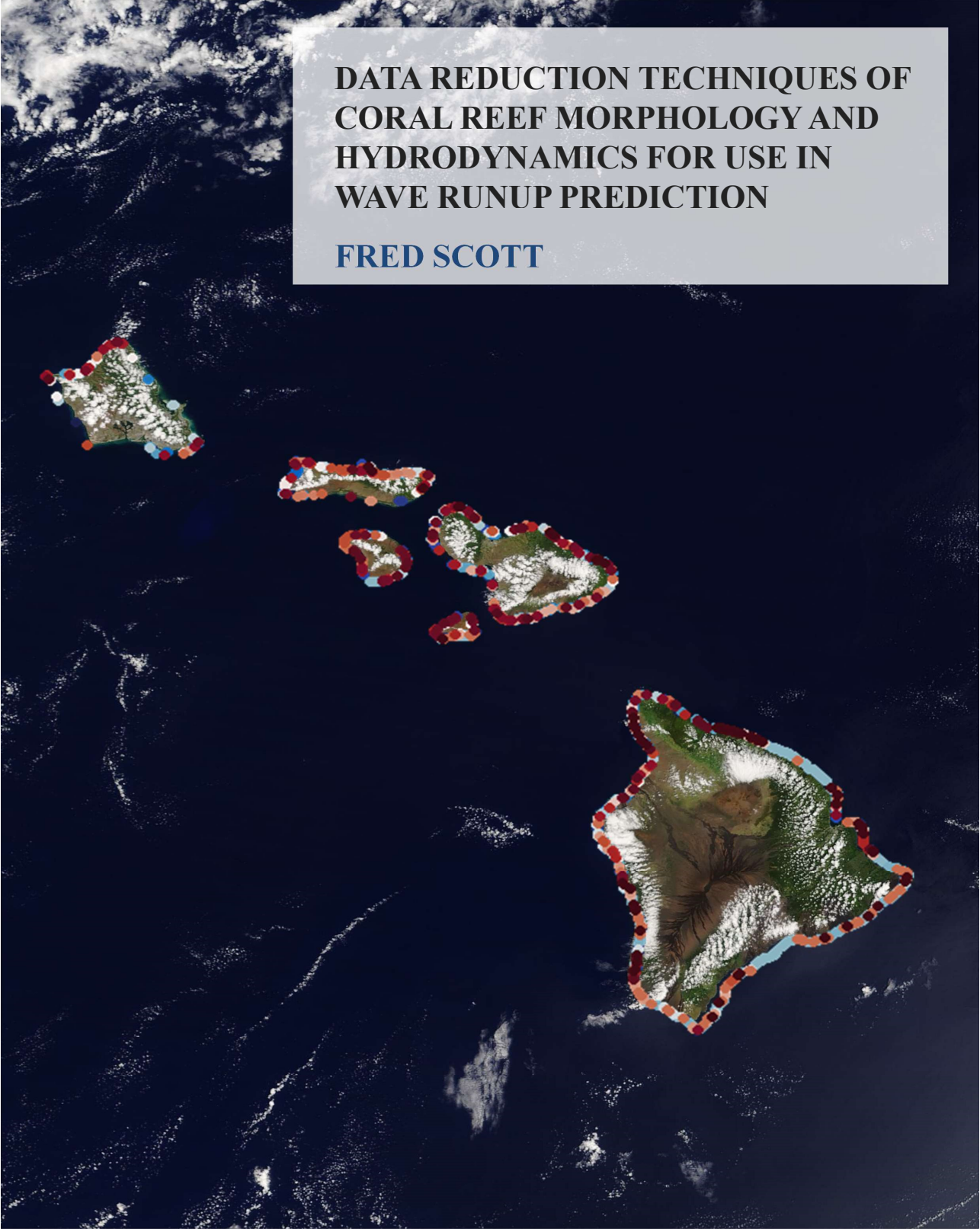
DATA REDUCTION TECHNIQUES OF CORAL REEF MORPHOLOGY AND HYDRODYNAMICS FOR USE IN WAVE RUNUP PREDICTION

FRED SCOTT



TU Delft Delft University of Technology

Challenge the future

ERASMUS +: ERASMUS MUNDUS MOBILITY PROGRAMME

Master of Science in

COASTAL AND MARINE ENGINEERING AND MANAGEMENT

CoMEM

DATA REDUCTION TECHNIQUES OF CORAL REEF MORPHOLOGY AND HYDRODYNAMICS FOR USE IN WAVE RUNUP PREDICTION

by

Fred Scott

Delft University of Technology

to be defended publicly on Wednesday July 17, 2019 at 10:30 AM at the Delft University of Technology

The Erasmus+: Erasmus Mundus MSc in Coastal and Marine Engineering and Management is an integrated programme including mobility organized by five European partner institutions, coordinated by Norwegian University of Science and Technology (NTNU). The joint study programme of 120 ECTS credits (two years full-time) has been obtained at two or three of the five CoMEM partner institutions:

- Norges Teknisk- Naturvitenskapelige Universitet (NTNU) Trondheim, Norway
- Technische Universiteit (TU) Delft, The Netherlands
- Universitat Politècnica de Catalunya (UPC). BarcelonaTech. Barcelona, Spain
- University of Southampton, Southampton, Great Britain
- City University London, London, Great Britain

During the first three semesters of the programme, students study at two or three different universities depending on their track of study. In the fourth and final semester an MSc project and thesis has to be completed. The two-year CoMEM programme leads to a multiple set of officially recognized MSc diploma certificates. These will be issued by the universities that have been attended by the student. The transcripts issued with the MSc Diploma Certificate of each university include grades/marks and credits for each subject.

Information regarding the CoMEM programme can be obtained from the programme coordinator:
Øivind A. Arntsen, Dr.ing.
Associate professor in Marine Civil Engineering
Department of Civil and Environmental Engineering
NTNU Norway
Mob.: +4792650455 Fax: + 4773597021
Email: oivind.arntsen@ntnu.no
CoMEM URL: https://www.ntnu.edu/studies/mscomem

CoMEM Thesis

This thesis was completed by:

Fred Scott

Under supervision of:
Prof. dr. ir. A.J.H.M. Reniers, TU Delft
Dr. ir. R.T. McCall, Deltares
Dr. ir. C. Storlazzi, USGS
Dr. ir. J.A.A. Antolínez, Deltares
Ir. S.G. Pearson, TU Delft/Deltares

As a requirement to attend the degree of
Erasmus+: Erasmus Mundus Master in Coastal and Marine Engineering and Management (CoMEM)

Taught at the following educational institutions:

Norges Teknisk- Naturvitenskapelige Universitet (NTNU)
Trondheim, Norway

Technische Universiteit (TU) Delft
Delft, The Netherlands

University of Southampton
Southampton, Great Britain

At which the student has studied from August 2017 to July 2019.

In collaboration with:

Keywords — Data reduction, cluster analysis, data mining, coral reefs, low-lying islands, coastal flooding, early warning systems, climate change, XBeach

Front cover — A satellite image of Hawaii. The colored dots represent the locations of coral reef profiles included in the analysis. The profiles with the same colors belong to the same cluster group, which were formed based on similar reef morphology. (Source: (Nasa(EOSDIS),2019))

Back cover — A world map to show the main locations of the coral reef profiles included in this study.

SUMMARY

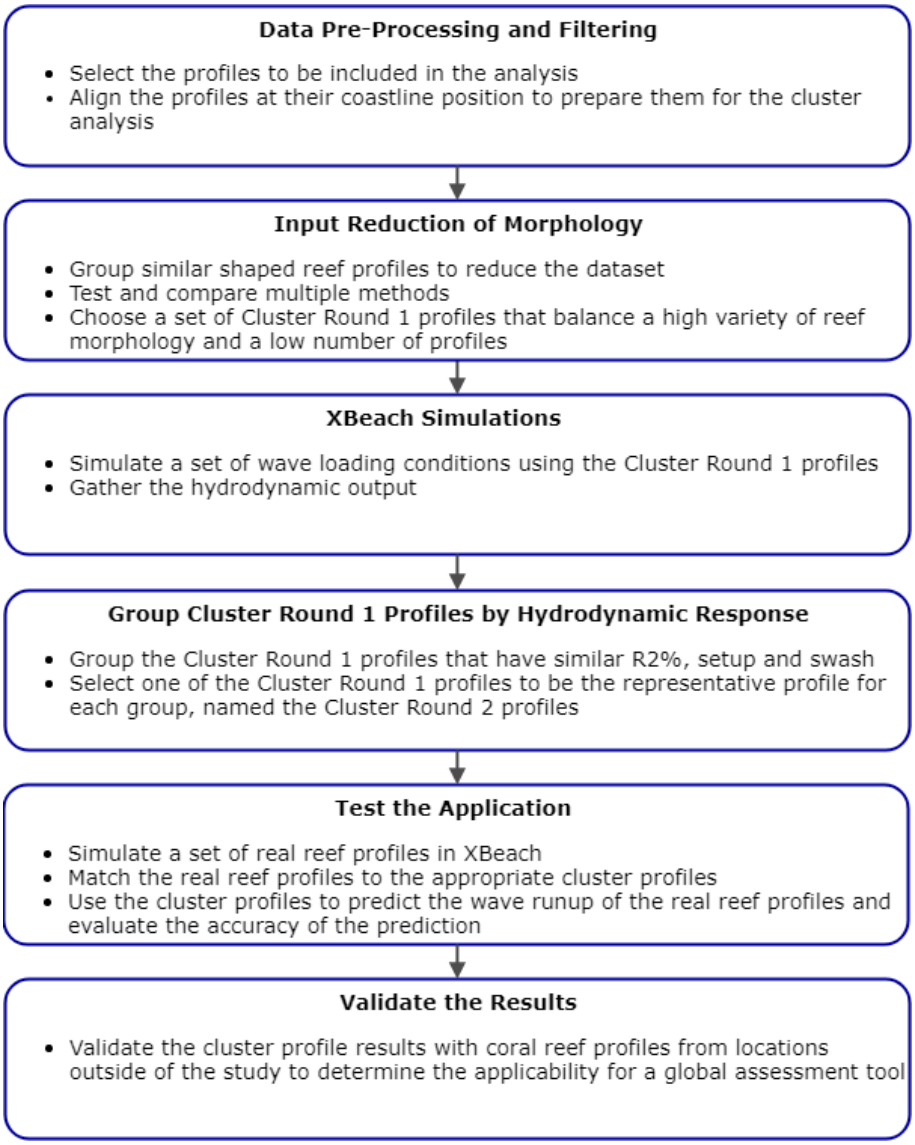
Many tropical, coral reef-lined coasts, are low-lying with elevations less than five meters above mean sea level. Climate-change-driven sea level rise, coral reef decay and changes in (storm) wave climate will lead to greater chance and impacts of wave-driven flooding, posing a heavy threat to these coastal communities. Early warning systems (EWS) are effective for risk management and disaster reduction, however, the vast majority of the world's inhabitants of coral reef-lined coasts have no such system in place. Unfortunately, the complex hydrodynamics and bathymetry of reef-lined coasts make it difficult to establish a global flood prediction model for these areas.

This thesis aims to develop a set of 'cluster profiles' that can be used to accurately represent coral reef-lined coasts around the globe. By representing an expansive variety of reef morphology, the cluster profiles are capable of predicting the wave runup over thousands of different coral reef profiles with a fraction of the number. The cluster profiles could be input into a tool such as a Bayesian probabilistic network which can be trained to provide real-time wave runup and flooding predictions given local bathymetry and offshore wave conditions, thus establishing a simplified global flooding EWS.

The methodology includes two stages of data reduction. First, cluster analysis techniques are used to group thousands of coral reef profiles into 500 clusters based on morphology alone. Second, agglomerative hierarchical clustering is used to further group the profiles with similar morphology and wave runup response, resulting in a final set of 311 to 45 cluster profiles.

Here we show that the cluster profiles are capable of predicting the wave runup for a set of 1000 reef profiles with a mean relative difference of approximately 10%. The comparison was done using the numerical wave model XBeach with four different wave conditions.

The methodology has been developed such that it could be expanded to other coastal environments. A summary of the methodology used in the study is illustrated on the following page.


Data Pre-Processing and Filtering
Select the profiles to be included in the analysis
Align the profiles at their coastline position to prepare them for the cluster analysis
Input Reduction of Morphology
Group similar shaped reef profiles to reduce the dataset
Test and compare multiple methods
Choose a set of Cluster Round 1 profiles that balance a high variety of reef morphology and a low number of profiles
XBeach Simulations
Simulate a set of wave loading conditions using the Cluster Round 1 profiles
Gather the hydrodynamic output
Group Cluster Round 1 Profiles by Hydrodynamic Response
Group the Cluster Round 1 profiles that have similar R2%, setup and swash
Select one of the Cluster Round 1 profiles to be the representative profile for each group, named the Cluster Round 2 profiles
Test the Application
Simulate a set of real reef profiles in XBeach
Match the real reef profiles to the appropriate cluster profiles
Use the cluster profiles to predict the wave runup of the real reef profiles and evaluate the accuracy of the prediction
Validate the Results
Validate the cluster profile results with coral reef profiles from locations outside of the study to determine the applicability for a global assessment tool

CONTENTS

List of Figures

LIST OF TABLES

LIST OF SYMBOLS

Symbol	Units	Description
g	ms^{-2}	Gravitational constant (9.81 m/s2)
H_s	m	Significant wave height
H_o	m	Offshore wave height
H_i	m	Wave height
L_o	m	Offshore wavelength
η_o	m	Offshore water level
$\frac{H_o}{L_o}$	-	Wave steepness
T	s	Wave period
f	s^{-1}	Wave frequency
R	(m)	Wave runup
$R_{2\%}$	(m)	Wave runup exceeded 2% of the time
S	m	Swash
η	m	Setup
c_f	-	Coefficient of friction
B_{beach}	-	Beach slope
B_{reef}	-	Fore reef slope
W_{reef}	m	Reef width
z_{beach}	m	Beach crest
h_i	m	Water depth
γ_i	-	Similarity parameter
σ	of data	Standard deviation
IC	-	Inconsistency coefficient

LIST OF ABBREVIATIONS

Acronym	Definition
EWS	Early Warning System
MSL	Mean sea level
SLR	Sea level rise
USGS	United States Geological Survey
IPCC	Intergovernmental Panel on Climate Change
BEWARE	Bayesian Estimator for Wave Attack in Reef Environments
BN	Bayesian Network
GMM	Gaussian Mixture Model
CH	Calinski-harabasz evaluation
DB	Davies bouldin evaluation
AIC	Akaike information criterion
BIC	Bayesian information criterion
LIDAR	Light Detection and Ranging
SDB	Satellite Derived Bathymetry

ACKNOWLEDGMENTS

This thesis concludes my two year journey, through three European countries, and four different apartments. I have had a wonderful experience while completing my CoMEM Master's degree, mainly because of all of the wonderful people that were involved. I'd first like to thank the CoMEM organizers, in particular, Øivind and Sonja. You welcomed us to Norway with open arms and warm waffles, fostering memories that will last a lifetime.

I would like to express my deep gratitude to my committee. Robert, thank you for taking me on as a thesis student. You were a great leader, always offering your time and assistance, and a pleasure to work with. José, thank you for your many dedicated hours and sharing your big data knowledge. Your contribution to this project was immense and I deeply appreciate your friendly support and creative ideas. Stuart, thank you for your enthusiasm towards the project. It was always very nice to get reminded by you about how cool or interesting a finding was, and this helped keep the research fresh and fun. I wish you all the best with the remainder of your P.h.D and hope to see you back in Canada at some point! Curt, thank you for your kindness and guidance. You were always able to steer us in the right direction when the path was not clear. Ad, thank you for your insightful ideas and challenging questions. Your experience and input were key contributors to the success of this project.

These two years would have been nothing without the great friendships developed along the way. I was so lucky to have such great classmates from all around the world to share this experience with. To my classmates that became housemates - Matteo, Flo, Nikos, Hassan and Ingrid, thank you for all of the laughs and good times. The Holyrood will never see a better bunch of people. For a chance at the Catan title, I'm afraid you'll have to visit me in Canada. I wish you all the very best, and I look forward to our reunions, as well as more surf and kite trips (Matteo and Flo)!

I would never have been able to do this program without the support from my family and friends back home. To my mom and dad, thank you for giving me every opportunity to pursue my passion for coastal engineering. Your quick visits to Delft recharged my batteries and always reminded me of the amazing support I have across the pond. To Miah, it was always the semester highlight to get a visit from you and go exploring, from the cabin in the mountains of Norway to the Roman Baths in England and the fries and chocolate in Belgium. Thank you for your endless care and support.

Fred Scott
Delft, July 2019

1

INTRODUCTION

CHAPTER SUMMARY
The following chapter details the motivation and purpose of the research. The research goals and approach are outlined, explaining how data reduction techniques will be applied to a large coral reef dataset. This is done to make predictions of wave runup and flooding for coral reef-lined coasts, including the extremely vulnerable coral atoll islands in the Pacific Ocean.

1

1.1. MOTIVATION

Coastal flooding from extreme weather events affects thousands of vulnerable coastal communities around the world. This is particularly true for many tropical, low-lying, reef-fronted islands know as coral atolls. Coral atolls are frequently claimed to be some of the most threatened coastal systems in the face of sea level rise (SLR) due to the combination of the various stresses on these natural systems, low human adaptive capacity and high exposure (McLean et al., 2001; Nicholls et al., 2007).

This study aims to aid the development of a global tool for estimating wave runup and flooding for reef-fronted coasts. There are millions of people living in areas at risk of coastal flooding (UNFPA, 2014), and the vast majority have no early warning system (EWS) in place. A global tool to better understand and predict wave runup will increase coastal resilience by providing timely information of potential flood events to local communities. The methodology developed in this study is designed to be transferable to other coastal environments, providing numerous opportunities and applications in large scale wave runup prediction.

1.1.1. CORAL ATOLL'S VULNERABILITY TO WAVE ATTACK

Atolls are defined as an annular mid-ocean reef around a central lagoon (Woodroffe, 2008). Most atolls have maximum elevations of less than 4 m and average elevations of less than 2 m above present sea level (Storlazzi et al., 2018). The elevation alone makes these islands extremely vulnerable to large waves and storm surge associated from tropical cyclones, as well as flooding that can occur from "blue sky events". Blue sky events refer to the idea that large waves and potential flooding can occur even when the weather seems calm due to the arrival of remotely generated long-wavelength wind-waves (swell). Swell results in wave setup, the elevation of the mean still water surface due to the breaking of the waves (Longuet-Higgins & Stewart, 1964), and typically causes the generation of infragravity waves (Pomeroy, Lowe, Symonds, van Dongeren, & Moore, 2012), causing wave runup at the shoreline, leading to flooding. Low land elevations increase the relative influence of the incoming swell waves and therefore makes these islands extremely vulnerable.

The Pacific Island countries have a population of almost 10 million people, and although the death toll and number of victims of natural disasters in these countries may appear low in comparison to worldwide statistics, they rank among the highest per number of inhabitants (ECHO, 2019). The European Commission and World Bank provide funding for disaster relief to many Pacific countries, including for flooding and storm surges (ECHO, 2019). It is estimated that the average annual direct losses caused by natural disasters in the South Pacific region are US$284 million (The World Bank, 2012).

There are also extreme consequences due to flooding that cannot be monetized. There are roughly 1,000 populated small islands in the Pacific Ocean, and for most of these islands, groundwater is the main source of freshwater (White & Falkland, 2010). The groundwater in these small islands occurs as "fresh groundwater lenses", which are thin veneers of fresh groundwater over top of seawater, in permeable and phreatic aquifers (White & Falkland, 2010) as shown in Figure 1.1. These freshwater lenses have a vital and increasing role to play in the future of public health, and environmental and

ecological stability of the island countries (Terry & Falkland, 2010). The quantity and quality of the freshwater are dependent upon the mixing and intrusion of seawater into the fresh groundwater, as well as human activities. The small size of the islands generally restricts the quantity to basic human needs but depends on atoll width, recharge rate and the ease of transmission of freshwater through the aquifers (White et al., 2007).

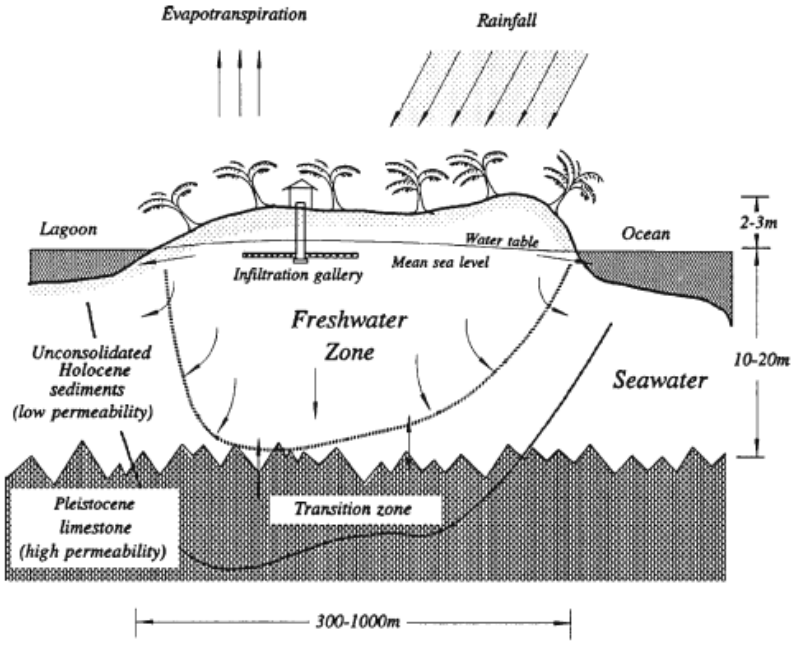


Figure 1.1: Schematized cross section of a groundwater lens on aa atoll. Source:(White & Falkland, 2010)

The main threat to freshwater lenses for low-lying islands is partial or complete overwash from storm waves and storm surge, as explained by Terry (2007), mainly associated with major tropical cyclones. Overwash of parts or all of some islands results in seawater intrusion into the freshwater lenses. Being able to predict which wave events will result in overwash could increase the time to prepare for such an event.

1.1.2. CLIMATE CHANGE IMPACTS

The overall threat to coral atoll islands increases dramatically with the addition of climate change effects. Although the dangers posed by climate change are well known, the world continues to act insufficiently to reduce emissions of greenhouse gases. At the time of writing this thesis, a measurement from the Mauna Loa Observatory in Hawaii recorded that the carbon dioxide concentration (CO_2) had reached 415 parts per million (ppm) which is the highest recorded concentration since measurements began in 1958 (CO2.Earth, 2019), and is calculated to be the highest the earth has seen in 3 million years (RTÉ, 2019). As emissions continue to leak and the greenhouse gas concentrations

continue to rise, the goal set by the Paris Agreement to limit global warming to no more than 1.5° above pre-industrial levels (Hulme, 2016) seems increasingly implausible.

The outcome is current global SLR at 3-4 mm/yr (Watson et al., 2015), which is expected to accelerate into the future (Church et al., 2013). Projections indicate that SLR in the tropics will be higher than the global average, thus increasing the expected stresses on atoll islands in the Pacific (Slangen et al., 2014). The majority of the coral reefs have vertical reef flat growth rates (2-6 mm/yr) up to an order of magnitude slower than rates of projected SLR (8-20 mm/yr) (Hall et al., 2016; Kopp et al., 2014; Storlazzi et al., 2018), which will result in a net increase in water depth over atoll reefs.

INCREASED WAVE RUNUP DUE TO SEA LEVEL RISE

Wave runup is the result of an interaction between waves, ocean level, and the morphology of coral reefs. Since the reefs will not be able to grow at the same rate as SLR (Storlazzi et al., 2018), the depth over the reef increases. Waves are dissipated by bottom friction and wave breaking, and with deeper water levels over the reef flat, there is simply less of an influence from the reef on the waves to dissipate energy before they reach the coastline.

Quataert, Storlazzi, Van Rooijen, Cheriton, and Van Dongeren (2015) showed the influence of different reef and hydrodynamic loading parameters on wave runup and other nearshore processes for reef-fronted coasts. Through a numerical model study, they altered the offshore water depth to represent SLR and concluded that the result of SLR is an increase in energy at the coastline due to less dissipation from the increased water depth. Therefore, in the future, there will be an increase in wave runup and thus total water levels at the shoreline (Cheriton, Storlazzi, & Rosenberger, 2016; Quataert et al., 2015), resulting in greater flooding consequences under the same wave conditions, and a higher frequency of flooding events. Vitousek et al. (2017) suggests that the 10 to 20 cm of SLR expected no later than 2050 will more than double the frequency of extreme water-level events in the Tropics, severely impairing the habitability of low-lying Pacific Island countries.

INCREASED WAVE RUNUP DUE TO CORAL REEF DAMAGE

Wave runup will also be increased due to ocean acidification and coral bleaching events. Decreasing seawater carbonate ion (CO_3^{2-}) concentrations because of rising atmospheric CO_2 are predicted to lower rates of calcium carbonate ($CaCO_3$) production of corals such that the rates of reef erosion will exceed rates of reef accretion across much of the tropics and subtropics (Pandolfi, Connolly, Marshall, & Cohen, 2011). Bleaching causes mortality of corals and reduces the energy available for growth and reproduction among survivors, therefore increases in bleaching frequency are expected to reduce coral cover (Pandolfi et al., 2011). The projected extent of the damage varies considerably, mainly due to the uncertainty associated with climate change projections and human efforts. The range of outcomes includes a complete collapse of coral cover by the middle of this century to similar levels of coral cover as present until the year 2100. The result in terms of wave runup is that a reduction in live coral coverage means reduced hydrodynamic roughness and increased water depths over the reef. This will further enhance wave energy propagating to the shoreline and wave-driven flooding (Cheriton et al., 2016; Quataert et al., 2015; Storlazzi et al., 2018).

INCREASE IN SALTWATER INTRUSION

The shoreline impacts translate to large consequences on the island. A major concern is the greater frequency of flooding events (Vitousek et al., 2017) leading to saltwater intrusion into the freshwater resources on the islands. A study on the freshwater lens of Pukapuka Atoll in the Northern Cook Islands (South Pacific Ocean) by Terry and Falkland (2010) provided the first field observations from two years of monitoring after a storm resulted in saltwater intrusion. They determined that the freshwater lens required 11 months to recover. On the barrier islands off the coast of North Carolina after overwash from Hurricane Emily in 1993, groundwater salinisation lasted up to three years, proving that the recovery process can be much longer. The effect of more frequent storms causing overwash on the islands will apply enormous pressure and pose great threats for the many populations depending on a fresh groundwater source.

FLOODING RISK

In the scientific community, risk is defined as a combination of the magnitude of the potential hazard, the exposure, and the vulnerability (Kron, 2005). The risk for atoll islands is high and will most likely only increase. The potential hazard, defined as the probability of occurrence of the threatening natural event, is most likely to increase with SLR and the more frequent number of storms (Nicholls et al., 2007). The exposure, defined as the values at risk including buildings and humans (Kron, 2005), will increase with developing populations on the islands, which are expected to grow for the vast majority of Pacific island countries (SDD, 2016). Lastly, the vulnerability, defined as the lack of resistance to damaging/destructive forces (Kron, 2005) of atoll countries is high due to the relatively low levels of physical infrastructure and their economic structure (Barnett & Adger, 2003). The coral reef acts as a natural defense to the destructive forces of waves, and as explained above in Section 1.1.2, the climactic impacts are posing a great risk on coral reefs and the protection they can provide (Storlazzi et al., 2018).

Understanding the risk of flooding for these islands is important for many reasons. Flooding will result in physical damages to property and infrastructure. It will cause economic destruction, causing financial burdens and inhibiting the ability to perform daily functions. In extreme cases, flooding can cause loss of life. Understanding the risk involved and finding ways to mitigate it is essential.

Overall, climate change effects will increase the risk of flooding on low-lying, reef-fronted islands. SLR, changes in the wave climate and reef degradation all lead to increased vulnerability (Quataert et al., 2015) and seem to be unavoidable (Kelman & Glantz, 2014). Therefore, the effort to fight climate change seems to be reverting to climate change adaptation. Schnieder et al. (2007) explains, “adaptation can significantly reduce many potentially dangerous impacts of climate change and reduce the risk of many key vulnerabilities.” However, a lack of technical, financial, and institutional capacity limits the implementation of effective adaptation strategies in many regions (Schnieder et al., 2007). Regions such as coral atoll islands that do not have the resources for structural flood protection measures need early warning systems.

1.1.3. THE NEED FOR FLOOD EARLY WARNING SYSTEMS

Early warning can be defined as 'the provision of timely and effective information, through identified institutions, that allows individuals exposed to a hazard to take action to avoid or reduce their risk and prepare for effective response' (UNISDR, 2004). It is an important part of a holistic approach to risk management of natural hazards (Alfieri, Salamon, Pappenberger, Wetterhall, & Thielen, 2012). The Flood Directive of the European Commission has specifically mentioned that early warning systems (EWS) are as an essential part of an effective preparedness towards natural disasters (European Union, 2007). On an international level, the Hyogo Framework for Action (United Nations, 2005) adopted at the United Nations World Conference on Disaster Reduction, emphasizes the need for building the resilience of society to disasters. Here, the cost-effectiveness of EWS is stressed, favoring the prevention rather than relying on post-disaster response and recovery. It is claimed that an EWS is an essential investment that protects lives and property, thus leading to a sustainable development. EWS has also been identified by the Intergovernmental Panel on Climate Change (IPCC) as an example of mitigation/adaptation technological innovation for disaster reduction and adaptation (de Coninck et al., 2018), as well as by the Sendai Framework for disaster risk reduction as an important aspect of understanding disaster risk and part of disaster preparedness for effective response (UNISDR, 2015).

Their effectiveness led the former United Nations Secretary-General Kofi Annan to call for the establishment of a global early warning system for all natural hazards (United Nations, 2006). Unfortunately, this call to action was was too late for the 2004 tsunami that struck the Indian Ocean region and killed thousands. The work proposed in this study aims to make progress towards a global EWS for coastal flooding, particularly for reef-fronted islands, which would serve a large population around the globe.

1.2. RESEARCH SIGNIFICANCE

The hydrodynamic behavior over coral reefs and the influence of the reef morphology on wave runup at the shoreline is well understood (Blacka, Flocard, Splinter, & Cox, 2015; Cheriton et al., 2016; Ferrario et al., 2014; Gourlay, 1994, 1996a, 1996b; Hardy & Young, 1996; Massel & Gourlay, 2000; Pearson, Storlazzi, van Dongeren, Tissier, & Reniers, 2017; Quataert et al., 2015; van Dongeren et al., 2013; Vetter et al., 2010; Young, 1989). This knowledge can be applied to create a world-wide EWS for reef-fronted coasts. To develop such a tool, the hydrodynamic response due to many different wave loading conditions over many variations of coral reef morphology must be simulated and gathered into a quick response network. The network could then provide an estimation of wave runup based on the pre-computed results. Choosing wave loading conditions is relatively straight forward and can be determined by wave hindcast models and local conditions, however trying to choose the appropriate reef morphology that can be representative of all coral reefs around the globe is difficult.

The goal of this study is to use the current knowledge of reef hydrodynamics as well as data reduction techniques to create a reduced subset of reef profiles that can be used for predicting wave runup and flooding on a global scale. Previous work on this subject has been done by Stuart Pearson who created the Bayesian Estimator for Wave Attack in Reef Environments (BEWARE) (Pearson, Reniers, van Dongeren, Tissier, & den Heijer, 2016)

during his Master's thesis. This model was developed to provide quick and reliable wave runup estimates for reef-fronted coasts, but was created using a schematized reef profile that represents one shape of common coral reefs. Using a subset of real reef profiles in the creation of such a tool would improve the effectiveness since a greater variety of coral reefs would be adequately represented. This would assist in large scale disaster preparedness.

1.2.1. BUILDING ON BEWARE

Pearson et al. (2017) developed a probabilistic model through a Bayesian Network (BN) to estimate wave runup on coral reef islands. A BN is a statistical tool which produces a probability distribution of likely outputs. The model is able to determine the probabilistic outputs because it is trained with similar data, and therefore uses its previous knowledge of associated inputs and outputs to calculate the outputs for a new set of inputs. For this case, the inputs into the model are the reef parameters (c_f, β_b,β_f, w_{reef}, z_{beach}), and the wave loading conditions (H_0, H_0/L_0, η_0) (see Figure 1.2). The outputs are the hydrodynamics over the reef including the wave runup at the shoreline. To generate the training dataset, many different reef profiles were created and simulated with several different wave loading conditions. Each different combination of loading conditions was paired with each different combination of reef profile shapes to build up roughly 400,000 XBeach simulations. With this training dataset, the model could then quickly interpret results of other loading scenarios and varying profile parameters.

The reef profiles that were used in this model are schematized reef profiles that are representative of many coral reefs in the Pacific Ocean. An example of the reef profile that Pearson et al. (2017) used is shown below in Figure 1.2.

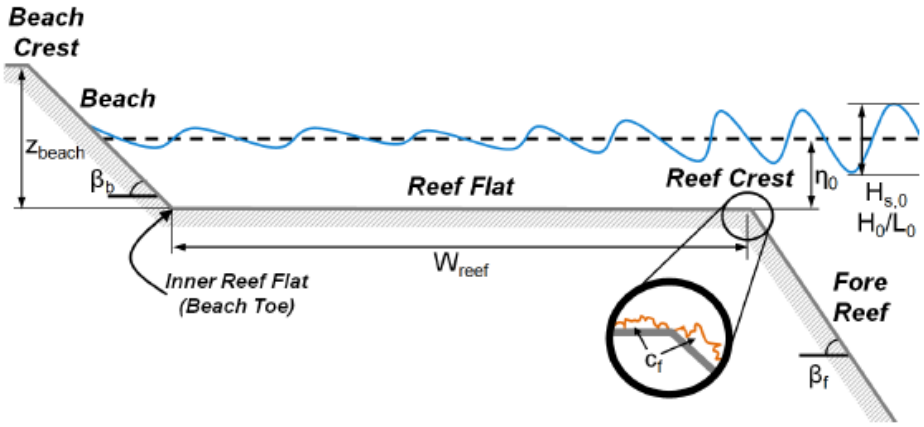


Figure 1.2: Simplified reef profile used in Stuart Pearson's thesis study to create BEWARE. Source: (Pearson, Reniers, van Dongeren, Tissier, & den Heijer, 2016)

The "Bayesian Estimator for Wave Attack in Reef Environments" (BEWARE) (Pearson et al., 2017) has been proven to be extremely valuable, but there are limitations that come

with it. Only profiles with the same input parameters can be used, and therefore each profile must have the same general shape, just with varying values. This is not representative of all reef profiles. Four reef profiles from Saipan are shown in Figure 1.3. These profiles all include features that make them quite different from the schematized profile that Pearson et al. (2017) used in Figure 1.2, and would, therefore, most likely be poorly represented by the BEWARE model.

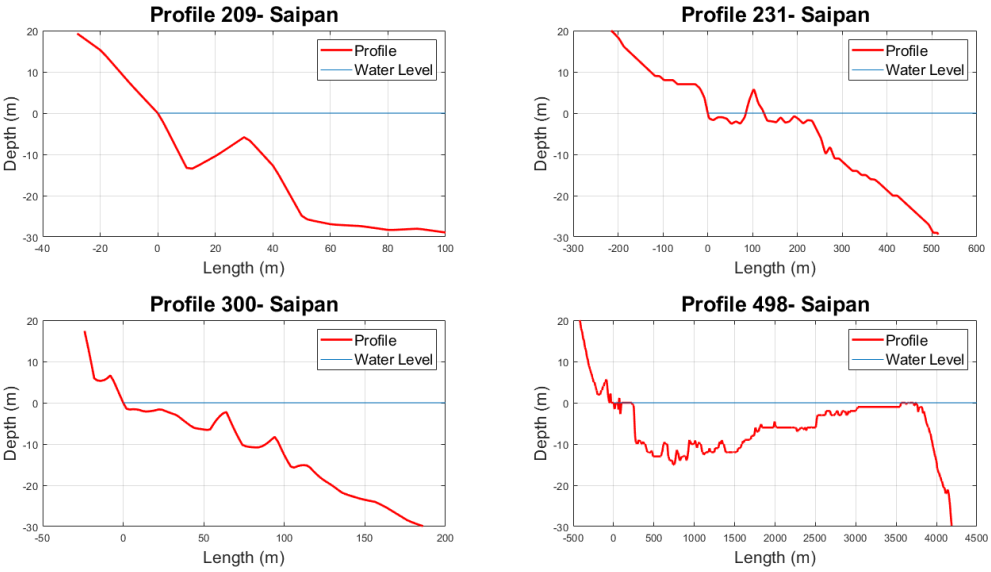


Figure 1.3: Examples of reef profiles included in the dataset that would not be represented well by the schematized reef profiles used to create BEWARE.

When using 3-7 values for each input, as was done to create BEWARE, 11,340 XBeach simulations are required. To expand the current system to be representative of one more general shape of coral reefs, say one with a lagoon, the number of parameters increases, drastically raising the number of XBeach simulations. For example, if two more parameters are included to define the lagoon, such as lagoon depth and lagoon width with 3 values for depth and 5 values for width, the number of XBeach simulations jumps to 170,100. Using a schematized profile, therefore, has severe limitations, and finding a way to efficiently represent a variety of reef morphology without drastically increasing the computational requirements for each new shape of coral reef would be extremely advantageous.

Now, data from the United States Geological Survey (USGS) are available with over 30,000 measured reef profiles from the East coast of the US as well as the US Pacific Islands. A dataset such as this has never been available before, and so to my knowledge, this will be the first reef classification study based on a robust dataset of this size. The challenge is to determine if there is a method that can effectively reduce the dataset to representative reef profiles that can be used in the XBeach model or BEWARE which will result in a more accurate and applicable wave runup prediction.

1.3. SCOPE AND RESEARCH QUESTIONS

1.3.1. SCOPE

The goal of this study is to develop a methodology to reduce a large dataset of reef profiles to a number of representative reef profiles that can accurately predict the wave runup and flooding for the full dataset, and other reef profiles from around the globe. The methodology could lead to many applications, but perhaps most significant would be a global early warning flood system to predict wave runup under multiple climate change scenarios, or be used towards other types of morphology. Essentially, if thousands of XBeach model runs were done with varying hydrodynamic loading conditions on the reduced dataset, it could be used as new data to input into an updated BEWARE model and lead to quick and accurate wave runup estimations for reef-fronted coasts around the globe.

1.3.2. RESEARCH QUESTIONS

This thesis aims to use input reduction and data mining techniques to generate a limited set of reef-profiles representative in terms of morphology and hydrodynamics of the entire dataset. In doing so, we aim to develop a methodology to effectively group coral reef cross-shore profiles. The goals can be broken down into the following research questions.

1. How can a large dataset of coral reef profiles be clustered such that the hydrodynamic response of grouped profiles is similar, and how should the cluster groups be represented?
2. What aspects of the reef profile are most important to consider for effective clustering in terms of wave runup?
3. What is the best approach to utilize the cluster profiles in order to predict wave runup of a natural reef profile?
4. How accurately can the selected cluster profiles predict wave runup and flooding over natural coral reefs?

1.3.3. RESEARCH APPROACH

To answer the aforementioned research questions, a set of objectives have been established. Points 1 and 2 are used to answer research question 1, and points 3 to 5 are used to answer the remaining questions in order.

1. Perform multiple methods of cluster analysis based on morphology to compare and deduce the one that groups profiles with the least morphological variance.
2. Establish a method to group the profiles based on runup values to reduce the dataset even further.
3. Compare the hydrodynamic response of profiles within the same cluster group to determine which features of similarly shaped profiles cause differences in wave runup.
4. Compare and evaluate methods to match a reef profile to the cluster profiles.
5. Compare the wave runup of random reef profiles and the matched cluster profiles to quantify the accuracy of wave runup prediction.

1.4. THESIS OUTLINE

Chapter 2 provides background information that is relevant for predicting runup on coral reef islands as well as the data reduction techniques used to reduce the dataset. Chapter 3 sets out the methodology and approach that was followed for the study, and Chapter 4 presents the corresponding results. In Chapter 5, the outcomes are discussed, as well as potential limitations and ideas for future research on this topic. Lastly, in Chapter 6, the findings of the report are summarized.

2

BACKGROUND

CHAPTER SUMMARY
In this chapter, relevant background information is provided, ranging from coral atoll island development, previous studies and findings of wave hydrodynamics over coral reefs, and details of the data reduction tools used in this study. Cluster analysis algorithms were the main tool used, and so they are elaborated on further, also touching on how cluster analysis has been used in the field of coastal engineering already.

2.1. CORAL ATOLL ISLANDS

2.1.1. CORAL ISLAND DEVELOPMENT

The first theories regarding coral reef development came from Charles Darwin in 1842 (subsidence model) (Ashton, Toomey, & Perron, 2013). He proposed that reefs go through stages of growth and subsidence, beginning as fringing reefs along the edges of a newly formed volcanic island. As the island sank due to land subsidence, a lagoon would form between the outer edge of the reef and the island, forming a barrier reef. Eventually, the island would completely subside below sea level, leaving an atoll (Darwin, 1832). This process is shown in Figure 2.1. Drill cores from the Eniwetok Atoll in the Pacific Ocean (Ladd, Ingerson, Townsend, Russell, & Stephenson, 1953) supported Darwin's theory and suggested that land subsidence plays a pivotal role in reef development, however more recent surveys suggest that the theory alone can not explain all of the reef forms that we see today. Work has been done to develop models to understand reef formations (Ashton et al., 2013), with the discovery that there are multiple factors, including coral growth, wave erosion, uplift and subsidence that lead to the variability observed in natural reefs around the globe.

Sandy beaches, on the other hand, can be characterized by an equilibrium profile, originally proposed by Bruun (1954) and later by Dean and Galvin Jr (1976). Sandy beaches reshape according to their forcing. The equilibrium profile of sandy beaches is a simplistic approach that can represent the shape of the coastline or its response to forcing through simple equations.

The stark differences between coral reef profiles and sandy beaches is what adds to the complexity of this study. Finding a method to generate a reduced subset of coral reef profiles that effectively represent the entire dataset when the natural variability of reef profiles is so high poses a great challenge.

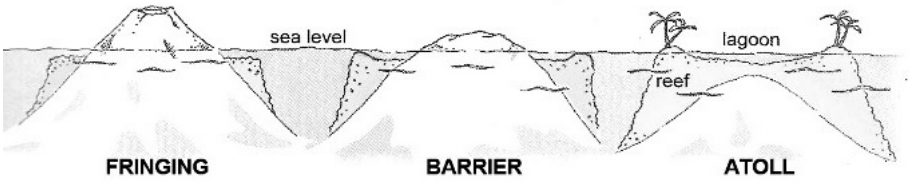


Figure 2.1: Coral reef formation due to land subsidence, showing the development pattern from fringing reef, to barrier reef to atoll. Source: (Tubbataha Reefs Natural Park, 2018)

The most at risk reef-fronted coasts are coral atolls. Coral atolls are defined as ring-shaped reefs surrounding a central lagoon. Atolls vary considerably in size and shape, and form islands that may be present along the entire rim of the atoll or in only few locations (Woodroffe, 2008).

2.1.2. LOCATION OF CORAL ATOLLS

Coral atolls are typically found in the mid-plate settings in the Pacific and Indian oceans (Woodroffe, 2008). A figure depicting the locations of the different types of coral reefs in the Pacific Ocean is shown in Figure 2.2, where atolls are shown in red. As noted in the figure, there are many throughout the Pacific Ocean.

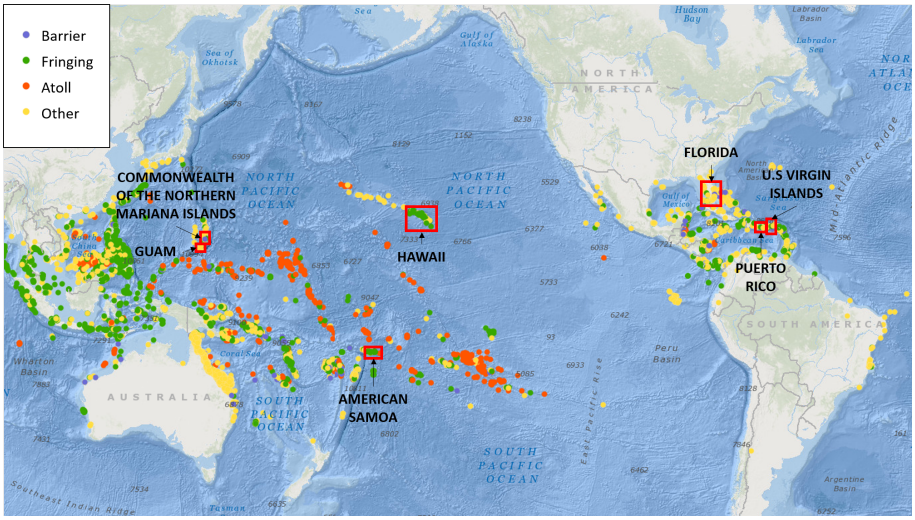


Figure 2.2: Global coral reef locations, highlighting locations of barrier, fringing and atoll reefs. Boxed in red are the US island territories from which the data of coral reef profiles for this study were measured. Adapted from (ReefBase, 2019; Storlazzi et al., 2019)

2.1.3. REEF PROFILE DATASET

The dataset used for this study is comprised of measured reef profiles provided by the US Geological Survey (USGS). The measurements come from seven different US regions in the Pacific Ocean and Caribbean. A map of US island territories with red boxes showing where the data originates can be seen in Figure 2.2.

The cross-shore transects were created by the USGS, spaced every 100 m alongshore using the Digital Shoreline Analysis System (DSAS) in ArcGIS. The transects were made in both the landward and seaward directions using the smoothed baseline cast method with a 500 m smoothing distance, applied perpendicular to the coastline. All transects vary in length in order to reach the -30 and +20 m elevation contours (Storlazzi et al., 2019). Further details of the dataset are provided in Appendix A.

The dataset is formed as a compilation of measurements from multiple sources ranging in date from 2001 to 2016 (Storlazzi et al., 2019). Since the data are gathered from different sources, it is expected that the accuracy of the measurements is variable. The varying accuracy in measurements required data pre-processing explained in Section 3.1. The number of profiles from each location used in this study is shown in Table 2.1 .

Table 2.1: Number of profiles from each measurement location included in the USGS dataset used for this study

Location	Number of Profiles
American Samoa	1198
Saipan, Tinian	1035
Guam	1295
US Virgin Islands	1664
Hawaii	13404
Puerto Rico	5531
Florida	6039
Total	30166

2.1.4. EFFECTIVENESS OF CORAL REEFS FOR COASTAL RISK REDUCTION

The coral reef surrounding atolls provide protection from the impact of large waves and storm surge (Ferrario et al., 2014). Reefs lead to energy dissipation through wave breaking and bottom friction (Lowe et al., 2005). Figure 2.3 illustrates this process, as shown by the significant decrease in wave heights over the reef flat compared to the offshore conditions.

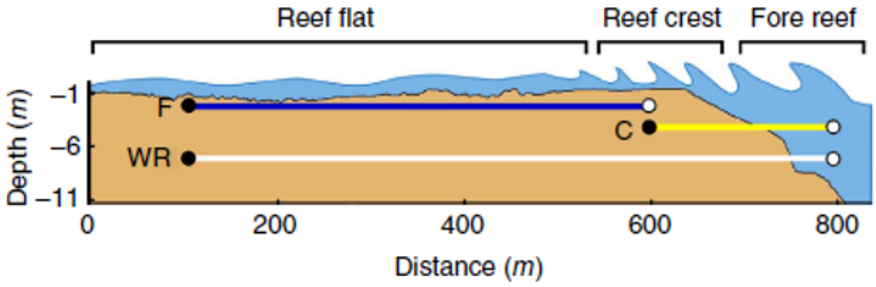


Figure 2.3: Illustration of wave height reduction at a coral reef due to wave breaking and frictional dissipation. *F* represents the reef flat, *C* represents reef crest, and *WR* represents the whole reef. Source: (Ferrario et al., 2014)

Meta-analyses have shown that coral reefs provide significant protection against natural hazards, reducing wave energy by an average of 97% (Ferrario et al., 2014). They actually perform better at wave attenuation than human designed structures such as submerged breakwaters. This natural protection serves roughly 100 million or more people (Ferrario et al., 2014). Based on these numbers, it is clear how important coral reef conservation and restoration is, and how coral reef restoration could be a cost-effective adaptation strategy.

Unfortunately, coastal risk is increasing with higher wave energy reaching the coast due to a combination of an increase in storms and coral bleaching, reducing reef wave attenuation effects (Sheppard, Dixon, Gourlay, Sheppard, & Payet, 2005).

2.2. REEF HYDRODYNAMICS

The processes that waves experience over a reef are different compared to the more well known and researched sandy beaches. Reefs experience complex bathymetry, high energy losses from breaking, higher frictional losses, and the highly non-linear behavior of impinging waves (Hardy & Young, 1996; Lee & Black, 1978; Young, 1989). The majority of reefs are depth limited, and therefore offshore water levels (tidal range) are extremely important for the ability of waves to propagate across the reef (Hardy & Young, 1996). Reef flat water levels are also dependent upon wave-induced setup, caused by offshore swell breaking over the shallow reef topography and resulting in a change in radiation stress, leading to a water level increase (Longuet-Higgins & Stewart, 1964).

The main processes associated with the transformation of offshore waves to runup at the beach are explained in this section.

2.2.1. WAVE BREAKING

As waves travel from offshore into the nearshore, they first experience shoaling, in which the waves begin to slow and increase in height, and then wave breaking when the depth reaches a certain ratio compared to the wave height. The morphology of a reef greatly influences how the waves will break and transform over it, and consequently the wave setup and wave induced flows across it (Gourlay, 1996a). Figure 2.4 demonstrates the reduction in wave height across a reef profile from waves breaking as the depth quickly decreases at the reef edge.

Tests done by Vetter et al. (2010) measured the extent to which waves are depth limited across the reef at Ipan, Guam, using the similarity parameter γ_i, where

$$\gamma_i = \frac{H_i}{h_i}$$

where H_i represents wave height and h_i is water depth. The tests found that the similarity parameter varied across the reef, yielding a value of γ_i=0.13±0.02 at the inner-reef sensor and γ_i=0.22±0.01 at the mid-reef. At the reef crest, there was a much higher value of γ_i=0.96±0.04. The difference in the values is due to the waves breaking at the reef crest and dissipation of the wave due to friction (Lowe et al., 2005; Vetter et al., 2010).

The parameter γ_b represents the similarity parameter at the location of breaking. Estimates of γ_b range from 0.91 for regular wave conditions and 1.13 for larger wave events. The differences are also most likely due to the difference in location of breaking on the steep reef face.

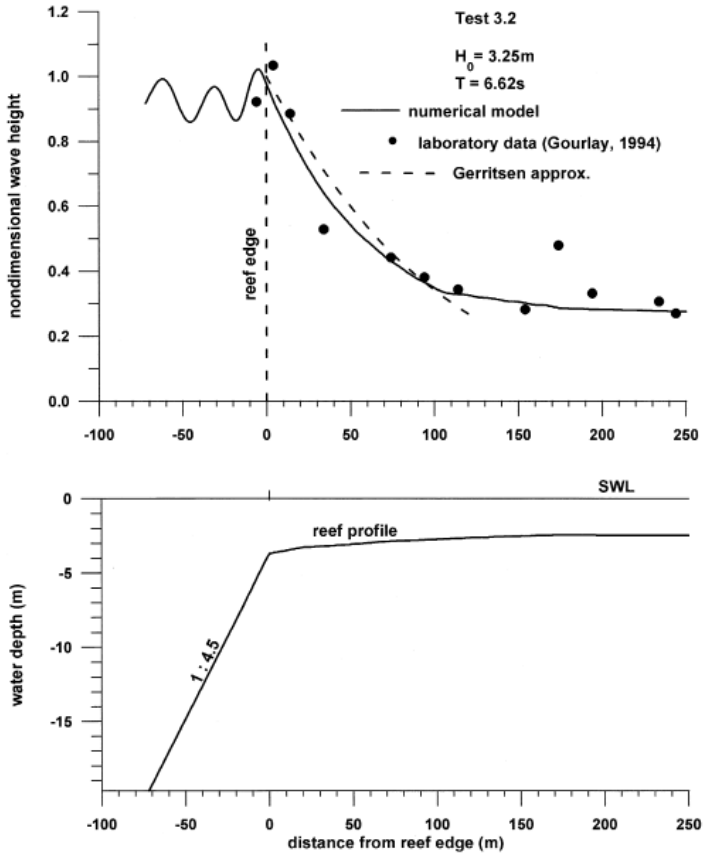


Figure 2.4: Model and experimental data from Massel and Gourlay (2000) of wave breaking at a coral reef. Source:(Massel & Gourlay, 2000)

2.2.2. WAVE SETUP

Breaking waves result in a water level increase on top of a reef, known as setup. Setup translates to higher water levels at the shoreline. Gourlay (1996a) used laboratory experimental data to investigate the wave setup and currents caused from breaking waves over a reef. He used an idealized horizontal reef, and then further compared his laboratory experiments with observations from natural reefs in a second paper (Gourlay, 1996b). The findings of the experimental data demonstrated that wave setup increases when:

1. Offshore wave height increases
2. Water depth over the reef decreases (low tide)
3. The lagoon is closed versus open
4. Wave period increases up to a certain point, then it remains constant

Wave setup can be caused from waves breaking on the reef top or the reef edge. The figure below shows the changes in radiation stress and hydrostatic pressure force that would be typical for waves breaking on the reef top, demonstrated by Longuet-Higgins and Stewart (1962). Location 1 is the reef edge and location 2 is the end of the surf zone. Through the simplification of the radiation stress and pressure force equations, the setup can be described as:

$$\frac{n_h}{H_0} = (0.135 \pm 0.042)(1 - 0.16(\frac{h_r}{H_0})^2)$$

This is appropriate when the relative water depth to wave height ratio is greater than 1 ($\frac{h_r}{H_0} > 1.0$) (Gourlay, 1996a).

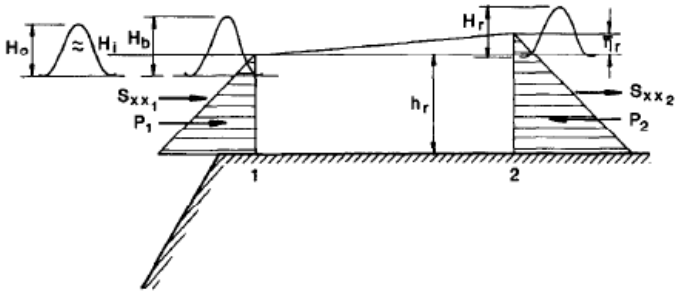


Fig. 12. Definitions for radiation stress theory for wave set-up on a reef.

Figure 2.5: Radiation stress diagram for wave setup on a reef. Source: (Gourlay, 1996a)

If the waves break at the reef edge, the situation can be represented by overtopping of a low breakwater with a very wide crest. The volume of water pumped onto the reef from the breaking wave uprush is equal to the volume of water discharged over the reef (Gourlay, 1996a). Gourlay (1996a) was able to prove that when waves are breaking on the reef-face, the hydraulics of the water discharging from the reef-top is critical for determining the wave setup and therefore the setup is likely to be significantly influenced by the reef morphology.

Essentially, with relatively large submergence, waves broke on the reef top and setup could be estimated theoretically from formulae based on radiation stress theory. Setup for this case was relatively small. With relatively low submergence, waves break on the reef face, essentially resulting in overtopping of the reef-edge, which could be estimated from continuity. This case produced relatively large setup.

From the Gourlay (1996b) experiments with various reef profiles, all tests show that setup only happens when offshore wave height is of the order of 0.4 h_r, the initial depth over the reef flat. Main conclusions were that wave setup is subject to the shape of the reef profile, particularly the slope of the reef rim, and the relative elevations of the reef-edge and reef-crest due to their effect on the amount of energy dissipated by the waves breaking on the reef-rim. Observations from islands in the Pacific by Vetter et al. (2010) showed that reef flat setup is highly correlated with incident wave height and consistent with the Longuet-Higgins and Stewart (1962) setup balance for localized breaking.

2.2.3. INFRAGRAVITY WAVES

Infragravity (IG) waves (also called low-frequency waves) are those with periods of 25 seconds to tens of minutes. They have been studied thoroughly on sandy beaches, but only more recently over reefs (Pomeroy et al., 2012). Although few studies have been done, IG waves have been recognized as being important in reef hydrodynamis for many years, mainly because they may make an important contribution to the water motion within the reef-lagoon systems (Hardy & Young, 1996; Lugo-Fernández, Roberts, Wiseman Jr, & Carter, 1998; Pomeroy et al., 2012).

IG waves can be formed from two mechanisms, known as “shoaling bound waves” (Longuet-Higgins & Stewart, 1962) and “breakpoint generated waves” (Symonds, Huntley, Symonds, & Bowen, 1982). The first mechanism is linked to the presence of short-wave groups which themselves are formed due to the superposition of two short-wave trains with similar wave length and frequency. The bound long waves are formed by the nonlinear interactions between the short-wave groups (Longuet-Higgins & Stewart, 1962). They travel from deep water to shallow water and are amplified as they enter the shoaling region. The second mechanism occurs within the surf zone by the time varying oscillation of the short-wave breakpoint (Symonds et al., 1982). Most studies show that, on mild sloping beaches, IG waves are mainly formed from shoaling bound waves; however when the relative slope at the breakpoint increases, the importance of surf zone generated waves significantly increases (Baldock, 2012).

Pomeroy et al. (2012) used field data and an IG wave-resolving numerical model (XBeach) to investigate how IG waves are formed and behave across reefs. They saw that there was a large reduction in short wave heights (and short-wave energy) going from deep water across the reef flat. IG waves were found across the reef, and it was noticed that IG wave heights were considerably smaller than the corresponding short wave heights at the fore reef, but that they picked up slightly at the reef crest, and then gradually dissipated across the reef flat at a much slower rate than the short waves did. The differences in dissipation rate results in the IG waves becoming increasingly important across the reef until they eventually dominate. Figure 2.6 demonstrates this, with the red line representing the IG wave height becoming greater than the short-wave height shown in blue at about x = 480 m. From then onwards towards shore, the IG wave main-

tains dominance over the short-wave.

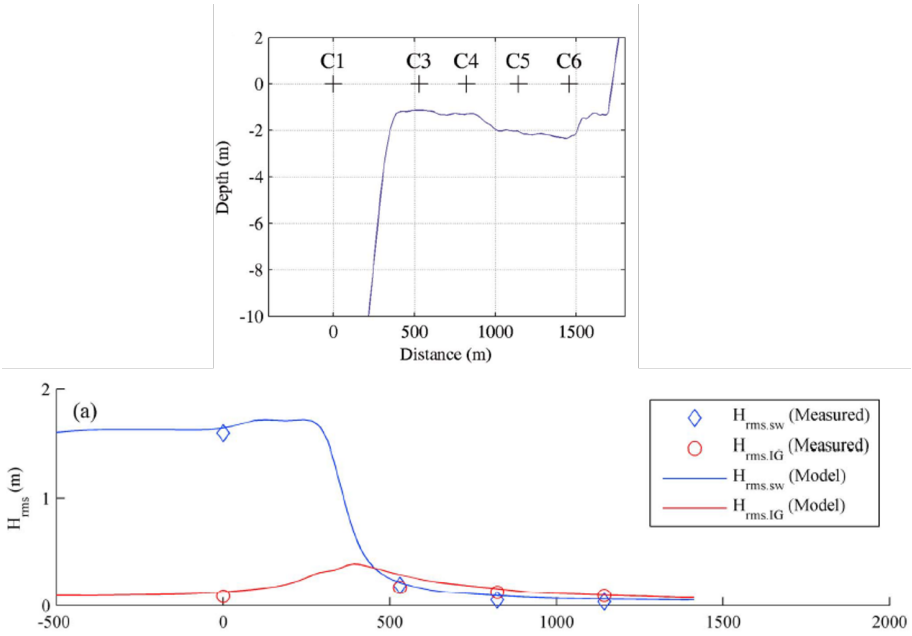


Figure 2.6: Comparison of the observed and modelled short-wave and IG wave heights done in the Pomeroy, Lowe, Symonds, van Dongeren, and Moore (2012) experiments. The red represents the IG wave and the blue represents the short-wave. Source: (Pomeroy, Lowe, Symonds, van Dongeren, & Moore, 2012)

Pomeroy et al. (2012) found that the generation of IG waves occurred over the fore reef slope, leading to the conclusion that IG waves on a reef are formed by breakpoint forcing. This can also be seen in Figure 2.6 where the IG wave height significantly picks up at the location of shortwave breaking and dissipation.

Finally, measurements of reflection were larger at the lagoon than at the seaward reef station, suggesting that there is some shoreline reflection of IG wave energy, but that there is also shoreline dissipation, since the reflection energy ratio was much less than 1. The observation of shoreline dissipation is consistent with other research of IG waves at sandy beaches. The remaining IG waves that do get reflected decay due to bottom friction, such that almost no IG wave energy reaches the seaward edge of the reef.

IG wave heights are important for the estimation of runup/flooding since they lead to varied and increased water level at the shoreline. IG wave heights can be influenced on the reef by two main factors. First, the tide level plays a role in that an increase in the tidal depth creates an increase in the IG waves over the reef. Secondly, work from Péquignet, Becker, Merrifield, and Aucan (2009) shows that the amplitude of IG waves can be significantly enhanced during periods of resonance, when the time scale of the offshore forcing matches the resonant mode of the reef morphology. .

2.2.4. BOTTOM FRICTION

Bottom friction refers to the measure of the coral reef's resistance to flow. A reef with a high friction value will result in greater frictional dissipation and less energy reaching the shoreline. Since coral reefs are naturally variable, the friction varies spatially, depending mainly on the biological and morphological zonation of the reef (Gourlay, 1996b). Reefs typically contain spur and groove features, which are a series of ridges and channels that generally are aligned with the direction of the dominant waves (Hopley, 1982). The spur and groove structures contribute to varying and unknown amounts of wave energy dissipation processes (Gourlay, 1996b). The complexity of the shape, friction and porosity of the reef makes it very difficult to model.

The difficulty is that the bottom friction is naturally extremely variable, and in models, the bottom friction value can drastically alter the results (Pearson et al., 2017). In this study, a one-dimensional reef profile is used in the XBeach model, and therefore the along shore variability of the reef is not taken into consideration. Instead, a constant friction value is used to simplify the analysis. The value selected is a conservative one of 0.05, which is the medium value used in Pearson et al. (2017) study. Pearson et al. (2017) based the friction values used in their study on nine other studies of coral reefs.

2.3. WAVE RUNUP AND OVERTOPPING

Once waves reach the shoreline, they result in swash, wave runup and potential overtopping. These processes are explained below.

2.3.1. SWASH

Swash, S, is generally defined as the time-varying location where the ocean meets the beach (Stockdon, Holman, Howd, & Sallenger, 2006). Miche (1951) came up with the concept that monochromatic waves consist of two parts: a progressive component that is dissipated during wave breaking as they approach shallower depths, and a standing component that has its maximum at the shoreline due to reflection. Swash represents the standing component.

From studies on sandy beaches, it was determined that the values of swash are dependent upon both the beach slope and wave period (Stockdon et al., 2006). Empirical equations can then be developed for such systems, but are not possible for the highly variable reef profiles. The information, however, is still important and transferable to coral reef coasts.

The swash signal can be broken down into the infragravity and incident frequency bands. This provides more information about the processes responsible for the water level at the shoreline. Stockdon et al. (2006) found that infragravity waves dominate the contribution to swash, however (Gawehn, 2015) found that incident frequency swash is also important to include in reef modelling of hydrodynamics.

2.3.2. RUNUP

The main statistic used in this study to compare the different reef profiles is wave runup. Stockdon et al. (2006) defines runup as the set of discrete water-level elevation maxima. An example of a water elevation time-series is shown in Figure 2.7, where the runup

values are highlighted with an asterisk.

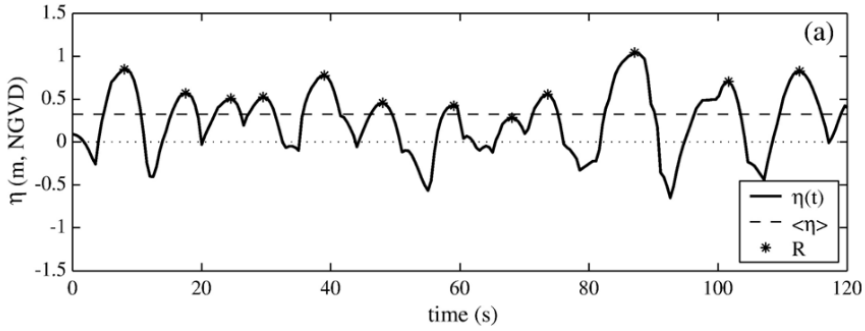


Figure 2.7: Example of a water level time-series, highlighting the runup values. Source: (Stockdon, Holman, Howd, & Sallenger, 2006)

Wave runup is a result of all of the processes occurring over the reef up until the waves finally reach the shoreline. The two main attributing factors are the maximum setup, $< n >$, and the swash. A common statistic for evaluating runup is the 2% exceedance value, known as $R_{2\%}$, calculated from the cumulative probability density function of runup elevations. This statistic is often used in engineering applications (Holman, 1986). Using data from previous studies, (Stockdon et al., 2006) developed an empirical equation for the elevation of $R_{2\%}$. In this study, the runup values used were calculated from the XBeach model output.

If the runup value is greater than the elevation of the land, a process called overtopping occurs, where water is able to flow over the land boundary. This may lead to flooding and severe damages, including property damage and erosion of the limited land on these small developing islands.

2.4. CLUSTER ANALYSIS

The goal of this study is to reduce the large dataset of reef profiles into a subset that are representative in terms of wave runup, using a technique called a cluster analysis. Cluster analysis is a method of grouping a collection of objects into smaller subsets or clusters. The goal is to have the objects within each group more closely related to one another than objects assigned to different clusters (Friedman, Hastie, & Tibshirani, 2001).

Cluster analysis with large databases can lead to severe computational requirements. Challenges associated with this led to the emergence of powerful broadly applicable data mining clustering methods (Berkhin, 2002). Common to all of them is the use of degree of similarity (or dissimilarity) between the objects.

2.4.1. OBJECT DISSIMILARITY

The clustering process revolves around the definition of similarity that is being applied. To define the similarity, a proximity matrix is used. The proximity matrix is an NxN matrix, where N is the number of objects, and each element of the matrix dii' records the

proximity between the ith and $i'th$ objects. For this study, N translates to the number of reef profiles, and the elements in the matrix are filled with the calculated difference between all profiles. An example of a proximity matrix is shown in Table 2.2.

Table 2.2: Example of a proximity matrix using cityblock distance with 10 reef profiles

	Cityblock Distance									
Profile	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	0.016	0.523	0.161	0.553	0.083	0.093	0.098	0.045	0.007
2	0.016	0	0.508	0.177	0.538	0.099	0.109	0.114	0.061	0.023
3	0.523	0.508	0	0.684	0.030	0.606	0.616	0.621	0.568	0.530
4	0.161	0.177	0.684	0	0.714	0.078	0.068	0.063	0.116	0.154
5	0.553	0.538	0.030	0.714	0	0.636	0.646	0.651	0.598	0.560
6	0.083	0.099	0.606	0.078	0.636	0	0.010	0.015	0.038	0.076
7	0.093	0.109	0.616	0.068	0.646	0.010	0	0.005	0.048	0.086
8	0.098	0.114	0.621	0.063	0.651	0.015	0.005	0	0.053	0.091
9	0.045	0.061	0.568	0.116	0.598	0.038	0.048	0.053	0	0.038
10	0.007	0.023	0.530	0.154	0.560	0.076	0.086	0.091	0.038	0

There are multiple methods to calculate the dissimilarity between observations. The selected method is subject to the conditions of the data (Friedman et al., 2001), as well as the specific conditions of the cluster analysis. In this study, squared-euclidean distance (SED) and cityblock distance were used for the hard partitioning algorithms, and mixture models were used for the probabilistic algorithms. Their details are explained in Appendix C.2.1.

2.4.2. CLUSTERING ALGORITHMS

The dissimilarity metrics are used in the clustering algorithms to separate the data into groups. In this study, five different clustering algorithms are tested and used to reduce the reef profile dataset. These are divided into partitioning relocation algorithms, probabilistic clustering, and hierarchical clustering. The methods are explained below.

KEY TERMS

For clarification, Table 2.3 provides the terms used to describe the cluster analysis algorithms and how they translate to the application of this study.

Table 2.3: Key cluster analysis terms and their associated meaning in this study

Cluster Analysis Terms	Definition	In This Study
Observation	Input into the algorithm	Reef profile
Variable	Contains the data of the observation	Cross-shore position & depth
Cluster	Group of observations	Grouped reef profiles
K	Number of cluster groups	Number of groups of reef profiles
Centroid	Represents the cluster group	Mean or median of profiles within the cluster
Inter-cluster	Between two clusters	-
Intra-cluster	Within a cluster	-

PARTITIONING RELOCATION CLUSTERING

Partitioning algorithms divide data into several subsets. Iterative optimization is used by means of different relocation schemes that iteratively reassign points between the clusters. The algorithms work to gradually improve the clusters until convergence, which results in high quality clusters (Berkhin, 2002). The iterative optimization partitioning algorithms are subdivided into K-medoids and K-means methods. An important feature for clustering algorithms is the initial centroid selection. The method used throughout this analysis was the K-means++ algorithm which is explained following the methods.

K-means

The K-means algorithm (J A Hartigan & Wong, 1979; John A Hartigan, 1975) is the most popular clustering tool used in scientific and industrial applications. The name is derived from representing each of the clusters by the mean (or weighted average) of its points, also called the centroid. It is intended to be used when all variables are of the quantitative type, and with squared-euclidean distance (SED) as the dissimilarity measure. Since the average of all points within the cluster is used to define the centroid, it can be negatively affected by outliers. An advantage of the method is the clear geometric and statistical meaning (Berkhin, 2002).

The steps in K-means clustering are as follows:

1. Initial cluster centers are selected. Multiple methods for the initial selection process are available, but K-means ++ is solely used throughout this thesis, as explained below.
2. Point to cluster centroid distances for each observation and each centroid are calculated using SED.
3. Each observation is assigned to the cluster with the closest centroid.
4. The average of the observations in each cluster is computed to determine the K new centroid locations.

5. Steps 2 to 4 are repeated until the cluster assignments do not change, or the maximum number of iterations is reached.

Figure 2.8 provides an example of the iterative process of K-means clustering. In the top left plot, initial centroids are selected, and in the top right plot, the first data partitioning is shown. The bottom left plot shows the improvement of the clustering after two iterations, and the bottom right plot shows the cluster groups at the end of twenty iterations. The lines separating the data are changed each iteration until convergence is reached.

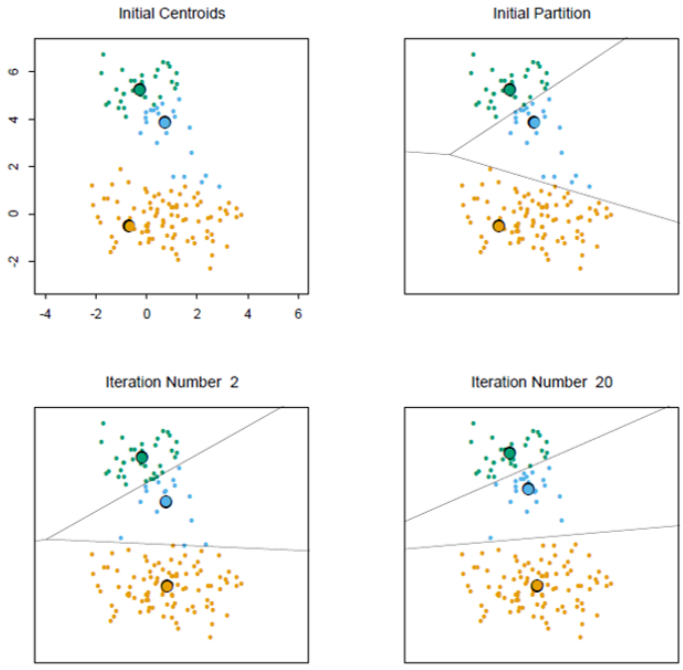


Figure 2.8: Illustrative example of K-means clustering. Source:(Friedman, Hastie, & Tibshirani, 2001)

***K*-medoids**

In K-medoids, the centroid is the most appropriate data point within the cluster. This method has two main advantages. First, it has no limitations based on attribute types, meaning that the data does not have to be quantitative. Second, since the medoid is an actual data point and is dictated by the location of a predominant fraction of the points inside the cluster, it is less sensitive to outliers (Berkhin, 2002). This can be seen in Figure 2.9

K-medoids is useful when the mean or median does not have a clear definition, since the medoid is an actual data point in the data set. Similar to K-means, the goal is to achieve the minimum sum of distances between the observations and centroid, which in this case is referred to as the medoid. There are multiple iterative algorithms that can be used to minimize the sum of distances from each object to its cluster medoid, but

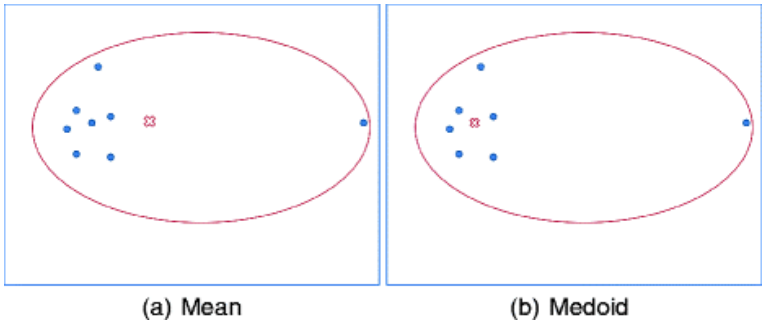


Figure 2.9: Illustrative example of how K-means clustering (a) differs from K-medoids (b).

the one used for this analysis is called partitioning around medoids (PAM) (Kaufman & Rousseeuw, 2009). The process follows two steps:

1. Build-step: Each K cluster is associated with a potential medoid. For this analysis this is initially done using K-means++.
2. Swap-step: Within each cluster, every point is tested as the medoid and the sum of within-cluster distances is calculated using the SED. If a point results in a lesser sum of distances, it is classified as the new medoid. Every point is then assigned to the cluster with the closest medoid.

These steps are repeated until the medoids are not changed.

***K*-medians**
The K-medians approach is a less common clustering technique that was achieved through modification of the K-means algorithm. The iterative process is the same as K-means, except for two main differences. First, the distance metric is changed to 'cityblock' (distance metrics explained in Appendix C.2.1), which calculates the absolute difference between each observation point and centroid point to determine the distance between observations. The second difference is that using cityblock distance results in the centroid being computed as the component-wise *median* of the points within that cluster. This means that the centroid definition is changed from the mean of the observations within the cluster to the median, which results in different outcomes.

***K*-means++**
Initial cluster centroids must be set to begin the iterative process for each clustering algorithm stated above. Normally, the clustering algorithms would select initial cluster centroids arbitrarily, leading to varying results. An improved selection algorithm known as K-means++ (Arthur & Vassilvitskii, 2007) has been proven to improve the running time and improve the quality of the final solution (Arthur & Vassilvitskii, 2007). If the number of clusters is K, K-means++ follows these steps:

1. Select an observation uniformly at random from the data set, X. This selected observation is the first centroid and is denoted c_1.

2. The distances from each observation to c_1 are computed. The distance between c_j and the observation m is denoted as $d(x_m, c_j)$.

3. The next centroid, c_2 is selected at random from X with probability

$$\frac{d^2(x_m, c_1)}{\sum_{j=1} nd^2(x_j, c_1)}$$

4. Choose the centre j by:
   a. Compute the distances from each observation to each centroid, assign each observation to its closest centroid.
   b. For $m = 1, ..., n$ and $p = 1, ..., j-1$, select centroid j at random from X with probability

$$\frac{d^2(x_m, c_p)}{\sum_{h; x_h \in C_p} d^2(x_h, c_p)}$$

   where C_p is the set of all observations closest to centroid c_p and x_m belongs to C_p. This means that each subsequent cluster centre is selected with a probability proportional to the distance from itself to the closest centre that is already selected.

5. Repeat step 4 until K centroids are selected.

PROBABILISTIC CLUSTERING

Unlike the clustering algorithms stated above which assign each observation to one cluster, in probabilistic clustering a distribution is fit to the data. This results in probabilities of each observation belonging to the clusters, and therefore the method is known as a soft-partitioning method, since each observation does not have to be fully part of one cluster.

Gaussian Mixture Model

In the Gaussian mixture model, parameters are fit using Gaussian distributions and the Expectation-Maximization (EM) algorithm. The Gaussian distribution (also known as the "normal distribution") is represented by the mean and standard deviation of the data. It is normalized so that the sum of all of the values of x gives a probability of 1.

An example of fitting Gaussian distributions to random data is seen in Figure 2.10. In subplot (a), one distribution is fit to the dataset and demonstrates that the distribution does poorly at representing the data. In subplot (b), two distributions are selected and the data is represented well within both. The EM algorithm works to find the combination of the set number of distributions that most accurately represents the entire dataset. Another example with scatter points in two-dimensions can be seen in Figure 2.11.

The EM algorithm is an iterative approach to determine the best fit of the distributions within the data. The algorithm proceeds with these steps:

1. In the expectation step, a soft assignment of each observation to each model is made. Posterior probabilities of cluster memberships are computed, which results in an nxK matrix, where element $(i\,j)$ contains the posterior probability that

observation i is from cluster j. The probabilities assigned are based on the location of the data within the Gaussian distribution. Observations close to the centre of a cluster will most likely get a probability near 1 for that cluster and near 0 for every other cluster. Observations between clusters will divide their probability accordingly.

2. The cluster-membership posterior probabilities are used as weights, and the algorithm estimates the cluster means, covariance, matrices, and mixing proportions by applying maximum likelihood.

These steps are iterated until convergence is reached. The mixture model also provides an estimate of the probability that an observation belongs to a component which can be useful for other purposes.

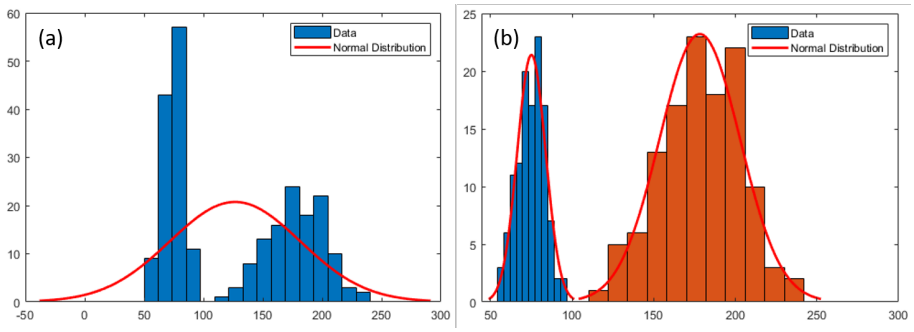


Figure 2.10: Example of 2 Gaussian distributions fit to random data.

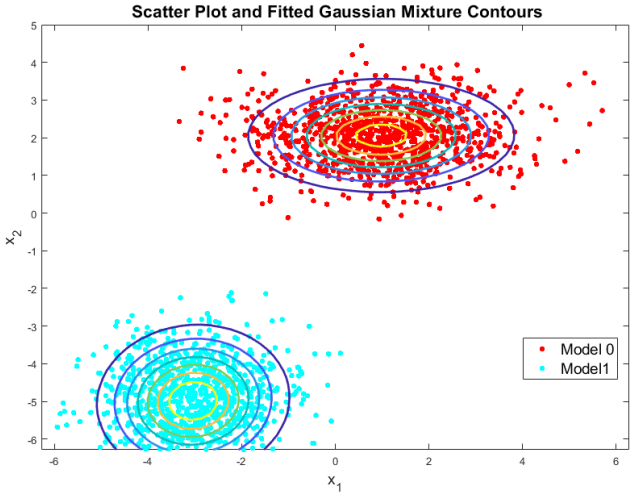


Figure 2.11: Illustrative example of a Gaussian Mixture Model with 2 clusters.

HIERARCHICAL CLUSTERING

In contrast to the partitioning methods such as K-means, as well as the probabilistic methods, hierarchical clustering does not require the user to specify a certain amount of clusters. Alternatively, they require the user to set a measure of dissimilarity between (disjoint) groups of observations, calculated using the pairwise dissimilarities among the observations in the cluster groups (Friedman et al., 2001). This method results in a hierarchical representation in which groups at each level of the hierarchy are formed by merging the groups at the level below. The lowest level of the hierarchy includes all of the observations treated as individual clusters, and at the highest level all observations are grouped together into one cluster (Friedman et al., 2001).

There are two distinct methods used in hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Agglomerative strategies begin at the bottom and at each level merge a selected pair of clusters into a single cluster. The pair chosen for merging consist of the two groups with the smallest intergroup dissimilarity. Figure 2.12 provides an example of agglomerative hierarchical clustering. Divisive strategies begin at the top with all observations grouped together, and at each level split one of the existing clusters into two new clusters. The clusters that are split are selected which have the largest between-group dissimilarity.

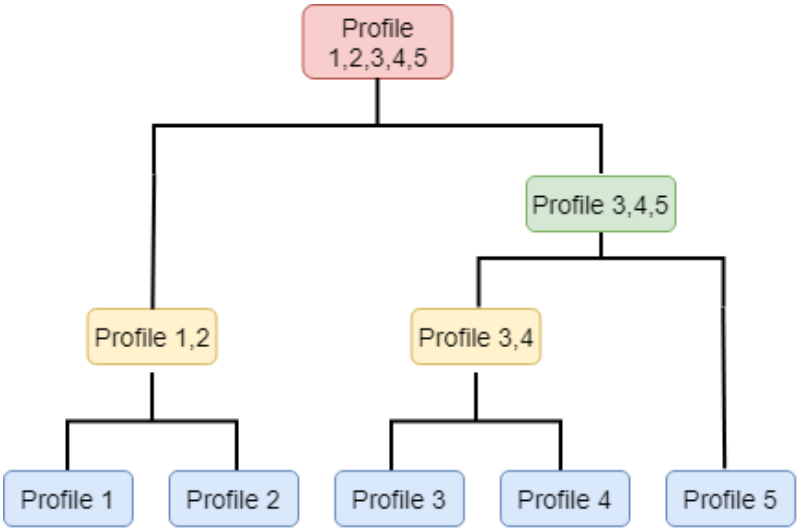


Figure 2.12: An example of how the agglomerative hierarchical clustering method groups profiles. To begin, all profiles are individual clusters. In the first step, profile 1 and 2, as well as 3 and 4 get grouped together. In the second step, the cluster with profiles 3 and 4 grousp with profile 5. In the last step, all profiles are grouped together to one cluster.

The method to calculate dissimilarity must be specified for the hierarchical clustering, and different methods can change the results drastically. There are several different measurement methods, including single, Ward, median, weighted, average, centroid, and complete. The details of these measurement methods can be found in Appendix C.3.1.

Hierarchical clustering is best represented through a dendrogram, shown in Figure 2.13. The links represent the merger of clusters, and the height corresponds to the distance between the merged clusters.

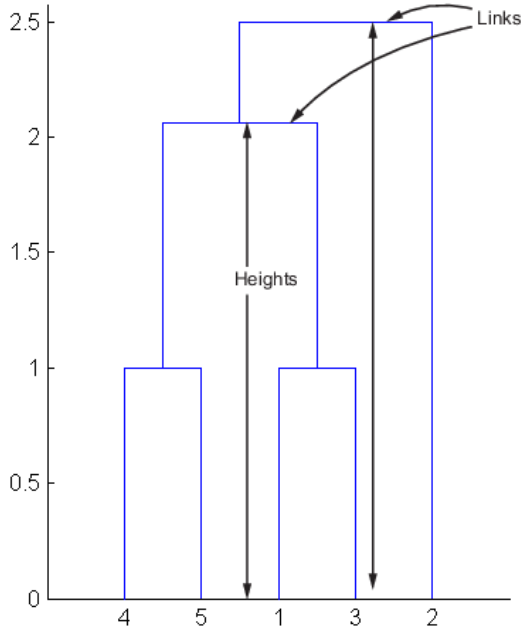


Figure 2.13: Example of a dendrogram, representing the hierarchical clustering process. Source: (MathWorks, 2019)

Cutoff Values

In hierarchical clustering, there are two ways to select how many final cluster groups are output from the method. First, one can set a limit in the height of the dendrogram (height represents intergroup dissimilarity), and the algorithm will finish once enough steps are complete in the hierarchy to reach the designated value. A second method, which is more well suited for finding natural divisions in the dataset, is the use of a cutoff value, which represents a limit of the differences of the merging cluster groups in comparison with previously formed clusters. The cutoff value in this study is specifically the inconsistency coefficient (IC).

The IC is a measure at each link of the hierarchical cluster tree that compares the height (difference within clusters) with the average height of other links up to two levels of the hierarchy below.

The inconsistency coefficient can be described mathematically as:

$$IC = \frac{D - \bar{D}}{\sigma}$$

in which IC is the inconsistency coefficient, D is the distance between the clusters being

merged, $\bar{D}$ is the mean distance of merged clusters included in the calculation, and σ is the standard deviation of distances included in the calculation.

The cutoff value limits how different the cluster groups can be in comparison to those already merged. A large IC means that the difference between the clusters being merged is high in comparison to the mean of previously merged clusters. A small IC means that the the difference between the clusters being merged is low in comparison to the mean, and the cluster groups are rather similar. Using the cutoff value, the user can restrict the relative dissimilarity between clusters. The lower the cutoff value, the more restrictive the method is and fewer links can be made, resulting in more final cluster groups. A higher cutoff value will be less restrictive and allow more links to be made, resulting in fewer final cluster groups.

MAXIMUM DISSIMILARITY ALGORITHM

Another tool used to reduce the dataset in this study is the Maximum Dissimilarity Algorithm (MDA). MDA, first developed by Kennard and Stone (1969), works to represent the full set of data by selecting the data points that are most dissimilar. It works very differently from the cluster-based and partition-based approaches previously presented, as it does not try to form subsets by grouping similar data, but rather defines subsets by selecting actual data points that are most different to the rest (Lajiness, 1997). The algorithm will finish with a subset that ideally represents the full range within the dataset. Willett (1999) describes variants of the MDA used in the field of combinatorial chemistry for selecting structurally diverse sets of compounds in chemical databases.

The basic maximum-dissimilarity algorithm is used to select a size-n subset from a size-N dataset. Similar to the other methods, there are different options for the choice of the initial observation, as well as the measure of dissimilarity. Each different definition of dissimilarity will result in a different version of the algorithm and a different final subset (Holliday & Willett, 1996). It follows an iterative approach, in which each iteration identifies the most dissimilar observation compared to the earlier selected observations. The final subset consists of a reduced number of data points that cover a wide range of the initial dataset.

An example of the MDA is shown in Figure 2.14. The numbers represent the iterations of the algorithm, the arrows show the distance to the furthest data point, and the circles show the selected data point from that iteration. The colours of the arrows match the colours of the circle showing how the next data point was selected.

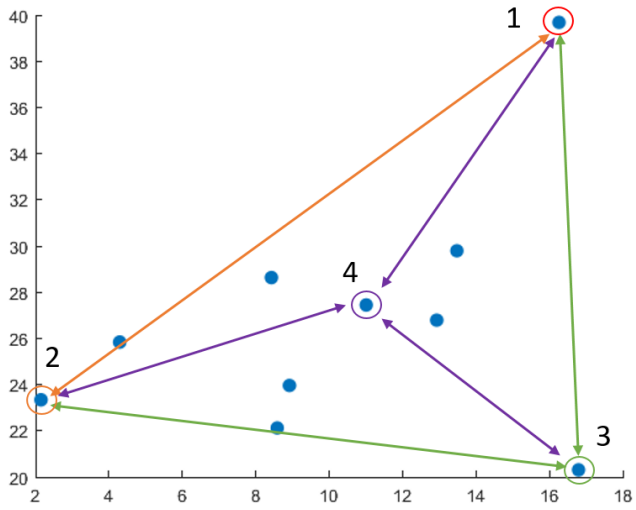


Figure 2.14: Illustrative example of the Maximum Dissimilarity Algorithm (MDA)

NUMBER OF REPLICATES

The final result of a clustering process is highly dependent upon the initial placement of cluster centroids. A clustering algorithm can be repeated multiple times with the same input conditions (different initial centroids since they are usually random or selected by another algorithm) and generate different results. To obtain the best possible result, one would repeat the clustering algorithm as many times as possible. This process is referred to as using replicates. Using a high number of replicates is beneficial in order to pick the best result from a greater number of simulations. Beginning the cluster analysis with a method such as K-means++ will reduce the effect of using a high number of replicates, as it initializes strategically, however there will still be a difference in results between the different simulations.

2.4.3. CLUSTER EVALUATION METHODS

Clustering a large dataset of coral reef profiles has rarely been done, and so one of the goals is to determine which clustering method works best for a dataset of this nature. The assessment of the effectiveness of the cluster analysis can be done via multiple methods. Direct evaluation proved to be the most successful for this study and is explained below. Other common techniques were also applied and are presented in detail in Appendix C.2.2.

DIRECT EVALUATION METHODS

Direct evaluation refers to assessing the internal quality of cluster groups in the application of interest (Manning, Raghavan, & Schütze, 2010). This was the main method used to compare clustering results in this study. In Cluster Round 1 the direct evaluation fo-

cused on the similarities in morphology within cluster groups, and in Cluster Round 2, the clusters were evaluated based on intra-cluster similarity in wave runup.

Direct evaluation is an effective way to quickly analyze the results based on the main requirement of the analysis, however, it lacks the inter-cluster assessment. For example, there could be two cluster groups with high intra-cluster similarity, but they themselves may be similar and should be merged. This comparison between cluster groups is not specifically analyzed and provides room for improvement in a further study.

2.5. CLUSTER ANALYSIS APPLICATIONS

Cluster analysis is a technique that is used often in many disciplines, and is beginning to be used more in engineering as more data and data mining techniques become available (Pham & Afify, 2007). In coastal engineering, few papers have been published that make use of cluster techniques. Those that have been published provide insight into how these techniques may work in coastal applications.

COASTAL ENGINEERING

Tomás et al. (2016) completed a study using clustering techniques to establish flood hazard and risk maps for the Spanish coast. Essentially, the study used a combination of extreme wave events and cross-shore profiles in a classification algorithm to establish representative cross-shore profiles. Tomás et al. (2016) obtained 30,000 profiles and 183 storms to include in their study. To reduce the number of modelling simulations, they applied clustering algorithms to the sea states and cross-shore profiles, using dimensionless variables to be able to relate the geometrical characteristics with the waves and sea level values. K-means was used to reduce the dataset to roughly 100 representative profiles of all of the representative profiles and storms along the Spanish coast.

The work done by Tomás et al. (2016) is the most similar to the goals of this study, but still contains many differences. The main difference is that the profiles used in that study are sandy beaches. For sandy beaches, empirical equations can be used to normalize the profiles, however this is not possible for reefs, due to their high variability. As a result, the morphology and hydrodynamics are treated separately in this study. Though the methodology is not the same, the desired outcome is similar and so the work done by Tomás et al. (2016) is useful to follow.

Costa, Araújo, Araújo, and Siegle (2016) used clustering algorithms to form characteristic reef profiles of Brazilian coast. Using 180 reef profiles spanning 18 kilometers of coastline, hierarchical cluster analysis was used to separate the profiles into two main groups with five subgroups. Costa et al. (2016) found that the groups were linked by location (north and south of the study area). This work is useful since it includes a very similar goal to this study, with the biggest difference being the size of the dataset. With only 180 reef profiles, the hierarchical method seemed to work well, but with roughly 30,000 profiles, it may not work as well. Hierarchical clustering requires high space and time complexity, and therefore is best not used for a very large dataset (Reddy, 2018). For this study, the hierarchical method is used once an initial input reduction is complete.

Lastly, Camus, Mendez, Medina, and Cofiño (2011) performed multiple clustering techniques to reduce trivariate time series of met-ocean parameters, including significant wave height, mean period, and mean wave direction. Their study was focused on

the ocean parameters and not the morphology. However, since their study compared different techniques, the pros and cons of the techniques in a coastal setting can be withdrawn and used in this study. They found that the k-means algorithm provided the most accurate clustering results of the three different algorithms tested, proving that the hard partitioning methods are capable of effectively working in high dimensional space. The MDA algorithm provided good performance in defining the boundaries of the data space.

2.6. XBEACH NON-HYDROSTATIC

XBeach is a coastal model developed by (Roelvink, 2009) to model nearshore processes. The model includes the hydrodynamic processes of short wave transformation, long wave transformation, wave-induced setup and unsteady currents, and overwash and inundation (Roelvink, 2009). The original application for the model was for assessing hurricane impacts on sandy beaches. Since then, the model has been extended to coral fringing and atoll reefs, mainly with the cooperation and funding from the University of Western Australia, USGS and the Asian Development Bank (Dano Roelvink et al., 2015).

An updated model incorporating a mode called non-hydrostatic plus (XBeach-nh+) (Smit et al., 2014) was used for this study. XBeach-nh+ uses a reduced two-layer approach to resolve intra-wave surface elevation and flow that provides many benefits. First, it allows the model to be used in deeper water since the dispersion relation is accurately modelled up to a kh value of 5, compared to the XBeach-nh mode which starts to lose accuracy past a kh value of 1 (De Ridder, 2018). Second, compared to the surf-beat mode, non-hydrostatic is a more complete model as it solves all processes, including short wave motions, long wave motions, currents and morphological change (Dano Roelvink, McCall, Mehvar, Nederhoff, & Dastgheib, 2018).

The depth-averaged flow due to waves and currents is computed using non-linear shallow water equations, as well as a non-hydrostatic pressure. Through manipulations of the dynamic pressure, the long waves obtain a dispersive behavior and the model can be used as a short-wave resolving model. Wave breaking is included by setting a maximum steepness value at which once it is passed, the non-hydrostatic pressure term is disabled and shallow water equations take over. This leads to one of the main advantages of the non-hydrostatic mode, which is that short-wave runup and overwashing are included.

For this study, it is important to keep in mind that XBeach-nh+ results in a spurious wave energy of the sub harmonics of about 5% (De Ridder, 2018). This means that even when the same hydrodynamic boundary conditions are input into the model for different simulations, there will be roughly 5% difference in conditions, making the comparison of simulations impossible to be 100% direct. This difference must be kept in mind when grouping profiles based on the hydrodynamic results.

3

METHODOLOGY

CHAPTER SUMMARY
The detailed methodology used to reduce the large reef profile dataset to a subset of representative profiles is explained in this chapter. The process follows two main steps of data reduction. First, the profiles are grouped based on similar shape. The median of the grouped profiles forms the 'cluster profiles'. Second, the cluster profiles are grouped based on similar shape and wave runup. The chapter also includes details of how the cluster profiles can be used for predicting wave runup of natural coral reefs.

The steps taken throughout this analysis can be summarized as:

1. Input reduction, using cluster analysis of reef morphology.
2. XBeach model simulations using the cluster profiles created in step 1.
3. Cluster analysis of reef hydrodynamics incorporating the wave runup results from the XBeach simulations.
4. XBeach model simulations using a set of random test profiles from the dataset.
5. Testing of the applicability of the method by matching test profiles to the cluster profiles and comparing the XBeach runup results.

As noted, there are two main steps to reduce the dataset. The first is done to create a reduced set of representative profiles that can then be simulated in XBeach. There are far too many profiles for all to be simulated in XBeach, and so clustering the profiles into groups of similar morphology saves a tremendous amount of computational time. The second data reduction then uses the XBeach results to group profiles based on hydrodynamics. The input reduction of morphology is labeled as 'Cluster Round 1', and the cluster analysis of the hydrodynamics is labeled as 'Cluster Round 2'.

The process is shown in Figure 3.1. The details of each aspect are explained throughout the rest of this chapter.

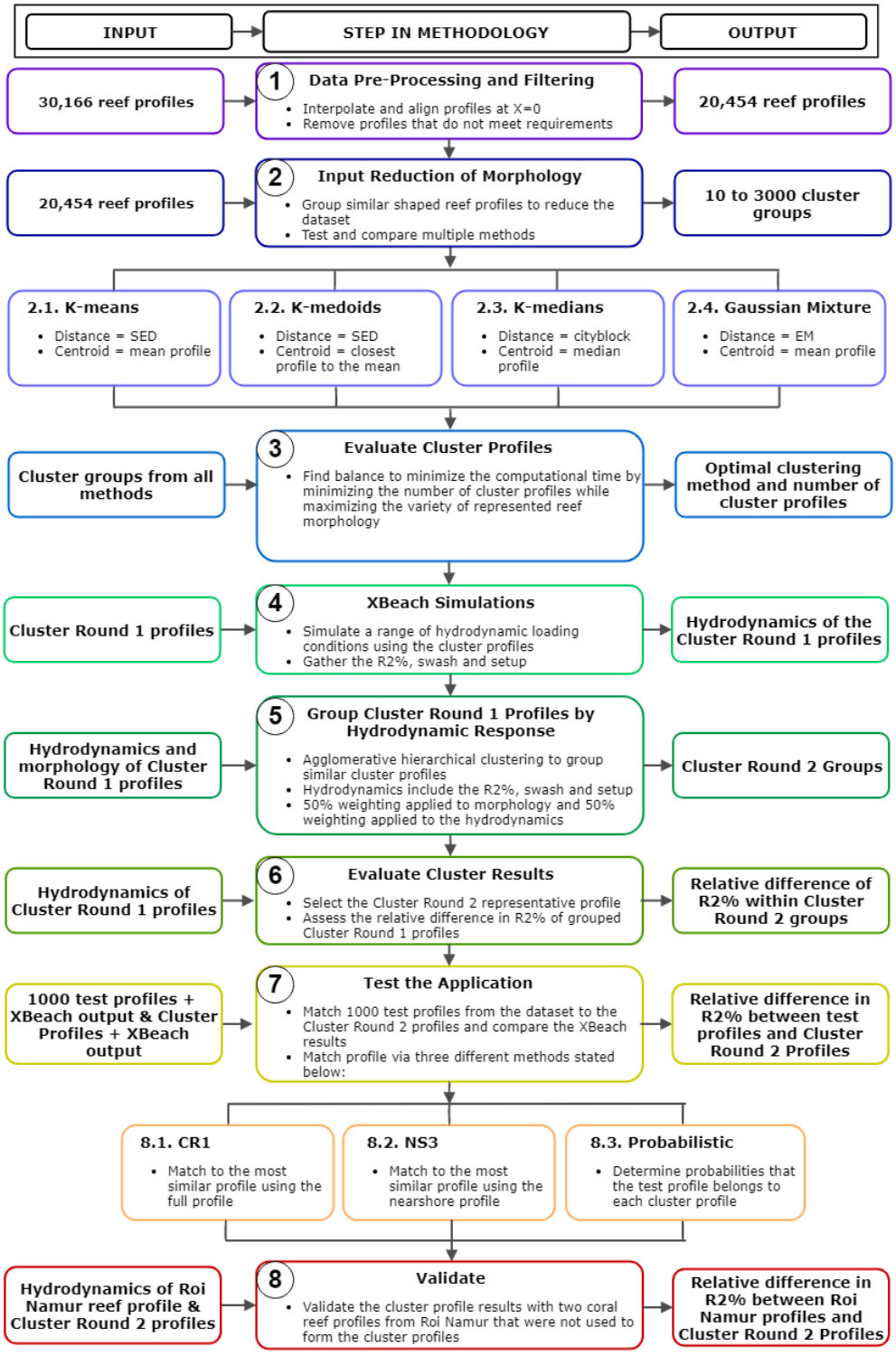


Figure 3.1: Detailed methodology including all steps of the analysis and the inputs and outputs of each step.

3.1. DATA PRE-PROCESSING

The data that was provided from the USGS contains 30,166 measured reef profiles, taken from American Samoa, Saipan and Tinian, Guam, the US Virgin Islands (USVI), Hawaii, Puerto Rico, and Florida. A distribution of the number of profiles from each location is shown in Table 3.1. The 'omitted profiles' refers to the number of profiles removed from the dataset after the pre-processing, explained in Section 3.1.1. Further details of the dataset are provided in Section 2.1.3 and Appendix A. All processes explained in this section are related to Step 1 in Figure 3.1.

Table 3.1: The number of profiles included in the analysis from each location, as well as the number of ommitted profiles.

Location	Profiles	Percent of Total Profiles	Omitted Profiles	Percent Omitted
American Samoa	1,198	4,0%	13	1,1%
Saipan and Tinian	1,035	3,4%	43	4,2%
Guam	1,295	4,3%	20	1,5%
USVI	1,664	5,5%	41	2,5%
Hawaii	13,404	44,4%	52	0,4%
Puerto Rico	5,531	18,3%	79	1,4%
Florida	6,039	20,0%	76	1,3%
Total	30,166	-	324	1,1%

3.1.1. MANIPULATING PROFILES

To begin the analysis, the raw profiles were slightly altered for use in the cluster analysis. This included aligning the profiles so that each depth measurement was from the same cross-shore distance from a reference location, and removing profiles that were deemed outliers or ineffective for this study. The methods are explained in this section.

SETTING A REFERENCE LOCATION

The reference point selected was X=0 (coastline) at depth = 0 m (MSL). Since there could be multiple locations along the profile where the elevation switches between positive and negative values, which would potentially indicate the coastline, land was determined to be where elevation reaches 2 m above MSL and remains above 0 m elevation for a minimum width of 100 m.

The 2 m elevation was selected based on statistics provided by Woodroffe (2008). He noted that oceanward shores are typically around 3 m above MSL, and generally are much lower at the lagoon side. Therefore, the 2 m cutoff should capture most atoll islands. The requirement of remaining above 0 m elevation for 100 m was somewhat arbitrary, but used to ensure that reef points that surpassed 2 m and then quickly fall to below MSL are not treated as land.

The search for the coastline location started from the most seaward point of the profile, and moved landward. If the 100 m width was not met, the same conditions were searched for again moving landward. This process repeated until the required 100 m width was met or the back of the profile was reached. An example is shown in Figure 3.2

of where the most seaward switch between positive and negative elevation values is not selected as the coastline position since the width of profile above 0 m elevation landward of the 2 m elevation is not 100 m.

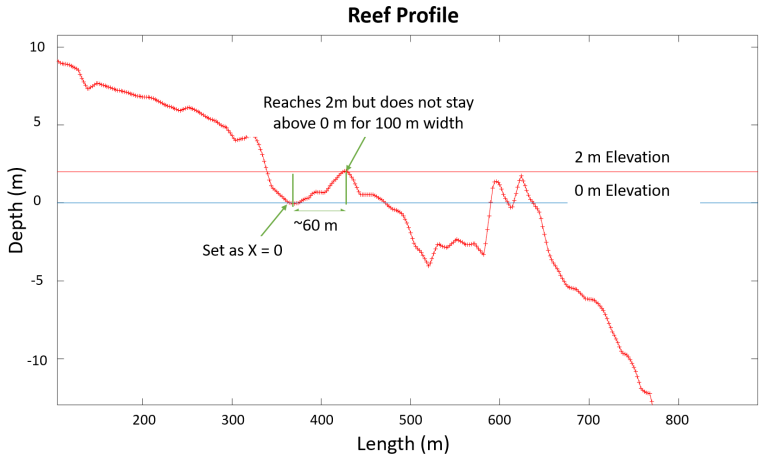


Figure 3.2: An example of a profile from the American Samoa illustrating the process of setting the reference location at depth = 0 m.

OMITTED PROFILES

A profile could be omitted for multiple reasons, all of which would inhibit the profile from contributing to the analysis. If there was a point of the profile that reached above 4 m elevation without above 0 m elevations for 100 m landward, the profile was omitted. These features would largely block wave propagation and behave as land, but without 100 m of above 0 m elevation landward, it was decided that this was too narrow to classify as land and therefore these profiles should be withheld from the analysis. An example of three profiles with this sort of spike is shown in Figure 3.3.

A profile was also skipped if the elevation never reached above 0 m elevation, or never reached below 0 m elevation. It is not possible to set an equal reference location to the others if a profile doesn't satisfy either of these requirements. All omitted profiles from each location are shown in Appendix B.

PROFILE INTERPOLATION

The cluster analysis is based on the similarities and equally the dissimilarities of the variables between each observation. In this case, each profile is an observation and each cross-shore point with a depth measurement is a variable. To compare the cross-shore measurements (variables), the profiles must be aligned. Once the reference point was selected, the profiles had to be slightly manipulated to truly align the profiles at depth = 0 m. Since the profiles have measurements at every 2 m spacing, it is highly unlikely a measurement will be done at exactly depth = 0 m. Therefore, linear interpolation was performed between the two nearest points to find where depth = 0 m. This would result

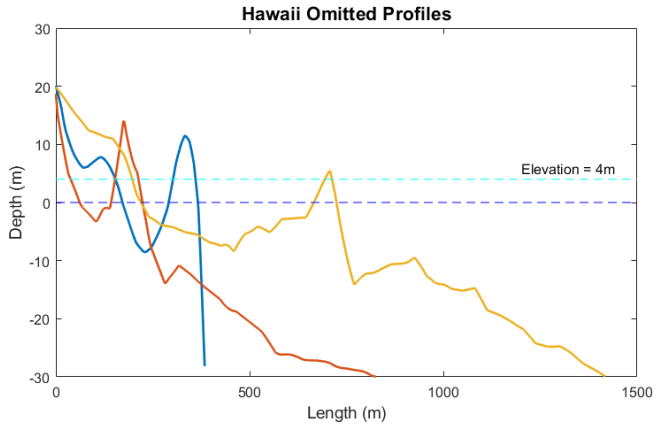


Figure 3.3: Three of the omitted profiles from Hawaii. Each of these profiles were omitted since they have peaks that reach above 4 m elevation that could not be labelled as land since the width of the profile above 0 m elevation is less than 100 m.

in a required 'shift' from one of the measurement points to the X=0 location. In order to maintain the same 2 m spacing between all profile points, linear interpolation was then performed between all points of the profile to obtain new depth values. Figure 3.4 demonstrates the required shift of the data points from setting X = 0 at depth = 0 m.

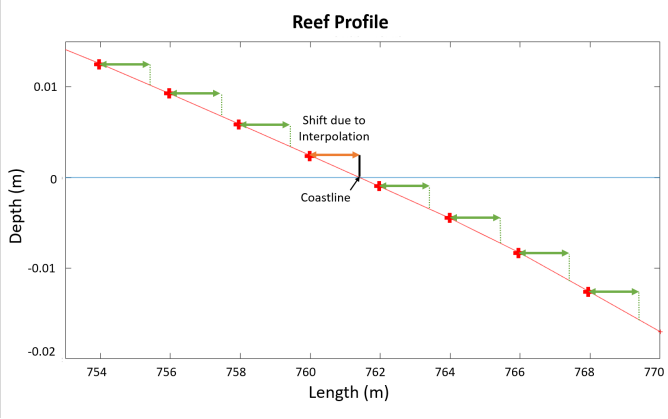


Figure 3.4: Example of the interpolation method applied to all profiles. Red plus marks are locations with measured depths. The orange arrow represents the shift that would be calculated initially from determining where X=0, and the green arrows represent the subsequent required shifts to all data points through interpolation to maintain 2m spacing.

The interpolation does not introduce any error or biases since the profiles were already interpolated in the creation of the dataset (Storlazzi et al., 2019). There is no way to justify that further interpolation will make the data any worse or better.

FILLING MISSING DEPTHS

Seaward Limits of Profiles

The provided dataset included profile points from 20 m elevation to -30 m elevation. The elevation limits for each profile were consistent, but due to the differences in profile shapes, the cross-shore lengths of the profiles varied considerably. A distribution of the profile lengths is shown in Figure 3.5. For the cluster analysis, each profile (observation) must have a value for each cross-shore position (variable). Therefore, each profile was extended with -30 m depth values to the length of the longest profile. The longest profile was from Florida, with a length of 18, 694 m seaward from X=0.

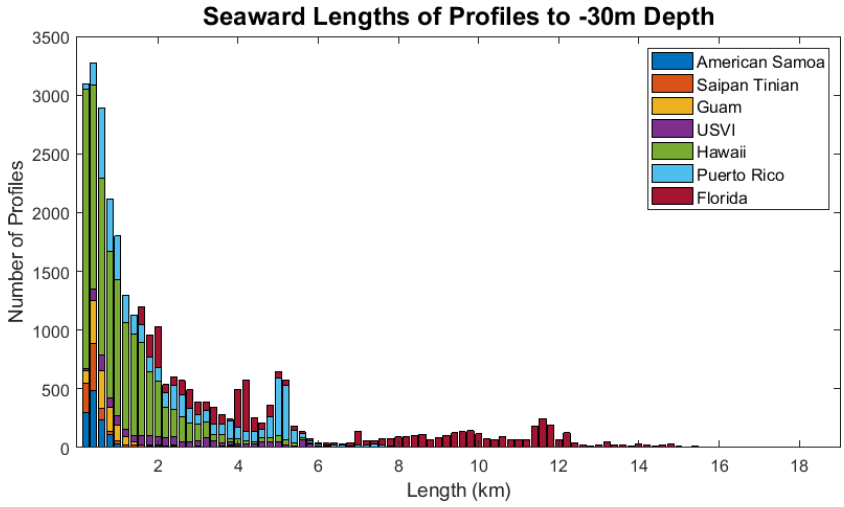


Figure 3.5: Histogram of the seaward length of profiles. Colors of the histogram represent the quantity of profiles from each region to that histogram bin.

Landward Limits of Profiles

By setting the reference location using the methods stated in Section 3.1.1, the landward length of each profile was modified. Some profiles could have X = 0 at its most landward measurement, and others left with kilometres of data landward of X = 0. The most landward value that could be included in the cluster analysis was X = 0 (where all profiles have measurements). It was determined that the important aspects of the profile for the cluster analysis are what occurs below MSL. The beach slope can easily be added to the profile before the XBeach simulations, and would have to be altered anyway to create a semi-infinite slope to gauge the runup values. Therefore, being limited of landward profile data at X = 0 was not a problem and the beach slopes were created later for the XBeach simulations.

ALIGNED PROFILES

Once the reference location was determined and the profile points were set at the equal 2 m spacing, the profiles were aligned and simple analysis of the profiles was possible. All profiles were plotted by location, showing the range of profile types that each location offers. An example is shown below in Figure 3.6, where all profiles from American Samoa

are plotted, as well as the mean profile, median profile, the standard deviation shaded in yellow, and percentiles of the profiles demonstrating the envelope. The dashed lines on the upper subplot, seaward of X=0 are at 2 m and 4 m elevation. The horizontal dashed blue line is at depth = 0 m and the vertical dashed blue line is at X = -100 m. No profiles are present within this area due to the set requirements for determining the coastline position. The same plots for the other locations can be found in Appendix B.

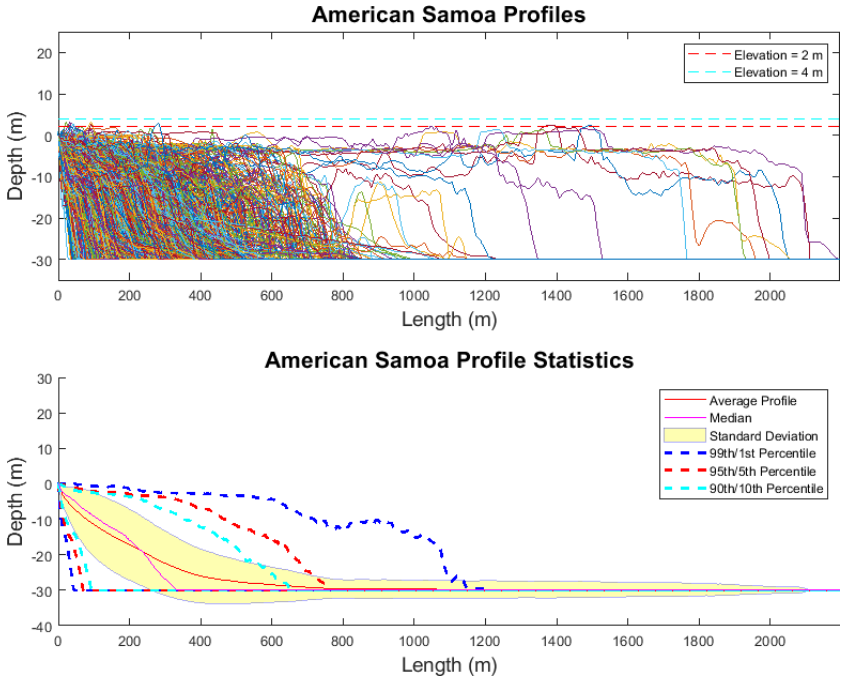


Figure 3.6: Top: All profiles from American Samoa after interpolation and aligning of the profiles. Bottom: The statistics of the profiles, including the mean, median, standard deviation shaded in yellow, and a range of percentiles to represent the profile envelope.

Once all profiles in each location were properly aligned, the cluster analysis methods could be used to begin grouping the profiles.

3.2. INPUT REDUCTION OF REEF MORPHOLOGY

This section refers to Step 2 in Figure 3.1. The main goal of the input reduction is to group similar shaped reef profiles. The hydrodynamics over a reef are known to be dependent upon the reef morphology. Therefore, profiles with similar morphology will have similar wave runup, and one representative profile can be used to estimate the wave runup of all grouped profiles. Multiple cluster analysis techniques were used to determine which worked best for grouping profiles with similar shape.

The input reduction of reef morphology was broken down into three sections. First, the initial runs were done with a wide range of cluster values and minimal replicates (Section 2.4.2) to get an idea of how each method works and which method, as well as

which number of clusters, should be selected for further review. The second round of cluster runs were then modified to concentrate on the methods and cluster values that provide the best results. One final round of clustering was then applied with a high number of replicates using the method and number of clusters that provided the best results.

3.2.1. INPUT REDUCTION INITIAL RUNS

Multiple clustering strategies were implemented to find the optimal method for the reef dataset. To begin, a wide range of cluster values were tested for each algorithm. An overview of the process for the initial cluster runs is shown in Figure 3.7. The main input parameters used for each method are shown in Table 3.2.

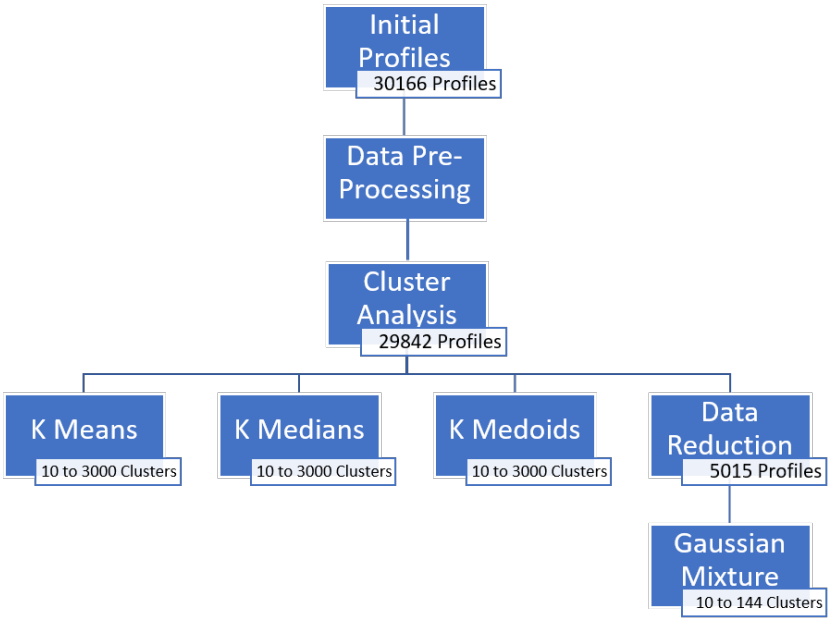


Figure 3.7: Overview of the initial runs in Input Reduction of Morphology.

The number of tested clusters was dependent upon the limitations of the method. As shown in Table 3.2, K-means, K-medoids, and K-medians all used the same range of cluster values between 10 and 3,000, whereas the Gaussian mixture was limited to a range of 10 to 144. The cluster values used for the Gaussian mixture are much less due to memory limitation (computer specifications provided in Appendix D.2). The estimation of a mixture model is computationally expensive for large datasets, such as the one used for this study with many observations and variables (Garcia, Nielsen, & Nock, 2009). It is also uncommon for mixture models to use a high number of clusters since it is difficult to fit so many different distributions to the data.

Few replicates were used for the initial cluster runs. As mentioned in Section 2.4.2,

Table 3.2: Inputs for the initial runs of the clustering algorithms.

	K-means	K-medoids	K-medians	Gaussian Mixture
Distance Metric	SED	SED	City block	EM
Initialization Method	K-means++	K-means++	K-means++	K-means++
Number of Clusters	10 to 3000	10 to 3000	10 to 3000	10 to 144
Replicates	1	3	1	1
Profiles Description	Full	Full	Full	Reduced MDA
Clusters 10 to 3000	10, 50, 100, 150, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1250, 1500, 2500, 3000			
Clusters 10 to 144	10, 21, 25, 49, 64, 81, 100, 121, 144			

it is best practice to use as many replicates as possible to have the best probability of finding the optimum solution. For the initial runs, however, it was more important to obtain the trends from the different methods and number of clusters, not a correct final solution. In order to save computation time, few replicates were used. If the results proved to be odd, higher replicates would be tried in the detailed runs.

FURTHER PRE-PROCESSING FOR GAUSSIAN MIXTURE

When applying the Gaussian Mixture Method, even with a reduced number of clusters, computer memory was a constraint. To be able to do the runs listed in Table 3.2, reduction of the dataset was necessary, as shown in Figure 3.7. To reduce the dataset, the number of profiles (observations) was reduced using the Maximum Dissimilarity Algorithm (MDA), and the number of cross-shore points (variables) was reduced by applying a moving average of the cross-shore profile.

Maximum Dissimilarity Algorithm

MDA was applied to each region individually so that the number of profiles being carried forward from each region could be selected manually and each region could be represented adequately. As seen in Table 3.3, the three locations with the most profiles, Hawaii, Puerto Rico, and Florida were set to find 1/10th of the number of profiles, whereas the four regions with less original profiles were set to gather 1/2 of the original profiles through the MDA. This resulted in 5,015 profiles being output from the MDA analysis and used in the Gaussian mixture method.

Moving Average

After the MDA was complete, the selected profiles were reduced in the cross-shore direction by applying a moving average over the data points. Each data point was averaged using itself, as well as the closest landward and seaward value. Once the landward and seaward values were used in the averaging, they were removed from the profile and the next three data points were used to calculate the next average. Since the original data points were spaced at 2 m intervals, the averaging resulted in a 6 m spacing, and the cross-shore values were reduced by a factor of 3. An image illustrating the process is shown in Figure 3.8. In the end, the red points would remain and the orange points would be extracted from the profiles to reduce the number of variables in the dataset. The averaging was not applied to the first data point (X = 0) to ensure that the reference

Table 3.3: Number of Profiles selected through the MDA for each region.

Location	Original Number	Fraction Used	Output From MDA
American Samoa	1,185	1/2	593
Saipan Tinian	992	1/2	496
Guam	1,275	1/2	638
USVI	1,623	1/2	812
Hawaii	13,352	1/10	1,335
Puerto Rico	5,452	1/10	545
Florida	5,963	1/10	596
Total	29,842	-	5,015

locations for each profile were maintained.

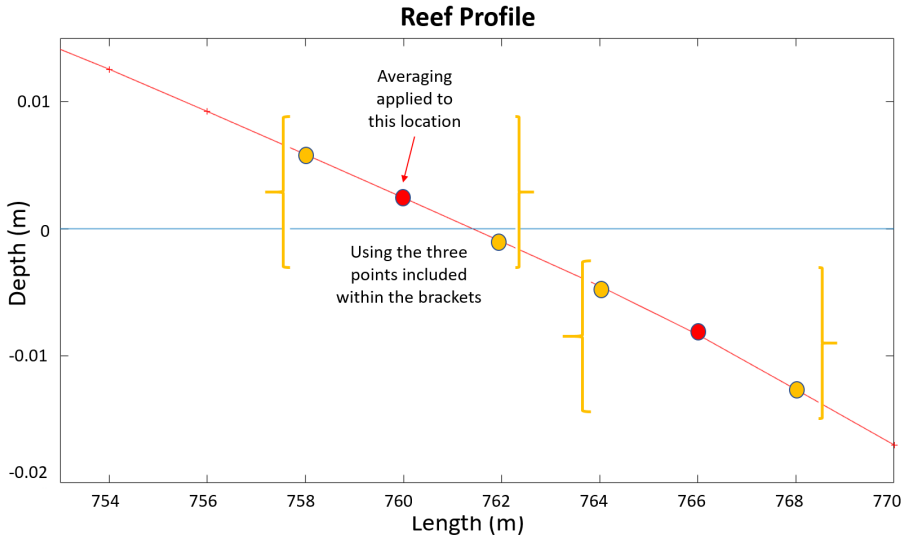


Figure 3.8: Example of how the moving average was applied to reduce the number of cross-shore points of the profiles.

3.2.2. INPUT REDUCTION DETAILED RUNS

The detailed runs consisted of focusing on two methods: K-medians and Gaussian mixture. For this step, more replicates were tested on a select set of cluster values. The details are shown in Table 3.4. The maximum number of clusters tested for the Gaussian mixture was reduced to 64 due to computer memory constraints, while using a higher number of replicates. Details of the computer specifications are provided in Appendix

D.2.

Table 3.4: Detailed cluster runs input

	K-medians	K-medians	Gaussian Mixture
Distance Metric	City block	City block	EM
Initialization Method	K-means++	K-means++	K-means++
Number of Clusters	10, 25, 50	300, 500	10, 25, 36, 49, 64
Replicates	50	100	10
Profiles Description	Full	Full	Reduced MDA

3.2.3. INPUT REDUCTION FINAL RUN

The *K*-medians method was ultimately used to develop the output from the first round of cluster analysis. A final change to the data was applied for this cluster analysis, in which the profiles input into the analysis were preselected by their lengths to exclude profiles that were deemed impractical to include the XBeach simulations. The input details for the final cluster analysis can be seen in Table 3.5, and an explanation of the profile length requirements is provided below.

Table 3.5: Final cluster run with the *K*-medians method and profiles reduced by the length requirement

	K-medians
Distance Metric	City block
Initialization Method	K-means++
Number of Clusters	500
Replicates	50
Profile Description	Length limited

PROFILE LENGTH REQUIREMENTS

A major issue regarding the profile data and setting up the next step of the analysis with the XBeach simulations was with regard to the lengths of the profiles. The longest profile measured from X = 0 going seaward to the first -30 m depth is a profile from Florida that reaches 18,694 m. The histogram of seaward lengths of the profiles, separated by region is shown in Figure 3.5.

The longest reef widths used by Pearson et al. (2017) were 1.5 km, which required a model spin up time of roughly 120 min. Going beyond this value would be computationally very expensive. Furthermore, the XBeach x-grid for the long profiles would require many nodal points that would lead to numeric dissipation of the wave energy and would as such provide inaccurate results. Suffice to say, XBeach is simply not an appropriate model to use for such long cross sections. Furthermore, the results of this analysis are meant to be included in an updated BEWARE network (Pearson et al., 2017). This system focuses on fringing reefs and atolls, and so focusing this study on similar reef

lengths makes most sense to generate meaningful contributions. The long profiles from Florida and Puerto Rico resemble barrier reefs, and so are not applicable for this study. Furthermore, with such long profiles, two dimensional processes and local wave generation (Nelson, 1997) may become important factors which XBeach does not incorporate. Assessing these longer profiles would be a suggested step for future work.

As a result of these factors, only profiles that met a length requirement were input into the final cluster analysis. The requirement was calculated based on the depth of the profile that would begin to impact the incoming waves. The largest wave height used in the XBeach simulations was set to be 7 m. According to Gourlay (1994), the breaking threshold $\gamma_b = H_b/h_b$ was proven through laboratory and field data to never exceed 0.55 for shallow water waves across a flat reef. Although the reefs in this study are not always flat, the estimate of $\gamma_b = 0.5$ was used to calculate at which depth the waves would begin to break on the reef. With a 7 m wave, this depth is calculated to be 14 m as shown below.

$$\gamma_b = \frac{H_b}{h_b};\ h_b = \frac{7m}{0.5};\ h_b = 14m$$

Through expert judgment by supervisor and XBeach developer Robert McCall (McCall, personal communication, 2019), it was assumed that roughly 10-20 wavelengths could be resolved properly in the XBeach model for the largest waves used in this study. The limiting wave condition has a wave period of roughly 10 s, and so simply assuming a depth of 10 m, the wave length is equal to 100 m. Therefore, 15 wavelengths is equal to 1.5 km.

The limit was set such that profiles were included if their length was less than or equal to 1.5 km at the most seaward -15 m depth point. A depth of 15 m was used to be slightly more conservative.

Using this profile length limit, 31% of the total profiles were removed, mostly from the regions of Florida and Puerto Rico, which are known to have very long profiles. The details of the number of profiles that satisfy the length requirement per region are shown in Table 3.6.

Table 3.6: Details on the number of profiles excluded from the cluster analysis by region, due to length restrictions.

Location	Original Number of Profiles	Number of Profiles that satisfy Length Limit	Percent Removed
American Samoa	1,185	1,177	1%
Saipan Tinian	992	909	8%
Guam	1,275	1,245	2%
USVI	1,623	1,132	30%
Hawaii	13,352	12,009	10%
Puerto Rico	5,452	2,927	46%
Florida	5,963	1,055	82%
Total	29,842	20,454	31%

EVALUATE CLUSTER RESULTS

Step 3 (Figure 3.1) involves evaluating the results to determine which clustering method groups profiles with the greatest intra-cluster similarity. At this step, the dissimilarity between all of the profiles and their centroid was used to compare the cluster groups. This was done by simply calculating the sum of absolute difference at each variable (cross shore position) between the observation (profile) and the centroid that it is assigned to, and dividing by the number of cross shore positions to obtain the average distance between the observation and the centroid. In mathematical terms:

$$IE = \frac{\sum_{i=1}^{n} |c_i - o_i|}{n}$$

where IE is the average individual error, c_i is the centroid value at variable i, and o_i is the observation value at variable i and n is the number of variables.

SELECT CLUSTER ROUND 1 PROFILES

To select the number of profiles, the cost-benefit of morphologic similarity vs XBeach computation time was assessed. When using more cluster groups, the similarity within groups is higher, but each cluster group adds valuable computational time. The incremental gain by adding more cluster groups was the ultimate indicator to determine the optimal number of profiles.

3.3. XBEACH MODELLING OF CLUSTER PROFILES

The Cluster Round 1 profiles were then modelled in XBeach. This section refers to Step 4 in Figure 3.1. Four different loading conditions were selected to cover a wide range of potential flooding conditions. The hydrodynamic output from the XBeach simulations was used to group the cluster profiles that have similar hydrodynamic responses.

A second set of XBeach simulations was done with the most dissimilar profiles within each group to the cluster profile. Comparing the runup results of the cluster profile and the dissimilar profiles provides quantitative results of how well each cluster profile represents all of the profiles within the cluster.

3.3.1. FORCING CONDITIONS

The forcing conditions were strategically chosen to encapsulate a variety of wave conditions with limited runs. The selected inputs for the four runs„ as well the coefficient of friction and beach slope applied to the profiles can be seen below in Table 3.7.

Only the wave height and wave steepness (H/L) vary between runs. The wave heights selected are 3 and 7 m. Anything much larger than 7 m can already be expected to cause significant flooding and damages, so 7 m was set as the maximum. The 3 m was set as the smaller wave because it is still a significant size that will lead to high runup, but is low enough to generate a much different response compared to the 7 m wave.

The wave steepness is the second varying input, which alternates between 0.05 to resemble wind waves, and 0.01 to resemble swell waves. Applying the different steepness values along with the varying wave heights further enhances the spread of conditions that can be modelled in limited runs.

Table 3.7: XBeach model wave loading conditions and additional reef profile parameters.

			Loading Condition			
Symbol	**Parameter**	**Units**	**1**	**2**	**3**	**4**
H_0	Wave Height	m	3	7	3	7
$\frac{H_0}{L_0}$	Wave Steepness	-	0.05	0.05	0.01	0.01
T	Wave Period	s	6.2	9.5	13.9	21.2
n_0	Offshore Water Level	m	1	1	1	1
c_f	Coefficient of Friction	-	0.05, 0.1	0.05, 0.1	0.05, 0.1	0.05, 0.1
B_{beach}	Beach Slope	-	1/10	1/10	1/10	1/10

3.3.2. ANALYSIS OF XBEACH RESULTS

INTRA-CLUSTER VARIABILITY

To select the most dissimilar profiles, a simple method was used in which the five profiles with the largest cumulative difference to the cluster profile in depth at all cross-shore points were selected. The difference is expressed as:

$$Difference = \sum_{i=1}^{n} |p_i - c_i|$$

where i is the cross-shore position, n is the number of cross-shore positions for the profile, p is a profile within the cluster group under consideration, and c is the cluster profile. Five profiles were selected to provide sound insight into the spread within each cluster group. For cluster groups with less than five profiles, the number of profiles within the group was used. For cluster groups of only one profile, there is no variability to assess. An example of one of the cluster groups, highlighting the cluster profile and one of the most dissimilar profiles is shown in Figure 3.9.

Since this study was restricted by time, only loading conditions 3 and 4 from Table 3.7 were used for modelling the most dissimilar profiles. These are the conditions with the 7 m wave height. They were used because the larger wave height will result in a greater spread in wave runup values, therefore providing the maximum difference in runup that can be expected from each cluster group.

ASSESSING VARIABILITY

To assess the variability within a cluster group, the relative $R_{2\%}$ difference between the dissimilar profiles and the cluster profiles was evaluated. The maximum difference from this method will provide the best estimation of the maximum spread in runup within the group, and will provide an idea of with what accuracy the cluster profiles can represent the runup of all profiles within the cluster.

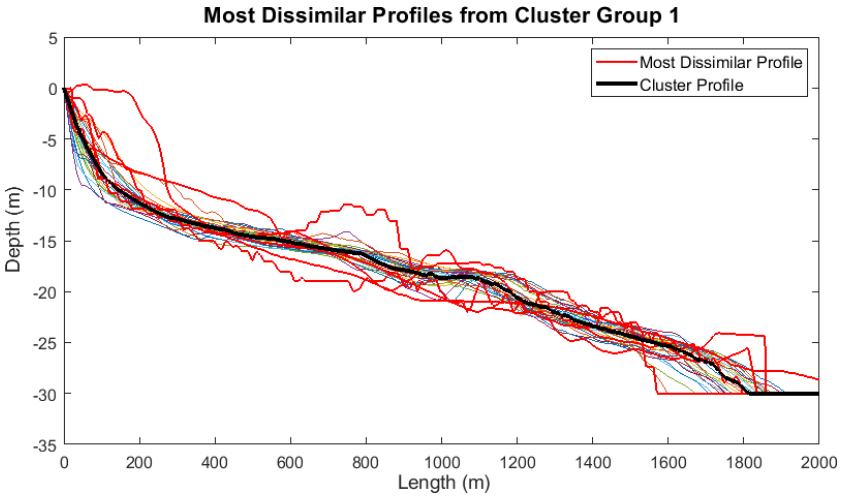


Figure 3.9: An example of one cluster group, highlighting the cluster profile and the most dissimilar profile. It is assumed that the most dissimilar profile in morphology will also lead to the most dissimilar runup results, leading to an approximation of the maximum spread in runup within the group. The multicolour lines are the rest of the profiles in the cluster group.

3.4. CLUSTER ANALYSIS OF REEF HYDRODYNAMICS

To finalize the data reduction process, a cluster analysis was applied to group profiles with similar wave runup results. This section refers to Step 5 in Figure 3.1. The wave runup output from the XBeach simulations on the Cluster Round 1 profiles can be used to further reduce the dataset, merging the groups that have similar responses to the different loading conditions. Figure 3.10 illustrates the process of moving from the initial thousands of reef profiles through the first data reduction in Cluster Round 1 to 500 cluster profiles, and the further reduction based on both morphology and hydrodynamics.

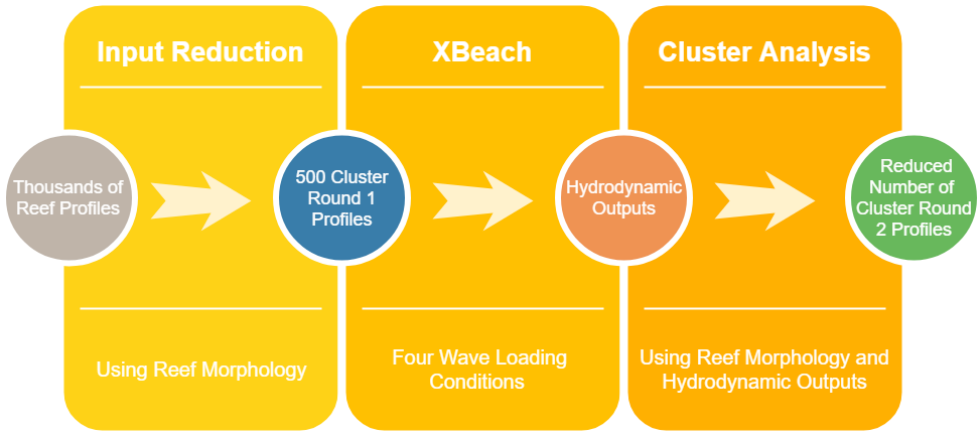


Figure 3.10: The general process of the cluster analysis, beginning with the thousands of raw reef profiles, and finishing with the final representative profiles as output from the second round of cluster analysis.

3.4.1. CLUSTER ANALYSIS INPUTS

For the second round of cluster analysis, both the reef morphology and the hydrodynamic outputs from the XBeach simulations were used to group similar profiles. The hydrodynamic values used were the $R_{2\%}$ values, the setup at the shoreline, and the swash values separated into infragravity and high frequency bands. A representation of how the inputs come together to be used for the second round of the cluster analysis is shown in Figure 3.11.

All hydrodynamic inputs were calculated using the runup gauge time series which recorded output every 0.5 s. The $R_{2\%}$ is calculated as the 2% exceedance value of all runup elevations. The setup is measured as the mean water level at the runup gauge. The swash values are determined from the spectra, PSD(F), of the runuo gauge water-level time series. Since the infragravity and high frequency swash was calculated separately, the spectrum was split at a frequency of 0.05. The swash value was then calculated as:

$$Swash = 4 * \sqrt{\sum PSD(f)df}$$

where PSD(F) was calculated using the trapezoidal numerical integration. All of these values are relative to the offshore water level, which was set to 1 m for all simulations.

The morphology and the hydrodynamics were given equal weighting. The morphology had already been used to group profiles in cluster analysis Round 1, and so using it as an input again is repetitive, however, it was included to ensure that profiles with the same runup values but different process to cause those runup values were not grouped together. If a profile has very different shapes, the processes occurring to generate the runup is most likely different. The underlying processes need to be similar to ensure that the runup will be similar for more than just the wave conditions tested in this study. Therefore, morphology was included with 50% weighting, and the three different hydrodynamic inputs shared the other 50% evenly. More details of the inputs are provided in

Appendix C.3.2.

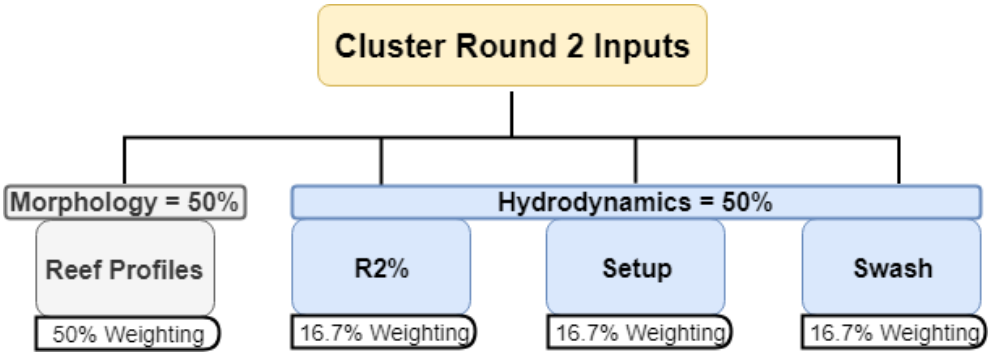


Figure 3.11: The inputs and weightings used for the second round cluster analysis.

3.4.2. HIERARCHICAL CLUSTERING

The Agglomerative Hierarchical Clustering method was used for grouping the Cluster Round 1 profiles. The process is explained in detail in Section 2.4.2, and works by starting with each profile as individual clusters and merging similar ones based on a set distance method and metric. The different distance methods are described in Appendix C.3.1, and each distance method can be computed with different distance metrics, such as euclidean distance and cityblock distance. The ward method with euclidean distance was determined to be the optimal method for this study, and therefore was used for grouping profiles and creating the final representative profiles. A detailed comparative analysis between the different hierarchical methods can be found in Appendix C.3.3.

ASSESSING CUTOFF VALUES

To establish the final number of cluster groups, multiple different cutoff values were tested and compared. The cutoff values are used to find natural divisions in the dataset (explained in Section 2.4.2), but what that means in terms of hydrodynamic response is unknown unless specifically evaluated. With each cutoff value tested, the newly formed clusters were compared based on the same hydrodynamic values that were used as inputs to the algorithm ($R_{2\%}$, setup, swash). Specifically, the intra-group variability in these values was assessed, since this provides a representation of the error that can be expected from the application of this method.

SELECTING THE REPRESENTATIVE PROFILE

Once the hierarchical clustering algorithm is applied, newly formed groups of profiles are created. In Cluster Round 1, when the new groups were formed, the median of the profiles was used as the cluster profile to represent the group. For Cluster Round 2, it was decided that the median profile is no longer applicable, since fewer profiles are being grouped together and it is no longer morphology alone being used to cluster the profiles. With only two or three profiles making up a cluster group, the median profile could be very different than either of the original profiles, which could also have very different

runup results compared to either profile part of the cluster group. An example is shown in Figure 3.12.

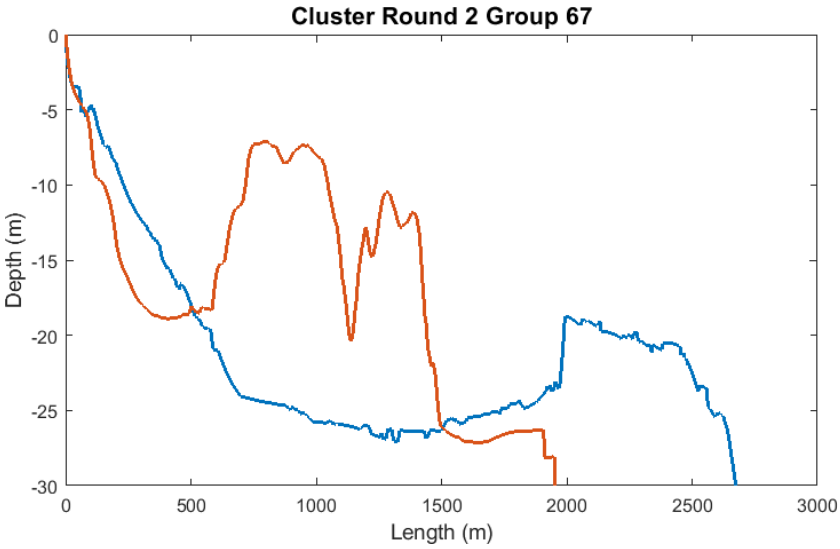


Figure 3.12: An example of two profiles merged together in Cluster Round 2 that would not be well represented by the median profile.

As a result, one of the profiles included in the cluster group was selected as the representative profile for the group. This way, the representative profile is guaranteed to force similar hydrodynamic results to all profiles part of the newly formed group, since they were grouped together based on their hydrodynamic results. The selection of which profile is to be the representative profile was based off the $R_{2\%}$ values. The profile with the $R_{2\%}$ values closest to the mean of the group was selected. For groups with one profile, or three or more, this method works fine. An example is shown in Figure 3.13 a. For groups with two profiles, this method does not work since each profile will be equal distance to the mean and so the one which represents more profiles from the first round of the cluster analysis was selected as the representative profile. An example is shown in Figure 3.13 b.

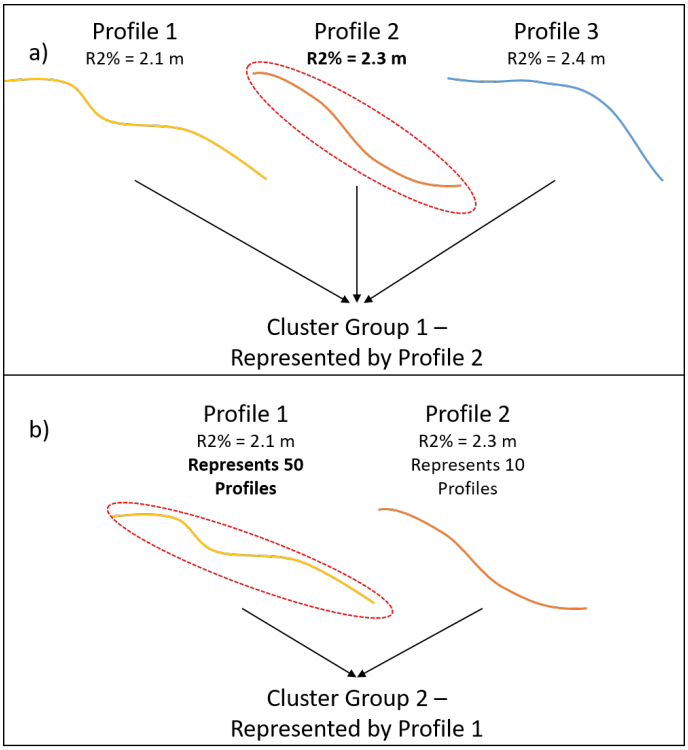


Figure 3.13: An example of how the representative profile for the cluster group is selected,a) when there are 3 or more profiles it is based on the profile with the $R_{2\%}$ value closest to the mean value of those grouped together, and b) when there are 2 profiles it is based on the profile that represents more profiles from the first round of cluster analysis.

ASSESSING RUNUP ERROR IN THE CLUSTER GROUPS

Step 6 (Figure 3.1) includes evaluating the Cluster Round 2 results. To establish the final number of cluster groups, multiple different cutoff values were tested and compared. The cutoff values (explained in Section 2.4.2) limit the allowable variance between merging clusters, meaning that each cutoff value will lead to a different number of final cluster groups. To compare them, the differences in $R_{2\%}$ of merged cluster profiles was evaluated. Specifically, the error between the most different profile in the group to the selected representative profile, shown in Figure 3.14 a. This error is a conservative estimate of the variability within the newly formed cluster groups and provides a great representation of the error that would be expected with the application of this method. It is not, however, always the maximum error since a profile that is part of one of the Cluster Round 1 groups could have a greater relative difference than any of the Cluster Round 1 profiles. An example is shown in Figure 3.14 b.

By assessing the average runup error that each cutoff value generates, the cutoff value and in turn the final number of cluster groups can be decided.

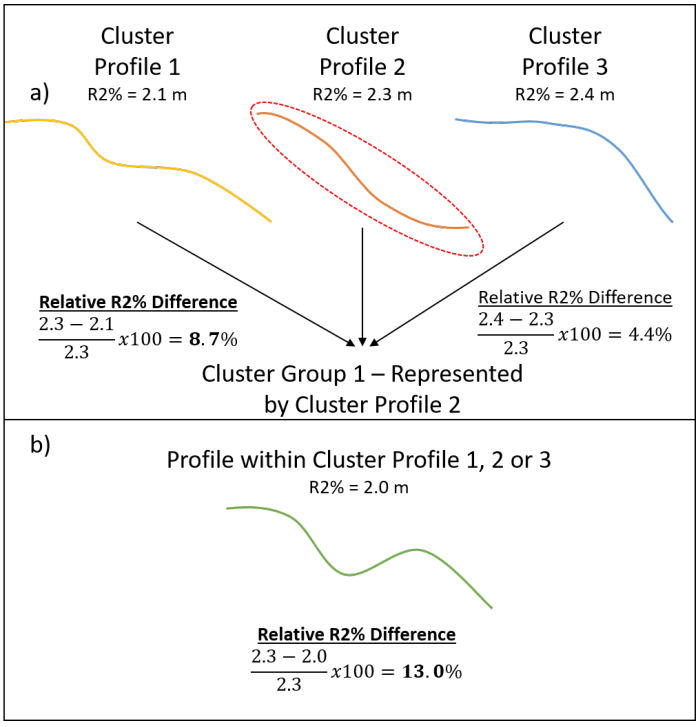


Figure 3.14: An example of how the relative error is calculated between Cluster Round 2 groups, and how the method is a conservative estimate since it calculates the highest error from the profiles being grouped (a), but that there could and most likely is another profile within the Cluster Round 1 profiles that is even further from the representative profile (b).

3.5. TESTING THE APPLICATION OF THE METHOD

The cluster profiles are intended to be used to estimate the runup over various different reef profiles from around the world. To test the application, 1,000 test profiles were extracted from the dataset and run in XBeach with the same loading conditions as the cluster profiles (Table 3.7). The test profiles were then be matched to one of the cluster profiles, and the runup values were compared. This section refers to Step 7 of the methodology (Figure 3.1), and the application process is shown in Figure 3.15.

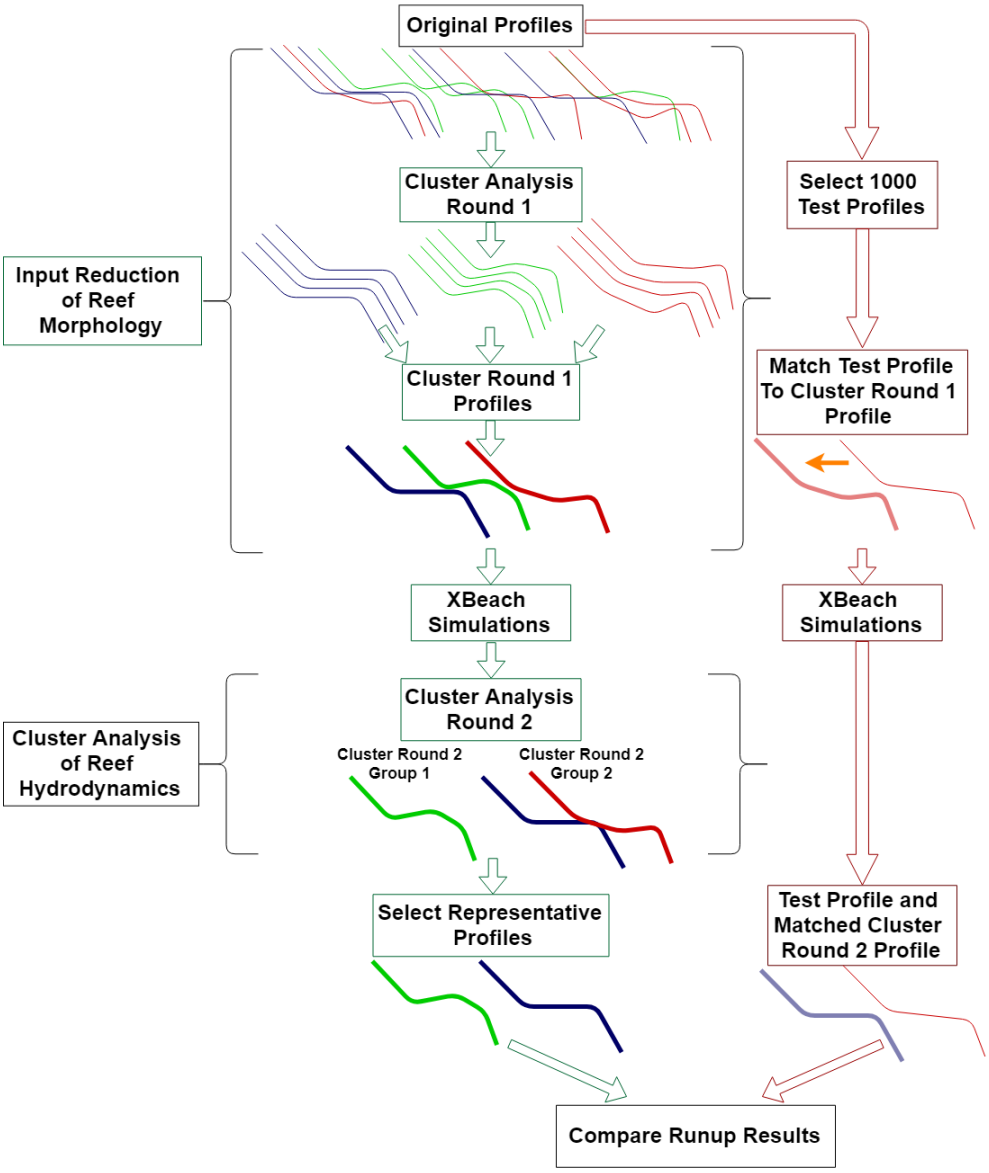


Figure 3.15: The process of how the clustering profile prediction is tested. A set of 1000 real profiles part of the initial dataset are simulated in XBeach, and the runup is compared to the matched cluster profile. The test profiles are matched to one of the Cluster Round 1 profiles (red), and through that association are then paired with one of the Cluster Round 2 profiles (blue). The runup results between the test profile and Cluster Round 2 profile are then compared.

3.5.1. MATCHING PROFILES TO CLUSTER PROFILES

The match between the test profile and cluster profiles is done by comparing morphology, and since the cluster analysis of hydrodynamics incorporates inputs other than morphology, matching a test profile to a Cluster Round 2 profile is not possible. The test profiles must be matched to the Cluster Round 1 profiles and through the association of the Cluster Round 1 profiles in the Cluster Round 2 groups, the match to the Cluster Round 2 profiles can be made. This is shown in Figure 3.16 in which the test profiles get matched to the Cluster Round 1 profiles, and based on the grouping of the Cluster Round 1 profiles the Cluster Round 2 profile can be determined.

Three different methods were tested to match the test profiles to the Cluster Round 1 profiles, explained below.

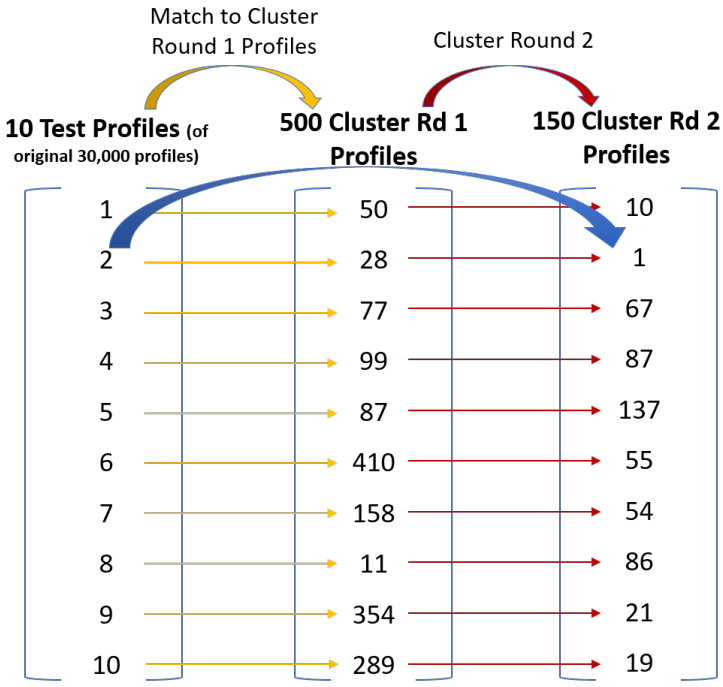


Figure 3.16: Example of 10 test profiles and how they are matched to the Cluster Round 2 profiles based on their association to the Cluster Round 1 profiles.

MATCH BASED ON FULL PROFILE

The first method includes computing the cityblock distance of the test profile to all Cluster Round 1 profiles, using the full profile measurements. The test profile is matched to the corresponding cluster profile with the minimum distance. Cityblock distance is used to stay consistent with the methods used in the data reduction techniques. Since the Cluster Analysis Round 1 was also done with the full profile and cityblock distance,

using this method, the test profile will be matched to the Cluster Round 1 profile that it was initially grouped to. Therefore, this method was given the name CR1 to represent the same match that would be formed from the Cluster Round 1 analysis.

MATCH BASED ON NEARSHORE DEPTHS

From previous findings during this study, it was concluded that the nearshore morphology (0 to -15 m depths) is critical in grouping profiles to obtain similar runup results. Therefore, matching the test profiles to the Cluster Round 1 profiles using a method focused on the nearshore depths was applied. The NS3 method includes three features for matching the profiles. First, the distances between the test profile and all Cluster Round 1 profiles is computed using cityblock distance for only the section of the profiles that is within the 0 to -15 m depths of the test profile. Second, the test profile is checked for a peak at some point along the profile above MSL, and is then limited to match with a cluster profile that also does or does not have a peak above MSL. Finally, the potential cluster profiles must not have a difference in total length (to -30m) of greater than 500 m compared to the test profile. Figure 3.17 illustrates the three requirements to match profiles based on the NS3 method.

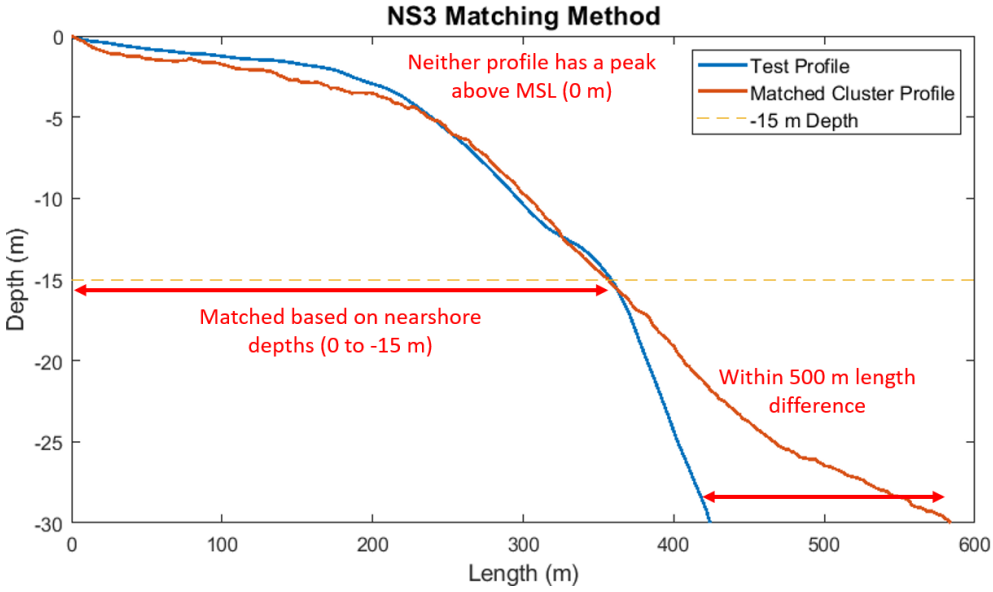


Figure 3.17: Example of the application of the NS3 match method with one of the test profiles.

MATCH BASED ON PROBABILITIES

The final matching method included matching each test profile to multiple cluster profiles based on probabilities. A probabilistic approach is beneficial when the test profile is not very similar to one cluster profile, but rather is in between cluster profiles. Using probabilities, also known as a soft clustering approach (Bauckhage, 2015), the test pro-

file belongs to several cluster profiles simultaneously to a certain degree. The probability reflects the degree to which the test profile belongs to each cluster profile.

To determine the match probability, the softmax function, described in Appendix C.4.1, was used. This transforms distances between profiles into a probability distribution that sums to 1. The distribution of probabilities can be influenced by a parameter, which when assigned a higher value, the closest match is given a larger share of the probability. This processes is explained in Appendix C.4.1.

This method has great potential when trying to incorporate the cluster profiles into a Bayesian Network such as BEWARE. Ideally, the associated probabilities could be input into the network, which would use them when calculating the uncertainty and estimation of the runup response. For this study, the runup estimation is calculated using a weighted ensemble mean.

In the weighted ensemble mean, each probability is used as a weight factor for calculating the runup value. The cluster profiles with the highest match probability will therefore contribute the most to the runup estimation, and the profiles with the lowest probabilities will contribute the least. Since the probabilities sum to 1, this method is valid. The process of going from probabilities to a runup estimation is demonstrated in Figure 3.18. An uncertainty can also be calculated using a weighted standard deviation.

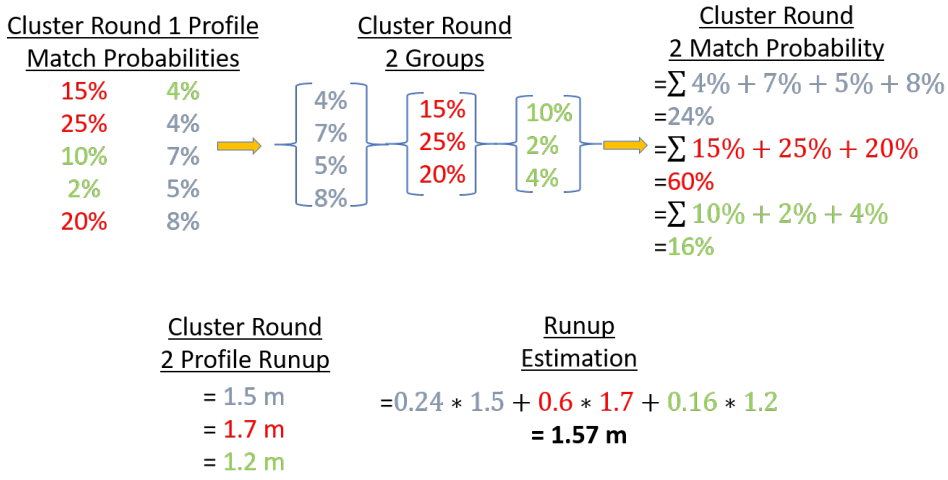


Figure 3.18: Example of the application of the probabilistic match method. First, the test profile is matched to the Cluster Round 1 profiles. Based on the Cluster Round 2 grouping, the probabilities are also grouped to form the Cluster Round 2 match probabilities. The probabilities are then used as weights in the weighted ensemble mean calculation to determine the runup estimation.

As mentioned above, the softmax function transforms distances into probabilities, and therefore, a method to calculate the distance is still necessary. The two distance methods mentioned in this section, including the distance between the full profile (CR1) and the nearshore only (NS3) were used in this study. An example of the probabilistic match between a test profile and the cluster profiles using the NS3 distance method is shown in Figure 3.19.

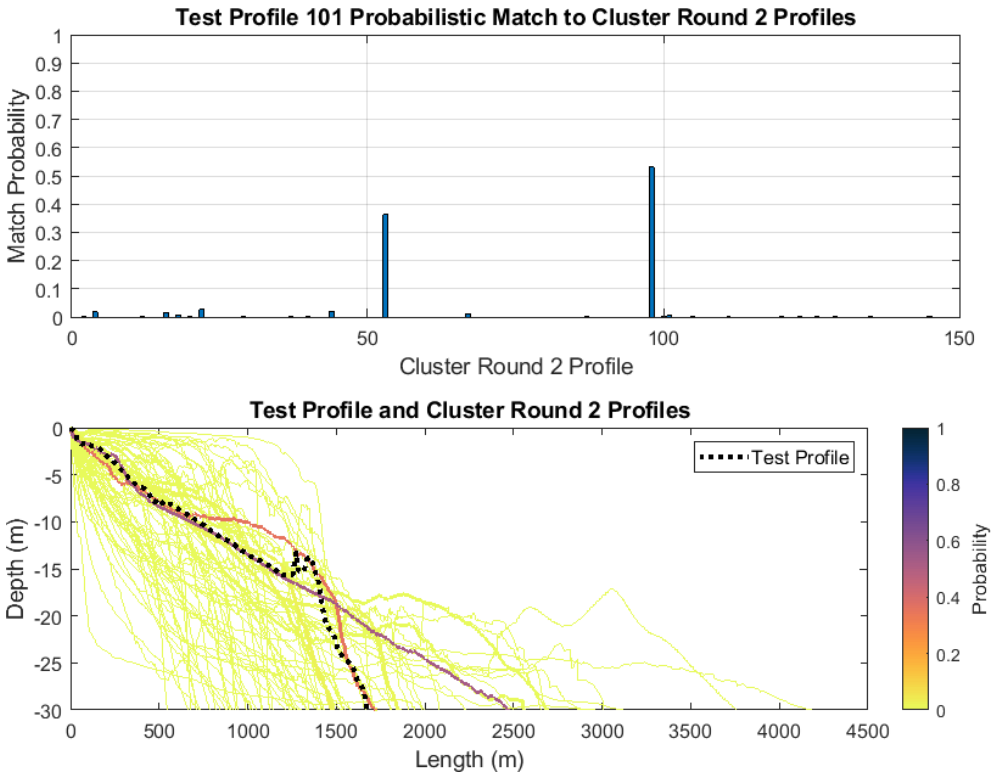


Figure 3.19: Example of the application of the probabilistic match method. First, the test profile is matched to the Cluster Round 1 profiles. Based on the Cluster Round 2 grouping, the probabilities are also grouped to form the Cluster Round 2 match probabilities. The probabilities are then used as weights in the weighted ensemble mean calculation to determine the runup estimation.

3.5.2. ASSESSING PERFORMANCE

Once the test profiles are matched to the cluster profiles, the XBeach output can be compared to assess how well the cluster profile matches the test profile. The cluster profiles and the test profiles were simulated under the same four loading conditions. Therefore, the relative difference in hydrodynamic outputs ($R_{2\%}$, setup, and swash) can be compared for each loading condition, as well as averaged to obtain a representative value of how well the cluster profile represents the test profile across all conditions.

3.6. VALIDATION

Step 8 (Figure 3.1) includes validating the results of the cluster profiles. This was done with two reef profiles from the Republic of the Marshall Islands, specifically from Roi Namur, an island in the north part of the Kwajalein Atoll. The validation is essential to determine how well the cluster profiles perform for locations that were not part of the study.

The two profiles were simulated in XBeach under the same four loading conditions

that the cluster profiles were subject to. The Roi Namur profiles then undergo the same treatment as the 1000 test profiles when testing the application of the method. First, they are matched to the cluster profiles, and second, the wave runup is compared to evaluate how well the matched cluster profile represents the true profile.

To match the Roi Namur profiles they must have depth measurements at the same cross shore positions as the cluster profiles, which is 2 m spacing from X = 0 at the coastline. This was achieved using the same interpolation method applied to all profiles in this study, explained in Section 3.1.1.

This process is an XBeach to XBeach comparison rather than comparing XBeach results to measured data. This is done to strictly validate the cluster profiles and the methodology rather than also dealing with XBeach validation against measured data.

The beach slope used throughout the analysis for all cluster profiles and test profiles has been 1/10. Since the focus is on the reef profile below MSL, the Roi Namur profiles were also modelled with a semi-infinite slope set to 1/10. Therefore, the validation is strictly of the cluster analysis method and does not include differences in features that were not analyzed in this study.

4

RESULTS

CHAPTER SUMMARY
This chapter presents the main results regarding the generation of the cluster profiles and their application in predicting wave runup. The results are separated in the order of the analysis, starting with the first input reduction, followed by the findings from the XBeach simulations, the cluster analysis of hydrodynamics, and concluding with the application and validation results. Following this approach, the dataset was reduced by two orders of magnitude to a final range of 311 to 45 cluster profiles, and the produced cluster profiles were capable of predicting wave runup with a mean relative difference of approximately 10%.

4.1. INPUT REDUCTION OF REEF MORPHOLOGY

To perform the input reduction of reef morphology, three different stages of clustering were completed, ranging in scope to decipher the optimal clustering technique. The results from the three different stages of clustering are presented in this section. Additional evaluation techniques including Calinksi Harabasz, Davies Bouldin, AIC and BIC can be found in Appendix C, as well as a description of the boxplot used to present the results.

4.1.1. INPUT REDUCTION INITIAL RUNS

The initial cluster runs consisted of multiple different algorithms and number of cluster groups. The results from these runs were used to determine which methods are most compatible with the dataset, as well as the numbers of cluster groups to focus on.

INDIVIDUAL AVERAGE ABSOLUTE DIFFERENCE

The average absolute difference details how well each profile is represented by its centroid. Explained in Section 3.2.3, this direct evaluation computes the average depth difference between the profile and its centroid. The results for each clustering method are shown in Figure 4.1.

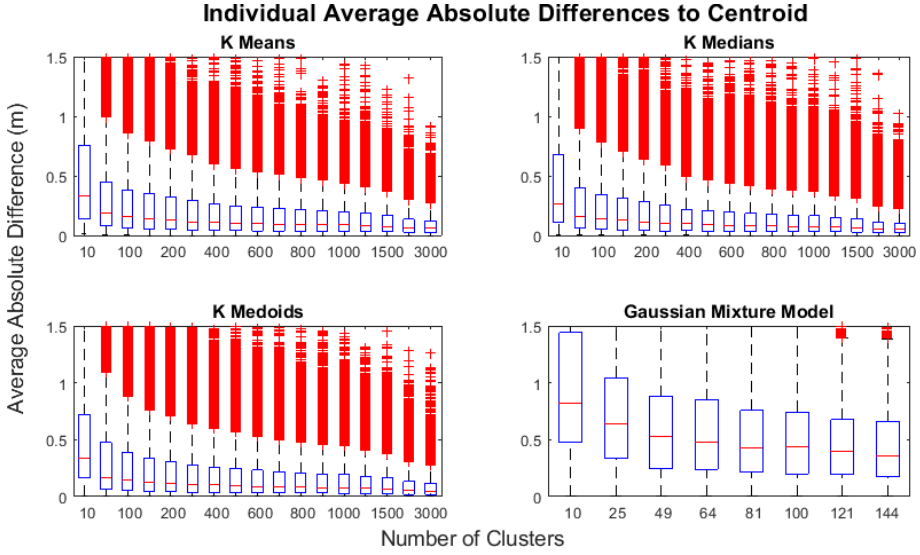


Figure 4.1: A comparison between the clustering algorithms and different number of cluster groups, analyzing the average absolute difference between each profile and its centroid.

The K-means, K-medians, and K-medoids methods show very similar distributions of error, whereas the error from the GMM is much higher. It is clear that the distance between the profiles and the centroids decrease as the number of clusters increases. At 10 cluster groups, the box representing the 25th and 75th percentiles shows a much greater spread compared to the box at 3000 cluster profiles.

For a more direct comparison between the methods, the results from the 10, 100, 500 and 1500 cluster groups from each method were plotted against each other, shown in

Figure 4.2. The blue dot represents the mean of all values above the 95th percentile. In the bottom two subplots, there is no boxplot for the GMM method because these cluster values were not performed with the GMM method.

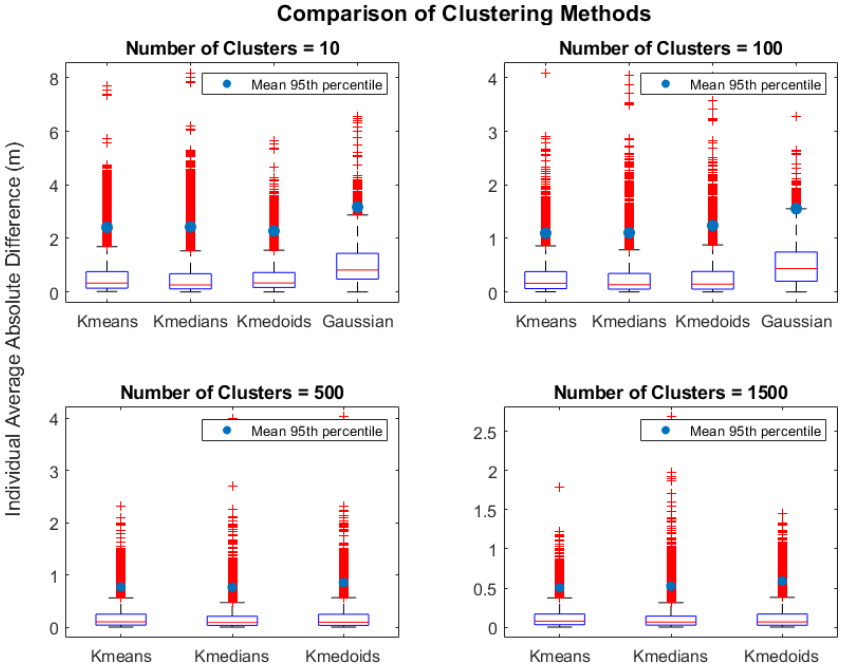


Figure 4.2: A comparison of the average individual profile difference to centroid between the different clustering methods for the case of 10, 100, 500 and 1500 cluster groups. The blue dot represents the mean of all values above the 95th percentile.

Although the results from the three hard partitioning methods are very similar, the spread between the 25th and 75th percentile from the K-medians method is slightly lower than the others, proving that the K-medians method produces the most similar cluster groups. There are more outliers with higher errors from the K-medians method compared to the others, especially in the bottom right plot displaying the results of 1500 cluster groups, however, these are only a handful of the approximately 30,000 profiles included in the analysis. The blue dot representing the mean above the 95th percentile is a better representation of the highest errors, and it is relatively constant between the three hard partitioning methods.

SELECTION OF CLUSTERING ALGORITHMS AND NUMBER OF CLUSTERS TO FOCUS ON

Upon completing the evaluation of the initial cluster results, K-medians and GMM algorithms were selected to be focused on further. K-medians was selected because it provided the best results from the three hard partitioning methods, and GMM because the sporadic results hinted towards more replicates being necessary to obtain a more accurate representation of the effectiveness of the method.

4.1.2. INPUT REDUCTION DETAILED RUNS

Once the two clustering algorithms were selected to investigate further, a slightly more detailed analysis was done to determine which method creates the cluster groups with the least variance in morphology. For this analysis, a greater number of replicates were applied for both the GMM and K-medians methods. As explained in Section 3.2.2, either 50 or 100 replicates were used for the K-medians method, and 10 replicates were used for the GMM method. Previously, both methods had used one replicate. The fewer number of replicates for the GMM method is due to computer memory limitations (see Appendix D.2), which also limited the maximum number of clusters evaluated for the GMM method to be 64. The results are presented below.

INDIVIDUAL AVERAGE ABSOLUTE DIFFERENCE

The average absolute difference between the profiles and their centroid for the K-medians and GMM are plotted together in Figure 4.3. The K-medians results were only slightly improved when using the higher number of replicates, whereas the GMM results improved more significantly. Both methods use the K-means ++ algorithm for strategic initialization. This proves that for K-medians, the initialization works very well, whereas for the GMM more replicates are necessary for a good result. This is most likely due to the fact that the K-means ++ algorithm is very similar to that of K-medians, and was designed for hard partitioning methods. The GMM method works differently and therefore the initialization from K-means ++ does not always lead the algorithm in the right direction.

Since the two methods were used with different numbers of clusters, in order to obtain a direct comparison, K-medians was also evaluated using 64 clusters. Figure 4.4 shows the average absolute differences for the two methods using 64 clusters, and here it is also shown that K-medians again has a lower median average difference between the profiles and the centroids.

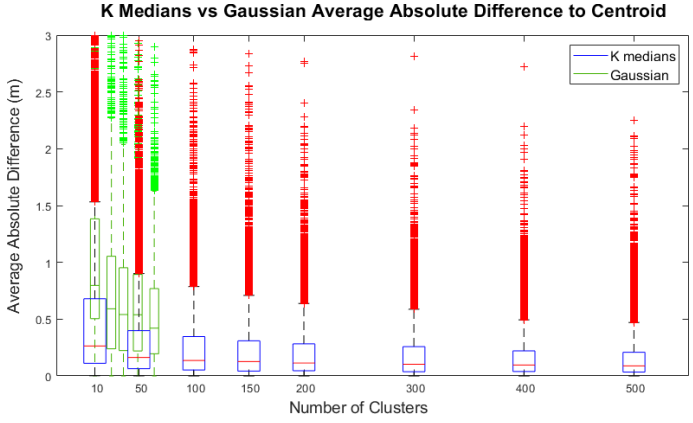


Figure 4.3: A comparison between the K-medians and GMM clustering results, showing the average differences between each profile and its centroid for the two algorithms and multiple different number of cluster groups.

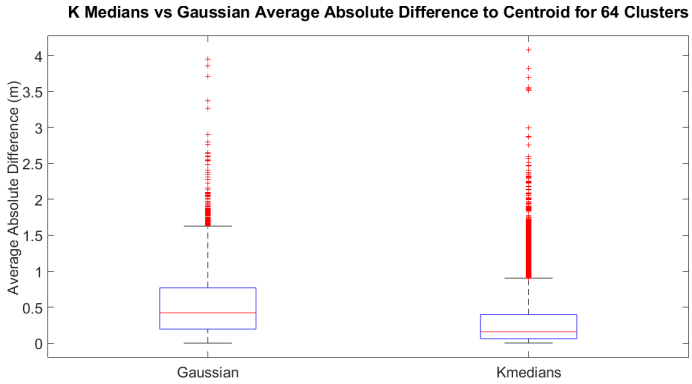


Figure 4.4: A comparison of the K-medians and GMM clustering algorithms when both are set to 64 clusters.

CLUSTER GROUP ANALYSIS

A visual inspection was also done for a set of the K-medians and GMM cluster groups. It is interesting to examine which profiles are grouped together, and to compare the spread of profiles within clusters groups. The standard deviation and maximum difference at each cross-shore point within the cluster group was also calculated and plotted to assess the fitting. An example of six of the cluster groups from the K-medians with 500 clusters can be seen in Figure 4.5 and Figure 4.6, and six of the Gaussian mixture cluster groups while using 64 clusters can be seen in Figure 4.7 and Figure 4.8. This is obviously not a viable direct comparison since the number of cluster groups is not the same, but 64 is the maximum number of clusters from the GMM method and therefore a comparison with the higher number of groups from K-medians is not possible.

The groups from K-medians are much tighter fitting, mainly due to having a higher number of clusters. The limitation of the Gaussian mixture to reach higher than 64 clusters restrains the method from providing similar results as the K-medians method.

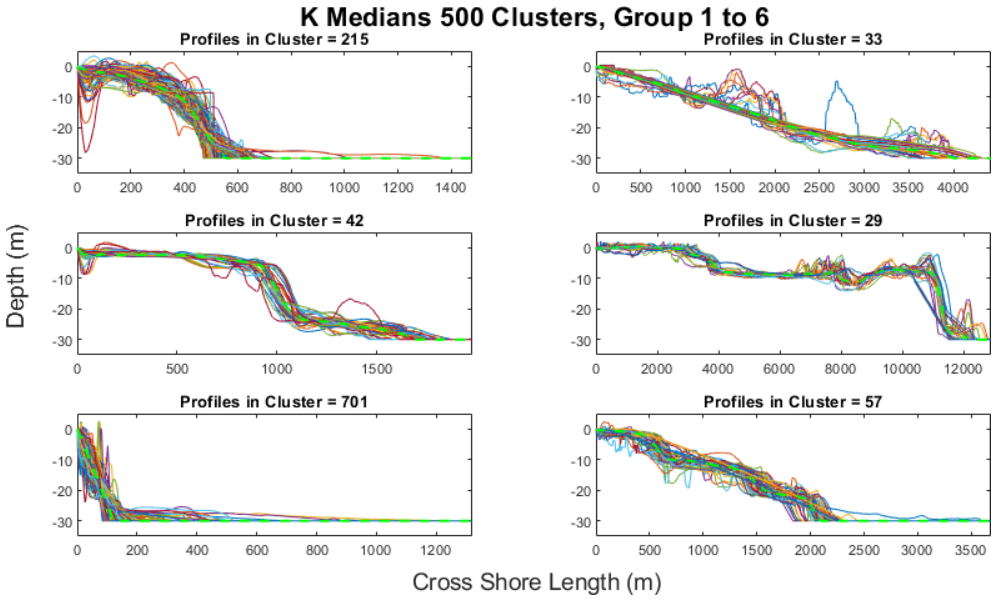


Figure 4.5: Example of 6 of the cluster groups from the 500 generated using K-medians.

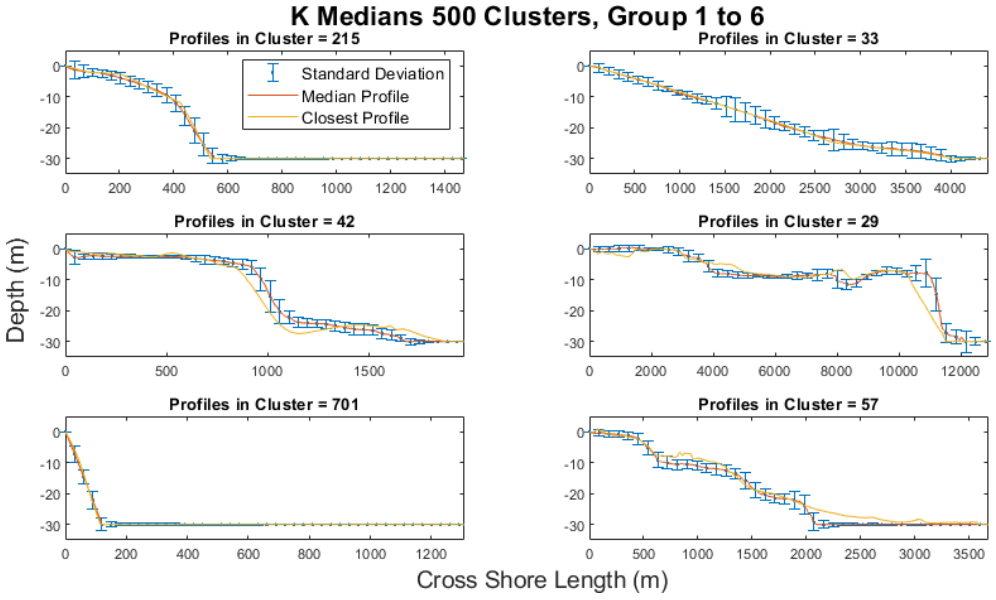


Figure 4.6: Example of 6 of the cluster groups from the 500 generated using K-medians, showing the cluster group statistics.

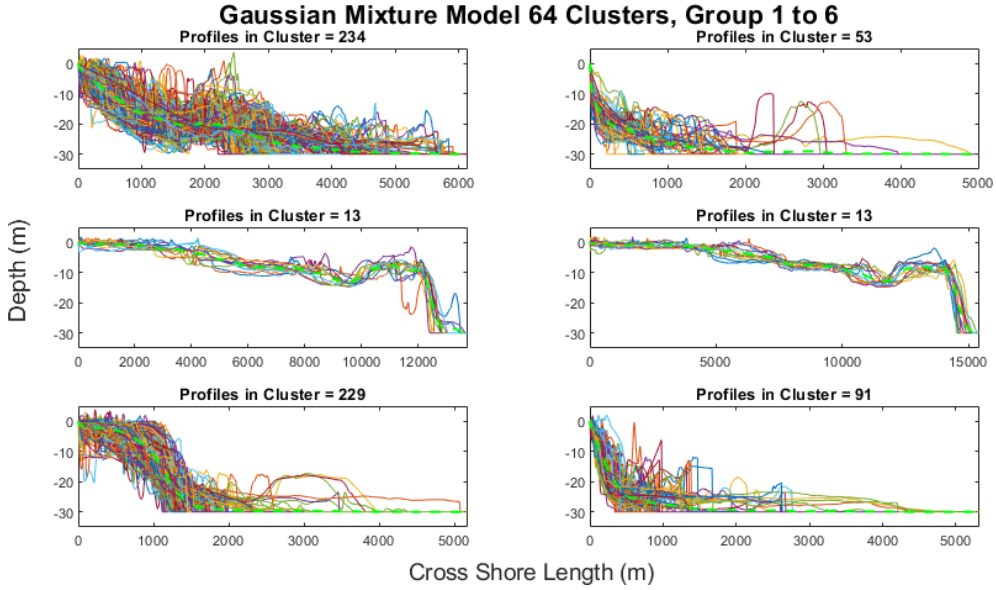


Figure 4.7: Example of 6 of the cluster groups from the 64 generated using Gaussian mixture.

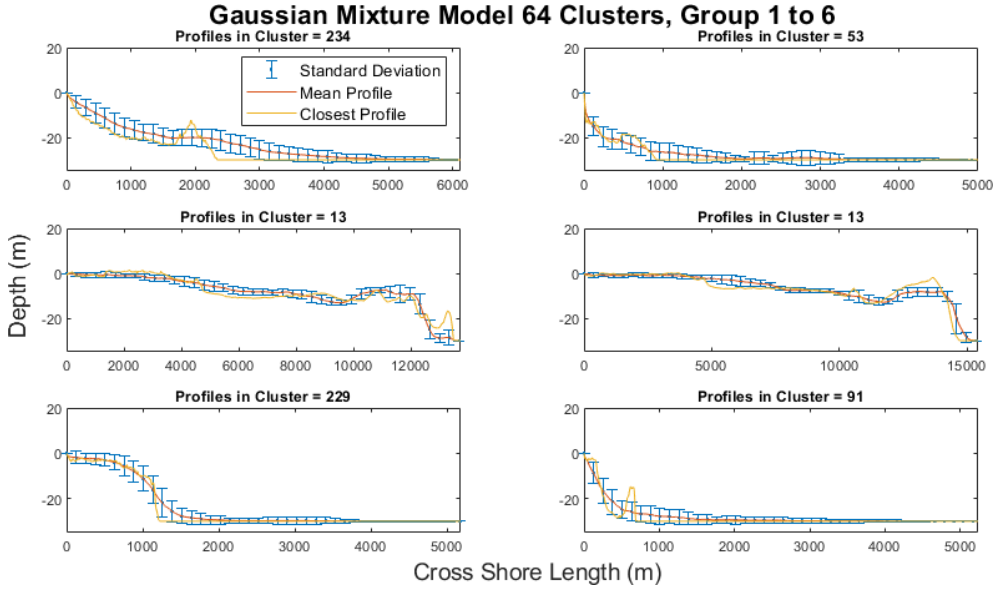


Figure 4.8: Example of 6 of the cluster groups from the 64 generated using Gaussian mixture, showing the cluster group statistics.

SELECTING THE METHOD AND NUMBER OF CLUSTERS FOR THE FINAL RUN

From the above mentioned analysis, K-medians was selected to be used for the input reduction of morphology. The error between the profiles and the centroids is the lowest of any of the methods, and the visual inspection also showed that K-medians resulted in tight cluster groups that should perform well for the remainder of the analysis. The appropriate final number of clusters requires a balance between a low number of clusters and high accuracy. This balance was found at 500 clusters since it is here that the average difference between profiles and the centroid plateaus, as shown in Figure 4.1 and 4.3, meaning that there is no longer a significant gain in error reduction when a higher number of cluster groups is selected, unless a much higher value is used. Keeping in mind the end goal of the project, a value of roughly 100 cluster profiles would be expected from the next round of cluster analysis, and therefore choosing a large number from this stage would make it more difficult to effectively cluster the profiles based on hydrodynamics. Therefore, 500 clusters was the selected number to move forward with.

4.1.3. INPUT REDUCTION FINAL RUN

For the final stage of the input reduction of morphology, two modifications were done to the analysis to enhance the output for the next steps. First, 50 replicates were used to ensure a successful grouping of the profiles, and second, a length limit was applied to restrict the profiles that could be included in the cluster analysis. Section 3.2.3 explains the methods used to set the length requirement.

With the addition of the length requirement, 20,454 reef profiles were included in the final cluster analysis (see Table 3.6). The 500 Cluster Round 1 profiles generated using the K-medians clustering algorithm are shown in Figure 4.9. The profiles are separated into four subplots for visualization purposes, as well as sorted by profile length.

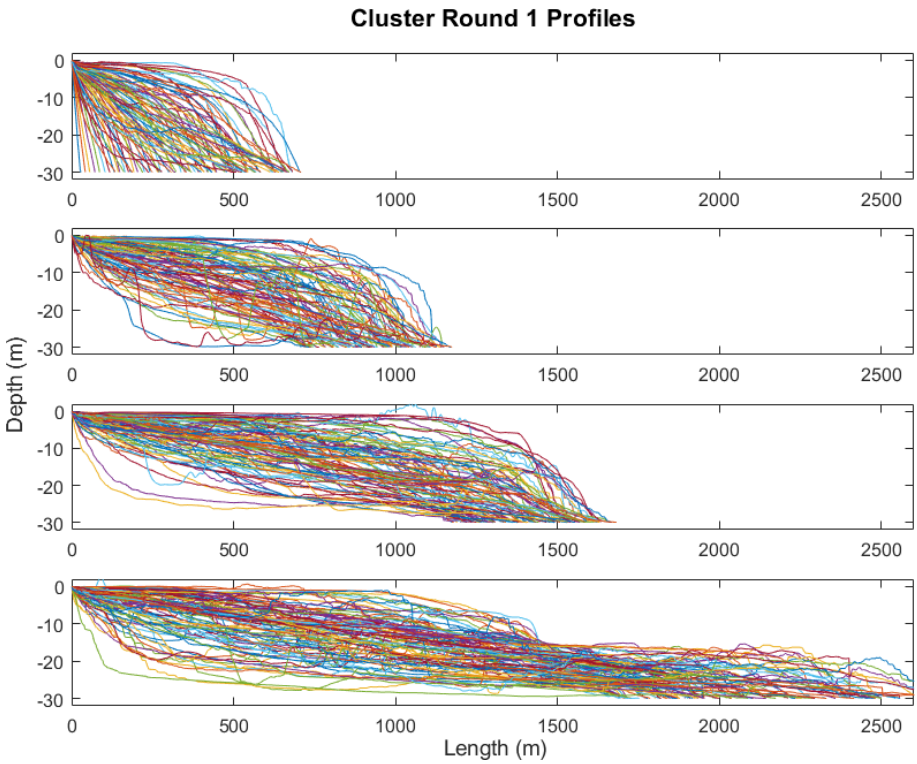


Figure 4.9: The 500 Cluster Round 1 profiles generated as the median of the cluster groups using the K-medians clustering algorithm.

4.2. XBEACH MODELLING OF CLUSTER PROFILES

XBeach simulations were done to determine the variability in hydrodynamics within the cluster groups, and to determine the hydrodynamics of the cluster profiles to be used in Step 5 (see Figure 3.1) for the cluster analysis of hydrodynamics (results presented in Section 4.3). To determine the variability within groups, five of the most dissimilar profiles within each cluster group were selected (Section 3.3.2) and simulated in XBeach to be able to do a direct comparison with the cluster profile and evaluate how well the cluster profiles represent the profiles within the group.

4.2.1. INTRA-CLUSTER VARIABILITY

The variability of runup within cluster groups was determined by comparing the XBeach results between the dissimilar profiles and the cluster profiles. For assessing the variability, XBeach model runs were done with loading conditions 3 and 4 (Hs = 7, steepness = 0.05 and 0.01, see Table 3.7). Only two loading conditions were used due to time constraints, and the two conditions with the largest wave heights were selected since they will most likely result in the largest variability of runup.

The maximum difference between the cluster profile and one of the most dissimilar

profiles was used to assess the spread in runup within each cluster. Figure 4.10 shows the spread of the maximum difference in $R_{2\%}$ between one of the most dissimilar profiles and the cluster centroid for each loading condition.

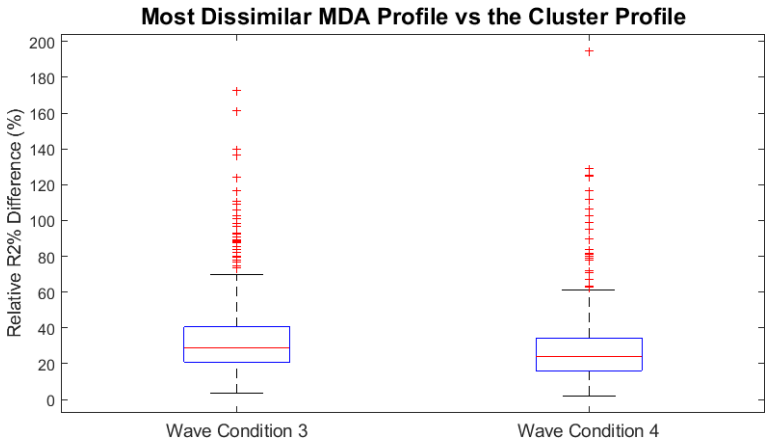


Figure 4.10: Maximum Difference of $R_{2\%}$ between the cluster centroid and one of the most dissimilar profiles for loading conditions 3 and 4.

The median of the largest difference in $R_{2\%}$ within the cluster groups is approximately 30% for both loading conditions. There are, however, a significant number of groups with a maximum relative difference greater than 100%. To understand what type of profiles result in such large differences, the groups above the 95th percentile in $R_{2\%}$ difference were analyzed further.

LARGEST RUNUP DIFFERENCES

Twelve cluster groups were in the highest 5% of error for both loading conditions. Figure 4.11 shows the centroid and the profile causing the high relative difference.

The most striking result is that ten of the twelve cases include a dissimilar profile that peaks above MSL, and the other two (as well as some of the others) are vastly different to the cluster profile in the nearshore area. These two features were investigated to understand what processes are occurring because of these features that lead to dissimilar wave runup.

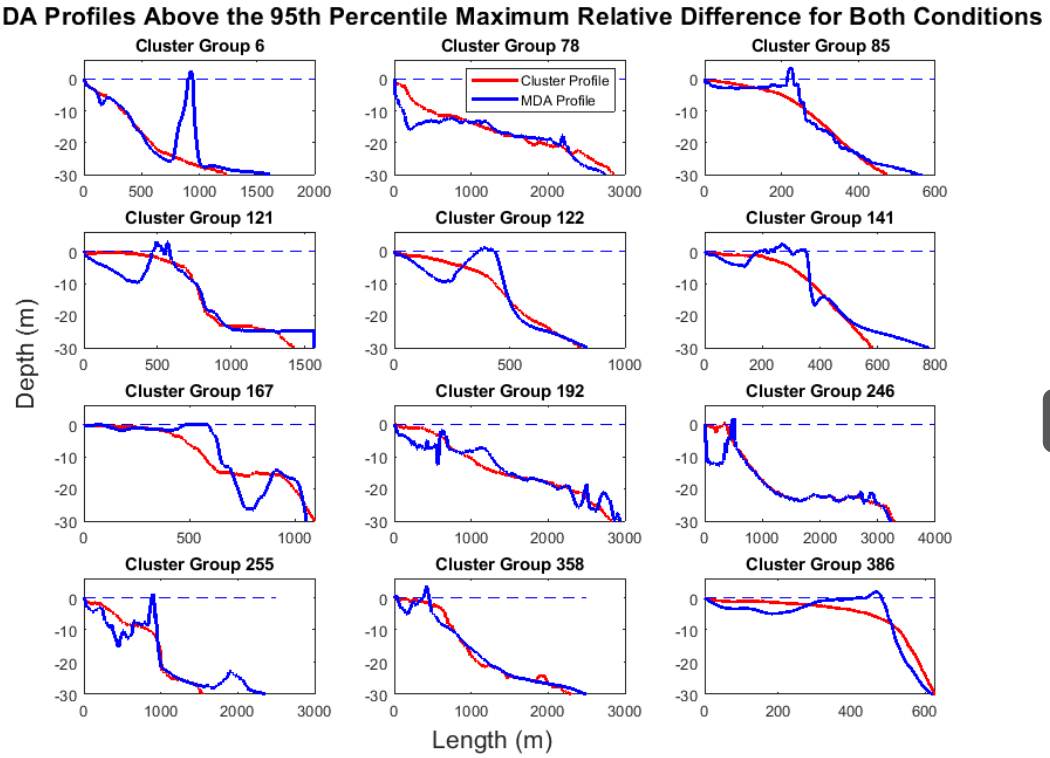


Figure 4.11: The cluster groups that are in the top 5% for maximum difference to centroid in $R_{2\%}$ for both loading conditions, showing the cluster profile and the dissimilar profile with the large difference in $R_{2\%}$.

4

ANALYSIS OF HYDRODYNAMICS

Peak Above MSL

The dissimilar profile causing the largest differences in wave runup in cluster group 6 (Figure 4.11) was selected to represent the scenario when a profile has a peak above MSL. The profile was compared to its cluster centroid under the wave loading condition 3, representing large wind waves. The results are shown in Figure 4.12.

As shown in subplot (d), the wave heights diverge between the two profiles at the peak of the dissimilar profile. The peak causes wave breaking, which the centroid profile would not have experienced yet. This is also demonstrated in subplot (b), where the spectrum for the cluster profile is still completely dominated by incident wave energy and the dissimilar profile has very little wave energy remaining, but is also higher in the lower frequencies. This difference carries forward towards the shoreline where the spectra for the dissimilar profile is now heavily dominated by infragravity (IG) waves, and the centroid follows a similar relationship but with much less IG dominance.

The setup across the profile is also very different. The peak causes the water to pile up on the landward side of it, causing higher setup all the way from the peak to the shoreline. This is also the main contributor to the wave runup as shown in subplot (e).

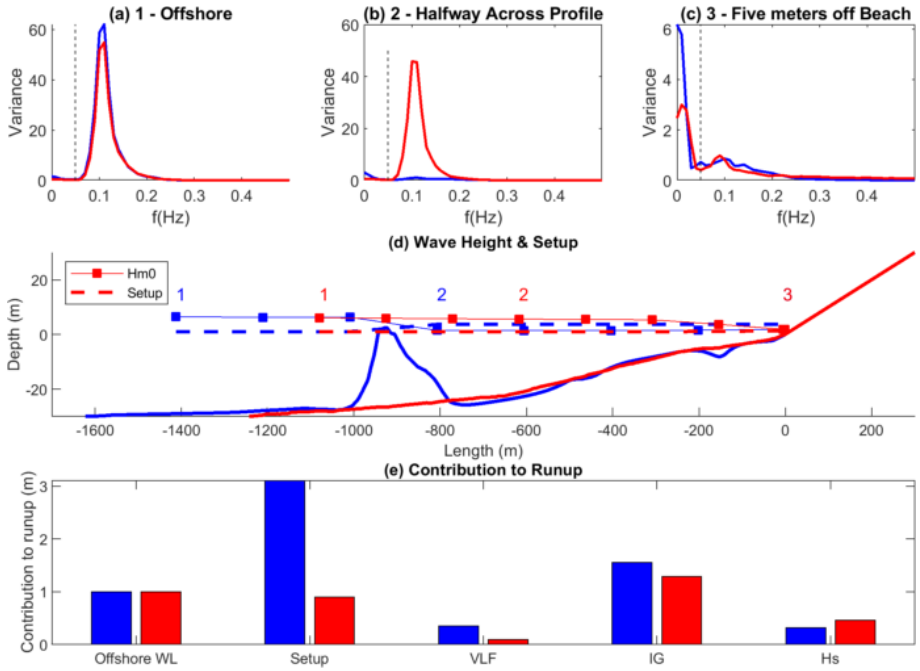


Figure 4.12: An analysis of the hydrodynamics of a dissimilar profile to the cluster centroid due to a large peak in the profile. The results are shown for loading condition 3, representing large wind waves. The dissimilar profile is shown in blue and the cluster centroid is shown in red. The wave spectra at three locations are shown in subplots (a-c). The locations are labeled with a number in the title that matches the location shown along the profile in subplot (d). Subplot (d) shows the two profiles, with the setup and the significant wave height across the profile. Subplot (e) shows the contribution to runup from five different components.

Large Nearshore Differences

Cluster group 78, shown in Figure 4.11, was selected to represent the scenario when a profile is similar to the cluster profile except in the nearshore. Wave loading condition 3 was again used for the analysis, and the results are shown in Figure 4.13.

Here we see that the profiles are extremely similar in length and shape right up until 500 m offshore at around -10 m depth, where the dissimilar profile drops off and remains deeper, resulting in a steeper slope leading to the coastline. The main difference that this type of dissimilarity seems to cause is difference in the wave energy at the nearshore. Subplot (c) shows that the wave spectra at 5 m from the shoreline is very different between the two profiles. The dissimilar profile still has almost all of its energy in the incident frequency band, whereas the cluster centroid, which would have had more wave breaking due to the shallower depths and more gradual nearshore slope, shows that most of the wave energy is now in the IG frequency band.

The effect of wave breaking is also shown in subplot (d) where the cluster centroid significant wave height drops considerably at the last point from the coastline, whereas for the dissimilar profile, the significant wave height is maintained and even grows slightly. This results in differences in the contributions to the wave runup, with the dissimilar

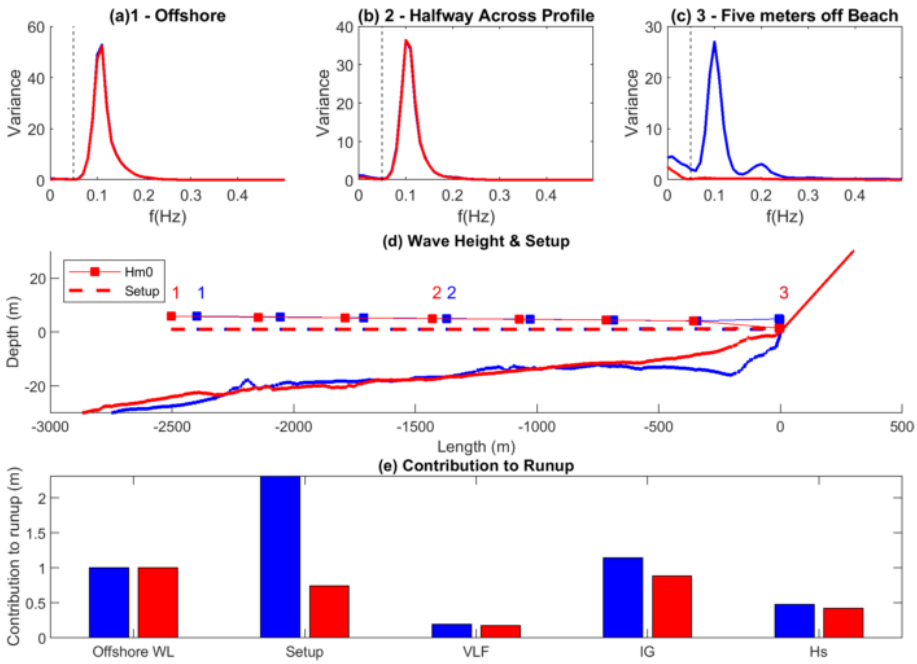


Figure 4.13: An analysis of the hydrodynamics of a dissimilar profile to the cluster centroid due to large nearshore differences. The results are shown for loading condition 3, representing large wind waves. The dissimilar profile is shown in blue and the cluster centroid is shown in red. The wave spectra at three locations are shown in subplots (a-c). The locations are labeled with a number in the title that matches the location shown along the profile in subplot (d). Subplot (d) shows the two profiles, with the setup and the significant wave height across the profile. Subplot (e) shows the contribution to runup from five different components.

profile having more of an influence from setup, most likely caused by a surge due to waves not being dissipated until very close to shore.

A profile that is deeper than the cluster centroid in the nearshore, and with a higher wave runup has been selected here as an example of differences in the nearshore. Conversely, there are examples of profiles with large differences in the nearshore in which the dissimilar profile becomes shallower than the centroid and the runup is less. This is due to the same relationship analyzed above, in which the shallower profile results in wave breaking and less energy reaching the shoreline. This result could be considered for future cluster analyses if a conservative estimate is required, where an extra limitation or input into the algorithm could include selecting a centroid that is deeper in the nearshore to ensure conservative estimates. However, the nearshore differences can also be removed by applying greater weight to the nearshore profile in the clustering algorithms, provided as a recommendation in Section 6.

PROFILE FEATURES THAT LEAD TO INEFFECTIVE GROUPING

The analysis presented above shows the two main features that lead to dissimilar wave runup of grouped profiles. Since the cluster analysis was based on the full profile, if the profiles are similar for the vast majority, they will most likely be grouped together. This

allows room for profiles with a sudden peak or sudden nearshore variation to be grouped with profiles that don't share that feature. To reduce the differences between grouped profiles, these two features should be focused on. This finding was included in Step 7 of the methodology (Figure 3.1) for matching observed profiles to the cluster profiles, and is stated as a recommendation in Section 6.2. Further analysis of profile features leading to similar and dissimilar wave runup is provided in Appendix E.

4.3. CLUSTER ANALYSIS OF HYDRODYNAMICS

The cluster analysis of hydrodynamics was done to merge the groups formed from the first round of cluster analysis that have similar wave runup results. This analysis included both morphology and the XBeach runup results as input for the clustering algorithm (see Section 3.4.1). The agglomerative hierarchical clustering algorithm was used with specified cutoff values that limit the intra-cluster variance. A large spread of cutoff values were used, resulting in a range of final numbers of cluster groups.

4.3.1. HIERARCHICAL CLUSTERING RESULTS

NUMBER OF CLUSTER ROUND 2 GROUPS

The agglomerative hierarchical clustering merged the Cluster Round 1 profiles into new groups, referred to as the Cluster Round 2 groups. Figure 4.14 shows the relationship between the cutoff value and resultant number of cluster groups formed, which ranges between 311 and 45.

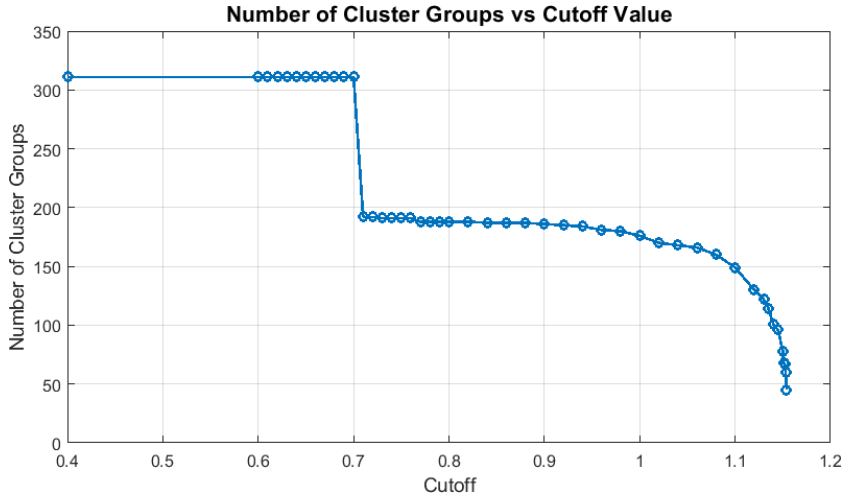


Figure 4.14: The relationship between the cutoff value and the number of Cluster Round 2 groups.

The dendrogram representing the hierarchical clustering is shown in Figure 4.15. The different colors show which leafs of the dendrogram are merged together up to a height of 0.5.

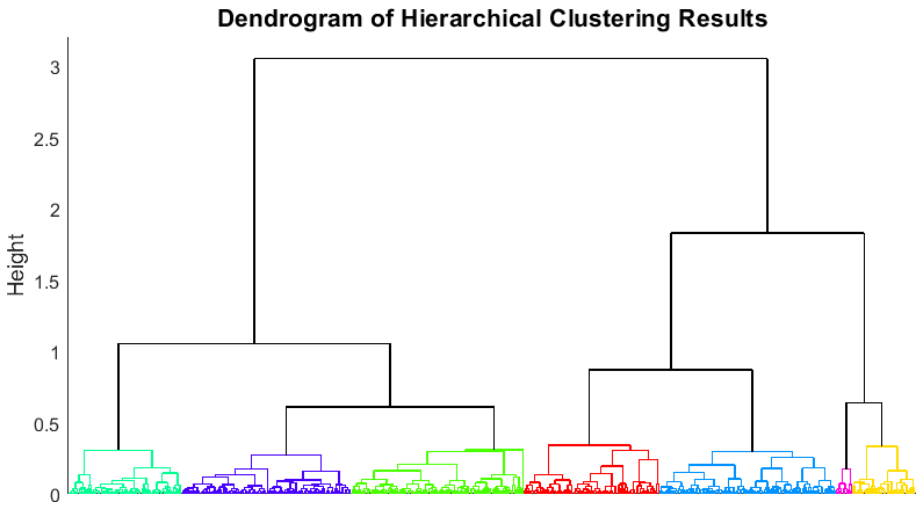


Figure 4.15: The dendrogram showing how the Cluster Round 1 profiles are grouped in the hierarchical clustering process during Cluster Round 2.

AVERAGE RUNUP ERROR WITHIN CLUSTER ROUND 2 GROUPS

The runup error is the greatest relative difference in $R_{2\%}$ between the selected representative profile of the Cluster Round 2 groups and the most dissimilar cluster profile within the group. This process is explained further in Section 3.4.2. The difference from the four loading conditions are averaged to obtain one value of error for each cluster group.

Figure 4.16 shows the distribution of the error via boxplots. The boxplots show that the median errors within the Cluster Round 2 groups are quite low, ranging from roughly 5% to 10%.

The errors within the 311 final cluster groups are the lowest, with the median error at approximately 5%. The next largest group of clusters is 192, a significant jump from 311. The large jump results in a noticeable increase in median error, to approximately 7%. This value is held relatively constant to the 160 cluster groups. From 160 to 45, the error consistently increases, and the spread between the 25th and 75th percentile does the same.

The maximum outlier is consistent for many of the boxplots. This shows that there is one group that is merged in each of these different scenarios that continuously results in the maximum error.

4.3.2. FINAL CLUSTER PROFILES

As an example of what the final cluster profiles could look like, Figure 4.17 presents the set of cluster profiles using a cutoff value of 1.145, which results in 96 different cluster profiles.

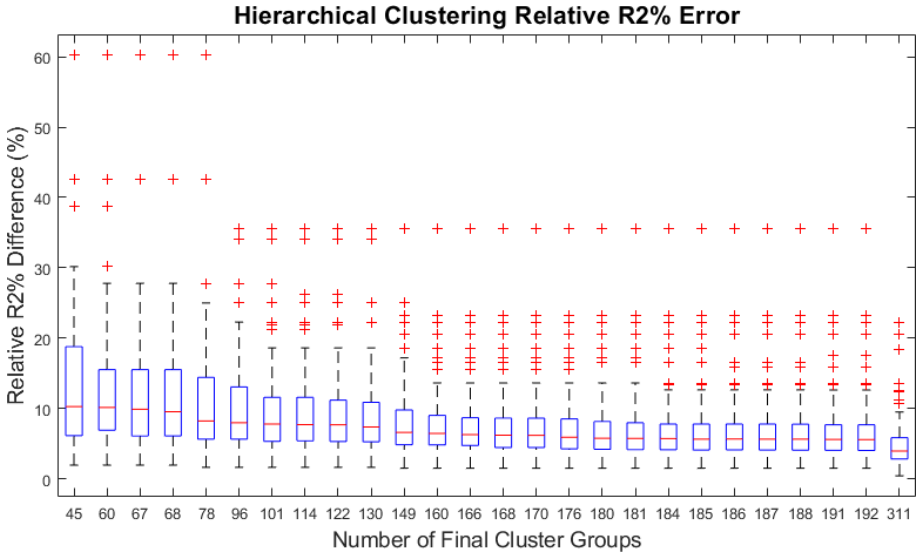


Figure 4.16: Boxplots for each set of Cluster Round 2 groups, representing the distribution of mean $R_{2\%}$ error within cluster groups.

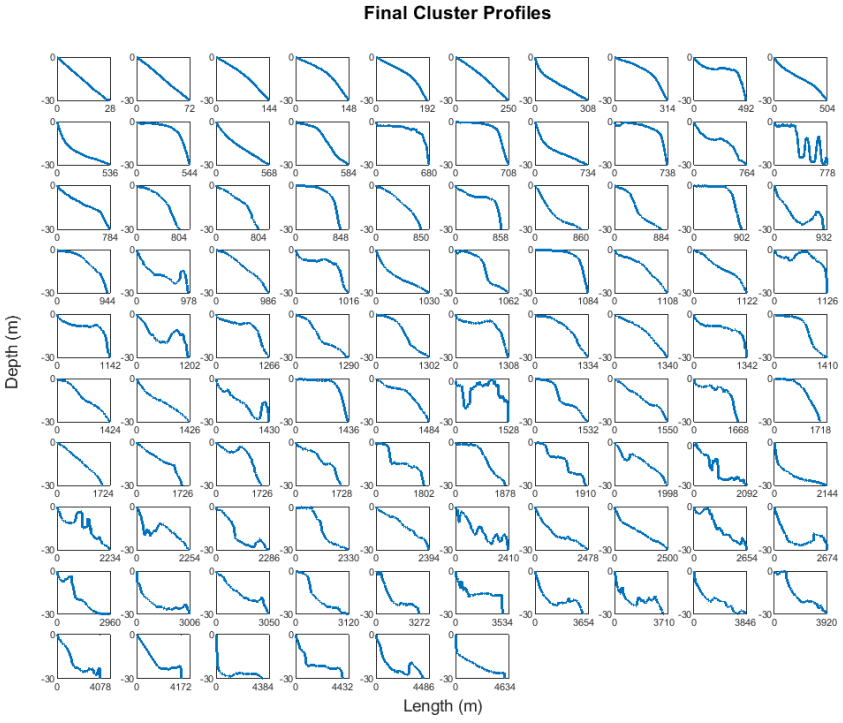


Figure 4.17: An example of a set of Cluster Round 2 profiles. Shown here are the cluster profiles selected while using a cutoff value of 1.145, resulting in 96 cluster profiles.

4.3.3. GROUPED PROFILE SIMILARITIES

Much research has been conducted to understand hydrodynamic response over coral reefs (Cheriton et al., 2016; Costa et al., 2016; De Ridder, 2018; Gourlay, 1994; Lashley, Roelvink, van Dongeren, Buckley, & Lowe, 2018; Massel & Gourlay, 2000; Pearson et al., 2017; van Dongeren et al., 2013; Young, 1989), and which parameters of the reef that effect wave runup has been determined. This study focuses on a different problem, which is grouping the profiles that have similar wave runup response, and in doing so, determining which features of the profile are important for grouping, and equally, determining which features are not of importance.

In Cluster Round 2, the cluster profiles are grouped with 50% weighting on full profile morphology and 50% weighting on wave runup. By analyzing which profiles get grouped together, an idea of the profile features that are important for similar wave runup can be determined. Figure 4.18 shows an example of a 15 Cluster Round 2 groups, when the cutoff value is set to 1.08 and there are 160 cluster groups. These 15 clusters were selected since they represent characteristic groups well.

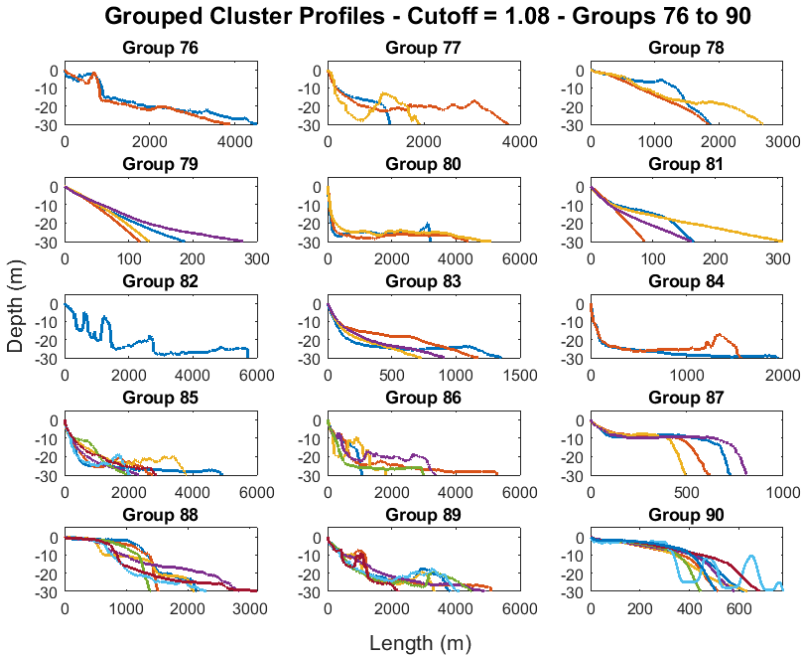


Figure 4.18: An example of 15 Cluster Round 2 groups when using a cutoff value of 1.08 that results in 160 cluster groups.

The profiles that are matched together in Cluster Round 2 can be seen to either be very similar along the entire profile (for example group 76, 80 and 84), or very similar in shallower depths. Group 81, 87, 88 and 90 are great examples, where the shallow water profile is extremely similar, and at a certain point they all break off into different shapes and to different lengths. Even Group 88 shows the profiles end up as either concave or convex, but since they are similar in the shallow depths, wave runup is similar and they

are grouped together.

This result is not unexpected since waves are known to be influenced by the bathymetry. First, the waves will begin to 'feel' effects from the bottom at a certain depth ratio to wavelength, followed by wave transformation across the reef that is also heavily influenced by depth (Quataert et al., 2015; Young, 1989). The profiles with similar nearshore shape are therefore grouped together since it is in the nearshore that the waves really start to feel effects from the profile, resulting in similar wave transformation over the reef and supplementary wave runup.

Figure 4.19 shows the relationship between grouped profile similarity and depth. The similarity is measured across depth bins, spaced every 5 m. The absolute difference is calculated between each profile to the mean profile of the cluster group, and then averaged within the depth bin. Four different examples of final cluster groups are shown, separated by the subplots.

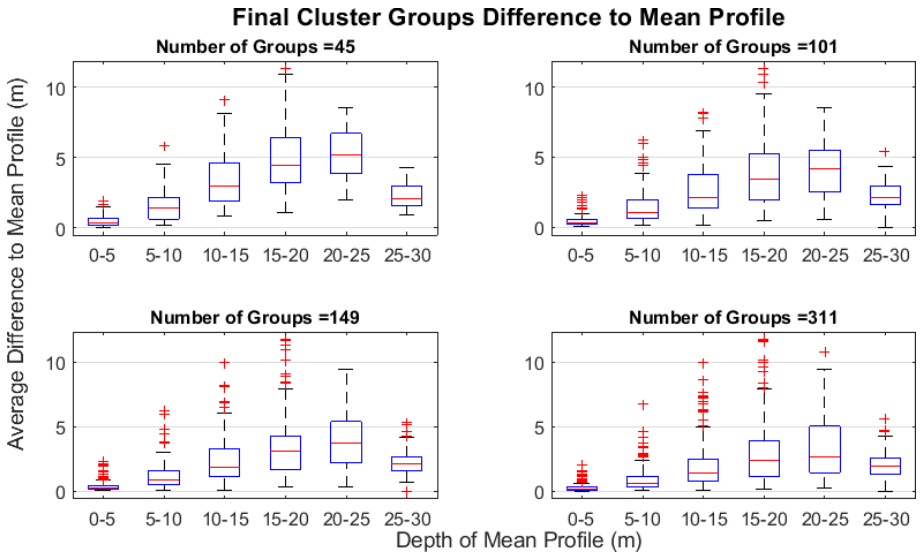


Figure 4.19: An analysis of the morphology differences between profiles matched together to form the Cluster Round 2 groups. It can be seen that the profiles that are matched together in Cluster Round 2 are much more closely related in the shallower depths.

The profiles that are grouped together through Cluster Round 2 are extremely similar in 0-5 m depth, and the dissimilarity grows with increasing depth. The average difference in the 25-30 m depth bin is rather low because the profiles are given a -30 m value to extend the profile in order for all profiles to be the same length, which will result in forced similarities.

ANALYSIS OF HYDRODYNAMICS

The same analysis as was done in Section 4.2.1 to assess features resulting in dissimilar wave runup is done here to evaluate the grouped Cluster Round 1 profiles with similar wave runup. Figure 4.20 shows the XBeach output for the three profiles grouped together

to form cluster group 77 (shown in Figure 4.18). This group was selected since the profiles within the group are quite dissimilar, varying drastically in length and deep water shape. The results are from loading condition 3, representing large wind waves.

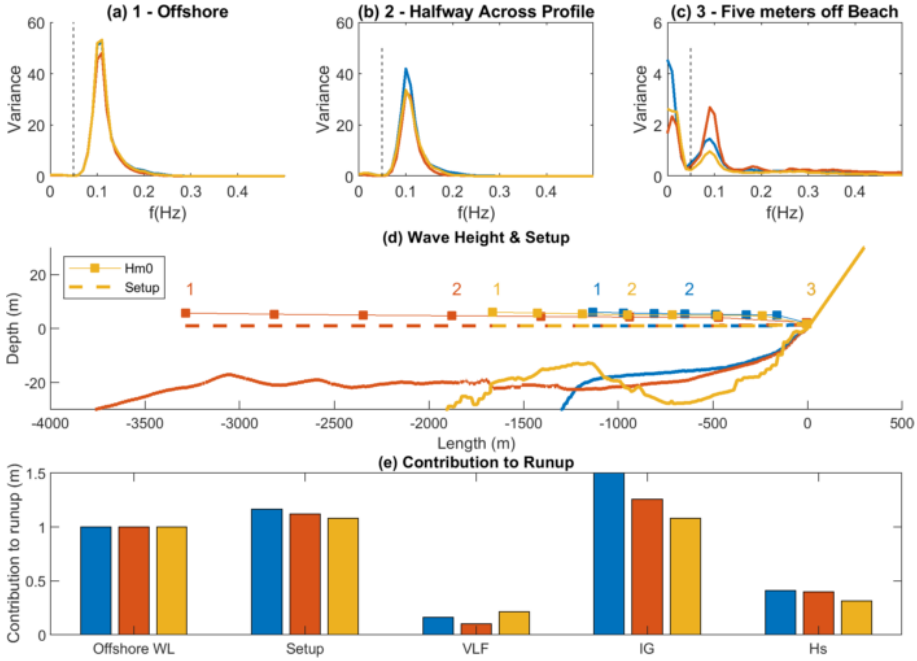


Figure 4.20: An analysis of the hydrodynamics of Cluster Round 2, group number 77, under wave loading condition 3. The three different colors used in all plots match the colors of the profiles shown in subplot (d). The wave spectra at three locations are shown in subplots (a-c). The locations are labeled with a number in the title that matches the location shown along the profile in subplot (d). Subplot (d) shows the three profiles in the cluster group, with the setup and the significant wave height across the profile. Subplot (e) shows the contribution to runup from 5 different components.

The three profiles result in similar wave characteristics along the entire profile. The spectra for the three profiles at each of the different locations all follow the same pattern and similar amounts of high frequency and low frequency components. The most notable difference between the profiles is the spike in infragravity spectra of the blue profile at the shoreline, displayed in Figure 4.20 subplot (c). This is caused by the steep fore-reef slope, resulting in breakpoint generated infragravity waves. The length of the profile, therefore, is another important feature for grouping profiles with similar wave runup, since similar lengths will help to ensure that the fore-reef slopes may not differ so significantly.

However, the significant wave height and wave setup is almost identical for all three profiles, shown in subplot(d). The contributions to the wave runup from the five different sources is also very similar, only showing minor differences between the profiles, most notably for the blue profile infragravity component, but all with almost equal amounts.

This example was selected because it is one of the Cluster Round 2 groups with the greatest variance in profile shapes, and therefore most other groups would likely demonstrate even greater similarities in hydrodynamics. This proves that by clustering profiles based on morphology and wave runup values, the grouped profiles have similar wave transformation across the entire reef, which ultimately leads to similar wave runup for many different wave loading conditions.

4.4. TESTING THE APPLICATION OF THE METHOD

Once the cluster profiles were finalized, their accuracy in predicting wave runup of natural reef profiles could be tested. A random set of 1000 reef profiles from the dataset were selected and run in XBeach with the same loading conditions that the cluster profiles had been subject to. No profiles with peaks above MSL were included as test profiles since the peaks have been determined to cause discrepancies that are not within the scope of this study. Also, the peaks are present only because of the method to select the coastline position and may not truly be offshore reef features. The test profiles were then matched to the appropriate Cluster Round 2 profiles, where the runup results were compared to determine how accurately the cluster profile represents the test profile (explained in Section 3.5).

The testing of the application was done with the 45, 101, 149 and 311 final cluster groups. These groups were selected because they span the complete range, as well as represent different levels of accuracy, as seen in Figure 4.16.

4.4.1. COMPARISON OF THE MATCHING METHODS

The different methods to match the test profile to the Cluster Round 2 profiles are explained in Section 3.5.1. It is important to remember that the matching of the test profile to cluster profiles is initially done with the Cluster Round 1 profiles. The association of the Cluster Round 1 profile in the Cluster Round 2 groups then allocates the test profile to the Cluster Round 2 profile. This is further explained in Section 3.5.1. The three matching methods include:

- CR1 - matching to the most similar Cluster Round 1 profile, based on the full profile morphology.
- NS3 - matching to the most similar Cluster Round 1 profile based on the morphology from 0 to -15 m depth. Cluster profiles are only considered for a match if the full length difference is less than 500 m. Also, if the test profile has a peak above MSL, the cluster profile must also, and vise versa.
- Probabilistic - using the softmax function to establish probabilities of matching to each cluster profile. The distance input to the softmax function is calculated from the NS3 method.

Figure 4.21 shows an example of nine test profiles matched to the Cluster Round 2 profiles by the different matching methods. The plots show the test profile, the Cluster Round 1 profile that the test profile was part of from the first cluster analysis, and the Cluster Round 2 profile based on the CR1 and NS3 matching methods. There is no profile

to represent the probabilistic matching since multiple profiles are incorporated in this method.

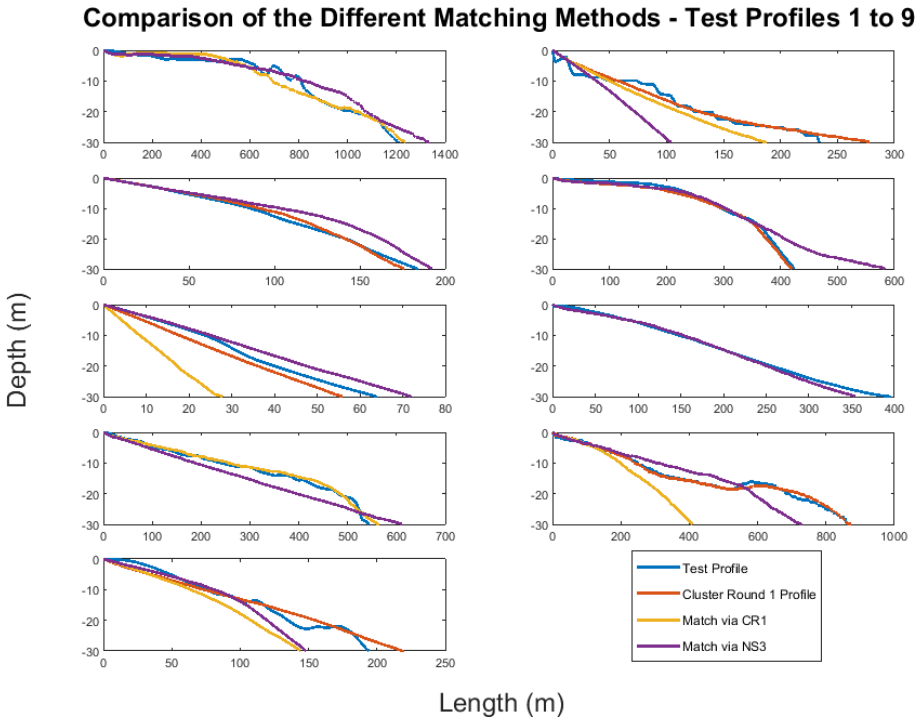


Figure 4.21: An example of 9 test profiles with their Cluster Round 1 profile, as well as the Cluster Round 2 profiles based on the NS3 and CR1 matching methods.

The different matching methods result in different accuracy of predicted wave runup. Figure 4.22 compares the mean $R_{2\%}$ relative difference between the 1000 test profiles and the Cluster Round 2 profiles while using the different matching methods. For all instances, the probabilistic matching method using the NS3 distance produces lower average errors.

Figure 4.23 shows the relative difference between the test profiles and the Cluster Round 2 profiles in violin plots. Violin plots are explained in Section C.1, but essentially show the distribution of data points while highlighting the mean and median. Here, it can be seen that the probabilistic matching method has the lowest maximum error of the three methods, and the greatest density of profiles with low errors.

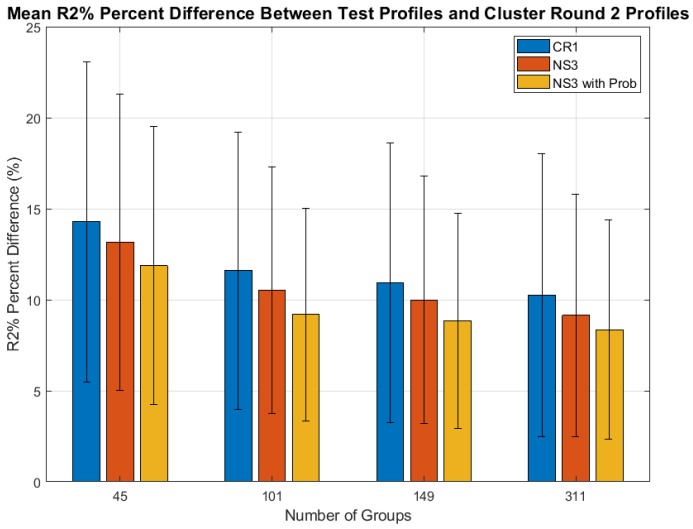


Figure 4.22: The mean relative difference in $R_{2\%}$ between the 1000 test profiles and the matched cluster profiles for four different cases of Cluster Round 2 groups. CR1 stands for the matching method based on the full profile, NS3 stands for the matching method using the nearshore depths, and NS3 with Prob refers to using the NS3 method to measure distances between profiles but applying the probabilistic matching technique. The error bars show the standard deviation.

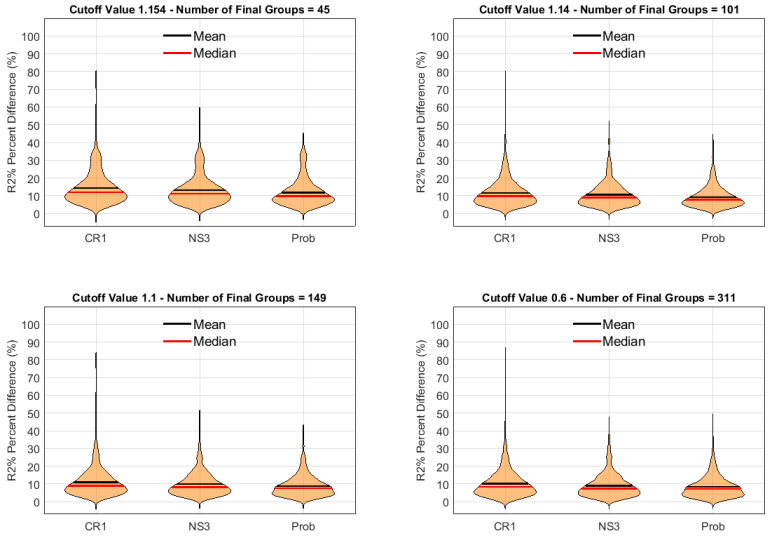


Figure 4.23: Distributions of the $R_{2\%}$ relative difference between the 1000 test profiles and the Cluster Round 2 profiles for different matching methods. The violin plots show the distribution of the results, as well as the mean and median value.

4.4.2. LOADING CONDITION ANALYSIS

For these four test cases, the errors from each loading condition were also examined individually to assess the predictive accuracy of the cluster profiles for the different types of wave conditions. The results from the probabilistic matching method are shown. Figure 4.24 shows the relative difference in $R_{2\%}$ between the 1,000 test profiles and the matched Cluster Round 2 profiles, separated by subplot based on different number of cluster profiles.

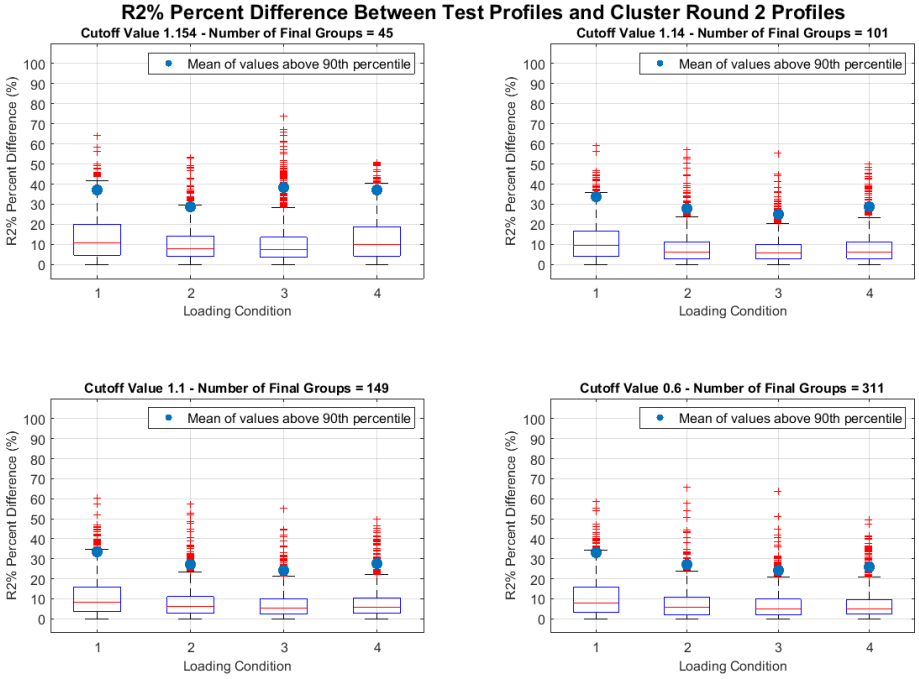


Figure 4.24: Boxplots of the $R_{2\%}$ relative difference between the 1000 test profiles and the matched Cluster Round 2 profiles for different number of cluster profiles. The test profiles were matched using the probabilistic method.

Loading condition 1 tends to have a slightly greater median relative difference, as well as a greater range between the 25th and 75th percentile values. This is because the relative difference from lower runup values caused by the smaller wave conditions in loading condition 1 will be more sensitive to differences. Loading condition 4 has the greatest outliers, since the most extreme waves cause the greatest discrepancies.

The blue dot on the boxplots shows the mean of all values above the 90th percentile. This was plotted to gain a better representation of the highest errors for each condition, rather than simply using the maximum value that could be a single outlier. This value is at around 40% for the case with 45 final profiles (top left), 25% for the case with 311 final profiles (bottom right), and in between for the other two cases. Although the median errors only differ slightly, the maximum errors do have quite a significant difference between the different number of cluster groups.

4.4.3. ACCURACY OF CLUSTER PROFILE PREDICTION

Figures 4.22 and 4.23 demonstrate the key finding that the cluster profiles are capable of predicting the $R_{2\%}$ of the 1000 test profiles with a mean relative difference of 8.4% when using 311 cluster profiles, or 11.9% when using 45 cluster profiles. The most accurate predictions were obtained while using the probabilistic matching method. This is a very exciting result, since in wave runup, a 10% error in prediction is highly acceptable. The empirical equation of $R_{2\%}$ for natural beaches, developed by Stockdon et al. (2006), is used by engineers around the world. From all of the data used to generate the equation, the mean difference between observed values and predicted values was -17 cm, and the mean runup elevation was 144 cm. This is equivalent to a mean relative difference of 11.8%. Therefore, the accuracy of the cluster profiles is in line with current wave runup predictions and would provide the necessary information for all purposes that the cluster profiles could be used for, including an early warning system and climate change impact analysis.

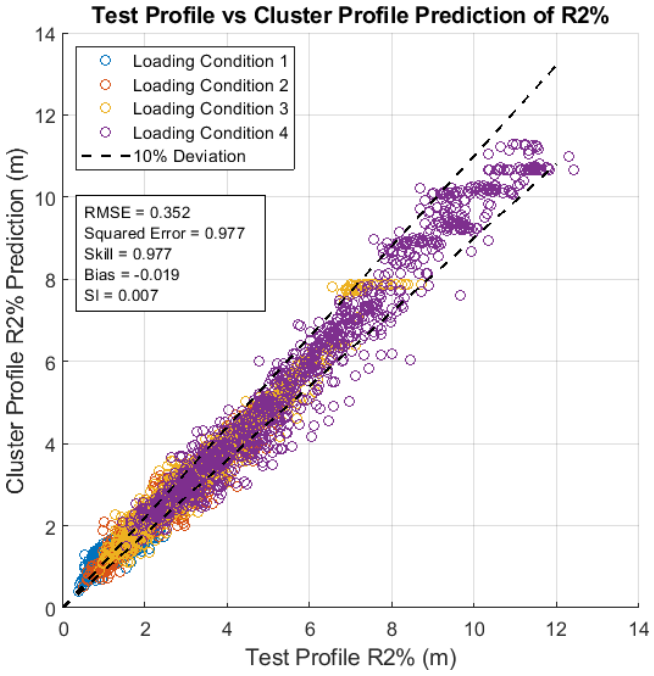


Figure 4.25: The 1000 test profiles vs the cluster profiles prediction of $R_{2\%}$, when using 149 Cluster Round 2 profiles and the probabilistic matching method. Each of the four loading condition results is matched individually, resulting in 4000 data points being compared.

Figure 4.25 presents the cluster profile prediction vs the test profile XBeach result when using the set of 149 Cluster Round 2 profiles. The different colors separate the results from the four different wave loading conditions. The linear relationship shows that the cluster profile prediction is highly accurate. The black dotted lines represent a 10% deviation. The majority of data points that lie outside of the 10% deviation are at low runup values. This is because percent differences at lower values are greater compared

to the same difference at a higher value. The fact that the greatest deviation is at lower wave runup is promising since the main use of the cluster profiles is for an EWS, which would be activated for large waves and high predicted runup. The linear relationship even at high runup values shows great promise for the cluster profiles. The points with the greatest runup differences are however from loading condition 4, in which the cluster profile under predicts the wave runup. Further analysis into these test profiles and the profile features that lead to this result is necessary.

4.5. VALIDATION

To validate the cluster profile results, two reef profiles from Roi Namur were simulated in XBeach under the same four loading conditions as the cluster profiles. They were then matched and the wave runup results were compared to the appropriate cluster profiles. The set of 149 Cluster Round 2 profiles was selected, which is one of the four that was tested, and it is in the middle of the range of potential Cluster Round 2 profiles. The NS3 and the probabilistic matching method were used. This section presents the validation results separated by the different matching methods. Further details of the matching methods and runup results for the validation are in Appendix F

4.5.1. NS3 MATCH

Figure 4.26 in subplot (a) and (b) shows the two Roi Namur profiles and the cluster profile that they match to. In subplot (c) and (d), the wave runup is plotted for each of the four loading conditions. The NS3 matching method, explained in Section 3.5.1, matches the profile to the Cluster Round 1 profile that is most similar in the nearshore depths (0 to -15 m), while ensuring that the lengths of the profiles do not differ by more than 500 m in length. Since the profile is matched to the Cluster Round 1 profile, the Cluster Round 2 profile does not have to meet the 500 m length requirement. This is apparent in Figure 4.26 (a), in which the Cluster Round 2 profile is about 600 m longer than the Roi Namur profile. .

Table 4.1 presents the $R_{2\%}$ comparison. Both profiles are well-represented by the cluster profiles, demonstrated by the average relative difference in $R_{2\%}$ of 16.4% and 10.6%. Roi Namur profile 1, however, is visibly more different to its matched cluster profile compared to Roi Namur profile 2, and the accuracy of the $R_{2\%}$ estimation is slightly lower as a result. The largest absolute difference in $R_{2\%}$ occurs at loading condition 4, which represents the largest waves. The cluster profiles underestimate the runup for both profiles. This is most likely because each of the cluster profiles are longer than the Roi Namur profiles, and the large waves would break further offshore on the cluster profile while also being subject to greater frictional dissipation.

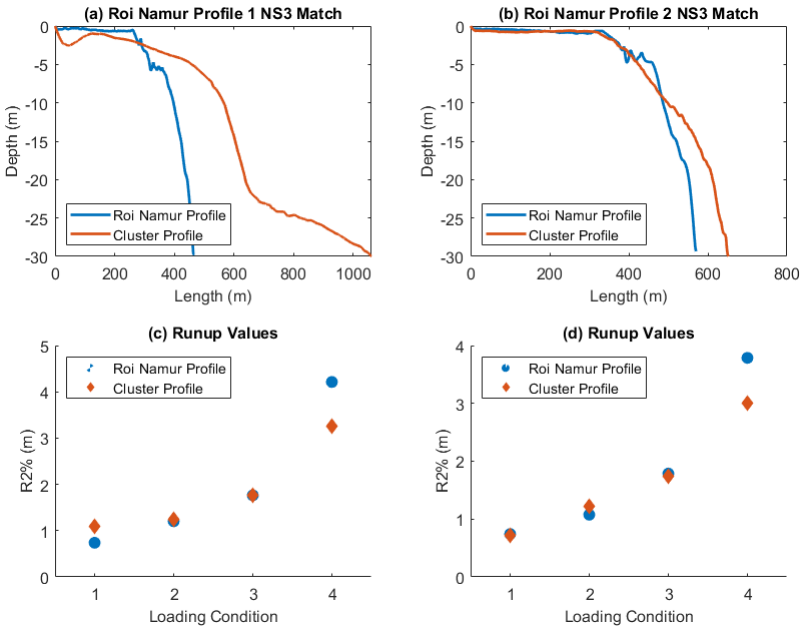


Figure 4.26: The two Roi Namur profiles, matched to the cluster profiles using the NS3 method. Subplot (a) and (b) show the reef profiles, and subplot (c) and (d) compare the $R_{2\%}$ values.

Table 4.1: Roi Namur profile validation results when using the NS3 match method. The relative difference in $R_{@\%}$ is compared for each loading condition, as well as averaged to determine the accuracy of the prediction.

Roi Namur Profile 1	**Loading Condition**	**Test Profile R2% (m)**	**Cluster Profile R2% (m)**	**Relative Difference (%)**
	1	0.74	1.09	32.3
	2	1.20	1.24	3.3
	3	1.76	1.75	0.6
	4	4.21	3.26	29.3
			Average R2% Error	**16.4 %**
Roi Namur Profile 2	**Loading Condition**	**Test Profile R2% (m)**	**Cluster Profile R2% (m)**	**Relative Difference (%)**
	1	0.74	0.72	2.3
	2	1.08	1.22	11.5
	3	1.79	1.74	2.6
	4	3.79	3.00	26.1
			Average R2% Error	**10.6 %**

4.5.2. PROBABILISTIC MATCH

The probabilistic matching method, explained in Section 3.5.1, essentially works by calculating the probability that a profile belongs to each cluster profile based on the distance between them. Therefore, a distance metric is required, and using the NS3 method to determine distance between profiles was applied since it provides the best results.

Figure 4.27 separates the two Roi Namur profiles by columns. Subplots (a) and (b) show the probabilities of belonging to each Cluster Round 2 profile. Subplot (c) and (d) plot the Roi Namur reef profiles as well as colored cluster profiles based on their associated probability, and subplot (e) and (f) compare the $R_{2\%}$ values for the four different loading conditions.

From the NS3 matching method, it was clear that the Roi Namur profile 1 did not match very well to its most similar cluster profile. Using the probabilistic approach, this results in no high probabilities, but rather a spread of lower probabilities because the profile is 'in-between' cluster groups. This is shown in subplot (a) and (c), where there are 5 cluster profiles assigned probabilities greater than 10%, but none greater than 25%. On the contrary, Roi Namur profile 2 did match quite well with its most similar cluster profile, resulting in a high probability of almost 60% to the closest match, with only one other cluster profile possessing greater than 10%.

Using the probabilistic approach has proven to generally provide the best results, but another benefit to using the probabilistic approach is the uncertainty that the method can provide. Compared to the direct matching methods such as NS3 where one cluster profile is used to determine the wave runup estimation, the probabilistic approach uses a weighted ensemble mean of multiple cluster profile results to calculate the runup estimation, and therefore an uncertainty associated with the result can be provided. In subplot (e) and (f) the weighted standard deviation, as well as the maximum and minimum runup value of cluster profiles with probabilities greater than 1% is provided as an example of such an uncertainty.

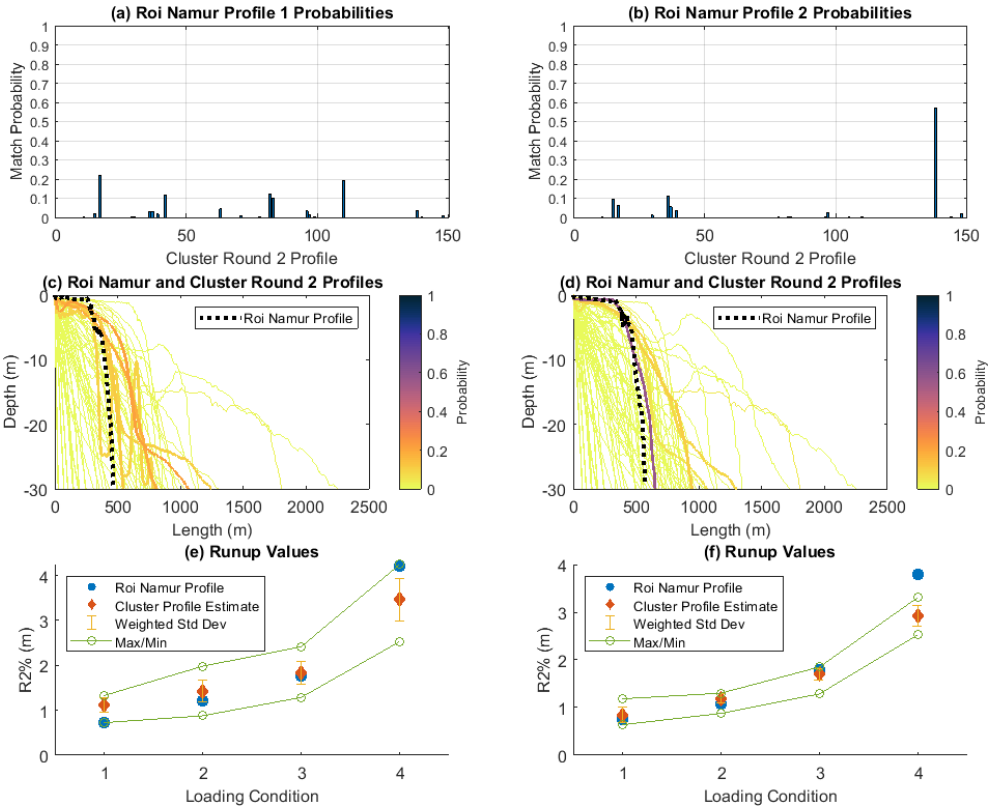


Figure 4.27: The two Roi Namur profiles, matched to the cluster profiles using the probabilistic match method. Subplot (a) and (b) show the match probabilities, and subplot(c) and (d) plot the reef profiles. All cluster profiles with match probabilities greater than 5% are plotted with a thicker line width. Subplot (e) and (f) compare the $R_{2\%}$ values.

The runup results are presented in Table 4.2. The mean relative difference between the cluster profiles estimate and Roi Namur profile 1 is 18.8%, and 14.2% for Roi Namur profile 2. These estimates are slightly worse compared to those obtained from the NS3 method. These profiles, therefore, are rare examples where the NS3 method produces better results compared to the probabilistic method. This is most likely due to the fact that most cluster profiles with significant probabilities, although similar in shape, are longer than the Roi Namur profile, therefore, almost all of the supplementary information from including more profiles in the runup estimate deters the estimate in one direction. In terms of wave runup, the profiles with significant probabilities have runup results that are either all higher or lower than the Roi Namur profile, instead of an even distribution on both sides which would bring the estimate closer to the true result.

The validation tests reveal that the cluster profiles can be used effectively for reef profiles from locations excluded from the analysis. The average difference in $R_{2\%}$ prediction from the cluster profiles for these two profiles is greater than the mean of the 1000 test

Table 4.2: Roi Namur profile validation results when using the probabilistic match method. The relative difference in $R_{@\%}$ is compared for each loading condition, as well as averaged to determine the accuracy of the prediction.

Roi Namur Profile 1	Loading Condition	Test Profile R2% (m)	Cluster Profile R2% (m)	Relative Difference (%)
	1	0.74	1.12	34.2
	2	1.20	1.43	15.7
	3	1.76	1.84	3.9
	4	4.21	3.47	21.4
			Average R2% Error	**18.8 %**
Roi Namur Profile 2	**Loading Condition**	**Test Profile R2% (m)**	**Cluster Profile R2% (m)**	**Relative Difference (%)**
	1	0.74	0.84	12.3
	2	1.08	1.19	9.7
	3	1.79	1.70	5.3
	4	3.79	2.93	29.4
			Average R2% Error	**14.2 %**

profiles previously evaluated, however the results are well within the expected range of accuracy.

4.6. XBEACH SIMULATIONS REDUCTION

The data reduction results in a reduced number of XBeach simulations required to effectively model the wide variety of coral reef morphology.

To estimate the number of XBeach simulations that would be required to effectively use the cluster profiles, it is assumed that the same parameters and values would need to be run as were done by Pearson et al. (2016) when creating BEWARE. Table 4.3 lists the parameters and values relevant from Pearson et al. (2016) that would be applied to the cluster profiles (removing the reef parameters used to create the schematized profile). The combinations of parameters requires 540 XBeach simulations per profile.

By reducing the dataset by two orders of magnitude, the required number of XBeach simulations is severely reduced, as pointed out in Table 4.4. After the first round of cluster analysis, when there are 500 cluster profiles, there is already a 97.6% reduction in simulations compared to the number of profiles that were included in the analysis.

Table 4.3: The values of the parameters used by Pearson, Reniers, van Dongeren, Tissier, and den Heijer (2016) to construct BEWARE. Using the same values would require a number of XBeach simulations equivalent to the number of cluster profiles multiplied by 540.

Symbol	Parameter	Units	Values
H_0	Wave Height	m	1, 2, 3, 4, 5
$\frac{H_0}{L_0}$	Wave Steepness	-	0.005, 0.01, 0.05
n_0	Offshore Water Level	m	-1, 0, 1, 2
c_f	Coefficient of Friction	-	0.01, 0.05, 0.10
B_{beach}	Beach Slope	-	1/5, 1/10, 1/20
	Number of Combinations		**540**

Table 4.4: Reduction in the required XBeach simulations as a result of the data reduction. 20,454 profiles is the number of profiles included in the analysis, signifying the result if no data reduction was done. The other rows resemble the effect of the data reduction, demonstrating significant reductions in XBeach simulations.

Stage of Analysis	Number of Profiles	Number of XBeach Runs	Percent Reduction
Initial dataset	20,454	11,045,160	
Cluster Round 1	500	270,000	97.6%
Cluster Round 2	311	167,940	98.5%
Cluster Round 2	45	24,300	99.8%

COMPARISON TO THE SCHEMATIZED PROFILE

A comparison between the number of XBeach runs for the cluster profiles and a set of schematized profiles is shown in Figure 4.28. The benefit of the cluster profiles is the wide range of morphology that they represent. For the schematized profiles to include a similar range, more parameters would need to be added. Therefore, for the comparison, the effect of adding more parameters to define the schematized profile was modelled. In red, the number of XBeach runs for a schematized profile starting with 1 parameter is shown. Each added parameter is modelled with five values associated with it. In blue, the current BEWARE schematized profile is modelled. It begins at 2 parameters since it currently is defined by the reef width (7 values) and fore-reef slope (3 values). Each added parameter also includes five values associated with it. In black, the number of XBeach runs for the cluster profiles is shown.

The current BEWARE system requires 11,340 XBeach simulations, with the two parameters. The lower bound of cluster profiles (45 profiles) requires 24,300. However, as Figure 4.28 shows, when one more parameter is added to the current BEWARE profile (with 5 values) the required number of XBeach simulations jumps to 56,700. The number of XBeach simulations for the cluster profiles follows a linear trend with increasing number of profiles, but for additional parameters for the schematized profile, each addition results in an exponential growth. Therefore, when trying to cover a greater variety of coral reef morphology, the cluster profile approach is much more beneficial.

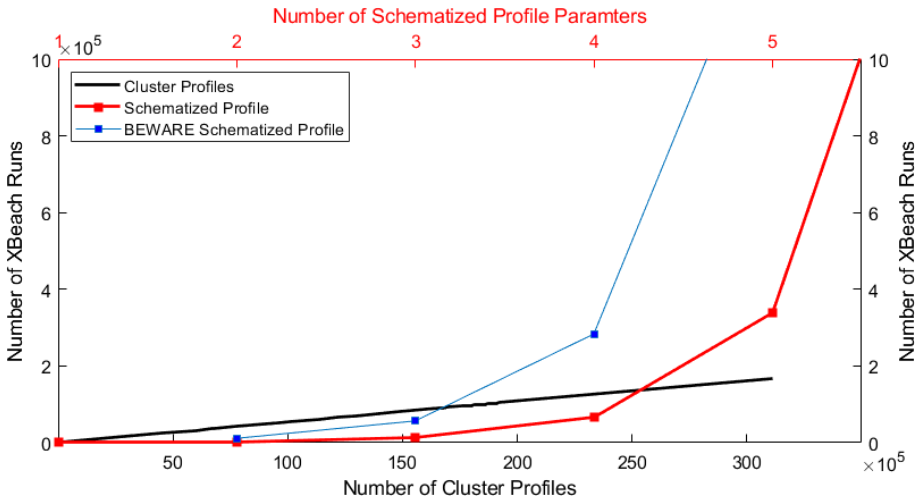


Figure 4.28: A comparison between the number of required XBeach simulations to create a BEWARE type model using a schematized profile and the cluster profiles. The schematized profiles relates to the top x-axis, comparing the increase in simulations with increasing parameters added to define the profile. The red line represents a schematized profile with 5 values for each added parameter. The blue line represents the current BEWARE setup, with 7 values for width and 3 values for fore-reef slope. Each added parameter is assumed to have 5 values. The bottom x-axis refers to the number of cluster profiles.

5

DISCUSSION

CHAPTER SUMMARY

This chapter includes a summary and the sensitivities of the data reduction process, a discussion regarding the applications of the cluster profiles and the methodology itself, ideas for next steps, and the main limitations of the findings. The key sensitivities of the methodology include the clustering algorithm selections, the number of cluster profiles, and the method to match the cluster profiles to observed reef profiles. The application of the cluster profiles for wave runup prediction is discussed in detail, as well as using the cluster profiles for climate change analysis and nature based solutions. The opportunities and changes required to use the methodology for other coastal environments are identified. Lastly, the next steps include updating the cluster profiles with the findings from this study before coupling them with a predictive tool such as BEWARE.

5.1. SUMMARY OF DATA REDUCTION

5.1.1. INPUT REDUCTION OF MORPHOLOGY

SELECTING THE CLUSTERING METHOD

For the input reduction of the reef profile dataset, multiple clustering algorithms were tested and compared in order to determine which could group the profiles with the least intra-cluster variance in morphology. The dataset was difficult to cluster as there are thousands of variables (cross-shore positions) for each observation (profile), requiring the algorithms to find similarities in many dimensions. The problems with high dimensionality result from the fact that a fixed set of data points becomes increasingly "sparse" as the dimensionality increase (Steinbach, Ertöz, & Kumar, 2004). The clustering algorithms operate around the degree of similarity between observations, but in such high dimensional space, the measure of similarity may become inadequate to properly form clusters. This situation refers to the 'curse of dimensionality', coined by Bellman (1961).

By testing different methods, however, a successful cluster analysis of the dataset was achieved. This was done using K-medians to form 500 cluster groups, which obtained a balance of low intra-cluster variance and a low number of cluster groups. This method may have worked best because the distance between the centroid and observation is computed by summing the distance of each individual variable. For a cross-shore profile, this seems to be a good approach, rather than using the squared euclidean distance computed in extremely high dimensional space.

CLUSTER ROUND 1 PROFILES

The input reduction resulted in 500 cluster groups of the initial 20,454 reef profiles included in the analysis. The median was determined to be the better representation of the cluster group compared to the mean, and did an adequate job in generating a smoothed out profile. A smoothed profile is required to remove the local disturbances. For example, if the cross-shore transect was taken only 10 m to the left or right, these small disturbances would most likely be different, but the main features (average slope of the profile, crest length, etc.) would most likely be the same or very similar.

There were 11 of the 500 cluster groups with only one profile, but these were deemed appropriate because these profiles are part of the dataset and therefore exist somewhere in the world and are representative of something, even if it is an odd and unique shape. Each profile included in the final set of cluster profiles, however, adds many hours of computation time, and so these single profile cluster groups could be removed if the use of including them is deemed not worth the supplementary computation time.

5.1.2. CLUSTER ANALYSIS OF HYDRODYNAMICS

The Cluster Round 1 profiles were then grouped together based on similar shape and wave runup. The agglomerative hierarchical clustering method was used, which takes more computation time compared to the clustering algorithms used in Cluster Round 1, but since the input is reduced time was not of concern. In this approach, the amount of grouping of the 500 Cluster Round 1 profiles varied. The output could include a range of 311 to 45 cluster groups depending on the cutoff value used. The more groups result in greater accuracy, but at the expense of more computational time. The output from the

full data reduction process is a set of cluster profiles that represent the entire dataset in terms of morphology and wave runup.

5.1.3. METHODOLOGY SENSITIVITIES

Although the results from this study are promising, each step of the methodology is sensitive to the decisions that were made. Having completed and tested the methodology with the formed cluster profiles, it is acknowledged that there is room for improvement and it is recommended that the cluster profiles be updated. The ideas for improvement are presented in Section 5.3. The methodology and the sensitivities are presented in Figure 5.1.

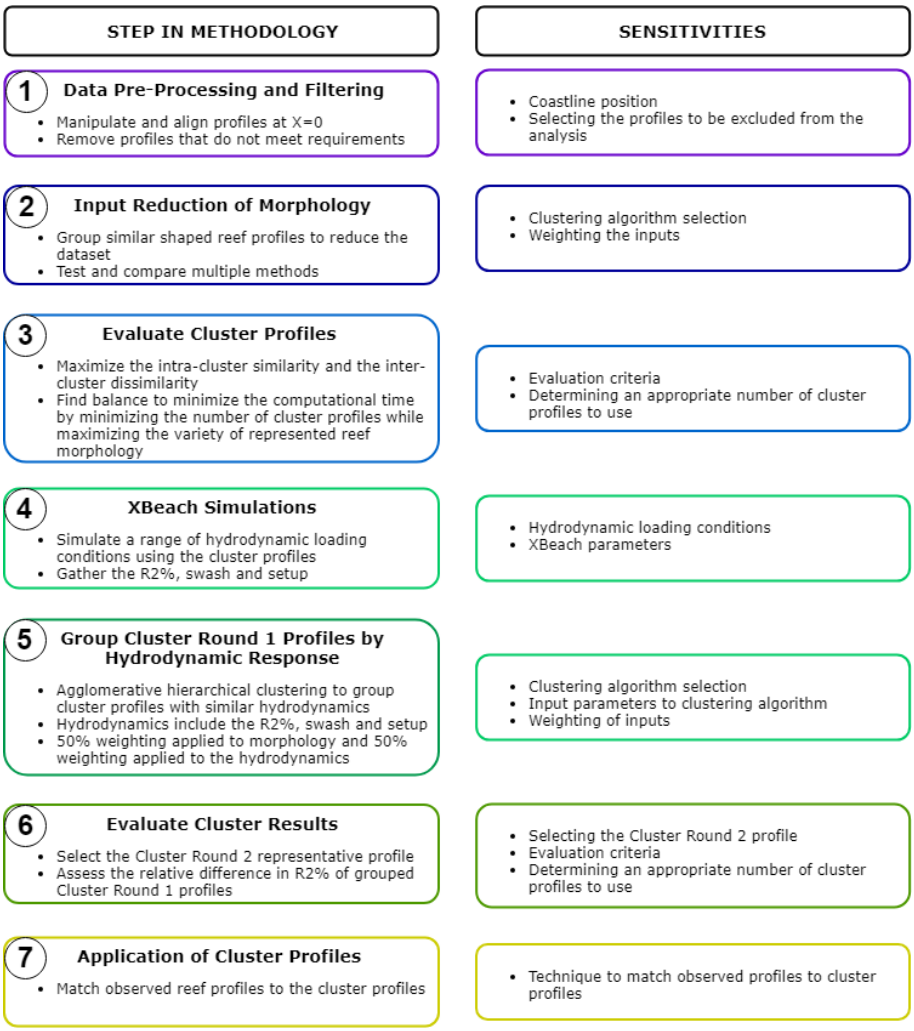


Figure 5.1: The steps of the methodology and the key sensitivities, detailing the decisions that had to be made at each step and what could be changed.

The methodology is mainly sensitive to the choice and application of the clustering algorithms (Step 2 and 5). As presented in Section 4.1, the clustering algorithms group the profiles differently. This includes the differences in methods to calculate distances between profiles, explained in Appendix C.2.1 and C.3.1. Not tested in this study, but recommended for future work in Section 6.2 includes using a weighting scheme for different profile depths in the clustering algorithms. This could provide more significant changes compared to using different clustering algorithms. Furthermore, the application of the cluster profiles is sensitive to the method used to match profiles to the cluster profiles (Step 7), as shown in Section 4.4.1. However, the mean accuracy of the cluster profile prediction varied by approximately 2% when using different matching methods, so the generation of the cluster profiles themselves seems to be the area where the greatest improvement could be made.

5.2. APPLICATION

The findings from this study can be used in many different applications. The methodology itself is a new development, using data mining techniques in the field of coastal engineering, and the cluster profiles can be used to generate reliable wave runup estimates for coral reef locations around the globe, among other useful applications explained in this section.

5.2.1. APPLICATION OF CLUSTER PROFILES

The cluster profiles formed in this study, which are representative of approximately 20,000 coral reef profiles from seven different locations throughout the Pacific Ocean, Caribbean and Florida, have the main purpose of being incorporated into an updated BEWARE model (Pearson et al., 2017). BEWARE has many useful applications, including being used as a global wave runup estimation system, to aid studies and planning for climate change scenarios, and to further understand coral reefs as nature-based flood defenses.

UPDATED BEWARE

The BEWARE model is a Bayesian Network which uses training data to predict an output given a set of inputs. The current BEWARE model created by Pearson et al. (2017) uses wave conditions and reef profile parameters to estimate wave runup and other important hydrodynamic output. However, since the input variables must remain constant, the tested reef profile must be matched to a highly schematized fringing reef profile, described only by the beach slope, reef flat width, reef crest depth and fore reef slope.

If the cluster profiles were used with the BEWARE model, observed reef profiles could be matched to the cluster profiles, as done in Step 7 of the methodology (see Figure 5.1), rather than the closest schematized profile as is currently being done. The cluster profiles have been developed to be more representative of natural reefs, as well as cover a much broader range of coral reef shapes, and therefore the applicability of BEWARE while using the cluster profiles is much greater. The BEWARE model would then be able to operate similarly as it does now, providing estimates of the desired outputs with associated ranges and probabilities, but with enhanced accuracy. The BEWARE tool has proven to be very useful and successful, having already been used by the World Bank for a study into runup-reduction through coral reef restoration in the Seychelles (S. Pearson,

van der Lugt, van Dongeren, Hagenaars, & Burzel, 2018). Adding the cluster profiles to the model would enhance its usefulness in dealing with the natural variability of coral reefs, and allow the system to be used by many more coastal communities.

To incorporate the profiles into the model, the XBeach output from a full range of wave loading conditions, as well as different beach slopes and bed friction values would be required for each cluster profile. The details of how the cluster profiles would be included in an updated BEWARE model are provided in Section 5.3.2.

GLOBAL WAVE RUNUP PREDICTION

The combination of the cluster profiles with BEWARE is a step towards the development of a global flood early warning system (EWS). If paired with a regional wave model and a database of reef morphology, the BEWARE model could generate custom flood forecasts. Essentially, with the real-time or forecasted wave conditions input into the model, along with the reef morphology, the BEWARE system could provide estimates of wave runup with uncertainties, signalling if flooding is expected. The same could be done with XBeach alone (Bosserelle, Kruger, Movono, & Reddy, 2015), however, Bayesian networks provide speed and uncertainty estimates that XBeach only models can not.

The benefit of this type of system is that the wave runup over all of the represented coral reef morphology would be pre-computed, removing both the model set up and run time. It is also a relatively cheap method, which for a lot of the Small Island Developing States is a necessary feature.

CLIMATE CHANGE SCENARIO MODELLING

Another use for the cluster profiles within the BEWARE model is estimating future impacts based on climate change scenarios. Since the model is a quick and accurate tool, many different loading conditions with varying offshore water levels and wave parameters could be analyzed to estimate the associated impacts that they would cause. Since the cluster profiles encapsulate such a wide variety of profiles, large scale climate change estimates could be made to assess the effects of different rates of SLR on multiple different types of coral reef profiles, providing valuable information about the types of profiles and islands that are most at risk.

NATURE BASED FLOOD DEFENSES

Nature based flood defenses are growing in popularity. The simple brilliance is that these designs are able to help solve a coastal problem while providing additional ecosystem services (Temmerman et al., 2013). Coral reefs are among the most effective nature based coastal defense, having proven to dissipate on average 97% of wave energy (Ferrario et al., 2014). An analysis of the most effective types and shapes of coral reef profiles as a natural coastal defense can be done with the cluster profiles. This would be very useful in further understanding the effects of wave attenuation over coral reefs and determining the most appropriate features to enhance during a coral reef restoration.

Pearson et al. (2016) found that narrower reefs are more vulnerable to coastal flooding and that increasing the friction of a narrower reef has greater benefit compared to a longer reef. Therefore, for cost-effective restorations, the focus should be on the narrower reefs. The cluster profiles show the same results, displayed in Figure 5.2. Here, the cluster profiles are colored to represent their wave runup rank, based on the four wave

conditions tested in this study. A connection can be made between the shape of the profile and its susceptibility to wave runup, most notably that shallower and longer reefs generally result in much less wave runup compared to steep and short reefs.

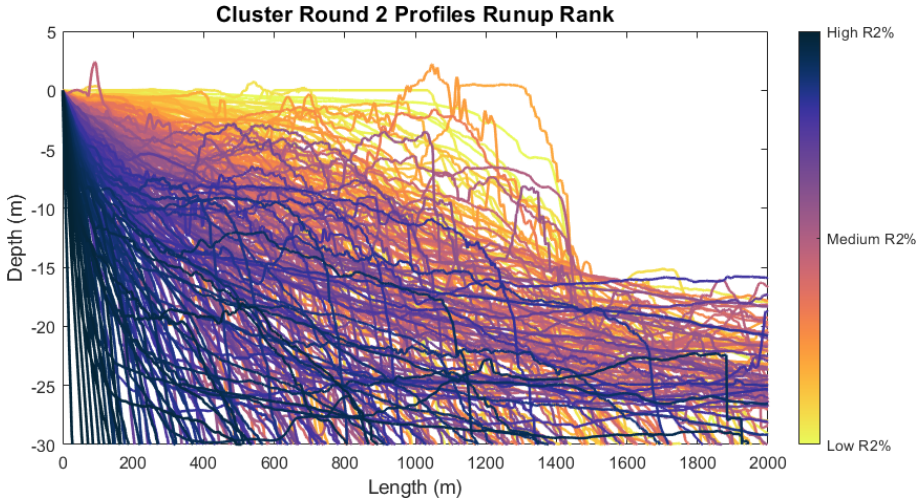


Figure 5.2: The set of 311 Cluster Round 2 profiles ranked from high to low $R_{2\%}$ based on the four loading conditions used in this analysis.

If planning a reef restoration, the local bathymetry could be matched to cluster profiles to determine the area that is most susceptible to runup and flooding risk, and therefore where the efforts should be focused. Also, the effects of increased friction could quickly be investigated. Potentially, the wave runup of similar cluster profiles could be compared to determine if artificially adding length or slightly modifying the shape would be beneficial.

5.2.2. APPLICATION OF DEVELOPED METHODOLOGY

The methodology includes using data mining techniques to reduce a large dataset of coastal profiles into a subset that is representative with regards to a certain output (in this case wave runup). This is useful because of the immense reduction in computational time required to simulate the hydrodynamics over such a range of morphology. In this study, the dataset was reduced by two orders of magnitude, which translates to a reduction in computational time of more than 98%.

The focus for this study was on coral reef profiles from islands in the Pacific Ocean, Caribbean Sea, and mainland Florida that mainly represent coral atolls and fringing reefs. A similar structure could be applied to other types of coastal environments, including coral reefs from the Atlantic and Indian Ocean, extending to barrier reefs, sandy beaches, gravel beaches and rocky shores. The output could then be used in a similar capacity as the cluster profiles formed in this study.

ATOLL AND FRINGING REEFS FROM OTHER LOCATIONS AND ROCKY SHORES

To apply the data reduction for atolls and fringing reefs from other locations, the methodology would not have to be changed. The cluster analysis techniques and hydrodynamic modelling would operate the same since their morphology and hydrodynamic response from these profiles will be in line with those already included in the study.

Although rocky shores were not included in this study, they are also fixed coastlines like coral reefs. For this reason, the methodology for rocky shores would also be very similar to the one proposed in this study.

BARRIER REEFS

Barrier reefs, however, were purposefully excluded from this study. This was mainly due to their length and the difficulties associated with modelling such long profiles in XBeach. The long lengths and lagoon feature of barrier reefs could result in wind generated waves on the reef that have a significant impact at the shoreline. In a model such as XBeach, the waves are input from the offshore boundary and locally generated waves are not included. However, barrier reefs could be included in such an analysis if the proper hydrodynamical modelling tools were used.

Step 4 in the methodology, therefore, would have to be altered. Perhaps something similar to the two-dimensional model used to model the Great Barrier Reef (Lambrechts et al., 2008) could be used. Step 5 could also include additional hydrodynamic parameters in the cluster analysis. The lagoon system may result in hydrodynamics other than setup, swash and $R_{2\%}$ that are relevant for grouping profiles that result in similar wave runup, such as wave reflection.

Step 7 would also need to be addressed, which involves how the cluster profiles are applied to represent an observed profile. Specifically, a new matching technique may be required. From the findings in this study, the nearshore profile should be heavily weighted, but the length of the lagoon and morphology of the barrier would most likely be of high importance as well. Sensitivity testing to different matching methods would again be necessary.

SANDY AND GRAVEL BEACHES

Sandy and gravel beaches have been classified in a broad sense as dissipative to reflective (Wright, Chappell, Thom, Bradshaw, & Cowell, 1979), but a more precise classification would enable the same type of BEWARE tool to be used for these coastal settings as well. The difficulty with sandy and gravel beaches is that they are not fixed as a coral reef is. The morphology adapts to the hydrodynamics and therefore the profile changes. This would require a slight change for Step 1 of the methodology, in which the data would have to include profile measurements from the same location at different times of the year in order to capture the full range of profiles that each beach can form to. Hydrodynamics are typically characterized by season (ex. stormy winter and calm summer). Obtaining measurements from the different seasons would capture a greater extent of the potential beach profiles to include in the analysis. Assuming that the initial profile shape is the greatest importance for estimating wave runup, including the full range of beach profile shapes would be beneficial.

The sediment properties pose another challenge for these coastal systems. This may not have to be addressed throughout the entire methodology, but potentially when form-

ing the XBeach database to train a model such as BEWARE. Each cluster profile would have to be modelled in XBeach with varying grain sizes and densities. This way the cluster profiles can be matched accurately to an observed profile with known grain properties, and predict the wave runup with high accuracy. This is similar to modelling different friction values for the coral reefs.

5.3. NEXT STEPS

Other additional analyses and ideas for future work related to this thesis are presented in this section. These include using the main findings of the sensitivities derived from this study to update and improve the methodology. Second is the combination of these profiles into an updated BEWARE model, followed by expanding to coral reefs for other locations around the world, as well as other coastal environments. Lastly, to focus on a satellite derived bathymetry which is a promising source for coastal measurements.

5.3.1. UPDATE CLUSTERS

The nearshore profile has been proven to be very important for grouping profiles with similar wave runup results (see Section 4.2.1, 4.3.3 and Appendix E). In this study, the input reduction of morphology worked to form clusters based on similarities of the whole reef profile, but this finding suggests that grouping the profiles with greater emphasis on the nearshore profile is more appropriate. Detailed recommendations are provided in Section 6.2.

Performing the methodology with the provided recommendations would form new Cluster Round 1 profiles that are specifically focused to capture and represent the aspects of the profile that are most important for wave runup. With more detailed clustering on the nearshore region, instead of 500 cluster profiles that represent the entire morphology generally, the 500 cluster profiles could represent the 500 different nearshore shapes specifically, providing much more accuracy and detail into the region that matters most. This would most likely increase the accuracy of the predictions.

5.3.2. COMBINING WITH BEWARE

Once a set of cluster profiles is ready for application, it can be incorporated into the BEWARE model. A range of wave loading conditions would have to be simulated in XBeach with all of the cluster profiles, forming a synthetic database that would be used as the training dataset for the Bayesian network.

If the same wave loading conditions, beach slopes and friction values were to be used as was done for the initial BEWARE model, the upper limit of 311 cluster profiles would require 167,490 XBeach simulations, and the lower limit of 45 cluster profiles would require 24,300 XBeach simulations. With access to high powered computers, these simulations could be complete in a few weeks. Also, with the gained knowledge from Pearson et al. (2017) study, the sensitivity to wave runup from the different parameters could be used to reduce the number of values for certain parameters. For example, the beach slope was proven to be one of the least influential parameters to the wave runup, and so instead of simulating three different beach slopes for each profile, perhaps two would be sufficient, reducing the number of XBeach runs by 33%.

The probabilistic matching method, which proved to be the best of the three methods tested, is used in Step 7 of the methodology (Figure 5.1) to pair observed reef profiles with the cluster profiles. It provides probabilities of belonging to each cluster profile, based on the distance between them. BEWARE uses a Bayesian probabilistic network (BN) to compute the outputs, but currently, the input to the BEWARE system is the closest profile and the required wave parameters. Providing probabilistic matches to the different cluster profiles as input to the BN could be a very nice addition to the model. This would enable the BN to calculate its output based on a probability distribution, which could lead to a more accurate range and uncertainty associated with the output.

5.3.3. FURTHER ESTABLISHING CORAL REEF DATASET

IMPROVE VARIETY OF CLUSTER PROFILES

The dataset used for this study is extensive, consisting of approximately 30,000 coral reef profiles from 7 different locations. However, all profiles are from the Pacific Ocean, the Caribbean Sea and Florida. For a more complete global coverage, reef profiles from the Indian and Atlantic Ocean would be beneficial to include in the analysis. One set of cluster profiles could then be used for all locations, or if clear differences are present between reefs from different oceans, multiple sets of cluster profiles could be made, and the appropriate set could then be used depending on the site of interest.

GEOMORPHOLOGICAL ANALYSIS

With a more global dataset, potential links could be made between reef profile shape and wave forcing, climate, and geographic area. If these connections could be formed, the cluster profile applicability could be improved. For example, if there is a project in Hawaii, and there is a clear connection to which cluster profiles represent the Hawaiian coastline based on the geography or similar wave loading conditions, then this could narrow down which cluster profiles to look at. The locations of the cluster profiles and the effect of the location on wave runup potential were examined only briefly. The findings can be found in Appendix G.

5.3.4. SATELLITE DERIVED BATHYMETRY

Bathymetry measurements are traditionally taken by singlebeam or multibeam echosounders, or airborne LIDAR (Light Detection and Ranging). Both methods are costly and have limitations. Echosounders will only gather data across travelled transects and are unable to reach very shallow waters. Airborne LIDAR can gather data up to 70 m depth, however, the resolution is typically coarse and requires clear waters (Cahalane, Hanafin, & Monteys, 2016).

Satellite Derived Bathymetry (SDB) is a great alternative since it can be used all around the globe and is relativity cheap. SDB has been used since the 1970s and is either implemented through analytical or empirical methods. Empirical methods develop statistical relationships between image pixel values and field measured water depths. The analytical approaches utilize the general principle that seawater transmittances at near-visible wavelengths are functions of a general optical equation which is dependent upon the intrinsic optical properties of seawater (Cahalane et al., 2016). Therefore, many external factors affect the accuracy of depth calculations such as the resolution of the imagery,

atmospheric effects, sunlight, and vegetation.

The effectiveness of the cluster profiles would be amplified with advancements in satellite imagery for detecting bathymetry. The current application of the cluster profiles for predicting wave runup relies on having detailed bathymetry measurements in order to know which cluster profile to match to, and so having a quick, cost-effective and accurate method to gather bathymetric data would be extremely beneficial.

Andrefouet et al. (2006) presented the Millennium Coral Reef Mapping Project, which uses a compilation of Landsat 7 ETM+ satellite images to provide unprecedented coverage of coral reefs worldwide. The spatial scale of the imagery is the difficult part of making this dataset most useful, but research is being done into increasing spatial resolution to the required accuracy for modeling coral reefs (Andrefouet et al., 2006).

With a global data set (at least of the Pacific Ocean, Indian Ocean and the Caribbean Sea) of coral reef bathymetry, the representative reef profiles developed in this study could be used to provide estimates of runup for all coral-reef lined coasts. Also, with a larger dataset, the methodology could be repeated with a greater selection of reef profiles. In either case, with increased efforts in remote sensing to gather detailed large scale bathymetric data, a tool such as the one developed in this study will become even more important and powerful. In addition, the measured cluster profiles could also be used to calibrate satellite estimates, aiding the accuracy and tuning of satellite derived bathymetry.

5.4. LIMITATIONS OF FINDINGS

ONE-DIMENSIONAL HYDRODYNAMIC ANALYSIS

Throughout the entire processes, the coral reef morphology and hydrodynamics over the reef have only been analyzed in the cross-shore. However, the hydrodynamics over a coral reef are known to fluctuate in the along-shore direction as well due to the complex reef bathymetry and directional wave propagation behavior (Su & Ma, 2018). In a numerical study done by Su and Ma (2018), it was found that with a fringing reef, the alongshore bathymetric variability promotes refraction and diffraction, resulting in circulation cells with onshore and offshore directed flow (Su & Ma, 2018). Rip currents developed, with the offshore flow located at the deeper part of the reef.

Another numerical modelling study done by van Dongeren et al. (2013) compared one and two-dimensional XBeach simulations on a fringing coral reef. It was found that the one-dimensional model was able to capture the gradients in the dominant hydrodynamic processes, but with a high friction coefficient. The two-dimensional model was capable of using a lower and more realistic bed friction coefficient which resulted in more optimum performance and more closely resembled measured values.

Suffice to say, using the one-dimensional XBeach model will not properly emulate all hydrodynamic processes. Along-shore variability in the reef causes wave focusing and de-focusing, which ultimately result in along-shore currents and variations in wave energy at the shoreline. However, a two dimensional model for the application in a tool such as BEWARE is not feasible. For the simplified EWS, the results obtained from the one-dimensional model should be sufficient to properly predict when a proper threat is approaching.

The case when using one-dimensional modelling would be most unreliable would be when there is high alongshore variability in the reef. For these cases, the cluster profile prediction would be least accurate since more alongshore processes would be present. A recommendation for such a scenario is provided in Section 6.2.

BOTTOM FRICTION

A similar profile shape is crucial to predict wave runup response, however, the bottom friction also is important to model correctly to obtain an accurate prediction. Bottom friction has been proven to be an important factor in wave dissipation and transformation over coral reefs (Lowe et al., 2005; Pearson et al., 2017; Quataert et al., 2015), in which higher friction coefficients lead to decreases in the wave runup. Pearson et al. (2017) found that with the schematized reef profiles, an increase in the coefficient of friction from 0.01 to 0.1 will reduce the $R_{2\%}$ on average by 23%. Therefore, even a profile with the exact same shape can have different wave runup depending on the friction value, and so even if an observed profile matches very well with a cluster profile, if the friction is not matched properly the wave runup estimation can be inaccurate.

In this study, bottom friction was held constant across the reef, and for all profiles. The bottom friction relates to the amount of coral on the reef, and the coral cover was not included in the analysis because it is highly dependent on local conditions (see Appendix B.3), and therefore when trying to create characteristic profiles of reefs from around the world, including such local effects is extremely tricky. Using a constant coefficient of friction can work well if in total the combined frictional effect from the constant value represents the spatially varying friction. When applying the cluster profiles, this will have to be the goal.

LIMITED WAVE CONDITIONS

In Step 4 of the methodology, XBeach simulations are done using the cluster profiles. The cluster profiles that have similar wave runup, obtained from the simulations, are grouped together in Step 5. Due to time constraints, only four wave loading conditions were tested. Although they were chosen strategically to cover a wide range of potential flooding conditions and different types of ocean waves, if two profiles have similar wave runup over these four conditions it does not necessarily mean that they always will. The cluster analysis and accuracy of cluster profile's prediction are therefore limited by the variety of tested wave conditions.

XBEACH MODEL LIMITATIONS

- Single-peaked JONSWAP spectra
- No directional spreading
- 1-D model
- Spatially uniform bottom fiction

6

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

- K-medians is the most effective clustering algorithm of those tested to cluster the coral reef profiles based on morphology, and agglomerative hierarchical clustering was successfully applied to group the cluster profiles based on morphology and hydrodynamics
- The nearshore profile is the most important for grouping profiles with similar wave runup
- Peaks above MSL and large nearshore differences cause vastly different reef hydrodynamics, leading to differences in wave runup response
- Using a probabilistic approach to match observed reef profiles with the cluster profiles results in the highest predictive accuracy
- The cluster profiles are capable of predicting wave runup with a mean relative difference of approximately 10%, based on completed XBeach computations
- Narrow, steep reefs result in much greater wave runup compared to long shallow reefs

RECOMMENDATIONS

- Perform rigorous data pre-processing to strategically select profiles for the analysis, as well as the coastline location
- Apply greater weighting to the nearshore profile throughout the analysis, including Steps 2, 5 and 7
- Research methods to match profiles to the cluster profiles
- Establish the value of depth that classifies 'nearshore'
- Validate the cluster profiles with other data sources and measured observations
- Ensure the cluster profiles are adequately smoothed to remove local disturbances
- Establish a warning system when 2-D processes will likely disrupt the runup prediction

Data reduction techniques were used to reduce an extensive dataset of coral reef profiles to a subset that can be used for wave runup predictions. This was done by first reducing the dataset into cluster groups based on full reef morphology, and secondly by grouping the cluster groups with similar wave runup. The key findings from this analysis are presented in this section, as well as recommendations for improving the analysis.

6.1. CONCLUSIONS

Four research questions were stated at the beginning of this study. The answers to the questions are listed below:

1. *How can a large dataset of coral reef profiles be clustered such that the hydrodynamic response of grouped profiles is similar, and how should the cluster groups be represented?*
   The dataset was reduced using two rounds of cluster analysis. The first round reduced the 20,454 reef profiles based on morphology alone. Multiple different clustering algorithms were tested to determine which was most effective. The K-medians method resulted in the lowest average distance between a profile and its cluster centroid. Selecting the number of cluster groups to output from this method was done by comparing the incremental gain in accuracy. 500 clusters were determined to be the optimal value to balance a low intra-cluster variance and high data reduction. To represent the cluster groups, the median of the grouped profiles was determined to represent the group better than the mean, as well as create an adequately smooth profile that removed the local variability.

   The second round of cluster analysis merged cluster profiles with similar shape and wave runup. Agglomerative hierarchical clustering worked well to combine morphological and hydrodynamic inputs, merging the 500 cluster profiles to a range of 311 to 45 Cluster Round 2 groups. These groups were represented by the cluster profile with the median wave runup. The two-step approach reduced the dataset by at least two orders of magnitude while capturing a wide range of coral reef morphology.

2. *What aspects of the reef profile are most important to consider for effective clustering in terms of wave runup?*
   The features of the profile that are important for effective clustering were mainly determined by assessing what leads to ineffective clustering. The features of profiles that result in the largest differences in wave runup between grouped profiles are large nearshore differences (0 to -15 m depth), as well as peaks in the profile above MSL. The nearshore differences cause large discrepancies in how the waves are dissipated while approaching the beach. A deeper profile will allow more wave energy to reach the shoreline compared to a shallower profile. The peak in the profile above MSL slightly resembles a barrier reef or a high offshore ridge. It causes the water to build up on the landward side of the peak, acting as a submerged breakwater, resulting in high setup landward of the peak that a profile without a peak would not generate. These key differences in the physical profile translate to severe variances in wave runup response and should be accounted for in future

analysis. The length of the profile was also found to be important for grouping profiles with similar wave runup, but less so than a very good match of the nearshore profile.

3. *What is the best approach to utilize the cluster profiles in order to predict wave runup of a natural reef profile?*
   Once the cluster profiles were formed, the method to apply the cluster profiles to predict wave runup of a set of real reef profiles was explored. The application relies on the real reef profile being matched to the appropriate cluster profile. Three approaches were evaluated to match a test profile to the cluster profiles. For a direct match, the NS3 method, which matches the profiles based on the nearshore morphology, produced the most accurate results. This coincides with the other finding during this study that the nearshore morphology is the most important to have well-matched in order to accurately predict wave runup. However, combining the NS3 criteria with a soft matching method that provides probabilities of belonging to each cluster profile lowered the mean error in predicted wave runup. The probabilistic approach incorporates multiple cluster profiles to calculate the runup estimation, thereby reducing the error when a real reef profile is in between cluster groups.

4. *How accurately can the selected cluster profiles predict wave runup and flooding over natural coral reefs?*
   The cluster profiles produced in this study were able to predict the wave runup of 1000 test profiles with a mean predictive error of approximately 10% when matching the test profiles to the cluster profiles using the probabilistic approach. This is true for all four of the wave conditions tested. The accuracy is dependent upon the number of final cluster groups selected, which range between 311 and 45. The accuracy and variety of reef morphology encapsulated in the cluster profiles come at the expense of a greater number of required XBeach simulations compared to the original BEWARE model, which this study aims to improve. However, the cluster profiles cover a much more diverse array of coral reef morphology compared to the schematized profile used in BEWARE, and therefore the additional XBeach simulations are justified. Further comparison and validation with measured data are required.

6.1.1. ADVANCES

The methodology developed in this study advances the use of data mining techniques in coastal engineering and also adds detail and accuracy to the BEWARE tool previously developed by Pearson et al. (2017). Cluster analysis has been used for wave climate data reduction (Camus et al., 2011; Olij, 2015) and for coastal bathymetry (Costa et al., 2016; Duce et al., 2016; Tomás et al., 2016), however an extensive method involving multiple different clustering algorithms and with the purpose of classifying reef profiles for wave runup prediction had not been done. This study opens the door for cluster analysis or other data mining strategies for other large coastal datasets, including other coastal environments. The USGS plans to use a similar strategy to classify and model the entire US coast, and this study aids in determining which strategies can be applied for such a

project. The BEWARE tool has been tested with measured data and has proven to be effective, however, one of the areas for improvement included the method for representing the reef profiles in the model. The cluster profiles developed in this study nicely fill that gap.

6.2. RECOMMENDATIONS

Grouping coral reef profiles had previously not been done with such a robust dataset. As such, trial and error were unavoidable, and although the methodology developed has been proven to be successful, there have been some key points that should be incorporated to update and improve the results. Several recommendations for future work related to this study are provided below:

1. *More rigorous data pre-processing*
   Several steps were taken to pre-process the data for the cluster analysis, however, some arbitrary values were selected, such as the defined width and height of land when selecting the coastline location along the profile. Referring to Step 1 in Figure 5.1, this step could be improved with a more detailed approach to determine the coastline location, as well as which profiles should be included and excluded from the analysis. The land was classified when the profile remained above MSL for 100 m width. The 100 m could be replaced by the average width of the land or islands for that location. Also, this method resulted in profiles with offshore peaks above MSL. These peaks could potentially be inhabited, but with the current method, there is no runup estimation for them. Using satellite imagery to validate these features could be beneficial to determine if they should be considered as land. If validation is not possible, the profile could be included in the analysis twice, in which one of the variations include the scenario that the peak is land. This would add the profile type to the cluster analysis and allow both situations to be represented.

2. *Applying higher weighting to the nearshore profile*
   It was concluded, as could be expected, that the shallower parts of the profile are of greater importance for estimating wave runup and for grouping profiles with similar wave runup. The idea of focusing on the nearshore profile was only applied in Step 7 of the methodology when matching test profiles to the cluster profiles. However, the cluster profiles themselves should also be formed with a greater emphasis on the nearshore. In Step 2 (the first cluster analysis), instead of applying the clustering algorithms with equal weighting across the entire profile, placing higher weighting to the measurements in shallower water would be beneficial, or perhaps even only clustering up to a certain depth limit. An approach similar to NS3, which is used to match the profiles, could be implemented in which profiles are grouped together based on their morphology up to -15 m, but also with certain length difference limitations. The same idea should also be used in Step 5 (the second cluster analysis).

   For all aspects of the methodology that focus on the nearshore profile, it is worth examining a non-uniform weighting scheme. In the current approach, each depth

measurement up to -15 m is treated with equal weighting, but the importance of the profile similarities is most likely not uniform along the profile from the beach to the -15 m depth. Surely, a 0.5 m difference between profiles at - 1 m depth is more significant than a 0.5 m difference at -15 m depth. Therefore, a weighted nearshore matching method in which the difference between profiles is weighted higher closer to the shoreline and lower as the depth increases may enhance the clustering and matching methods.

3. *Improve the matching method*
In Step 7, three different methods were established to match an observed profile to the cluster profiles. Each time a new method was developed, the accuracy of the cluster profile prediction increased. The most successful method developed in this study is the probabilistic approach, using the NS3 distance measurement as input. Although the accuracy with this method is high, further testing and development of others could improve the results. For example, the NS3 method limits the matching of profiles to those with full lengths within 500 m of each other. This value was somewhat arbitrarily selected. Perhaps the 500 m is excluding too many cluster profiles or including too many. On top of this, the proposed probabilistic method uses the softmax function with a relatively high B value of 4. A higher B value assigns a higher probability to the closest match, whereas a lower B value will include more cluster profiles with significant probabilities, as shown in Appendix C.4.1. Testing different B values to determine which approach provides the best results is recommended. Potentially there should not be one fixed method to match the profiles, but rather allow the matching to be a function of the conditions relevant to the observed profile. This could result in the best cluster profile to be selected based on the specific wave conditions that the observed profile requires runup estimates for. This is not a priority over the cluster profile generation but could be another source of error to improve upon.

4. *Determine the depth value that is most important for grouping profiles with similar wave runup*
The recommendation to apply greater weighting to the nearshore profile has already been noted, but there needs to be a clear definition of nearshore. In this analysis, -15 m depth was used to classify nearshore, mainly because at this depth the largest waves would begin to break on the reef. By testing different depth values, the depth of the profile that is most important for grouping profiles with similar wave runup could be established. For the cluster analysis in both Step 2 and 5, this value would be useful for determining up to which depth higher weighting should be applied. In Step 7, to predict the wave runup of real reef profiles, this value would be useful for improving the matching method. Similar to the NS3 matching method, the profiles could then be matched to the cluster profiles up to the set depth value that has been determined to be most critical.

5. *Perform the second round of cluster analysis with hydrodynamics only*
In the second round of cluster analysis, the inputs into the algorithm included morphology and hydrodynamics, each with 50% weighting. This was done to ensure grouped profiles do not have vastly different shapes even though they have

similar wave runup. However, it would be interesting to determine if including the morphology is necessary and to determine what types of profiles are matched together when morphology is excluded. This could lead to additional findings into what types of reef profiles should be grouped together and the features of reef profiles that cause certain hydrodynamic responses.

6. *Validate the cluster profiles with measured data*
   The main comparisons and quantification of runup accuracy of the cluster profiles has been done using reef profiles included in the dataset, which were used to generate the cluster profiles. Two profiles from Roi Namur were used for a small validation, but two profiles are much too little for proper validation. Also, all comparisons between real reef profiles and the cluster profiles were done with XBeach to XBeach results.

   Therefore, a two-step validation is necessary. First, a validation of the method must be complete using data from other sources. This will determine how applicable and accurate the cluster profiles truly are. More profiles from the Marshall Islands with measured wave runup data, for example those used by Cheriton et al. (2016), as well as profiles from the Indian Ocean, Caribbean, and other atoll islands in the Pacific Ocean would be very useful. Second, validation of XBeach runup results vs observed runup results is necessary to ensure that the model predictions are accurate.

7. *Smooth the Cluster Round 1 profiles*
   Following the proposed method, the Cluster Round 1 profiles are the median of the profiles within the cluster group. The median naturally provided a smoothed profile when the cluster group is highly populated, but in a group with few profiles, the local perturbations of the profile can be carried forward to the median cluster profile. Therefore, a slight smoothing to the profiles could help eliminate some of the strictly local features and produce cluster profiles more representative of a greater region. This could be done with a moving average smoothing function which removes the fine disturbances.

8. *Assess nearby reef morphology to aid in wave runup prediction*
   One of the limitations of the study is the one-dimensional approach. When there is significant alongshore variability in the reef, the wave runup will most likely differ compared to the output from a one-dimensional model. One method to overcome this, or at least set a warning for such a case, would be to examine the nearest cross-shore profiles to the profile of concern. If the nearest profiles differ greatly compared to the profile of concern, it can be assumed that the alongshore variability will have a significant impact on the wave runup and a warning of the accuracy of the cluster profile prediction can be supplied.

In conclusion, the proposed methodology provides a means of reducing large datasets of coral reef profiles into a subset of representative profiles that can be used for wave runup and flooding prediction. Sea level rise and fierce wave conditions pose many risks for the vulnerable coral atoll islands, but by making use of powerful data mining techniques, tools can be developed to help weather the storm.

BIBLIOGRAPHY

Akaikei, H. (1973). Information theory and an extension of maximum likelihood principle. In *Proc. 2nd int. symp. on information theory* (pp. 267–281).

Alfieri, L., Salamon, P., Pappenberger, F., Wetterhall, F., & Thielen, J. (2012). Operational early warning systems for water-related hazards in Europe. *Environmental Science & Policy*, *21*, 35–49.

Andrefouet, S., Muller-Karger, F. E., Robinson, J. A., Kranenburg, C. J., Torres-Pulliza, D., Spraggins, S. A., & Murch, B. (2006). Global assessment of modern coral reef extent and diversity for regional science and management applications: a view from space. In *Proceedings of the 10th international coral reef symposium* (Vol. 2, pp. 1732–1745). Japanese Coral Reef Society Okinawa, Japan.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Ashton, A. D., Toomey, M., & Perron, J. T. (2013). Profiles of ocean island coral reefs controlled by sea-level history and carbonate accumulation rates. *Geology*, *41*(7), 731–734. doi:10.1130/g34109.1

Baldock, T. E. (2012). Dissipation of incident forced long waves in the surf zone—Implications for the concept of "bound" wave release at short wave breaking. *Coastal Engineering*, *60*, 276–285. doi:10.1016/j.coastaleng.2011.11.002

Barnett, J., & Adger, W. N. (2003). Climate dangers and atoll countries. *Climatic change*, *61*(3), 321–337.

Bauckhage, C. (2015). *Lecture Notes on Data Science: Soft k-Means Clustering*. doi:10.13140/RG.2.1.3582.6643

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, New Jersey: Princeton University Press.

Berkhin, P. (2002). A Survey of Clustering Data Mining Techniques. (pp. 25–71). doi:10.1007/3-540-28349-8_2

Blacka, M., Flocard, F., Splinter, K. D., & Cox, R. J. (2015). *Estimating Wave Heights and Water Levels inside Fringing Reefs during Extreme Conditions*. Retrieved from https://www.researchgate.net/publication/294799669

Bosserelle, C., Kruger, J., Movono, M., & Reddy, S. (2015). *Wave inundation on the Coral Coast of Fiji*. Retrieved from https://www.researchgate.net/publication/293987211

Bruun, P. (1954). *Coast erosion and the development of beach profiles*. US Beach Erosion Board.

Cahalane, C., Hanafin, J., & Monteys, X. (2016). Improving Satellite-derived Bathymetry. Retrieved June 15, 2019, from https://www.hydro-international.com/content/article/improving-satellite-derived-bathymetry

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Camus, P., Mendez, F. J., Medina, R., & Cofiño, A. S. (2011). Analysis of clustering and selection algorithms for the study of multivariate wave climate. *Coastal Engineering*, *58*(6), 453–462.

Cheriton, O. M., Storlazzi, C. D., & Rosenberger, K. J. (2016). Observations of wave transformation over a fringing coral reef and the importance of low-frequency waves and offshore water levels to runup, overwash, and coastal flooding. *Journal of Geophysical Research: Oceans*, *121*(5), 3121–3140.

Church, J. A., Clark, P. U., Cazenave, A., Gregory, J. M., Jevrejeva, S., Levermann, A., … Nunn, P. D. (2013). *Sea level change*. PM Cambridge University Press.

CO2.Earth. (2019). Daily CO2. Retrieved from https://www.co2.earth/daily-co2

Costa, M. B. S. F., Araújo, M., Araújo, T. C. M., & Siegle, E. (2016). Influence of reef geometry on wave attenuation on a Brazilian coral reef. *Geomorphology*, *253*, 318–327. doi:https://doi.org/10.1016/j.geomorph.2015.11.001

Darwin, C. (1832). The structure and distribution of coral reefs: being the first part of the geology of the voyage of the Beagle, under the command of Capt. *Fitzroy, RN during the years*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227.

De Ridder, M. P. (2018). *Non-hydrostatic wave modelling of coral reefs with the addition of a porous in-canopy model* (Doctoral dissertation, TU Delft Repository).

de Coninck, H., Revi, A., Babiker, M., Bertoldi, P., Buckeridge, M., Cartwright, A., … Sugiyama, T. (2018). *Strengthening and implementing the global response. In: Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthen*. Intergovernmental Panel on Climate Change. In Press.

Dean, R. G., & Galvin Jr, C. J. (1976). Beach erosion: causes, processes, and remedial measures. *Critical Reviews in Environmental Science and Technology*, *6*(3), 259–296.

Deshpande, M. (2017). Clustering with Gaussian Mixture Models. Retrieved from https://pythonmachinelearning.pro/clustering-with-gaussian-mixture-models/

Duce, S., Vila-Concejo, A., Hamylton, S. M., Webster, J. M., Bruce, E., & Beaman, R. J. (2016). A morphometric assessment and classification of coral reef spur and groove morphology. *Geomorphology*, *265*, 68–83. doi:https://doi.org/10.1016/j.geomorph.2016.04.018

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Runze, L. (2012). Sensitivity and Specificity of Information Criteria. *Technical Report Series*, *12*(119), 1–30. doi:10.7287/PEERJ.PREPRINTS.1103V2

ECHO. (2019). *European Civil Protection and Humanitarian Aid Operations - Pacific Region*. European Civil Protection and Humanitarian Aid Operations (ECHO).

European Union. (2007). Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks. *Official Journal of the European Union*, *288*, 27–34.

Ferrario, F., Beck, M. W., Storlazzi, C. D., Micheli, F., Shepard, C. C., & Airoldi, L. (2014). The effectiveness of coral reefs for coastal hazard risk reduction and adaptation. *Nature communications*, *5*, 3794.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York, NY, USA:

Galarnyk, M. (2018). Understanding Boxplots – Towards Data Science. Retrieved May 13, 2019, from https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Garcia, V., Nielsen, F., & Nock, R. (2009). *Levels of details for Gaussian mixture models*. Retrieved from http://www.lix.polytechnique.fr/%7B~%7Dnielsen/MEF/Garcia%7B%5C_%7D2009%7B%5C_%7DACCV.pdf

Gawehn, M. A. (2015). Incident, infragravity and very low frequency wave motions on an atoll reef platform.

Geurts, M., Box, G. E. P., & Jenkins, G. M. (2006). Time Series Analysis: Forecasting and Control. *Journal of Marketing Research*, *14*(2), 269. doi:10.2307/3150485

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. (Chap. Numerical, pp. 78–95). MIT Press. Retrieved from http://www.deeplearningbook.org

Gourlay, M. R. (1994). Wave transformation on a coral reef. *23*(1-2), 17–42. doi:10.1016/0378-3839(94)90013-2

Gourlay, M. R. (1996a). Wave set-up on coral reefs. 1. Set-up and wave-generated flow on an idealised two dimensional horizontal reef. *Coastal Engineering*, *27*(3-4), 161–193. doi:10.1016/0378-3839(96)00008-7

Gourlay, M. R. (1996b). Wave set-up on coral reefs. 2. set-up on reefs with various profiles. *Coastal Engineering*, *28*(1-4), 17–55. doi:10.1016/0378-3839(96)00009-9

Hall, J. A., Gill, S., Obeysekera, J., Sweet, W., Knuuti, K., & Marburger, J. (2016). *Regional sea level scenarios for coastal risk management: Managing the uncertainty of future sea level change and extreme water levels for Department of Defense coastal sites worldwide*.

Hardy, T. A., & Young, I. R. (1996). Field study of wave attenuation on an offshore coral reef.

Hartigan, J. A. [J A], & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. doi:10.2307/2346830

Hartigan, J. A. [John A]. (1975). Clustering algorithms.

Holliday, J. D., & Willett, P. (1996). Definitions of"dissimilarity"for dissimilarity-based compound selection. *Journal of Biomolecular Screening*, *1*(3), 145–151.

Holman, R. A. (1986). Extreme value statistics for wave run-up on a natural beach. *Coastal Engineering*, *9*(6), 527–544. doi:10.1016/0378-3839(86)90002-5

Hopley, D. (1982). *The Geomorphology of the Great Barrier Reef*. Wiley New York.

Hulme, M. (2016). 1.5 °C and climate research after the Paris Agreement. *Nature Climate Change*, *6*(3), 222–224. doi:10.1038/nclimate2939

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. *Englewood Cliffs: Prentice Hall, 1988*.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kelman, I., & Glantz, M. H. (2014). Early warning systems defined. In *Reducing disaster: Early warning systems for climate change* (pp. 89–108). Springer.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*(1), 137–148.

Kopp, R. E., Horton, R. M., Little, C. M., Mitrovica, J. X., Oppenheimer, M., Rasmussen, D. J., … Tebaldi, C. (2014). Probabilistic 21st and 22nd century sea-level projections at a global network of tide-gauge sites. *Earth's Future*, *2*(8), 383–406.

Kron, W. (2005). Flood Risk = Hazard • Values • Vulnerability. *Water International*, *30*(1), 58–68. doi:10.1080/02508060508691837

Ladd, H. S., Ingerson, E., Townsend, R. C., Russell, M., & Stephenson, H. K. (1953). Drilling on Eniwetok atoll, Marshall islands. *AAPG Bulletin*, *37*(10), 2257–2280.

Lajiness, M. S. (1997). Dissimilarity-based compound selection techniques, Perspect. *Drug Discov*, 7–8.

Lambrechts, J., Hanert, E., Deleersnijder, E., Bernard, P.-E., Legat, V., Remacle, J.-F., & Wolanski, E. (2008). A multi-scale model of the hydrodynamics of the whole Great Barrier Reef. *Estuarine, Coastal and Shelf Science*, *79*(1), 143–151.

Lashley, C. H., Roelvink, D., van Dongeren, A., Buckley, M. L., & Lowe, R. J. (2018). Nonhydrostatic and surfbeat model predictions of extreme wave run-up in fringing reef environments. *Coastal Engineering*, *137*, 11–27. doi:https://doi.org/10.1016/j.coastaleng.2018.03.007

Lee, T. T., & Black, K. P. (1978). The energy spectra of surf waves on a coral reef. In *Coastal engineering 1978* (pp. 588–608).

Legendre, P., & Legendre, L. (2012). *Chapter 8 - Cluster analysis*. doi:10.1016/B978-0-444-53868-0.50008-3

Longuet-Higgins, M. S., & Stewart, R. W. (1962). Radiation stress and mass transport in gravity waves, with application to 'surf beats'. *Journal of Fluid Mechanics*, *13*(4), 481–504.

Longuet-Higgins, M. S., & Stewart, R. W. (1964). Radiation stresses in water waves; a physical discussion, with applications. *Deep Sea Research and Oceanographic Abstracts*, *11*(4), 529–562. doi:https://doi.org/10.1016/0011-7471(64)90001-4

Lowe, R. J., Falter, J. L., Bandet, M. D., Pawlak, G., Atkinson, M. J., Monismith, S. G., & Koseff, J. R. (2005). Spectral wave dissipation over a barrier reef. *Journal of Geophysical Research: Oceans*, *110*(C4).

Lugo-Fernández, A., Roberts, H. H., Wiseman Jr, W. J., & Carter, B. L. (1998). Water level and currents of tidal and infragravity periods at Tague Reef, St. Croix (USVI). *Coral Reefs*, *17*(4), 343–349.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100–103.

Massel, S. R., & Gourlay, M. R. (2000). On the modelling of wave breaking and set-up on coral reefs. *Coastal Engineering*, *39*(1), 1–27. doi:10.1016/s0378-3839(99)00052-6

MathWorks. (2019). Hierarchical Clustering - MATLAB. Retrieved May 27, 2019, from https://www.mathworks.com/help/stats/hierarchical-clustering.html

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(12), 1650–1654.

McCune, B., Grace, J. B., & Urban, D. L. (2002). *Analysis of ecological communities*. MjM software design Gleneden Beach, OR.

McLean, R., Tsyban, A., Burkett, V., Codignott, J., Forbes, D., Mimura, N., … Ittekkot, V. (2001). *Coastal Zones and Marine Ecosystems*. Intergovernmental Panel on Climate Change. Retrieved from http://papers.risingsea.net/IPCC.html

Miche, A. (1951). *{Le pouvoir r'efl'echissant des ouvrages maritimes expos'e}s à l'action de la houle*.

NASA. (2019). Earth Observing System Data and Information System (EOSDIS). Retrieved from https://worldview.earthdata.nasa.gov/

Nelson, R. (1997). *Height limits in top down and bottom up wave environments*.

Nicholls, R. J., Wong, P. P., Burkett, V., Codignotto, J., Hay, J., McLean, R., … Arblaster, J. (2007). Coastal systems and low-lying areas.

Olij, D. J. C. (2015). *Wave climate reduction for medium term process based morphodynamic simulations* (Doctoral dissertation, TU Delft repository).

Pandolfi, J. M., Connolly, S. R., Marshall, D. J., & Cohen, A. L. (2011). Projecting coral reef futures under global warming and ocean acidification. *Science*, *333*(6041), 418–422.

Pearson, S., van der Lugt, M., van Dongeren, A., Hagenaars, G., & Burzel, A. (2018). *Quick-scan runup reduction through coral reef restoration in the Seychelles*.

Pearson, Reniers, A., van Dongeren, A. R., Tissier, M. F. S., & den Heijer, C. (2016). *Predicting Wave-Induced Flooding on Low-Lying Tropical Islands*. TU Delft. Delft.

Pearson, Storlazzi, C. D., van Dongeren, A. R., Tissier, M. F. S., & Reniers, A. (2017). A Bayesian-based system to assess wave-driven flooding hazards on coral reef-lined coasts. *Journal of Geophysical Research: Oceans*, *122*(12), 10099–10117.

Péquignet, A. C. N., Becker, J. M., Merrifield, M. A., & Aucan, J. (2009). Forcing of resonant modes on a fringing reef during tropical storm Man-Yi. *Geophysical Research Letters*, *36*(3), n/a–n/a. doi:10.1029/2008gl036259

Pham, D. T., & Afify, A. A. (2007). Clustering techniques and their applications in engineering. *Journal of Mechanical Engineering Science*, *221*(11), 1445–1459. doi:10.1243/09544062JMES508

Pomeroy, A., Lowe, R., Symonds, G., van Dongeren, A., & Moore, C. (2012). The dynamics of infragravity wave transformation over a fringing reef. *Journal of Geophysical Research: Oceans (1978–2012)*, *117*(C11). doi:10.1029/2012JC008310

Quataert, E., Storlazzi, C., Van Rooijen, A., Cheriton, O., & Van Dongeren, A. (2015). The influence of coral reefs and climate change on wave-driven flooding of tropical coastlines. *Geophysical Research Letters*, *42*(15), 6407–6415. doi:10.1002/2015gl064861

Reddy, C. (2018). Understanding the concept of Hierarchical clustering Technique. Retrieved May 26, 2019, from https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

ReefBase. (2019). A Global Information System for Coral Reefs. Retrieved from http://www.reefbase.org

Roelvink, D. (2009). *XBeach Model Description and Manual Delft*. Unesco-IHE Institute for Water Education, Deltares and Delft University of Technology.

Roelvink, D. [Dano], McCall, R., Mehvar, S., Nederhoff, K., & Dastgheib, A. (2018). Improving predictions of swash dynamics in XBeach: The role of groupiness and incident-band runup. *Coastal Engineering*, *134*, 103–123. doi:10.1016/J.COASTALENG.2017.07.004

Roelvink, D. [Dano], van Dongeren, A., McCall, R., Hoonhout, B., van Rooijen, A., van Geer, P., … Nederhoff, K. (2015). Xbeach Manual, 135. doi:10.1590/S0100-54052009000400003

RTÉ. (2019). Global CO2 levels at highest ever recorded levels. Retrieved from https://www.rte.ie/news/2019/0514/1049289-co2%7B%5C_%7Drecord%7B%5C_%7Dlevel/

Schnieder, S., Patwardhan, A., Semenov, S., Burton, I., Magadza, C., Oppenheimer, M., … Suarez, A. (2007). Assessing key vulnerabilities and the risk from climate change. *Climatic change*, 779–810.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

SDD. (2016). Population Statistics. Retrieved from https://sdd.spc.int/en/stats-by-topic/population-statistics

Sheppard, C., Dixon, D. J., Gourlay, M., Sheppard, A., & Payet, R. (2005). Coral mortality increases wave energy reaching shores protected by reef flats: Examples from the Seychelles. *Estuarine, Coastal and Shelf Science*, *64*(2), 223–234. doi:https://doi.org/10.1016/j.ecss.2005.02.016

Slangen, A. B. A., Carson, M., Katsman, C. A., Van de Wal, R. S. W., Köhl, A., Vermeersen, L. L. A., & Stammer, D. (2014). Projecting twenty-first century regional sea-level changes. *Climatic change*, *124*(1-2), 317–332.

Smit, P. B., Stelling, G. S., Roelvink, D., van Thiel de Vries, J., McCall, R., van Dongeren, A., … Jacobs, R. (2014). XBeach: Non-hydrostatic model. *Report, Delft University of Technology and Deltares, Delft, The Netherlands*.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The challenges of clustering high dimensional data. In *New directions in statistical physics* (pp. 273–309). Springer.

Stockdon, H. F., Holman, R. A., Howd, P. A., & Sallenger, A. H. (2006). Empirical parameterization of setup, swash, and runup. doi:10.1016/j.coastaleng.2005.12.005

Storlazzi, Gingerich, S. B., van Dongeren, A., Cheriton, O. M., Swarzenski, P. W., Quataert, E., … McCall, R. (2018). Most atolls will be uninhabitable by the mid-21st century because of sea-level rise exacerbating wave-driven flooding. *Science Advances*, *4*(4), eaap9741. doi:10.1126/sciadv.aap9741

Storlazzi, Reguero, B. G., Cole, A. D., Lowe, E., Shope, J. B., Gibbs, A. E., … Beck, M. W. (2019). *Rigorously valuing the role of U.S. coral reefs in coastal hazard risk reduction*. doi:10.3133/ofr20191027

Su, S.-F., & Ma, G. (2018). Modeling two-dimensional infragravity motions on a fringing reef. *Ocean Engineering*, *153*, 256–267. doi:https://doi.org/10.1016/j.oceaneng.2018.01.111

Symonds, G., Huntley, D. A., Symonds, G., & Bowen, A. J. (1982). Two-Dimensional Surf Beat: Long Wave Generation by a Time-Varying Breakpoint. *Geophysical Research Atmospheres*, *87*(C1), 492–498. doi:10.1029/JC087iC01p00492

Teknomo, K. (2015). Similarity Measurement. Retrieved from https://people.revoledu.com/kardi/tutorial/Similarity/CityBlockDistance.html

Temmerman, S., Meire, P., Bouma, T. J., Herman, P. M. J., Ysebaert, T., & De Vriend, H. J. (2013). Ecosystem-based coastal defence in the face of global change. *Nature*, *504*(7478), 79.

Terry, J. P. (2007). *Tropical cyclones: climatology and impacts in the South Pacific*. Springer Science & Business Media.

Terry, J. P., & Falkland, A. C. (2010). Responses of atoll freshwater lenses to storm-surge overwash in the Northern Cook Islands. *18*(3), 749–759. doi:10.1007/s10040-009-0544-x

The World Bank. (2012). Pacific Islands: Disaster Risk Reduction and Financing in the Pacific. Retrieved from http://projects-beta.worldbank.org/en/results/2012/04/01/pacific-islands-disaster-risk-reduction-and-financing-in-the-pacific

Tomás, A., Méndez, F. J., Medina, R., Jaime, F. F., Higuera, P., Lara, J. L., … Álvarez de Eulate, M. F. (2016). A methodology to estimate wave-induced coastal flooding hazard maps in Spain. *Journal of Flood Risk Management*, *9*(3), 289–305.

Tubbataha Reefs Natural Park. (2018). Tubbataha Reefs Natural Park. Retrieved from http://www.tubbatahareef.org/wp/formation

UNFPA. (2014). *Population and Development Profiles: Pacific Island Countries*. United Nations Population Fund.

UNISDR. (2004). Terminology: basic terms of disaster risk reduction. United Nations International Strategy for Disaster Reduction (UNISDR) Geneva.

UNISDR. (2015). *Sendai Framework for Disaster Risk Reduction 2015 - 2030*. United Nations Office for Disaster Risk Reduction. Retrieved from https://www.preventionweb.net/files/43291%7B%5C_%7Dsendaiframeworkfordrren.pdf

United Nations. (2005). Hyogo framework for action 2005–2015: building the resilience of nations and communities to disasters. *World Conference on Disaster Reduction in Kobe, Japan*.

United Nations. (2006). *Global Survey of Early Warning Systems: An assessment of capacities, gaps and opportunities towards building a comprehensive global early warning system for all natural hazards*. Retrieved from www.unisdr.orgwww.unisdr-earlywarning.org

van Dongeren, A., Lowe, R., Pomeroy, A., Trang, D. M., Roelvink, D., Symonds, G., & Ranasinghe, R. (2013). Numerical modeling of low-frequency wave dynamics over a fringing coral reef. *Coastal Engineering*, *73*, 178–190. doi:https://doi.org/10.1016/j.coastaleng.2012.11.004

Vetter, O., Becker, J. M., Merrifield, M. A., Pequignet, A., Aucan, J., Boc, S. J., & Pollock, C. E. (2010). Wave setup over a Pacific Island fringing reef. *Journal of Geophysical Research: Oceans*, *115*(C12).

Vitousek, S., Barnard, P. L., Fletcher, C. H., Frazer, N., Erikson, L., & Storlazzi, C. D. (2017). Doubling of coastal flooding frequency within decades due to sea-level rise. *Nature*, *7*(1), 1399.

Watson, C. S., White, N. J., Church, J. A., King, M. A., Burgette, R. J., & Legresy, B. (2015). Unabated global mean sea-level rise over the satellite altimeter era. *Nature Climate Change*, *5*(6), 565.

White, I., & Falkland, T. (2010). Management of freshwater lenses on small Pacific islands. *Hydrogeology Journal*, *18*(1), 227–246.

White, I., Falkland, T., Perez, P., Dray, A., Metutera, T., Metai, E., & Overmars, M. (2007). Challenges in freshwater management in low coral atolls. *Journal of Cleaner Production*, *15*(16), 1522–1528.

Willett, P. (1999). Dissimilarity-Based Algorithms For Selecting Structurally Diverse Sets Of Compounds. *Journal of Computational Biology*, 447–457. Retrieved from http://eprints.whiterose.ac.uk/77603/8/WRRO%7B%5C_%7D77603.pdf

Woodroffe, C. D. (2008). Reef-island topography and the vulnerability of atolls to sea-level rise. *Global and Planetary Change*, *62*(1-2), 77–96.

Wright, L. D., Chappell, J., Thom, B. G., Bradshaw, M. P., & Cowell, P. (1979). Morphodynamics of reflective and dissipative beach and inshore systems: Southeastern Australia. *Marine Geology*, *32*(1-2), 105–140.

Young, I. R. (1989). Wave transformation over coral reefs. *Journal of Geophysical Research: Oceans*, *94*(C7), 9779–9789. doi:doi:10.1029/JC094iC07p09779

A

Reef Profile Dataset

A

A.1. REEF PROFILE LOCATIONS

The dataset used for this study was provided by the USGS. A recent report which used the dataset for the generation of a coastal hazard tool (Storlazzi et al., 2019) provides details of the coral reef dataset. Table A.1 shows the number of profiles included in the dataset from each main location, as well as separating the number of profiles via sub location.

Table A.1: Number of measured cross-shore transects for each location. Source: (Storlazzi et al., 2019)

Location	Sublocation	Number of cross-shore transects
American Samoa	Tutuila	1,004
American Samoa	Ofu and Olosega	196
American Samoa	Ta'ū	275
Northern Mariana Islands	Saipan	585
Northern Mariana Islands	Tinian	450
Guam	Guam	1,295
Florida	Dry Tortugas	300
Florida	Key West	545
Florida	Florida Keys	1,127
Florida	Miami	1,139
Florida	Palm Beach	1,168
Hawai'i	Island of Hawai'i	4,582
Hawai'i	Maui	2,087
Hawai'i	Lāna'i	759
Hawai'i	Moloka'i	2,886
Hawai'i	Kaho'olawe	456
Hawai'i	Kaua'i	1,455
Hawai'i	Ni'ihau	677
Hawai'i	O'ahu	1,997
Puerto Rico	Isla de Puerto Rico	4,588
Puerto Rico	Isla de Culebra	244
Puerto Rico	Isla de Vieques	687
Virgin Islands	Saint Croix	803
Virgin Islands	Saint John	396
Virgin Islands	Saint Thomas	466

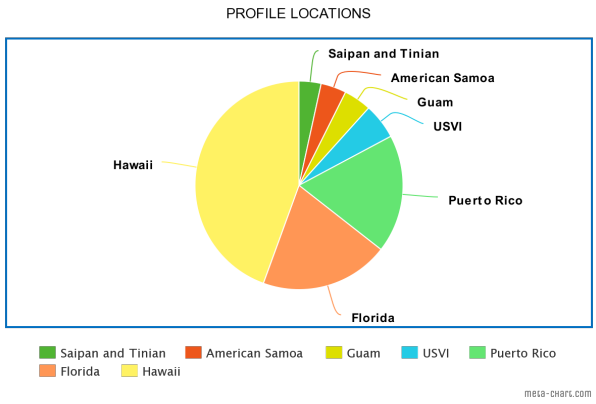


Figure A.1: Distribution of profile locations included in the initial dataset provided by Storlazzi et al. (2019).

A.2. REEF PROFILE SOURCES

The sources of the coral reef profile measurements are also provided by Storlazzi et al. (2019). Table A.2 gives the year and source for each set of measured profiles that together make up the entire dataset used in this study. The oldest source is from 2001, and the most recent is from 2016.

Table A.2: The sources of the measured reef profiles used in this study. Source: (Storlazzi et al., 2019)

[NGDC, National Geophysical Data Center; NOAA, National Oceanic and Atmospheric Administration; PacIOOS, Pacific Islands Ocean Observing System; PIBHMC, Pacific Islands Benthic Habitat Mapping Center]

Location	Sublocation	Data source
American Samoa	Tutuila	Carignan and others, 2013
American Samoa	Ofu, Olosega, and Ta'ū	Lim and others, 2010
Northern Mariana Islands	Saipan	PIBHMC, 2007a Amante and Eakins, 2009 PacIOOS, 2016a
Northern Mariana Islands	Tinian	PIBHMC, 2007b Amante and Eakins, 2009 PacIOOS, 2016b
Guam	Guam	Chamberlin, 2008
Florida	Dry Tortugas	NGDC, 2001
Florida	Key West	Grothe and others, 2011
Florida	Florida Keys	NGDC, 2001
Florida	Miami	Carignan and others, 2015
Florida	Palm Beach	NGDC, 2001
Hawai'i	Island of Hawai'i	NGDC, 2005
Hawai'i	Hilo	Love and others, 2011a
Hawai'i	Kawaihae	Carignan and others, 2011a
Hawai'i	Keauhou	Carignan and others, 2011b
Hawai'i	Maui Nui	NGDC, 2005
Hawai'i	Maui	NOAA, 2016 Taylor and others, 2008a
Hawai'i	Lāna'i	NGDC, 2005
Hawai'i	Moloka'i	NGDC, 2005
Hawai'i	Kaho'olawe	NGDC, 2005
Hawai'i	Kaua'i	Friday and others, 2012
Hawai'i	Ni'ihau	Friday and others, 2012
Hawai'i	O'ahu	Love and others, 2011b
Puerto Rico	Arecibo	Taylor and others, 2008c
Puerto Rico	Isla de Culebra	Taylor and others, 2008b
Puerto Rico	Fajardo	Taylor and others, 2008d
Puerto Rico	Guayama	Taylor and others, 2008e
Puerto Rico	Mayagüez	Taylor and others, 2008f
Puerto Rico	Ponce	Taylor and others, 2008g
Puerto Rico	San Juan	Taylor and others, 2008h
Puerto Rico	Isla de Vieques	Taylor and others, 2008b
Virgin Islands	Saint Croix	Love and others, 2014a
Virgin Islands	Saint Thomas	Love and others, 2014b
Virgin Islands	Saint John	Love and others, 2014b

B

Data Pre-Processing

B.1. OMITTED PROFILES

The profiles that did not satisfy the requirements explained in Section 3.1.1 and were omitted from the analysis are shown here, separated by region. These profiles either have a peak in the profile above 4 m elevation seaward of the coastline, or have all data points above or below MSL.

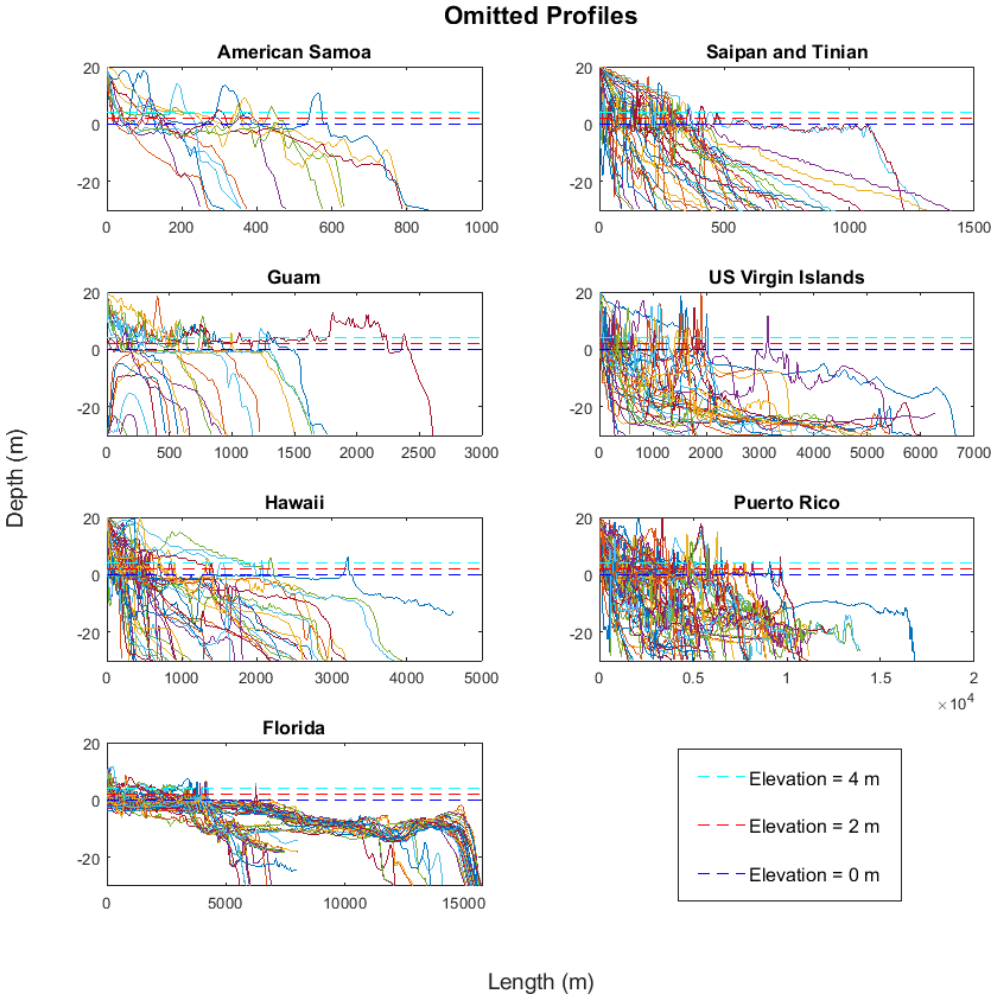


Figure B.1: The skipped profiles from all locations included in this study. The profiles are omitted because they either never reach above or below MSL, or have a peak in the profile above 4 m elevation that is not considered land since the above MSL elevation is not maintained for 100 m in width.

B.2. PROFILE STATISTICS BY LOCATION

To gain an understanding of the typical characteristics of the reef profiles from each region, the profiles were plotted together, with the mean and median profiles, as well as the standard deviation and the envelopes. This is shown in Figure B.2.

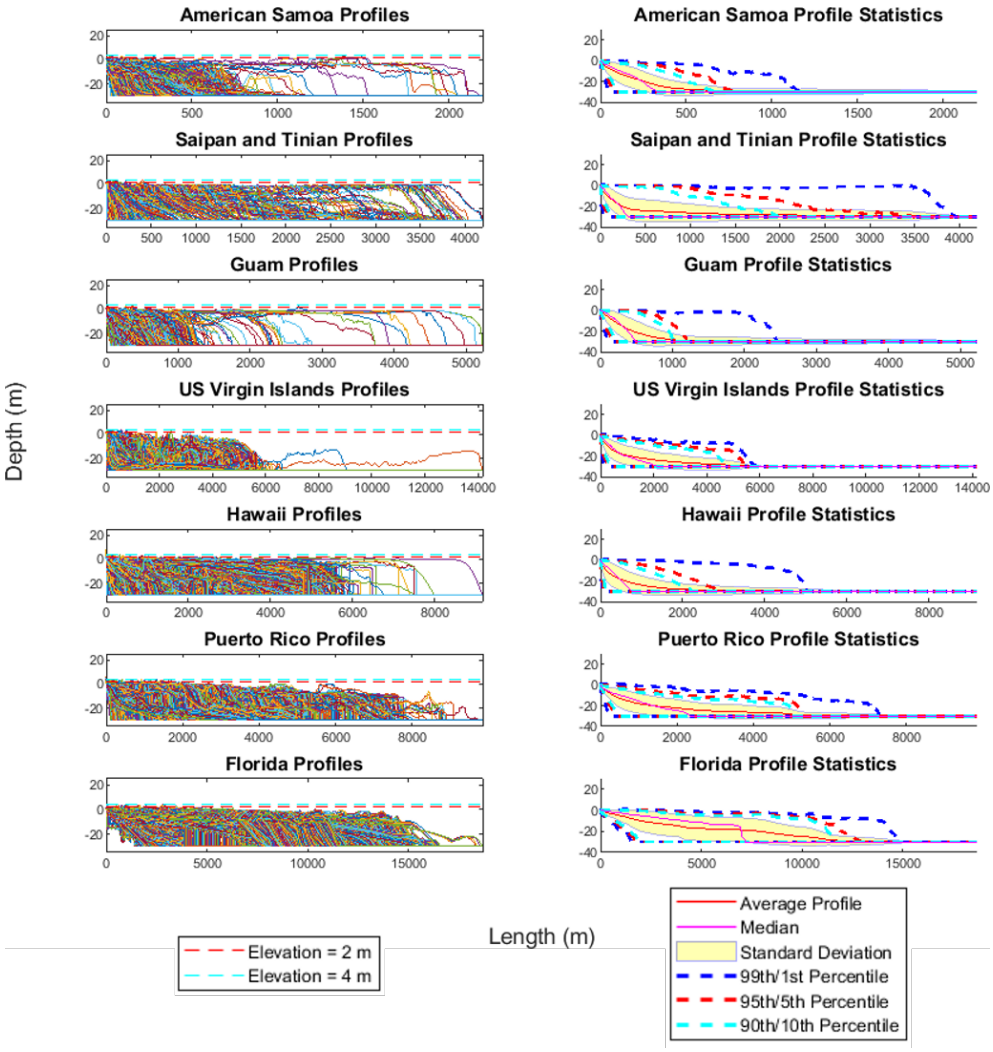


Figure B.2: Profile Statistics from all locations

B.3. CORAL COVER

The dataset of reef profiles contains information on spatial coral cover along the profile. This information was analyzed to determine if there is relationship between coral cover and depth. Specifically, the depth corresponding to the last documented coral cover along the profile was checked, to see if there is a depth to which a higher friction coefficient could be used in the XBeach simulations to represent the coral. Unfortunately, not useful relationship was found, as shown in Figure B.3. The histogram shows that a large portion of the profiles have no documented coral cover, and those that do are uniformly distributed across the depth values. Therefore, a conclusion could not be drawn as to where coral most likely is on the profile and where it is not, resulting in a uniform coefficient of friction to be used in the XBeach simulations.

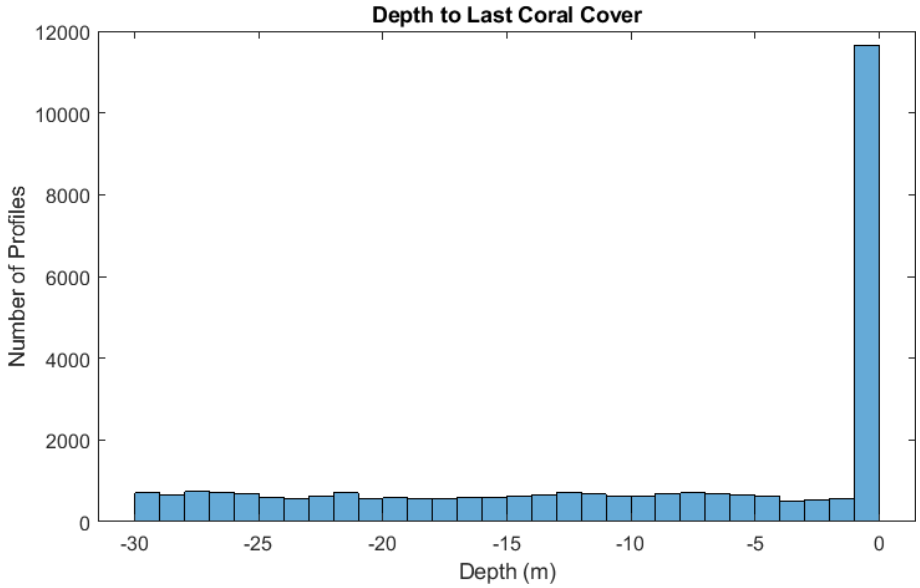


Figure B.3: Depth to the last coral cover along all profiles in the dataset.

C

Cluster Analysis

C

C.1. PLOTTING METHODS

BOXPLOTS

A boxplot is used to provide visualization of summary statistics for a set of data. These are used in Section 4 to compare the results of the different clustering algorithms. An example of a boxplot is shown in Figure C.1 subplot (a).

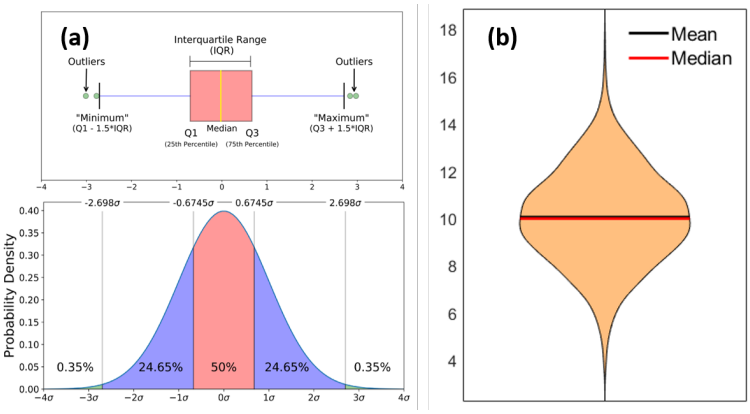


Figure C.1: An example of a boxplot with a normally distributed dataset. Source: (Galarnyk, 2018)

The middle line represents the median, and the box represents the 25th and 75th percentiles of the dataset, which is known as the interquartile range (IQR). Whiskers are extended to a value of 1.5 times the IQR on either end of the box. For a normally distributed dataset, the whiskers extend to cover 99.3% of the data points. Any values that fall outside of the whiskers are deemed outliers.

VIOLIN PLOTS

A violin plot is used in similar situations as boxplots, but also have the ability to show the distribution of the data points and the probability density. An example of a violin plot of a random set of normally distributed data is shown in FigureC.1 subplot (b). The black line shows the mean and the red line shows the median. The shape of the shaded region represents the distribution of data points at each y-axis value.

C.2. CLUSTER ANALYSIS OF REEF MORPHOLOGY

The first round of cluster analysis was done using reef morphology, and the clustering algorithms K-means, K-medians, K-medoids and Gaussian Mixture Models were used. Additional information on these algorithms and evaluation methods are presented in this section.

C.2.1. DISTANCE METHODS

SQUARED EUCLIDEAN DISTANCE

The squared Euclidean distance (SED) is popular method used for estimating parameters of statistical models. As its name implies, it is simply the square of the euclidean distance, which is the Pythagorean theorem applied to multiple dimensions (McCune, Grace, & Urban, 2002). Figure C.2 a shows an example of euclidean distance in 2D. SED is the main method to be used with K-means clustering (Berkhin, 2002) since K-means works by minimizing square errors (within-cluster variance) and therefore the SED corresponds directly. It can also be used for many other clustering methods. The equation to describe SED is shown here:

$$d(x,c) = (x-c)(x-c)'$$

where x is an observation and c is a centroid. A weighted Euclidean distance can also be used by redefining the x values.

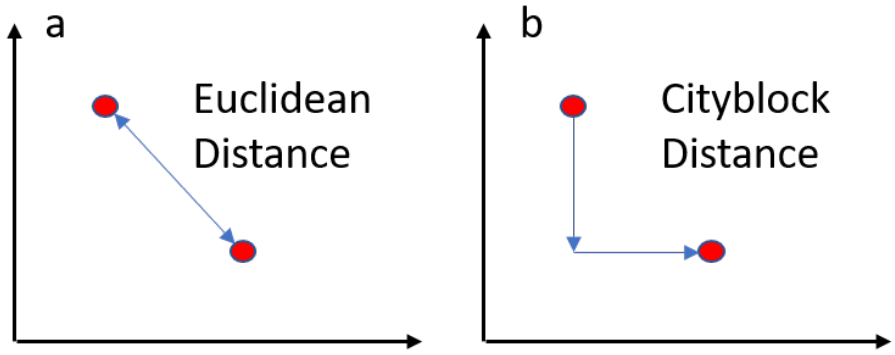


Figure C.2: Illustration of euclidean distance vs cityblock distance in 2D.

CITY BLOCK DISTANCE

City block distance, also known as Manhattan distance, boxcar distance, and absolute value distance, represents distance between points in a city road grid. It compares the absolute differences between a pair of objects (Teknomo, 2015). The equation to describe city block distance is shown here:

$$d(x,c) = \sum_{j=1}^{p} |x_j - c_j|$$

where x is an observation, c is a centroid, and j is one of the p variables that make up the observation. Figure C.2 b shows an example of cityblock distance in 2D. This distance metric was applied when using K-medians. For this dataset, since all profile measurements are aligned in the X coordinate system, the cityblock distance in this case is simply the sum of absolute difference in depth between two profiles at each cross-shore point. Figure C.3 represents the application of cityblock distance using two profiles from this study, in which the object dissimilarity for these two profiles would be the sum of all of the red lines representing the cityblock (absolute) difference between the profiles at each measured cross-shore point.

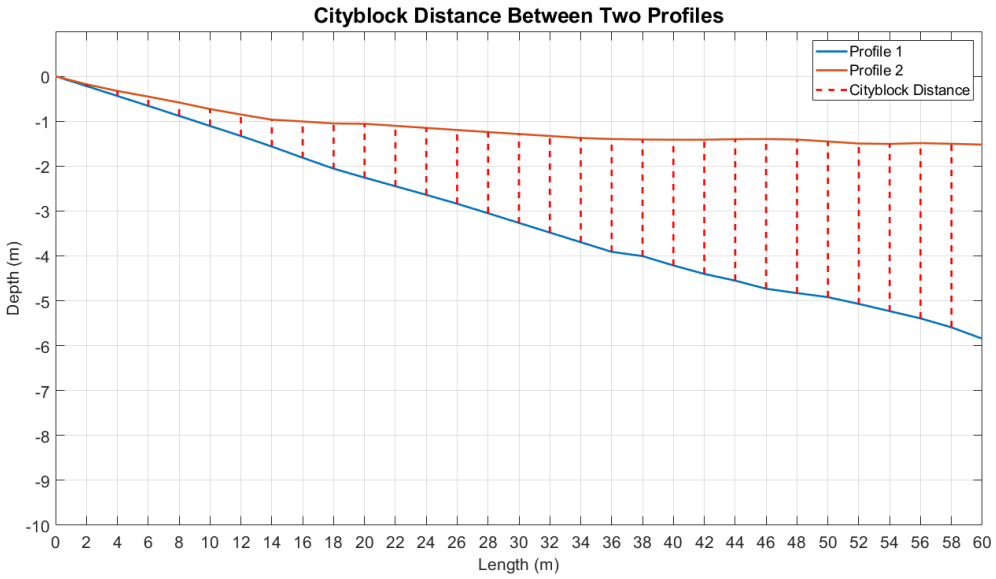


Figure C.3: Example of the cityblock distance metric with two reef profiles.

MIXTURE MODELS

Different to the two methods stated above, mixture models are used for probabilistic clustering techniques. In these techniques, statistics and probabilities are used to allocate a percentage that the observation belongs to each cluster rather than a hard division. Therefore, the calculation of dissimilarity is done with probability distributions.

For probabilistic clustering techniques, the mixture model is a useful tool for density estimation. The general concept is fitting M probabilistic distributions to a dataset to collectively make a mixture distribution $f(x)$. Mixture models can use any component densities instead of Gaussian, but the Gaussian mixture model is by far the most popular (Friedman et al., 2001). The Gaussian mixture model has the form

$$f(x) = \sum_{m=1}^{M} a_m \phi(x; \mu_m, \Sigma_m)$$

with mixing proportions a_m,$\sum_m a_m = 1$ and each Gaussian density has a mean μ_m and a covariance matrix $\sum_m$. The use of mixture models for clustering purposes is explained in Section 2.4.2. An example of data grouped with two Gaussian distributions is shown in Figure C.4.

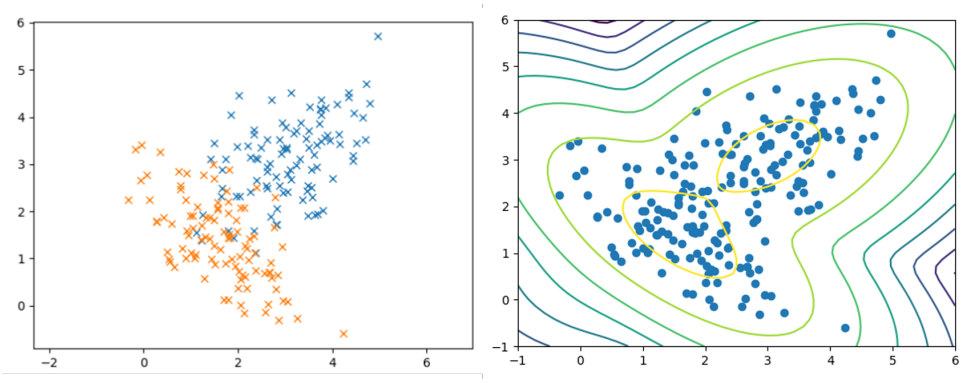


Figure C.4: Example of a mixture model being used to group data into 2 clusters using a Gaussian mixture model. Source: (Deshpande, 2017)

C.2.2. CLUSTER EVALUATION

EVALUATION METHODS

Typical methods to evaluate clusters include the Calinski Harabasz (CH) and Davies-Bouldin (DB) evaluations, as well as the AIC and BIC criterion. The CH and DB evaluations are intended to apply towards the hard partitioning methods (K-means, K-medians, K-medoids) and the AIC and BIC criterion work best for the probabilistic method (GMM).

The evaluation methods used for each clustering technique are shown in Table C.1.

Table C.1: Clustering evaluation methods used for each type of clustering algorithm

	Evaluation Criterion			
Method	**Calinski Harabasz**	**Davies-Bouldin**	**AIC and BIC**	**Average Errors**
K-means	✓	✓	x	✓
K-medoids	✓	✓	x	✓
K-medians	✓	✓	x	✓
Gaussian Mixture	x	x	✓	✓

A definition of each of the methods is provided below.

CALINSKI HARABASZ EVALUATION

The Calinski-Harabasz (CH) criterion is referred to as the variance ratio criterion (VRC). It is defined as

$$VRC_K = \frac{SS_B}{SS_w} \times \frac{(N-K)}{(K-1)}$$

where SS_B is the total between-cluster variance, SS_W is the total within-cluster variance, K is the number of clusters, and N is the number of observations. Successful clusters will have a large between-cluster variance (SS_B) and a small within-cluster variance (SS_W), which translates to a high VRC_K ratio. This evaluation is noted to be best suited for use with K-means clustering with squared Euclidean distance (Caliński & Harabasz, 1974).

The total between-cluster variance SS_B is defined as

$$SS_B = \sum_{i=1}^{K} n_i ||m_i - m||^2$$

in which n_i is the number of observations within the cluster i, m_i is the centroid of cluster i, m is the total mean of the sample data, and $||m_i - m||$ is the Euclidean distance between the two vectors. The total within-cluster variance SS_W is defined as

$$SS_W = \sum_{i=1}^{K} \sum_{x \in c_i} ||x - m_i||^2$$

where x is a data point, c_i is the ith cluster, m_i is the centroid of the cluster i, and $||x - m_i||$ is the Euclidean distance between the two vectors.

DAVIES-BOULDIN EVALUATION

The Davies-Bouldin (DB) index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation (Maulik & Bandyopadhyay, 2002). A large value of $D_{i,j}$ means that the within-to-between cluster ratio for cluster i is bad, and therefore the optimal clustering solution will have the smallest DB index value (Davies & Bouldin, 1979).

The DB index is defined as

$$DB = \frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} \{D_{i,j}\}$$

where $D_{i,j}$ represents the within-to-between cluster distance ratio for the ith and jth clusters. $D_{i,j}$ is defined as

$$D_{i,j} = \frac{(\bar{d}_i + \bar{d}_j)}{d_{i,j}}$$

$\bar{d}_i$ is the average distance between the points within the ith cluster and the ith cluster centroid. $\bar{d}_j$ is the average distance between the points within the jth cluster and the jth cluster centroid. Finally, $d_{i,j}$ is the Euclidean distance between the centroids of the ith and jth clusters.

BIC AND AIC OF GAUSSIAN MIXTURE

The Akaike's Information Criterion (AIC) (Akaikei, 1973) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) are tools used to assess the fit of a model. They consist of a goodness-of-fit term, plus a penalty to test over-fitting, providing a standardized method to balance sensitivity and specificity (Dziak, Coffman, Lanza, & Runze, 2012).

The two models search for the optimized log likelihood function value, but penalize in different ways. AIC and BIC both penalize more complex models, but BIC does so with a function of the sample size, and therefore it typically penalizes more than AIC (Geurts, Box, & Jenkins, 2006). Models that minimize AIC and BIC should be used. The AIC is described as

$$AIC = -2(logL) + 2(numParam)$$

and BIC is described as

$$BIC = -2(logL) + numParam * log(numObs)$$

where $logL$ is the log likelihood function value, $numParams$ is the number of parameters associated with the fitted model, and $numObs$ is the number of observations (sample size) associated with the fitted model.

EVALUATION RESULTS

The Calinksi Harabasz, Davies Bouldin, AIC and BIC evaluation methods provided interesting results, but due to the nature of the dataset and vast range of the number of tested cluster groups, they provided less information on the most effective clustering technique compared to the direct evaluation methods. The results from these methods are presented in this section.

INITIAL RUNS

Calinski Harabasz Evaluation

The Calinski Harabasz (CH) criterion (Section C.2.2) was used to evaluate the K-means, K-medoids, and K-medians methods. The Gaussian mixture results were excluded from this method since it is intended for hard clustering methods. A good clustering output will have a high CH score. The results are shown in Figure C.5.

The results suggests that the fewer the number of clusters, the better the result which is opposite to what was expected. This is due to the nature of the equation which applies a heavy weighting to the results that use fewer clusters. The results follow the same pattern of the weighting function on its own, leading to the finding that the variance within clusters and between clusters is not changing with the same proportion as the weighting function when the number of clusters increases. Therefore, the CH method favours the fewest number of clusters. However, the efficiency is not the only aspect of the result that is of interest. Perhaps a more important value is a certain limit of variance within each group, and therefore the CH method does not provide very meaningful results for this study.

Davies-Bouldin Evaluation

The Davies-Bouldin (DB) criterion (Section C.2.2) was also used to evaluate the K-means, K-medoids, and K-medians methods. For similar reasons, the Gaussian mixture results were excluded from this method. The results are shown in Figure C.6. A low DB value suggests that the result is good.

Similar to the CH method, the DB compares within cluster variance to between cluster variance. However, there is not the same weighting applied favouring fewer clusters. The result suggests that using 10 clusters is better than 50, 100 and 150, but then it follows a pattern that was expected in which more clusters provides a better result, as the

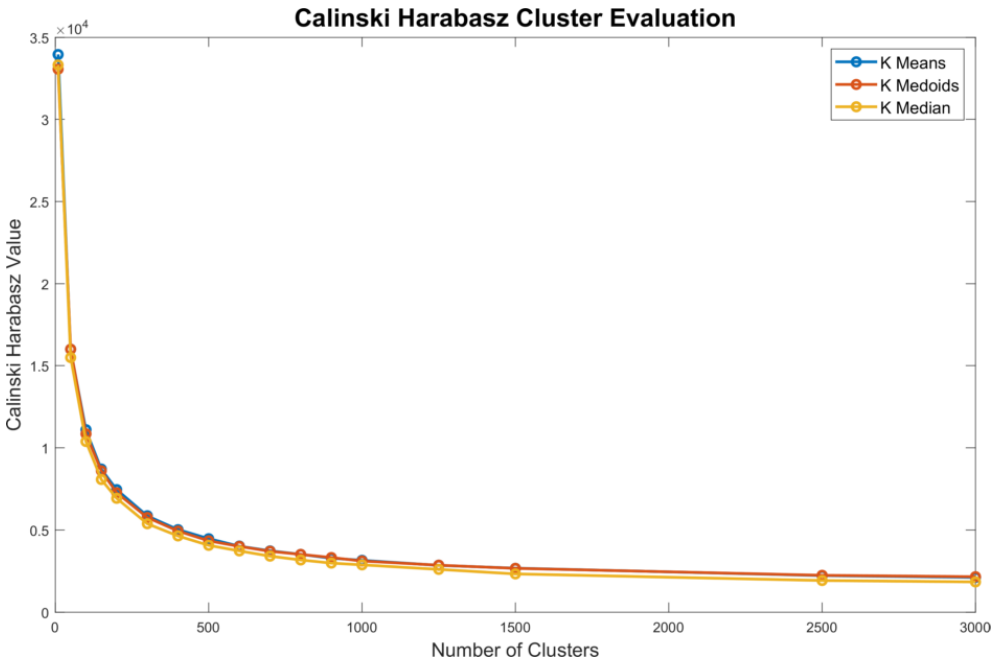


Figure C.5: Calinski Harabasz evaluation criterion for each number of clusters tested for K-means, K-medoids, and K-medians. A higher number suggests the best fit.

criterion decreases with a higher number of clusters. The peak at 150 shows that at this value, the ratio of within cluster variance to between cluster variance in relation to the number of clusters used is at its worst, and that it is much better when using fewer clusters as well as when using many more clusters. K-medoids appears to be the method that provides the best results, followed by K-means and K-medians. This result provides useful information comparing these three methods, but basing the final value of K based on this result is not possible since the ratio of within to between cluster variance is not the most important criteria for this study.

AIC and BIC of Gaussian Mixture

A similar type of evaluation for the Gaussian mixture method is done by evaluating the AIC and BIC (Section C.2.2) values. These measure the goodness of fit via the log-likelihood function as well as provide penalties based on over-fitting. A lower value means that the cluster number is more optimal. The values for the two criteria are shown in Figure C.7.

Both the AIC and BIC follow the trend that the fewer number of clusters is more optimal, opposite to what was expected. This is most likely due to ... JOSE Notes.

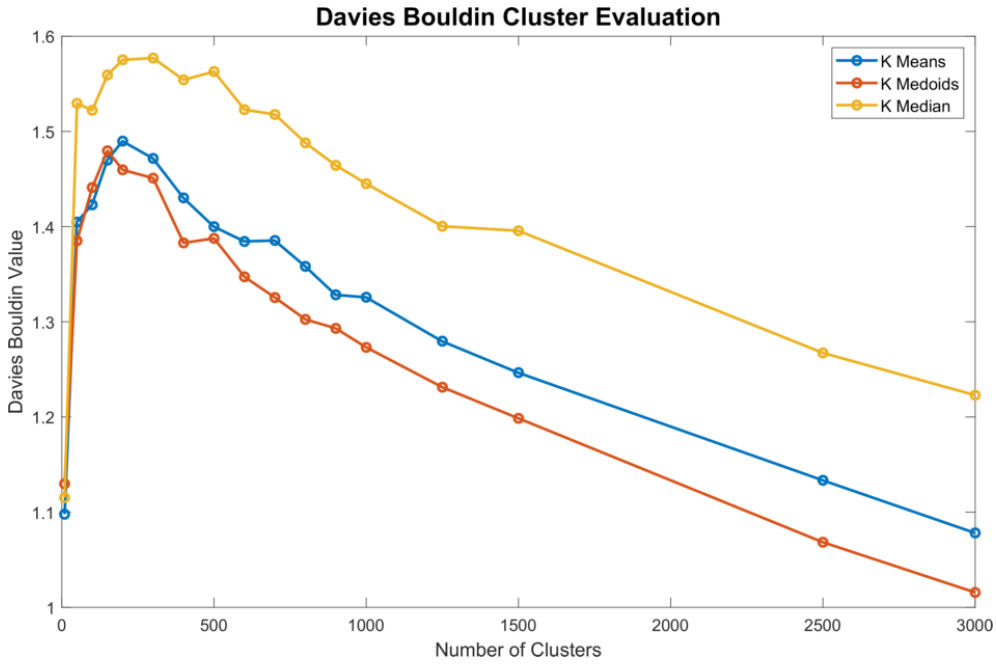


Figure C.6: Davies-Bouldin evaluation criterion for each number of clusters tested for K-means, K-medoids, and K-medians. A lower number suggests the best fit.

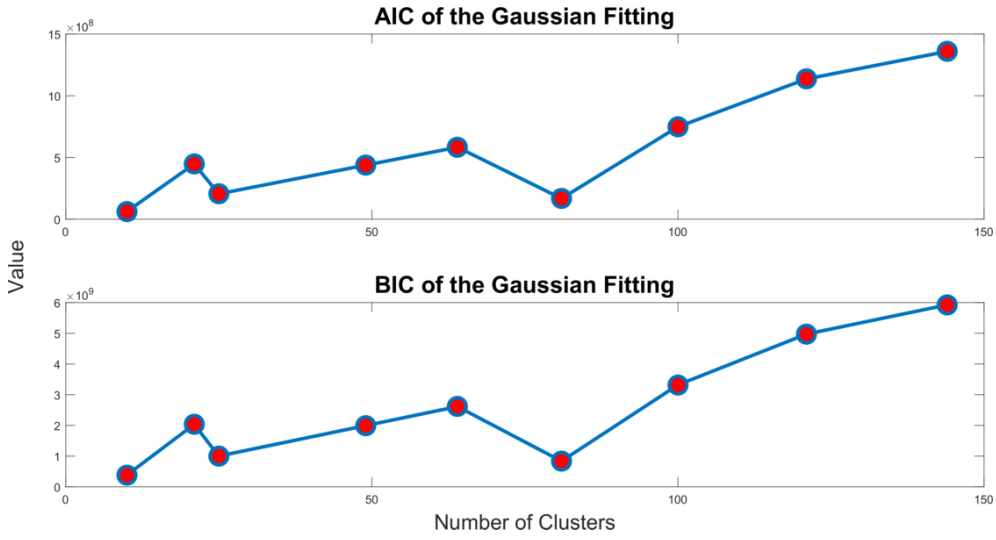


Figure C.7: AIC and BIC of Gaussian Mixture initial runs. A lower values suggests the best fit.

DETAILED RUNS

AIC and BIC of Gaussian Mixture

The addition of more replicates did not change the output of the AIC and BIC. The results follow the same trend of higher AIC and BIC values at higher number of clusters. The results are shown in Figure C.8.

Figure C.8: AIC and BIC of Gaussian Mixture for the detailed runs

C.2.3. CLUSTER ROUND 1 GROUPED PROFILES

The individual average absolute errors for the final cluster run are shown in Figure C.9. The histogram shows the distribution of the average difference between a profile and its cluster centroid. Almost all profiles have a mean difference less than 1 m to the cluster profile, and the vast majority have a mean difference of less than 0.2 m. This calculation includes the - 30 m depth values used to extend each profile to the longest profile included in the analysis, which would lower the average difference heavily for shorter profiles.

DISTRIBUTION OF LOCATIONS WITHIN CLUSTER ROUND 1 GROUPS

Figures C.10 to C.14 present the distribution of reef profile locations within each of the 500 Cluster Round 1 groups. The cluster groups were sorted by size (number of profiles within the group) for an easier representation. The figures show that generally, profiles from many locations are grouped together to form the cluster groups. There is a dominance from the Hawaii profiles, simply because there are many more included in the analysis compared to the other locations.

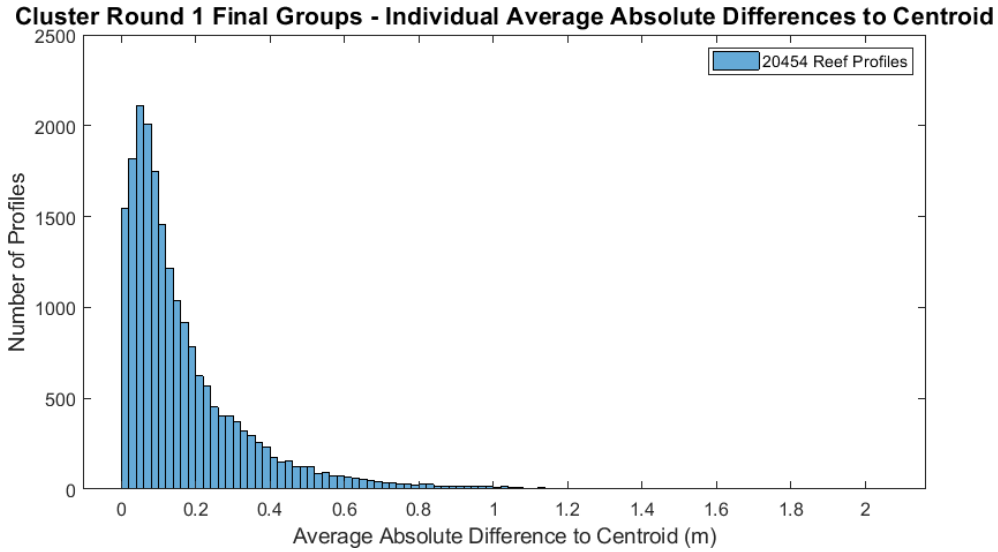


Figure C.9: The average difference between the profiles and their centroid for the final cluster run using K-medians and 500 cluster groups.

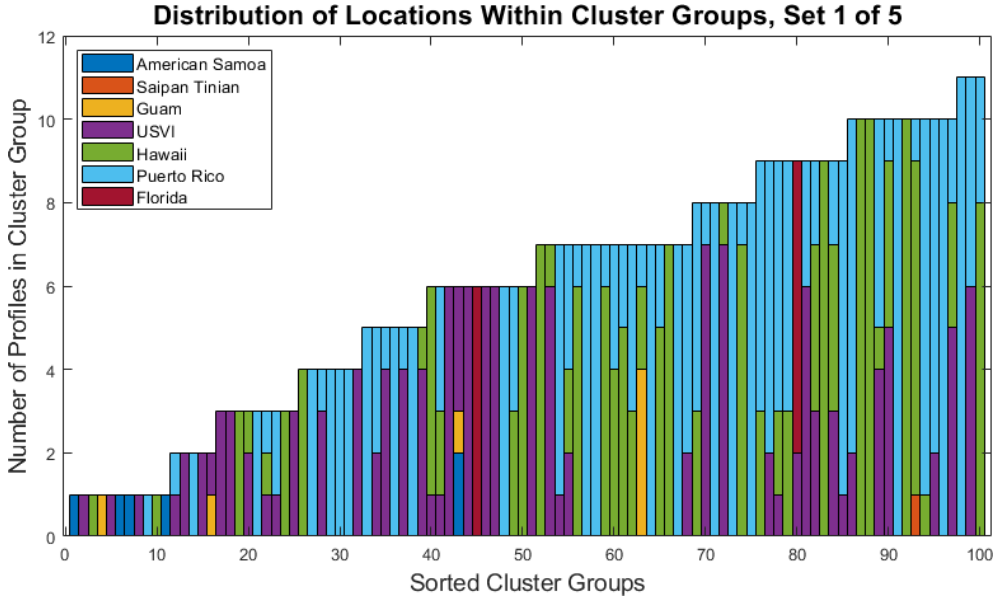


Figure C.10: Distribution of locations of the profiles within each cluster group, set 1 of 5.

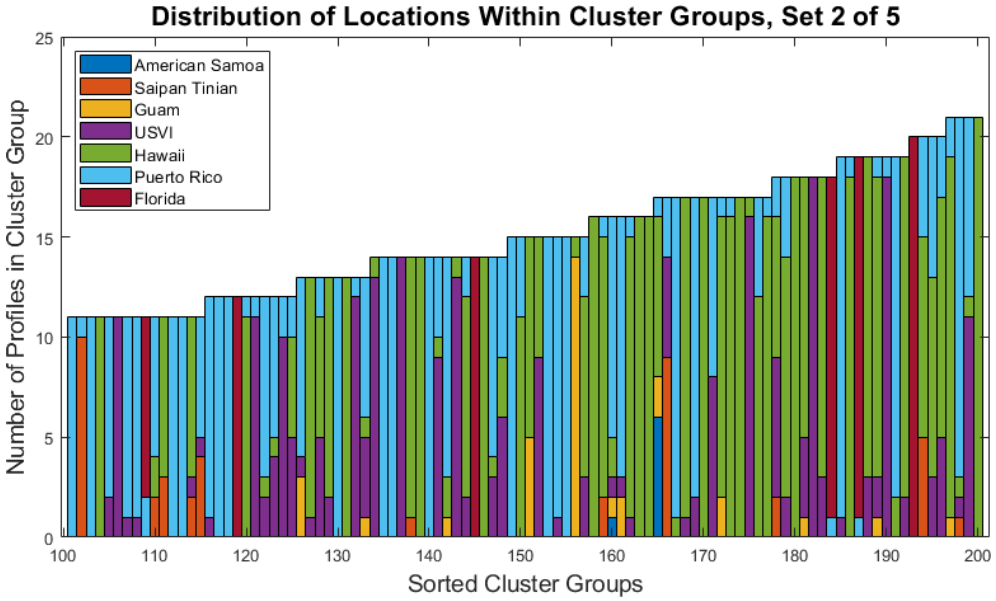


Figure C.11: Distribution of locations of the profiles within each cluster group, set 2 of 5.

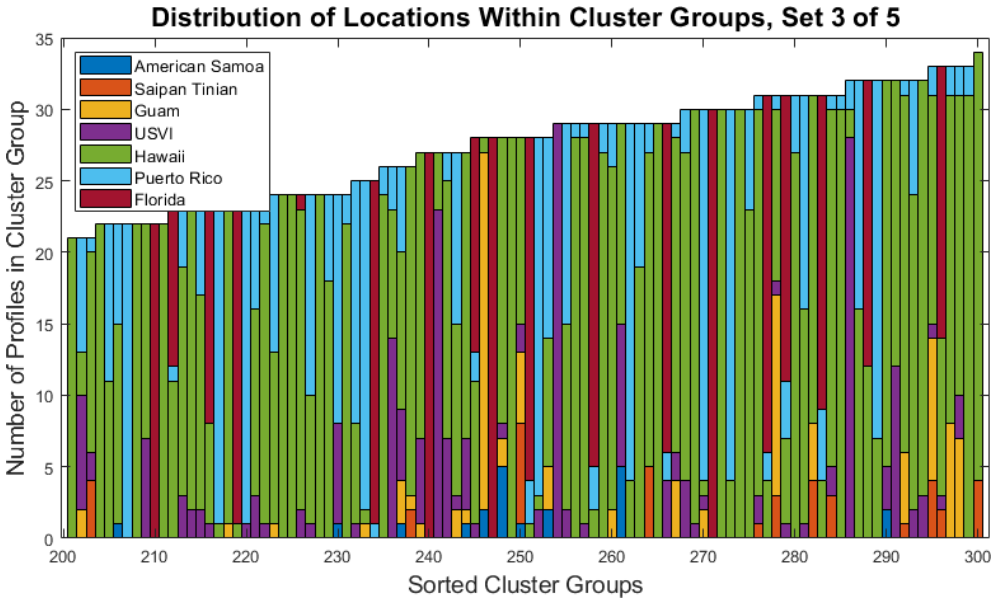


Figure C.12: Distribution of locations of the profiles within each cluster group, set 3 of 5.

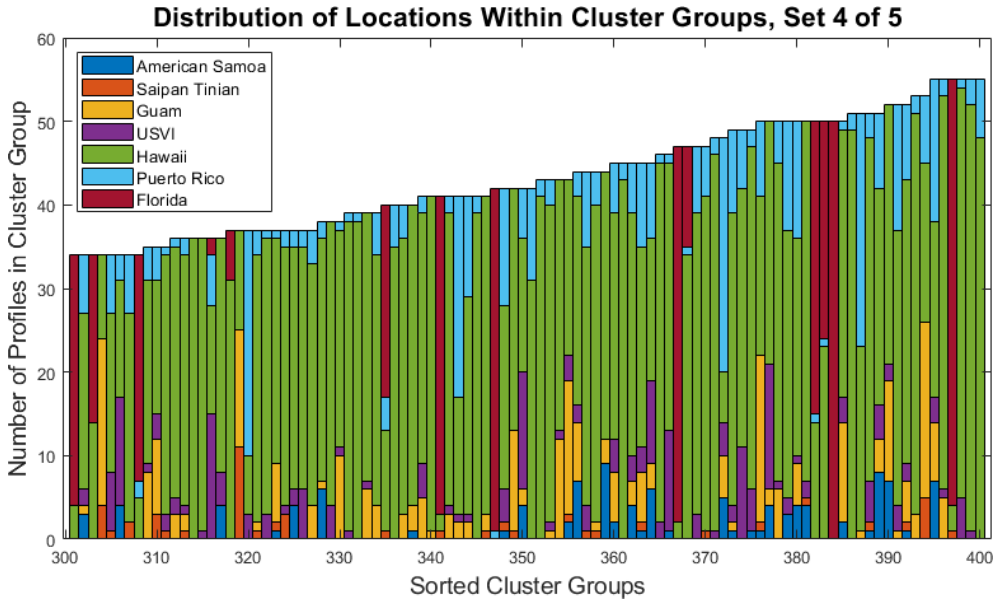


Figure C.13: Distribution of locations of the profiles within each cluster group, set 4 of 5.

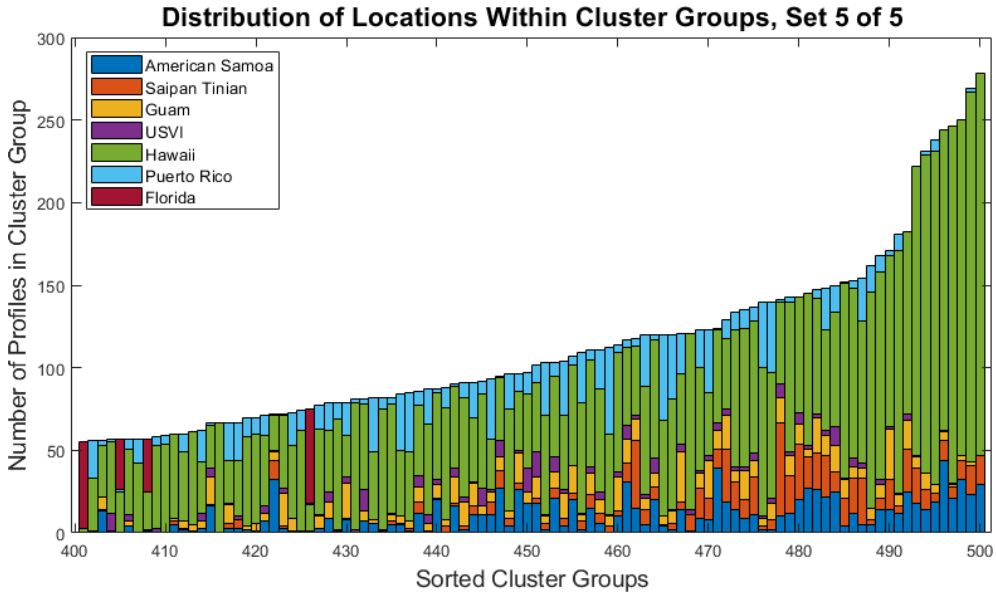


Figure C.14: Distribution of locations of the profiles within each cluster group, set of 5.

C.3. CLUSTER ROUND 2

C.3.1. HIERARCHICAL CLUSTERING DISTANCE METHODS

In agglomerative hierarchical clustering, at each stage of the hierarchy, clusters are merged together. If there are N individual observations, there will be $N-1$ levels in the hierarchy. At each of the $N-1$ steps the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level. Therefore, a measure of dissimilarity between two clusters (groups of observations) must be defined (Friedman et al., 2001).

Single

The single method is also known as the nearest neighbour technique since it measures the intergroup dissimilarity to be that of the closest (least dissimilar) pair of observations (Legendre & Legendre, 2012). An example is shown in Figure C.15 in the top right corner.

Complete

The complete distance method uses the most dissimilar pair of observations (furthest-neighbour technique) to measure dissimilarity (Legendre & Legendre, 2012). An example is shown in Figure C.15 in the bottom right corner.

Average

The average distance method, or the unweighted average distance (UPGMA) simply uses the average dissimilarity between groups. The method computes the arithmetic average of similarities or distances between all members of both clusters attempting to be merged (Legendre & Legendre, 2012). An example is shown in Figure C.15 in the top left corner.

Centroid

The centroid distance method, also known as unweighted centroid clustering (UPGMC), compares the centroids ("mean point") of the two groups and uses the distance between them to measure intergroup dissimilarity (Legendre & Legendre, 2012). An example is shown in Figure C.15 in the bottom left corner.

Median

The median distance method is also called the weighted centre of mass distance (WPGMC), where equal weights are given to the two clusters on the verge of merging, independent of the number of objects in each cluster (Legendre & Legendre, 2012). This method is meant to rid the problem of too much weight being applied to the types of samples which are included most in the sample set, since a cluster with 5 observations will be given equal weighting compared to a cluster with 100 observations.

Weighted

The weighted distance method is also referred to as the weighted arithmetic average clustering (WPGMA). It was designed to provide equal weighting to a smaller group of observations compared to larger groups of observations, since in the UPGMA method, the average distance could be distorted due to differences in cluster size. The method gives equal weights to the two branches of the dendrogram when computing similarities, which is equivalent to down-weighting the largest group (Legendre & Legendre, 2012).

Ward

The ward distance method is commonly referred to as the minimum variance method. It is related to the centroid method since it also leads to a geometric representation using

the cluster centroid. The main goal of the method is to minimize the squared error. At the beginning of the method, each observation is in a cluster of its own, and the distance to the centroid is 0. Once the method proceeds and observations are merged together, the centroids move away from initial positions and the sum of of the squared distances between the observations and the centroids increase. At each step, the pair of clusters or observations will be merged that result in the smallest increase as possible of the sum, over all observations, of the squared distances between observations and the cluster centroid (Legendre & Legendre, 2012).

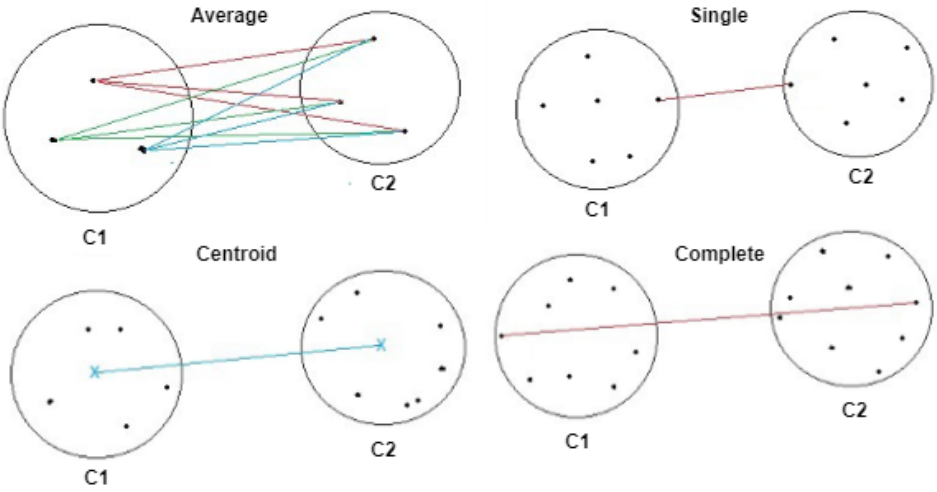


Figure C.15: Illustrative example of different distance methods used in hierarchical clustering. The average method is shown in the top left, the single method in the top right, the centroid method in the bottom left, and complete method in the bottom right. Source: (Reddy, 2018)

C.3.2. CLUSTER ANALYSIS ROUND 2 INPUTS

In Cluster Round 2, both the hydrodynamics and morphology were included as inputs into the clustering algorithm. Each were weighted with 50%. To apply the weighting, the inputs were normalized and then multiplied by the appropriate weight.

MORPHOLOGY

For all profiles there were 3,066 cross-shore positions of depth measurements. These points were normalized by subtracting the mean and dividing by the standard deviation (Jain & Dubes, 1988). To apply the 50% weighting, each cross shore point was multiplied by 0.00016 (0.5/3066).

HYDRODYNAMICS

The hydrodynamics consisted of the $R_{2\%}$, setup and swash. The swash was divided into infragravity and incident frequency bands. There were four values for each parameter because of the four wave loading conditions, totaling 16 hydrodynamics variables to be used as input into the clustering algorithm. These parameters were normalized the same

way, by subtracting the mean and dividing by the standard deviation. The 50% weighting was equally spread between the three different types of hydrodynamics, meaning that the $R_{2\%}$, setup and swash all were given 16.67% weighting. The swash then divided this evenly into two for the two different components.

The normalized and weighted morphology and hydrodynamics inputs were then combined into one matrix to be used in the clustering algorithm.

C.3.3. HIERARCHICAL DISTANCE METHODS AND METRICS ANALYSIS

The application of hierarchical clustering requires the user to select the method to evaluate similarity (distance methods) between cluster groups, as well as the distance metric that the method uses in calculating distances. The different distance methods are mentioned in Section C.3.1, and each were applied and compared to assess which method works best for this dataset. For this analysis, the euclidean, squared-euclidean, cityblock, and correlation distance metrics were applied with each distance method, resulting in 28 different combinations of hierarchical clustering solutions.

ANALYSIS OF HIERARCHICAL DISTANCE METHODS

To assess the effectiveness of each method/metric combination, the formed groups were evaluated by intra-cluster differences in runup statistics, including R2%, setup, and swash values. A broad range of cutoff values were input for each metric to also determine which cutoff values should be investigated further with the chosen distance method. The results of the relative R2% difference, relative setup difference, relative infragravity (IG) swash and high frequency (HF) swash difference are shown in Figure C.16 through Figure C.20. The methods were evaluated calculating mean or maximum intra-cluster values for these metrics. Within each newly formed cluster, the maximum and minimum value was used to calculate the relative difference for that cluster group. Once the relative difference for each cluster group was calculated, these values were averaged to obtain one average relative difference.

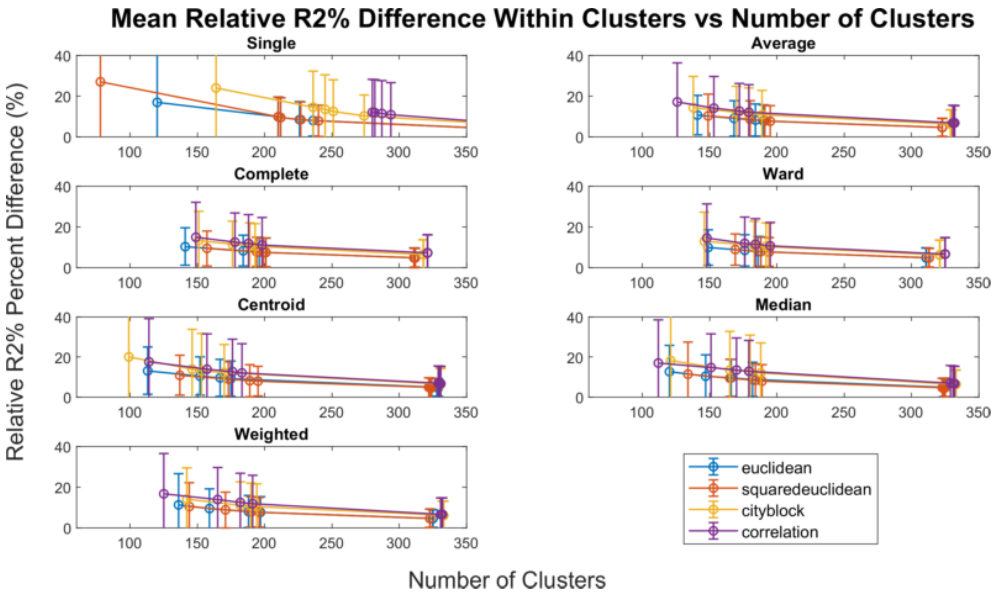


Figure C.16: Mean relative R2% difference within cluster groups for each distance method and distance metric. Standard deviation bars are attached to each data point.

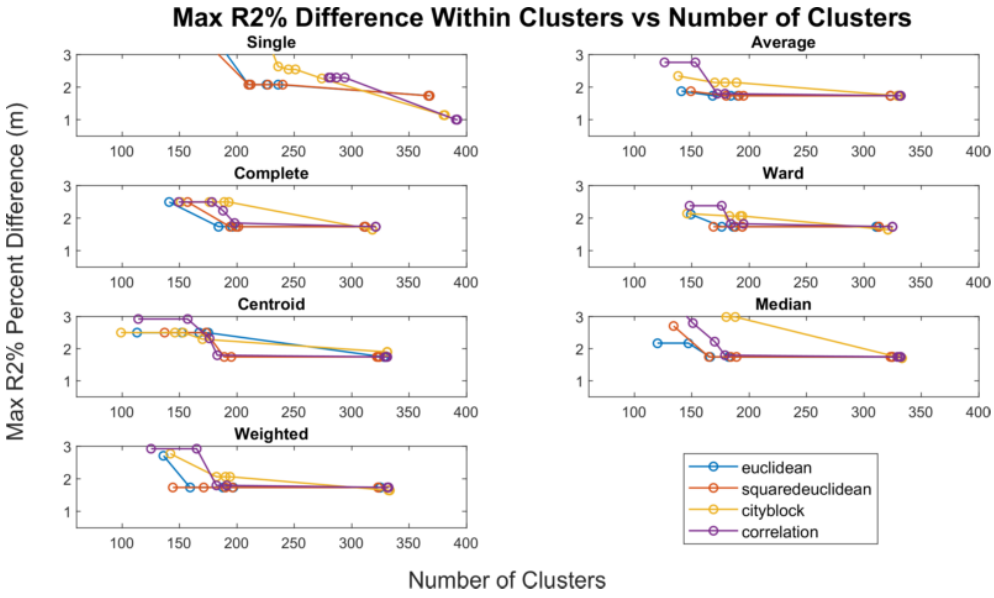


Figure C.17: Maximum relative R2% difference within cluster groups for each distance method and distance metric.

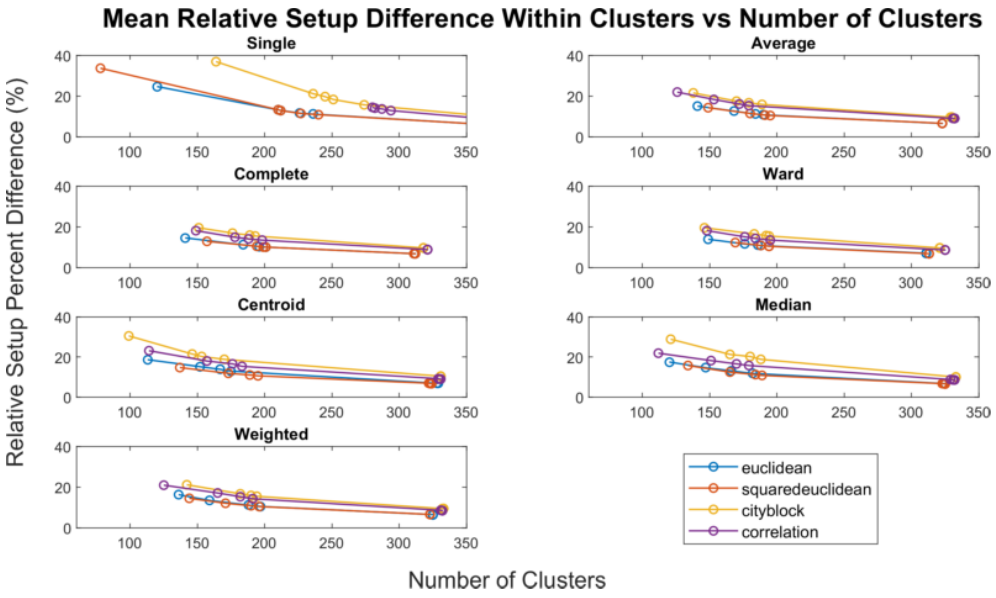


Figure C.18: Mean relative setup difference within cluster groups for each distance method and distance metric.

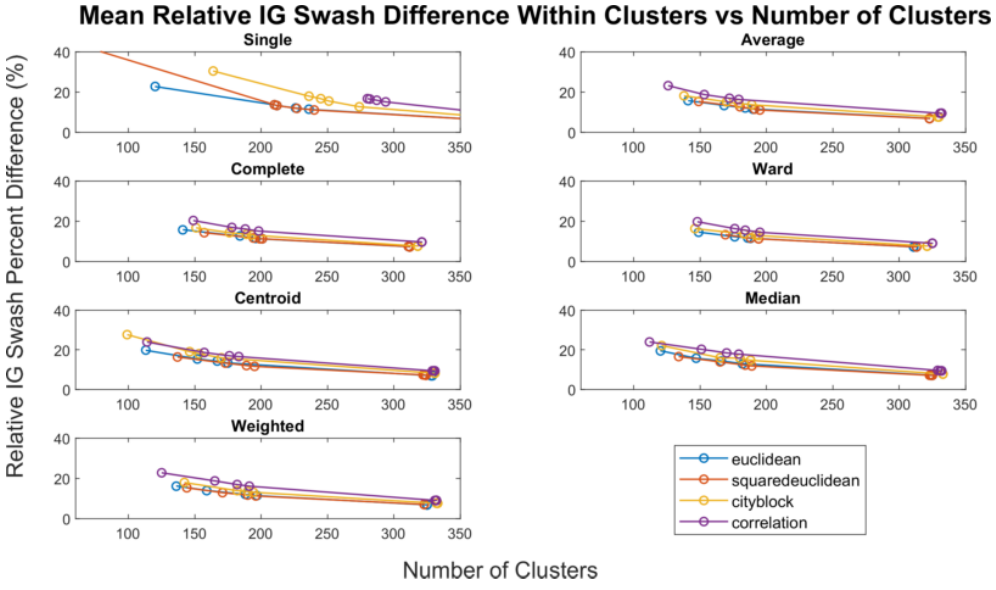


Figure C.19: Mean relative IG swash difference within cluster groups for each distance method and distance metric.

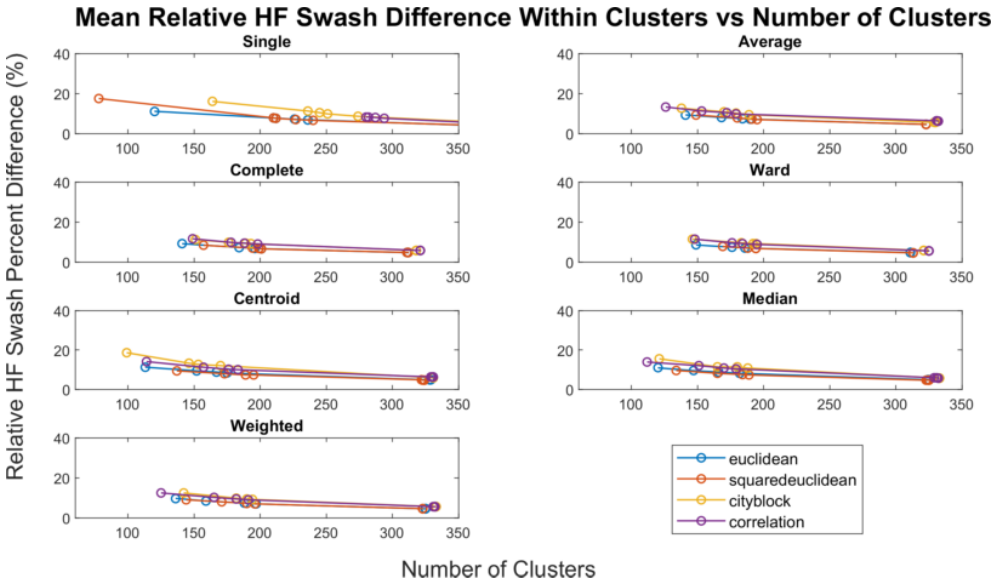


Figure C.20: Mean relative HF swash difference within cluster groups for each distance method and distance metric.

The main finding from this analysis is that the single distance method provides the worst results, whereas the other distance methods all provide similar results. Also, the euclidean and squared-euclidean distance metric seems to work best for each distance method. However, the centroid, median, and ward methods are all designed to be used with the euclidean distance only, and since euclidean distance is the more common and appropriate measure compared to squared euclidean for hierarchical clustering, it was used to compare all methods. The result of the relative R2% difference (the most important metric for this study) output from the different distance methods using the euclidean distance metric is shown in Figure C.21.

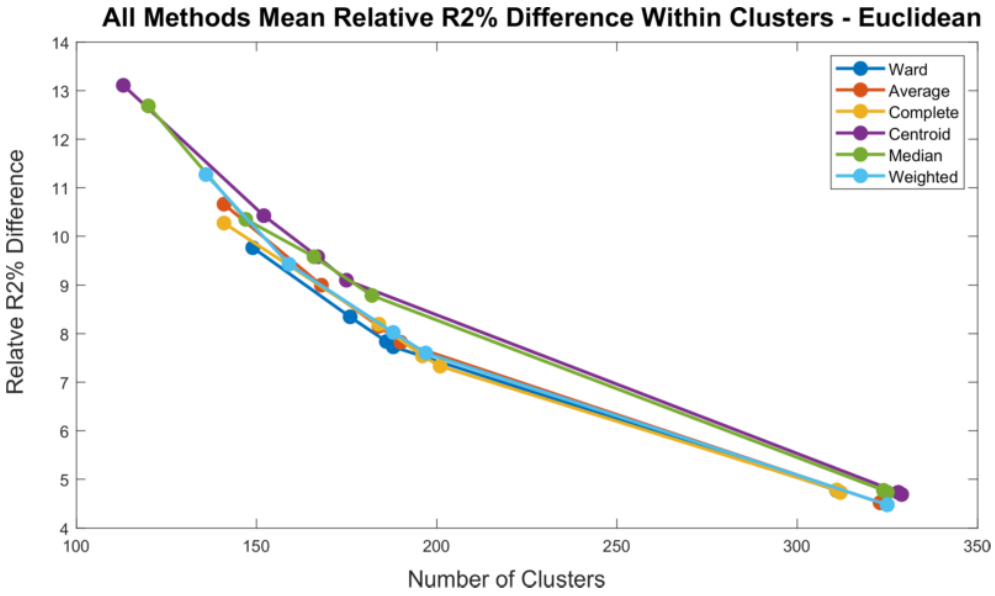


Figure C.21: Comparison of the mean relative R2% difference within cluster groups for each distance method with euclidean distance.

It can be seen that the Ward method provides slightly lower R2% difference within the cluster groups compared to the other distance methods for similar numbers of clusters. Through this analysis, it was decided that Ward provides the best results and therefore the Ward distance method with euclidean distance was selected to be used to form the final cluster groups.

C.4. TESTING THE APPLICATION

C.4.1. PROBABILISTIC MATCH

SOFTMAX FUNCTION DESCRIPTION

A method used in neural networks to predict the probabilities associated with a multinoulli distribution is the softmax function (Goodfellow, Bengio, & Courville, 2016). The function maps a vector of inputs to a posterior probability distribution (Goodfellow et al., 2016). For the application with the method proposed in this study, once the distances have been calculated between a test profile and all cluster profiles, if the distances are input into the softmax function, each distance is attributed a probability ranging between 0 and 1. This process is schematized in Figure C.22.

The equation is defined as:

$$S(x)_i = \frac{exp(-B * x_i)}{\sum_{j=1}^{n} exp(-B * x_j)}$$

where $S(x)$ is the probability of matching to cluster profile i, x_i is the distance between the test profile and cluster profile i, x_j is the distance between the test profile and cluster profile j, B is the stiffness parameter, and n is the number of cluster profiles. Due to the exponents, $S(x)$ is always positive, and since the numerator appears in the denominator summed with other values above 0, $S(x)$ will be less than 1. Therefore, the output is always between 0 and 1.

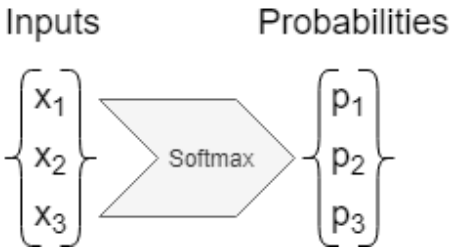


Figure C.22: Illustration of how the softmax function operates, transforming inputs into posterior probabilities.

PROBABILISTIC MATCH SENSITIVITY TO BETA

The B value essentially acts as an inverse variance, such that larger values of B will cause the distribution to be narrower so that probabilities of points far away from the centroid will become small (Bauckhage, 2015). Figure C.23 shows an example of ten distances increasing from one to ten and the associated probabilities with two different B values. This simple example shows how the B value changes the distribution of the probabilities.

Figure C.24 shows histograms of the maximum match probability between the 1000 test profiles and the 149 Cluster Round 2 profiles for different B values. When the B value is higher, there are more profiles with a high maximum probability value, meaning the profiles are matched with greater weighting to the closest profiles. When the B value is 1, the maximum probabilities are very low, showing that the test profiles are matched more evenly among many cluster profiles.

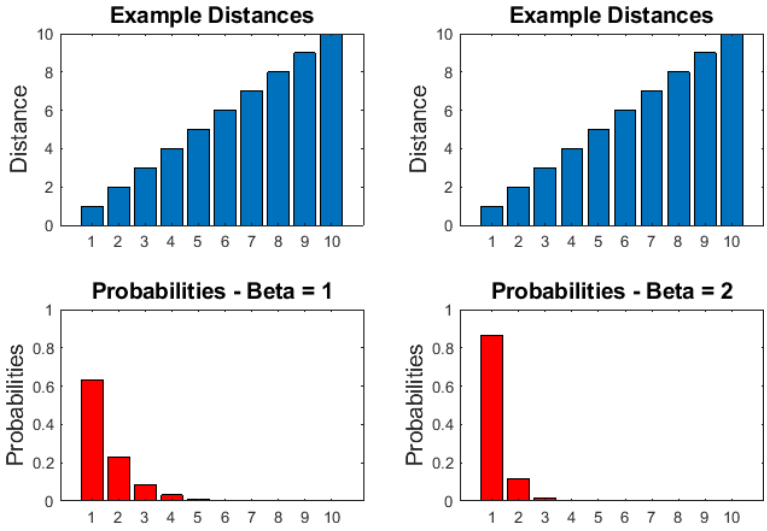


Figure C.23: An example of the output probabilities of a dataset from 1 to 10, with different B values. As the Beta increases, the closest distance gains a greater share of the probability.

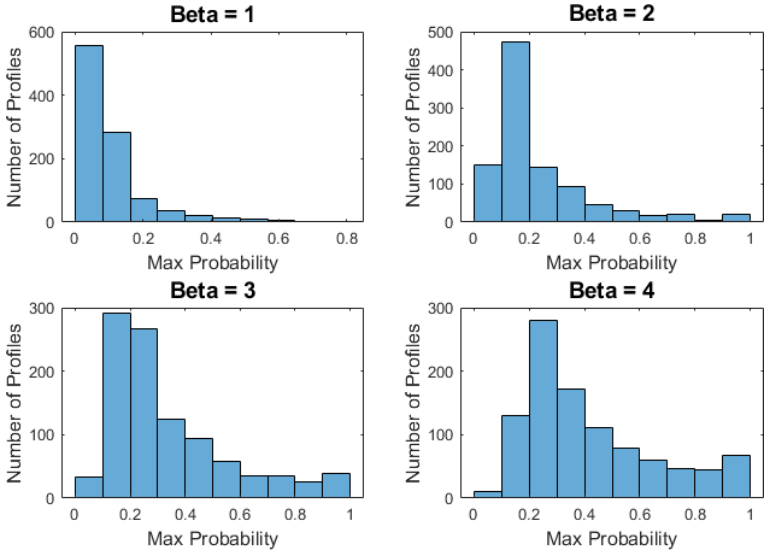


Figure C.24: The maximum probabilities for the 1000 test profiles for different values of Beta.

D

XBeach Model Pre and Post Processing

D.1. XBEACH INPUTS

The XBeach simulations were completed using a very similar method as Pearson et al. (2016). A file named the *batman* file contains all of the main information for the model runs, including the wave loading parameters, model duration times and grid settings. An example of the *batman* file is shown in Figure D.1.

```
%% BATCH MANAGER

% Run directory names
V3_BC,          % bcdir - folder with boundary condition runs
V3_Runs,        % runsdir - folder with full set of runs
1,              % mkbct - flag for making boundary conditions (1=yes; 0=no)
1,              % mkinp - flag for setting up full input parameter files (1=yes; 0=no)

% Initialize parameter ranges
1,          % wl - water level above reef flat (when reefheight = 0) [m]
3,7,        % hs - significant wave height [m]
0.05,0.01,  % steep - H0/L0 steepness [-] % tp - peak period [t]
3.3,        % specShape - spectral shape (JONSWAP peak enhancement factor) [-]
10,         % slope2=[1V/(slope2)H] - slope 2 (beach) [-]
30,         % beachCrest - max elevation of beach crest [m]
0.05,       % cf - coefficient of friction [-]

% Model duration and output intervals
7200,       % runDur - total run duration without spinup [s] (1 hr = 3600s)
3600,       % model spinup time [s]
1800.0,     % tintm - mean output interval [s]

% Grid resolution settings
0.25,   % dxmin - minimum dx for main model runs - BC gen dmin is larger by default
4,      % dxmax - maximum dx for main model runs - BC gen dmax is larger by default
64,     % np - number of gridpoints per wavelength
270,    % mainang - primary wave direction
10,     % s - directional spreading - default is 10
0.005,  % dfj - step size frequency used to create JONSWAP spectrum

```

Figure D.1: Sample XBeach batman file, which contains the main parameter information to input to the XBeach simulations.

The *batman* file is accessed and used to create all of the *params* files for each reef profile. An example of the general *params* file before it is filled in with information specific to one profile and wave condition is shown in Figure D.2. XBeach non-hydrostatic was used, and to maintain simplicity many default parameters for this model were utilized. The parameters that were altered between runs are highlighted with the word KEY.

Each XBeach simulation with different profiles required unique depth, x grid, and friction files. The observation points were also set at a fraction of the profile length and therefore they were also unique to each simulation. The model run times were also set up as a variable to be prepared to alter them if need be, but were held constant for each simulation. The spin up time was set to one hour, and the total duration of the simulation was three hours, resulting in two hours of simulation time with output for analysis for each profile.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% XBeach parameter settings input file                                       %%%
%%%                                                                            %%%
%%% date:     16-April-2019 12:53:50                                           %%%
%%% function: xb_write_params                                                  %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

bedfriction = cf
bedfricfile = fric.txt
sedtrans    = 0
morphology  = 0
taper       = 600
nonhq3d     = 1
wavemodel   = nonh
swave       = 0
front       = nonh_1d
back        = abs_1d

%%% Grid parameters %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

nx       = NXKEY
ny       =  0
vardx    =  1
depfile  = profile.dep
xfile    = x.grd
posdwn   = -1

%%% Model time %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

tstop        = TSTOPKEY

%%% Tide boundary conditions %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

tideloc      = 0
zs0          = 1
%%% Wave boundary condition parameters %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

wbctype      = WBCKEY
bcfile       = BCFILEKEY

%%% Output variables %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

outputformat = netcdf
tintg        = 900
tintm        = TINTMKEY
tstart       = TSTARTKEY
tintp        = 0.5

npoints = 9
OBSPT1KEY 1. 7/8th
OBSPT2KEY 1. 6/8th
OBSPT3KEY 1. 5/8th
OBSPT4KEY 1. 4/8th
OBSPT5KEY 1. 3/8th
OBSPT6KEY 1. 2/8th
OBSPT7KEY 1. 1/8th
OBSPT8KEY 1. 5m off beach
OBSPT9KEY 1. 1m off beach

npointvar = 2
zs
u

nrugauge     =1
RUGAUGEKEY 1
rugdepth     = 0.1

nglobalvar   = 2
zs
zb

nmeanvar = 1
zs
```

Figure D.2: XBeach params file example to demonstrate the XBeach inputs for the model simulations, highlighting the variables that are changed for each simulation.

BOUNDARY CONDITIONS

Each profile simulated in XBeach was tested with four wave loading conditions. These conditions were reused to obtain the most accurate comparison. An example of the *jonstable* that was used in the XBeach simulation to create one of the wave boundary conditions, with a significant wave height of 3 m and wave steepness of 0.01, is shown in Table D.1.

The six rows of the table cause six sets of the boundary condition to be input throughout the simulation, one after another, providing some variability. The total time of the conditions adds to three seconds past three hours (three hours is the total XBeach simulation time). This is done to ensure that the boundary conditions do not end before the simulation.

Table D.1: Example of the 'jonstable' used as input to generate the boundary conditions

Hm0 (m)	Tp (s)	MainAng (degrees)	Gammajsp	s	Duration (s)	dtbc (s)
3	13.8617	270	3.3	10	1800	1
3	13.8617	270	3.3	10	1800	1
3	13.8617	270	3.3	10	1800	1
3	13.8617	270	3.3	10	1800	1
3	13.8617	270	3.3	10	1800	1
3	13.8617	270	3.3	10	1803	1

D.2. COMPUTER SPECIFICATIONS

The XBeach simulations and clustering algorithms required high computing power in order to work with such a large dataset. The specifications of the computers used during this study are presented in Table D.2.

Table D.2: Specifications of the computers available for this study.

Type	Number of Computers	Memory (GB)	Number of Cores	Speed (GHz)	Capacity (GB)
WCF	1	16	4	2.6	200
WCP	5	16	8	2.6	100

Apart from the high memory and speed, the WCP computers with eight cores allowed eight XBeach simulations to be performed simultaneously, drastically reducing the required time to complete them compared to operating on an average laptop equipped with 2 cores.

D.3. XBEACH SIMULATION RUN TIMES

CLUSTER PROFILES

The actual XBeach simulation times for the 500 cluster round 1 profiles are displayed in boxplots, separated by the wave loading conditions in the top subplot of Figure D.3. The drastic difference between loading condition 1 and all others is because the WCF computer was used while simulating the first loading condition and the WCP computers were used for all others. The power of the WCP, detailed in Table D.2, is fully displayed here. The mean of all run times is displayed with the black dotted line and is approximately 15 minutes. Without considering the first loading condition, the mean run time is approximately 9 minutes.

MOST DISSIMILAR PROFILES

While assessing the intra-cluster variability, the most dissimilar profiles for each cluster group were also simulated in XBeach to compare the hydrodynamic results with the cluster profile. In total, 4810 XBeach simulations were done on these profiles, using only two wave boundary conditions. Only the WCP computers were used for these runs. The results of the run times for these XBeach simulations are shown in the bottom subplot of Figure D.3. The mean run time for these simulations was approximately 10 minutes. Therefore, while using the highest powered computers available for this study, the XBeach simulations took approximately 10 minutes per run.

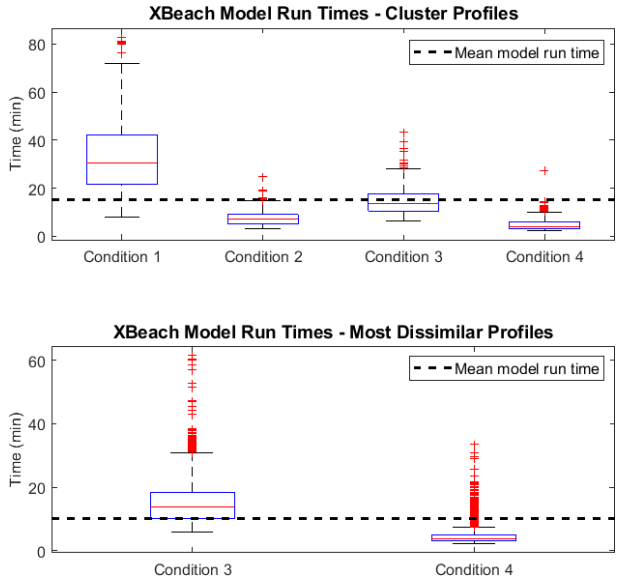


Figure D.3: XBeach simulation run times for the 500 cluster profiles (top), and for the most dissimilar profiles (bottom). Wave condition 1 for the cluster profiles was computed with a combination of the WCF and WCP computers, whereas all other simulations were computed with the WCP computers.

E

Profile Features and Wave Runup

E.1. FEATURES LEADING TO DISSIMILAR WAVE RUNUP

Assessing the differences between grouped profiles (based on full profile morphology) with dissimilar wave runup resulted in interesting findings. As shown in Section 4.2.1 profiles with a peak above MSL causes a setup between the peak and the shoreline that results in differences in runup behavior. Nine of the eleven observed profiles had this peak. The second observation was that the nearshore similarities of the profiles is much more important than further offshore in deeper water. To solidify the second point, all profiles with a peak above MSL were removed and the profiles causing the largest differences in wave runup were analyzed again. The result is shown in Figure E.1.

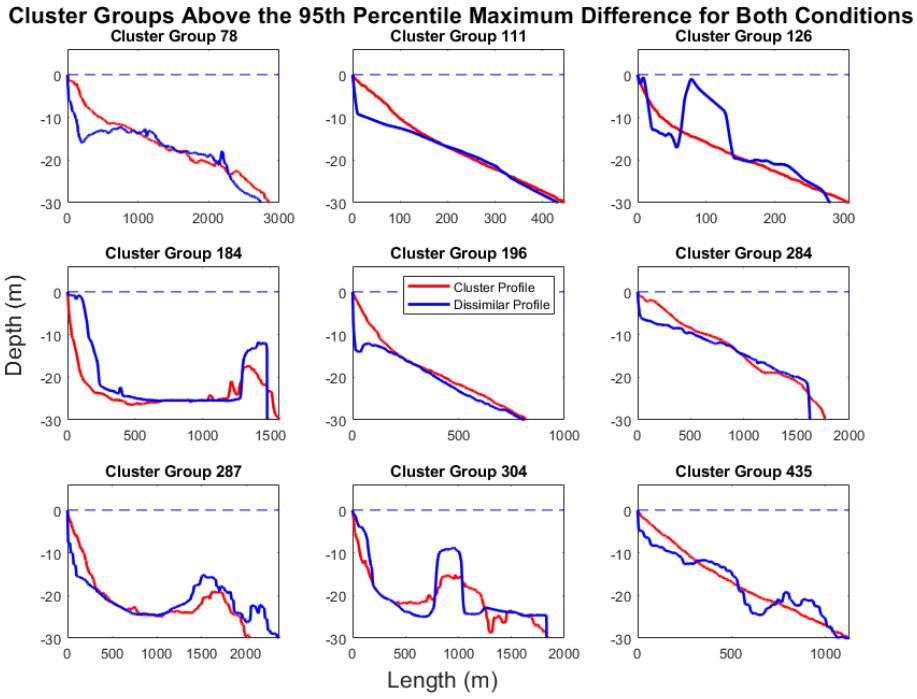


Figure E.1: The cluster groups that are in the top 5% for maximum difference to centroid in R2% for the two tested wave loading conditions. All profiles with peaks above MSL are removed. The cluster profile is shown in red, and the dissimilar profile within the group with large differences in R2% are in blue.

After removing all profiles with a peak above MSL, it is clear that the other profile feature that causes large wave runup discrepancies is deviations in the nearshore region. None of the cases in Figure E.1 have a tight connection in the nearshore. The profiles seem to connect again between depths of -10 and -20 m. The second obvious feature in Figure E.1 is some sort of large deviation, as shown in Cluster Group 126 and Cluster Group 384. Otherwise, the profiles are well matched and there is no striking differences that would lead to differences in wave runup.

E.2. FEATURES LEADING TO SIMILAR WAVE RUNUP

The features that cause similar wave runup were also investigated. To do this, the 500 Cluster Round 1 profiles were matched to another cluster profile based on similar wave runup. An example of ten cluster profiles and their match are shown in Figure E.2.

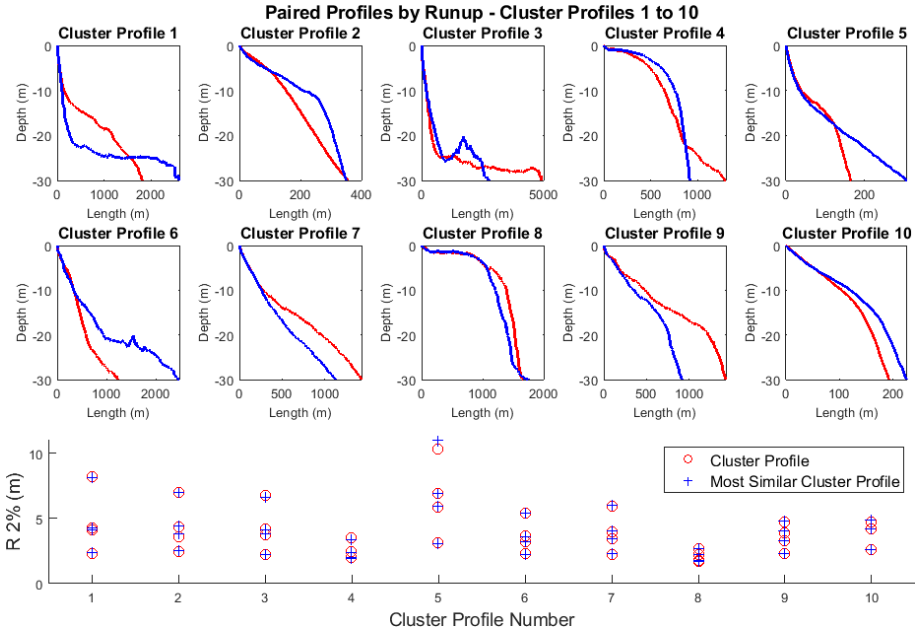


Figure E.2: An example of ten Cluster Round 1 profiles matched to another cluster profile based on the most similar wave runup, averaged over the four wave loading conditions. The wave runup for the four loading conditions is plotted along the bottom subplot.

The most obvious similarity between matched profiles is in the nearshore region. From depths of 0 to approximately -15 m the profiles are very similar, and then at greater depths some of the profiles diverge, for example cluster profile 3 which goes on to have a length difference of two kilometers. The nearshore profile was therefore noted to have great importance in grouping profiles with similar wave runup and was incorporated into the methods to match a profile to the cluster profiles. It is also included as a recommendation to be focused on more heavily in the cluster analysis for future work.

F

VALIDATION

F.1. MATCHING METHOD COMPARISON

This appendix provides additional analysis of the main hydrodynamics for the two Roi Namur profiles that were used for validation. The validation is separated by the two different methods to match the Roi Namur profiles to the cluster profiles. The validation was done with the case when there are 149 Cluster Round 2 profiles.

F.1.1. NS3 MATCH

The cluster profiles selected to represent the Roi Namur profiles when using the NS3 matching method are shown in Figure F.1. The Roi Namur profile, matched Cluster Round 1 profile, and Cluster Round 2 profile are shown. The bottom subplot, displaying Roi Namur profile 2 can be seen to not show the Cluster Round 1 profile. This is because the Cluster Round 1 profile is also the Cluster Round 2 profile. The Cluster Round 2 profiles were selected as the Cluster Round 1 profile within the group that has the median wave runup, explained in Section 3.4.2. This results in a slightly better match for Roi Namur profile 2 compared to Roi Namur profile 1.

Figure F.1: Roi Namur profiles 1 and 2 (in blue) matched to the Cluster Round 1 and Cluster Round 2 profiles. Roi Namur profile 2 is matched to a Cluster Round 1 profile that is also selected as a Cluster Round 2 profile.

Figure F.2 and F.3 present the comparison of $R_{2\%}$, setup, and swash between the Cluster Round 2 profile and the Roi Namur profile. separated by loading condition. For Roi Namur profile 1, the setup and high frequency swash vary the most, whereas for Roi Namur profile 2, the results are very similar, generally varying most heavily for loading condition 4.

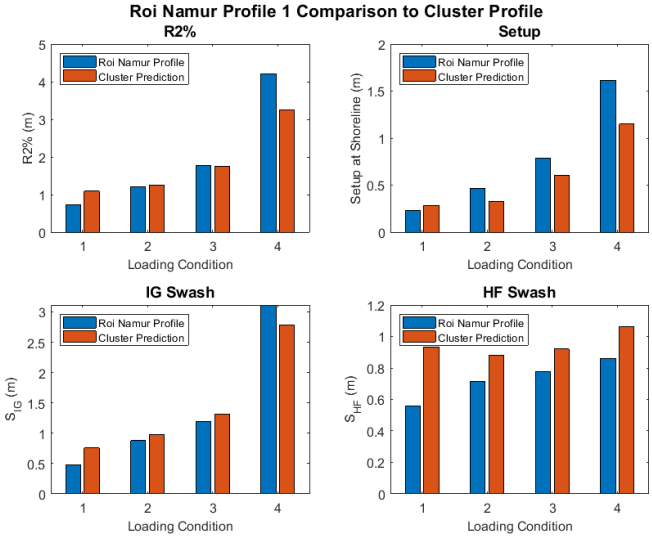


Figure F.2: Comparison of the $R_{2\%}$, setup, and swash for the four different wave loading conditions between Roi Namur profile 1 and the matched cluster profile using the NS3 matching method.

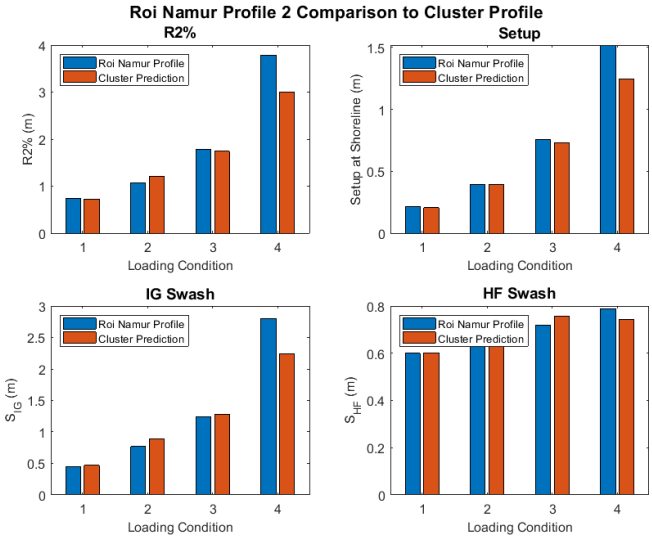


Figure F.3: Comparison of the $R_{2\%}$, setup, and swash for the four different wave loading conditions between Roi Namur profile 2 and the matched cluster profile using the NS3 matching method.

F.1.2. PROBABILISTIC MATCH

The probabilistic match results for the Roi Namur profiles are shown in Figure F.4. Subplots (a) and (b) show the probability of the Roi Namur profile belonging to the 149 cluster profiles. Subplots (c) and (d) plot each of the cluster profiles, colored to represent their match probability, and the Roi Namur profile in black. Roi Namur profile 1 has a greater spread of association with the cluster profiles since it is not very similar to one cluster profile. Conversely, Roi Namur profile 2 has a high probability (roughly 60%) of belonging to one cluster profile since it does match very well, as seen from the NS3 results shown in Figure F.1.

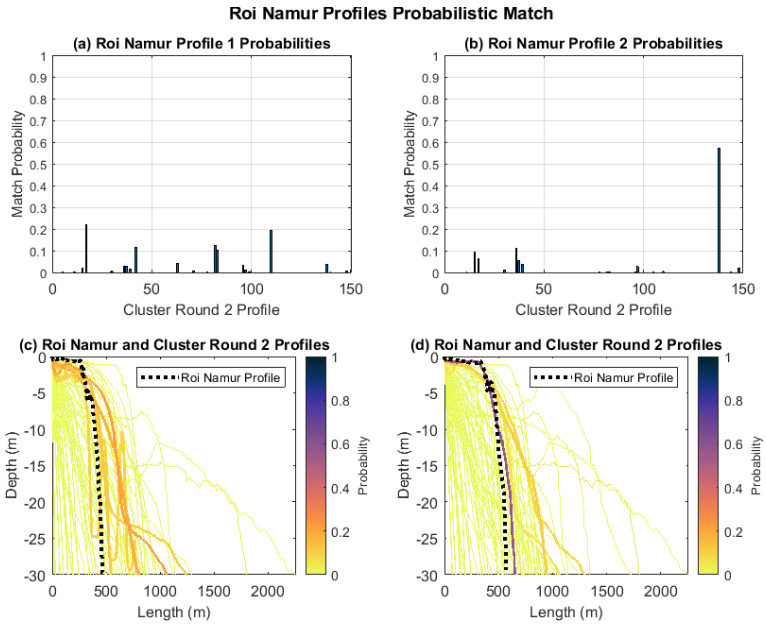


Figure F.4: The two Roi Namur profiles probabilistic match results to the Cluster Round 2 profiles.

The runup comparison results from the probabilistic match are very similar to the NS3 match results, but with slightly greater difference between the Roi Namur profile and the cluster prediction. These two profiles demonstrated the rare case where the probabilistic method does not improve the prediction. This is most likely due to the fact that the cluster profiles with significant probabilities are not surrounding the Roi Namur profile, but rather all slightly longer, thereby forcing the prediction to be less accurate compared to the NS3 method.

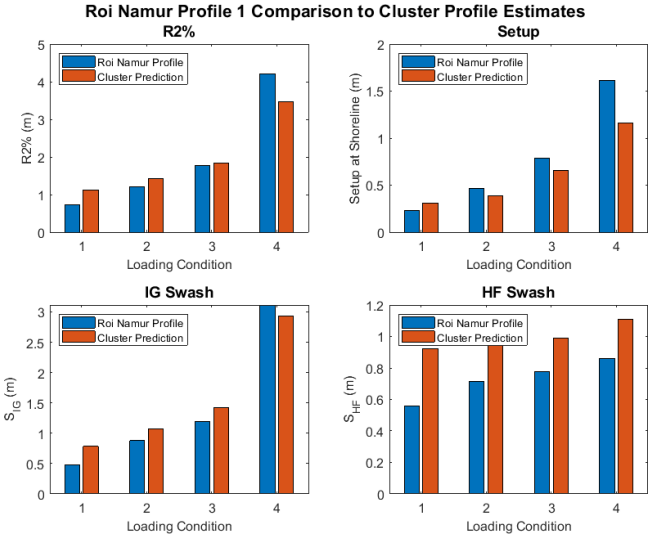


Figure F.5: Comparison of the $R_{2\%}$, setup, and swash for the four different wave loading conditions between Roi Namur profile 1 and the matched cluster profile using the probabilistic matching method.

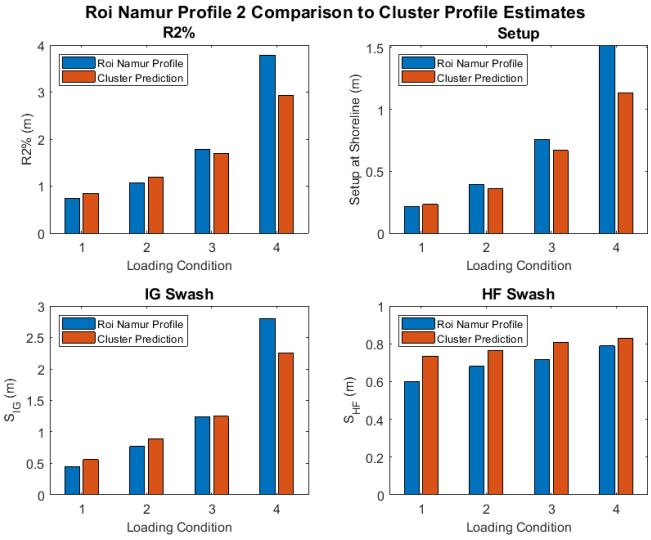


Figure F.6: Comparison of the $R_{2\%}$, setup, and swash for the four different wave loading conditions between Roi Namur profile 2 and the matched cluster profile using the probabilistic matching method.

G

Geographic Analysis

G.1. LOCATIONS OF GROUPED PROFILES

The locations of the profiles that are grouped together provides information on which coastlines around the world are similar to each other. This information can be used to draw conclusions between wave forcing, geography and coral reef morphology. Figure G.1 shows the locations of the grouped profiles in the Cluster Round 2 groups. Four different cases of cluster groups are analyzed, including the 45, 101, 149 and 311 cluster groups. The figure shows that for each case, there is one profile that represents a great majority (almost one quarter) of the profiles included in the analysis. This cluster group is mainly made up of profiles from Hawaii, but also includes profiles from all locations other than Florida.

In general, the cluster profiles are well mixed, meaning that coral reef morphology is shared among the different locations.

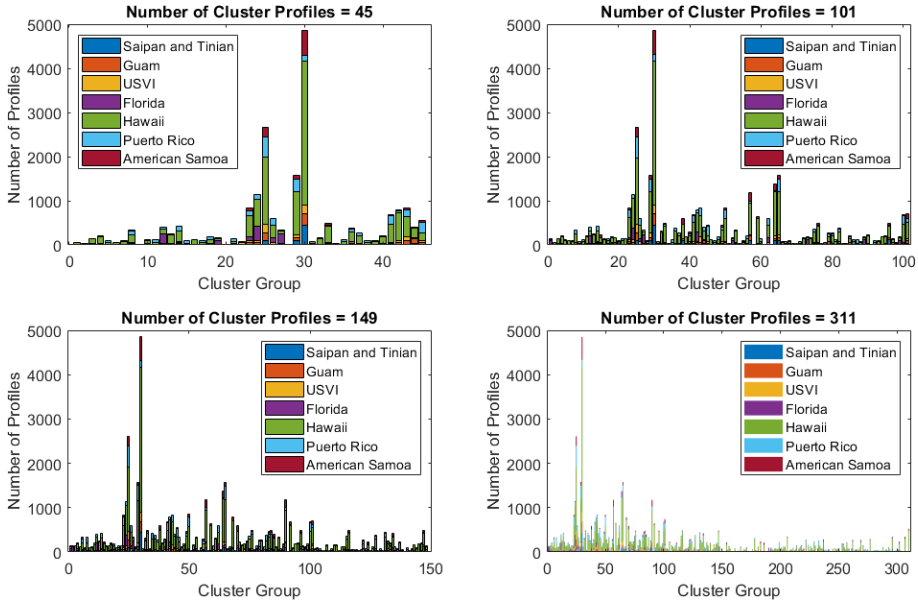


Figure G.1: The number and distribution of profiles that make each Cluster Round 2 group, analyzing the cases of 45, 101, 149 and 311 cluster groups. There is always one group that represents the vast majority of profiles included in the analysis.

G.2. LOCATION OF CLUSTER PROFILES

One possible application of the cluster profiles is to research the external factors causing certain profile shapes. If more is known about why the profiles are shaped as they are, then this can be used to aid the prediction of what a profile will most likely look like, and which cluster profiles will most likely be similar. For example, if one wants to predict the wave runup on the North shore of Guam, if there is a clear connection between that geography, as well as the wave conditions and profile shape, this may help in choosing the appropriate cluster profile to model that area.

To provide an idea of how this could work, the closest match between a profile in the dataset to the 500 Cluster Round 1 profiles was found. This is done since the cluster profiles do not actually exist (they are the median of the cluster group), meaning that the cluster profile location must be that of the closest real profile. These were then plotted on a map shown in Figure G.3, colored to show the rank in wave runup in comparison to all of the cluster profiles. The colored circles are the matches to the 500 Cluster Round 1 profiles, and the blue '+' signs show the locations of all the profiles that were included in the analysis.

This map provides an idea of where the cluster profiles are around the world, as well as how susceptible these coastlines are to wave runup. A link can start to be made between the location of the profile and its runup rank. For example, Puerto Rico, which is much more susceptible to large swells from the North, clearly has profile shapes along its North shore that are more susceptible to wave runup (darker colors).

A plot of the 500 cluster profiles, colored by wave runup rank is shown in Figure G.2. The most dominant relationships displayed in this figure are of the profile steepness and total length. The top subplot, with the 125 profiles with the lowest runup rank are all relatively long, wide and shallow until they drop off to the deeper depths. The bottom subplot, displaying the 125 profiles with the highest wave runup are much more steep and narrow, with limited lengths.

Connections can begin to be made between geographic location, wave forcing, and profile shape. Further analysis into these relationships could be very beneficial.

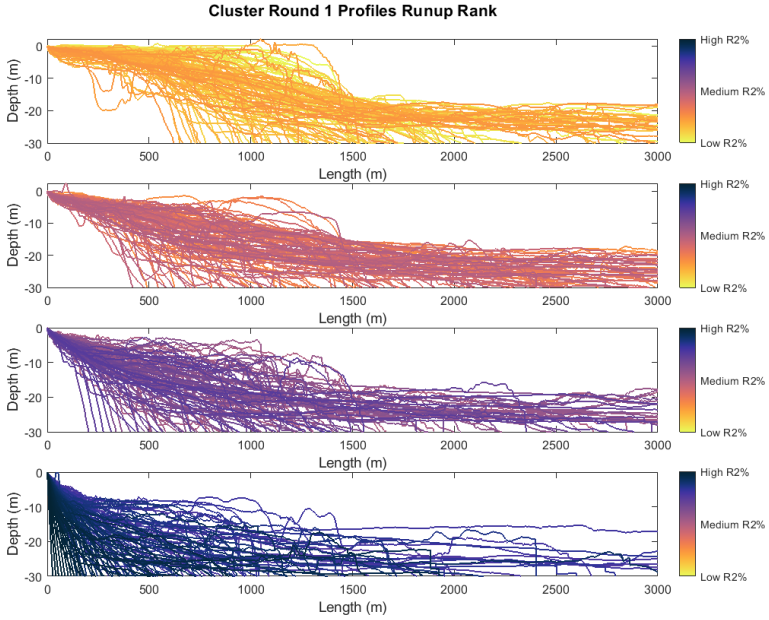


Figure G.2: The 500 Cluster Round 1 profiles, sorted and colored by their runup rank. The top subplot shows the profiles with on average the lowest wave runup, which are generally long and shallow. The bottom subplot shows the profiles with the highest runup, which are generally steep and narrow.

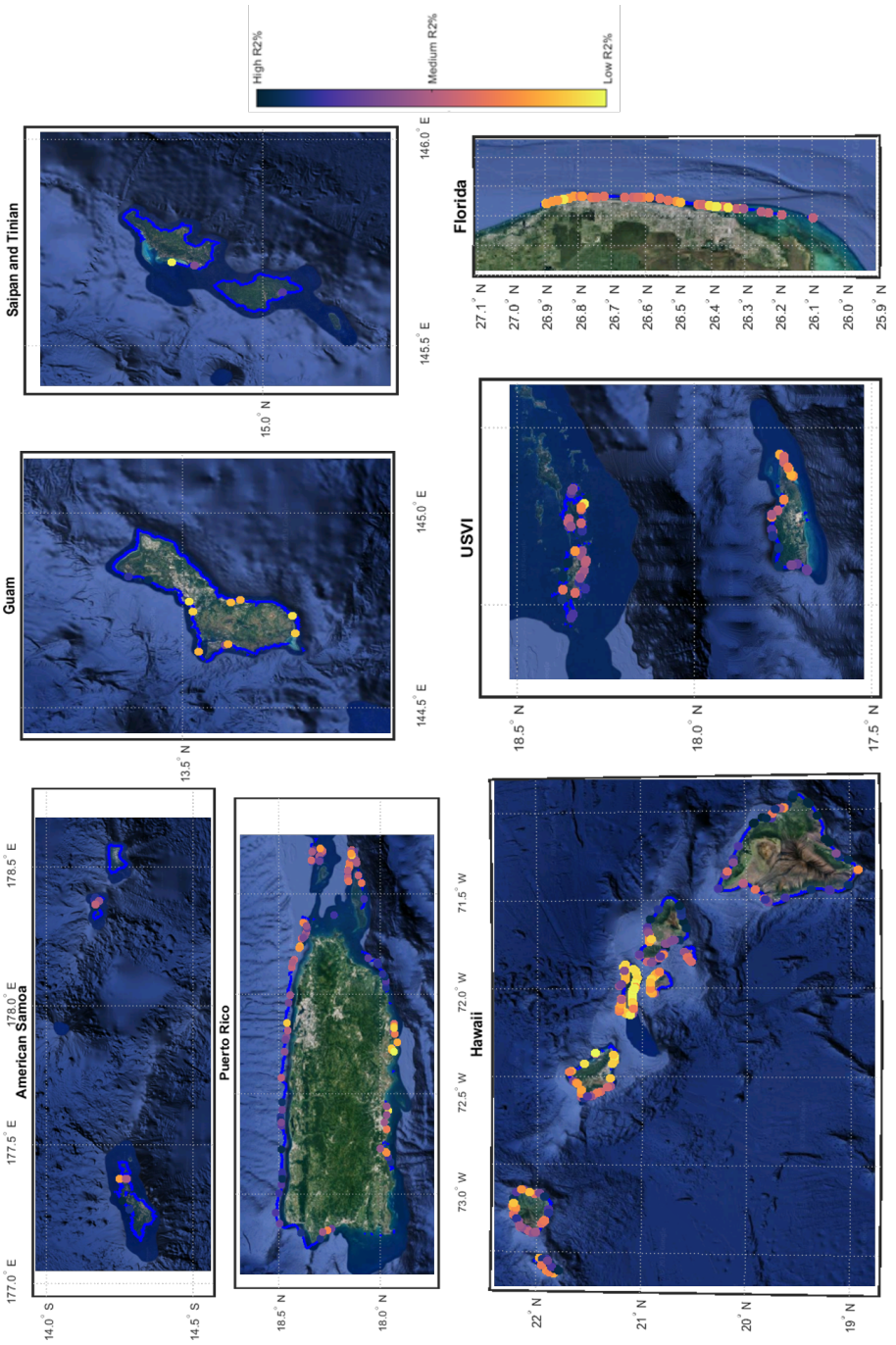


Figure G.3: The locations of the closest profiles to the 500 Cluster Round 1 profiles, plotted on a satellite image. Colored dots show the closest match to the cluster profiles, with the respective runup rank, and blue '+' signs show the locations of profiles included in the study. Source: (NASA, 2019)

G