17\,\text{cm} \tag{6.3}$$

Hence, despite using the same data, parameters, and approach for the preprocessing step, the bias is 17 cm bigger than that for the DBSCAN model.

The significance of the bias compared to actual bathymetry in the area is challenging to address as the heights are not compared to in situ measurements. Instead, the EOMAP data was used to evaluate the models' performances. As the EOMAP data is also a product of a model with results made from satellite data, there are uncertainties tied to the data, cf. Figure 3.5b in Sec. 3.3, and the biases for both the DBSCAN model and the empirical model are within that uncertainty.

Generally, both of the models are good at estimating the bathymetry from the ICESat-2 data in the Heron reef area, where the waters are generally very clear and the bottom is mostly made of sand. The area has great conditions for ICESat-2 to receive a clear bathymetry signal. The major differences in the performance of the models are that the DBSCAN model results contain more sea surface noise, cf. Figure 5.5 and contain less of the low density bathymetry, cf. Figure 5.3. However, the empirical result seems to have a significantly bigger bias than the DBSCAN model when a comparison is made with the EOMAP data, cf. Figure 5.17 and Eq. 6.3. Furthermore, it appears that both the DBSCAN model and the empirical model estimate the lower heights, $< -10\,\text{m}$ differently compared to the EOMAP heights, cf. Figure 5.15. The different heights can be due to different approaches to correcting for refraction. As previously mentioned, there are uncertainties related to the computation of the sea surface. The uncertainties also lead to uncertainty in the refraction correction as it depends on depth, which depends on the sea surface.

Judging from the results and the literature that has been discussed in this chapter, there is great potential for using machine learning to determine bathymetry in shallow water regions with clear waters as there is in the Heron reef. More data will be needed with different water and seabed conditions to make a model that can be implemented globally, as dirty water will result in more noisy data. Furthermore, more data will be required to supplement the ICESat-2 data both to increase the amount of data and because the ATLAS instrument has a limited range to $\approx 40\,\text{m}$ depth.

Chapter 7

Conclusion

The objectives of this project are to determine bathymetry in a particular area, create a machine learning model that can determine bathymetry, compare the machine learning model with the empirical model, and evaluate machine learning as a tool for determining bathymetry.

An unsupervised machine learning model called DBSCAN has been implemented to accommodate the challenge of finding bathymetry from ICESat-2 LiDAR data. The critical parameters for using the DBSCAN algorithm, MinPts, and ε, are obtained from a priori knowledge about how reflections from the sun interfere with the amount of noise in the LiDAR data. The estimation of the parameters is inspired by similar studies using DBSCAN to determine bathymetry and from generic methods suggested for finding the parameters.

Generally, the DBSCAN model performed well and succeeded in finding bathymetry. However, when compared to high accuracy SDB data from EOMAP, bias plots revealed that the model had difficulty omitting the sea surface photons. Furthermore, the model was challenged with finding the photons in low-density bathymetry areas.

The empirical model, from [4], which is used to compare with the DBSCAN model, is a statistical interpolation method that classifies photons as high, medium, and low confidence bathymetry. In this study, only the high confidence bathymetry is considered. A smoothing filter equivalent to the one applied to the DBSCAN model results have been used to ensure that a comparison is made on the same basis.

A comparative study of the two models shows that both the DBSCAN model and the empirical performed well at determining bathymetry in the Heron reef area, which is part of the great barrier reef in Australia. Results show that the empirical model was better at omitting the sea surface photons, whereas the results from the DBSCAN have included several. Furthermore, both models are challenged in the areas where the bathymetry signal has a low point density. Furthermore, the empirical model has a more significant bias than the DBSCAN model.

The RMSE for the overall results from the two models, are $\text{RMSE}_{\text{EM}} = 0.46\,\text{m}$ and $\text{RMSE}_{\text{DBSCAN}} = 0.96\,\text{m}$, and the coefficients of determination, R^2, are $R^2_{\text{EM}} = 0.97$

and $R^2_{\text{DBSCAN}} = 0.84$, respectively. Hence, the heights from the empirical model are generally closer to the EOMAP heights than the heights from the DBSCAN model. The reason for this is mainly that the DBSCAN results include a lot of the sea surface.

However, statistics of residuals between the model results and the high accuracy SDB from EOMAP reveal that the results from the empirical model have a bias of -21 cm. The bias indicates that the resulting heights from the empirical model tend to be 21 cm closer to the sea surface than the EOMAP heights. The preprocessing of the raw ICESat-2 data includes removing the geoid, masking the area and signal confidence, removing the sea surface, and refraction correction. The origin of the bias has caused great wonder, and after hours of inspecting the scripts, the reason for the bias was not found.

The model performances were compared to models in similar studies using ICESat-2 data, conducted in St. Thomas, in the Caribbean, and Yongle Atoll in the South China Sea, respectively. The comparison showed that the RMSE of the empirical model is performing slightly better than both models in the similar studies, and the RMSE for the DBSCAN model is slightly poorer than both models in the studies. In similar studies, only the results in one of the studies were compared to actual in situ data. In the other study, the RMSE value was estimated indirectly from the ICESat-2 data.

In this project, no in situ data were available. Thus, the EOMAP was used instead, and throughout the evaluation of the DBSCAN model and the empirical model, the uncertainty of the EOMAP data is considered.

Based on the performance of the DBSCAN model and the models from the literature showcasing different types of machine learning models for the particular purpose of determining bathymetry, it can be concluded that machine learning has great potential for this specific topic.

Chapter 8

Future Work

Throughout the project, the models have been compared in the Heron reef, a small area with clear water. To test the model performance further, future work can include expanding the area of interest to areas with more challenging conditions.

If more time were provided for finishing the project, the main focus would be to improve the DBSCAN model to account for the low-density areas, and the included sea surface noise. A way to do that can be by implementing another method of calculating the sea surface, e.g., as in [19]. Furthermore, the model might be improved by segmenting the data such that there will be a unique model for each segment. The sizes of the segments and whether they should be in the vertical or horizontal direction or by some points would require methodically moving forward as it appears from the literature that all three have been done previously [19][20][21].

An alternative to applying the smoothing filter after using the DBSCAN algorithm can be to apply the DBSCAN algorithm twice or combine it with other machine learning algorithms. Other machine learning models could be; long short term memory (LSTM), random forest or support vector regression (SVR) as in [21]. However, as these are all supervised machine learning algorithms, they will require ground truth labeled data and significantly more data, which can be obtained by expanding the area. In situ, data would be required to ensure sufficient accuracy of the results from these supervised models. If the ground truth data comes from a model, the machine learning model will not be better than the ground truth.

Ultimately, a contribution to the Seabed 2030 project could be made from the resulting heights. However, making a global bathymetry model will require a combination of data as the ICESat-2 data is limited to $\approx 40\,\text{m}$ and only has a repeat cycle of 91 days. The combination could be including multi-spectral Sentinel-2 data as in [19] and [21], Landsat-8 as in [23], or Worldview-2 which is used to generate the EOMAP data [14].

Bibliography

[1] D. Monahan. “Bathymetry”. eng. In: *Encyclopedia of Ocean Sciences* (2019). Ed. by JK Cochran, HJ Bokuniewicz, and PL Yager, pp. 45–52. DOI: 10.1016/B978-0-12-409548-9.11290-4.

[2] S. M. Smith. “Seabed 2030: A Call to action”. eng. In: *Hydro International* 22.1 (2018), pp. 22–23. ISSN: 13854569.

[3] M. Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. eng. In: *Kdd-96 Proceedings. Second International Conference on Knowledge Discovery and Data Mining* (1996). Ed. by E. Simoudis, J. Han, and U. Fayyad, pp. 226–231.

[4] H. Ranndal et al. “Evaluation of a Statistical Approach for Extracting Shallow Water Bathymetry Signals from ICESat-2 ATL03 Photon Data”. eng. In: *Remote Sensing* 13.17 (2021), p. 3548. ISSN: 20724292. DOI: 10.3390/rs13173548.

[5] C. E. Parrish et al. “Validation of ICESat-2 ATLAS bathymetry and analysis of ATLAS’s bathymetric mapping performance”. eng. In: *Remote Sensing* 11.14 (2019), p. 1634. ISSN: 20724292. DOI: 10.3390/rs11141634.

[6] National Oceanic and Atmospheric Administration (NOAA) Coastal Services Center. *What is lidar?* Last accessed 15. April 2022. 2021. URL: https://oceanservice.noaa.gov/facts/lidar.html.

[7] P. F. McManamon. *LiDAR technologies and systems*. eng. SPIE Press, 2019. ISBN: 1510625399.

[8] National Oceanic and Atmospheric Administration (NOAA) Coastal Services Center. *Lidar 101: An Introduction to Lidar Technology, Data, and Applications*. 2012.

[9] R. Garcia-Dias et al. “Clustering analysis”. eng. In: *Machine Learning: Methods and Applications To Brain Disorders* (2019), pp. 227–247. DOI: 10.1016/B978-0-12-815739-8.00013-4.

[10] H. Belyadi and A. Haghighat. “Chapter 4 - Unsupervised machine learning: clustering algorithms”. In: *Machine Learning Guide for Oil and Gas Using Python*. Ed. by H. Belyadi and A. Haghighat. Gulf Professional Publishing, 2021, pp. 125–168. ISBN: 978-0-12-821929-4. DOI: https://doi.org/10.1016/B978-0-12-821929-4.00002-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128219294000020.

[11] T. A. Neumann and A. Brenner. “ATLAS/ICESat-2 L2A Global Geolocated Photon Data, Version 5”. In: *Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center* (2021). DOI: https://doi.org/10.5067/ATLAS/ATL03.005.

[12] A. L. Neuenschwander and L. A. Magruder. “Canopy and terrain height retrievals with ICESat-2: A first look”. eng. In: *Remote Sensing* 11.14 (2019), p. 1721. ISSN: 20724292. DOI: 10.3390/rs11141721.

[13] T. Neumann. Last accessed on 21 June 2022. URL: https://icesat-2.gsfc.nasa.gov/space-lasers.

[14] Dr. K. Hartmann and P. Klinger. *Survey Report: Satellite-derived Bathymetry.* Nov. 2019.

[15] *Scipy.Interpolate.CloughTocher2DInterpolator documentation.* Last accessed 15. May 2022. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.CloughTocher2DInterpolator.html#id2.

[16] P. Alfeld. “A trivariate Clough-Tocher scheme for tetrahedral data”. eng. In: *Computer-aided Geometric Design* 1.2 (1984), pp. 169–181. ISSN: 18792332, 01678396. DOI: 10.1016/0167-8396(84)90029-3.

[17] *NOAA Solar Calculator: Sunrise, sunset, and noon for any place on Earth.* Last accessed 12 May 2022. 2020. URL: https://gml.noaa.gov/grad/solcalc/index.html.

[18] Z. Zhang et al. “Signal photon extraction method for weak beam data of icesat-2 using information provided by strong beam data in mountainous areas”. eng. In: *Remote Sensing* 13.5 (2021), pp. 1–29. ISSN: 20724292. DOI: 10.3390/rs13050863.

[19] Y. Ma et al. “Satellite-derived bathymetry using the ICESat-2 lidar and Sentinel-2 imagery datasets”. eng. In: *Remote Sensing of Environment* 250 (2020), p. 112047. ISSN: 18790704, 00344257. DOI: 10.1016/j.rse.2020.112047.

[20] Y. Ma et al. “Estimating water levels and volumes of lakes dated back to the 1980s using Landsat imagery and photon-counting lidar datasets”. eng. In: *Remote Sensing of Environment* 232 (2019), p. 111287. ISSN: 18790704, 00344257. DOI: 10.1016/j.rse.2019.111287.

[21] V. V. A. K. Surisetty et al. “Synergistic Fusion of ICESat-2 Lidar and Sentinel-2 Data to Leverage Potential Mapping of Bathymetry in Remote Islands Using SVR”. eng. In: *Journal of the Indian Society of Remote Sensing* (2022), pp. 1–9. ISSN: 09743006, 0255660x. DOI: 10.1007/s12524-022-01537-4.

[22] C. Xie et al. “Improved filtering of icesat-2 lidar data for nearshore bathymetry estimation using sentinel-2 imagery”. eng. In: *Remote Sensing* 13.21 (2021), p. 4303. ISSN: 20724292. DOI: 10.3390/rs13214303.

[23] T. Sagawa et al. “Satellite derived bathymetry using machine learning and multi-temporal satellite images”. eng. In: *Remote Sensing* 11.10 (2019), p. 1155. ISSN: 20724292. DOI: 10.3390/rs11101155.

Appendix A

Data

A.1 List of Data Files

Table A.1: List of ICESat-2 data files.

ATL03_20190317220620_12130214_005_01.h5
ATL03_20190408085308_01540308_005_01.h5
ATL03_20190616174556_12130314_005_01.h5
ATL03_20190915132546_12130414_005_01.h5
ATL03_20191215090534_12130514_005_01.h5
ATL03_20200405153209_01540708_005_01.h5
ATL03_20200614002507_12130714_005_01.h5
ATL03_20200705111156_01540808_005_01.h5
ATL03_20201004065141_01540908_005_01.h5
ATL03_20201212154448_12130914_005_01.h5
ATL03_20210612070430_12131114_005_01.h5
ATL03_20211002133119_01541308_005_01.h5

Appendix B

Results

B.1 Figures

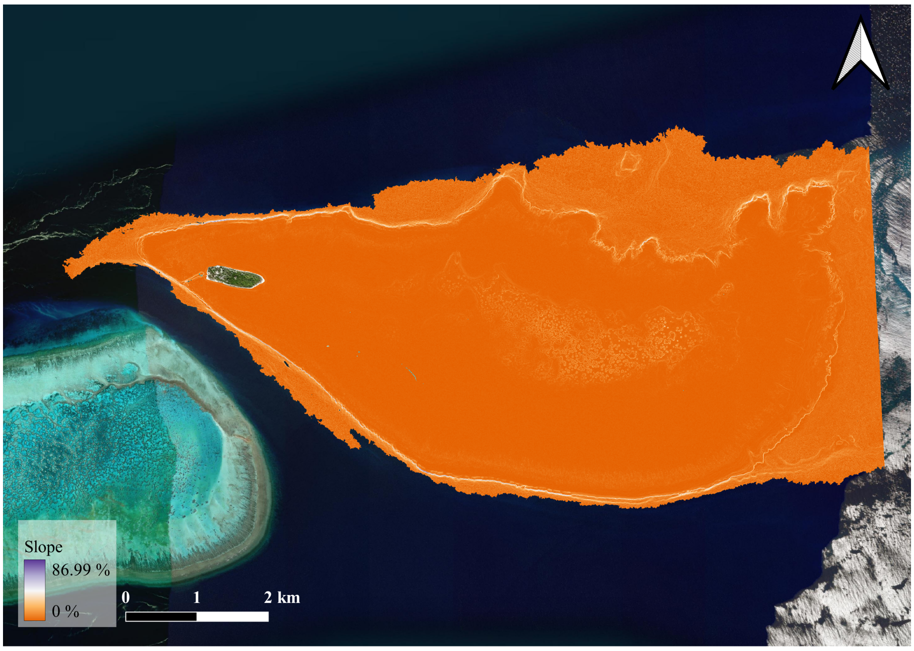


Figure B.1: Slope map computed in QGIS from the EOMAP data.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Technical
University of
Denmark

Elektrovej, Building 328
2800 Kgs. Lyngby
Tlf. 4525 1700