Emulators in the investigation of sensitivities and uncertainties in tsunami models

Andria Sarri

Institute for Risk and Disaster Reduction

University College London

A thesis submitted for the degree of

Doctor of Philosophy

November 2014

Statement of originality

I, Andria Sarri, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Chapter 2 is based on a published work in the Natural Hazards and Earth System Sciences [Sarri et al., 2012] and it is joint work with Dr Serge Guillas (University College London) and Prof. Frederic Dias (University College Dublin). It was the first application of statistical emulation on tsunami modelling.

Chapter 3 is joint work with Dr Simon Day (Institute for Risk and Disaster Reduction). His extensive knowledge about seabed deformation in case of an earthquake is provided to me in order to investigate how the coseismic displacement representation affects the resulting tsunami waves.

Chapter 4 is collaboration with Dr Serge Guillas and Dr Simon Day. Part of the results for this Chapter have been obtained during a project funded by UCL-Business and NERC Follow on Fund grant, which are the initial steps for the development of the UCL Cascadia tsunami catastrophe risk model. The bathymetry data for the analysis in this Chapter have been provided by Xiaoyu Liu, that converted the public available NOAA/NGDC data into a mesh that can be used for evaluating the tsunami model.

Abstract

This thesis consists of three parts with the unifying theme of applying emulators and other statistical methods to the investigation of sensitivities and uncertainties in tsunami models. The first part contains the statistical emulation of the wave time series at fixed locations that are obtained using an analytical landslide-generated tsunami computer model. Leave-one-out diagnostics are used to validate the emulator, showing excellent agreement in predictions and model evaluations. The emulator is used for sensitivity and uncertainty analyses, where many evaluations are necessary. The computational time required for both analyses is at least 20,000 times lower using the emulator instead of the computationally expensive model.

The second part investigates the significance of the seabed deformation representation on tsunami wave predictions of the advanced VOLNA tsunami model. A novel realistic representation of the tsunami source deformation using quadratic curves is compared with simpler representations used in existing tsunami models: single block, four blocks, sixteen blocks and "piano keys". Factor-of-two differences in wave elevations obtained for different representations of seabed deformation.

The last part presents an investigation of tsunamis produced by different fault rupture scenarios on the Cascadia Subduction Zone. Polygons are used to model the source geometry and quadratic curves the deformation in different source event scenarios that are chosen using experimental design. The VOLNA model is evaluated for these scenarios to explore how the source characteristics affect the wave amplitude and coastal inundation predictions. Statistical emulation is applied to a large number of locations in order to investigate the use of emulators in place of expensive tsunami simulators in practical applications. The leave-one-out diagnostics show that the application of Registration and Functional Principal Components techniques to the emulation process leads to improved predictions. Finally, a limiting event is investigated in order to obtain the worst-case scenario inundation along the whole margin in a single simulation.

Acknowledgements

First of all I would like to sincerely and gratefully thank my PhD supervisor Dr Serge Guillas. This research project would not be possible without his supervision. Throughout my PhD he gave me lot of opportunities and he was always very motivating with great positive energy. He was always there when I was needing his help, not only for guidance and ideas about my research, but also with general support and advices about life. Our discussions were always very helpful for me.

Also, I want to express my gratitude to my second supervisor, Dr Simon Day. I learned a lot from him about tsunamis and earthquakes. His huge experience and deep knowledge on his topic was very inspiring for me. I always knew that whenever I had any difficulties on aspects of geology, he would kindly explain to me in the best possible way.

Furthermore, I am particularly thankful for his support to the IRDR Director, Professor Peter Sammonds. He is a great academic, but the most important is that he is a great person. He was always friendly and approachable and he was giving me confidence and encouragement to keep going with my research.

I would like to thank Professor Frederic Dias for his great ideas and support, especially at the beginning of my research that I was finding very helpful.

Also, I am thankful to the Institute for Risk and Disaster Reduction, as well as the UCL Department of Statistical Science for supporting financially my study.

A huge "thank you" to my parents and sister for their love, support, trust and encouragement. Moreover, I cannot forget to thank Yiannis Christodoulou. I would like to express my deepest gratitude to him. He is my family in London and the person that I trust the most. He was always there for me whenever I was needing him.

Finally, I want to thank my colleagues and friends in the IRDR and the Statistical Department for creating a nice environment and for the lovely time we had outside university as well. Special thanks for being so amazing friends to Giulia, Katerina, George M., George P., Menelaos and Nayia.

List of Figures

List of Tables

Contents

---

[1] The term "Validation" in this chapter is used for the process of assessing the emulator's predictions and not for checking the mathematical model, as it was the case for Chapter 2

Chapter 1

Introduction

1.1 Tsunami events

Catastrophic tsunami events lead to the loss of human lives, damages on buildings and infrastructure, and generally they cause tremendous panic and pain in the affected area. The tsunami events before the 1990s were the motivation for the academic community to start investigating the behaviour of tsunamis, in order to develop early warnings that can help in the reduction of damages due to future disastrous tsunami events. These initial investigations improved the general knowledge about tsunamis and specifically about their propagation.

The Boxing Day 2004 tsunami event at the Indonesian island of Sumatra, as well as the March 2011 event at Japan, were more recent wake up calls and proved to the world how dangerous and catastrophic tsunamis can be. Additionally, these dreadful tsunami events highlighted the fact that the existing knowledge about some tsunami features is still limited. Even in Japan, which is currently the most tsunami-ready country in the world, the existing awareness and prediction tools available were insufficient to prevent the large disaster caused by the 2011 event.

Although the loss of life in these events was tragic, one positive side effect has been the improvement of the global tsunami safety for future events, since they increased the scientific awareness and extended the research related to tsunamis. In general, the scientific research focuses on the warning and forecasting, as well as the generation of efficient defences and inundation maps that can help in hazard mitigation for future events as well as for the risk assessment for specific locations.

A better understanding and modelling of tsunami events can result in more accurate forecasting as well as better preparedness and elimination or even complete prevention of similar disasters in the future. Nevertheless, one of the main problems in tsunami event investigations is the lack of data, which makes both the understanding and modelling of the events challenging. In order to be able to study past tsunami events, the use of computer models, also known as simulators, is necessary. Unfortunately, in most of the cases, the computer models used for tsunami predictions are computationally expensive to be evaluated. This makes very difficult further analyses that require a large number of model evaluations; for example the calculation of the uncertainties in predictions. Additionally, in cases that the computer models are used for real-time predictions, their computational demand leads to significant delays in predictions. A solution to this issue comes with the use of statistical surrogate models, known as statistical emulators, to approximate the computer models. The statistical emulators are much less computationally expensive compared to simulators and they can be used for time- and resource-demanding analyses of the computer models.

The primary objective of the analysis in this research is to solve the problem of high computational demand of the computer models that are used for tsunami predictions and investigations. The process to achieve this is to generate statistical emulators that accurately approximate the computer models. The significant benefits of using the statistical emulators in place of the simulators are demonstrated in this thesis.

1.2 Summary of Thesis Investigations and Results

The research focuses on the statistical analysis of tsunami models, where both landslide- and earthquake-generated tsunamis are investigated. Statistical emulation has been applied to approximate the expensive-to-run computer model with fast and accurate statistical surrogates. The substantial benefits resulting from this particular approximation are the fast predictions and the capability to perform computationally demanding analyses. Generally, the analysis in the thesis aims to address the issue of time and resources expensive computer models which make the investigation of tsunami events challenging. Due to the computational burden of a large number of model evaluations, the further analysis and improvement of the current tsunami models is difficult to be performed. Therefore, the thesis demonstrates how the statistical emulator can be used in place of the simulator for further analyses of the tsunami models, specifically for sensitivity analysis and uncertainty quantification.

Chapter 2 introduces an extensive introduction to various general concepts that have been investigated in this research as well as some general definitions. Following the Background chapter, Chapter 3 focuses on the application of statistical emulation on a landslide-generated tsunami model. Additionally, this chapter includes a demonstration of how the emulator can be used for sensitivity and uncertainty analyses. This chapter is based on the published work
"Sarri, A., Guillas, S. and Dias, F. (2012). Statistical emulation of a tsunami model for sensi-

tivity analysis and uncertainty quantification, *Natural Hazards and Earth System Sciences* **12**, 2003-2018 ”.

An extensive investigation on the sensitivity of an earthquake-generated tsunami model output, which is the wave elevation, on the earthquake source description is performed in Chapter 4. The approach followed for this analysis is to create an advanced realistic coseismic deformation representation and compare it with simple deformation cases. The realistic deformation presented is novel and improves on the way the earthquake source is described compared to existing literature since it is more flexible and avoids unnecessary assumptions. The question answered with this particular investigation is whether or not a clever, advanced and close-to-reality coseismic deformation representation is necessary in order to obtain accurate predictions of the resulting tsunami wave elevation. The ability of the computer models used in the analysis to realistically represent tsunami events is not investigated and the models were assumed to be accurate and reliable.

Finally, possible earthquake events at the Cascadia Subduction Zone are explored in Chapter 5. Different event scenarios of coseismic deformation are generated and the resulting tsunami wave elevations are investigated using statistical emulation. Also, the emulator is used for making feasible the demanding uncertainty and sensitivity analyses. Moreover, a “limiting” case event is investigated in order to obtain the worst-case-scenario of coastal inundation which helps to generate inundation maps for the area. This helps with the development of efficient defences and also raises people’s awareness to cope with future tsunami events.

Chapter 2

Background

2.1 Tsunami events

The thesis presents a statistical analysis focusing on tsunami events. The life of a tsunami can be separated in three phases, which are the generation, propagation and inundation. All the phases of the tsunami lifetime are going to be investigated in this research. The propagation phase is the one that is the most well understood, whereas the other two phases, generation and inundation, are highly uncertain. During the inundation phase, as a tsunami wave approaches the coast its amplitude increases and the propagation speed decreases. The water flow at this phase can be strong and can result in large damages over the coastal areas.

A brief definition for tsunamis is that they are series of powerful water waves that can be generated by either earthquakes, volcanic eruptions, underwater landslides or landslides along the coast. Generally speaking, a tsunami can be generated by the displacement of a large volume of water. Tsunami waves are long waves, with the order of magnitude of their wavelength and amplitude $10\mathrm{km}$ and $1\mathrm{m}$, respectively. They are characterised by high propagation speed and they are able to transfer energy to long distances. In many cases they have travelled through oceans, as for the case of the Japan tsunami event in 2011 [Stephanakis, 2013]. Tsunamis are rare events; the average return period of large earthquakes, which are the most common cause of tsunamis, is between 300 and 3000 years at specific locations.

2.1.1 Comparison of landslide- and earthquake-generated tsunamis

The most common source of tsunamis is earthquakes triggered by seafloor displacement. In this case, a subduction[1] of convergent adjoint tectonic plates occurs, displacing a large amount of water, resulting in tsunami waves generation. The tsunamis caused by this source are often catastrophic. The second most common source of tsunamis are landslides, which are a very dangerous tsunami generation mechanism due to the large vertical seabed deformation and the

---

[1] Subduction is when one lithosphere plate descends below the other

fact that their generation is close to the coast. These two aspects leads to large amplitude tsunami waves that can reach the coast very fast allowing short time for reaction. Several past tsunami events, such as the Papua New Guinea tsunami in 1998, are attributed to coseismic[2] landslides. Both earthquake- and landslide-generated tsunamis are going to be investigated in the analysis in this thesis.

A fault rupture slip is able to generate both of these kinds of tsunami events. In particular a slip along the fault can lead to either seabed deformation or seismic shaking. In the first case, an earthquake-generated tsunami might be obtained. The second case might cause a coseismic landslide, which moves down a seabed slope, disturbing the water surface and resulting in tsunami waves generation, see Wright and Rathje [2003]. The earthquake ground motion characteristics are important for predicting landslide-generated tsunamis [Bardet et al., 2003]. However, in the analysis in Chapter 3, where landslide-generated tsunamis are investigated, a simple generation mechanism is going to be considered and no attention to the source of the landslide, whether it comes from a seismic shaking or not, is going to be given. On the other hand, the coseismic seabed deformation case is going to be extensively investigated in Chapters 4 and 5.

There are significant differences between tsunamis generated by landslides and the ones generated by earthquakes. The first significant difference is that the landslide-generated tsunamis have shorter periods and wavelengths compared to earthquake-generated tsunamis. Additionally, the whole life cycle of the landslide-generated tsunami waves takes place near the source, since they do not travel as long distances as the earthquake-generated tsunamis do. Tsunami Warning Systems are not very efficient in the case of landslide-generated tsunamis, since they are not fast enough to alert the coast evacuation in the short available time [Bardet et al., 2003]. A solution to these challenges associated with landslide-generated tsunamis is the generation in advance of hazard maps for the coastal areas that might be affected from potential underwater landslide, in order to mitigate the hazard [Bardet et al., 2003].

Landslide-generated tsunamis can reach high amplitudes and can be extremely harmful [Synolakis et al., 2002; Bardet et al., 2003; Tinti et al., 2008]. Watts [2004] showed that approximately 30% of the Pacific Basin tsunamis have been induced by underwater landslides and the resulting run-ups were larger than the expected coastal run-up in an earthquake tsunami case. Furthermore, Bardet et al. [2003] pointed out that high and gathered tsunami waves are usually observed along the shoreline in the case of submarine mass failures, whereas more regular wave amplitudes are observed for earthquake-induced tsunamis.

---

[2] Coseismic is an event occurs at the time of the earthquake

The distinct characteristics of the tsunami waves occurred by landslides and earthquakes result from the differences in the seafloor deformation associated with each type of events. Specifically, for the case of landslides, the seafloor vertical deformation is much larger, and can reach hundreds of meters, compared to the earthquakes case, where the vertical deformation is usually only several meters, even for large earthquake cases [Bardet et al., 2003].

Since the most common source of tsunamis are the earthquakes, earthquake-generated tsunamis have been extensively investigated, whereas landslide-generated tsunamis have been underestimated for years and therefore they are researched less. Additionally, the generation of landslide-generated tsunamis depends strongly on how the shape of the seafloor changes with time, where there is limited knowledge and understanding [Liu et al., 2005] and this makes their investigation challenging.

A big challenge is to recognise whether a tsunami event resulted from an earthquake or from a landslide source. An example of this difficulty can be illustrated by two catastrophic tsunami events in 1990s; the Indonesia 1992 and Papua New Guinea (PNG) in 1998. The existing knowledge by the time of their occurrence was not enough to explain some of the special characteristics of these particular events [Bardet et al., 2003], which are the following: unexpectedly large wave amplitude, unpredictably long propagation time required for the first wave to reach the coast and also limited coastal inundation.

Looking specifically at the case of the PNG tsunami event, an earthquake of magnitude around 7 initially occurred and some minutes later, a tsunami wave approximately $10\mathrm{m}$ hit the shore and destroyed a relatively small coastal area. The delay observed between the earthquake occurrence and the time the tsunami reached the coast has been attributed to the delayed sediment failure. This phenomenon was firstly highlighted by Murty [1979]. Furthermore, Imamura and Hashi [2002] concluded that the tsunami that can be generated exclusively by the magnitude 7 earthquake are too small to be conspicuous. In conclusion, the three special characteristics of the two events give a strong indication that they are landslide-generated tsunamis and not earthquake-generated tsunamis [Bardet et al., 2003].

2.1.2 Landslide-generated tsunamis

Tsunami events generated by landslides are investigated in Chapter 3, with Chapter 4 and 5 to be focused on earthquake-generated tsunamis. At this point, it is worthy to mention that only some of the submarine landslides are able to generate catastrophic tsunamis. Investigations for the circumstances under which a tsunami can be induced by a landslide have been carried out in the literature and resulted in important conclusions. Wiegel [1955] performed the first experiments for landslide-generated tsunamis. At these experiments a sliding mass moves down an incline and the conclusion is that the wave amplitude is proportional to the slope of the incline and inversely proportional to the submergence of the sliding mass. More recently, Murty [2003] showed that the slide's volume is the most influential factor for landslide-generated tsunamis and that the dependency on other factors, such as the sloping beach angle, the water depth, the slide's density and the speed and duration of the slide's motion is much weaker. Ruff [2003] came to the conclusion that in the case that the ratio of the initial water depth above the landslide to the vertical distance of motion is less than one, large amplitude tsunamis occur. Additionally, Ruff [2003] declared that the wave amplitude strongly depends on the landslide's height and therefore submarine landslides are able to generate large tsunamis since their height ranges from few meters to several kilometres.

Liu et al. [2005] developed a three-dimensional numerical model to study the waves amplitude and run-up generated by a sliding mass. To validate this model, large-scale experiments in a wave tank were performed and the time series of the wave free surface elevation at many locations have been observed. The authors noticed that after the release of the slide, a wave propagation in the offshore direction was generated, while the water surface at the initial position of the slide was depressed, resulting to shoreline retreat. The water depression was followed by its "rebound" elevation wave. Furthermore, they observed that by increasing the initial submergence, the amplitude of both the leading subsidence and the following elevation waves increases. The initial slide's acceleration, which is obtained by the velocity time series and described by the incline's angle to the horizontal, the slide's mass and shape and also the Coulomb friction, it was shown to be a very significant parameter affecting the waves run-up.

As described by Liu et al. [2005], the initial acceleration is relatively constant for the case of a submerged slide, whereas for a subaerial slide, the acceleration increases as the distance above the water level increases. The initial acceleration of a submerged landslide is smaller due to larger inertial and viscous effects. Furthermore, for submerged slides, the resulting run-up increases by either decreasing the submergence or increasing the initial acceleration. In the case where the submergence depth is more than three times the slide's height, the slide becomes ineffective to generate any wave motion. Also, as the incline's slope decreases, the initial acceleration decreases as well. For significantly small incline's slope, only underwater landslides with very large dimensions are able to generate catastrophic tsunami waves. Liu et al. [2005] observed that generally the subaerial slides result in larger run-up compared to submerged slides and that the run-up and run-down can be controlled by the slide's size, submergence and it's initial motion. Another important observation was the significant effect of the sidewall on the resulting maximum run-up. This is because in cases where the waves reflected on the tank boundaries, the measured wave elevations at different locations were larger than for the cases where no reflection observed.

Panizzo and De Girolamo [2005] performed experiments where symmetric solid non-deformable slides start from rest and slide down an incline, accelerating due to gravity. The authors pointed out that a subaerial landslide falling into water produces waves that travel long distances, cause disaster far away from the generation area and result to high wave run-ups on the shoreline. At the experiments performed the water viscosity was neglected, whereas the landslide density, porosity and angle between he landslide front and the incline were kept constant. They observed that the landslide motion above the water surface was described by constant acceleration with little friction. At the time the velocity reached the maximum value, a sudden deceleration was noticed, which is the instant referred to as the impact time. After the impact, the landslide motion under the water was decelerated until it reached the tank bottom. The duration of the underwater motion increases by decreasing the impact velocity and the landslide's volume. Additionally, the maximum wave elevation depends on both the duration of the underwater motion and the front shape of the slide.

Lynett and Liu [2005] performed numerical studies about tsunamis generated by a three-dimensional sliding mass travelling down a slopping beach. The whole life cycle of the tsunami have been investigated looking both at the cases of subaerial and submerged slides. Unfortunately, landslide-tsunami experiments usually face the problem of reflection on the tank walls, leading to non representative wave amplitude measurements, which is the case for Lynett and Liu [2005]'s study. Their approach was to consider a weighted average of the aerial and submerged velocities as the slide's velocity, with the weighting to be the fraction of the landslide's submergence. They investigated the dependency of the resulting tsunami wave amplitude on six dimensionless parameters: the slide's thickness ϵ, the ratio of the front to back lengths of the slide κ, the slide's steepness μ, the slide's horizontal aspect ratio (length/width) A, the specific gravity of the slide's material γ and also the slope of the beach $S = \tan\theta$. It is concluded that for submerged slides, the dimensionless wave elevation decreases with decreasing ϵ, which is not the case for subaerial slides. For the case of a subaerial landslide, there is a dependency on γ. A weak dependency of the wave elevation with the parameter κ is observed. Additionally, the authors noticed that there is no dependency on either μ, A or S for both subaerial and submerged slides.

Enet et al. [2005] performed large scale laboratory experiments and three-dimensional fully nonlinear simulations to investigate tsunami generation by underwater landslides. A Gaussian shaped solid underwater body was released on a rail down a plane slope from an initial submergence depth 80 mm and at the end of the slide's motion, a significant deformation and spreading was noticed. The tsunami wave surface elevation was measured using wave gauges placed at strategic locations. On another study Watts and Grilli [2003] showed that the slide's length increases and thickness decreases as the travelled distance increases. They concluded that the tsunami amplitude dependency on the centre of mass motion is much stronger compared to the dependency on the landslide's deformations. Using this fact, Enet et al. [2005] assumed that the slide's motion is equal to the centre of mass motion and concluded that there is a strong correlation between the submergence depth and the resulting coastal run-up.

A tsunami modelling strategy resulting to the GEOWAVE tsunami model was proposed by Watts et al. [2003]. GEOWAVE is able to simulate run-up and inundation simultaneously. The landslide's deformation has been neglected, following the conclusion reached by Watts and Grilli [2003], and the slide's motion is considered to be the same with its centre of mass. Also the viscous drag forces are neglected. Watts [1998], Watts [2000] and Watts et al. [2005] concluded that the initial acceleration is the most important parameter in tsunami generation. The foremost concluded that the duration of the tsunami generation is equal to the duration of the landslide's acceleration. The particular modelling strategy used by Watts et al. [2003] is based on the three steps of tsunami's life cycle: generation, propagation and inundation. GEOWAVE model resulted by combining three different models. The first one is a model for the centre of mass motion of the slide, the so-called wavemaker model. The second is the tsunami generation model for specified submarine mass failure shape and motion and it is based on both two- and three-dimensional full nonlinear models. Finally, a model for the tsunami propagation and inundation was also considered, which is based on the fully nonlinear Boussinesq equations. The tsunami propagation and inundation have been simulated using the long wave propagation model FUNWAVE based on the fully nonlinear Boussinesq equations. The reason of choosing a Boussinesq wave propagation model instead of a nonlinear shallow water wave model is to give the flexibility for the horizontal velocities to varying over depth. The GEOWAVE model was finally validated with case studies, proving its ability to give accurate wave elevation and horizontal inundation at the same time.

Another research on landslide-generated tsunamis carried out by Watts [2004] where it is concluded that the tsunami waves amplitude is independent on the incline's slope and is strongly dependent on the maximum slide's thickness. Moreover, the authors state that approximately every 75 years a landslide-generated tsunami with an amplitude greater than $1\mathrm{m}$ is expected to reach Southern California. Furthermore, they concluded that subaerial landslides have significantly higher ability of generating tsunamis compared to submerged ones. Finally, they pointed out that stronger sediments create both larger landslides and tsunamis. Watts et al. [2005] found out that the tsunami waves amplitude is approximately proportional to the landslide's thickness.

Large-scale three-dimensional experiments to study landslide-generated tsunamis performed by Enet et al. [2007], where the experimental results used further for validation of a three-dimensional numerical model. Their experiments highlighted the strong dispersive and directionality nature of the landslide-generated tsunamis. The coastal run-up measurements showed that there is a strong correlation between the run-up and the initial slide submergence depth. Additionally, even the fact that the deformation effects are more important for 3-D slides. Moreover, they considered very small friction between landslide and include, which was negligible compared to hydrodynamic drag. They also considered the motion of the rigid slide down the slope to be equal to its centre of mass motion. Similarly to Watts et al. [2005], the authors concluded that the landslide's initial acceleration is the most important parameter for landslide-generated tsunamis, whereas the landslide velocity plays a secondary role. From the laboratory experimental results, it was observed that at the beginning of the water motion, a subsidence wave generated initially above the slide's initial location and then the wave rebounded to form a main elevation wave propagating offshore. The rebounded wave also propagating in the shore direction, reflected on the slope, resulting to run-up and some smaller waves.

Tappin et al. [2008] made an extensive overview of the Papua New Guinea tsunami of 17 July 1998 and highlighted the major impact of this event on tsunami science, that is possibly the most studied Submarine Mass Failure (SMF) tsunami. The particular tsunami was not earthquake-generated, but it was created by an underwater slump probably triggered by an earthquake. Tsunami investigations took place some years ago supported that underwater landslides are not able to create significant tsunamis. These beliefs have been refuted by the PNG tsunami event. The event initiate by a 7.1 magnitude earthquake. The specific magnitude was small compared to the induced approximately $10\mathrm{m}$ high tsunami that hit and completely destroyed three villages. Additionally, the earthquake epicentre was very close to the shore, which usually results in an almost immediate tsunami impact. However, this was not the case for the PNG tsunami, that reached the coast approximately 20 minutes after the earthquake event. Moreover, the coastal run-up measurements are not indicating an earthquake source. Therefore, there was no evidence that the earthquake was the tsunami source, with the most likely alternative to be a SMF. Early investigations of the event considered translational landslide source as the most likely mechanism of failure and tsunami generation, despite the fact that the first published results in 1999 concluded that a slump was the most likely source. Tappin et al. [2008] confirmed that the mass failure was a slump. Using the equation of motion described by Grilli and Watts [2005], the authors calculated a characteristic time of slump motion, an initial acceleration and a maximum velocity. Finally, they concluded that landslides initiated in shallow water and traveling down the continental slope can generate tsunamis over considerable distances, whereas slumps do no travel as far as landslides, because of basal friction, and so their potential to generate tsunami is reduced.

Due to the fact that both the shoreline movement and the wave field are induced by the landslide interaction with the water, to study further landslide-generated tsunami models, an appropriate two-horizontal dimension model is necessary. Unfortunately, the number of the existing two-horizontal dimension analytical landslide-generated tsunami models still remains limited. Sammarco and Renzi [2008] made an important contribution by developing an analytical three-dimensional model for tsunamis originated by landslides based on the forced linear long-wave equation of motion, considering a plane beach with a constant slope. The inputs of the model are the initial position, speed and spread ratio of the landslides. The output is the free surface elevation at specific times and locations. Sammarco and Renzi [2008] concluded that the landslide generates a wave field that is composed by two components, the oscillatory and the evanescent. The evanescent component decays rapidly with time, with the smaller the landslide's width or steeper the sloping beach, the faster the decay.

Furthermore, it was observed that at the beginning of the wave motion, the landslide pushes water in the offshore directions and a leading elevation wave is generated. At the same time, a subsidence wave occurs moving landwards. Later on, the wave motion is mainly along the shoreline, while no motion is noticeable around the origin. For the case of landslide-generated tsunamis along the sloping beach, the large waves are not in the front of the wave-train, as for the case of a constant depth, but they are located in the middle of the wave-train. By comparing to available experimental data, Sammarco and Renzi [2008] showed that the analytical model represents the overall behaviour of the wave with acceptable accuracy. Neverheless, an important drawback of the model is that it neglects energy dissipation. This leads to overestimated predictions of run-ups and run-downs.

2.1.3 Subduction Zone Earthquakes

Chapters 4 and 5 consists of an extensive investigation on the source and specifically the seabed deformation in case of Subduction Zone earthquakes. Hence, at this point it is worth providing a brief definition for Subduction Zones. First of all, the upper layer of Earth, known as the crust, together with the upper part of the mantle layer underneath made up the lithosphere. The lithosphere is broken into a number of large tectonic plates[3] that fit around the Earth like puzzle pieces and meet at the plate boundaries, known as the faults. The lithosphere converges, resulting in plates motion toward each other at the convergent boundaries, where depending on the plates' nature they either form a subduction zone or a continental collision. Subduction zones are the boundaries between the plates and they are marked by trenches. A subduction zone is generated when an oceanic plate meets and moves beneath another plate, which can be consisted of either oceanic or continental crust. The subduction process at these boundaries commonly results to geohazards, such as earthquakes and volcanic eruptions [Yeats et al., 1997].

Earthquakes occur when energy stored in the elastically strained plates is suddenly released. Even though that an earthquake can be extremely catastrophic in just a few seconds, the strain generation is very slow. Therefore, earthquakes at the same location occur at intervals of hundreds to thousands of years [Yeats et al., 1997], according to the plates convergence rate and the percentage of this convergence that is accumulated in the locked section of the plate boundary. The rate of convergence is measured in centimetres per year, with an average of approximately 2 to 8 centimetres per year.

Earthquakes originate from the seismogenic zone[4], which is a layer of Earth's crust. In the time intervals between earthquakes, the plates become "locked" at this zone. This leads to large amounts of strain to build up and it is released at some point in one or more earthquakes. A significant sudden vertical deformation of the seabed typically occurs during a large subduction zone earthquake at the plate boundaries. This results in a vertical displacement of the ocean water above. As the displaced water regains its equilibrium, catastrophic tsunami waves are generated.

Generally, the most catastrophic tsunamis are generated from large and shallow earthquakes at the plates interface. This type of earthquakes are known as "megathrust" earthquakes, they occur every 300-500 years and they are the world's largest earthquakes. It is noteworthy that no strain is able to build up somewhere outside this zone, since the plates move always smoothly and hence no large earthquake can be generated. The zone can be locked for many years as in the case of the Cascadia Subduction Zone which hasn't experienced any megathrust earthquake since 26 January, 1700 [Satake et al., 1996]. The magnitude of this last earthquake was estimated as 9.0 and it resulted in a tsunami event.

Earthquakes occur either within the overriding plate, along the plate interface or within the plate being subducted (Figure 2.1). For almost 100 years since 1900, the total seismic

---

[3] The tectonic plates are able to move and are composed of oceanic and continental lithosphere, each topped with oceanic and continental crust, respectively

[4] Seismogenic is a fault that is able to generate earthquakes of significantly high magnitude (greater than 5)

activity worldwide has been dominated by large shallow earthquakes on subduction zone plates. Nevertheless, in most of the cases these earthquakes didn't cause large disaster due to the fact that they were located offshore and hence the largest risk was from a resulting tsunami instead of earthquake shaking.

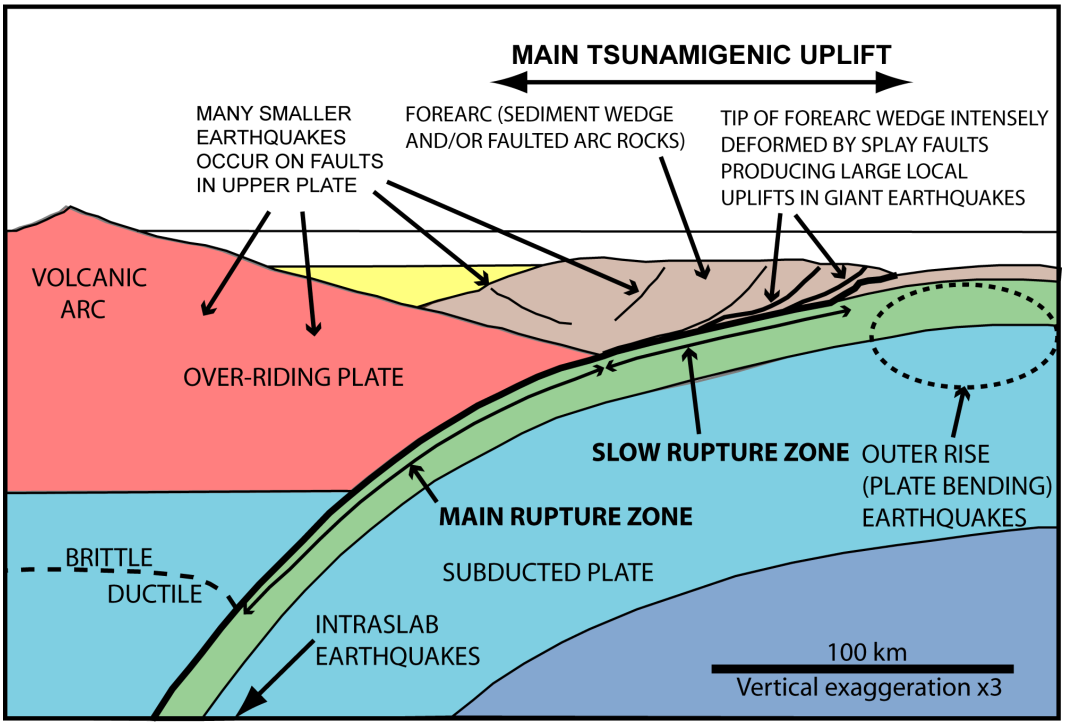


Figure 2.1: Illustration of the shallow part (approximately the upper 50 km) of a subduction zone. Source: Diagram courtesy of Dr Simon Day, based upon Byrne et al. [1988].

2.1.3.1 Cascadia Subduction Zone

In Chapter 5, earthquake events on the Cascadia Subduction Zone are investigated. The Cascadia Subduction Zone (CSZ) is located at South-Western Canada and North-Western United States and is shown in Figures 2.2 and 2.3. At the CSZ the oceanic crust of the Juan de Fuca plate is subducting below the continental crust of the North America plate. The seismogenic zone has landward limit extended slightly beneath the coasts at some locations and it is away from the coast at the rest of the locations. This means that the larger cities that are located 100 to 200 km inland, e.g. Victoria, Vancouver and Seattle, do not experience ground motion from large subduction zone earthquakes. As stated earlier, the specific subduction zone has very low seismicity, with the last event taking place 300 years ago [Atwater et al. [1995], Yeats et al. [1997], Atwater et al. [2006], Satake and Atwater [2007]]. There are no historical records of large earthquakes for almost 200 years in the CSZ. Therefore, recurrence intervals[5]

---

[5] Recurrence interval is the average of a number of interevent times

are uncertain because of difficulties in identifying and dating earthquakes.

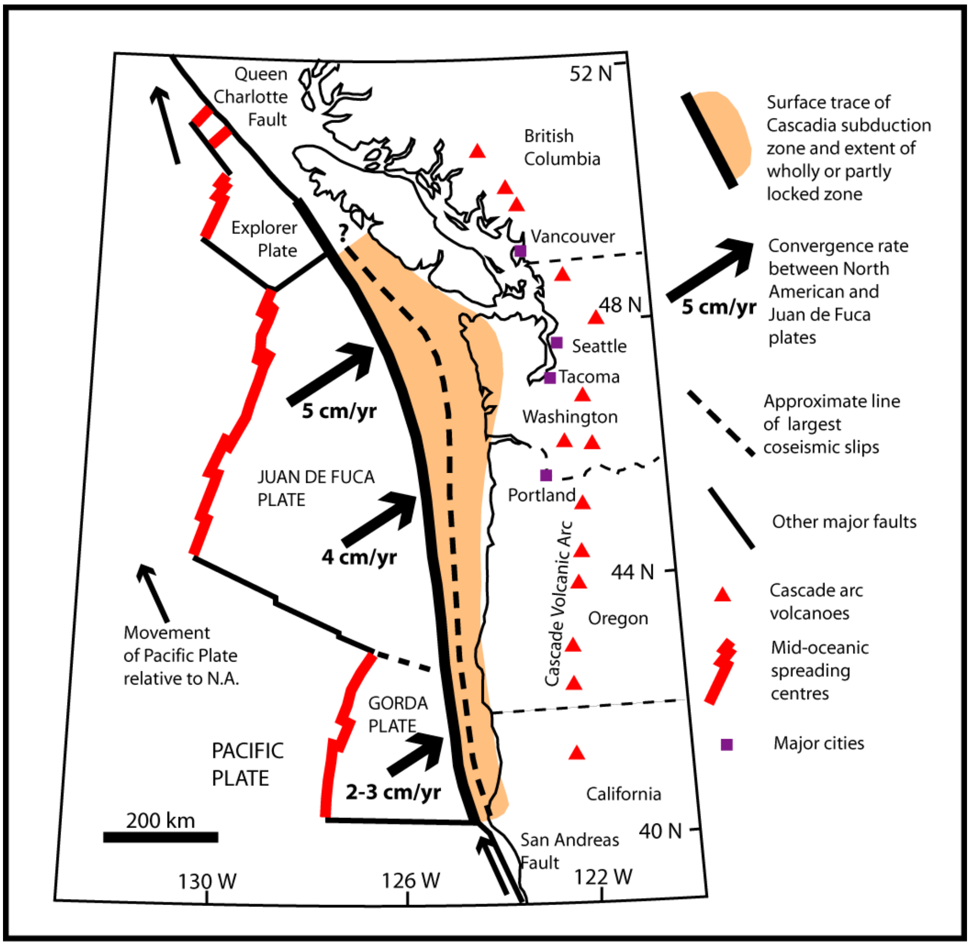
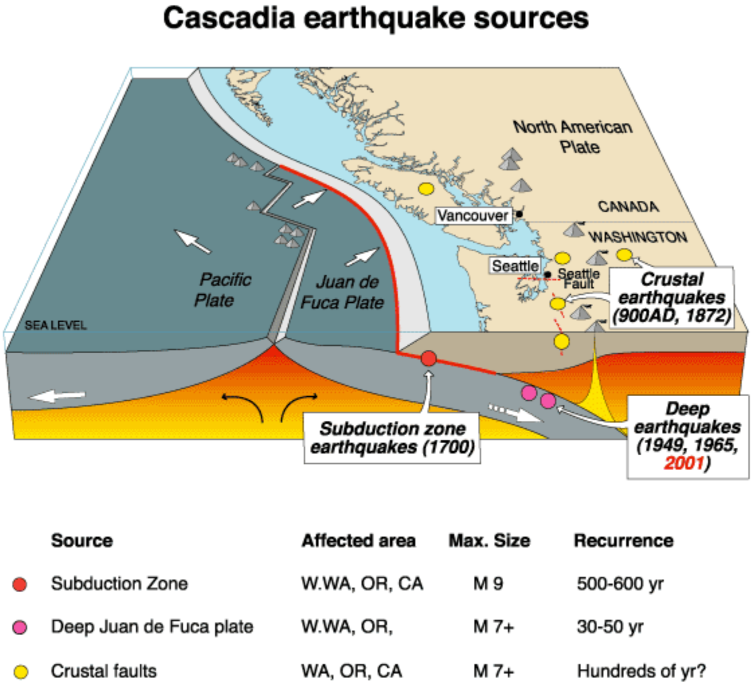


Figure 2.2: The Cascadia margin of western North America. Source: Diagram courtesy of Dr Simon Day, based upon Hyndman and Wang [1995].

However, there are strong geological evidences onshore and offshore, such as costal layers indicating sudden land level changes [Atwater et al. [1995], Satake et al. [1996], Clague [1997]] and turbidite deposits [Goldfinger [2011], Goldfinger et al. [2012]], respectively. As Atwater et al. [1995] noted, the coastal displacement can be explained by a rupture at the plate boundary. Also there are biological evidences, such as tree rings, as well as human records, of past earthquakes. Paleoseismicity data [6] from many locations along the coast give an indication that earthquakes have occurred at the CSZ at irregular time intervals approximately every 600 years on average. Similar margins at other regions experienced very damaging earthquake events. From geological reports it is known that approximately 40 significantly large earthquakes occurred in total in the last 12,000 years. Goldfinger et al. [2003] stated that it is feasible to construct reliable event records for the last 10,000 years for CSZ by carefully correlating the

---

[6] Paleoseismology is the study of past earthquake events using geological methods



	Source	Affected area	Max. Size	Recurrence
●	Subduction Zone	W.WA, OR, CA	M 9	500-600 yr
●	Deep Juan de Fuca plate	W.WA, OR,	M 7+	30-50 yr
●	Crustal faults	WA, OR, CA	M 7+	Hundreds of yr?

Figure 2.3: Earthquake sources in the CSZ (U.S. Geological Survey)

turbidite and land records.

As noted by Rogers [1988], it is typical to have very little or even no seismic activity between large earthquake events. This indicates a potential for large earthquakes in the future in the CSZ region. From the investigations performed so far focusing on the CSZ, a generally accepted conclusion is that it is a locked thrust. Geodetic measurements and geophysical modelling results indicate that the locked thrust accumulates elastic strain[7] that is going to be released during a large earthquake in the future.

The plate-boundary rupture at the CSZ is 50 km wide and 200 km long, with the whole rupture length to be around 1000 km. The size of the rupture suggests magnitude-8 earthquakes, as these magnitude earthquakes causes ruptures of an area approximately 10,000 km^2. During the last earthquake, 300 years ago, the rupture covered the whole length of the subduction zone. Therefore, in the case that it was a single event, the earthquake's magnitude was probably as large as 9.0 [Hyndman and Wang, 1995]. The narrow width of the subduction zone limits the size of the earthquakes. However, magnitude over 8.0 earthquakes are possible as the maximum depends on the margin's length. Worldwide [2013] support that the CSZ is capable to produce

---

[7] Strain is a deformation of a body due to an applied stress. The elastic strain is the deformation that disappear completely instantly when this force is removed.

powerful earthquakes that create catastrophic long-period seismic waves and these earthquakes can result in tsunamis.

Goldfinger et al. [2012] collected an earthquake record for 10,000 years at the CSZ, with clusters of great earthquakes separated by gaps of many hundreds of years. Kulkarni et al. [2013] statistically analyse Goldfinger et al. [2012]'s record in order to calculate time-dependent recurrence intervals and probabilities. Specifically, Monte Carlo simulations have been used to check whether the events follow a homogeneous Poisson process. A Poisson process refers to an event that happens randomly in time, with the time between each pair of consecutive events follows an exponential distribution. Additionally, any inter-events time is assumed to be independent of other inter-event times. In the case that there is an evidence of a non-Poisson process, it means that the process is time dependent. They concluded that there is strong evidence that the earthquakes do not follow a homogeneous Poisson process. Furthermore, Kulkarni et al. [2013] performed cluster analysis in order to identify potential clusters.

2.2 Computer models for tsunami events investigation

A deep knowledge about tsunamis is required in order to be able to accurately predict the wave inundation and run-up and produce helpful early warning notices to the regions that might be affected by a possible tsunami event. To obtain this knowledge, an investigation of both the geological evidence and impacts of past tsunamis, as well as the generation source, is necessary. Most of the time, it is difficult or even impossible to collect data for a tsunami event due to the fact that tsunamis do not frequently occur. The inadequate number of observations available is the reason of the high uncertainty in the description of the source at the generation phase, even after many years of scientific investigations in this area.

The small number of data collected from past events prevents the extended investigation of the events. Therefore, computer models, also referred to as simulators, are used to study these events. Tsunami simulations can represent past events and model potential hazard scenarios in order to compute tsunami waves inundations. The computer models can be employed in advance to generate evacuation maps that can help in mitigating the hazard. To produce evacuation maps, the maximum possible coastal inundation for the specific location have to be computed. Additionally, the computer models can be used for insurance applications, where the annual probability of exceedance of a particular water elevation height is required, see Dominey-Howes et al. [2010] and Wood and Goof [2004].

Computer models are extensively used for analyses of complex physical phenomena as they can simulate real-world physical processes. Unfortunately, due to the uncertainties about many tsunami features, it is challenging to simulate and analyse the water flow. The reliability standards for real-time predictions for natural hazards are very high and the development of an accurate model that can represent real process is very challenging. For the development of the computer models, a number of choices is required, where these choices depend on the available expert knowledge and on budget constraints. The latter affects the number of experimental data that can be obtained from laboratory tests. Each of these choices affect in a positive or negative way the computer model performance on representing reality and hence the amount of uncertainty in the predictions.

Computer models are used by the Tsunami Warning Centres (TWCs) at locations that are in high danger of tsunami events, for example at Indonesia, at Pacific Ocean. The TWCs announce real-time warnings and provide information to the local communities about threatening tsunamis that they might possibly experience. Therefore, the TWCs help generally in the risk reduction of future tsunami event. The most efficient way of hazard mitigation is achieved through a combination of accurate forecasting models and proper education of the local population. Unrealistic estimates and unnecessary fake alarms are not desirable, as they can cause needless panic and evacuations that can be costly. A combination of computer model evaluations and historical data of past events is used by the TWCs to create precomputed event scenarios. However, the historical data are usually incomplete and unreliable and this makes the TWCs' job very challenging. The precomputed scenarios are combined with the new observations from the event itself in order to obtain the predictions and fast warnings and also to generate inundation maps [Titov et al., 2005; Titov, 2009].

Nevertheless, even the most accurate computer models do not represent reality perfectly. The model input parameters are usually uncertain and in most cases, approximations are used within the computer model to represent mathematical concepts and make the model tractable [NRC, 2012]. These approximations introduce error and result in discrepancies between model and reality. George Box observed that “all models are wrong, but some may be useful”. In other words, no computer model can predict with 100% accuracy the real process, with error and uncertainties in predictions being unavoidable. However, they can predict with acceptable tolerance. The report by NRC [2012] investigates ways to generate the most useful models by quantifying how wrong they are. By understanding the model limitations and the uncertainties in the predictions, the results can be used with confidence. In the cases where the discrepancy between predictions and model evaluations cannot be quantified, the computer model evaluations are not useful.

It is challenging to understand the impact of the possible sources of model limitations and uncertainties, such as input uncertainty and limited model evaluations, on the resulting model predictions. Nevertheless, the assessment of the models' predictions before being employed is absolutely necessary and it is demanded in case that the computer model is used for natural hazard events, real-world forecasting, emergency planning and development of tsunami impact elimination plans, which means that important decisions are made based on these predictions. For tsunami events investigations, additionally to the demand of high accuracy in the predictions of the computer models, large computational domains are also necessary. The high accuracy and large domains impose time and resources constraints, that leads to delays in predictions.

Additionally, a proper estimation of the computer model uncertainties is required. Indeed, an accurate representation of the uncertainty in the predicted tsunami characteristics corresponding to uncertain trigger features, e.g. either position, shape and speed of the landslide, or sea floor deformation associated with earthquakes, can result in a positive contribution to the hazard mitigation. However, the assessment and further investigation of the computer models requires a large number of model evaluations. Usually the computer models are computationally expensive and time consuming to evaluate, when high accuracy and large domains are required. The computational cost of the simulations increases with the complexity of the model and the spatiotemporal resolution required. Therefore, further analyses of the computer model are difficult to be accomplished due to the high computational cost.

2.2.1 Available models for tsunami investigations

The shallow water equations, either nonlinear, or linear in the simplest case, are commonly used for tsunami modelling, since the tsunami waves' horizontal length scale is greater than the vertical length scale. More specifically, tsunamis are long waves and therefore their wavelength is significantly larger and their amplitude significantly smaller compared to the water depth. Therefore, a 2-dimensional approximation is fine. The shallow water equations are derived from the Navier-Stokes equations, that represent mass and energy conservation principles. Unfortunately, the swallow water equations neglect dispersion. But this is a not a big issue and usually does not have any significant consequences, as explained by Dutykh et al. [2011].

The Nonlinear Shallow Water Equation (NSWE) system can capture significantly well the long waves' features. Therefore, it is frequently used for tsunami modelling. This system is given by the following equations:

$$H_t + \nabla \cdot (H\vec{u}) = 0 \tag{2.1}$$

$$(H\vec{u})_t + \nabla \cdot (H\vec{u} \otimes \vec{u} + \frac{g}{2}H^2) = gH\nabla h \tag{2.2}$$

where h is the bathymetry height, η is the water elevation, $H = h + \eta$ is the total water depth and $\vec{u} = (u, v)(\vec{x}, t)$ is the depth averaged horizontal velocity.

The numerical solution of the NSWEs is a research area that is extensively studied and many computational codes that solve these equations currently exist, such as Anastasiou and Chan [1997], Garcia-Navarro and Vazquez-Cendon [2000] and Audusse et al. [2004]. However, these codes are not generated to be used for tsunami modelling and real-world applications and they can be very computationally expensive. On the other side, there are many codes that can be used for tsunami modelling, but their disadvantage is that they use old-fashioned numerical methods to solve the NSWEs [Dutykh et al., 2011]. Two of the tsunami numerical models that are commonly used are the Method of Splitting Tsunami (MOST) [Titov and Synolakis, 1998] and the Cornell Multi-grid Coupled Tsunami Model (ComCot) [Liu et al., 1998].

The VOLNA tsunami model was developed by Dutykh et al. [2011] to overcome this problem, since it combines both modern numerical methods for solving the NSWE and also it can be used efficiently for tsunami modelling in realistic environments. VOLNA can handle the complete life-cycle of a tsunami: generation, propagation and inundations. VOLNA was verified using benchmark cases by Dr Joakim Beck, Research Associate at UCL, where the predictions were compared with the analytical solutions, concluding that VOLNA's predictions are very accurate [Beck and Guillas, 2014]. VOLNA is used for the analysis in Chapters 3 and 4, where earthquake-generated tsunami events are investigated.

2.2.2 Validation and Verification

In order to find out how accurate the computer models are, the processes of validation and verification have to be employed. These processes are explained in detail by Sargent [2005] and NRC [2012]. The validation process determines how well the mathematical model represents reality. Generally the reliability of the mathematical model's predictions depends on the quality of the available information, that is collected from experiments and expert knowledge, and it always contains uncertainties. The verification process assesses how well the mathematical equations are solved by the computer model and whether the computer models' predictions are accurate, i.e. the verification consists of the analysis assessing the correctness of the code. Figure 2.4 displays which process assesses each steps when the reality is represented by computer predictions.

Sargent [2005] describes different approaches and techniques for model validation and verification. NRC [2012] illustrates validation and verification and discusses additionally the uncertainty quantification process. All the three processes, validation, verification and uncertainty quantification, can help with understanding the impact of error and uncertainties in the

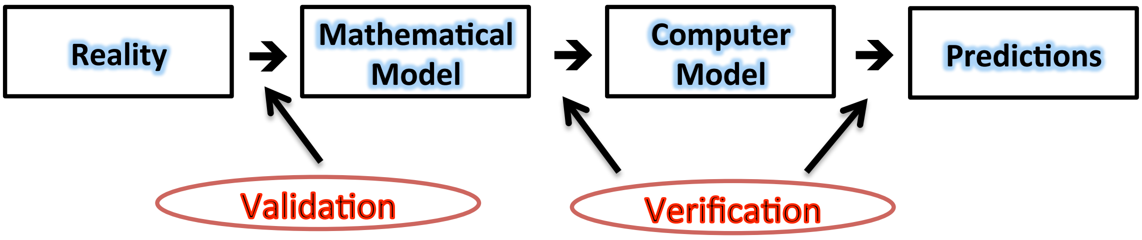


Figure 2.4: Scheme illustrating the use of Validation and Verification in the process of representing reality with computer models. Source: Based upon Babuška et al. [2007].

predictions. Specifically for tsunamis, Synolakis et al. [2008] present analytical and laboratory tests that can be used for validation and verification of tsunami models. They emphasise that the model validation must be a continuous process, and the computer models have to be additionally tested when new knowledge and additional data are available.

The concept of predictions that include uncertainties is also proposed by Bayarri et al. [2007]. The authors present a computer evaluation process that addresses the question of whether the computer model represent reality adequately, based on Bayesian and likelihood methodology. Their process produce tolerance bounds for the model predictions that take into account the key uncertainties in the problem. The approach of Bayarri et al. [2007] does not focus on investigating whether the model is correct or not and if it adequately represents reality. However, the main focus is to check if the model's predictions are sufficiently accurate for the intended use.

The tolerance bounds for the predictions give the chance that the corresponding true process value would lie within the specified range. Bayarri et al. [2007] suggest these tolerant bounds to be routinely provided along with the predictions. Large tolerant bounds indicate that the model is considerably less accurate at the specific input values compared to input values where the bounds are smaller. Bayarri et al. [2007] additionally focused on the computational expense in performing model validation when the computer model is demanding to be evaluated and suggested an approximation of the model with a statistical model that can be used for the validation process. Indeed it is too costly and time consuming to determine that a model is valid, i.e. to perform verification and validation, over the whole domain of its intended applicability, especially when high confidence is required. Furthermore, Bayarri et al. [2007] showed some ideas for the application of the verification and validation processes when only limited data are available.

Goldstein and Rougier [2009] demonstrated an approach for relating the behaviour of the mathematical models with the inferences about the physical systems that are represented by the model. The authors named this approach “reified analysis”. It is concerned with the logical framework within which a proper uncertainty analysis have to be constructed and offers an approach for assessing all the uncertainties arising in relating the simulators to the system that those simulators represent.

2.3 Scientific investigations on earthquake-generated tsunami events

The most common approach to define the source characteristics is to find the ones that best match available observations of tsunami inundations or other signals such as ground shaking. This approach is known as inverse modelling and it usually requires strong assumptions that cannot be justified by the existing data available after an event. Following that, the seabed displacement is estimated using these earthquake source characteristics, with the motion usually considered as instantaneous.

However, Titov et al. [2005] observed that small-scale features of the tsunami source may not be critical when the tsunami propagation away from the source is investigated, since it is not sensitive in such details. The authors concluded that most critical source parameters determining characteristics of the far-field tsunami propagations is the magnitude of the seafloor displacement and its location. Also, Titov et al. [2005] demonstrated the ability of localised earthquakes to transport energy at very large distances. They focused on the devastated tsunami event of Sumatra, Indonesia in 2004 and they concluded that large tsunamis can propagate substantial and damaging wave energy to distant coasts, including different oceans. Grilli et al. [2007] also focused on the Sumatra tsunami event and developed a calibrated tsunami source in terms of the coseismic seafloor displacement and rupture timing along the trench, based not the available seismological and geological data. They specified the tsunami source as a series of discrete, discontinuous dislocation source segments located along the trench. Seismic inversion models are used to predict the rupture propagation direction, duration and speed. The source obtained is used to build a numerical model of tsunami generation, propagation and coastal inundation and run-up for the December 26, 2004 event. Their work aimed to achieve a better understanding through modelling of this catastrophic event, that can help the scientific communities in more accurate predictions and mitigation of similar future disasters.

A large number of source inversion analyses have been performed after the Tohoku 2011 tsunami, in order to derive the coseismic slip distribution for the event. These analyses use strong motion data [Suzuki et al., 2011], tsunami data [Lay et al., 2011], geodetic data [Yue and

Lay, 2013] and joint inversion of multiple data sets [Simons et al., 2011]. Strong motion data are recordings of how damaging earthquake ground motions affect the buildings, with monitors placed on or close to the ground surface, in areas where earthquakes are more likely to occur in order to get as much information as possible when an event occurs. From strong motion data one can derive how much the ground motion increases as the magnitude increases, at given distance, as well as how much the ground motion decreases as the distance increases at a given magnitude. Tsunami data for the event can be obtained through coastal wave gauges and GPS buoys typically positioned around 20 km away from the coast. Geodetic data are measurements of the coseismic displacement. Finally, the joint inversion is the method where a solution is obtained with multiple types of data sets involved and inverted simultaneously.

The most common tsunami generation mechanism is the sea bottom deformation due to underwater earthquakes. The largest of these are subduction zone earthquakes, that often cause significant seabed displacements resulting in large tsunami events. For the case of earthquake generated tsunamis, three main characteristics are required for tsunami wave predictions. The first is the earthquake source description. The second is the coseismic seabed deformation as changes over time, known as dynamic or active generation, and the final is the presence of sediments at the source [Dutykh and Dias, 2010]. The most common approach in the literature was to assume that the seabed deformation is entirely transferred to the water free surface as an initial condition, known as the Okada solution [Okada, 1992]. This approach is known as passive generation. It ignores the seafloor characteristics and simply assigns an initial condition estimated using seismic inversion methods. It is noteworthy that this approach is applicable only for constant or nearly constant depth cases. In reality, since tsunamigenic seafloor displacements occur repetitively and build up some of the largest scale ocean terrains, such as ocean trenches and continental slopes, tsunamis produced by tectonic displacements invariably occur in regions with large scale bathymetry relief and so the water depth in regions of tsunami generation is far from constant. Therefore, the passive generation approach is not fully valid for real subduction zone zones.

Dutykh et al. [2006] focused on the comparison of passive and active tsunami generation. At the active tsunami generation the coseismic displacement is considered, instead of the "frozen" seabed deformation to be transitioned to the water water, which is the passive generation case. They concluded that there are significant differences between the two type of deformation. Specifically, they showed lower wave amplitudes resulting from active generation compared to the ones resulting from the passive generation, with differences of the order of 20%. They even looked at the case of a very fast seabed motion, for which they concluded that the resulting tsunami wave elevation is still lower than the passively generated one.

One of the main problems in tsunami modelling, similarly to other types of complex numerical modelling, is the trade-off between the accuracy of model inputs and the physical process representation on the one hand, and on the other hand the model computational expense. Computational expense, mainly time expense, is especially critical in real time tsunami warning applications. This chapter will investigate the dependency of the induced tsunami waves time series, in the case of an earthquake-generated tsunami, on the seabed deformation representation. According to the sensitivity of the model output upon the input source characteristics, the computational effort put into the numerical modelling may or may not be justified.

A novel coseismic displacement representation is proposed. In this representation, quadratic curves over polygon shapes are used to describe the deformation. This allows both continuous and stepwise (at polygon boundaries) variations in seabed displacement as indicated by geological evidence. An important advantage of this approach is that it enables a more flexible representation of both uplift and subsidence than is possible with the standard Okada solution. This allows in particular, localised uplifts due to irregular fault geometries as well as variations in the distribution of fault slip near the surface trace of the subduction zone at the trench. Therefore, this is a more advanced representation compared to what is currently used in literature, e.g. Lay et al. [2011]. The current investigations use continuous surfaces to describe the source deformation, but they assume no displacement at the trench. Therefore, they use strict assumptions and they are less flexible compared to our approach.

2.3.1 Current tsunami source modelling and limitations

The adaptivity of the source representation employed in this analysis contrasts with the existing tsunami source description models that mostly represent the source as displacements produced by elastic deformation over rectangular fault surfaces with uniform slip, commonly referred to as Okada models, as presented by Shuto [1991], or with varying degrees of resolution, using a number of blocks to represent sub faults with uniform slop displacements in each one. The latter representation is derived typically from finite fault model representations used to inverse-model seismic records of the source fault ruptures, described by Lay et al. [2011] and Simons et al. [2011]. These models present inherent limitations due to the assumptions of uniform elastic shear modulus and the use of simplified fault geometry. They have difficulty representing the varying rock properties and complex fault geometries that lead to development of large localised seabed uplifts, especially in the near-trench area of subduction zone systems and particularly in cases where the fault rupture breaks the surface at or near the trench axis.

These types of localised but large seabed displacements are considered responsible for the largest amplitude components of major earthquake-generated tsunamis on the basis of post-event inverse modelling, see for example Saito et al. [2011] and Satake et al. [2013] for the 2011 Tohoku tsunami. Another example of a more sophisticated post-event model using a seismically determined or otherwise constrained fault movement history of the source event is the study of the 2006 Java earthquake and tsunami by Dutykh et al. [2013]. The authors proposed a novel computationally cheap simple model to describe the coseismic deformation based on the Finite Fault solution and dynamic seabed deformation and they divide the deformation area into 147 equal sub-faults. The dynamic seabed motion is considered as a function of the sub-fault's activation time, that is estimated by dividing the distance between the each sub-fault and the epicentre locations by the rupture velocity, which is assumed constant.

The post-event studies of recent giant earthquakes noted above have the advantage of seismic and geodetic evidence[8], such as finite-fault models and real time ground movement evidence from continuous-recording GPS receivers, that is used to constrain the history of fault slips during source events. This additional data has enabled the consideration of dynamic effects in these studies. This is in contrast with probability hazard studies and real-time tsunami warning applications, where the absence of such data leads to the assumption of simplified displacement histories, and often instantaneous displacements, neglecting dynamic seabed displacement effects.

During the Tohoku 2011 tsunami, large near-trench slips of up to 69m [Satake et al., 2013] occurred in relatively small areas and time intervals. These slips were spatially and to some extend temporally separated from the region of the fault rupture in which most seismic energy is released. This raises the question of what features of a seabed displacement produced by a subduction zone fault rupture are the most important in order to obtain reasonably accurate predictions of the resulting tsunami waves. For example, is it important to accurately represent the maximum displacement of any area within the region of seabed displacement, or is the broad shape of the overall region of displacement more important?

2.4 Statistical surrogate models

The best solution to the problem of computationally expensive simulations is to use a statistical surrogate model, the so-called statistical emulator, in place of the computer model. The statistical emulator gives perfect predictions at the input points that are used in its generation process, and it enables a non-linear interpolation to predict the model output at any other input point that is within the computational domain. This is demonstrated in Figure 2.5, where x represent the

---

[8] Geodetic evidence is the one determined by geodesy, where geodesy is the science involved with defining Earth's size and shape as well as exact locations of points on its surface

input points and y the model output.

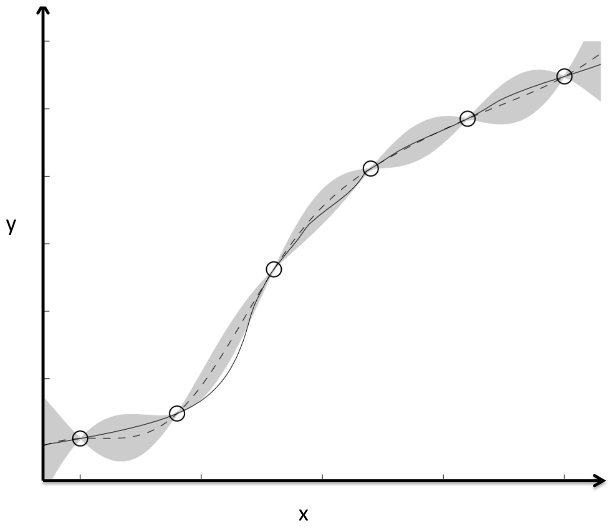


Figure 2.5: Demonstration of emulator predictions for 6 input points x. The continuous line represents the simulator's evaluations, the dashed line the mean value of the emulator's predictions and the shaded area the 95% Credible Interval. It is obvious that the emulator predictions for the computer model output are perfect at the inputs location. Source: Based upon $mucm.aston.ac.uk/MUCM/MUCMToolkit$

The significant advantage of using the emulator is that it is much less computationally demanding to be evaluated and therefore it can be employed to carry out fast predictions and expensive analyses, such as sensitivity and uncertainty analyses. The sensitivity analysis describes how significantly the model outputs are affected by changes in inputs and the uncertainty analysis estimates the uncertainties in the predictions, due to uncertain inputs. The use of the emulator in place of the simulator adds an extra source of uncertainty, since it is an approximation. However, the uncertainty induced by using it can be estimated and included into the overall assessment of uncertainty.

Sarri et al. [2012] presented an application of Gaussian Process statistical emulation for the accurate approximation of a landslide-generated tsunami model. The emulator was built from the combination of selections for regression and residuals covariance functions and parameters with a significantly small number of model evaluations. The resulting emulator was validated using Leave-One-Out diagnostics, concluding that the emulator's predictions highly agree with the computer model's evaluations. Additionally, the emulator was used further for the computational demanding sensitivity and uncertainty analyses. Sarri et al. [2012] demonstrated the potential of reducing significantly the computational cost using the emulator in place of the computer model. The results from this paper are presented in Chapter 3.

A very recent study on the quantification of uncertainties for a tsunami computer model using another statistical surrogate is performed by Sraj et al. [2014]. The authors investigate the uncertainties in the resulting wave elevation predictions, due to the uncertainty in a specific parameter, which is the Manning's friction coefficient. This coefficient is a parametrisation of the effect of bottom friction. Sraj et al. [2014] use Polynomial Chaos to build a surrogate model that is a computationally cheap approximation of the computer model. The Polynomial Chaos surrogate model is used for efficient quantification of uncertainty in the predicted wave elevations, as well as for sensitivity analysis. Following that, the authors apply Bayesian inverse modelling to estimate the posterior distribution of the Manning's coefficient using data collected during a particular tsunami event.

Additionally, for the investigation of the expensive-to-run computer model, a significant aspect is the careful choice of the model's runs that are going to be performed. The goal is to obtain the most information in the most efficient way, which means the less runs possible. Experimental design can be used for the optimal selection of these runs.

2.4.1 Predictions with uncertainties for tsunami events

In the case of natural hazards events, it is desirable to obtain predictions that involve the amount of uncertainty. The quality of the predictions, the appropriateness of the assumptions incorporated and also the computer model's capability for accurate predictions, can be assessed. This can result in more trustful and representative risk assessment. Unfortunately, even nowadays, the warnings from the Tsunami Warning Centres come without uncertainty estimates.

A careful choice of the computer model that is going to be used as well as the choice of method for the uncertainty estimation in the predictions are necessary in order to obtain reliable real-time forecasting and inundation maps. The ideal computer model for tsunami events investigation and forecasting must produce accurate and fast predictions. The predictions need to be obtained in a sufficiently short time that can give a chance to the population to evacuate the area before the tsunami waves reach the threatened coast. Additionally, the computational time required for obtaining model's evaluations must be short in order to enable many runs and therefore high resolution for the analysis. Therefore, the use of statistical emulators for real-time tsunami events is recommended.

2.4.2 Probabilistic Tsunami Hazard Analysis

PTHA is commonly used for tsunami investigations. It is implemented either by using the tsunami historical catalogues or by employing deterministic "scenario-based" models. Since the number of the available observations is usually limited, the first method can be applied only on isolated specific locations. On the other hand, the second method does not always give realistic PTHA. This is because in order to have reliable results, the probability of occurrence of each of the different scenarios has to be assigned and this is not an easy job. Additionally, the second method can be computationally demanding as it requires the simulation of a large number of scenarios, and this prevents the whole range of possible sources of uncertainties to be considered.

PTHA involves three steps; first is to define the source. The source is describe by some parameters, where each combination of the source parameters define a hypothetical event. Second is to obtain the resulting tsunami waves propagation and inundation for the different events, using computer model simulations. The final step are the probabilistic calculations, where the rate of occurrence of each event is used to calculate the rate of this particular source to result in tsunami waves exceeding particular amplitudes. Also, the probability of occurrence of a tsunami with amplitude greater than a specific height in the next T year period can also be obtained from the analysis. This requires an assumption for the distribution of the tsunami occurrence.

Looking at the case of “scenario-based” PTHA, Power et al. [2007] focused on the tsunami hazard along the coast of the New Zealand generated by earthquakes along the South American Subduction Zone. The available historical data for New Zealand may be not adequate to represent the tsunami hazard, since they cover a relatively short period of 200 years and they may be incomplete. Hence the authors concluded that the catalogue-based PTHA is not applicable. They developed simplistic statistical and physical models for the tsunami generation and propagation phases and they used them for hazard estimations using Monte-Carlo simulations. They considered a range of source locations and seismic movements as the set of possible earthquakes that can generate tsunami events. Using simple appropriate relationships they estimate first the surface deformation and slip distribution, which are used further as the initial input to a tsunami propagation and impact model. The estimated maximum wave heights are then combined with a statistical model of the occurrence of each of the different events. These combinations are used for generation of a synthetic catalogue of sources and impacts and can be used for hazard assessment for any location around New Zealand coastline.

However, the models used by Power et al. [2007] are very simple and may result in an unreliable PTHA. The only way to achieve an accurate PTHA is by taking into account all sources of uncertainty. To do so, it is necessary to take a large number of simulations of the tsunami generation, propagation and inundation model, where the tsunami model must be realistic and comprehensive. Generally, the larger the number of sources of uncertainty, the larger the number of model simulations required. To make this feasible the use of relatively low computational demanding models is necessary. But this models are usually not realistic.

Grezio et al. [2010] presented a Bayesian procedure for PTHA applied to the probability of an event per unit time (years), where an event is described as overcoming a particular threshold of a selected parameter, such as wave height or coastal runup. The posterior probability distribution is obtained by combining the prior probability distribution based on the physical knowledge and the likelihood based on the available historical data. They investigated all sources of uncertainty from the stage of tsunami generation to coastal impact. They modelled separately each part of the tsunami process - generation, propagation, impact, which allows to look separately the uncertainties of each step of the process. The authors highlighted the requirement of a large number of simulations in order to capture in a statistical significant way all the sources of uncertainty, which is computationally demanding in the case that expensive realistic models are used. In order to make the computations faster, the authors used unrealistic empirical laws. Also, they did not address the uncertainty in propagation and impact stages since they assumed that the imposed epistemic uncertainties are negligible compared to the aleatory uncertainties, in order to simplify the problem.

The recent PTHA application of Lorito et al. [2014] focused on tsunamis generated by sea floor deformation due to earthquakes, known as Seismic PTHA (SPTHA). The authors present an approach for reducing the huge number of model evaluations required for incorporating the full expected variability of the seismic sources for the SPTHA. Specifically, they introduced a semi-automatic filtering procedure for the results of the SPTHA in order to select a reduced set of sources. This procedure can be easily repeated for any location of interest. Lorito et al. [2014] highlighted the strength of their approach by comparing the results of the filtered SPTHA and the ones with the full set of sources. They concluded that their approach can results in a significant reduction (75-80%) of the number of simulations required and at the same time to maintain reasonable accuracy of probabilistic inundation maps.

Chapter 3

Statistical Emulation of a landslide-generated tsunami model

Tsunami events caused by landslides or earthquakes are capable of causing tremendous catastrophe: they are able to destroy buildings, roads and generally the infrastructure is seriously affected. However, the most tragic part of the disaster is the loss of human lives. In order to eliminate the hazard, accurate predictions of the wave elevation and coastal inundation as well as early warning notices for the population that is going to be affected, are necessary. In the case of natural hazard events, such as tsunamis and earthquakes, the predictions have to be accurate and reliable, since false alarms and non representative predictions can lead to unnecessary panic for the case of event overestimations, and on the other hand they can result in unpreparedness to cope with the situation in the case of underestimation of the event. Therefore, accurate realistic computer models are necessary for predictions.

Unfortunately, computer models that are used for tsunami modelling, such as the Sammarco and Renzi [2008]'s tsunami model, are in most cases very time and resources consuming to run. This leads to significant delays in predictions and makes further demanding analyses, such as uncertainty quantification and sensitivity analysis, impractical. In almost all the natural hazard events, the source characteristics, which could be either position, shape and speed of a landslide, or sea floor deformation associate with earthquakes, are uncertain. Therefore, there is a need to represent the uncertainty in the predictions of the computer model corresponding to the uncertain trigger features. A solution to the problem of the delays and for making expensive analyses feasible is the approximation of the computer model with an accurate statistical surrogate model.

An extensive literature review about landslide-generated tsunamis is introduced in Section 2.1.2. This chapter presents a proof-of concept case study for the statistical analysis of a landslide-generated tsunami model, by employing the analytical model constructed by Sammarco and Renzi [2008]. The main strategy followed in the analysis is to build a statistical emulator that accurately represents the analytical model [Sarri et al., 2012]. Specifically, an Outer Product Emulator (OPE) [Rougier, 2008] is generated, which is an efficient convenient approximation of the standard emulator. Following that, the resulting emulator is used for fast predictions, as well as uncertainty and sensitivity analyses. Initially, an overview of the definition of the statistical emulator as well as the OPE is provided. An analytical description of the appropriate parameter selections and calculations required to build the emulator are also presented and the concept of the experimental designs and its implementation is described. Additionally, the application of the OPE, its validation and then the very efficient uncertainty and sensitivity analyses for the Sammarco and Renzi [2008] tsunami model are presented.

3.1 Statistical Emulator

Complex computer models like the ones used for tsunami modelling, that investigate large computational domains and demand accuracy in predictions, are usually very computational expensive to be evaluated. Therefore, limited number of model evaluations can be obtained. The use of emulators comes as a solution to this problem, since emulators run almost instantaneously. A statistical emulator is a simple statistical model that approximately represents the simulator. Specifically, for any input, the emulator gives probabilistic predictions of the output that the simulator would give if it was used for the specific input. Given some inputs $\mathbf{x}$, the simulator output is given by $\mathbf{y} = f(\mathbf{x})$ and the emulator is denoted by $\hat{f}(\mathbf{x})$, where the hat indicates an approximation. The principal advantage of using an emulator is that it is very fast to be evaluated and so can be used in place of the expensive simulator. Also, the fact that the emulator runs almost instantaneously, makes feasible analyses that would otherwise be impossible because of computational intensity.

The emulator is created by employing a number of simulator evaluations. The error in its predictions is inversely related to the number of simulator evaluations. A significantly large number of evaluations can make this error negligible, but this case is unusual due to the simulator computational complexity. Since the emulator represents a deterministic input-output model, it is also a deterministic model. This means that by running the simulator at the same inputs for more than one time will always result to exactly the same outputs. Therefore, the emulator can perfectly predicts, with zero error, the output at input points that have been used for emulator's generation. At new points the emulator gives a distribution for $f(\mathbf{x})$ with mean value $\hat{f}(\mathbf{x})$ and standard deviation representing the error in prediction, that is how close it is likely to be to the true simulator output $f(\mathbf{x})$. Hence, the emulator gives probabilistic predictions, not just numbers. Due to the fact that emulators are approximations of the computer model, some error is introduced by using them. Therefore, emulators are recommended to be used only in the case when the simulator is expensive to evaluate. The error amount introduced by using the emulator in place of the simulator, the so-called code uncertainty, can be estimated since they make probabilistic predictions of the output that the simulator would produce if it was exercised over certain regions of the input space.

Bayesian statistical analysis, through the emulators, can be much more efficient than other methods to quantify uncertainties, e.g. the standard Monte Carlo method, for which computer model simulations are necessary repeatedly [Currin et al., 1991]. The idea of creating a statistical emulator using a Bayesian approach, specifically Gaussian Processes, was proposed in the 1980s. The main aspect of the Bayesian approach is that the emulator is created first and then it is used for further demanding analyses. Analyses such as uncertainty and sensitivity analyses, as well as calibration, require a large number of evaluations of the expensive simulator and this means that they can become impractical. An emulator can be built and used to make such demanding analyses more efficient. Currin et al. [1991], Oakley and O'Hagan [2002] and O'Hagan [2006] focused on a Bayesian approach for performing uncertainty and sensitivity analyses. The uncertainty analysis provides knowledge about the distribution of the simulator output due to uncertain inputs, whereas the sensitivity analysis investigates how much and in which way each of the inputs affect the output. They concluded that Bayesian method is more efficient than the Monte Carlo method as it requires significantly fewer model evaluations. Therefore, an advantage of this method is that the model can be evaluated at higher resolution. Additionally, Kennedy and O'Hagan [2001] presented a Bayesian approach for the calibration of a computer model, where the calibration is the process of fitting a model to the available observations by adjusting its parameters. More recently, Sraj et al. [2014] investigated the uncertainties in the predictions of a tsunami model using a statistical surrogate model in place of the computer model, that was generated using Polynomial Chaos.

The form of the emulator used in the analysis in this thesis is the Gaussian Process (GP). GP is one of the most popular families of stochastic processes and it is used for modelling data that are observed over time or space or both time and space [Rasmussen and Williams, 2006]. A GP can be called an "extension" of the familiar and popular Normal or Gaussian distribution, with nice mathematical properties of the Normal distribution to carry over to the GP. More specifically, one of the most significant properties of GP that makes it popular is the fact that just its mean and covariance functions are required to completely determined it. The GP is the principle tool for creating an emulator to represent our prior knowledge about the simulator. It is worthy to say that “prior knowledge” describes the initial beliefs about the simulator before the available data are used. An unknown function $f(\cdot)$ has a GP distribution if for any set of n inputs $\{x_1, \ldots, x_n\}$, the set of outputs $\{f(x_1), \ldots, f(x_n)\}$ follows a multivariate Normal distribution. The simulator is represented by a GP with mean function $m_0(\cdot)$ and covariance function $V_0(\cdot, \cdot)$, i.e.

$$f(\cdot)|\beta, \sigma^2, B \sim GP(m_0(\cdot), V_0(\cdot, \cdot)) \tag{3.1}$$

where the mean function is described by

$$m_0(x) = h(x)^T \beta, \tag{3.2}$$

in which $h(\cdot)$ is the set of regression functions and β is the vector of the unknown coefficients. The functions $h(\cdot)$ are chosen to represent the main form of the actual simulator $f(\cdot)$. The simplest choice for $h(\cdot)$ is $h(\mathbf{x}) = 1$ resulting in $m_0(\mathbf{x}) = \beta$. This choice indicates no prior knowledge about the outputs response in input’s variations. Another simple choice is $h(\mathbf{x}) = (1, \mathbf{x})$ resulting to $m_0(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$. This choice indicates that there are prior expectations that the simulator’s output will show a linear trend on inputs’ variations. When prior beliefs support that there is a non-linear trend, then quadratic or higher order polynomial terms have to be used [Bastos and O’Hagan, 2009; Bastos, 2010].

The covariance function describes the prior covariance between the output at two different inputs. When these two inputs are the same, the covariance function defines the prior variance of the output, i.e. the prior uncertainty about the simulator output. It is given by

$$V_0(x, x') = \sigma^2 C(x, x'; B) \tag{3.3}$$

where $C(., .; B)$ is a correlation function whose shape is known but with unknown correlation parameters B, also called hyperparameters. A common choice for $C(., .; B)$ is

$$C(x, x'; B) = \exp\{-(x - x')^T B (x - x')\} \tag{3.4}$$

where B is a diagonal matrix of the so-called smoothing parameters b_{ii}. The inverse square roots of these parameters, $1/\sqrt{b}_{ii}$, are known as the correlation length scales. The b_{ii} (or the correlation length scales) describe how rapidly the output responds to changes in each input; the correlation lengths scales give an indication of the distance in the input space for which correlation between the simulator outputs is either significant or negligible.

There are two methods that are commonly used for estimating the hyperparameters B. The first one, which leads to fully Bayesian analysis, known as the hierarchical Bayes, does this in a properly Bayesian manner: it uses the data in combination with prior distributions that are assigned to the hyperparameters. The second approach is to estimate the hyperparameters from the data by either maximizing the marginal likelihood, i.e. the b_{ii} is estimated by determining the most probable values, or by other approaches, such as REML or cross-validation [Santner et al., 2003; Hodges, 2013]. The second approach, known as empirical Bayes, is not fully Bayesian and has the advantage that is computationally cheap and therefore it is very commonly used.

3.1.1 Outer Product Emulator

In the case where the simulator has multiple outputs, the creation of a statistical surrogate model is much more complicated. The simplest approach to this is to build separate independent emulators for each output. However, this method has a major drawback: it ignores the correlations between the outputs. Rougier [2008] proposed an approximate emulator, named the "Outer Product Emulator" (OPE) that creates one emulator for all outputs, simplifying the process by using separable functions for inputs and outputs. This means that no interactions between inputs and outputs are considered in the analysis, i.e. they are assumed to be independent.

The main advantage of the OPE is that it is much more efficient with the building cost to be significantly smaller compared to a general multivariate emulator. The construction and use of an OPE can be fast, even in the case where many simulator evaluations and large number of outputs exist. This property of the OPE is very important for the case investigated in this work, where the wave shape is not oscillating periodically and hence the frequency of the oscillation is not constant. Therefore, it is necessary to run the simulator at small time steps in order to capture the wave behaviour and so a large number of simulator's evaluations is collected to describe the outputs.

The OPE has the form [Rougier et al., 2009]:

$$f_i(r) = \sum_{j=1}^{\nu} \beta_j g_j(r, s_i) + \epsilon(r, s_i) \tag{3.5}$$

where $f_i(r)$ is the i^{th} simulator output at input r, g_j is the set of regressors, β_j are the unknown coefficients and ϵ is the residual. Additionally, s_i represents the output domain - e.g. time and space - corresponding to the i^{th} simulator run.

In order to build an emulator, appropriate prior distributions for β and ϵ must be chosen. A convenient choice is the Normal Inverse Gamma distribution that enables the use of conjugacy (so posterior estimates can be computed explicitly without resorting to Markov Chain Monte Carlo as in more standard fully Bayesian emulators), described by

$$\beta|\tau, B \sim N(m, \tau V) \tag{3.6}$$

$$\epsilon|\tau, B \sim GP(0, \tau\kappa_\lambda(\cdot)) \tag{3.7}$$

$$\tau|B \sim IG(a, d) \tag{3.8}$$

where $B = \{m, V, a, d, \kappa_\lambda(\cdot)\}$ is the set of the hyperparameters, $\kappa_\lambda(\cdot)$ is the covariance function of the residuals with correlation lengths λ and IG denotes the Inverse Gamma distribution. Summing up,

$$\{\beta, \epsilon\} \sim NIG(m, V, a, d) \tag{3.9}$$

where the hyperparameters a and d denote the degrees of freedom and the scale, respectively.

Furthermore, a choice for the appropriate regression functions, $g_j(\cdot)$, and covariance functions of the residual, $\kappa_\lambda(\cdot)$, is required. There are two main characteristics that distinguish the OPE from a standard multivariate emulator. The first is that the covariance function of the residuals is separated in inputs r and outputs s. This property can be represented by the equation

$$\kappa_\lambda(r, s, r', s') = \kappa_\lambda^r(r, r') \times \kappa_\lambda^s(s, s') \tag{3.10}$$

The second characteristic is that the set of the regressor functions, G, is the outer product of the set of regressors for inputs, $G^r \triangleq \{g_j^r(r)\}_{j=1}^{j=\nu_r}$, with the set of regressors for outputs, $G^s \triangleq \{g_j^s(s)\}_{j=1}^{j=\nu_s}$, or in other words, the functions g_j are given by products

$$g_j(r, s_i) = g_j^r(r) \otimes g_j^s(s) \tag{3.11}$$

where $\otimes$ is the outer product symbol and $j = \{1, \ldots, \nu\}$, where $\nu = \nu_r \times \nu_s$.

3.1.1.1 Maximizing the marginal likelihood

To find the most accurate representation of the simulator, appropriate values for the correlation lengths and other parameters can be estimated using the empirical Bayes method of maximizing the marginal likelihood, that is extensively demonstrated by Rasmussen and Williams [2006]. A brief summary of the method is follows.

Consider first the general equation of the emulator, given by

$$\begin{aligned}\mathbf{y} = f(\mathbf{x}) &= h(\mathbf{x}) + \epsilon(\mathbf{x}) && (3.12)\\ &= g(\mathbf{x})^T\beta + \epsilon(\mathbf{x}) \\ &= Q(\mathbf{x})\beta + \epsilon(\mathbf{x}), && (3.13)\end{aligned}$$

where

$$\epsilon \sim GP(0, \tau\kappa_\lambda), \tag{3.14}$$

$$\beta \sim N(0, \tau V), \tag{3.15}$$

assuming that the mean value of the unknown coefficients is zero and also that V can be defined as $V = \sigma^2 I$, with the common multiplier parameter τ to be described by

$$\tau \sim IG(a, d) \tag{3.16}$$

Therefore, the product of the regression functions with the unknown coefficients β, i.e. the function $h(\cdot)$, has a Normal prior distribution given by

$$h|B \sim N(0, \tau QVQ^T) \tag{3.17}$$

The likelihood function is described as follows:

$$y|h, B \sim N(h, \tau\kappa_\lambda) \tag{3.18}$$

The marginal likelihood can be obtained from the integral of the likelihood times the prior, i.e.

$$p(y|B) = \int p(y|h, B)p(h|B)dh \tag{3.19}$$

Hence, the marginal likelihood has a Normal distribution described by

$$y|B \sim N(0, \tau\kappa_\lambda + \tau QVQ^T) \tag{3.20}$$

Consequently, the log marginal likelihood function is given by

$$\Lambda = \log(p(y)) = -\frac{1}{2}f^T C^{-1} f - \frac{1}{2}\log|C| + constant \tag{3.21}$$

where $C = \tau\left(\kappa_\lambda + QVQ^T\right)$. The derivative, with respect to the correlation lengths, of the log marginal likelihood is given by

$$\nabla\Lambda = \frac{1}{2}f^T C^{-1}\frac{\partial C}{\partial\lambda}C^{-1}f - \frac{1}{2}tr(C^{-1}\frac{\partial C}{\partial\lambda}) \tag{3.22}$$

In order to calculate C^{-1}, the Cholesky decomposition is used, where C have to be decomposed into a product of a lower triangular matrix and its transpose, i.e. $C = L^T L$, where L is a lower triangular matrix. Optimization methods can be used to help with the maximization of the marginal likelihood function in order to find correlation lengths.

3.1.1.2 Hyperparameters selection

The final step in the process of building the prior emulator is the selection of hyperparameters $\{m, V, a, d\}$. To determine adequate hyperparameters, the simple approximation method presented by Rougier et al. [2009] is used. The idea is to average the simulator output $f_i(r)$ over the inputs r and output i, which means that $f_i(r)$ is replaced by $f(x^U)$, where x^U is assumed to have a Uniform distribution on the joint space of simulator's inputs and output. Using the mean and variance of the averaged simulator output, $f(x^U)$, the hyperparameters are estimated. The hyperparameter a represents the degrees of freedom or the strength of the prior information, in terms of the equivalent number of evaluations. Completing the selection of hyperparameters yield the prior emulator.

The prior emulator resulting from the prior choices of functions and hyperparameters can be updated using a sample of simulator's evaluations, called the training sample, resulting to the GP emulator, which has a Student-t distribution. The error in emulator's predictions is inversely related to the number of training sample points. Hence, a significantly large number of evaluations can make this error negligible. However, due to the fact that simulators are usually computational demanding, it is unusual to obtain such a large number of evaluations.

3.1.2 Experimental Design

One of the most important steps in every statistical analysis is the experimental design, which is the process of selecting the input points by finding a space filling design that covers the input space sufficiently. Due to the fact that the input points are selected strategically, the amount of useful information passed to the emulator can be maximized. Hence, the required number of simulator runs for an accurate emulator can be reduced, resulting in a more efficient procedure. The use of not space-filling designs may yield to poor predictions in input space areas that are sparsely covered [Santner et al., 2003].

Many different experimental designs exists. The simplest one is the regular grid, where equally space points are selected for each parameter. However, some drawbacks exists by using this method. The most important one is the “collapsing” property of the specific design, where multiple grid points have the same coordinate value when they projected onto a parameter axis. This results to limited information from these points. For example, for a three-dimensional input space, in order to obtain n distinct evaluations for each of the three parameters, the total number of required simulator evaluations is n^3, which is obviously highly inefficient.

A more advanced experimental design for choosing simulator inputs have been proposed by McKay et al. [1979], commonly referred to as the Latin Hypercube Design (LHD). This design was constructed to avoid the “collapsing” property of the grids. The LHD selects n different sample points from each k variables $X_1, \ldots, X_k$ using the following process. Firstly, the range of each variable is divided into n equal probability and non-overlapping intervals. Then, one value from each interval is randomly selected with respect to the probability density of the interval. The n values selected for X_1 are paired randomly with the n values for X_2. These n resulting pairs are then combined randomly with the n values of X_3 giving n triplets. The same process continues until the n k-tuplets are formed, which is the LH sample.

Urban and Fricker [2010] made a comparison of the Latin Hypercube with the regular grid design for the multivariate emulation. They concluded that the emulators generated using the LHD make significantly improved predictions relative to an emulator created using a regular grid training sample. Furthermore, they have shown that the LH emulator is more accurate compared to the regular grid emulator in sensitivity analysis of a single-parameter model.

However, only a subset of LH designs are space filling. To ensure a space filling input selection, a very popular design within the class of LH designs is commonly used, named the “maximin” LHD. The specific design follows exactly the same process as the LHD to choose the sample points, although it has an additional constraint that is to maximize the minimum distance between points. Therefore, a maximum coverage of the input space is achieved by using the specific experimental design.

Additionally, in a randomly generated LHD the variables may be highly correlated. Joseph and Hung [2008] proposed a method to find LHDs that combine both minimum pairwise correlations between the variables and also ensure maximum space coverage, by maximizing the inter-site distances. However, by minimizing the correlation the points spreads out, whereas by maximizing the distance between the points the correlation reduces. To achieve both minimum correlations and maximum space coverage an optimization approach was recommended by the authors.

3.1.3 Validation [1]

A GP emulator sometimes gives poor prediction of the simulator output. The reasons for this can be two; either the assumption of stationary GP distribution with the selected mean and covariance is incorrect or a poor experimental design is used. For the case of tsunami modelling where human lifes can be affected, a non-representative emulator is obviously undesirable. Hence, a validation process must be performed before using an emulator as a surrogate of a simulator, in order to check that the assumptions used to build the emulator are reasonable. However, always the validation of an emulator must be performed using cross-validation methods, or new data, since the GP emulator is a deterministic function. Therefore, for the validation, emulator's predictions and simulator's evaluations are compared for a new data set.

Even the fact that GP emulators are extensively used to represent simulators and also that the validation of the simulators is extensively investigated, the existing research on emulators validation is limited. Bastos and O'Hagan [2009] and Bastos [2010] presented various diagnostic tools, both numerical and graphical, for validating a GP emulator, i.e. for qualify how accurate a GP emulator predicts simulator's outputs. These diagnostics concentrate on comparison between simulator's output and GP emulator's outputs for some test data, known as validation data. The validation input points have not been used for building the emulator. Since the emulator gives perfect predictions only at the points that have been used for its generation, the difference between the emulator's predictions and the actual simulator's outputs at these inputs is not zero, but instead can be used for validating emulator's capability.

First of all, Bastos and O'Hagan [2009] and Bastos [2010] have defined the individual prediction errors to be the differences between the observed simulator's outputs and the predicted mean output at the same validation input points. For a set of inputs $\mathbf{X}^* = (x_1^*, \ldots, x_n^*)$ and output given by $\mathbf{y}^* = f(\mathbf{X}^*)$, where $\mathbf{y}^* = (y_1^*, \ldots, y_m^*)$, the prediction errors are given by

$$y_i^* - E[f(x_i^*|\mathbf{y}^*)] \tag{3.23}$$

for $i = 1, 2, \ldots, n$. The first proposed diagnostics are the standardized prediction errors given by

$$\frac{y_i^* - E[f(x_i^*|\mathbf{y}^*)]}{\sqrt{V[f(x_i^*|\mathbf{y}^*)]}} \tag{3.24}$$

The standardized prediction errors can be considered to follow a standardized Normal distribution, in the case of a relatively large training data set and hence a large number of degrees of

---

[1] The term "Validation" in this chapter is used for the process of assessing the emulator's predictions and not for checking the mathematical model, as it was the case for Chapter 2

freedom. The authors pointed out that individual large errors give an indication of disagreement between simulator's and emulator's evaluations. In general, the errors described by Equation (3.23) are useful in cases where the emulator is used as point predictor of the simulator, whereas the standardized errors, described by Equation (3.24), are more relevant in cases where the uncertainty resulted by using the emulator is considered in addition to any other uncertainty in the process. Bastos and O'Hagan [2009] and Bastos [2010] have also proposed graphical methods, which are very powerful and efficient tools for emulator's validation and also for testing the assumptions used for the emulator's generation, such as plot of the individual errors agains the emulator's predictions and Quantile-Quantile plots for testing the normality assumption.

Rougier et al. [2009] demonstrated a cross-validation method called the "leave-one-out" diagnostic. The process for applying this method is to leave out one input point and the corresponding model's evaluations at this point and build the emulator with the remaining input points. Following that, the model's output at the leave-out input is predicted using the resulting emulator. Then, since the actual simulator's output is know at the particular input point, it can be compared with the emulator's prediction. This process have to be repeated for all the input points, resulting to an efficient quantification of the emulator's capabilities to represent the simulator. This validation method is particularly useful for cases that no real evaluations are available to compare with emulator's predictions, which this is the case in the analysis in this chapter. Therefore, after building the emulator, the "leave-one-out" diagnostic method is going to be applied for testing how accurately the emulator represents the simulator.

3.2 Application of statistical emulation on a landslide-generated tsunami model

In this section, the methods described earlier are applied to find an accurate statistical representation of the analytical landslide-generated tsunami model of Sammarco and Renzi [2008], abbreviated as the SR model from now on.

3.2.1 Landslide generated tsunami model

The SR landslide-generated tsunami model takes as inputs the initial position x_0, the initial speed u_0 and also the spread ratio or shape c of the landslide, where the "spread ratio" is defined as the ratio of the landslide's characteristic length over the characteristic width. Figure 3.1 illustrates this specific analytical model set up.

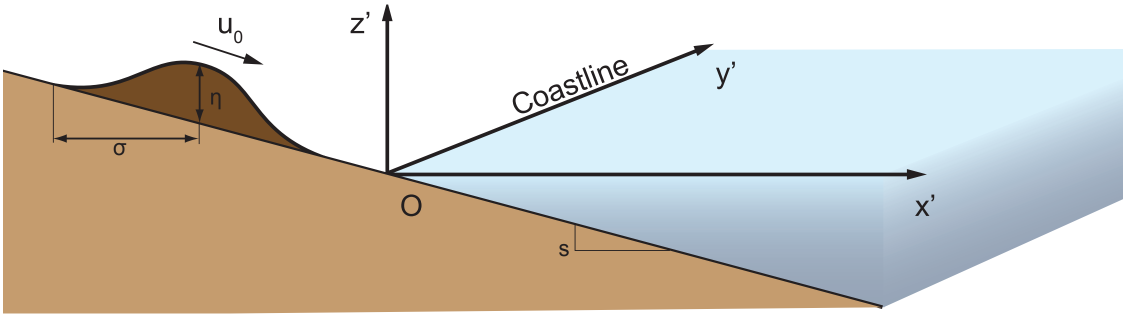


Figure 3.1: Sketch illustrating the landslide's motion as considered in Sammarco and Renzi's analytical model. The y-axis represents the shoreline, while the x-axis is perpendicular to it.

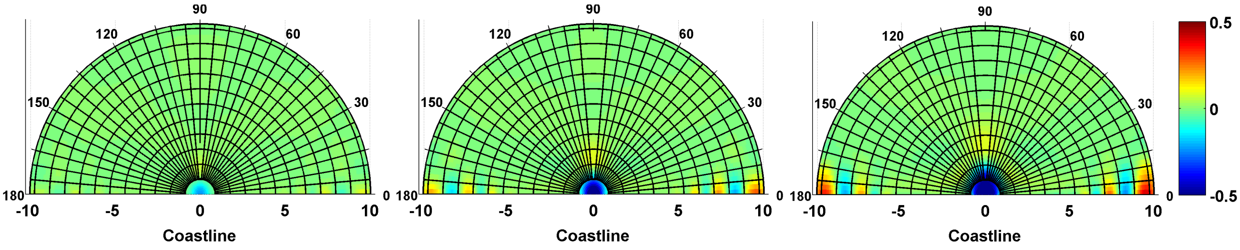
All the coordinates, functions and parameters used in the model are non-dimensional:

$$\begin{aligned} &x = \frac{x'}{\sigma}, \quad y = \frac{y'}{\sigma}, \quad t = \sqrt{\frac{gs}{\sigma}}t', \quad \zeta = \frac{\zeta'}{\eta}, \\ &u_0 = \frac{1}{\sqrt{\sigma g s}}u_0', \quad c = \frac{\sigma}{\lambda} \end{aligned} \tag{3.25}$$

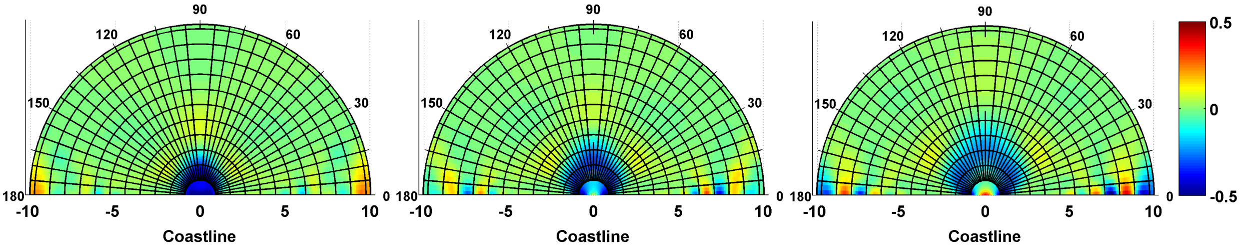
where the primes denote dimensional values, σ is the landslide's characteristic horizontal length, s is the beach slope, η denotes the landslide's maximum vertical thickness, ζ is the the non-dimensional free-surface elevation, λ is the landslide's characteristic width, t is the time and g is the acceleration due to gravity.

When the landslide starts its motion from the origin, which is the position where the water surface meets the sloping beach, x_0 is equal to zero. Also, negative values of x_0 indicate that the landslide initiates from a subaerial position, whereas positive values of x_0 indicate initially submerged slides. The output of the SR model is the free-surface elevation of the water wave at given times and locations. A plane beach with constant slope is considered and it is assumed that the landslide continues the motion even after it falls into the water. This causes the existence of high wave elevations even at large times. By considering this model, Sammarco and Renzi [2008] came to the conclusion that the landslide generates a wave field that is composed by two components, oscillatory and evanescent.

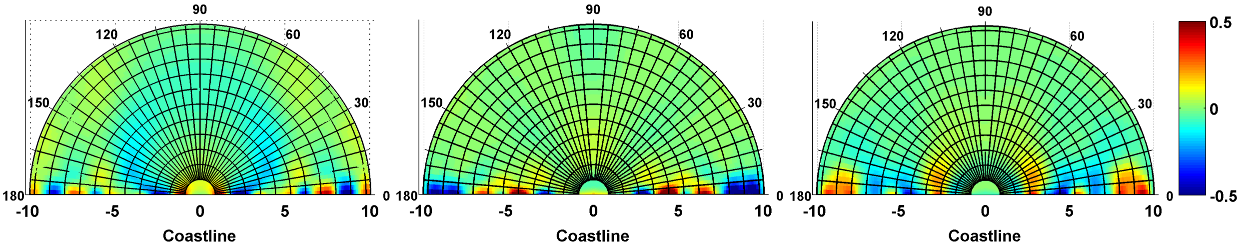
The life cycle of the wave can be visualized in Fig. 3.2, where the free-surface elevation of the landslide-generated tsunami wave is shown in polar coordinates at times $t = 0.5, 1, 1.5, 2, 2.5, 3, 5, 10, 20$. At the specific case the initial position of the landslide is considered to be the origin, the initial speed is taken equal to 1 and the spread ratio of the landslide is assumed equal to 2, which means that the characteristic length is twice the size of the characteristic width. When the landslide occurs, it displaces water forward and an elevation wave is generated, that propagates mostly in the offshore direction. Also a depression wave occurs



(a) t=0.5 (b) t=1 (c) t=1.5



(d) t=2 (e) t=2.5 (f) t=3



(g) t=5 (h) t=10 (i) t=20

Figure 3.2: Free-surface elevation of the landslide-generated tsunami waves observed at different times with non-dimensional inputs $(x_0, u_0, c) = (0, 1, 2)$. The horizontal axis represents the shoreline and the vertical axis points to the offshore direction.

near the origin [Fig. 4.6a]. Later on, the elevation wave spreads along the shoreline, while the depression wave extends around the origin [Fig. 4.6b, 4.6c, 4.6d]. At larger times, a second elevation wave is generated at the origin and the depression wave spreads out [Fig. 3.2f, 3.2g]. Finally, at even larger times, the wave motion is dominated by edge waves propagating along the shoreline, with no motion around the origin [Fig. 3.2h, 3.2i]. From this study, it is concluded that the first generated waves are not those with the larger amplitude. This indicates that in order to capture the maximum elevation, the model has to be evaluated up to a significantly large time t.

3.2.2 Training Sample

A statistical emulator has been constructed looking at specific locations, i.e. its output is only time-depended. Specifically, seven locations along the shoreline ($x_l = 0$) at $y_l = 2, 4, 6, 7, 8, 8.38$ and 10 have been investigated. The time domain is selected to be between 0 and 35. Small time steps are required in order to have sufficient information to capture the wave shape with adequate detail: specifically $dt = 0.2$ was chosen for the analysis.

The first step of the analysis is the experimental design. Using the “maximin” Latin Hypercube design method, as detailed in Section 3.1.2, forty points, (x_0, u_0, c), are chosen to cover the three-dimensional input parameter space. This is a compromise in order to have a significantly good coverage of the design space as well as a small computation cost. The input domain is chosen to be the following: $x_0 \in [-3, 1]$, $u_0 \in [1, 2]$ and $c \in [0.5, 3]$. The positions of the forty inputs in the parameter space are shown in Fig. 3.3. The colour at each point indicates the maximum free-surface elevation, for the specific location $x_l = 0$ and $y_l = 8.38$, i.e. along the shoreline and far away from the source.

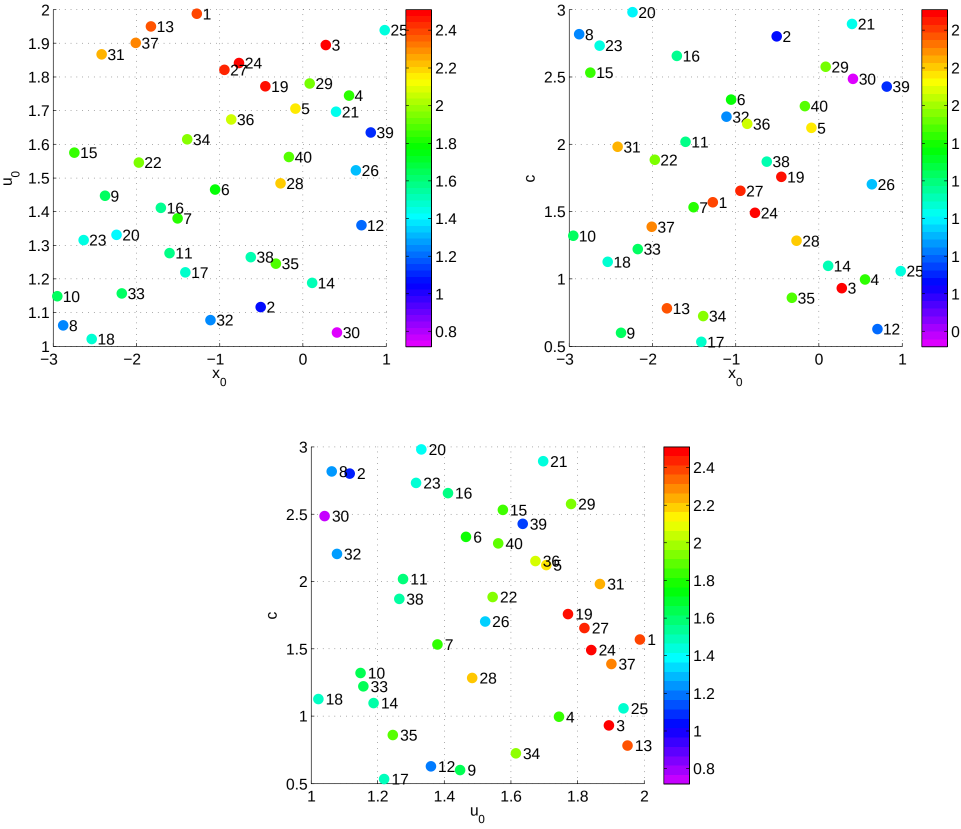


Figure 3.3: Maximum free-surface elevation at the location $(x_l, y_l) = (0, 8.38)$ for time between 0 and 35 for each of the 40 design input points selected using the “maximin” LHD method. Three quantities are varied: the initial landslide’s speed, its initial location and its shape, that are given in non-dimensional form as in Eq. 3.25.

The Figure 3.3 shows that the maximum wave elevation significantly depends on the initial speed of the landslide: the larger the initial speed u_0, the larger the maximum elevation. Furthermore, it can be observed that the maximum wave elevation shows higher amplitudes when the landslide starts from a subaerial close to the origin position and also when the landslide’s spread ratio is less than 2. However, the dependence of the maximum elevation on the initial position and spread ratio of the landslide is not as obvious as that on the initial speed. For example consider points 13 and 25. They both represent landslides starting with high initial speed and spread ratio close to one. However point 13 is a subaerial case while point 25 is a submerged one. This yields a significant difference in the maximum free-surface elevation, with the subaerial case being much higher. The simulator's evaluations for the other six locations along the shore, yields similar conclusions about the dependency of the maximum free-surface elevation to the input parameters.

3.2.3 OPE prior choices

The next step in the analysis involves the appropriate prior choices for the regression and residuals covariance functions for inputs r and outputs s. In the case of the SR model, r is equal to (x_0, u_0, c) and s is time t. The set of input regression functions, $G^r \triangleq \{g_1^r, \ldots, g_{\nu_r}^r\}$, where ν_r is the number of input regressors, consists of a linear combination of appropriate choices of polynomials for each of the three input parameters. Since the simulator's output variation with respect to r is smooth, combinations of a constant, a linear and a quadratic polynomial are used to represent each input's regressors. These polynomials are shifted into the unit interval $[0, 1]$ and their coefficients are selected so that the two functions for each input parameter are orthonormal with a uniform weighting function. Combining all the inputs' functions, the set of chosen input regressors is the following:

$$\begin{aligned} G^r &= \{1, \sqrt{3}\frac{(x_0+3)}{4}, -3\sqrt{5}\frac{(x_0+3)}{4} + 4\sqrt{5}(\frac{x_0+3}{4})^2, \\ &\quad \sqrt{3}(u_0-1), -3\sqrt{5}(u_0-1) + 4\sqrt{5}(u_0-1)^2, \\ &\quad \sqrt{3}\frac{(c-0.5)}{2.5}, -3\sqrt{5}\frac{(c-0.5)}{2.5} + 4\sqrt{5}(\frac{c-0.5}{2.5})^2\} \end{aligned} \tag{3.26}$$

After choosing the regression functions for the inputs, an appropriate choice for the regression functions for the output, $G^s \triangleq \{g_1^s, \ldots, g_{\nu_s}^s\}$, is required, where ν_s is the number of output regressors. Fourier terms are chosen of the form $\sin(\frac{2\pi t}{T})$ and $\cos(\frac{2\pi t}{T})$, where T is the period of the oscillation. However, since the free-surface elevation waves do not oscillate with constant period, this selection is challenging. To make this selection, the range of oscillating frequencies present in the wave is considered and using a trial and error method, it is concluded to the most representative set of frequencies, which is $\{\frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$. Therefore, the selected set of output regression functions is given by

$$\begin{aligned} G^s &= \{1, \sin(\pi t/3), \cos(\pi t/3), \sin(2\pi t/5), \cos(2\pi t/5), \\ &\quad \sin(\pi t/2), \cos(\pi t/2), \sin(2\pi t/3), \cos(2\pi t/3), \\ &\quad \sin(\pi t), \cos(\pi t)\} \end{aligned} \tag{3.27}$$

Power exponential functions are chosen for input and output residuals covariance functions, κ^r and κ^s [Stein, 1999]:

$$\begin{aligned} \kappa^r &= \exp\left(-\left(\frac{|x_0 - x_0'|}{\lambda_x}\right)^{3/2}\right) \times \exp\left(-\left(\frac{|u_0 - u_0'|}{\lambda_u}\right)^{3/2}\right) \\ &\quad \times \exp\left(-\left(\frac{|c - c'|}{\lambda_c}\right)^{3/2}\right) \end{aligned} \tag{3.28}$$

and

$$\kappa^s = \exp\left(-\left(\frac{|t_1 - t_2|}{\lambda_t}\right)^{3/2}\right) \tag{3.29}$$

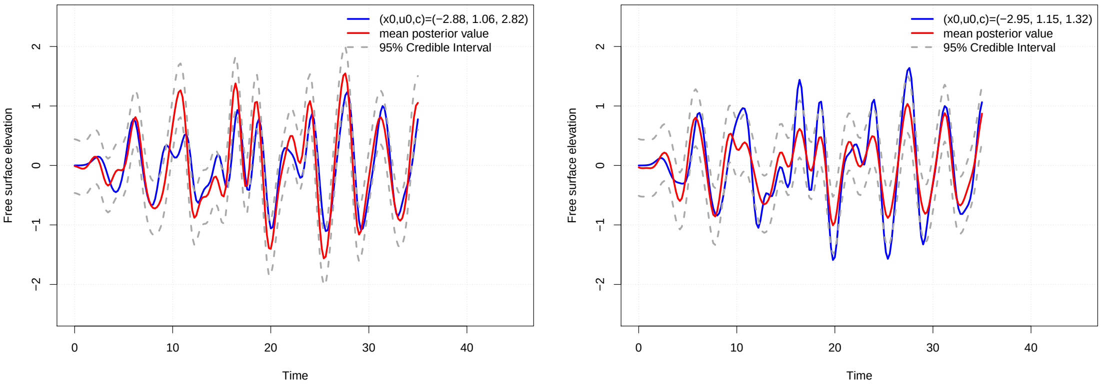
respectively, where λ_x, λ_u, λ_c represent the correlation lengths for inputs and λ_t denotes the output (i.e. time) correlation length. The values of the correlation lengths can be varied in order to adjust the fit of the emulator. The correlation lengths are chosen by maximizing the marginal likelihood. This process also yields an appropriate value for the common coefficient τ. Note that the $3/2$ exponent is chosen so that the covariance is smooth enough, but not too much as the usual choice of square power is infinitely smooth and hence may not be realistic for such a complex simulator [Rasmussen and Williams, 2006; Rougier et al., 2009].

The last step for the creation of the prior emulator for the SR model is to make a choice for the values of the hyperparameters $\{m, V, a, d\}$. To do so the method described by Rougier et al. [2009] is followed. It has been already assumed that $m = 0$. The hyperparameter a, which is equal to the number of degrees of freedom, takes the value 3 in the case of the SR model. Also, after the simple calculations recommended by Rougier et al. [2009], it is concluded that $\sigma^2 = 0.257$ and $d = 0.058$. Hence, V can be easily obtained from $V = \sigma^2 I$.

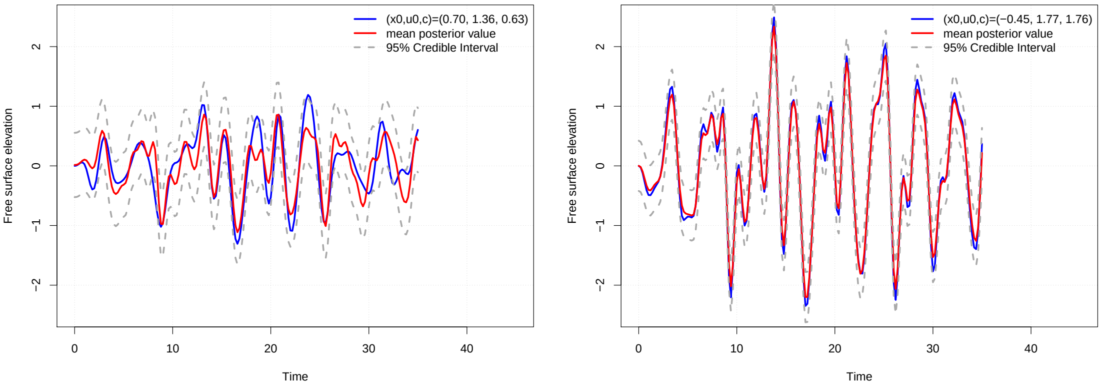
By fixing these parameters, the creation of the prior emulator is completed. Using the evaluations of the 40 selected design points, the prior emulator is updated to obtain the posterior, which is the statistical emulator. Evaluating the statistical emulator at a given input point, (x_0, u_0, c), results in predictions of the output's distribution for all the points in the time domain, in this case from 0 to 35, every 0.2 time step, i.e. 176 prediction distributions.

3.2.4 Emulator's validation

After the generation of the emulator, the LOO validation method is applied, resulting in 40 LOO diagnostic plots. These diagnostics give information about the predictive power, capabilities and shortcomings of the emulator, since the amount of the error induced by using the emulator instead of the simulator can be estimated. Some of the diagnostic plots for the location $(x_l, y_l) = (0, 8.38)$ are shown in Fig. 3.4. Similar diagnostic plots are created for all the other locations investigated. In general, the LOO diagnostics allow us to conclude that in most of the cases the emulator predicts very well the simulator evaluations, capturing both the shape and the maximum wave elevations (peaks). Additionally almost always the simulator's evaluation line is within the 95% prediction credible interval (ideally it should be within this interval 95% of the time).



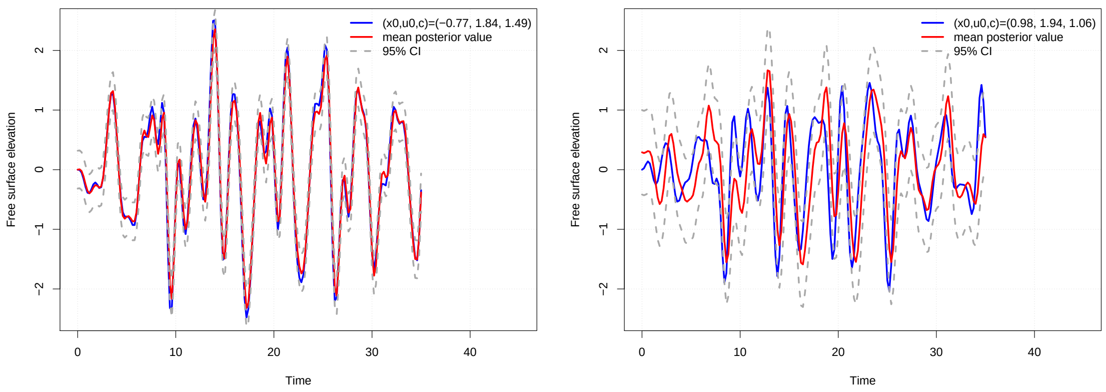
(a) (x0,u0,c)=(-2.88,1.06,2.82) – point no. 8 (b) (x0,u0,c)=(-2.96,1.15,1.32) – point no. 10



(c) (x0,u0,c)=(0.70,1.36,0.63) – point no. 12 (d) (x0,u0,c)=(-0.45,1.77,1.76) – point no. 19

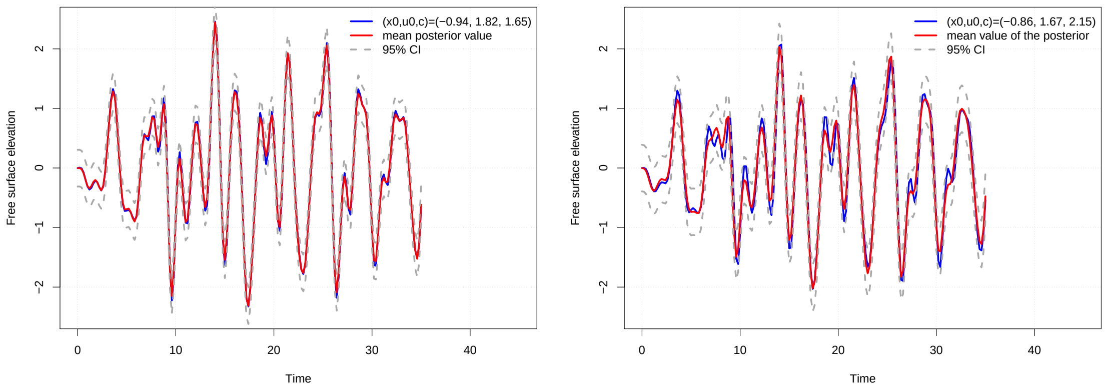
Figure 3.4: Diagnostic plots for some of the input points looking at $(x_l, y_l) = (0, 8.38)$. Blue line is the simulator's evaluation, red is the mean value of the posterior distribution and dotted grey is the 95% credible interval of the posterior distribution.

However, on some of the diagnostic plots, the prediction is not very accurate. One of the



(e) (x0,u0,c)=(-0.77,1.84,1.49) – point no. 24

(f) (x0,u0,c)=(0.98,1.94,1.06) – point no. 25



(g) (x0,u0,c)=(-0.94,1.82,1.65) – point no. 27

(h) (x0,u0,c)=(-0.86,1.67,2.15) – point no. 36

Figure 3.4: Diagnostic plots for some of the input points looking at $(x_l, y_l) = (0, 8.38)$. Blue line is the simulator's evaluation, red is the mean value of the posterior distribution and dotted grey is the 95% credible interval of the posterior distribution.

fundamental reasons affecting the emulator performance is the position of the point at which is tried to be predicted in the input space. Generally, it is expected to obtain more accurate predictions in the cases where the points at which are tried to be predicted are surrounded closely by other design points, compared to the cases where the points are located in a sparsely covered region, since more information can be obtained by the neighbouring points. The behaviour at each point is significantly linked to the behaviour at the points close to it and this influence decays rapidly with the distance separating the two points.

To quantify this, the Euclidean distances in the three-dimensional input space between a point and the other 39 points are obtained. Then the mean value of these distances (MED) for each of the 40 input points is calculated:

$$MED = \frac{\sum_{i=1}^{39} \sqrt{(x_1 - x_2)^2 + (u_1 - u_2)^2 + (c_1 - c_2)^2}}{39} \tag{3.30}$$

Figure 3.5 displays the mean Euclidean distances for all the design input points. It is noticeable that the points 8, 10, 12 and 25 show a large MED from the rest of the 39 points. Looking at the LOO diagnostics of these four points in Fig. 3.4a, 3.4b, 3.4c, 3.4f, can be easily observed that the predictions are not very accurate. However, the prediction of the maximum wave elevation, which is the most important measurement, is still satisfactory and almost everywhere the simulator evaluation lines are within the 95% credible intervals. This indicates that, even for the design points that are isolated from the neighbouring points, the emulator predictions are still usable.

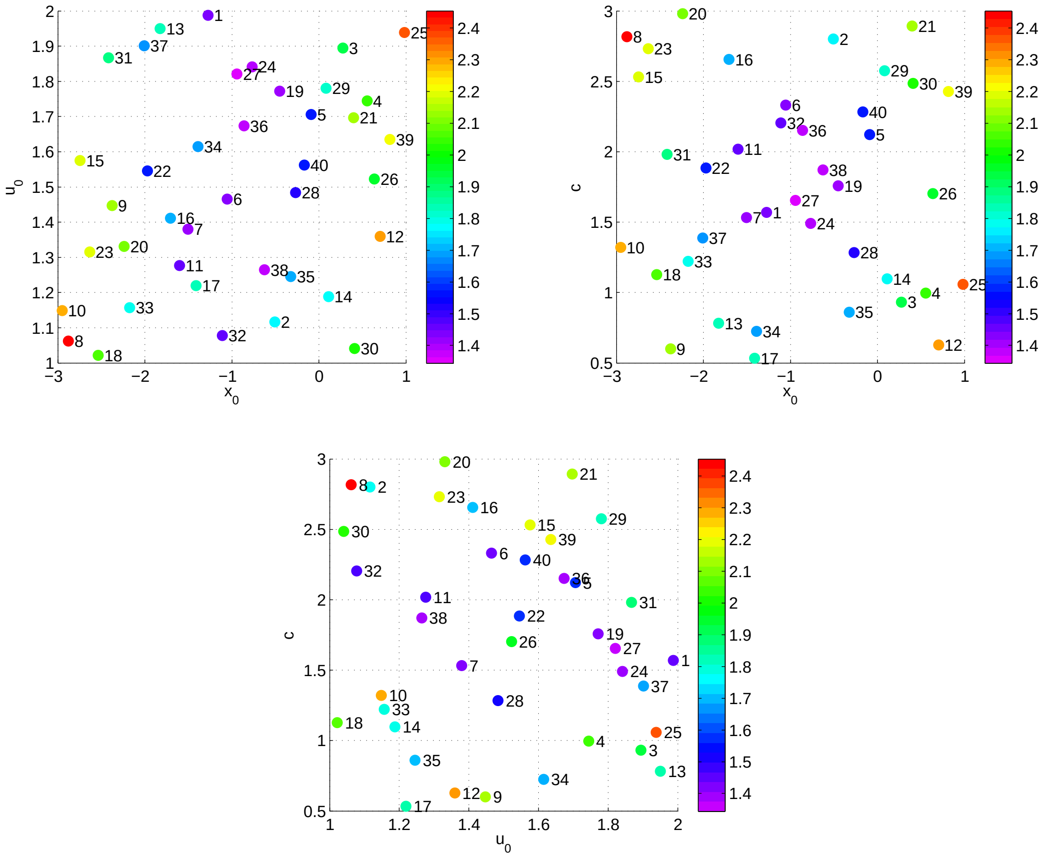


Figure 3.5: Euclidean distances between each of the points and the other 39 points in the three-dimensional parameter space.

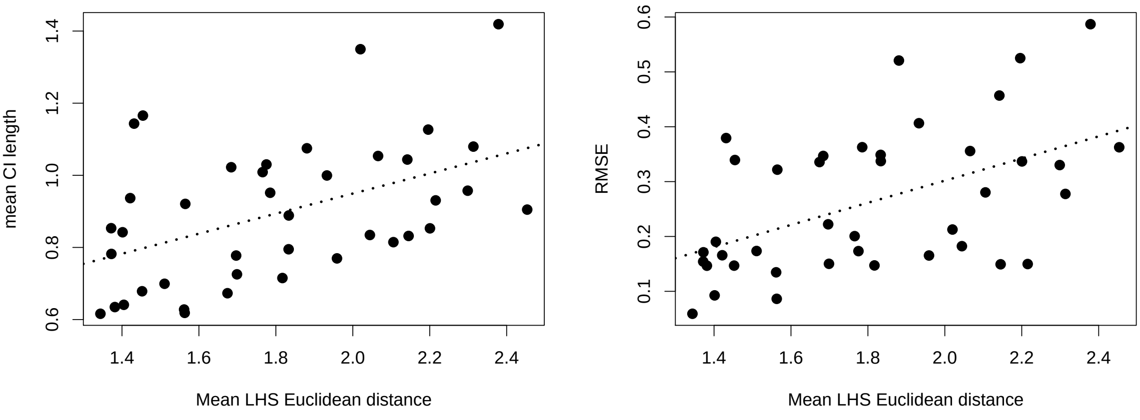
On the other hand, points such as 19, 24, 27 and 36 are affected significantly by the other points, separated by small Euclidean distances from the rest of the 39 points in space. Looking at the diagnostic plots of these points (Fig. 3.4d, 3.4e, 3.4g, 3.4h), it is obvious that the emulator does an excellent job in prediction, since all the features of the wave are predicted accurately by the emulator.

Two measures that can be used to quantify the emulator's accuracy are the mean credible interval length (MCIL) and the root mean square error (RMSE) between the observed and the predicted evaluations. For each of the 40 LOO diagnostics, individual RMSEs and MCILs can be calculated. The RMSE is given by the equation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{3.31}$$

where y_i and $\hat{y}_i$ are the observed and predicted values at each time step i, respectively, and n is the total number of time steps.

Figure 3.6 displays both the MCIL and the RMSE versus MED for all the input points, looking at the case of the location $(x_l, y_l) = (0, 8.38)$. A positive correlation between the MED and both the MCIL and the RMSE can be observed. This confirms that the distance separating the points in space is a fundamental factor that affects the predictive power of the emulator and hence this highlights the importance of a good experimental design. A non-optimal experimental design can result in both larger confidence intervals as well as largest distance between the emulator's prediction and the simulator's output. This positive correlation is also satisfied for the other locations examined.



(a) (b)

Figure 3.6: Mean Euclidean distance vs. (a) mean 95% credible interval length and (b) RMSE for the location $(x_l, y_l) = (0, 8.38)$, where the dotted line is the linear regression.

In Fig. 3.7, the RMSE with respect to MCIL is presented for all the 40 diagnostics for the seven locations along the shoreline investigated, in order to compare the emulator's performance when applied to different locations. A combination of both small RMSE and MCIL is desirable, indicating both small error and small uncertainty in emulator's predictions. The figure clearly shows that the emulator performs similarly for all the locations investigated. Therefore, the emulator can be applied to different locations along the shoreline, resulting in accurate enough

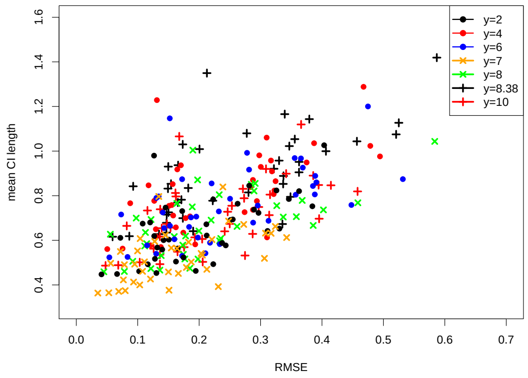


Figure 3.7: Root Mean Square Error vs. mean CI length. Different types and colors represent different locations along the shoreline.

representations of the simulator output. The reasons that there are slightly better predictions at some locations compared to others is an area of further investigation. Nevertheless, the location along the shoreline with $y_l = 8.38$ shows the worst results in Fig. 3.7. Therefore, the predictions of the emulator for the other locations are better than the ones given in Fig. 3.4. This reinforces the confidence of using the statistical emulator.

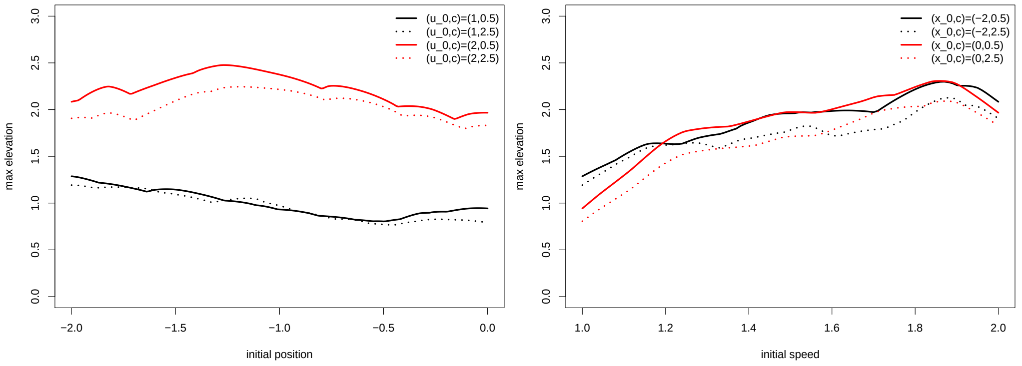
3.3 Sensitivity and Uncertainty Analyses of the SR model

The process of generating a statistical emulator that can predict the simulator's output with sufficient accuracy is presented for different locations along the shoreline in this chapter. The emulator can be used further in place of the expensive-to-run simulator to efficiently perform analyses that require a large number of evaluations, in order to save time without sacrificing accuracy. Sensitivity and uncertainty analyses using the resulting emulator are performed for the SR model and they are demonstrated in this section.

3.3.1 Sensitivity Analysis

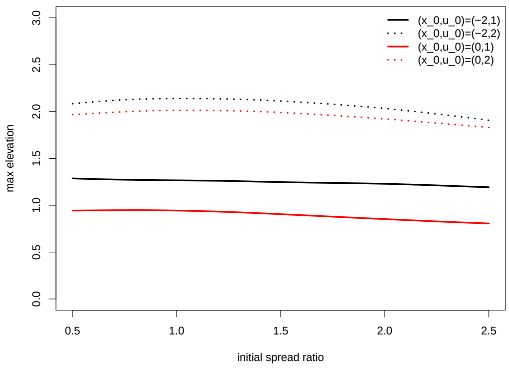
The statistical emulator is used to carry out a sensitivity analysis of the model, where the sensitivity of the maximum wave elevation for $t \leq 35$ to changes in inputs is investigated. Additionally, it is examined whether the different individual locations along the shoreline present consistent sensitivity to inputs' variation.

Fig. 3.8 displays the results from the analysis for the location $(x_l, y_l) = (0, 8.38)$. In each of the three plots, the maximum elevation is plotted against the initial position x_0, initial speed u_0 and initial spread ratio c of the landslide, respectively, with the other two input parameters being kept constant. To ensure maximum emulator's accuracy and keep RMSE to the minimum, the input domain in sensitivity analysis is chosen to be the subset of the whole domain where the mean Euclidean distance between the points are small as presented in Fig. 3.5. Specifically, the analysis is concentrated on the input domain $x_0 \in [-2, 0]$, $u_0 \in [1, 2]$ and $c \in [0.5, 2.5]$.



(a) max. elevation w.r.t. initial position for two different initial speeds and spread ratios

(b) max. elevation w.r.t. initial speed for two different initial positions and spread ratios



(c) max. elevation w.r.t. initial spread ratio for two different initial positions and speeds

Figure 3.8: Maximum free-surface elevation with respect to (a) initial position, (b) initial speed and (c) initial shape, for the time interval $[0, 35]$ and position $(x_l, y_l) = (0, 8.38)$.

From Fig. 3.8a an obvious relationship between the landslide's initial speed and the maximum elevation can be observed. Specifically, a landslide with a larger u_0 gives larger maximum free-surface elevations. No strong dependency of the maximum elevation on initial position and spread ratio can be observed. Figure 3.8b highlights the positive relationship between u_0 and the maximum elevation, with the larger the u_0, the larger the maximum elevation. Finally, Fig. 3.8c shows that a landslide initiating from a subaerial position shows larger maximum free-surface elevations compared to a landslide starting from the origin. So, a relationship between the x_0 value and the maximum elevation is indicated. Also, a landslide starting with a larger speed yields larger maximum elevations. Moreover, it cannot be concluded that the spread ratio is a significant factor at the specific range investigated. The same conclusions result by repeating the sensitivity analysis for the other six locations. Similar analyses in which the output is another important aspect of the tsunami, different from the maximum elevation, can be easily performed.

A comparison of how sensitive is the maximum wave elevation at different locations to changes in the input parameters is showed in Fig. 3.9, 3.10 and 3.11. Each of the figures illustrate the change in maximum free-surface elevation with respect to variations in one of the input parameters, keeping the other two constant. Four different combinations of the constant parameters are investigated. It can be concluded that the sensitivity of maximum elevation is very similar for all the investigated locations along the shoreline.

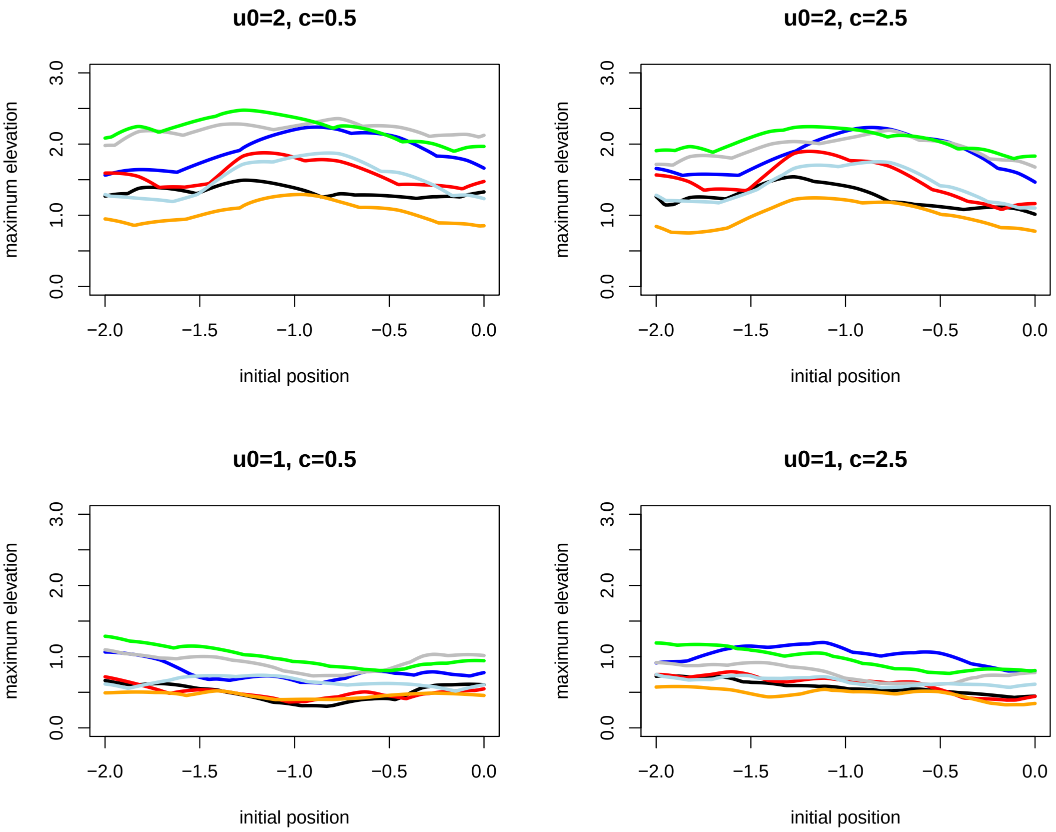


Figure 3.9: Maximum free-surface elevation with respect to the initial position x_0 for different combinations of u_0 and c. The different colours correspond to the different locations along the shoreline: black (y=2), blue (y=4), red (y=6), orange (y=7), grey (y=8), green (y=8.38), light blue (y=10).

Overall, the conclusions reached by using the emulator are the same as those obtained by using the simulator as shown in Fig. 3.3. However, the emulator has the fundamental advantage that it is much faster compared to the simulator. Therefore, it can be evaluated at a much larger number of inputs, leading to higher resolution and smoother plots. For example, the plots in Figure 3.8 have required a large number of emulator's evaluations, specifically 2012. Importantly, the required emulator's running time was very short. A total time for this entire analysis for a specific location was around 186.6 seconds on a Dual Core 3.06GHz computer.

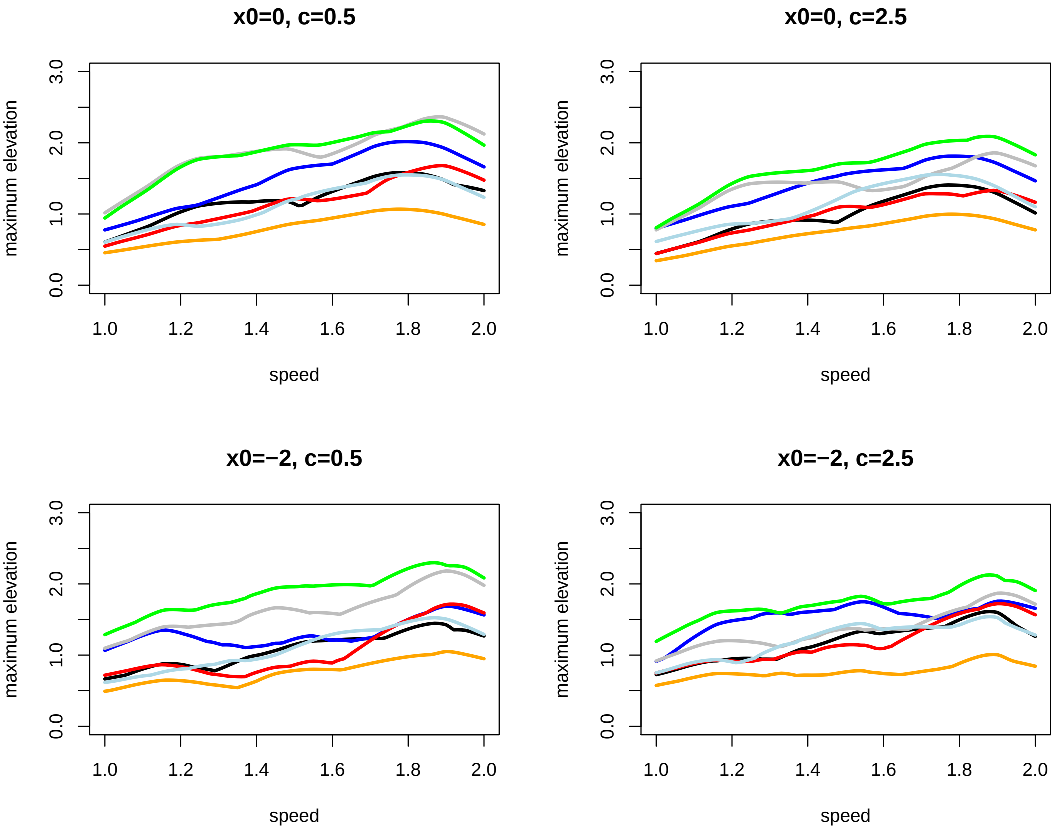


Figure 3.10: Maximum free-surface elevation with respect to the speed u_0 for different combinations of x_0 and c. The different colours correspond to the different locations along the shoreline: black (y=2), blue (y=4), red (y=6), orange (y=7), grey (y=8), green (y=8.38), light blue (y=10).

Using a simulator to perform the same analysis would take much longer, as a single run to reconstruct the free-surface elevation time series up to time 35 with the SR analytical model takes about 30 minutes.

3.3.2 Uncertainty Analysis

Usually the largest amount of uncertainty induced in simulator's evaluations comes from the high uncertainty of tsunami trigger features. It is impossible to know exactly the initial position, speed and spread ratio of the landslide that cause the tsunami. Since, as it is shown earlier, the emulator can provide accurate enough predictions of the simulator's outputs, an uncertainty analysis is performed by employing the emulator in the place of the simulator. The uncertainty analysis will give the amount of uncertainty in the predictions that is due to the uncertain inputs, as well as from the use of emulator in place of the simulator. Usually experts have some knowledge about the most likely distribution of the inputs. Using these distributions, one can draw a number of random input samples, that can be given to the emulator in order to estimate

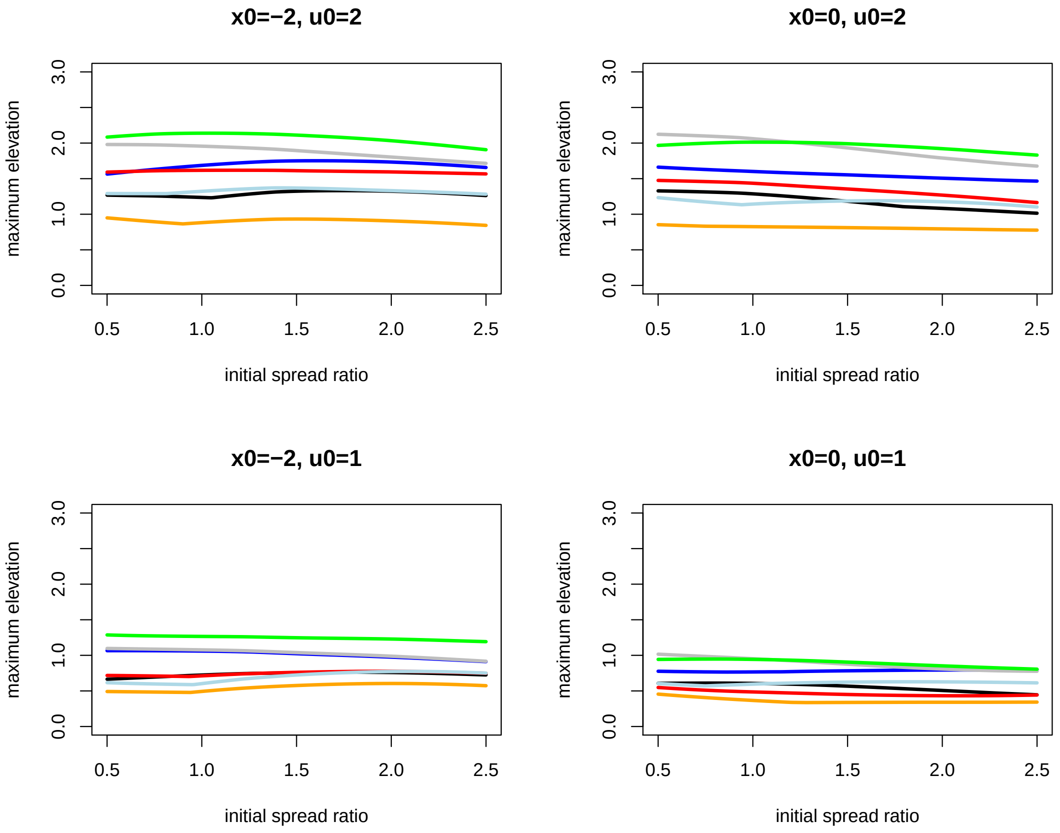


Figure 3.11: Maximum free-surface elevation with respect to the spread ratio c for different combinations of x_0 and u_0. The different colours correspond to the different locations along the shoreline: black (y=2), blue (y=4), red (y=6), orange (y=7), grey (y=8), green (y=8.38), light blue (y=10).

the posterior distribution of key tsunamis features, e.g. maximum elevation.

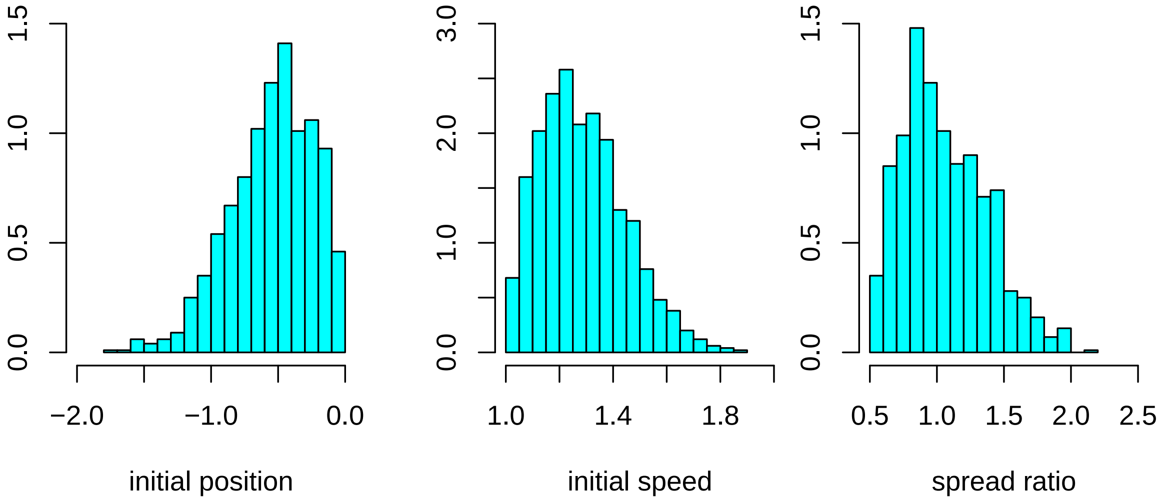
It is assumed that some collection of emergency management experts come to the conclusion that the inputs follow a beta distribution with some skewness and that the input domain is the same as with the sensitivity analysis. The beta distribution is a flexible distribution over a finite interval that can enable experts to express their believes. The distributions of input parameters are given by

$$x_0 \sim Be(5,2) \quad \text{for} \quad x_0 \in [-2,0] \tag{3.32}$$

$$u_0 \sim Be(2,5) \quad \text{for} \quad u_0 \in [1,2] \tag{3.33}$$

$$c \sim Be(2,5) \quad \text{for} \quad c \in [0.5,2.5] \tag{3.34}$$

Therefore, the initial position of the landslide follows a distribution which indicates that a starting position near the origin is more likely. Additionally, both the initial speed and spread ratio distributions are skewed to the left, in order to highlight initial speeds most likely close to one and initial characteristic length and width of the landslide to be most likely of similar dimensions.



(a) (b) (c)

Figure 3.12: Histograms showing the prior expectations about the distribution of input points.

For this analysis, one thousand random samples is drawn for the inputs from the distributions given in (3.32), (3.33), (3.34), resulting in the prior input distributions shown as histograms in Fig. 3.12. The emulator is evaluated using the selected inputs. Therefore, one thousand predictions for the wave elevation at a fixed position along the shoreline for times up to 35 at 0.2 time intervals are obtained. From each of these time series, the maximum elevation and the mean CI length are estimated, resulting in one thousand estimates for each one. The variation among the thousand values are quantified using quantiles. The same process is repeated for all the examined locations along the shoreline. The quantiles for the case of $(x_l, y_l) = (0, 8.38)$ are summarized in Table 3.1. The posterior distribution of the maximum elevation is plotted in Fig. 3.13. This information summarizes the expected tsunami wave elevation and the associated uncertainty in prediction due to the uncertain inputs and the use of the emulator in place of the simulator.

Therefore, for a tsunami wave caused by the postulated landslide features, there is a 95% confidence that the resulting tsunami wave will have maximum elevation less than 2.18, and 99% confidence that it will be less than 2.35, looking at a location along the shoreline and far away from the source ($y_l = 8.38$). The same analysis could be performed similarly for other locations along the shoreline. Again the ability of the emulator to make predictions almost

	1%	5%	50%	95%	99%
maximum elevation	0.92	1.03	1.66	2.18	2.35
mean CI length	0.28	0.40	0.66	0.90	1.03

Table 3.1: Maximum elevation and mean CI length percentiles for the position $(x_l, y_l) = (0, 8.38)$.

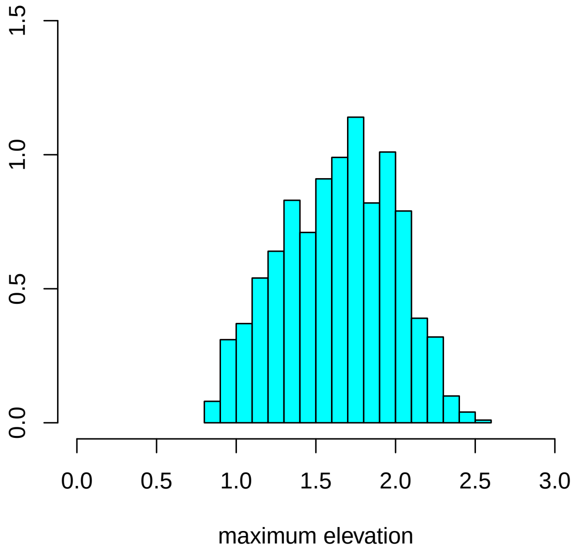


Figure 3.13: Output distribution for maximum wave elevation at the location $(x_l, y_l) = (0, 8.38)$.

immediately is highlighted in this case, since the total running time was just 83.9 seconds for 1000 runs at each of the locations compared to 30 minutes on the same computer for a single run of the SR tsunami model.

3.4 Spatial-temporal emulator

The next logical step after the generation of an accurate statistical surrogate for the SR model is to look at space-time variations of the tsunami waves, instead of looking at the wave motion over time at fixed locations. This step requires computationally intensive analyses and strong statistical expertise. The consideration of the water waves behaviour in space additionally to the behaviour in time might improve the predictions of the statistical emulator that considers only time variations and is restricted to a single locations in space. This possible improvement can be attributed to the fact that a spatial-temporal emulator can borrow strength across both space and time. A huge number of simulator outputs have to be obtained in order to have a sufficient knowledge of the waves' behaviour both in time and space and they have to be used for the emulator's generation. For this reason, Outer Produce Emulators are the most appropriate, compared to standard emulators, since they can handle efficiently both the large number of input parameter and also the extremely large number of evaluations.

The input parameter space is the same with the fixed-location application; similarly the prior choices for the input regressors and residuals covariance functions are the same. However, the output regressors and residuals covariance functions need to be modified. These functions are required to describe with the best way possible the behaviour of the output along the space coordinates. This again highlights the need of a dense enough computational mesh in order to have a large enough number of evaluations, as it would be impossible to find representative functions in the case of insufficient resolution.

3.4.1 Separable and Non-separable covariance functions for time and space

In order to make the analysis simpler and computational feasible, the residual covariance function for the outputs can be assumed to be separable in time and space. This means that it can be written as the product of a purely spatial and a purely temporal covariance functions. Separable covariance functions indicates that the space and time variations are independent. This assumption has the significant convenient advantage that the covariance matrix can be expressed as the Kronecker product of two smaller matrices that are purely temporal and purely spatial. Therefore, the determinant and the inverse of this matrix can be easily obtained. Nevertheless, this assumption is not realistic, as it imposes a strong constraint by ignoring the interactions between time and space. Physically there is a clear association between the spatial and the temporal dimensions and a realistic statistical representation should take this aspect into account. Non-separable covariance functions model space-time interactions and in many cases non-separable models are physically more realistic.

Genton [2007] developed techniques for finding approximate separable covariance matrices for cases where the space-time covariance matrices are non-separable. In this case, two matrices, one purely temporal and one purely spatial, that give the ‘best’ separable (Kronecker product) approximation have to be determined. The solution to this problem is to describe the non-separable matrix by the nearest Kronecker product (NKP) approximation. The NKP for space-time covariance matrix (NKPST) is described as follows: Let Σ be the space-time covariance matrix of a stochastic process $Z(\mathbf{s}, t)$, where $\mathbf{s}$ represents the spatial location and t the time. The NKPST problem is to find two matrices, S and T that minimize the Frobenious norm $||\Sigma - S \otimes T||_F$, where for a matrix $A = (a_{ij})$, $||A||_F = (\sum_i \sum_j \sigma_{ij}^2)^{\frac{1}{2}}$.

The use of separable covariance functions for time and space is supported by Rougier [2011], where he highlights some significant advantages of using separable covariance functions, such as the fact that the variance matrix has a Kronecker product form which makes the inversion, that is required in the emulation process, much more easier. Rougier [2011] concluded that the use of separable covariance functions for the prior emulator is safe. This is because by updating the prior emulator using computer model evaluations, the separability of the covariance function is erased.

On the other hand, Fricker et al. [2013] contradicts the opinion of Rougier [2011] that using separable covariance structures for the outputs is sufficient and results in accurate enough predictions. The authors state that wrong representations of the joint uncertainty can result in considering the simulator outputs as independent. At their analysis, they generate non-separable covariance structures for GP emulators and they compare the performance of the emulator resulting with this non-separable covariance with the performance of one that uses separable covariance structures and also with emulators that assume independence between the outputs. The conclusion of Fricker et al. [2013] is that only the non-separable emulators are able to give accurate predictions and represent appropriately the joint uncertainty about the model outputs.

To investigate this further, an analysis examining whether or not the use of a separable covariance structure for time and space actually improves the emulator's predictions was performed. For this analysis, the predictions of three different emulators were compared (results not shown here). The difference between the three emulators was only the selection of the output covariance function, with all the other functions and parameters to be exactly the same. The statistical emulation is applied to data generated from a non-separable covariance function and also to the Irish wind speed data set [Haslett and Raftery, 1989]. This data set is extensively used for covariance investigations due to its non-separable structure in time and space.

For the first emulator, a non-separable output covariance function, proposed by Gneiting [2002], was used. For the second emulator, the technique of Genton [2007] was employed for approximating the non-separable covariance function of the first emulator with a separable one. Finally, for the last emulator, a separable covariance function for time and space was used. The conclusion from this investigation was that a non-separable output covariance function does not improve the emulator's predictions. On the contrary, the best predictions for both the data sets resulted from the separable output covariance emulator. Similar conclusions were obtained by Bowman and Woods [2012].

3.4.2 Subemulators

Spiller et al. [2014] highlighted the fact that in some cases, the use of a single emulator for predicting variations in space as well as variations in time may be unnecessary challenging and damage the analysis. Some of the reasons for this is the complicated spatial correlations from topography and also the fact that in the case of tsunami events, some waves do not reach most of the locations and hence the simulated output is zero for many input scenarios. Another reason is the need for large matrices inversions which is the computational problem of using GP emulators. Therefore, the authors support that the generation of a single emulator at each location is more than enough, since there are not going to be any significant advantages by generating a single emulator that takes into account spatial-temporal correlations. Spiller et al. [2014] present an approach of building one emulator for each location using automating methods. Specifically, one location is considered at a time and an emulator is fitted specifically to the model output at the particular locations. The authors called these emulators “subemulators”. This process is automated and it is repeated in parallel for all the locations. Consequently, it can be concluded that the generation of a spatial-temporal emulator is not necessary in case that an efficient automated process of simultaneously building emulators at each of the locations of interest is introduced.

3.5 Discussion and Conclusions

Due to the catastrophic consequences of tsunamis, early warnings need to be issued quickly in order to mitigate the hazard. Additionally, there is a need of uncertainty and sensitivity analyses of the computer model used for predictions. In most of the cases, the more accurate and realistic the computer model is the more computational expensive is. Therefore, in these cases there are delays in predictions and also analyses that require a large number of model evaluations are impractical. In this chapter, based on results of Sarri et al. [2012], a statistical surrogate model is generated that approximates the landslide generated tsunami model developed by Sammarco and Renzi [2008]. This surrogate model, known as statistical emulator, runs almost instantaneous and represents very accurately the computer model, as it is proved using leave-one-out validation diagnostics. A convenient approximate emulator, known as the Outer Product Emulator, is used which assumes separable input and output regressors and covariance functions of the residual. The emulator was generated in a Bayesian framework, using combination of prior knowledge about the computer model properties, appropriate choices of functions and parameters and a limited number of simulator evaluations. The simulator is computationally expensive to evaluate, while the emulator produces estimates almost instantaneously. However, since the emulator is an approximation, an additional error was introduced in predictions. Nevertheless, this amount of error can be estimated, since the emulator gives predictions as statistical distributions. Additionally, using an accurate enough emulator, the computer model is represented with almost negligible error in the central region of the input space.

The emulator is used further for sensitivity and uncertainty analyses of the computer model. These analyses are almost impossible to be performed using the expensive simulator. In this chapter, these analyses are demonstrated on the landslide-generated tsunami model and the potential for reducing significantly the computational time is highlighted. In the case where the emulator requires 83.9s to get a thousand evaluations, the simulator requires 30 min for a single simulation. Therefore, in critical situations where early warnings are necessary, an emulator can be a lifesaver by providing accurate prediction in a very short time. The analysis in this chapter can be improved further by the introduction of a more realistic landslide source description. Additionally, in order to get better quantifications of the water free surface elevations as well as more accurate run-ups on the shore for the case of a tsunami event, more detailed simulations using a more advanced physical-based model and complex realistic bathymetry are required.

Chapter 4

Sensitivity of tsunami models to different representations of the coseismic uplift

4.1 How clever and computationally demanding do we have to be in coseismic seabed displacement modelling?

The Boxing Day tsunami in 2004 motivated an extensive research to improve the existing knowledge about tsunamis in order to be prepared and avoid similar disasters in the future. The research was focused on the forecasting as well as the design of efficient defences, such as sea walls, to protect people and infrastructure. Nevertheless, a few years later, the 2011 Tohoku tsunami catastrophic event, that was much larger than anticipated, highlighted the existence of huge gaps in understanding tsunamis and showed the weakness of the large number of numerical models available for tsunami propagation and inundation predictions. The Tohoku disaster therefore underlined the critical importance of a proper understanding of uncertainties in hazard models in order to avoid the underestimation of the hazard in disaster reduction planning based upon those models.

Extensive scientific research has been performed for earthquake-generated tsunamis, as it is the most common type of tsunamis. However, even all the investigations and the knowledge developed, many aspects of them are still unknown, with the description of the source and generally the tsunami generation phase to be highly uncertain. Consequently the predictions about the resulting tsunami waves are uncertain. A literature review about earthquake-generated tsunamis and the current modelling literature and limitations are presented earlier in Section 2.3 in the Background chapter.

4.1.1 Novel seabed deformation representation

In this chapter a new advanced approach for realistic seabed representation using quadratic curves over polygons to define the subsidence and uplift of the earthquake source is introduced.

This representation is compared with four simpler cases of different combinations of vertically-displaced blocks in the same displacement area. Specifically the “blocks” representations investigated are the following: single block, four blocks, sixteen blocks and four “piano keys”. The resulting tsunami wave elevation time series of all the cases are compared. The question to be answered is whether a naive inaccurate representation of the coseismic uplift and subsidence can result in inaccurate tsunami wave predictions. The most important aspect that is going to be examined is the prediction of the maximum tsunami wave amplitude anywhere in the model output wave train; it is the most significant for tsunami warnings and hazard mitigation, due to the high impact at shoreline. Furthermore, the aspect of whether or not blocks representations create high frequently waves artefacts in the water surface associated with the edges of the blocks is going to be investigated. This analysis gives us an insight on how clever it is required to be in representing coseismic uplifts as the tsunami source and also helps to understand the sensitivity of the tsunami model evaluations to the specification of the source deformation.

4.2 Tsunami model simulation for the comparison of different source representations

Using a simple slope bathymetry in a rectangular computational domain, five different tsunami event cases of coseismic deformations are generated. The bathymetry used for the analysis is the so-called canonical case which is a sloping beach followed by a constant depth region and it is shown in Figure 4.1. More specifically, the slope starts from the sea surface and goes down to a depth of 3000m and from that point it is considered constant. The sloping part is divided into three subparts, where each one has a different slope, with the first one to have the fastest decay and the middle one to be the only part that is displaced during an earthquake event. The first subpart that has length 100km and goes from the sea surface to 2500m below the water surface. The second covers the distance from 100km to 170km away from the coastline and the bathymetry decays by 300m. The last part goes up to 240km away from the coast and the bathymetry slowly goes down to the maximum depth considered that is the 3000m. Note that the shape of the bathymetry is considered constant along the y direction, which is parallel to the trench. The total width of the rupture displacement is 100km.

4.2.1 Advanced tsunami model VOLNA

An advance numerical tsunami model, called VOLNA [Dutykh et al., 2011], has been evaluated to obtain the resulting tsunami waves time series for different events generated using simple bathymetry and different coseismic displacement cases. VOLNA solves the nonlinear shallow water equations (NSWEs) in two horizontal dimensions and can be used for real-world opera-

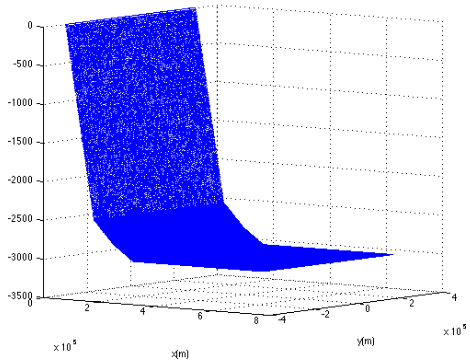


Figure 4.1: Simple canonical case bathymetry used for the analysis. It is divided into three parts of different decay rates and in the case of a tectonic earthquake event, seabed displacement is assumed to take place only at the middle subpart.

tional purposes. VOLNA code can cover the complete lifetime of a tsunami event: generation, propagation and inundation. The choice of the same model for the three stages of tsunami's life has a large computational advantage, since many computational problems, such as transferring large data sets from one model to another, are avoided. VOLNA uses a cell centred finite volume method and can function in an unstructured triangular mesh, which is considered as the computational domain.

The unstructured computational mesh makes it practicable to handle realistic environments for the bathymetry and geometry. Indeed, computational cost saving can be achieved, as the mesh can be designed to be more refined over areas of particular interest, such as around the coastline. On the other hand, the mesh can be made to be less unnecessarily fine at other locations, such as the boundaries of the computational mesh, where no detail is required for the analysis. Nevertheless, a careful choice for the boundaries of the computational domain is necessary in order to avoid undesirable reflections at the boundaries resulting in inaccurate evaluations.

A drawback of VOLNA is that it neglects dispersive and wave breaking effects. Dutykh et al. [2011] support that there are no significant consequences if the dispersive effects are neglected, since short waves don't significantly affect the coastal run-up and inundation. Additionally, Dutykh et al. [2011] highlighted the fact that no operational numerical model exists that considers dispersion and that the ones that do are very computationally expensive, and therefore impractical for real-time operation purposes. Therefore, VOLNA cannot model short-period tsunami waves, which means that it cannot be used for landslide generated tsunamis in case that the landslide causes short period waves.

Due to the fact that the tsunami waves amplitude is much smaller than the water depth, when they are far away from the coast, the NSWEs are assumed to describe them even the fact that they assume non-dispersive waves. These equations are practical due to the fact that they have nice mathematical properties and therefore, they are selected to model tsunami generation, propagation and inundation. The use of NSWE for tsunami modelling has been validated by Kervella et al. [2007] concluding that they do excellent work in the case of non-dispersive waves as they are able to capture very well the long waves behaviour.

Additionally, it is worth mentioning that VOLNA allows as an input time-dependent bathymetry which is a significant feature for the modelling of tsunami events generated by earthquakes or submarine landslides. Moreover, VOLNA can handle well the wet/dry transition and is able to run uniformly in the whole computational domain. This is very important aspect for the accuracy of the run-up computations. The VOLNA numerical code has been quantitatively validated by Dutykh et al. [2011] using realistic test cases. Specifically, two kinds of test have been used by the authors. The first one is a comparison of VOLNA evaluations with analytical solutions for testing the accuracy of the model. The analytical solutions are appropriate solutions of the linear shallow water equations. Dutykh et al. [2011] concentrated on tsunami generation and inundation phases, considering the case of a tsunami generation by a two-dimensional landslide. Specifically, the landslide is considered as a Gaussian shaped mass moving down an incline and starting from a position along the shoreline. They reach to the conclusion that there is a good agreement between VOLNA evaluations and analytical solutions.

The second kind of test is the comparison of the model evaluations with laboratory experiments results. This allows testing the ability of the code to reproduce actual events and also the validity of the nonlinear shallow water equations for tsunami modelling. For this analysis, Dutykh et al. [2011] investigate the tsunami coastal run-up in the case of a complex three-dimensional beach. They perform an experiment in a 205m long wave tank that reproduces at $\frac{1}{400}$ scale the tsunami event that took place in 1993 in Monai valley in Okushiri, Japan. The VOLNA evaluations have been compared with experimental measurements taken at three water gauges established in the water tank. In general, it can be concluded that VOLNA does a very good job in the predictions of the experimental data. Specifically, it is observed that the main wave is accurately described and additionally that the maximum run-up is adequately captured by the model. Therefore, the numerical model can reproduce the laboratory experiments accurately, even without bottom friction modelling. Both of these tests highlighted the high accuracy of VOLNA predictions. Stephanakis [2013] further validated VOLNA using the Catalina benchmark problems [Synolakis et al., 2007], which are test cases for analytical, experimental and field observations and similarly to Dutykh et al. [2011] VOLNA code's accuracy is confirmed.

The VOLNA code has been used by Poncet et al. [2010] to take evaluations for modelling tsunami events induced by two submarine landslides from the St. Lawrence Estuary. Particularly, they have considered a Gaussian shaped mass moving with constant acceleration [Liu et al., 2003] to describe seafloor motion due to a landslide. Also, they used an unstructured triangular mesh of approximately 150,000 triangles, which has been adjusted to be more detailed near the landslide. The authors pointed out that a next step for VOLNA in order to give accurate wave inundation predictions and so to be able to be used for coastal risk analysis is to work on refined bathymetric and topographic data. Additionally, they highlighted that is necessary to perform a sensitivity analysis for the landslide's parameters in order to find out which of them are significantly affecting the induced waves behaviour. These two aspects recommended by Poncet et al. [2010] are investigated in this thesis.

Looking at the technical aspects of VOLNA simulations, it is worthy mentioning that with help from University of Oxford, VOLNA can be evaluated on the Emerald GPU supercomputer. This makes the process faster and high-resolution modelling computationally feasible, as there is a possibility of using multiple GPUs and run many jobs in parallel, i.e. it is possible to perform simulations for all the locations investigated simultaneously. Additionally, VOLNA's running time is reduced significantly due to the fact that the outputs of the new version of VOLNA are written in Hierarchical Data Format (HDF5) format[1] files as a compressed matrix for all the gauges, instead of generating txt files for each gauge separately, which was initially the case. By generating HDF5 files instead of txt files, the total reduction in memory is about 36 times, although a lot more time reduction comes from having just one file as opposed to thousands. These technical improvements performed for VOLNA enable us to obtain evaluations for the forty different combinations of source characteristics for a large number of different locations.

4.2.2 Mesh generation

For the analysis, an unstructured triangular mesh is generated with size larger than the deformation area. Generally the larger the size of the mesh, the more the delay in reflections at the domain boundaries. Reflections at the boundaries are undesirable, since when reflections take place, the resulting wave elevations are not only the ones induced by coseismic seabed deformation or landslide, but they are combined with the waves that have already reached the boundaries and have reflected back. However, the size of the mesh has to be large at a limit that it is computationally affordable, since the larger the computational domain, the larger the computational cost implied. Therefore, both the reflections and the computational cost issues have

---

[1] HDF5 format is a file structure that can handle large amount of numerical data.

to be taken into consideration for the selection of the mesh size. For the analysis in this chapter, an unstructured triangular mesh is generated composed of 711,394 triangles in total. The width of the computational domain is selected to be three times the width of the deformation area, i.e. $300\mathrm{km}$ symmetrically around the origin and the length is chosen to be two times the length of the deformation, i.e. $480\mathrm{km}$ starting from the coastline.

4.2.3 Different coseismic displacement representations

Using the computational mesh generated and the different bathymetry deformation cases, VOLNA tsunami model simulations are performed, with the wave elevation time series collected at 63 gauges equally spaced at the whole domain. Specifically, looking at the direction perpendicular the coast, the gauges are located every $50\mathrm{km}$ between $50\mathrm{km}$ and $450\mathrm{km}$ from the shoreline, whereas for the direction parallel to the coast, gauges are placed at the origin and also two at each side of the origin every $40\mathrm{km}$. The location of the 63 gauges as well as the location of the deformation area in the whole computational domain is presented in Figure 4.2. The gauges are placed at the water surface level and records of the wave elevation are obtained every 0.1 seconds for up to 1000 seconds from the beginning of the seabed deformation motion.

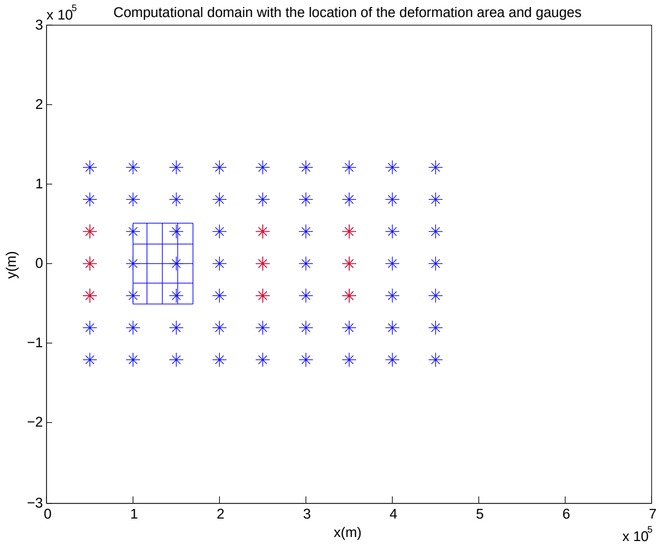


Figure 4.2: Computational domain with the rectangle indicating the location of the deformation area, which is divided into 16 rectangles as shown in Figure 4.3. The stars indicating the gauges location, with the red ones showing the gauges where the resulting wave elevation time series are investigated further in the analysis.

The wave elevation time series resulted using the advanced seabed representation, in which the source is described using quadratic curves, are compared with the evaluations obtained using the four simpler coseismic deformation representations. For the simpler representations the displacement is represented using blocks - one block, four blocks, sixteen blocks and four “pi-

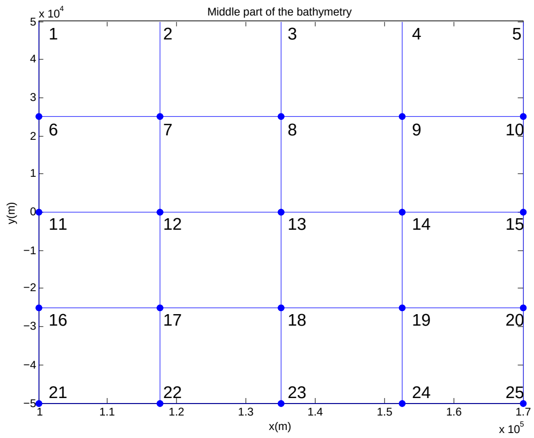


Figure 4.3: The middle part of the bathymetry is divided into 16 equal rectangles in order to be able to model the different seabed deformation cases.

ano keys" - instead of curves. The quadratic curves representations is more realistic compared to the blocks representation. This comparison enables us to look at the sensitivity of the model output to input representations of seabed deformation.

In all the five bathymetry representation cases only the middle subpart of the initial bathymetry is moving. More specifically, this subpart moves for 100 seconds with constant speed, which means that the rupture propagation's overall duration is 100 seconds, and this is the time when the maximum bottom deformation is achieved. For the generation of the novel seabed representation using quadratic curves the middle subpart is divided into 16 equal rectangles, as shown in Figure 4.3. The shape of the deformation is considered constant at the direction parallel to the coastline, whereas the amplitude is varied as it is explained in details below. On the other hand, the deformation shape varies along the direction perpendicular to the coastline, with initial subsidence followed by uplift. Appropriate magnitudes for the final seabed deformation are assigned at the nodes of the 16 rectangles in order to achieve a realistic representation of coseismic deformation events. The seabed starts from flat at all the nodes and moves to the final uplift or subsidence positions, with the nodes points to be jointed together with smooth surfaces. Specifically, the nodes are joint together with quadratic curves in the direction perpendicular to the shoreline and with straight lines in the direction perpendicular to the shoreline. The maximum uplift and subsidence of the deformation are considered to be $25m$ and $3.75m$, respectively. The shape of the deformation is the same for each row of rectangles in Figure 4.3. However, each one has have different amplitude.

Looking at the middle subpart which is displaced in Figure 4.3, the right edge, which consists of nodes 5, 10, 15, 20 and 25, is considered as the trench. The left hand side column of nodes from the trench, which are the numbers 4, 9, 14, 19 and 24 in Figure 4.3, is assumed to be the location where the maximum uplift occurs, denoted by j in the rest of the analysis in this chapter. It is noteworthy to mention that the uplift height is not constant across the length of the deformation part, with the uplift at the middle part of the area to be the highest with the uplift at node 14 is the maximum and is reaching the 2.5m. The uplift height decreases by moving from the centre to the boundaries, taking values up to 1.7m and 2m at nodes 9 and 19, respectively, and also 0.2m and 1.2m at nodes 4 and 24, respectively. The nodes at the trench are assumed to take uplift heights half the corresponding heights at the maximum, i.e. $j/2$. At the middle of the deformation part, which consists of nodes 3, 8, 13, 18 and 23, the seabed does not undergo any deformation, with the nodes height equals to zero during the deformation. Similarly, the bathymetry at the nodes of the left boundary of the middle subpart (1, 6, 11, 16, 21 in Figure 4.3) is not displaced. The bathymetry goes to the minimum at the second from the left column of points in Figure 4.3, with numbers 2, 7, 12, 17 and 22. It is assumed that the subsidence magnitude is always 15% of the trench height, i.e. $0.15j$. This percentage could range between 5% and 30% for realistic representation. Therefore, the bathymetry is uplifted at the two right columns of rectangles, whereas it subsides at the two left columns, with both the uplift and subsidence to represented with quadratic curves that have the maximum and minimum at the middle.

Even though the nodes are joint together with smooth surfaces, a big discontinuity is generated on purpose at the right boundary of the deformation, which is the trench location. This is in order to represent the trench fault that breaks the surface. The final form of the seabed deformation is clearly shown from the top view in Figure 4.4. The total displacement volume is estimated by adding the subsidence and uplifted volumes and this volume estimation is used for the generation of the four simpler deformation cases. In order to make the comparison fair, the same volume of displacement is selected for all the five cases investigated. Additionally, the four simpler representations are designed to “mimic” the more advanced representation whenever possible by matching the positions of uplifts and subsidences as much as is possible given the resolution that they provide.

The first simple bathymetry deformation case assumes that the whole middle subpart is uplifted, with all the 16 rectangles moving together as a single block up to a certain height. This height is selected so that the total volume of displacement is equal to the volume of displacement of the quadratic curve representation described before. Therefore, for this representation only uplifted displacement is able to be considered. The second simple representation assumes that the middle subpart is divided into four equal rectangles and each one have different dis-

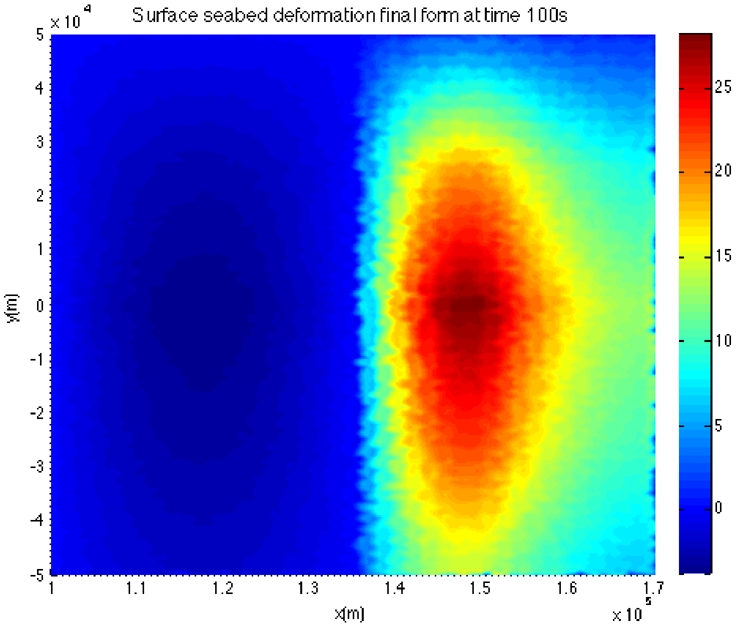
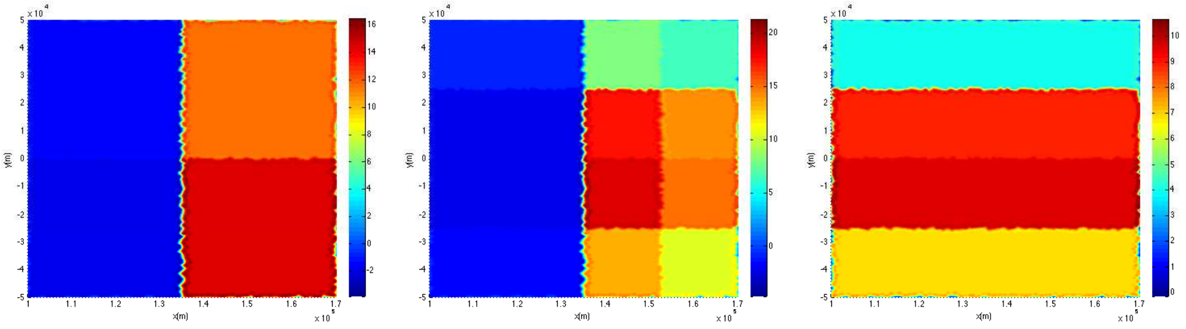


Figure 4.4: Final form of the deformation for the advanced quadratic curves representation. The left part of the seabed subsides and the right half is uplifted to a maximum.

placement amplitude. These amplitudes are selected so that the motion mimics the advanced seabed representation case, with respect to the volume of displacement as well as whether they subside or uplifted. More specifically, for each of the four rectangles the volume of displacement for the case of the quadratic curves representation is calculated and used to estimate the displacement amplitudes for the four blocks deformation case. The two rectangles of the left column subsided and the two right ones uplifted, as it is shown in Figure 4.5a.

The third representation investigated is similar to the second, but it considers sixteen instead of four rectangles and it is presented in Figure 4.5b. The displacement amplitudes are selected as for the second case described earlier. Finally, the fourth type of simple deformation is the four “piano keys”, where the middle block is divided into four rectangles parallel to the slope, which are the four rows of rectangles shown in Figure 4.4. The displacement height of each piano key is estimated as for the second and third simple cases. As with the single block representation, this time the uplift and subsidence parts of the quadratic curve representation cannot be mimicked, as each ”piano key” covers both uplift and subsidence parts. Hence, either uplift or subsidence displacement has to be assigned to each of them and since the uplift is much larger than the subside in total for each row of four rectangles, all the ”piano keys” are considered to be elevated. The final case is shown in Figure 4.5c. The five representations in the whole computational domain are shown in Figure 4.6 with the maximum uplift and subsidence



(a) four blocks (b) sixteen blocks (c) ”piano keys”

Figure 4.5: Bathymetry deformation middle part for the (a) four blocks, (b) sixteen blocks and (c) ”piano keys” representations.

Case	Max uplift (m)	Max subsidence (m)
1	7.33	
2	14.44	1.93
3	18.75	2.25
4	9.56	
5	25	3.75

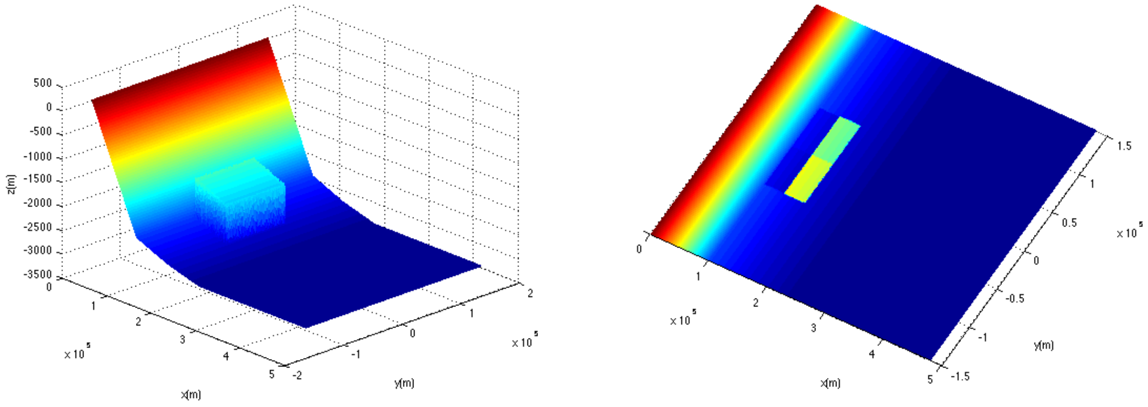
Table 4.1: Maximum uplift and subsidence for each of the five coseismic deformation cases. Case 1: Single block, Case 2: Four blocks, Case 3: Sixteen blocks, Case 4: “Piano keys”, Case 5: Quadratic curves representation.

amplitudes for each one summarised in Table 4.1.

4.2.4 Resulting wave elevations comparison

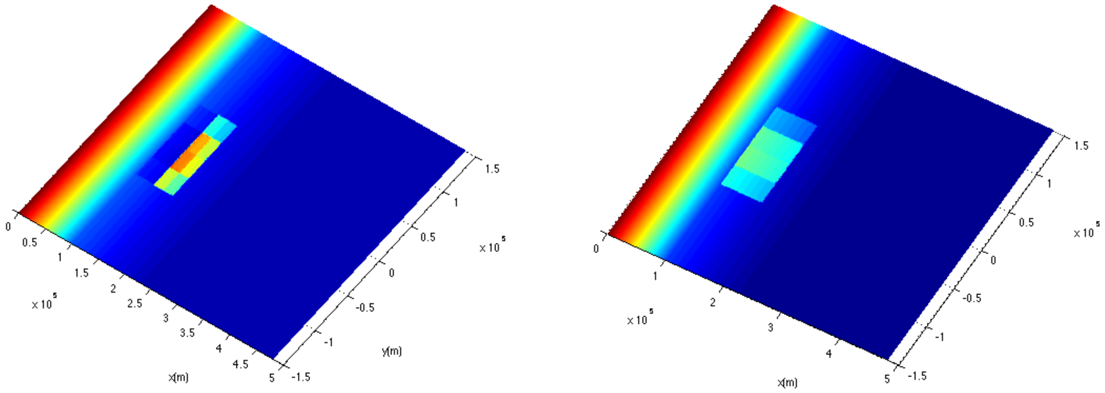
In order to compare the five deformation representation, VOLNA runs are carried out and the resulting wave elevation time series are investigated. Each of Figures 4.7, 4.8 and 4.9 presents the wave elevation at three gauge locations. Specifically, Figure 4.7 consists of three gauges evaluations along the width of the domain, located $50\mathrm{km}$ away from the shoreline; one for the middle of the displacement width, and two at $40\mathrm{km}$ at each side. Similarly, Figures 4.8 and 4.9 contain evaluations for the same three locations at $250\mathrm{km}$ and $350\mathrm{km}$ away from the shore, respectively. The wave elevation is recorded every 1s time step up to 250s after the beginning go the displacement. Remember that the seabed displacement requires 100s to take its final form.

From the time series, it can be concluded that the coseismic deformation representation significantly affects the resulting wave elevation, especially for the first wave, with up to factor-of-two differences can be obtained. Specifically, the first wave amplitude for the cases of single block and “piano keys” is similar and much smaller than the other three representations, which



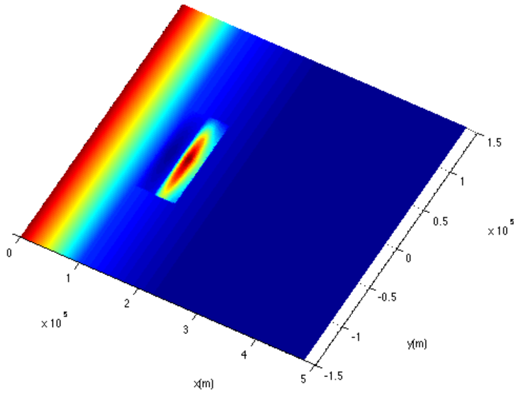
(a) single block (case 1)

(b) four equal blocks (case 2)



(c) sixteen equal blocks (case 3)

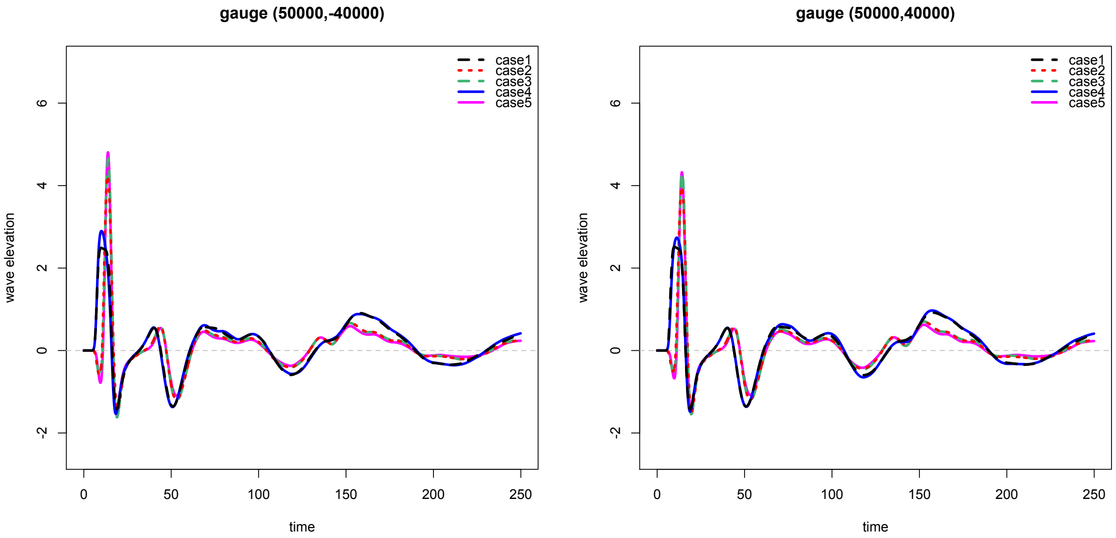
(d) piano keys (case 4)



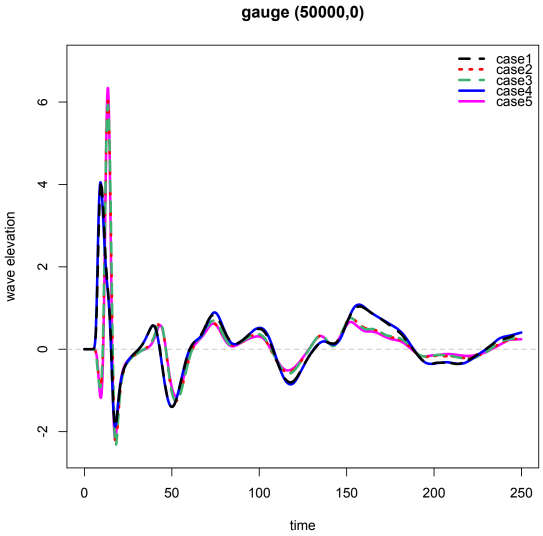
(e) quadratic curves representation (case 5)

Figure 4.6: Final form of different bathymetry representations compared, with the uplift and subside to be scaled up by 100 for demonstration purposes.

have similar amplitude as well. Additionally, the frequency of the first waves for the cases of single block and "piano keys" is much smaller than the other three cases. Also, for the case of the gauge located at $50\mathrm{km}$ away from the shore, the quadratic curves and the four and



(a) $(x_l, y_l) = (50, -40)\mathrm{km}$ (b) $(x_l, y_l) = (50, 40)\mathrm{km}$

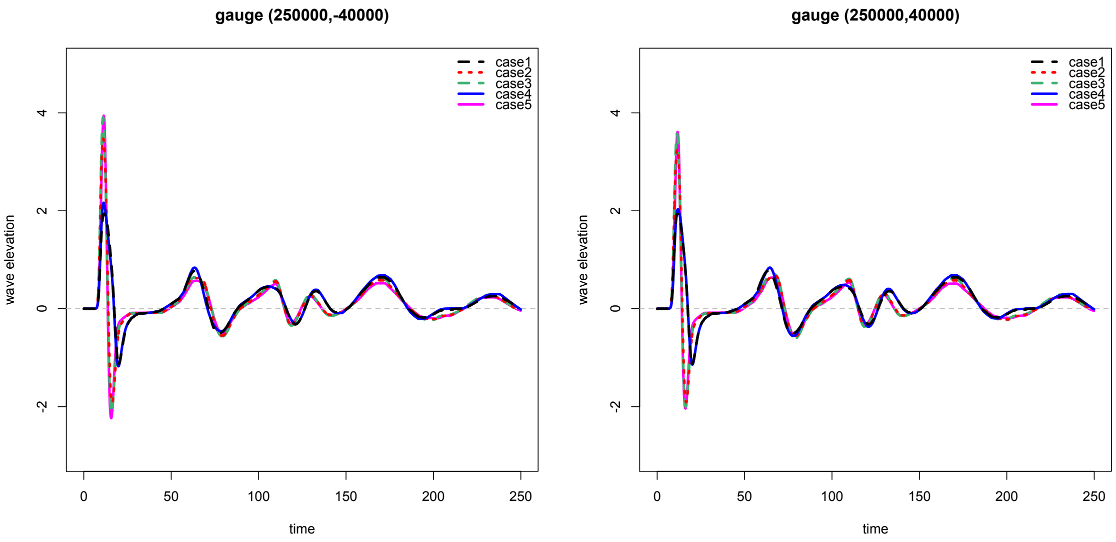


(c) $(x_l, y_l) = (50, 0)\mathrm{km}$

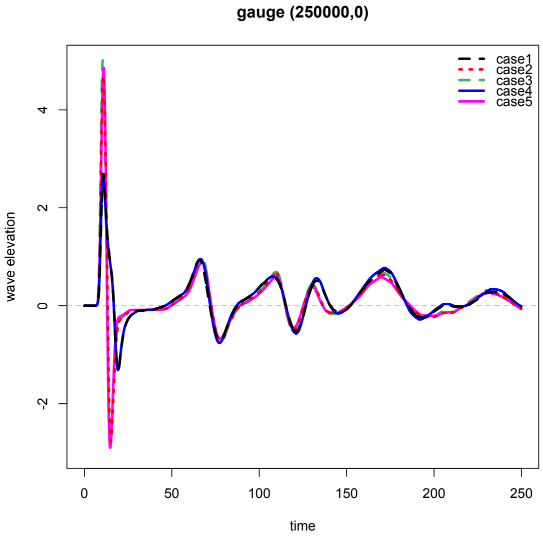
Figure 4.7: Tsunami wave elevation time series comparison for five different seabed deformation cases looking at three gauges $50\mathrm{km}$ away from the coastline. Case 1: Single block, Case 2: Four blocks, Case 3: Sixteen blocks, Case 4: "Piano keys", Case 5: Quadratic curves representation.

sixteen blocks cases present a small depression wave before the first wave elevation, which this is not the case for the other two seabed representations. This is attributed to the subsidence that is introduced in the deformation of the three cases, whereas for the other two cases, the deformation is assumed to only be uplifted.

After the first wave, all the cases have very similar wave behaviour, especially for the location $250\mathrm{km}$ away from the shoreline. Generally, slightly higher wave amplitudes are obtained



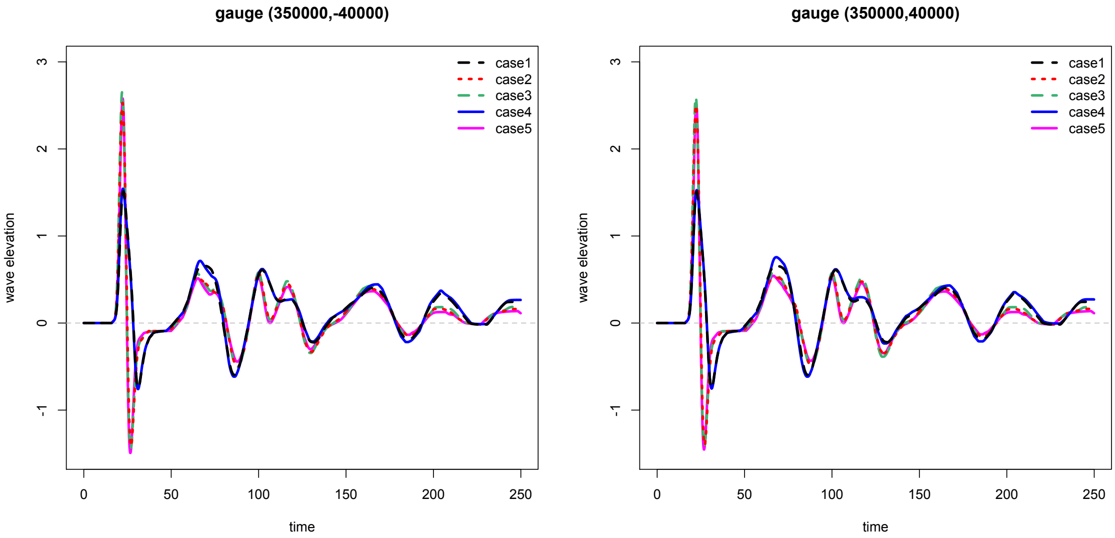
(a) $(x_l, y_l) = (250, -40)\mathrm{km}$ (b) $(x_l, y_l) = (250, 40)\mathrm{km}$



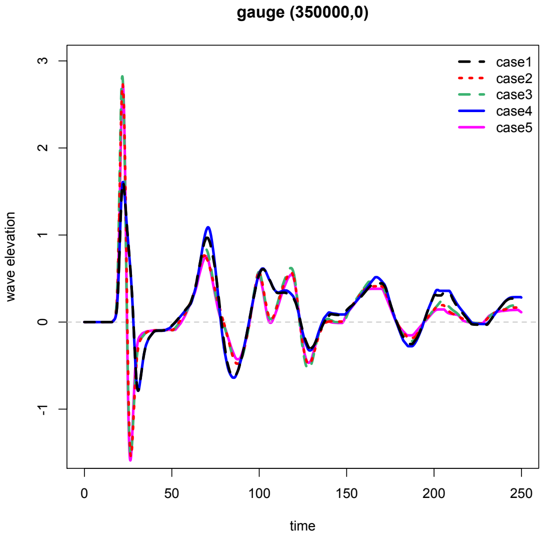
(c) $(x_l, y_l) = (250, 0)\mathrm{km}$

Figure 4.8: Tsunami wave elevation time series comparison for five different seabed deformation cases looking at three gauges $250\mathrm{km}$ away from the coastline. Case 1: Single block, Case 2: Four blocks, Case 3: Sixteen blocks, Case 4: “Piano keys”, Case 5: Quadratic curves representation.

after the first wave for the single block and “piano keys” representations compared to the other three cases. At all the gauges investigated, the waves time series behaviour is divided into groups, with the single block and “piano keys” cases to behave similarly, and the other three cases, four blocks, sixteen blocks and quadratic curve, to be the second group that has similar behaviour. Since the four and sixteen blocks representations have alike time series to the quadratic curve representation, this indicates that these two representations are “better” com-



(a) $(x_l, y_l) = (350, -40)\text{km}$ (b) $(x_l, y_l) = (350, 40)\text{km}$



(c) $(x_l, y_l) = (350, 0)\text{km}$

Figure 4.9: Tsunami wave elevation time series comparison for five different seabed deformation cases looking at three gauges 350km away from the coastline. Case 1: Single block, Case 2: Four blocks, Case 3: Sixteen blocks, Case 4: “Piano keys”, Case 5: Quadratic curves representation.

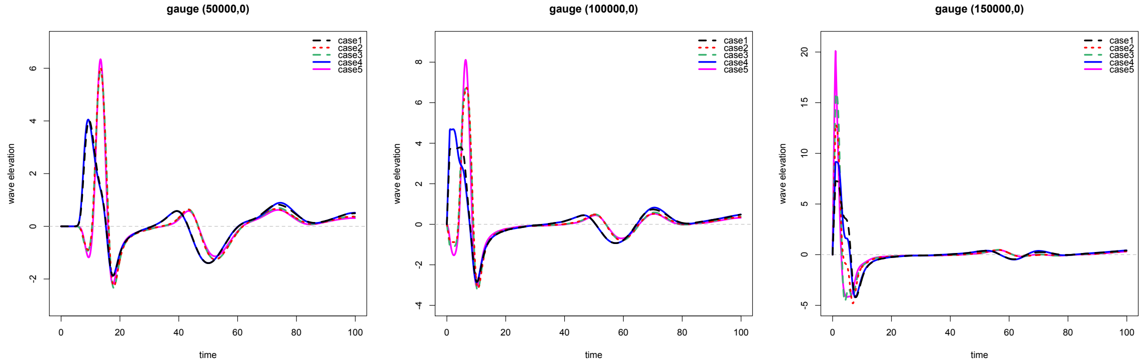
pared to the other two, as they are closer to the realistic representation. This suggests that the locations of the subsidence and uplifted parts of the deformation area are important for the wave elevation predictions, since the cases that are doing better (four and sixteen blocks) are the cases where it is feasible to mimic the locations of the quadratic curve representation uplift and subsidence. Nevertheless, the use of the realistic quadratic curve representation is recommended, since still some differences are obtained. However, a good alternative could be a representation where the area of the deformation is stated correctly, together with the accurate maximum uplift height and subsidence, since the exact shape of the deformation is not that important.

By comparing the time series of the gauges along the origin with the ones located at the left and right sides of the origin, it can be obtained that the first wave amplitudes for all the representation cases are larger for the gauges along the origin compared to the side ones, and this is more obvious for the gauges $50\mathrm{km}$ and $250\mathrm{km}$ away from the shore. In general, the time series for gauges along the origin and off shore are very similar. Figure 4.10 presents the wave elevation time series up to time 100s, which is the duration of the deformation, for all the gauges along the origin. For the figure, it is noticeable that the first wave amplitude is larger for gauges closer to the deformation area, with the second wave maintain its amplitude.

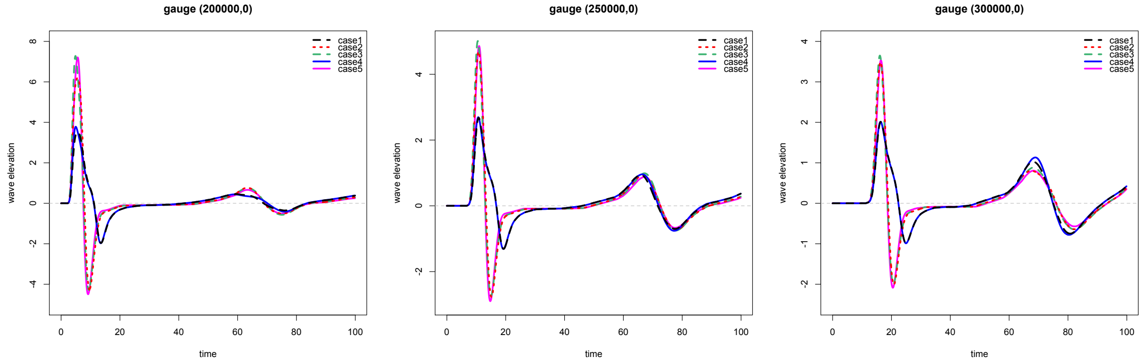
4.3 Discussion

In this chapter, an advanced representation of the coseismic deformation using quadratic curves is presented. This representation does not require any data for past events or any strong assumptions such as flat bathymetry or passive generation, which are commonly used so far in the literature. Generally, the simpler the deformation is described and the more assumptions used the less realistic it is. The advanced deformation is compared with four simpler cases that they are extensively used in source modelling: single block, four blocks, sixteen blocks and four "piano keys". The comparison is performed by observing the resulting tsunami wave elevation time series obtained by evaluating VOLNA tsunami simulator, assuming simple sloping bathymetry, with three subparts of different slopes, that is extended to flat bathymetry until the end of the domain. The computational domain used is an unstructured triangular mesh, which is selected to have significantly large size in order to avoid reflections at the boundaries.

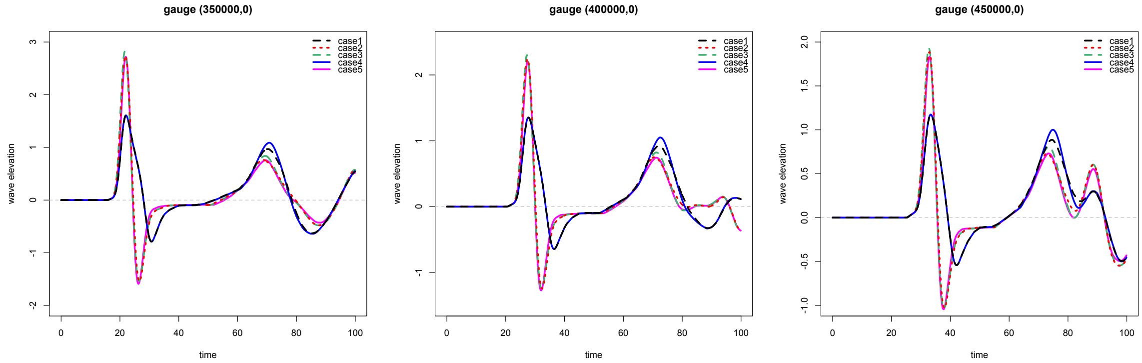
The analysis explores the sensitivity of resulting tsunami wave elevations to the seabed deformation representation. The conclusion from this investigation is that the seabed representation is an important aspect in tsunami modelling. Large biases in the predictions can result from a simplistic coseismic representation. Across the five cases, differences by a factor of two in the tsunami wave elevations of the first incoming wave can be observed, which is the most important parameter for tsunami warnings, predictions and specification of defences.



(a) $(x_l, y_l) = (50, 0)$km (b) $(x_l, y_l) = (100, 0)$km (c) $(x_l, y_l) = (150, 0)$km



(d) $(x_l, y_l) = (200, 0)$km (e) $(x_l, y_l) = (250, 0)$km (f) $(x_l, y_l) = (300, 0)$km



(g) $(x_l, y_l) = (350, 0)$km (h) $(x_l, y_l) = (400, 0)$km (i) $(x_l, y_l) = (450, 0)$km

Figure 4.10: Wave elevation time series for all the gauges along the origin $y = 0$.Case 1: Single block, Case 2: Four blocks, Case 3: Sixteen blocks, Case 4: “Piano keys”, Case 5: Quadratic curves representation.

Chapter 5

Cascadia study using VOLNA evaluations

This chapter details an extended investigation on the coseismic deformation that could cause tsunami events, focused on the Cascadia Subduction Zone. The seabed displacement is realistically modelled and hypothetical earthquake events have been generated. The geometry of the deformation area is represented using arbitrary shaped 4-sided polygons and quadratic curves are used to describe the subsidence and uplift shapes, which is a realistic representation as explained in detail in Chapter 4. The coseismic representation in this analysis is novel, more advanced and closer to reality compared to other studies.

Simulations of the highly accurate tsunami model VOLNA give evaluations at many locations for the tsunami wave propagation and coastal inundation for the different scenarios, which are generated by varying three of the source characteristics. These evaluations have been statistically analysed using statistical emulation. Registration and Functional Principal Components techniques have been employed in order to improve the predictions of the emulator. The improvement results from the fact that these methods are able to remove phase variations in the data and also they can help in the selection of more representative regression functions.

Following that, a limiting event has been investigated. Specifically, the trench amplitude was assumed to be three times the maximum middle uplift. Moreover, the uplift amplitude was considered as constant along the whole margin and the deformation was simultaneous lengthwise the displacement area. This event has minor chances to occur in reality. Nevertheless, it is a relatively computationally inexpensive way to draw conclusions about the maximum possible inundation at the Cascadia Subduction Zone, which could help in generation of inundation maps and effective defences in order to be prepared and eliminate disaster from future events in the area.

However, in order to be able to produce reliable tsunami warnings leading to hazard mitigation, as well as, obtaining tsunami hazard intensity that helps in tsunami exposures definition, high accuracy is necessary for eliminating uncertainties in the analysis as much as possible. To attain this, high resolution and a large number of tsunami model simulations are required. These can be achieved by combining VOLNA tsunami model evaluations with statistical emulation.

A general introduction with definitions about the Subduction Zones and also specifically about the Cascadia Subduction Zone (CSZ) can be found in the Background chapter and specifically in Section 2.1.3 and 2.1.3.1.

5.1 Mathematical Methods

At this point, a brief introduction and literature of some mathematical approaches applied at this chapter are going to be presented in this section. Specifically, the concepts of Principal Components, Functional Principal Component Analysis and Registration are introduced. By incorporating these methods in the process of building the statistical emulator, the resulting emulator's predictions can be improved further.

5.1.1 Principal Components (PCs)

The process of building a statistical emulator to represent a computer model is demonstrated in detail in Chapter 3 and is going to be applied again at this chapter, focusing on earthquake-generated tsunami events at the Cascadia Subduction Zone. As it is described earlier, in order to build an emulator, appropriate prior choices for the regressors and the covariance functions are necessary together with a choice of some parameters. In this chapter, a different approach for building an emulator is introduced, where PCs are used in place of the standard choice of Fourier basis functions for the output regressors. The emulator's predictions from the two different approaches are compared.

The largest advantage of this replacement is the the number of functions required to sufficiently describe the wave elevation time series is smaller, which can reduce the computational cost. Furthermore, the use of PCs is a data-driven process. It does not require expert knowledge and time spending in observing a large number of time series and deciding on the selection of Fourier terms that describe the behaviour of the wave elevations.

Similar to the approach introduced in this chapter, a methodology for emulation of computer models using GP models and PC analysis is described by Chang [2014] [Chang et al., 2014]. Their approach is first to describe the computer models evaluations as PCs and then to apply statistical emulation using GP. However, due to the fact that the PCs are not correlated with each other, they are modelled separately using independent GPs. This means that the dependency between the PCs is ignored. Nevertheless, the emulation based on this assumption provides very accurate approximations of the computer model in most of the cases. Additionally, Chang [2014] applied calibration to the model parameters using the resulting GP emulator for the PCs, named as the PC emulator.

Furthermore, Bowman and Woods [2012] present a methodology for modelling the multivariate computer model outputs. Their approach takes into account the spatial structure in the responses. The authors propose the use of a thin-plate spline (TPS)[1] to capture the spatial structure and fit a GP emulator to the coefficients of the resulting basis functions. They show that TPS can be used successfully for dimension reduction and it results in more accurate emulator predictions compared to the use of PCs. Bowman and Woods [2012] compare three different emulation techniques which are the following: fully Baysian approach with a PC basis, fully Baysian approach with a TPS basis and "plug-in" Bayesian approach with a TPS and separable covariance. The comparison of the three techniques highlight the significant outperform of the TSP emulator compared to the PC emulator. The second technique assumes that the basis coefficients are independent, whereas the last one, accounts for the basis coefficients dependency. The last technique performance is slightly better and it results in significantly more realistic uncertainty quantification and is generally a more computationally tractable method, allowing fast emulation. Therefore, the results Bowman and Woods [2012] support the use of a separable emulator.

5.1.2 Landmark Registration

Functional data are evaluations of functions varying over a continuum, which is commonly time. They are observed usually as discrete n pairs (t_j, y_j), with y_j to be the output of the function at time t_j. Usually it is assumed that there exists a function x that can produce the observed data. The function x is smooth in the case that a pair of adjacent outputs, y_i and y_{i+1}, are similar to each other. Commonly, the data y_j, $j = 1, \ldots, n$, are used to infer the function x. The function can be described by the equation

$$y_j = x(t_j) + \epsilon_j \tag{5.1}$$

where ϵ_j is the observational error or noise. The variance-covariance matrix for $\mathbf{y} = \{y_1, \ldots, y_n\}$ is equal to the corresponding matrix for $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n\}$, since the values $x(t_j)$ are considered as fixed with variance 0.

A system of basis functions is a set of known functions ϕ_k. An appropriate choice of basis functions can be used to approximate any function by a linear combination or weighted sum of a sufficiently large number K of these functions. The function x is built up from a basis

---

[1] TPS is a technique for data interpolation and smoothing that obtains a "minimally bended" surface that is smooth and passes through all given points.

functions system $\phi_1(t), \ldots, \phi_K(t)$, as the linear combination

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{5.2}$$

where c_k are the coefficients. Therefore, the functional data are generated by assigning a set of basis functions and a set of coefficients. When $K = n$, an exact representation can be obtained, as the coefficients can be selected to get $x(t_j) = y_j$ for each j. Hence, the degree of smoothness can be determined by the number K of basis functions. This means that K is a parameter that is chosen according to the characteristics of the data. The larger the number of the basis functions K, the better the fit to the data. However, a large number of K might result in fitting unnecessary variations. On the other hand, small K can result in missing some important aspects of the function x.

One of the most popular basis function system is the Fourier series which is defined by the basis $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin r\omega t$ and $\phi_{2r}(t) = \cos r\omega t$. The specific basis is periodic and ω defines the period $2\pi/\omega$. Therefore, for cases where the data are periodic, then the Fourier basis is useful. For non-periodic functional data, the most common choice is spline basis functions, also known as B-splines. In order to define a spline, the time interval has to divided into L subintervals, which are separated by values τ_l, $l = 1, \ldots, L-1$, known as the breakpoints or knots. The spline is a combination of polynomials and has specific order m over each interval[2]. At the knots, adjacent polynomials are smoothly join up, which means that the functional values are constrained to be equal at the knots locations. The number of degrees of freedom is equal to the order of the polynomials plus the number of the interior knots. A spline function is determined by the order of the polynomials m and the knots sequence τ. Hence, each basis function, $\phi_k(t)$, is a spline as it is determined by the polynomial order and knots. For any given set of knots, the sum of the B-spline basis function values at any time point is equal to one.

Since in the case of VOLNA evaluations wave elevation time series do not fluctuate periodically, the use of Fourier basis functions to describe the output regression functions might be not the best idea. The steps described below are followed in order to apply registration to the data and eliminate phase variations. Then Functional Principal Component Analysis is employed to obtain the main modes of variation that describe the time series. The steps are [Ramsay [2006], Ramsay et al. [2009]]:

1. The generation of a B-spline basis. For this step, the interval over which the functional

---

[2] The order of the spline is one more than the degree, which is the higher power or the number of interior knots, and defines the smoothness.

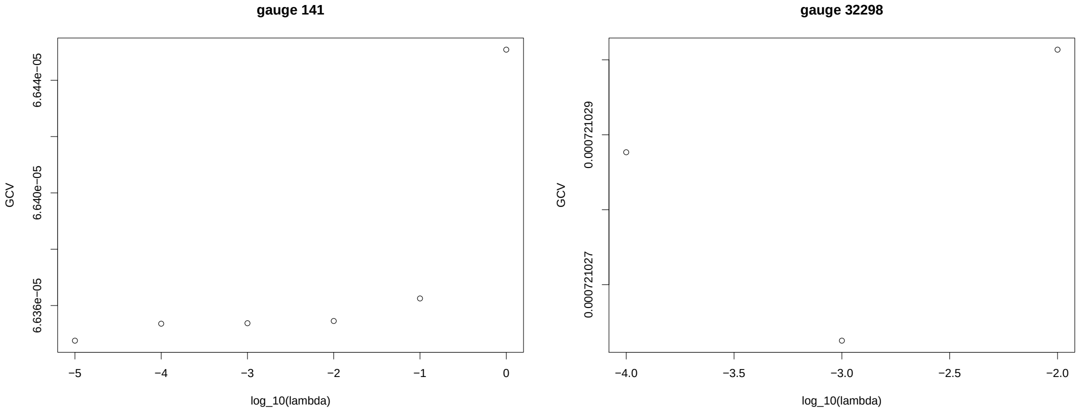
data are obtained is defined. For this analysis, the whole time domain is used, where evaluations are available every 10 seconds, up to 3000 seconds for some gauges and up to 7000 seconds for others. Additionally, the number of knots has to be selected. The knots are selected at every time steps, i.e. every 10 seconds. Furthermore, the order of the B-spline has to be defined. It is chosen to be equal to 4, which corresponds to a cubic spline and gives flexibility for capturing the wave shapes. The number of basis functions can then be calculated using:

$$no.\,basis = no.\,knots + order - 2 \tag{5.3}$$

2. Following that, the basis functions are evaluated at each time step. This results in a matrix of evaluations, with the number of rows to be equal to the number of time steps and the number of columns to the number of basis functions, and each entry is the corresponding evaluation.

3. Define the amount of smoothness that should be applied to the data. Different values are tried in a range defined by some trial-and-error initial tests in order to identify the best choice. For each value that is tested, the fitting criterion is the weighted least squares and the smoothness and it is defined in terms of the roughness penalty λ. As λ increases, the curvature decreases, where a very large λ implies a straight line. At the other extend, as λ goes to zero, the function is free to fit the data as closely as possible. Departure from smoothness trades off against fit to the data. To decide this trade off, the resulting value of the generalised cross-validation (GCV) criterion is obtained, which is proposed by Wahba and Craven [1978]. Specifically, the smoothing parameter λ is chosen so that the GCV is minimum. The process for this selection is to plot the GCV with respect to $\log_{10}\lambda$ and choose the value of $\log_{10}\lambda$ which corresponds to the minimum GCV. However, usually the GCV values vary slowly with $\log_{10}\lambda$ close to the minimum. This means that more than a single value of λ corresponds to the same GCV value and therefore an exact value of λ cannot be selected. The plots of GCV versus $\log_{10}\lambda$ for the two gauges number 141 and 32298 are presented in Figure 5.1.

   From the Figure 5.1, it can be concluded that the value of λ that corresponds to the smaller GCV criterion and hence works better for the data is equal to 0.00001 for gauge 141 and 0.001 for gauge 32298. For the case of gauge 141, the value of $\log_{10}\lambda$ equals to -6 leads to a much larger GCV value. Hence -5 is the minimum, even though the curvature of the GCV curve is small for $\log_{10}\lambda$ between 10^{-2} and 10^{-5}.



(a) gauge 141

(b) gauge 32298

Figure 5.1: GCV vs $\log_{10}\lambda$. The value of $\log_{10}\lambda$ which corresponds to the lower GCV is selected.

4. Landmark registration is then applied in order to remove the phase variation that can be observed in the data. This means that certain features, such as peaks, do not occur at the same time among the wave elevation time series for different inputs. In this case the mean curve of the non-registered curves is not representative, since the amplitude and the slope of the curves cannot be captured. On the other hand, the mean curve of the aligned curves is more representative. The registered curves are aligned and vary only in amplitude. The transformations, called the warping functions, capture phase variations of the original data set. In the case of functional data, phase variations are recommended to be eliminated first. To do so, the first step is the selection of landmarks.

   The landmarks timing is used to align the curves, by estimating a strictly increasing transformation of the time, $h_i(t)$. For a time interval $[0, T]$, the time warping functions must satisfy the constraints $h(0) = 0$ and $h(T) = T$. The registered functions are described with the equation

   $$x_i^*(t) = x_i[h_i^{-1}(t)] \tag{5.4}$$

   where the aligning function, $h^{-1}(t)$, satisfies $h^{-1}[h(t)] = t$

   Landmark registration is the simplest curve alignment procedure. However, in cases where no visible landmarks exist, another registration process has to be used. Even when the landmarks are clear, identifying the timing for all the different curves might be time consuming and a fully automatic method might be preferred. Nevertheless, the landmark registration can become automatic using clever selection of landmarks. Furthermore, another negative aspect of landmark registrations is the fact that using only few landmarks, the curves might be left unregistered at other locations.

Another form of registration is the continuous registrations, which uses the whole time series curve for alignments, not just specific features. Kneip and Ramsay [2008] present a registration process, where an algorithm is generated combining registration with PCA. They apply it to real data showing that it works very well. However, the authors state that their approach can be improved further by obtaining a way to estimate simultaneously the PCs and the warping functions and also by avoiding dependency of the solution on the smoothing parameter λ.

The simple landmark registration is used for the analysis in this chapter. The reason for this is that there are clear landmarks and also that the process in applied only a few specific locations. Furthermore, the results after applying landmark registration are more than satisfactory and so there is no clear need for another form of registration at the moment. Two landmarks are selected looking at all the 40 different time series, one for each LHD points. The landmark registration gives functional registered curves and the coefficients of these curves are used and combined with the B-splines to obtain the registered functional data.

5. Following the landmark registration, Functional Principal Component Analysis (FPCA) has been applied to the functional registered data and it is described in the next section. The FPCA can capture the primary modes of wave elevation variations without loosing essential information. These modes of variation are known as Principal Components (PCs) or harmonics. For the two gauges presented here, only the first two PCs are enough to describe the data, as they are able to characterise a large percentage of the data variability. As a result, data-driven (empirical) basis functions are obtained with their coefficients to be the eigenfunctions or harmonics of the FPCA. It is worthy noticing that if FPCA is carried out using the unregistered data, the first principal components account for a low percentage of variation.

5.1.3 Functional Principal Components Analysis

Principal Component Analysis (PCA) can be used to obtain the primary modes of variations in the data. The main concept is firstly to find the direction that corresponds to the largest variance between the data points. This can be achieved by calculating the eigenvectors and eigenvalues of the data's covariance matrix. The eigenvector with the largest eigenvalue is the direction of greatest variation. This eigenvector is the first PC. Following the selection of the first PC, the data are then projected on an orthogonal to that direction hyperplane, where a new data set and a new direction of maximal variation are obtained. The direction which corresponds to the second largest eigenvalue is the one with the next highest variation and so on. The specific process is repeated and it stops when enough directions are selected. The eigenvalues of the data's covariance matrix indicate the importance of each of the PCs and they are used to decide how many PCs are required to sufficiently describe the data.

The corresponding analysis for functional data is the Functional Principal Components Analysis (FPCA). The advantage of the FPCA compared to the PCA is that it allows us to take into account the smoothness of the data. The steps for the FPCA are first to define the basis functions, where typically B-splines are selected with K coefficients. Following that, using this basis, functional representations are obtained. Then, the GCV criterion is used to select the smoothness λ. The smoothing parameter is applied to account for the smoothness of the curves in the process of the PCs computation. In the case that the smoothing parameter was not included, the PCs would be very noisy. Afterwards, smooth basis functions are obtained and registration is applied to align the curves. Finally, PCA is applied on the registered functional data, with the PCs to be described by the following equation:

$$\hat{x}(t) = \sum_{l=1}^{K} d_l \psi_l(t) \tag{5.5}$$

where d_l are the coefficients or scores and $\psi_l(t)$ is the B-spline basis at t.

5.2 Realistic Seabed Deformation Representation

The large challenge to accurately represent the source characteristics and displacement is investigated in this analysis. The research around the numerical modelling of a tsunami event started around 50 years ago and a large progress have been achieved in these years. Numerical simulations are fundamental for tsunami warnings and generation of defences in order to mitigate the disaster and save human lives and infrastructure. Progress in this area has been achieved by the improvement in the estimations for the fault rupture mechanism and the sea bottom displacement, as well as the appearance of high-speed computers which made feasible the simulation of more computationally demanding models.

Even in the most extensive investigations, there are still large uncertainties in these simulations. Nevertheless, as stated by Shuto [1991], the initial numerical simulation research performed resulted in satisfactory predictions and reliable enough for practical use, as the total error in predictions of the maximum run-up height is within 15%. However, this does not indicate that the whole tsunami wave time series can correctly reproduced. A review of the early numerical models of tsunami generation by coseismic deformation has been provided by Shuto [1991]. Shuto [1991] concluded that it is not feasible to calibrate the computed wave profiles due to the limited number of tsunami gauges and their poor resolution ability.

Focusing on the Cascadia Subduction Zone and using realistic data for the initial bathymetry, different hypothetical scenarios of events that can result in a tsunami are generated by changing three of the source characteristics. Specifically, the source maximum uplift, the ratio of the trench to the middle maximum heights, as well as the source propagation have been used to describe the source and they are going to described in detail later. Different combinations of these characteristics describe different event cases.

The seabed deformation at the location of the source is added at the initial bathymetry. The deformation shape is represented by a set of arbitrary shaped 4-sided polygons, where the arbitrary shape allows the realistic representation of the geometry. The polygons are chosen so that the deformation is allowed to vary parallel and perpendicular to the trench. For the Cascadia Subduction Zone, the exact location of the trench is unknown due to the lack of information about land-level movements from past events. Therefore, the width of the fault rupture, which is an important aspect of the earthquake magnitude, is highly uncertain [Nelson et al., 2008]. As marked by Nelson et al. [2008], in order to determine the width of past earthquake ruptures at Cascadia, an estimation of the coseismic coastal subsidence is required.

Also, quadratic curves are used to represent the deformation shape of the uplift and subsidence, as described in Chapter 4. This is very flexible as no assumptions are required. Also it avoids discontinuities at the surfaces, except as indicated by geology. Specifically, the only discontinuity that is introduced in this representation is at the trench, which is a realistic discontinuity where the subduction zone cuts the seabed. These facts makes the deformation novel as it is more advanced and close to reality compared to previous similar probabilistic hazard studies, where typically uniform slop are used rather than localised large slip patches, especially near the trench.

5.2.1 Refined computational mesh

For the analysis in this chapter, accurate seabed bathymetry data are used and combined with the seabed deformation representation to run VOLNA and obtain the wave elevations for different hypothetical event cases. Information about the bathymetry at each node of a very dense detailed unstructured triangular mesh, which is shown in Figure 5.2 is available, where the number of nodes in the mesh is 1,197,384 corresponding to 2,392,352 triangles.

The public available National Oceanic and Atmospheric Administration and National Geophysical Data Centre (NOAA/NGDC) bathymetry and topography data sets have been con-

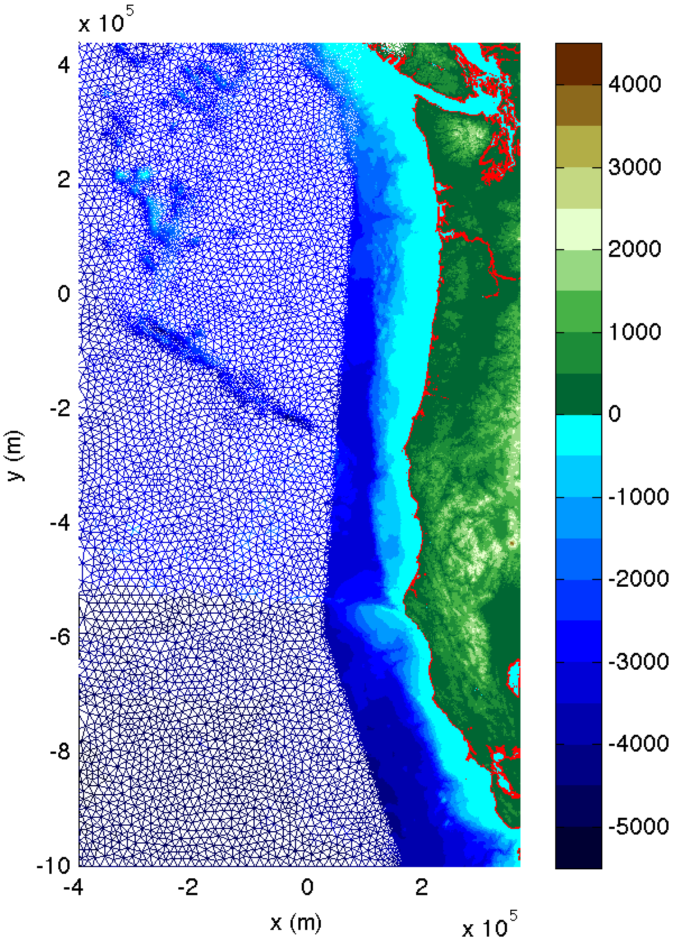


Figure 5.2: Triangular mesh of the Cascadia Subduction Zone, with the coastline indicated by the red colour. The colour scale presents the public available elevation data for the bathymetry and topography.

verted to an adaptive mesh that can be used in VOLNA tsunami model. This conversion was performed by UCL PhD student Xiaoyu Liu and the bathymetry data were kindly provided to me. The NOAA/NGDC data sets are the best available bathymetric and topographic data and they are extracted from the US Geological Survey (USGS) National Elevation Dataset (NED) [Gesch et al., 2014]. NED consists of elevation information[3] and it is widely used for earth science studies as well as mapping applications. NED is derived from the highest quality USGS digital elevation models (DEMs) or three-dimensional representation of the land's surface. NED is generated using land elevation data for the United States, Alaska, Hawaii, Mexico and Canada. DEMs can be obtained for any place in Earth at several resolutions. NED provides national US coverage at a grid spacing of 1 arc per second, which is approximately 30 metres.

---

[3] Elevation or geometric height is the height above or below a fixed reference point, usually the Earth's sea level. The elevation data are collected by aerial or bathymetric surveys and they contain information for the Earth's surface on land and under water.

When new data become available, NED is updated. Also, it is worthy mentioned that NED data are in public domain and they are used for mapping and visualisation, global change research, resource monitoring as well as many other applications [Gesch et al., 2014].

For the generation of the mesh used in the analysis in this chapter, different resolutions of DEMs have been used and they are merged to cover the whole Cascadia Subduction Zone area. Specifically, five different data sources are used, which are the following:

1. ETOPO1 Global Relief Model 1 arc-minute
2. Strait of Juan de Fuca, WA 30 arc-second MHW DEM
3. Strait of Juan de Fuca, WA 5 arc-second MHW DEM
4. US Coastal Relief Model - Northwest Pacific 3 arc-sec
5. US Coastal Relief Model - Central Pacific 3 arc-sec

The cell sizes in the mesh have been calculated using a specific rule which combines the water depth and the slope angle so that the mesh triangles sizes to be smaller at the locations where the waves are travelling more slowly and therefore greater computational accuracy is required. Additionally, for areas where only coarse bathymetry data are available, the mesh size has to be relatively large, whereas in cases where higher resolution data are available, the mesh size is much smaller.

Looking at the generation of the mesh in more detail, the processed followed by Xiaoyu Liu was the following:

1. For areas covered by high resolution DEMs, which are the Strait of Juan de Fuca, WA 5 arc-second, US Coastal Relief Model - Northwest Pacific 3 arc-sec and US Coastal Relief Model - Central Pacific 3 arc-sec, a much more dense mesh is generated compared to lower resolution DEMs areas, which are the ETOPO1 Global Relief Model 1 arc-minute and Strait of Juan de Fuca, WA 30 arc-second MHW DEM.
2. For the case of high resolution DEMs areas:
    - For the water surface area:
        - if the water depth is greater than 250m, which is the area far away from the shoreline and close to the boundaries of the domain, then the mesh cell size is set to be 1000m
        - if the water depth is smaller than 10m, which is the area close to the shoreline, the cell size is set to be 200m

    - if the water depth is between 10m and 250m, the cell size is obtained by linear interpolation
  - For the coast area:
    - if the topography height is greater than 50m above the sea level, then the cell size is set to be 1000m
    - if the height above the sea level is less then 10m, which is for locations around the coastline, the cell size is set to be 200m
    - if the topography height is between 10m and 50m above the sea level, then linear interpolation is used to obtain the mess cell size

3. Similarly, for the areas covered by low resolution DEMs:
   - For the sea area:
     - if the water depth is greater than 3000m, the mesh cell size is set to be 15000m
     - if the water depth is smaller than 150m or when the bathymetry slope is greater than the 90% upper percentile of all the slopes, then the cell size is set to be 3000m
     - if the water depth is between 150m and 3000m, the cell size is obtained by linear interpolation
   - For the coast area:
     - if the topography is greater than 100m above the sea level, the cell size is set to be 15000m
     - if the topography height is less then 50m, the cell size is set to be 3000m
     - if the topography height is between 50m and 100m above the sea level, then linear interpolation is used to obtain the mess cell size

Therefore, the mesh is very dense close to the coastline and by moving away from the coast to the boundaries it is getting more sparse, in order to avoid unnecessary detail at locations that are not important. Hence, the particular way of mesh generation saves computational cost. On the other hand, the adaptive mesh gives high resolution at coastal areas and hence more accuracy in inundation predictions, where is the locations we are interested more as they are densely populated.

However, since there are no available data for the coastline location, the coastline coordinates have been identified by selecting the triangles that include both sea and coastline. This is done by checking the bathymetry and topography at all the nodes and collect the triangles that have nodes which have both positive and negative bathymetry. From these triangles, the centres coordinates are obtained and plotted with red nodes in Figure 5.2. This indicates that the coastline location algorithm is not 100% accurate, but it is still very close and it is enough for the purpose of the analysis as no further detail is necessary now. Note that the triangular mesh defines the computational domain for VOLNA. Therefore, it has too be significantly large in order to avoid wave reflections at the boundaries of the domain.

5.2.2 Domain extension to avoid reflection effects

At the early stages of the analysis, the computational mesh used was smaller in dimensions and then it was replaced by a significantly larger one. The small dimension of the initial mesh used was generating problems with reflections at the boundaries of the domain. In order to avoid this issue, a sloped bathymetry was generated. The slope was able to cause delay at the reflections by slowing down the propagation waves. This is because for shallow water the tsunami waves speed of propagation is proportional to $\sqrt{gh}$, where g is the acceleration of gravity and h is the water depth. This means that the smaller the water depth, the smaller the speed of propagation. Therefore, no reflections were obtained for the time the wave elevation has been investigated and used for further analysis. This extension in the mesh is shown in Figure 5.3.

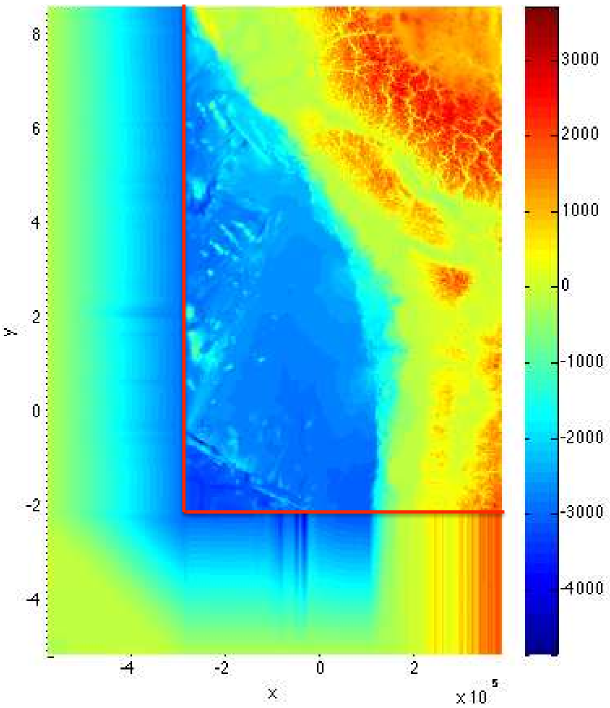


Figure 5.3: The initial small mesh used for the analysis is shown in the red box. This mesh is extended to avoid the reflection issues at the domain boundaries, as it is shown in the Figure.

The bathymetry is extended with a synthetic slope at the sides where the reflection occurs, which are the West and South boundaries. The process for the domain extension was the following: first of all, the mean value of the bathymetry at these two boundaries, excluding coastal bathymetry, is calculated. Then the bathymetry at the West and South boundaries has been replaced with the mean value. Additionally, these boundaries have been extended further and the new boundaries are given a fixed bathymetry amplitude 200m below the water surface. Specifically, the bathymetry is extended from $-296\mathrm{km}$ to $-580\mathrm{km}$ at the West boundary and from $-233\mathrm{km}$ to $-518\mathrm{km}$ at the South boundary. Following that, a straight line slope is created for representing the bathymetry between the "old" and "new" boundaries, which is basically a linear interpolation between the bathymetry at the "old" boundary. It takes the mean bathymetry value at the "old" boundary and the bathymetry equal to $-200\mathrm{m}$ at the "new" boundary.

The larger mesh that is used further in the analysis and have replaced the smaller extended mesh is shown in Figure 5.2. This mesh is large enough in order to avoid boundary reflections for the whole time where the wave elevation was observed. Nevertheless, in the case that a much larger time domain was necessary, probably reflections at the boundaries would have been observed and in this case the use of a sloping bathymetry method would be necessary.

5.2.3 Polygons for accurate realistic representation of displacement geometry

In addition to the realistic bathymetry and topography data, the realistic representation of the coseismic deformation is a very important aspect for modelling an earthquake that is capable of generating a tsunami hazard event. For the investigation performed, the trace of the deformation area, that is shown in Figure 2.2, is defined using arbitrary shaped 4-sided polygons. Specifically, 56 polygons have been used, which are set up into fourteen rows of four polygons as it shown in Figure 5.4.

It is assumed that the deformation consists of both uplifted and subsided parts, which are represented using quadratic curves as described in Chapter 4. Each row of polygons is considered separately, having exactly the same shape but different uplift and subside heights. The middle rows of the deformation, as shown in Figure 5.4 have larger uplift and subside heights compared to the upper and lower rows. The surface of neighbourhood rows are combined with straight lines in order to have smooth surfaces and avoid discontinuities in the uplift between the polygons. The deformation modelling in this chapter, using uplift and subsidence surfaces, attempts to mimic as close as possible the real source deformation considering sediment's displacement in the case of an earthquake event, which is presented in Figure 2.1.

One of the first investigations on the effect of sediment displacement near the trench was performed by Tanioka and Seno [2001]. The authors explore the effect of an additional uplift

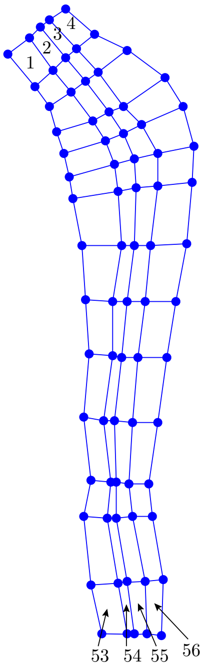


Figure 5.4: The real shape of the deformation area and the locations of the seabed uplift and subsidence are captured by 56 polygons.

of the sediment, focusing on the Sanriku event, 1896. The specific earthquake event resulted in a tsunami larger than expected from the weak seismic waves occurred. Generally, earthquakes like this are known as "tsunami earthquakes". Tanioka and Seno [2001] concluded that this additional sediments uplift near the trench, due to a large seabed coseismic horizontal movement, has a large effect on the tsunami generation and that the combination of the uplift with the slow rupture probagation results to tsunami earthquakes.

Unfortunately, apart for the limited information about the deformation region and the sediment displacement, no other aspects that could help with the description of the source's shape are known. Therefore, the analysis at this chapter attempts to define the source using a simple deformation in a way that represents the reality in the best possible way. The highest uncertainty on the source description is the trench uplift and location. In order to account for this uncertainty, the trench height is allowed to vary a lot and no assumption is imposed, contrarily to recent studies in this area. The sediment displacement is represented by an uplift and subsided with both described by quadratic shaped curves. The uplift starts from the trench location and goes to the maximum, with the location of the maximum to be known, and it is shown with the dotted line in Figure 2.2. The subduction part occurs at the location where the mantle, which is located in high temperatures and therefore can be described as an elastic material, moves downwards, when the tectonic plate moves upwards.

Apart from the shape, the propagation of the displacement is totally uncertain. For the analysis, the motion of the seabed deformation is assumed to start from North and moves to the South, with the four polygons of each row to move together, i.e. trench parallel rupture propagation. However, time difference between the initiation of motion at the first row at North and the last row at South varies, with one possibility to be for them and therefore for all the polygons to move together simultaneously. Even the fact that there are many other ways to represent the rupture propagation, for example the middle rows could start moving first, with the North and South parts to follow, the North to South propagation is selected for this analysis. This choice is largely arbitrary as there are no data available so far to justify or disapprove it. It is assumed that each row of blocks move for 150 seconds and the first row from North starts at time zero and then followed by the second from the North row after some time, where this time is the same for each consecutive row and is varied in order to generate different scenarios. When all the polygons complete their motion and stop moving, the rupture takes its final shape. Lay et al. [2005] and Guilbert et al. [2005] described the Sumatra 2004 earthquake coseismic displacement with trench parallel rupture propagation. Similarly, Simons et al. [2011] employed parallel to the trench deformation propagation for the Tohoku-Oki, Japan 2011 earthquake.

The two left columns of polygons are assumed to be uplifted, whereas the two right ones, which are closer to the coastline, are assumed to be subsided. The uplifted part has quadratic curve shape with the maximum being at the middle, which is the intersection of the two left polygons. The left edge is considered as the coseismic trench location. The height at the trench is assumed to take values between zero and the maximum height of the quadratic curve's middle part. The right edge of the quadratic uplift, which is at the right side of the second polygon, is equal to the seabed level. The subsidence is assumed to be equal to 15% of the maximum uplift of the specific row. This percentage could varied between 0% and 30% but the specific percentage is selected as it is considered realistic. Therefore, the row with the largest maximum uplift has also the largest subsidence.

At the beginning of the motion, the polygons $1, 2, 3$ and 4, as shown in Figure 5.4, start to move simultaneously, followed by the next row of polygons which is the below neighbour. Then the below row follows and so on, with the last row of polygons, numbers $53, 54, 55$ and 56 in Figure 5.4 to be the last one initiating motion. By the time the last row of polygons completes the motion, the seabed deformation obtains its final shape. It is assumed that all the rows of polygons move for 150 seconds and by this time they take their final form. The shape of the deformation for each row of polygons is presented in Figure 5.5, where the flexibility of the trench height is presented.

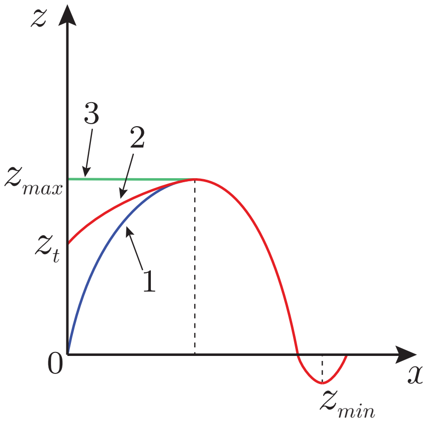


Figure 5.5: Different cases representing bathymetry final form. The uplifted part corresponds to the two columns at the West side of the deformation shown in Figure 5.4, whereas the subsidence part to the two columns of polygons at the East side. Both the maximum and minimum locations are at the intersections of the two West and East columns, respectively.

It is worthy to notice that the shape of the deformation is related to the polygons composition shown in Figure 5.4 in the sense that for each row of polygons, the first polygons to the left $(1, \ldots, 53)$ start from the trench height and goes to the maximum, as shown in the left half of the uplift in Figure 5.5, i.e. from z_t to z_{max}. Similarly, the second polygons from the left $(2, \ldots, 54)$ take the final form as the right half of the uplift in the Figure 5.5, with the uplift height to go from z_{max} to the water surface level. Finally, the left edge of the third $(3, \ldots, 55)$ column of polygons and the right edge of the forth $(4, \ldots, 56)$ column of polygons are always at the water surface level, and the bathymetry subsides to the minimum in the middle, which is the intersection between the two columns of polygons. The shape of the seabed displacement can be observed clearly in Figure 5.6.

5.3 Hypothetical event cases

A large uncertainty in tsunami predictions is induced by using an imprecise description of the coseismic source. It is impossible to accurately represent the source and to be 100% confident about its characteristics. The earthquake source is not deterministic and differs between events. For this reason, different hypothetical earthquake sources that can cause tsunami events are generated by different combinations of three source characteristics in order to represent a range of possibilities. Specifically, forty different combinations are selected using experimental design methods and specifically the “maximin” LHD method, which is described in detail in Section 3.1.2. VOLNA has been run for these forty events and the resulting tsunami wave elevation time series have been used to draw conclusions about how sensitive is the wave elevation at

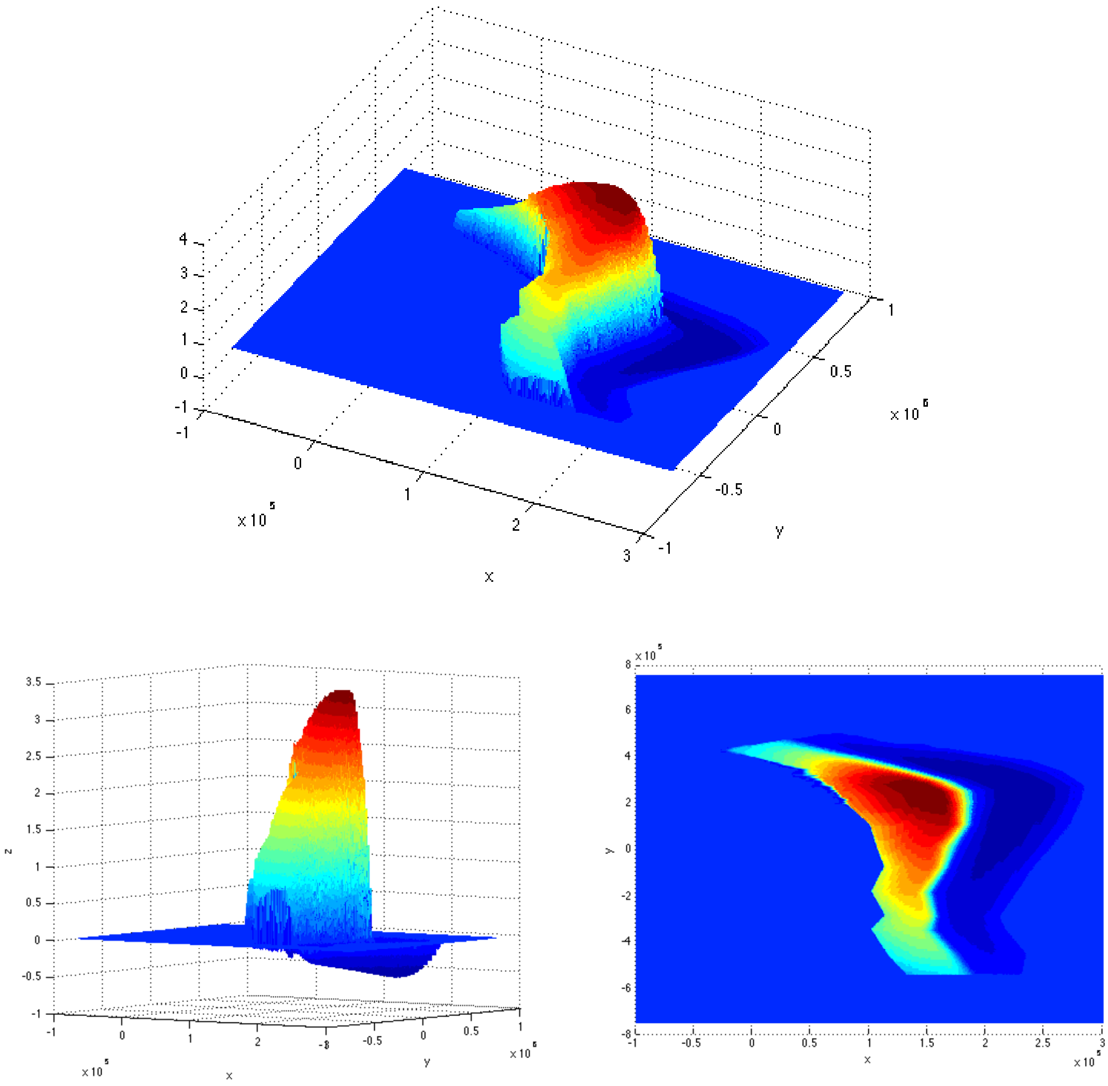


Figure 5.6: Bathymetry deformation final shape presented at three different angles. The specific deformation is for one of the 40 cases of bathymetry deformation investigated in the analysis. Specifically, it is the case where $z_{max} = 1.52$, $z_t/z_{max} = 0.67$ and $t_p/t_r = 6$, where the three parameters are described in detail in Section 5.3.1. The particular event case describes a slow propagated coseismic deformation, with a relatively low maximum uplift and where the trench uplift height is more than half the middle maximum uplift.

each of the source characteristics or at combinations of them.

Furthermore, a statistical emulation has been applied to these evaluations to create an accurate and fast surrogate of the VOLNA computer model, which can be used for fast predictions for any combination of the three characteristics. Hence, an uncertainty analysis can be performed, where a distribution for each input characteristic can be given to the emulator and the distribution of the resulting tsunami wave elevation can be predicted. Therefore, the uncertainty in predictions due to uncertain trigger features can be obtained, with results giving a percentage of confidence for the wave elevation height. Additionally, the emulator can be used for sensitivity analysis. The use of the statistical emulation can make these analyses feasible. The specific analyses are very computationally demanding due to the need of a large number of model evaluations at each of the combination of the source characteristics.

5.3.1 Input parameters generating the different events

Different event cases describing an earthquake that can cause a tsunami event are investigated. The different events are generated by varying three of the source characteristics, which are the following:

1. The maximum source elevation factor, z_{max}, that takes values between 1 and 5. z_{max} is multiplied by the fixed maximum uplift height of each row of polygons, which is located in the middle of the uplifted part. These heights are provided by Dr Simon Day and realistically represent an event based on Wells and Coppersmith [1994].

2. The ratio of the trench elevation height, which is at the left boundary of the source, to the middle maximum height, z_t/z_{max}, which is assumed to be in the range between 0 and 1. It is equal to 0 when the trench height does not change throughout the deformation and stays at the water level. On the other hand, it is equal to 1 when the trench has the same height with the middle maximum. It is possible the ratio to take values greater than 1 that is the case where the trench height is greater than the middle maximum. Nevertheless, it is only likely to be greater than 1 locally and it is almost impossible to take values greater than one at the whole rupture. Since for this analysis the whole rupture propagation scenario is investigated, the specific range for z_t/z_{max} is considered.

3. The ratio of the rupture propagation duration from North to South, t_s, to the rupture propagation time of each row of polygons, t_r, t_p/t_r, where t_r is assumed to be fixed and equal to 150 seconds. t_p is equal to the difference between the starting time of motion of the first row of blocks at North and the initial time of propagation of the last row at South. The minimum value of this parameter is zero and is the case where all the polygons move together. The maximum value of this parameter is selected to be equal to 6, which is the case where the whole rupture propagation takes place in 900 seconds. The particular selection of rupture propagation duration is justified by past event. Specifically, it is known that the rupture propagation times for the Sumatra 2004, Chile 2010 and Tohoku 2011 events were approximately 1000, 100 and 200 seconds, respectively.

5.3.2 Experimental Design

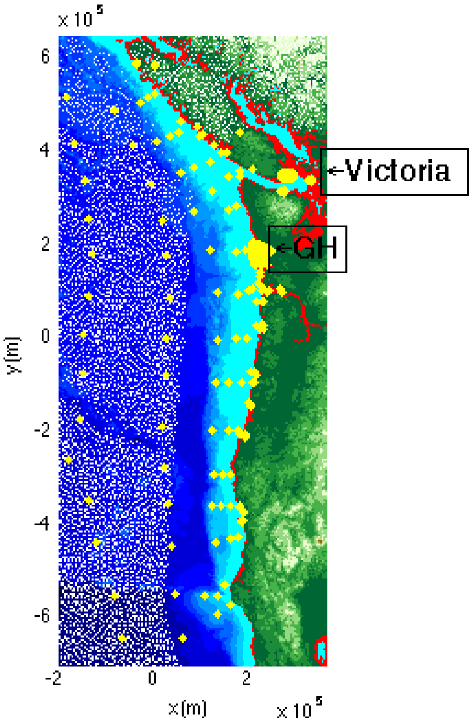
Using the “maximin” Latin Hypercube Design, which is extensively described in Section 3.1.2, 40 combinations of the three input parameters are selected to cover the three-dimensional input space. Each combination corresponds to a different hypothetical seabed deformation event. VOLNA simulations were performed for each of these 40 cases resulting to evaluations for the induced tsunami wave elevation time series at a large number of gauges. The analysis is focused on Victoria and Grays Harbor areas, which are important areas potentially affected by the Cascadia Subduction Zone. They are significant areas due to the fact that Victoria is the capital city with a large population, around 100,000, and Grays Harbor bay area consists of the population with the highest vulnerability.

5.3.3 VOLNA evaluations sensitivity analysis

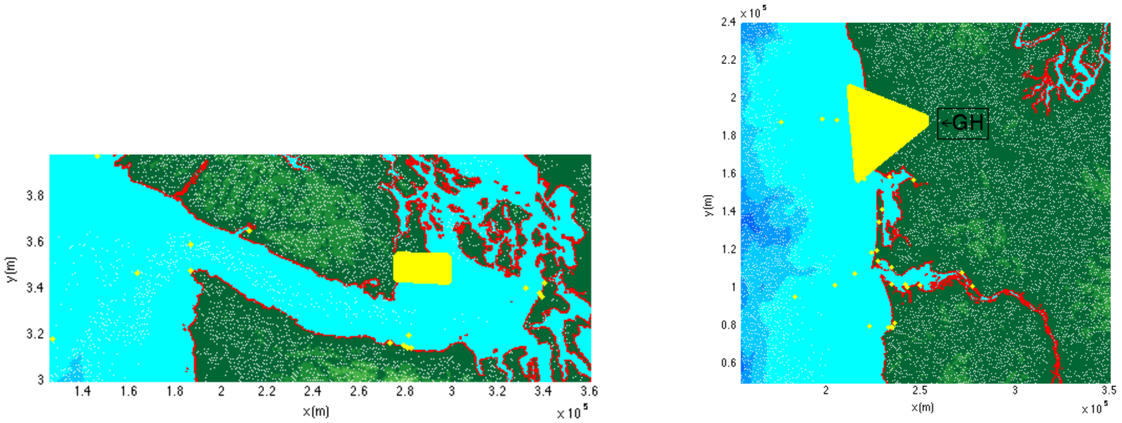
The realistic tsunami model VOLNA [Dutykh et al., 2011] has been evaluated using as inputs the computational mesh, the bathymetry and topography and the dynamic seabed deformation. VOLNA gives as output the tsunami wave elevation time series at a number of gauges. Specifically, the number of gauges considered in the analysis is 33431, where the larger percentage of them are located in Grays Harbor bay area (26879) and at Victoria city area (6379). The gauges at Grays Harbor and Victoria area are located every 200 m. Also, a small number of additional gauges (173) covering the whole area of the Cascadia Subduction Zone are also investigated. The location of the gauges can be obtained in Figure 5.7.

Using VOLNA evaluations, first of all, the impact of the source characteristics on resulting wave elevations has been investigated in order to draw conclusion on which of these characteristic and in which way they are affecting the outputs. The comparison between the different combinations of the three source characteristics for two of the gauges, numbers 141 and 32298, are presented in Figures 5.8 and 5.9, respectively. Also, for each of the two gauges, Figures 5.10 and 5.11 present the wave elevation time series for all the 40 LHD points for different combinations of small/medium/large z_{max} and $\frac{z_t}{z_{max}}$. The ranges for small, medium and large for the two inputs have been obtained from the estimation of the 33% and 67% quantiles. Specifically, a value z_{max} is considered small if it is in the range $[1, 2.33]$, whereas the medium and large values correspond to the ranges $(2.33, 3.7]$ and $(3.7, 5]$, respectively. Similarly, the corresponding ranges for small, medium and large for the input $\frac{z_t}{z_{max}}$ are the following: $[0, 0.34]$, $(0.34, 0.67]$ and $(0.67, 1]$.

From the four figures, it can be concluded that the wave amplitude depends significantly on z_{max}. Specifically, the larger the value of z_{max} the larger the induced wave amplitude. However, the input parameter z_t/z_{max} also affects the output in cases where z_{max} values are



(a) location of all gauges



(b) location of Victoria gauges

(c) location of Grays Harbor gauges

Figure 5.7: Gauges locations shown in yellow colour.

similar. When two sources have very similar maximum uplift height, then the case with the larger z_t/z_{max} is going to result in a larger wave elevation. This indicates that when there is a large discontinuity at the trench, high wave amplitudes are obtained. Finally, in respect to the third input parameter, it can be concluded that the duration of source propagation does not influence crucially the tsunami wave amplitude, whereas, it significantly affects the wave propagation time, in the sense that it moves the location of the maximum elevation (peak) of the time series. More specifically, the smaller the value of t_p/t_r the faster the wave generated

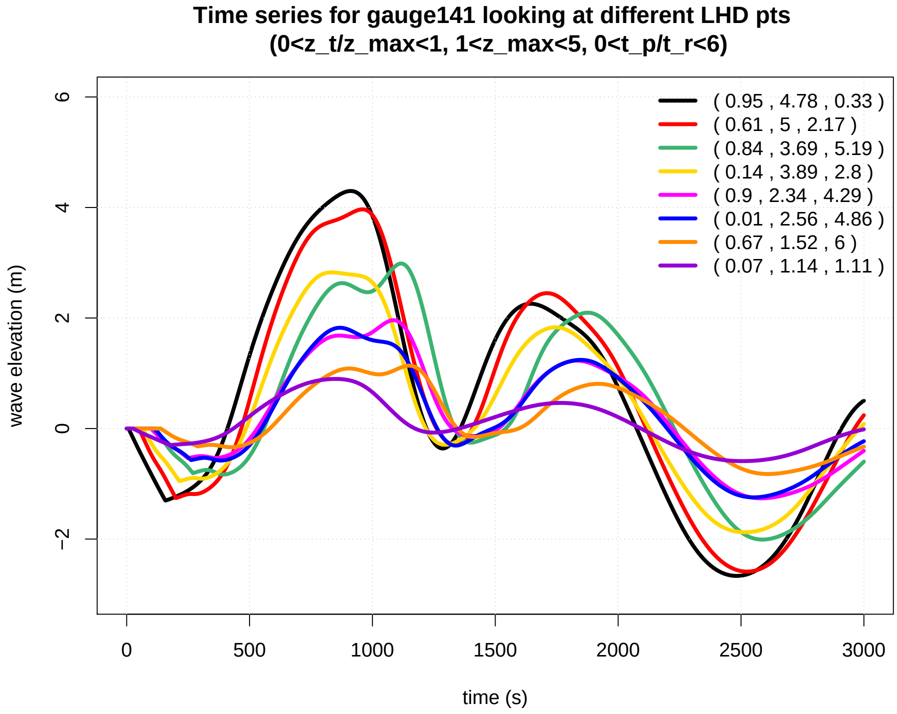


Figure 5.8: Comparison of wave elevations for different sources characteristics for gauge 141 for different combinations of the source characteristics $(z_t/z_{max}, z_{max}, t_p/t_r)$.

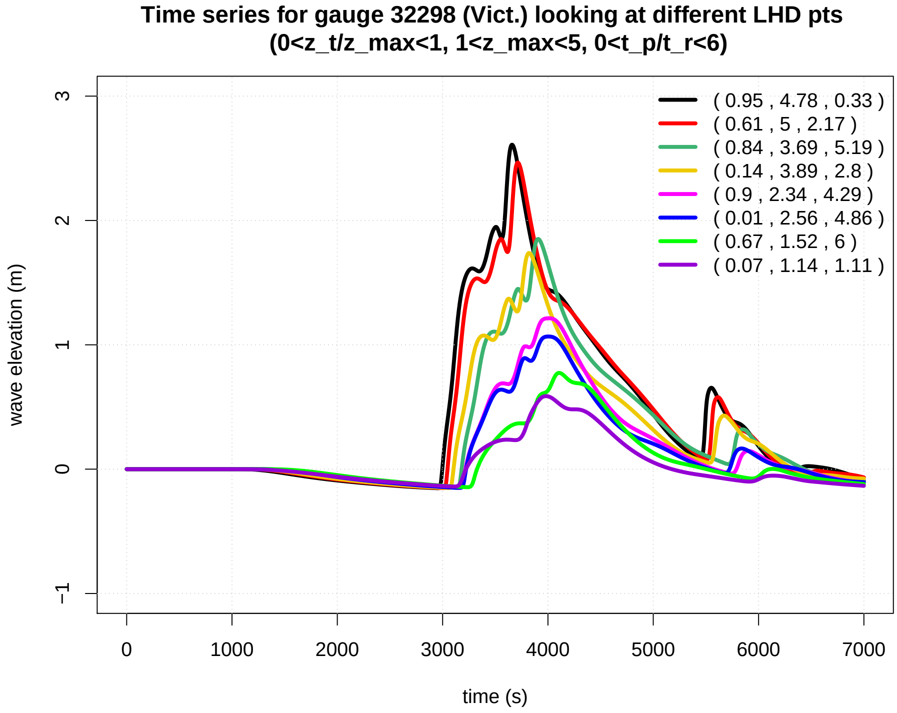


Figure 5.9: Comparison of wave elevations for different sources characteristics for gauge 32298 for different combinations of the source characteristics $(z_t/z_{max}, z_{max}, t_p/t_r)$.

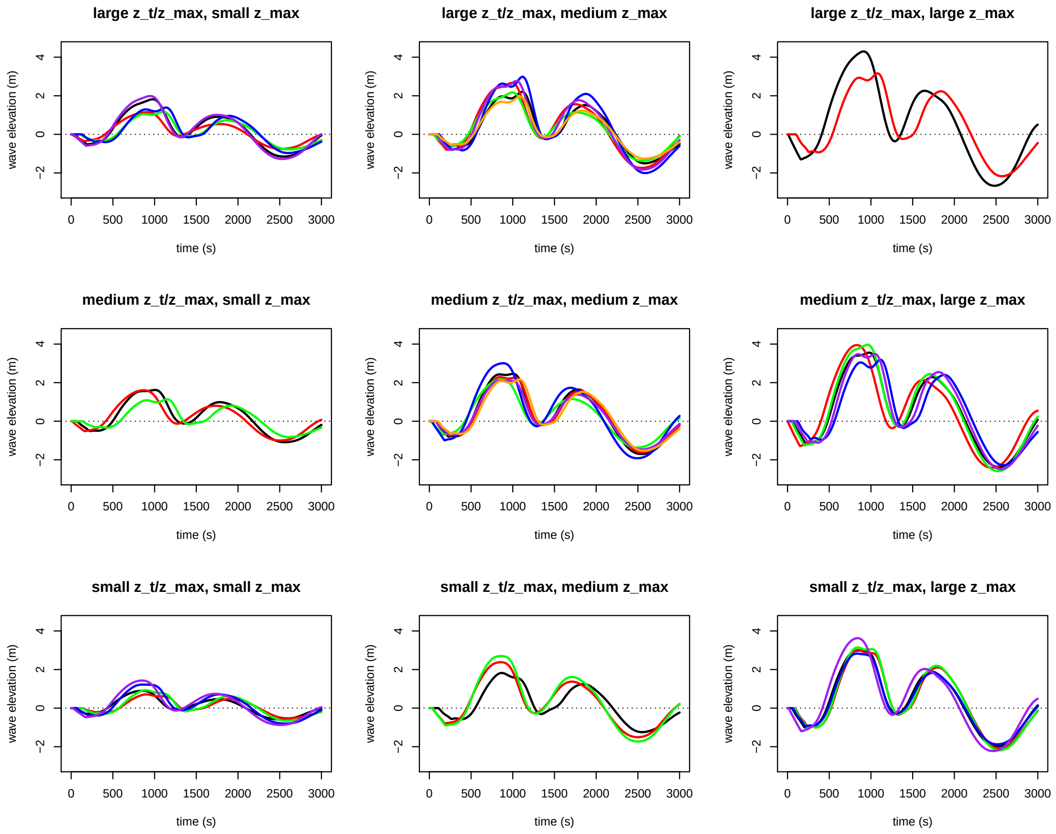


Figure 5.10: Wave elevation time series for all the LHD points investigated, looking at different combinations of small/medium/large z_{max} and $\frac{z_t}{z_{max}}$ for gauge 141. Moving from left to right the value of z_{max} increases, whereas moving from bottom to top the value of $\frac{z_t}{z_{max}}$ increases. The different lines in each plot correspond to different values of the third input parameter t_p/t_r.

arrives at the gauge. This means that in the case where all the blocks consisting the source are moving together, with no delays between North and South parts of the source, then the tsunami waves arrive faster. This conclusion applies to all the gauges where evaluations are available. Additionally, from Figures 5.8, 5.9, 5.10 and 5.11, it is noticeable that the water level significantly decreases after the arrival of the main wave, which is also observed in the analysis in Chapter 3 and it is known from real tsunami event cases.

The next step of the analysis is the application of statistical emulation to the model evaluations. This is going to be described in the next sections, where the high agreement of the resulting emulator's predictions with the actual VOLNA evaluations is presented. Additionally, it is worth mentioning that the time required for just a single evaluation of the VOLNA model is approximately 20 minutes. This highlights the importance of approximating the computer model with a statistical emulator when a large number of locations is going to be investigated.

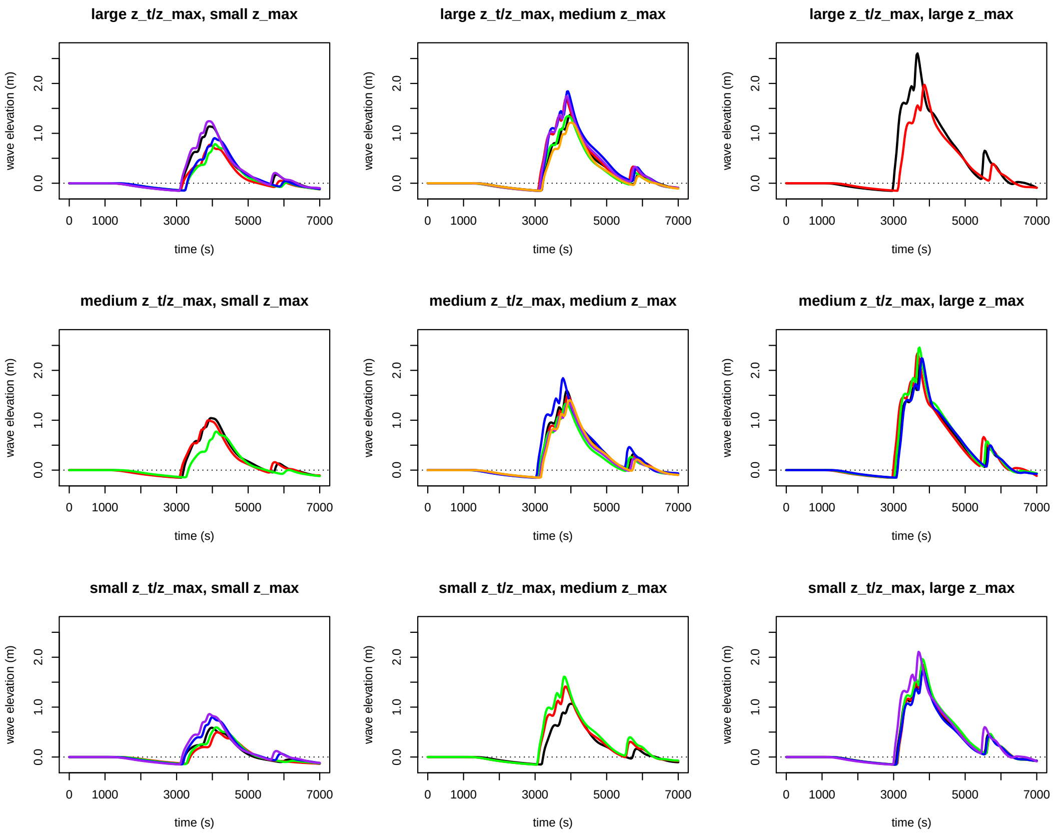


Figure 5.11: Wave elevation time series for all the LHD points investigated, looking at different combinations of small/medium/large z_{max} and $\frac{z_t}{z_{max}}$ for gauge 32298. Moving from left to right the value of z_{max} increases, whereas moving from bottom to top the value of $\frac{z_t}{z_{max}}$ increases. The different lines in each plot correspond to different values of the third input parameter t_p/t_r.

5.4 Statistical Emulation looking at a large number of locations

The dramatic consequences of tsunami events highlight the need to obtain practical measures for mitigating the risk. The prediction of the tsunamigenic risk can be either deterministic or probabilistic. The complexity as well as the limited knowledge of the waves behaviour along the coast, and the difficulty to describe accurately the tsunami source, make the deterministic prediction impossible. This makes the use of Probabilistic Tsunami Hazard Analysis (PTHA) beneficial. Section 2.4.2 in the Background chapter contains more information as well as literature review related to PTHA.

For PTHA a significantly large number of simulations is required. To make this feasible, simplistic models that are far away from being realistic are unfortunately commonly used in literature. A solution to this computational problem is the use of statistical emulators, since they can give a large number of evaluations in a very short time and are able to approximate accurately expensive computer models. A realistic accurate tsunami model, such as VOLNA, would give precise and trustworthy tsunami hazard estimations and hence the analysis and the application of statistical emulation can be focused on realistic cases for hazard mitigation.

For this analysis VOLNA evaluations are obtained for a large number of gauges for realistic hypothetical events. Therefore, a statistical emulator is going to be generated to obtain fast predictions that can be used further for the final step of PTHA. Specifically, an Outer Product Emulator has been created resulting in a very satisfactory performance in predictions. The novel step in this analysis is the application of Landmark Registration to the wave elevation evaluations and the use of Functional Principal Components Analysis in order to select representative regression functions for the output that are able to capture the behaviour of the tsunami wave with respect to time for all the gauges.

5.4.1 Comparison of different techniques of emulation

This section demonstrates the application of statistical emulation to the induced tsunami wave elevation time series. Specifically, four different techniques of emulations have been performed and compared. The first one was exactly the same as it is described in Chapter 3 where statistical emulation has been applied to Sammarco and Renzi [2008] landslide-generated tsunami model evaluations. In this case, Fourier terms are used for the output regression functions. The second technique of emulation, uses exactly the same prior selections for functions and parameters as in the first one, but with the only difference that the prior choices are combined with the registered model evaluations, where Landmark Registration has been applied initially to transform the data. The third method uses the unregistered data, similarly to the first method. However, Functional Principal Component Analysis has been applied and the first two principal components are selected and used in place of the Fourier basis for the output regression functions. The final technique of emulation uses a combination of Principal Components and registered data. The predictions of the four emulation techniques are compared looking at the diagnostic plots and some statistics. In conclusion, the combination of data registration with the use of Principal Components (PCs) for the regression functions leads to improvements in the predictions.

Earlier in the Mathematical Methods section, an extensive description and literature review related to the methods of Landmark Registration, PCs as well as FPCA that are used in this section, are introduced. The comparison of the three techniques of emulation in this section is presented for two specific gauges in the Cascadia Subduction Zone. Following that, the statistical emulation using both registration and Functional Principal Components analysis, which it is proved to be the best from the four, is applied simultaneously to a large number of gauges over the Victoria area giving satisfactory predictions for all the locations. This research is still in progress with plans to be able to apply statistical emulation for a large number gauges covering the whole domain. This will give us the opportunity to perform uncertainty analysis in the whole domain.

5.4.2 Application of Registration and FPCA to the emulation process

Statistical emulation has been applied to the wave elevation time series at different gauges and it is presented here for gauges number 141 and 32298, which are shown in red and green dots in Figure 5.12. These specific gauges are selected as their wave elevation time series have very different shape, even though they are both located in the Victoria area.

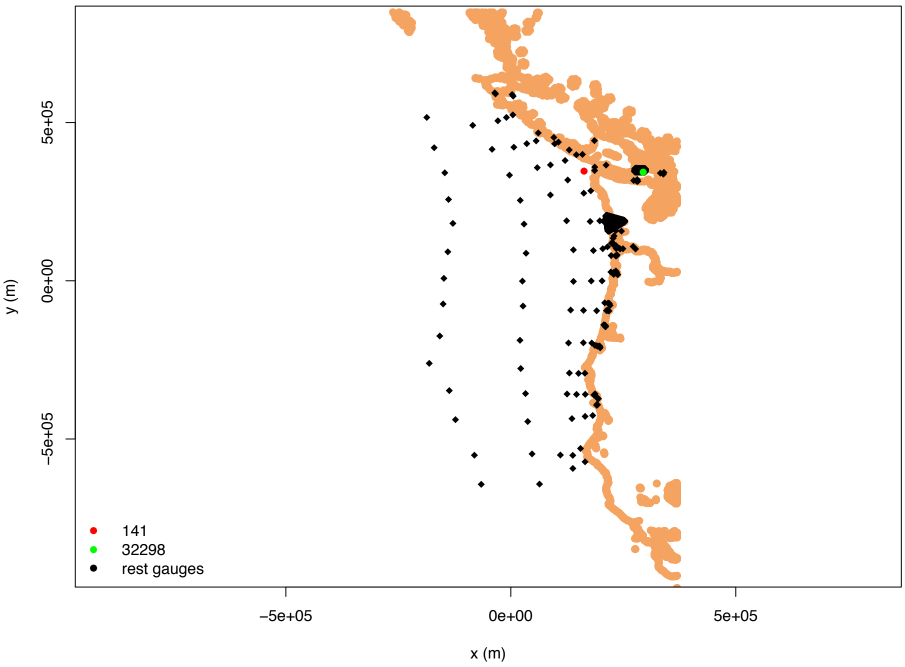


Figure 5.12: Location of the 33431 gauges where VOLNA evaluations obtained. The two gauges with green and red colours are the ones that we are going to focus on in the rest of the analysis in this chapter presenting the results for these two locations.

Outer Product Emulation is used, which was described in detail in Chapter 3. Four different variations of the emulation process are performed and their predictions are compared. The first one uses Fourier basis functions for the regression functions and is similar to the one applied at Sammarco and Renzi wave elevation time series in Chapter 3. The second process uses Fourier basis functions as well, but on registered model's evaluations. The third and forth processes investigated use PCs in place of the Fourier basis functions on unregistered and registered data, respectively. Since the wave elevations do not oscillate periodically, the use of Fourier basis describing the wave behaviour might not be the best idea, but is nevertheless a commonly use approximation of wave elevation phenomena.

Model evaluations are available for 40 combinations of the three source characteristics described above, $r = (z_t/z_{max}, z_{max}, t_p/t_r)$, with each one representing a hypothetical earthquake event that is able to generate a tsunami. These 40 points are chosen to cover the three-dimensional input parameter space using "maximin" LHD, with the following domain to selected: $\frac{z_t}{z_{max}} \in [0, 1]$, $z_{max} \in [1, 5]$ and $\frac{t_p}{t_r} \in [0, 6]$. The location of these 40 input points in the parameter space is shown in Figure 5.13.

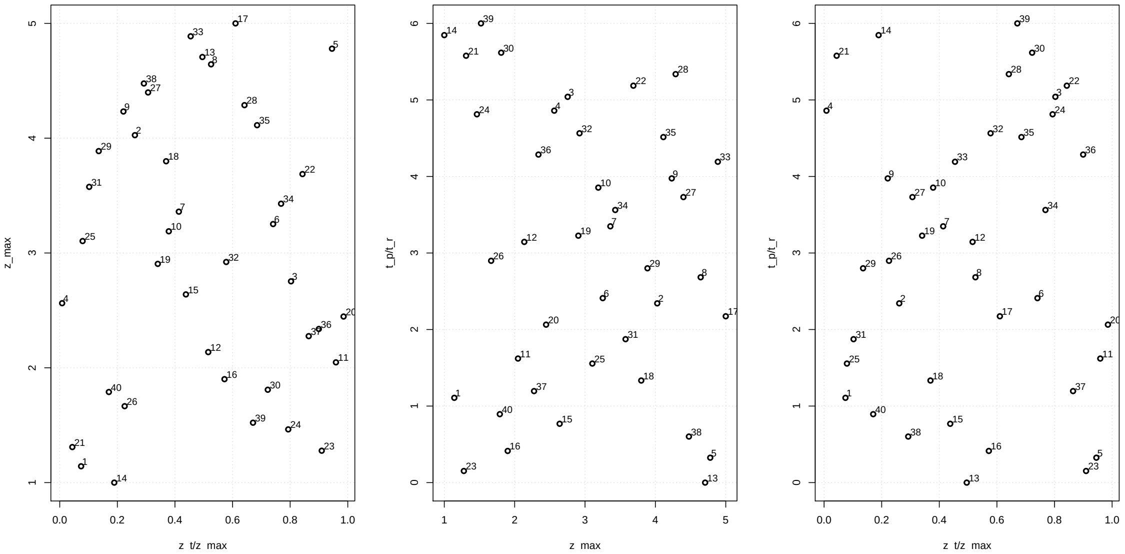


Figure 5.13: Location of the 40 inputs that are selected using "maximin" LHD method. They are combinations of three source characteristics.

For gauge 141 evaluations are obtained for time up to 3000 seconds with 1 second time steps, whereas for gauge 32298 for time up to 7000 seconds with 1 second time step as well. However, not all the evaluations are used and given to the prior emulator, since a large number of time steps would make the emulation process computationally expensive. Specifically, evaluations every 9.9 and 27.9 seconds are considered for gauges 141 and 32298, respectively. On the other hand, a sufficient number of time steps is required to capture the wave behaviour. The emulator is constructed looking at fixed specific locations, which means that the output is only time-dependent.

For the emulation process, appropriate choices of regression and covariance functions for inputs r and output s, where s equals to time, are made, as described in Chapter 3. The set of input regression functions, $G^r \triangleq \{g_1^r, \ldots, g_{\nu_r}^r\}$, where ν_r is the number input regressors, contains appropriate choices of polynomials for each of the three input parameters. The number of input regressors is seven; a linear and quadratic polynomial for each input is selected, plus an additional constant term. Higher order polynomials are not required since the wave elevation is smooth with respect to the inputs. As in the case of Sammarco and Renzi model evaluations, the polynomials are shifted into the unit integral $[0,1]$, with the selection of the coefficients to make the two functions for each input to be orthonormal with the uniform weighted function. The set of the selected input regressors for both the gauges presented is the following:

$$\begin{aligned}G^r = \{&1, \sqrt{3}(z_t/z_{max}), -3\sqrt{5}(z_t/z_{max}) + 4\sqrt{5}(z_t/z_{max})^2, \sqrt{3}(\frac{z_{max}-1}{4}),\\ &-3\sqrt{5}(\frac{z_{max}-1}{4}) + 4\sqrt{5}(\frac{z_{max}-1}{4})^2, \sqrt{3}(\frac{t_p/t_r}{6}), -3\sqrt{5}(\frac{t_p/t_r}{6})\\ &+ 4\sqrt{5}(\frac{t_p/t_r}{6})^2\}\end{aligned} \tag{5.6}$$

After the selections of inputs regression functions, the next step is the appropriate choice for the output regression functions, $G^s \triangleq \{g_1^s, \ldots, g_{\nu_s}^s\}$, where ν_s is the number of output regressors. For the first technique of emulation, Fourier terms are used for G^s in addition to a constant term. The Fourier terms are of the form $\sin(\frac{2\pi t}{T})$ and $\cos(\frac{2\pi t}{T})$, where T is the period of the oscillation. Observing the wave time series for the different inputs and using some trial-and-error tests, the set of frequencies that leads to good predictions and it is as small as possible is selected. The selection of these frequencies is different for each gauge as they oscillate differently. Specifically, for gauge 141 the set of frequencies is the following:

$$\{\frac{1}{1600}, \frac{1}{1400}, \frac{1}{1200}, \frac{1}{1000}, \frac{1}{800}, \frac{1}{600}, \frac{1}{400}\} \tag{5.7}$$

Hence, the set of output regression functions is the following:

$$\begin{aligned}G^s = \{&1, \sin(\pi T/800), \cos(\pi T/800), \sin(\pi T/700), \cos(\pi T/700),\\ &\sin(\pi T/600), \cos(\pi T/600), \sin(\pi T/500), \cos(\pi T/500),\\ &\sin(\pi T/400), \cos(\pi T/400), \sin(\pi T/300), \cos(\pi T/300),\\ &\sin(\pi T/200), \cos(\pi T/200)\}\end{aligned} \tag{5.8}$$

The corresponding selection of frequencies for gauge 32298 is

$$\{\frac{1}{6400}, \frac{1}{5800}, \frac{1}{5200}, \frac{1}{4600}, \frac{1}{4000}, \frac{1}{3400}, \frac{1}{2800}, \frac{1}{2200}, \frac{1}{1600}, \frac{1}{1000}, \frac{1}{400}\} \tag{5.9}$$

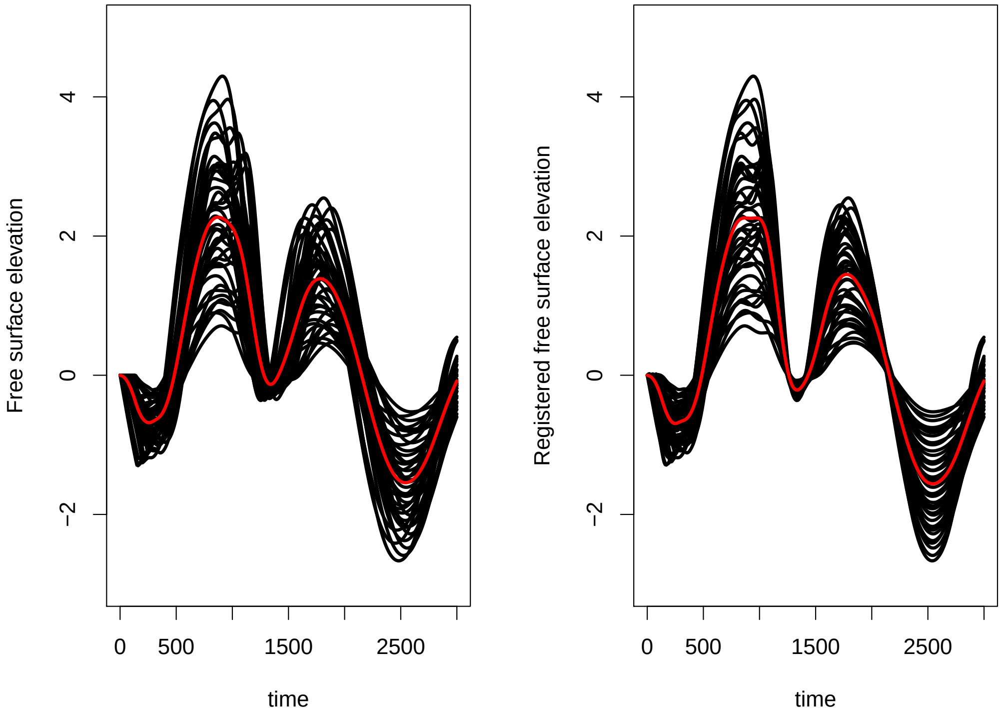
resulting to the following set of regression functions for the output

$$
\begin{aligned}
G^s = \{&1, \sin(\pi T/3200), \cos(\pi T/3200), \sin(\pi T/2900), \cos(\pi T/2900),\\
&\sin(\pi T/2600), \cos(\pi T/2600), \sin(\pi T/2300), \cos(\pi T/2300), \sin(\pi T/2000), \cos(\pi T/2000),\\
&\sin(\pi T/1700), \cos(\pi T/1700), \sin(\pi T/1400), \cos(\pi T/1400), \sin(\pi T/1100), \cos(\pi T/1100),\\
&\sin(\pi T/800), \cos(\pi T/800), \sin(\pi T/500), \cos(\pi T/500), \sin(\pi T/200), \cos(\pi T/200)\}
\end{aligned}
\tag{5.10}
$$

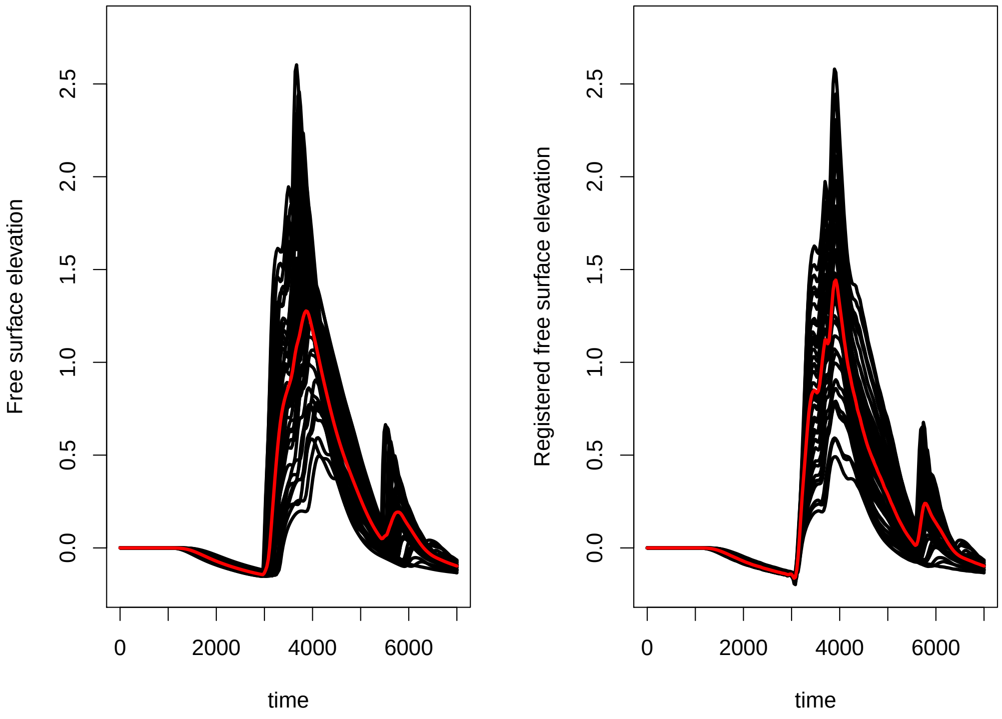
Generally, the application of registration to the data in order to remove all the phase variations can possible improve the analysis of the data. Therefore, a landmark registration is applied to the data and the same selection of function is used combined with the registered data. Looking in more detail the application of landmark registration at the two gauges presented here, for the first one, gauge 141, two landmarks are selected. Specifically, the locations of the second and third minimum of the times series are selected as landmarks. For the other gauge number 32298, two landmarks are selected as well, which are the locations of the first maximum and first minimum of the time series. The effect of the landmark registration at the two gauges is presented in Figure 5.14. It can be concluded that more detail of the wave elevation time series is captured by the mean curves (red lines) after the registration.

After the landmark registration, for both the gauges, the frequencies that describe the wave elevation oscillations are selected and there are the same as for the non-registered data. This is because the registration does not change a lot the frequency of oscillations in the curves, but it moves slightly the landmarks of all the curves to be at the same location. The emulation process is then carried out using Fourier functions with registered data.

However, since the wave elevations vary a lot among the gauges as well as among the different scenarios, where amplitude variations are obvious at many cases, the choice of a fixed set of basis functions for the output regression functions might not be the best idea. Specifically, the choice of Fourier basis functions is not that representative for the wave elevation evaluations, as in some cases the whole time series or some parts of them do not have wave shape form. Also, the choice of Fourier basis functions require manual care and the process cannot be automatic in case where simultaneous statistical emulation at many locations is desired. Furthermore, usually the number of Fourier basis functions selected is quite large in order to capture all the range oscillation frequencies. To overcome these issues, Principal Components are selected and used in place of the Fourier basis for the input regression functions, for another technique of emulation examined in the analysis. But having in mind that the PCA gives better results



(a) gauge 141



(b) gauge 32298

Figure 5.14: Effect of landmark registration on the wave elevation time series for (a) gauge 141 and (b) gauge 33298. The time series are aligned at selected landmarks. The red line shows the mean value of all the curves for different inputs.

for phase aligned curves, a landmark registration is applied first and then the PCs are used. However, the case where PCs are used in place of the Fourier term with unregistered data is also investigated. Similarly to this analysis, Bayarri et al. (2007) and Higdon et al (2008) applied emulation and calibration on functional data that are outputs of computer models using wavelet and Principal Components, but not Functional Principal Components and also no registration was incorporated.

Functional PCA is performed on both the unregistered and registered wave time series evaluations. The principal components that describe the main variations in the curves are selected and together with the mean value curve, which is the mean value of all the curves (the 39, since the wave elevation time series that is tried to be predicted is not included in the estimation of the mean value), at all the time steps, they replace the Fourier terms for the output regression function. Note that the mean value curve is added in the set of output regression functions at the other two emulation techniques in order to check whether or not the predictions of the resulting emulator improved. The conclusion is that by including the mean value curve in the Fourier regression functions set, the emulator predictions are not improved, but they are even getting worse.

For both gauges, it is observed that the first two principal components collected are enough to describe the variations in curves. For the gauge number 141, the two eigenfunctions or harmonics are shown in Figure 5.15. Since the principal components represent variations around the mean, the mean curve is plotted with the +'s and -'s that correspond to the consequences of adding and subtracting a small amount of each principal component. Specifically, the +'s and -'s curves are described by the following equation:

$$mean \pm PC \times se(PC) \tag{5.11}$$

where large differences between the + and - curves indicates large variations of the 39 curves at these time steps.

For the first principal component, larger variations are obtained at the maxima and minima. The second harmonic shows phase variations and also indicates variations only after 750 seconds from the beginning of the motion. The first two PCs capture 94.3% of the variations in the resulting tsunami wave elevation with the second PC accounting for a significantly large percentage of the variations (27.7%).

Looking at the second gauge, the two eigenfunctions are shown in Figure 5.16. The first principal components accounts for a large percentage of variation in the data, 97.6%, and shows

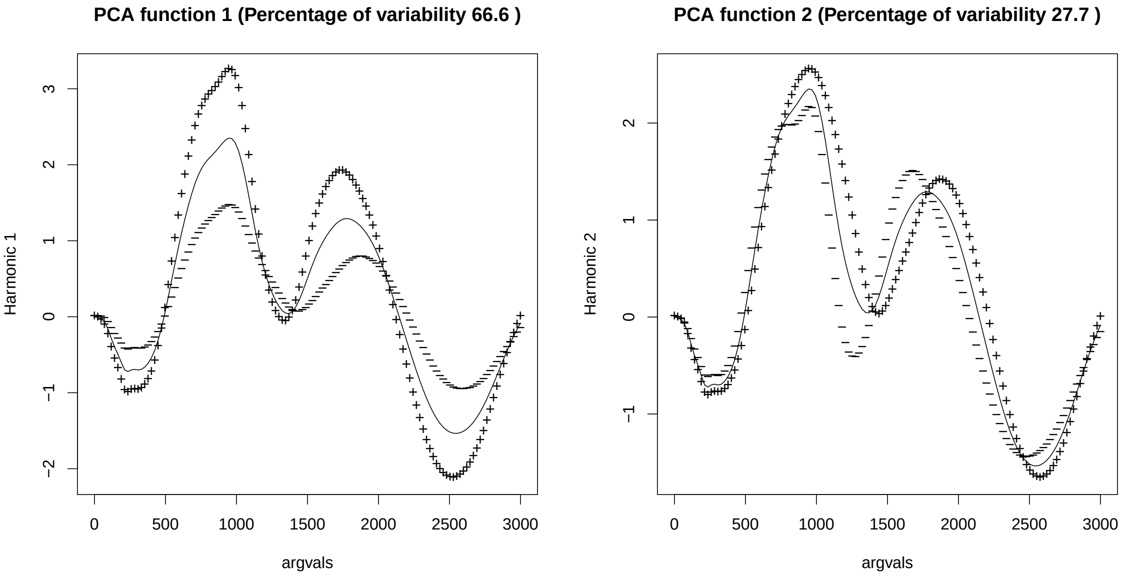


Figure 5.15: The solid line in the two panels is the mean wave elevation for gauge 141. Each principal component is plotted in terms of its effect when a multiple of it is added (+) and subtracted (-) from the mean curve.

very small variability initially up to time 300 seconds, where the wave amplitude varies very little, and then when the first incoming wave arrives and there are large variations at the maximum positions. The second principal components represents just 1% of the variation and does not provide lot of information.

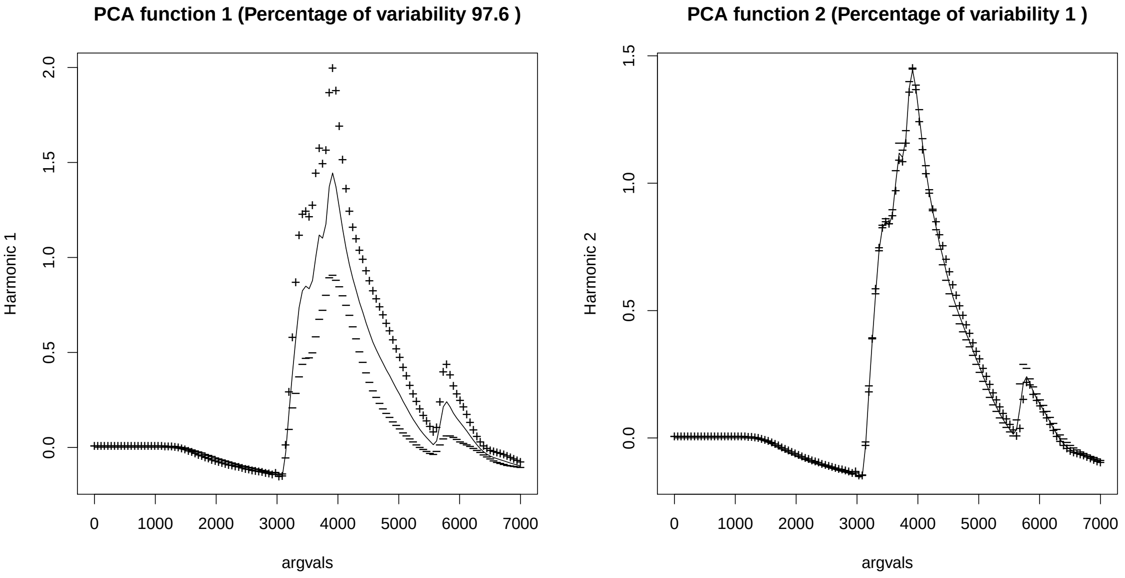


Figure 5.16: The solid line in the two panels is the mean wave elevation for gauge 32298. Each principal component is plotted in terms of its effect when a multiple of it is added (+) and subtracted (-) from the mean curve.

Following the selection of regression functions, the input and output residuals covariance functions, κ^r and κ^s, have to be selected. The choice is power exponential functions is the same for all the four techniques of emulation and for the two gauges. They are the following:

$$\begin{aligned}\kappa^r &= \exp(-(\frac{|\frac{z_t}{z_{max}} - \frac{z_t}{z_{max}}'|}{\lambda_1})^{3/2}) \times \exp(-(\frac{|z_{max} - z'_{max}|}{\lambda_2})^{3/2}) \\ &\times \exp(-(\frac{|\frac{t_p}{t_r} - \frac{t_p}{t+r}'|}{\lambda_3})^{3/2})\end{aligned} \tag{5.12}$$

$$\kappa^s = \exp(-(\frac{|t - t'|}{\lambda_t})^{3/2}) \tag{5.13}$$

where λ_1, λ_2, λ_3, represent the correlation lengths for inputs and λ_t the output correlation length. The choice of the $3/2$ power is as explained in Chapter 3, to have enough smoothness.

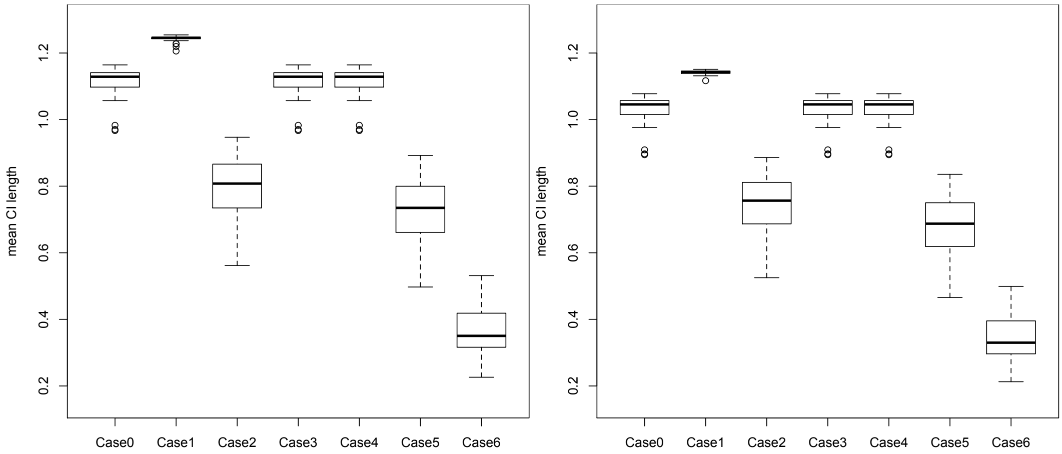
5.4.2.1 Selection of correlation lengths

The values of the correlation lengths are able to adjust the fit of the emulator and for this analysis they are selected using trial and error methods. Specifically, different combinations of λ_1, λ_2, λ_3 and λ_t have been tried and the resulting emulator's predictions are validated and compared using the total mean CI length and RMSE. These comparisons have been applied to the gauges 141 and 32298 for all the four techniques of emulators concluding to a set of correlation length values that is commonly used for the emulation at all the gauges investigated.

For this investigation, the correlation lengths λ_1, λ_2 and λ_3 are considered to take the same values for all the combinations of the four correlation lengths. Specifically, the following cases of correlation lengths combinations have been investigated:

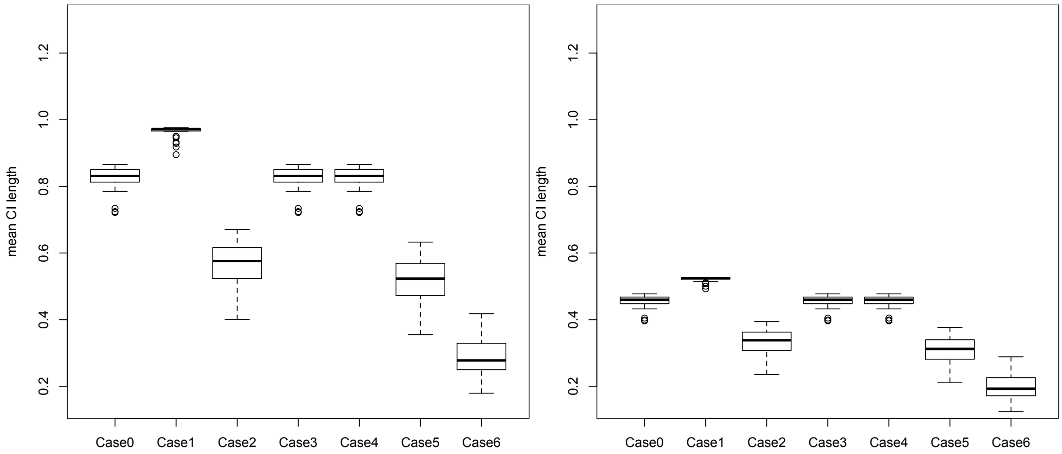
1. <u>Case 0</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$ and $\lambda_t = 0.5$
2. <u>Case 1</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$ and $\lambda_t = 0.5$
3. <u>Case 2</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 0.9$ and $\lambda_t = 0.5$
4. <u>Case 3</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$ and $\lambda_t = 0.1$
5. <u>Case 4</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 0.5$ and $\lambda_t = 0.9$
6. <u>Case 5</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 1$ and $\lambda_t = 0.5$
7. <u>Case 6</u>: $\lambda_1 = \lambda_2 = \lambda_3 = 2$ and $\lambda_t = 0.5$

The mean CI length and RMSE for the predictions of all the four emulation techniques for the different correlation lengths combinations are shown in Figures 5.17 and 5.18, respectively, for the gauge 141. Similarly, the corresponding Figures for the gauge number 32298 are shown in Figures 5.19 and 5.20.



(a) Fourier basis regression functions on unregistered data

(b) Fourier basis regression functions on registered data

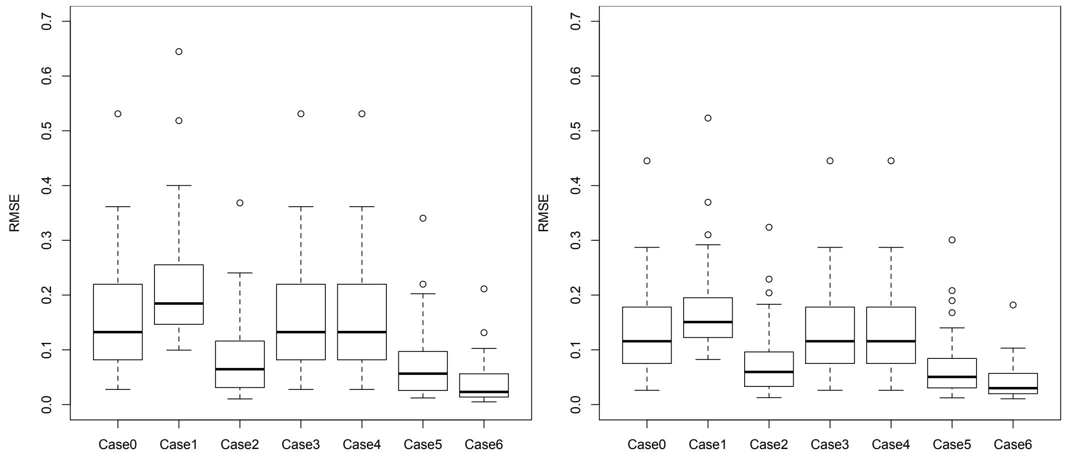


(c) PCs as regression functions on unregistered data

(d) PCs as regression functions on registered data

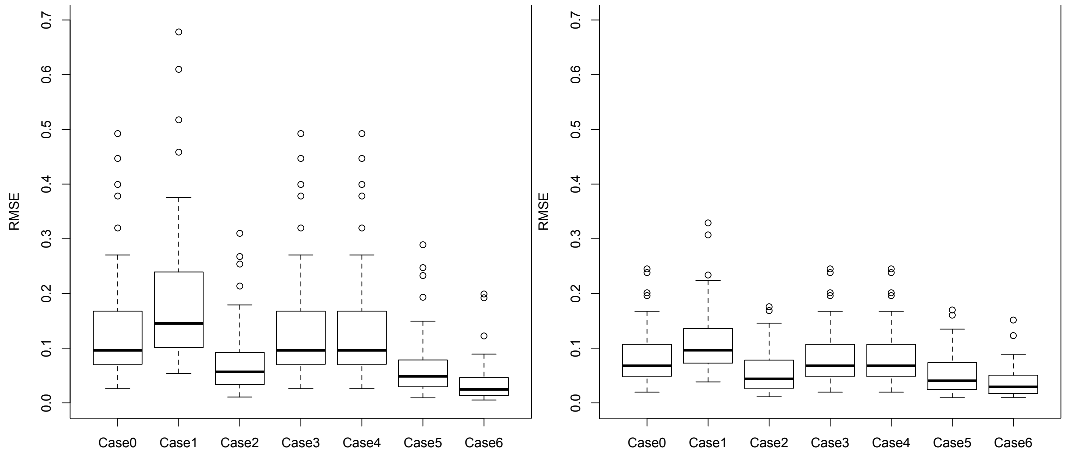
Figure 5.17: Comparison of the mean CI length for the different combinations of correlation lengths, for gauge 141, looking at the four different kind of emulations separately.

From Figures 5.17 to 5.20 it can be obtained that by keeping fix the λ_1, λ_2 and λ_3 and changing just the value of λ_t, the emulator's predictions do not change. Also, for the same values of λ_t, by increasing the values of λ_1, λ_2 and λ_3, the predictions are improved, whereas by decreasing these values, the predictions are getting worse. One can get a sense on the influence if these estimates of correlation lengths on the output. From the comparisons of the RMSE and mean CI length figures, the final selection for the input and output correlations lengths that are used for the analysis in this chapter is the following:



(a) Fourier basis regression functions on unregistered data

(b) Fourier basis regression functions on registered data



(c) PCs as regression functions on unregistered data

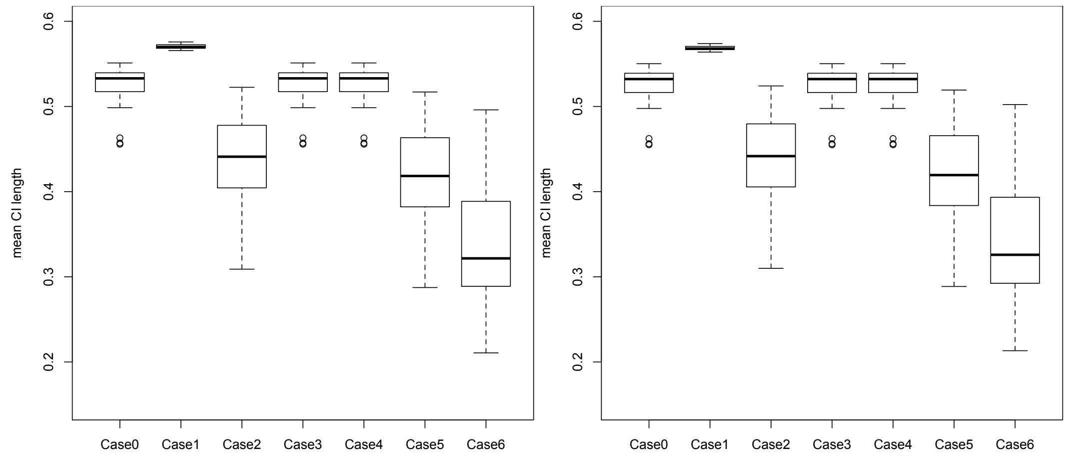
(d) PCs as regression functions on registered data

Figure 5.18: Comparison of the RMSE for the different combinations of correlation lengths, for gauge 141, looking at the three different techniques of emulations separately.

$$\begin{aligned} \lambda_1 = \lambda_2 = \lambda_3 &= 1 \\ \lambda_t &= 0.5 \end{aligned} \tag{5.14}$$

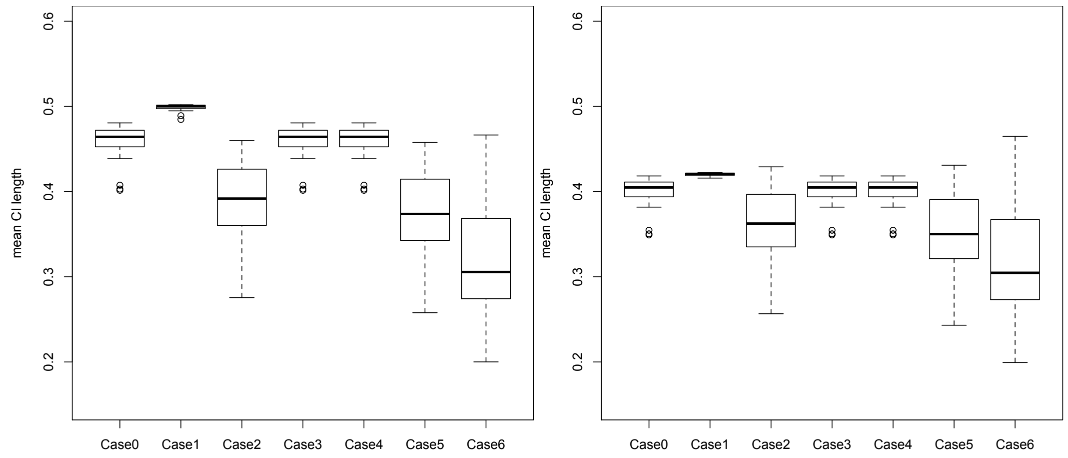
The choice of $\lambda_1 = \lambda_2 = \lambda_3 = 2$ results in slightly better predictions, with smaller RMSE and mean CI length for gauge 141. However, for gauge 32298, where the emulator's job is harder due to the more complicated wave shape, the choice $\lambda_1 = \lambda_2 = \lambda_3 = 1$, both with smaller RMSE but also with smaller range of the mean CI length. Therefore, we select the correlation lengths shown in Equation 5.14.

The selection of values for the correlation lengths could be performed by employing marginal likelihood, as in the analysis in Chapter 3. However, the maximising the marginal



(a) Fourier basis regression functions on unregistered data

(b) Fourier basis regression functions on registered data



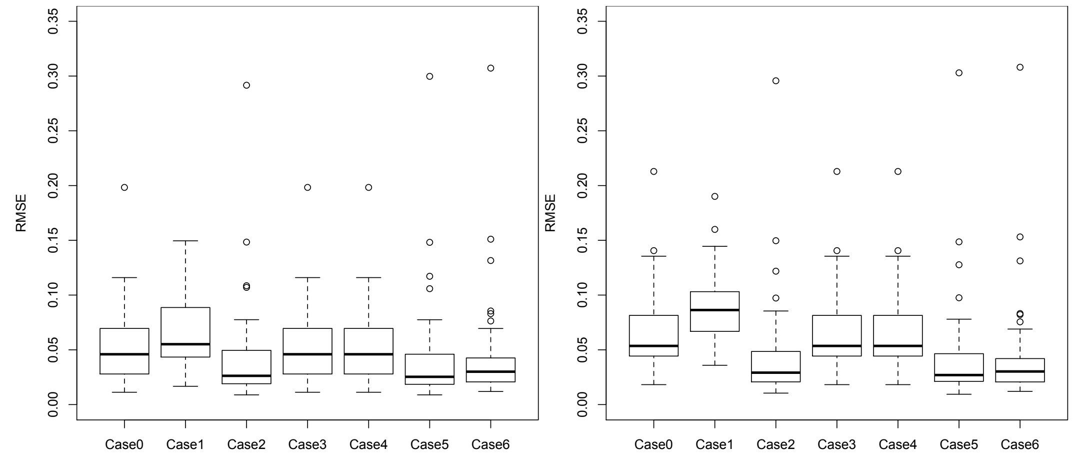
(c) PCs as regression functions on unregistered data

(d) PCs as regression functions on registered data

Figure 5.19: Comparison of the mean CI length for the different combinations of correlation lengths, for gauge 32298, looking at the four different kind of emulations separately.

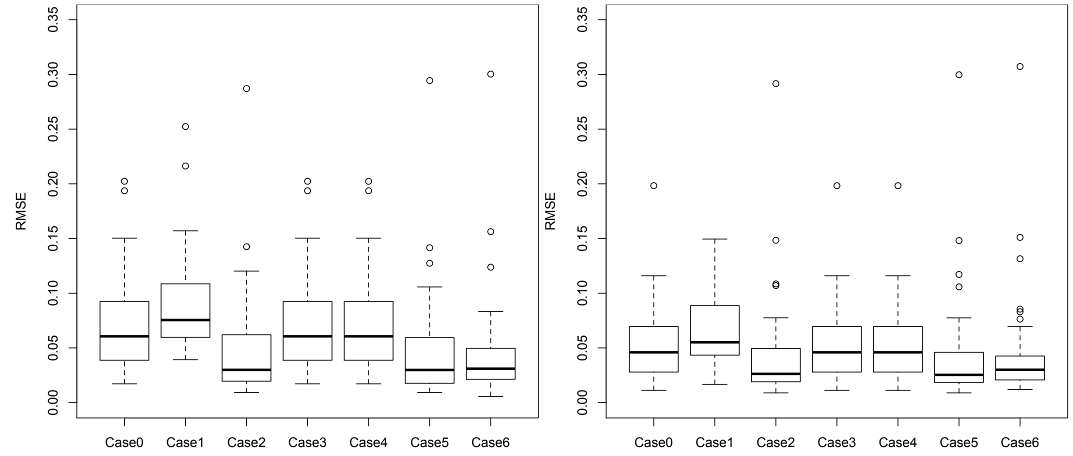
likelihood method requires matrix multiplications and inversions, which makes the process time and resources demanding. For the analysis in this section the selection of the correlation lengths is carried out using trial and error methods for simplicity and for avoiding time consuming computations. The resulting emulation predictions using the correlation lengths selected with this simple method, are still very reasonable and acceptable. Nevertheless, it is planed the repetition of this analysis using correlation lengths selected with the maximising marginal likelihood process in the future.

The last step in statistical emulation is always the choice of hyperparameters $\{m, V, a, d\}$ described in detail in Chapter 3. The same method is repeated in this chapter twice, one for the unregistered and one for the registered data. With these selections of parameter, the prior emulator is constructed and is combined with the model evaluations. The resulting emulators



(a) Fourier basis regression functions on unregistered data

(b) Fourier basis regression functions on registered data



(c) PCs as regression functions on unregistered data

(d) PCs as regression functions on registered data

Figure 5.20: Comparison of the RMSE for the different combinations of correlation lengths, for gauge 32298, looking at the four different techniques of emulations separately.

are able to give probabilistic predictions of the wave elevation for 10 seconds time steps for any combination of the three source characteristics in the three-dimensional design space.

5.4.2.2 Emulator's predictions validation

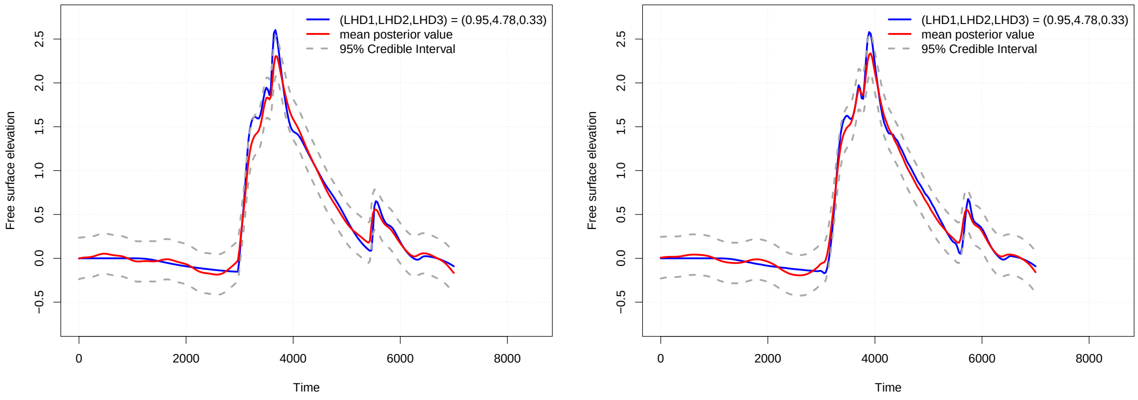
The LOO validation method described in Chapter 3 is used for conclude how accurate the emulator predicts the model output. Therefore, 40 LOO diagnostic plots are obtained for each location and each techniques of emulator. For most of the cases, the LOO diagnostics allow the conclusion that the emulator predictions are very satisfactory, with both the shape and the peaks to be accurately predicted. Additionally, the 95% credible intervals capture the computer model evaluations for most of the time. The mean credible interval length (MCIL) and the root-mean-square error (RMSE) between the observed and the predicted values for each of the 40 inputs are calculated and used to assess the predictions.

5.4.3 Comparison of emulator's predictions before and after the application of Registration and FPCA

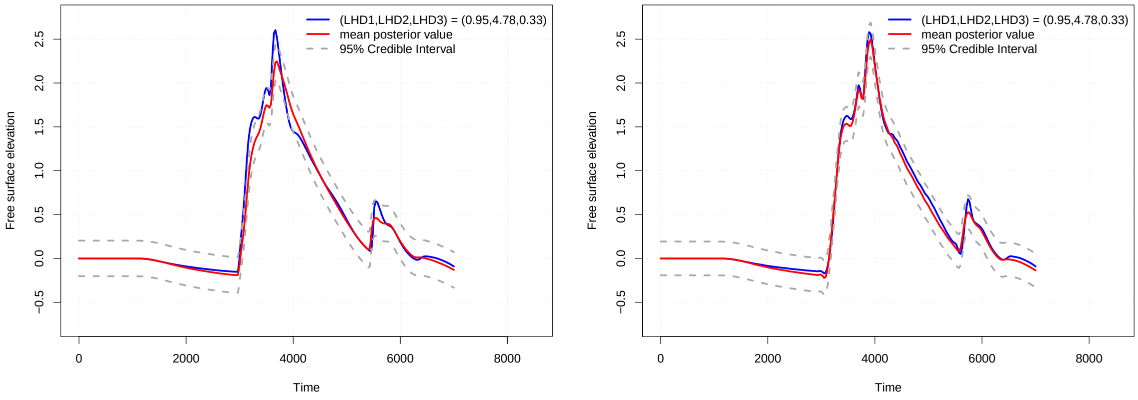
The four techniques of emulation validations are compared. First the diagnostics for gauge 32298, where the waves shape is more complicated compared to gauge 141, are observed. The more complicated time series shape makes the emulator's job to give accurate predictions harder. This gives the opportunity to have more significant differences between the predictions of the four techniques of emulations and to finally draw conclusions whether or not the application of registration and FPCA improves the emulator's performance.

First, the LHD input point number 5 is observed. The specific case has relatively large, close to 1, ratio of the trench height to the middle maximum height, z_t/z_{max}. This indicates close to the maximum trench uplift and hence large discontinuity at the trench. Also, the maximum height is large with z_{max} close the the upper limit of the range. Hence, the resulting tsunami waves have large amplitudes and variations, since it is proportional to z_{max}, which makes the emulator's job more difficult. The predictions for the four variations of emulation are shown in Figure 5.21. Similarly for the same gauge the corresponding predictions for another LHD input point, number 36, are shown in Figure 5.22. The specific point has relatively small z_{max} value and hence the waves amplitude and generally the variations are small compared to the previous input point presented. This gives higher chances to the emulator for good predictions.

For both input points it can be observed that using Fourier basis functions the emulator predictions cannot escape from being wave shaped even at locations where the wave elevation is almost flat, which is at the initial 3000 seconds. However, it can be concluded that the emulator performs rather well even with the unregistered data and the selection of Fourier basis functions. The comparison of the mean CI length and RMSE statistics for the four emulation processes examined are shown in Figure 5.23. From 5.23(b) it can be obtained that the use of PCs instead of Fourier basis functions for the unregistered data does not improve the predictions; in fact it makes the emulator's predictions worse. From Figure 5.21 it is obvious that the use of PCs eliminate the initial wrong wave-shaped fluctuations of the predictions that are result of the use of Fourier terms for the output regression functions. However, even this improvement, for the case of the unregistered data, the use of PCs causes the predictions to be unable to catch the amplitude as well as the details in the shape of the wave time series evaluations. This is also the case for some other LHD points at this location, where the wave amplitude is large and there are many details in the wave oscillation. Nonetheless, after the application of landmark registration, the use of PCs improve further the emulator's predictions.



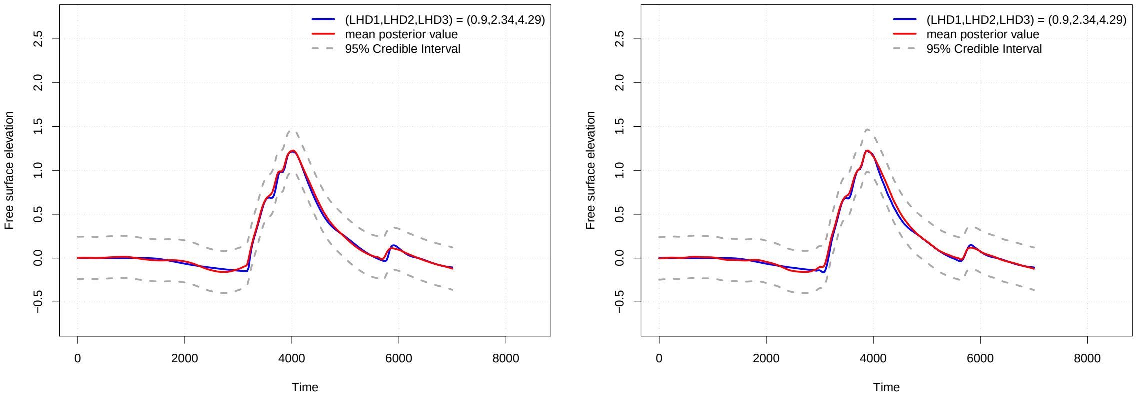
(a) Fourier basis regression functions on unregistered data (b) Fourier basis regression functions on registered data



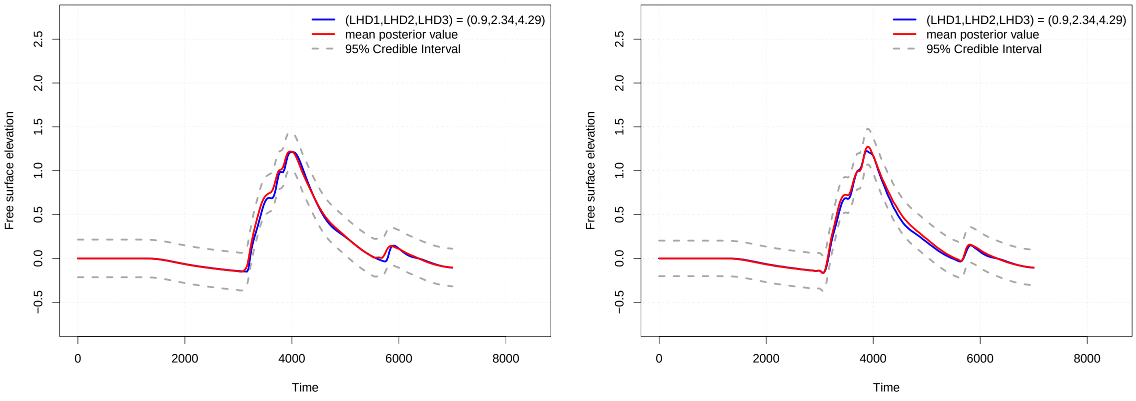
(c) PCs as regression functions on unregistered data (d) PCs as regression functions on registered data

Figure 5.21: Emulator's predictions for gauge 32298 for the coseismic deformation case described the the source characteristics $z_{max} = 4.78$, $z_t/z_{max} = 0.95$ and $t_p/t_r = 0.33$ (LHD point 5). LOO diagnostics plots are shown comparing how well the four techniques of emulation do on predicting a specific input resulting wave elevation.

Since two of the emulation processes are performed using the unregistered data and the other two to the registered data, it is impossible to directly compare the predictions of the four processes by plotting the mean value of the predictions on the same figure. However, the comparison of the diagnostics of the Fourier basis regression functions and the PCs applied on registered data are presented in Figure 5.24 for both LHD points 5 and 36. Note that these predictions cannot directly compared with the actual simulator's evaluations. However, having in mind the shape of the actual evaluations curve, it can be concluded that the predictions using PCs as output regression functions are much more accurate since they avoid the unnecessary initial fluctuations on the time series before the first wave elevation. The reason for these fluctuations is the use of Fourier functions that automatically assume periodic wave oscillations. Additionally, the FPCA based technique of emulation captures better the peaks of the oscillation.

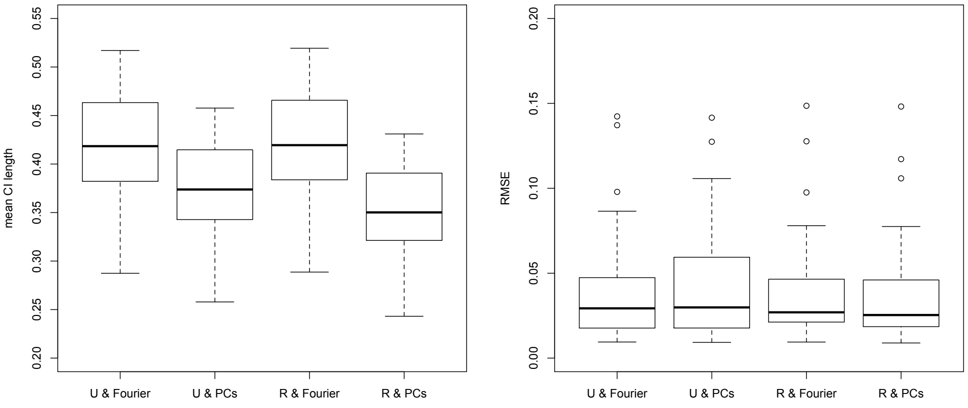


(a) Fourier basis regression functions on unregistered data (b) Fourier basis regression functions on registered data



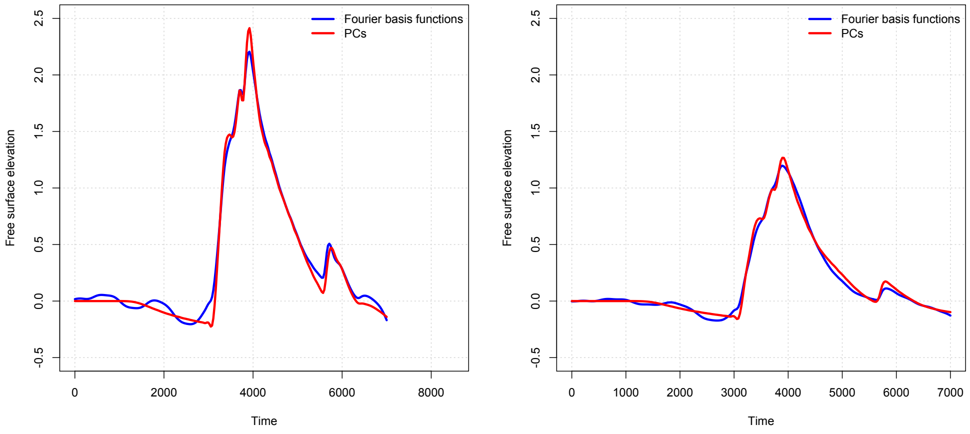
(c) PCs as regression functions on unregistered data (d) PCs as regression functions on registered data

Figure 5.22: Emulator's predictions for gauge 32298 for the coseismic deformation case described the the source characteristics $z_{max} = 4.78$, $z_t/z_{max} = 0.95$ and $t_p/t_r = 0.33$ (LHD point 36). LOO diagnostics plots are shown comparing how well the four techniques of emulation do on predicting a specific input resulting wave elevation.



(a) Mean CI length (b) RMSE

Figure 5.23: Total Mean Credible Interval length and RMSE for emulation prediction for the three techniques of emulation looking at gauge 32298.



(a) LHD point 5

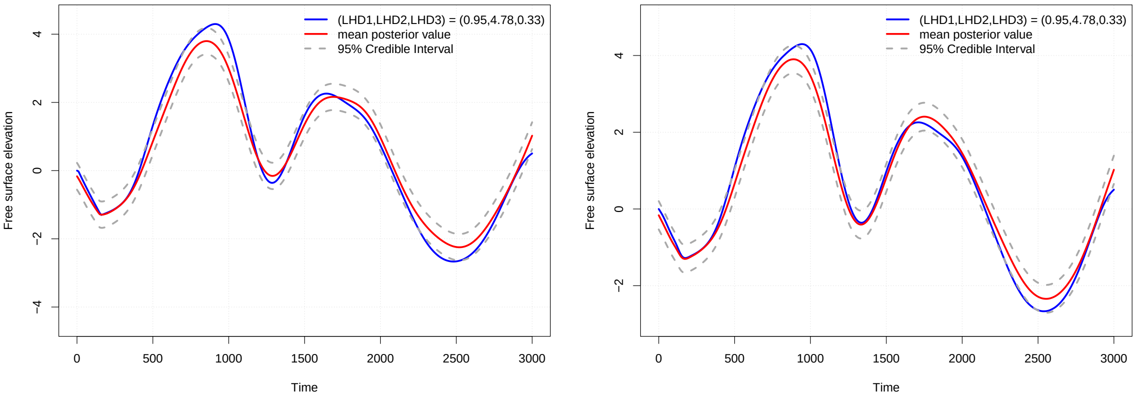
(b) LHD point 36

Figure 5.24: Comparison of the predictions after the application of landmark registration for the Fourier basis regression functions and the PCs for gauge 32298, looking separately at LHD points number 5 and 36.

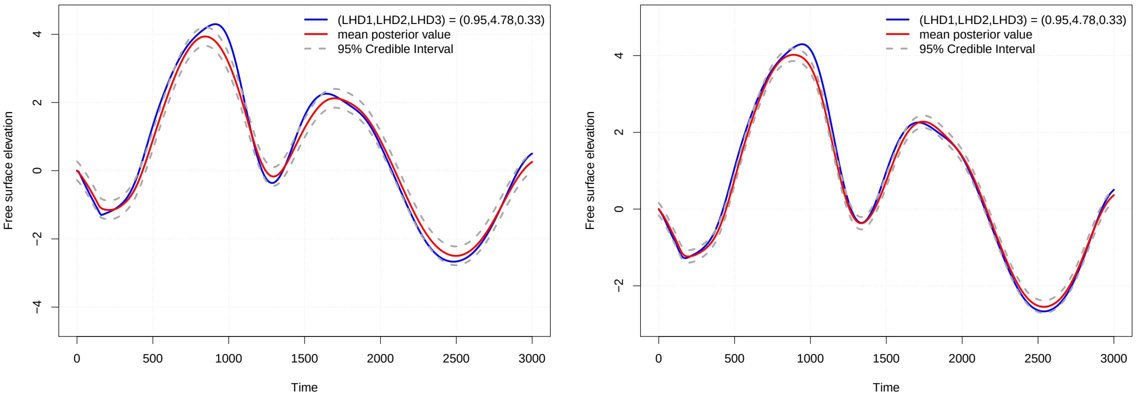
The same diagnostic plots for the same LHD inputs as well as the same statistics are presented in Figures 5.25, 5.26 and 5.27, respectively, for gauge 141. In the case of this gauge, contrary to gauge 32298, the use of PCs in place of Fourier basis functions improve the predictions for both the unregistered and registered data cases. The reason for this is that the wave elevation oscillations at gauge 141 are smoother without lot of fluctuations compared to gauge 32298. At this point, it is worth mentioning that LHD point 5 is an input point that is located at the boundaries of the computation domain, as is shown in Figure 5.13. However, the predictions of the emulator are still excellent.

Finally, it is worthy mentioned that another possible choice for output regression functions could be the wavelets. A wavelet transform is very similar to the Fourier transform, but with the function to be represented by wavelets instead of sine and cosine functions. Wavelets are very effective at extracting frequency information from the data. Specifically, only a small number of selected wavelet coefficients is necessary to describe complicated signals which makes it very computationally efficient. Also, the wavelets can be used to overcome the significant disadvantage of the Fourier which is that it has only frequency resolution and not time resolution. This means that using Fourier regression functions can determine all the frequencies presented in the signal. However, it is not possible to know when they are present. On the other hand, wavelets can be used to represent a signal simultaneously at time and frequency domain [Chui, 2014; Nason, 2010].

As a conclusion to the analysis of this section, comparing the four different methods investigated to obtain a representative statistical emulator, the emulator's performance is impressively



(a) Fourier basis regression functions on unregistered data (b) Fourier basis regression functions on registered data



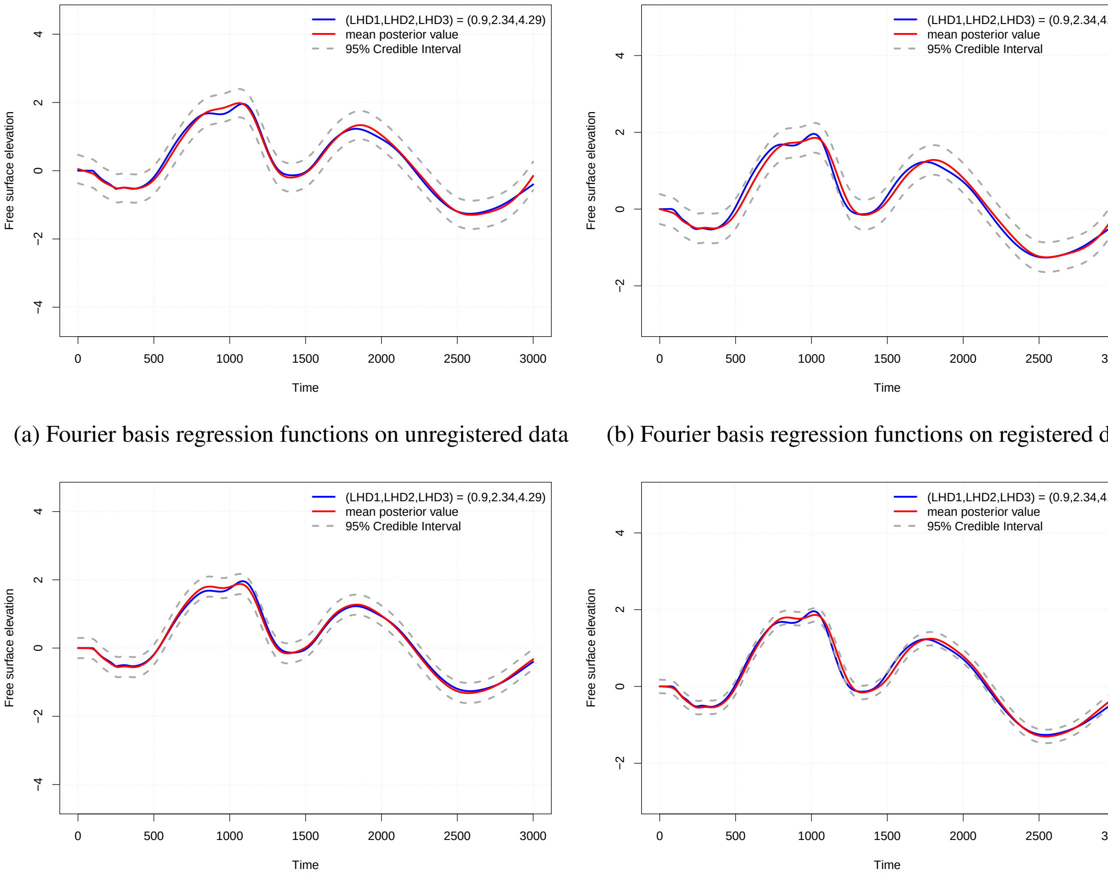
(c) PCs as regression functions on unregistered data (d) PCs as regression functions on registered data

Figure 5.25: Emulator's predictions for gauge 141 for the coseismic deformation case described the the source characteristics $z_{max} = 4.78$, $z_t/z_{max} = 0.95$ and $t_p/t_r = 0.33$ (LHD point 5). LOO diagnostics plots are shown comparing how well the four techniques of emulation do on predicting a specific input resulting wave elevation.

good for all the cases. However, the use of a landmark registration together with PCs for the output regression functions can improve further the emulator's predictions.

5.4.4 Sensitivity and Uncertainty Analyses

After building an accurate statistical emulator, it can be used further for sensitivity and uncertainty analyses. The analysis performed in Section 3.3 for the landslide-generated tsunami model is repeated in this section for the realistic representation of earthquake events which can result in tsunamis in the Cascadia Subduction Zone. The sensitivity analysis and quantification of uncertainties are presented for the two gauges number 141 and 32298, but they can be repeated for any gauge location. These analyses are very high computationally expensive, since they require a huge number of model evaluations. Therefore, the use of the statistical emulator in place of the simulator is necessary. The computational time required for these demanding analyses using the emulator is impressively small.



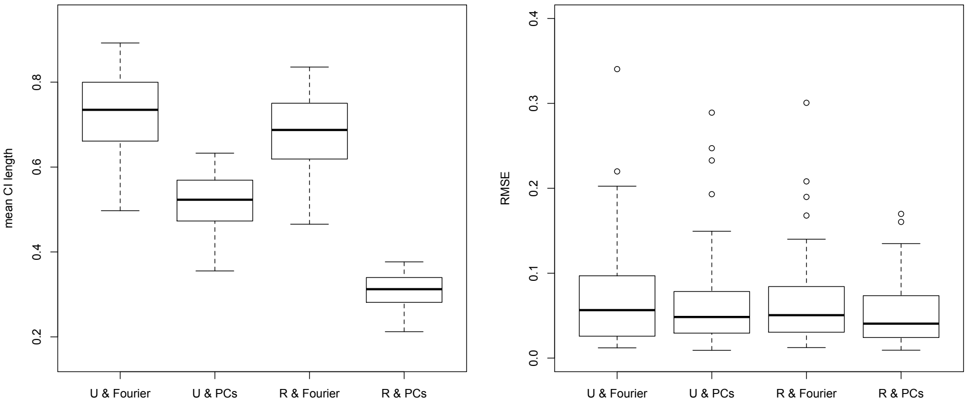
(a) Fourier basis regression functions on unregistered data

(b) Fourier basis regression functions on registered data

(c) PCs as regression functions on unregistered data

(d) PCs as regression functions on registered data

Figure 5.26: Emulator's predictions for gauge 141 for the coseismic deformation case described the the source characteristics $z_{max} = 4.78$, $z_t/z_{max} = 0.95$ and $t_p/t_r = 0.33$ (LHD point 36). LOO diagnostics plots are shown comparing how well the four techniques of emulation do on predicting a specific input resulting wave elevation.



(a) Mean CI length

(b) RMSE

Figure 5.27: Total Mean Credible Interval length and RMSE for emulation prediction for the four techniques of emulation looking at gauge 141.

5.4.4.1 Sensitivity analysis

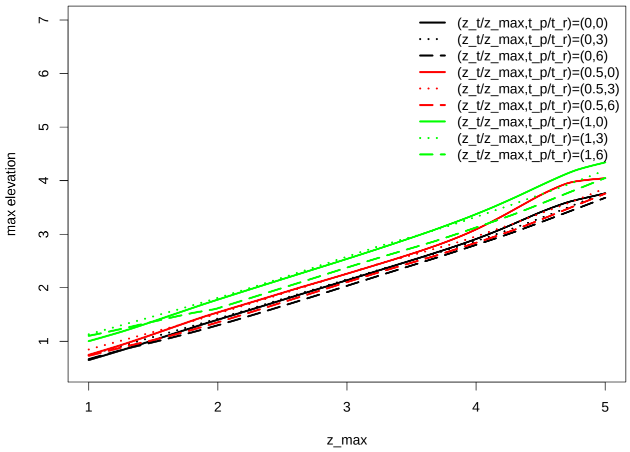
The use of the computationally expensive computer model for sensitivity analysis is not the best idea, since a large number of model evaluations is required. This analysis can become feasible and efficient with the use of the statistical emulator in place of the computer model. In this section, the sensitivity of the maximum wave elevation for the whole time series to changes in the three source characteristics is investigated. The results of the analysis are presented in Figures 5.28 and 5.29.

For the sensitivity analysis for z_{max}, emulator predictions are obtained at combinations of $(z_{max}, z_t/z_{max}, t_p/t_r)$, for every 0.01 steps for the parameter z_{max} ($z_{max} = [1, 1.01, 1.02, \dots, 5]$) and for the maximum, middle and minimum values of the range of the two other parameters, z_t/z_{max} and t_p/t_r. This equals to 3609 different scenarios of combinations of the source. Similarly, for the sensitivity analysis for z_t/z_{max}, 909 source deformation events are investigated in total, resulting from combinations of 0.1 steps for z_t/z_{max} and the maximum/middle/minimum values of z_{max} and t_p/t_r. Finally, the process is repeated for the analysis of the sensitivity of model output on the parameter t_p/t_r, with 5409 input combinations in total ($t_p/t_r = [0, 0.01, 0.02, \dots, 6]$, $z_{max} = [1, 3, 5]$ and $z_t/z_{max} = [0, 0.5, 1]$).

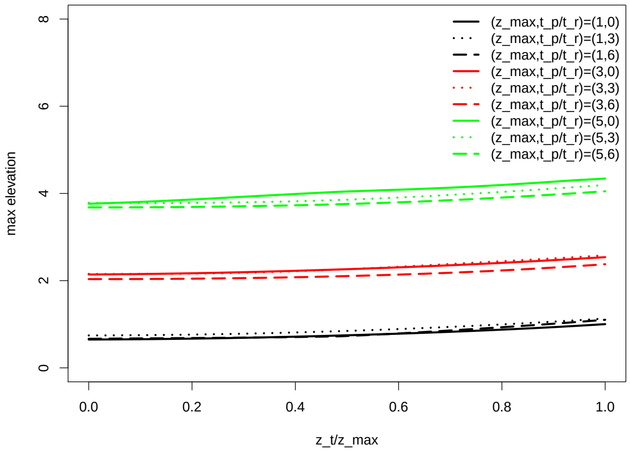
Looking first at gauge 141, from the Figure 5.28a it can be concluded that there exists a strong positive correlation between the maximum wave elevation and the z_{max}. Also, there is a slight increase in the wave elevation when the trench height z_t is closes to the middle maximum height z_{max}. Additionally, no strong dependency of the model output on the rupture propagation time is obtained. Similarly, from Figure 5.28b the clear positive dependency of the maximum wave elevation on the z_{max} values is again highlighted. Also, a small positive slope of the lines can be obtained which indicates small positive correlation between the z_t/z_{max} and the model output. However, no obvious relationship on the third input parameter is shown. Same conclusions can be drawn from the last Figure 5.28c.

For the gauge 32298, the same dependencies of the maximum wave elevation on the three source characteristics are valid. In conclusion, the model output is strongly positively correlated with the parameter z_{max}, with the large the middle maximum uplift and subsidence of the deformation, the large the resulting wave amplitude. Additionally, a not that significant positive relationship of the output on the parameter z_t/z_{max} is obtained. This means that the higher seabed elevation at the trench results in higher tsunami waves. Finally, no obvious dependency of the output on the last parameter t_p/t_r is observed. Hence, the time required for the deformation to be completed is not significantly affecting the wave elevation amplitude.

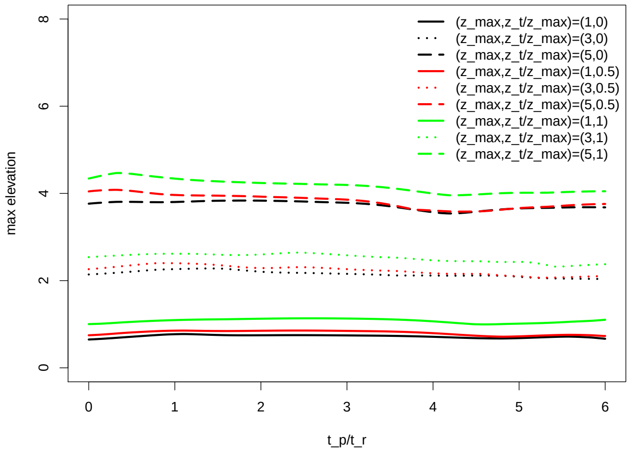
The sensitivity analysis is then repeated looking at the sensitivity of the model output on



(a) max wave elevation with respect to z_{max} for different combinations of z_t/z_{max} and t_p/t_r

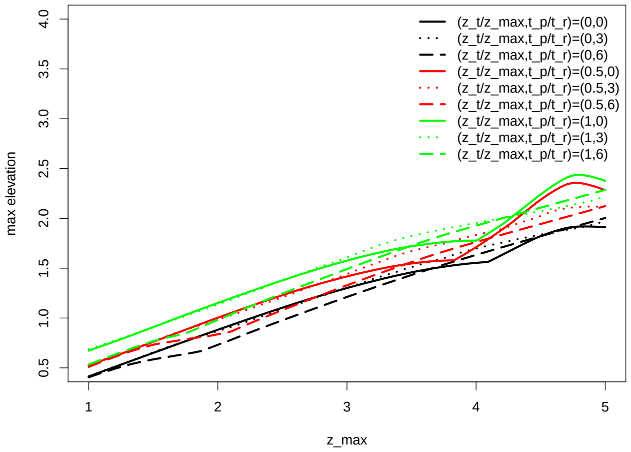


(b) max wave elevation with respect to z_t/z_{max} for different combinations of z_{max} and t_p/t_r

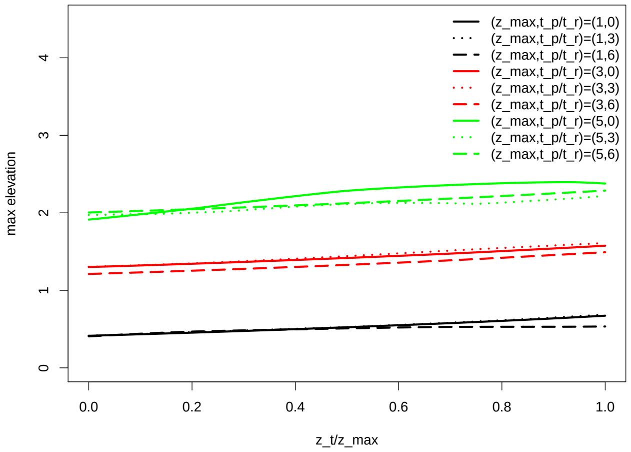


(c) max wave elevation with respect to t_p/t_r for different combinations of z_{max} and z_t/z_{max}

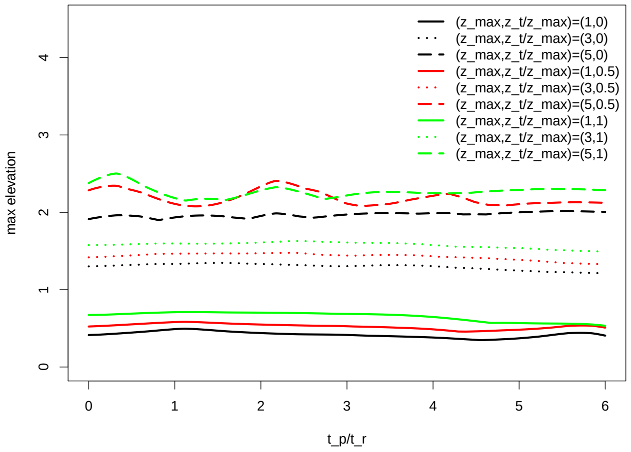
Figure 5.28: Sensitivity analysis of maximum wave elevation with respect to (a)z_{max}, (b)z_t/z_{max} and (c)t_p/t_r for gauge 141.



(a) max wave elevation with respect to z_{max} for different combinations of z_t/z_{max} and t_p/t_r



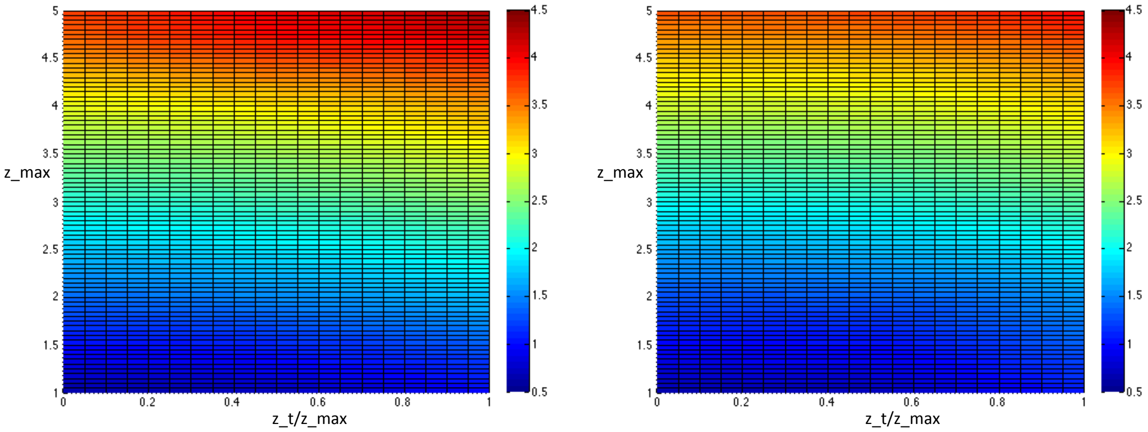
(b) max wave elevation with respect to z_t/z_{max} for different combinations of z_{max} and t_p/t_r



(c) max wave elevation with respect to t_p/t_r for different combinations of z_{max} and z_t/z_{max}

Figure 5.29: Sensitivity analysis of maximum wave elevation with respect to (a)z_{max}, (b)z_t/z_{max} and (c)t_p/t_r for gauge 32298.

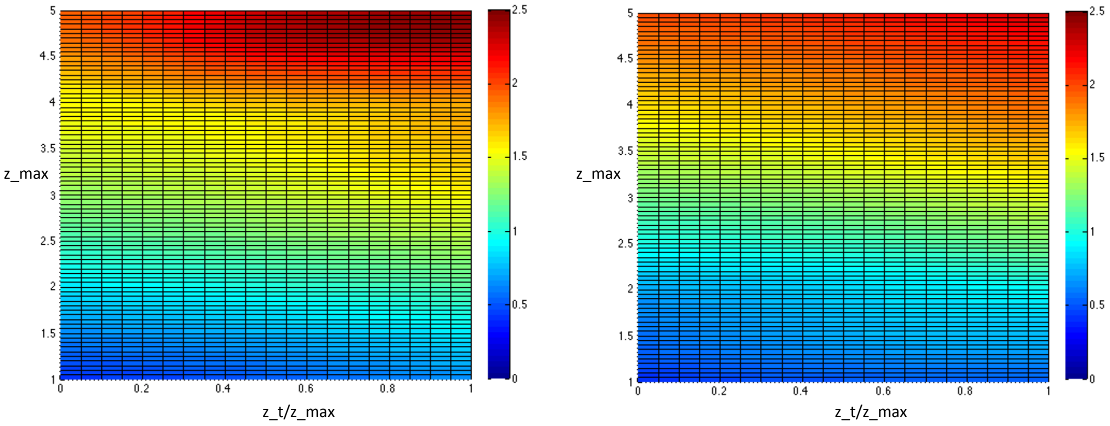
combination of two inputs instead of one. Each time the emulator is evaluated at different combinations of the inputs, at every 0.05 steps for the two of them, and at the maximum and minimum values for the third one. Therefore, for each gauge, three different combinations of inputs are considered. Specifically, for the first combination that investigates the sensitivity of the maximum wave elevation on z_{max} and z_t/z_{max}, 3402 different input combinations are considered (for $z_{max} = [1, 1.05, \ldots, 5]$, $z_t/z_{max} = [0, 0.05, \ldots, 1]$ and for $t_p/t_r = [0, 6]$). The results are presented in Figures 5.30 and 5.31 for gauges 141 and 32298, respectively. Similarly, for combinations of z_{max} and t_p/t_r, and z_t/z_{max} and t_p/t_r, 19602 and 5082 different event scenarios have been investigated. Figures 5.32 to 5.35 result from these analyses.



(a) $t_p/t_r = 0$

(b) $t_p/t_r = 6$

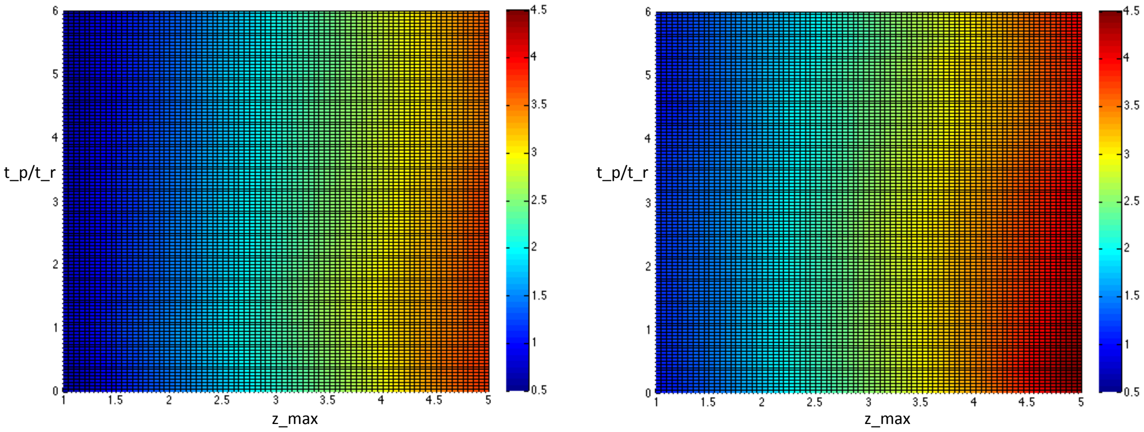
Figure 5.30: Sensitivity analysis of maximum wave elevation with respect to changes in $zmax$ and z_t/z_{max} for (a)$t_p/t_r = 0$ and (b)$t_p/t_r = 6$, for gauge 141.



(a) $t_p/t_r = 0$

(b) $t_p/t_r = 6$

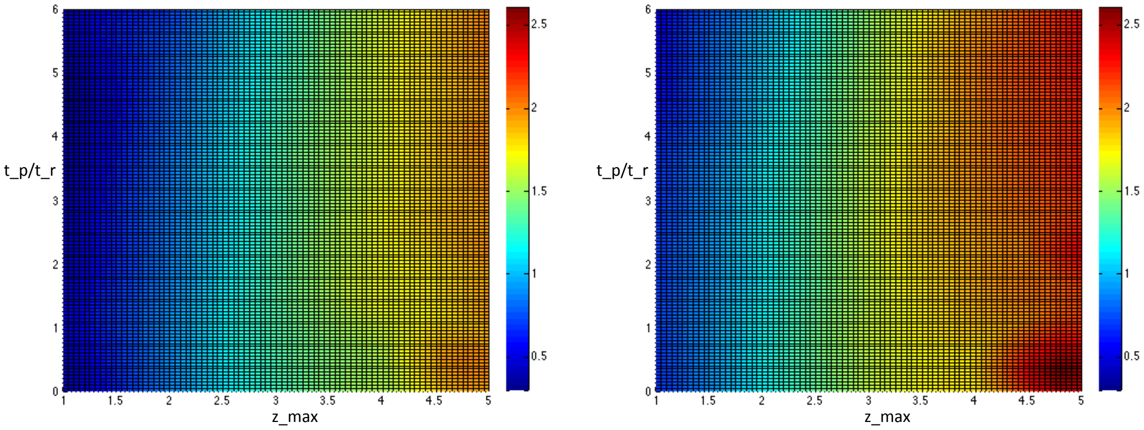
Figure 5.31: Sensitivity analysis of maximum wave elevation with respect to changes in $zmax$ and z_t/z_{max} for (a)$t_p/t_r = 0$ and (b)$t_p/t_r = 6$, for gauge 32298.



(a) $z_t/z_{max} = 0$

(b) $z_t/z_{max} = 1$

Figure 5.32: Sensitivity analysis of maximum wave elevation with respect to changes in z_{max} and t_p/t_r for (a)$z_t/z_{max} = 0$ and (b)$z_t/z_{max} = 1$, for gauge 141.



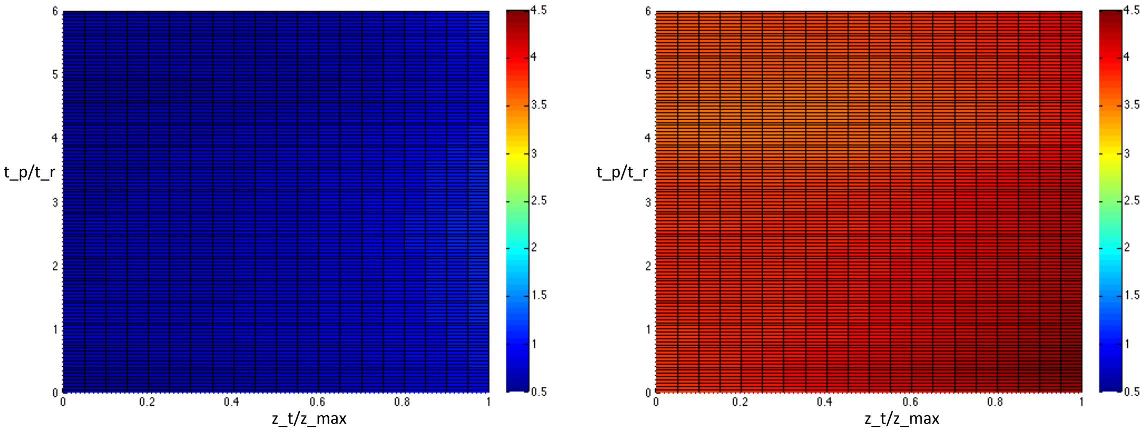
(a) $z_t/z_{max} = 0$

(b) $z_t/z_{max} = 1$

Figure 5.33: Sensitivity analysis of maximum wave elevation with respect to changes in z_{max} and t_p/t_r for (a)$z_t/z_{max} = 0$ and (b)$z_t/z_{max} = 1$, for gauge 32298.

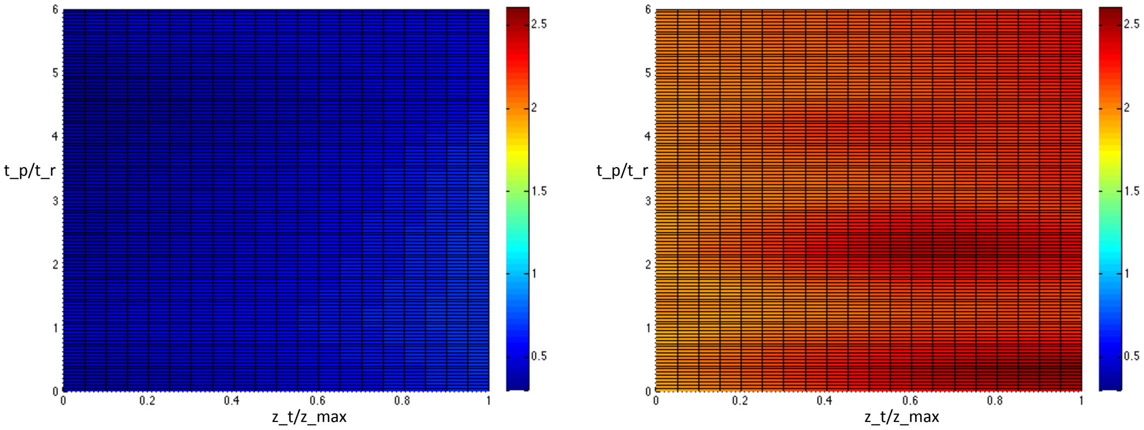
Again from these Figures, the lack of any dependency of the resulting wave elevation amplitude on the rupture propagation time is highlighted for both of the gauges. Only in Figure 5.31 a small decrease in the wave amplitude by increasing t_p/t_r can be observed. This means that the faster the propagation, the slightly higher the maximum wave elevation. Also, the strong dependency on the middle maximum uplifted height z_{max} is very obvious.

Additionally, from Figures 5.34 and 5.35, it can be concluded that for fixed ratio of trench to middle maximum heights, the larger waves result in the case that the middle maximum height is large and the propagation is fast. Furthermore, from Figures 5.32 and 5.33, for fixed z_{max} the larger/smaller output obtained when the ration $\frac{z_t}{z_{max}}$ is close to one/zero and the propagation is fast/slow.



(a) $z_{max} = 1$ (b) $z_{max} = 5$

Figure 5.34: Sensitivity analysis of maximum wave elevation with respect to changes in z_t/z_{max} and t_p/t_r for (a)$z_{max} = 1$ and (b)$z_{max} = 5$, for gauge 141.



(a) $z_{max} = 1$ (b) $z_{max} = 5$

Figure 5.35: Sensitivity analysis of maximum wave elevation with respect to changes in z_t/z_{max} and t_p/t_r for (a)$z_{max} = 1$ and (b)$z_{max} = 5$, for gauge 32298.

In conclusion, the statistical emulator is used in place of the expensive-to-run simulator to perform high resolution sensitivity analysis of the maximum wave elevation on the source characteristics, as well as on combinations of these characteristics. For the analysis, a large number of different source deformation cases are investigated. Specifically, 31,696 different input combinations are investigated for each of the two locations investigated, with the total time required to be approximately 2 hours for each one. This time is impressively small for such a demanding analysis, especially if it is compared to the time required for a single run of VOLNA on 3 GPUs, that is around 20 minutes. Hence, the importance of statistically representing the computer model is highlighted.

5.4.4.2 Uncertainty analysis

The input parameters that describe the coseismic deformation source are always uncertain for real event cases. This makes the uncertainty analysis necessary in the investigation of natural hazard events. In this section, the uncertainty analysis aims to predict the amount of uncertainty on the resulting maximum wave elevation due to the uncertain input characteristics z_{max}, z_t/z_{max} and t_p/t_r. To do so, a distribution is assigned to each of the three input parameters that describes in the best way the existing knowledge about them. Using the literature, as well as Dr Simon Day's opinion about the most possible values for these parameters, a distribution is assigned to each of them and these selections are reasoned in detail below. Based on these distributions, 2000 random input combination samples are drawn and the statistical emulator is used to get the resulting maximum wave elevation for all the 2000 cases. As it is shown earlier in this chapter, the statistical emulator can give accurate and reliable predictions and allows demanding analyses to be performed efficiently.

The distributions that are given to each of the inputs are the following:

1. The first parameter z_{max} is assumed to have a Beta distribution[4] given by

$$z_{max} \sim Be(2,3) \; for \; z_{max} \in [2,5] \tag{5.15}$$

   The parameter z_{max} values relate closely to the seismic magnitude. Additionally, z_{max} depends on the coastal subsidence. The latter is supported by Goldfinger et al. [2012] who state that the coseismic uplift is the proxy turbidite volume. Leonard et al. [2010] describe the distribution of the coastal subsidence and therefore help in defining the distribution of z_{max}. The dependency of z_{max} on the coastal subsidence is particularly valid for cases where the ratio of the maximum uplift to the maximum subsidence is considered as constant, which is the case of the analysis in this chapter.

   For the whole margin rupture case investigated the values of z_{max} are considered to be more likely between 2 and 4, and less likely between 4 and 5. However, it is assumed to be improbable to take values less than 2 or more than 5. This range of values for z_{max} is supported by Wells and Coppersmith [1994] who observe that there is a range of slips associated with a fixed fault rupture length that does not go to zero. Two thousand random samples for z_{max} are drawn from the distribution (5.15), resulting in the prior distribution shown in Figure 5.36.

---

[4] The probability density function of a beta distribution, $X \sim Beta(\alpha, \beta)$, for $0 \leq x \geq 1$ and shape parameters α, β> 0, is given by $f(x; \alpha, \beta) = constant\ x^{\alpha-1}(1-x)^{\beta-1}$.

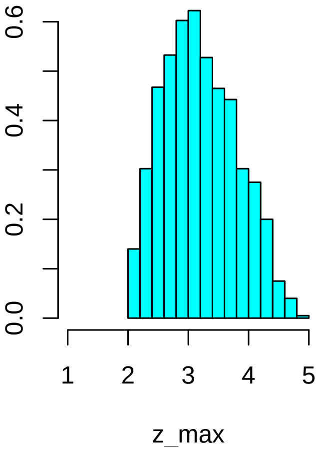


Figure 5.36: Histogram of the prior distribution assigned to the input parameter z_{max}.

2. The ratio of z_t with z_{max} is considered to follow a Truncated Normal distribution[5] described by

$$\frac{z_t}{z_{max}} \sim TN((0,1),1,0.09) \; for \; \frac{z_t}{z_{max}} \in [0,1] \tag{5.17}$$

z_t/z_{max} is assumed to be in the range $[0,1]$, even the fact that is possible to be greater than 1. However, since for the analysis the whole rupture propagation case is investigated, it is not possible this ratio to take values larger than 1 all along the margin. Later in the analysis in this chapter, a limiting event case is going to be investigated where this ratio is assumed to be equal to 3 for the whole margin. This case is definitely not realistic, but the purpose was to define an envelope for the maximum plausible coastal inundation.

Before the Tohoku 2011 tsunami event, most authors supported that the trench uplifted slightly compared to the middle maximum elevation, which means that they believed that this ratio is close to zero. However, post-Tohoku, many authors support that z_t/z_{max} is closer to one, i.e. the trench uplift height is close to the middle maximum uplift height [Masterlark and Hughes [2008], Grilli et al. [2007]], with the ratio to be much higher than one locally. The latter opinion is also supported by Dr Simon Day and this is going to be assumed for the uncertainty analysis. As for the first input parameter, 2000 random samples are drawn from the distribution 5.17, with the prior distribution of z_t/z_{max} to

---

[5] A random variable following a Truncated Normal distribution is normally distributed and its values are bounded above or below or both. For a random variable $X \sim N(\mu, \sigma^2)$, which has a Normal distribution and lies within the interval (a,b), $-\infty \leq a < b \leq \infty$, then X has a Truncated Normal distribution, conditional on $a < X < b$. For $a \leq x \leq b$ the pdf is given by

$$f(x;\mu,\sigma,a,b) = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \tag{5.16}$$

and for $x < a$ or $x > b$, then $f = 0$. In the pdf equation, $\phi(\xi)$ is the pdf of the standard Normal distribution, which is given by $\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp(\frac{1}{2}\xi^2)$ and $\Phi(\cdot)$ is its cumulative distribution function.

presented with a boxplot in Figure 5.37.

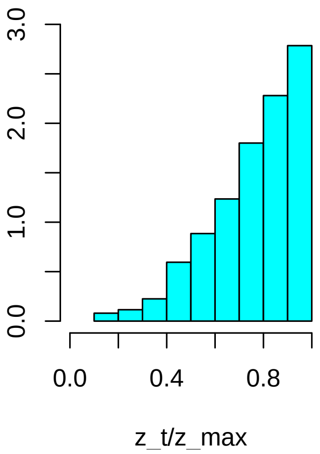


Figure 5.37: Histogram of the prior distribution assigned to the input parameter z_t/z_{max}.

3. For the last input, the limited knowledge about the way and duration of propagation leads to the selection of Uniform distribution between 0 and 6[6]:

$$\frac{t_p}{t_r} \sim U(0,6)\ for\ \frac{t_p}{t_r} \in [0,6] \tag{5.18}$$

The selection of the value 6 for the maximum possible value of t_p/t_r is assumed earlier in the analysis and it results in maximum possible rupture propagation of 900 seconds, since each raw of polygons is considered to complete its motion, t_r, in 150 seconds. Since the propagation time for most of the past events is unknown, this value is selected based on the knowledge for few giant earthquake events. Specifically, Sumatra, 2004, Chile, 2010 and Tohoku, 2011, events are considered, where the propagation times are approximately 1000, 100 and 200 seconds, respectively. On the other hand, the value 0 for this parameter indicates instantaneous rupture propagation. Therefore, for the analysis, 2000 random samples are drawn from the Uniform distribution, resulting to the prior distribution for t_p/t_r shown in Figure 5.38.

The three sample sets are combined to generate 2000 different coseismic source deformation events. Since, there is a positive correlation between the first two inputs, i.e. the larger the z_{max} value is, the larger the z_t/z_{max} [Wells and Coppersmith, 1994], the inputs combinations are arranged accordingly. Specifically, the values of z_{max} and z_t/z_{max} are combined after they are placed in ascending order and then these combinations are combined with the values of t_p/t_r in random order.

---

[6] For $X \sim U(a,b)$, the pdf is the following: $f(x) = \frac{1}{b-a}$ for $a \leq x \geq b$. For $x < a$ or $x > b$, $f(x) = 0$.

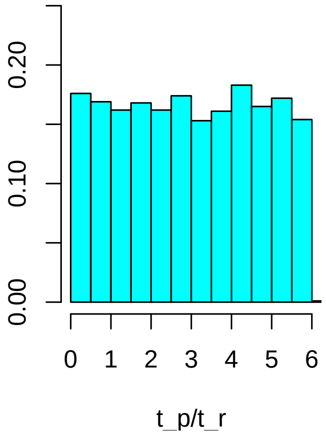


Figure 5.38: Histogram presenting the distribution assigned to the input parameter t_p/t_r.

The statistical emulator is evaluated for the 2000 different input combinations for the gauges 141 and 32298 for times up to 3000 and 7000 seconds, respectively, every 9.9 seconds for gauge 141 and every 27.9 seconds for gauge 32298. For each of the time series, the maximum elevation and the mean CI length have been estimated for each combination of the input parameters. The variations are quantified using quantiles, which are summarised in Tables 5.1 and 5.2, for gauges 141 and 32298, respectively. Additionally, the distribution of the resulting maximum wave elevation is shown in Figure 5.39 for both of the gauges investigated.

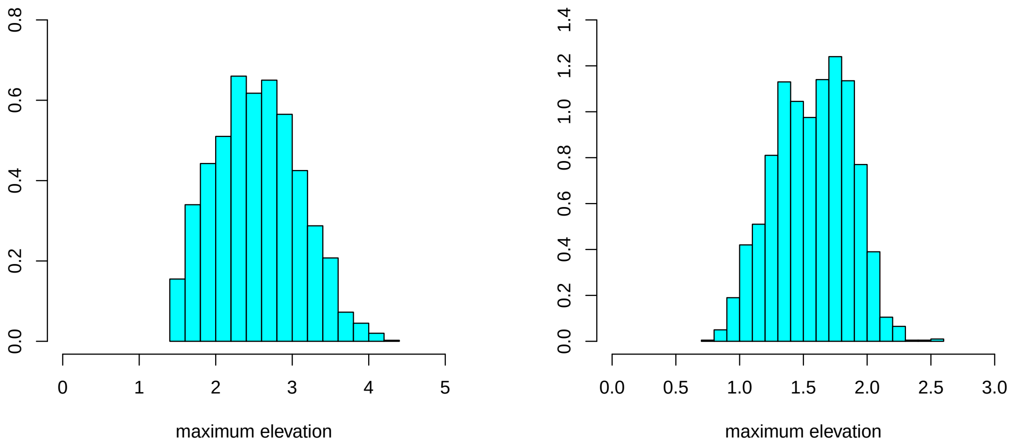
	0.1%	1%	5%	25%	50%	75%	95%	99%	99.9%
max Elevation	1.42	1.48	1.66	2.12	2.52	2.93	3.50	3.82	4.14
mean CI length	0.22	0.31	0.39	0.55	0.64	0.71	0.77	0.79	0.80

Table 5.1: Maximum wave elevation and mean CI length quantiles for the predictions for gauge 141.

	0.1%	1%	5%	25%	50%	75%	95%	99%	99.9%
max Elevation	0.82	0.92	1.07	1.35	1.59	1.80	2.02	2.15	2.42
mean CI length	0.10	0.13	0.17	0.24	0.27	0.31	0.33	0.34	0.34

Table 5.2: Maximum wave elevation and mean CI length percentiles for the predictions for gauge 32298.

From the Tables and Figures it can be concluded with 95% confidence that the resulting tsunami wave is going to have maximum wave elevation less than $3.50m$ at gauge 141 and less than $2.02m$ at gauge 32298. Additionally, from Figures 5.39 it is noticeable that for gauge 141 the histogram is positively skewed, whereas for gauge 32298 it is negatively skewed. For gauge 141 the maximum wave elevation is more likely to be between 2 and $3m$, whereas for the gauge 32298, the between 1 and $2m$. This analysis can be repeated for any location. The total



(a) gauge 141

(b) gauge 32298

Figure 5.39: Maximum wave elevation predictions distribution for (a) gauge 141 and (b) gauge 32298.

time required was approximately 26 minutes for 2000 emulator predictions. This highlights one more time the importance of replacing expensive computer models with accurate statistical surrogate models.

5.5 Full scale emulation

The generation of an overall or full scale emulator that can be evaluated at all the locations investigated is challenging. There are many reasons that make this challenging. One of them is the spatial complications in the topography, where points located close to each other can have very different heights. Additionally, at some of the coastal locations, tsunami waves are unlikely to reach due to the elevation far inland. Furthermore, another issue that makes the full scale emulation challenging is the inversion of large matrices that is necessary for the generation process of GP emulators. To overcome these difficulties, a large number of separate emulators, one of each location, is generated, instead of a full scale one.

This approach is also followed by Spiller et al. [2014], where many statistical emulators are constructed simultaneously by considering only one map point at a time and fitting an emulator for the output at each point. The authors named these emulators “subemulators” and they described an automatic process that allows the process to be repeated in parallel for all the points, using only the initial set of computer model runs. Therefore, their approach is computationally cheap and allows fast and flexible uncertainty quantification for multiple sources of uncertainty in probabilistic hazard mapping. Finally, they perform an assessment of each sub emulator using leave-one-out-tests.

Similarly to Spiller et al. [2014], an automatic statistical emulator process is applied for a large number of gauges in Victoria area for the investigation in this section. The 40 inputs describing different seabed deformation cases, which have been used earlier in the analysis, are considered here as well. The first step is to remove all the gauges where water is not arriving. To do so, the choice of gauges was constrained only for ones that presenting at least 0.5m wave amplitude variations from maximum to minimum at some point in the whole time series. Furthermore, the gauges located more than 10 meters above the sea level, which are gauges in land and far away from the coast, are not taken into consideration. The total number of gauges at Victoria area satisfy these two criteria is 3023. Statistical emulation is applied to these gauges time series simultaneously. In order to get the best predictions and the process to be automatic, landmark registration is applied to the simulator evaluations, using an automatic selection process for the landmarks, and also PCs are used for the output regression functions in the emulation process.

As described earlier in Section 5.1.2, the first step is to generated the cubic B-splines with a general common smoothing parameter λ that is used for all the gauges. Different values of λ are tested for some gauges, and it is concluded that $\lambda = 0.0001$ works the best for most of the cases. The exact value of λ that is the best choice is hard to be obtained, due to the fact that the GCV curve is relatively flat around the minimum in most of the cases. Therefore, the choice of $\lambda = 0.0001$ for all the 3023 gauges is adequate. The next step in the analysis is the choice of landmarks. To do so, for each locations, two landmarks are selected automatically. The first one is the location of the first time each of the curve cross the horizontal axis from negative to positive. The second one is the location where each of the 40 time series reach the maximum amplitude.

Following the landmark registration, PCA is applied to the registered data and the first two principal components are selected, which are then used together with the mean value curve as output regression functions for the emulation. Moreover, the input regressors as well as the input and output covariance functions are selected, followed by the selection of the hyperparameters, which is described in Section 3.1.1.2. An automatic procedure for the selection of hyperparameters is applied and therefore it is performed simultaneously for all the gauges. The selection of the input regression functions is given by Equation 5.6, whereas the input and output residuals covariance functions selection is shown in Equations 5.12 and 5.13, respectively. The correlation lengths selected are the ones that work the best for the two gauges investigated earlier, which are described in (5.14). The choices of functions and parameters are combined with the registered wave elevations to obtain statistical emulator predictions at all the gauges.

However, the posterior distribution for the emulator is eventually obtained only for 695 out of the 3023 gauges, since the process stops with an error for the rest of the gauges. The explanation for the errors is the selection of landmark points. Specifically, the automatic landmark selection employed could not detect the specific landmarks at all of the gauges time series, resulting to an error. This prevents the process to continue to the generation of the statistical emulator. A more careful choice of landmarks or a better method of registration, such as continuous registration, which does not require the selection of landmarks, could avoid these errors and result in successful statistical emulation for all the 3023 gauges. Nevertheless, the specific analysis in this sections presents some initial steps that can be further improved with better landmark registration. Unfortunately, due to the limited time, the improvements of this analysis did not take place at the end but it is going to be performed in the future.

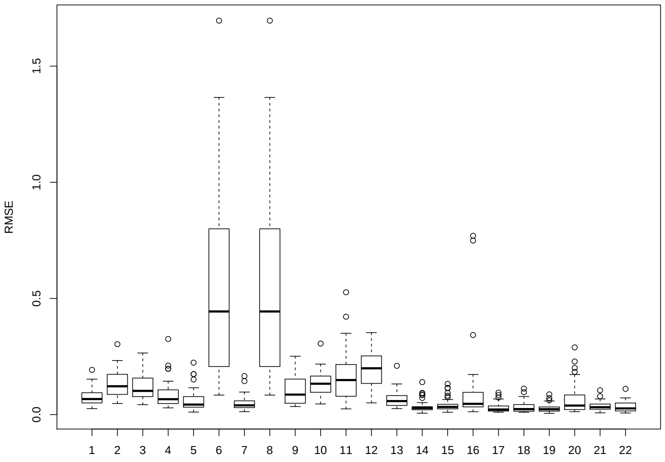
Considering the emulators predictions at the 695 gauges, LOO diagnostics methods are applied to validate them concluding that the predictions are satisfactory at almost all the locations. The RMSE and mean Credible Interval length are calculated for 22 locations, where the two of them are the ones with the maximum RMSE and mean Credible Interval length and the rest 20 are randomly selected, and their boxplots are presented in Figure 5.40. Also the maximum RMSE and mean CI length at all the 695 gauge are obtained and presented in Figure 5.41.

The location with the largest RMSE value as well as the largest mean CI length is the one presenting the worst emulator prediction and it is the gauge number 145. For this gauge, the LHD input points 20 and 24 give the largest RMSE and mean CI length, respectively. The leave-one-out diagnostic plots for the specific two inputs for gauge 145 are shown in Figure 5.42. On the other hand, the gauges 435 and 430 present the best emulation predictions, with the lowest RMSE and mean CI length, respectively. For both of these locations, the best predictions are obtained for input characteristics described by LHD number 7 and the corresponding LOO diagnostic plots are shown in Figure 5.43.

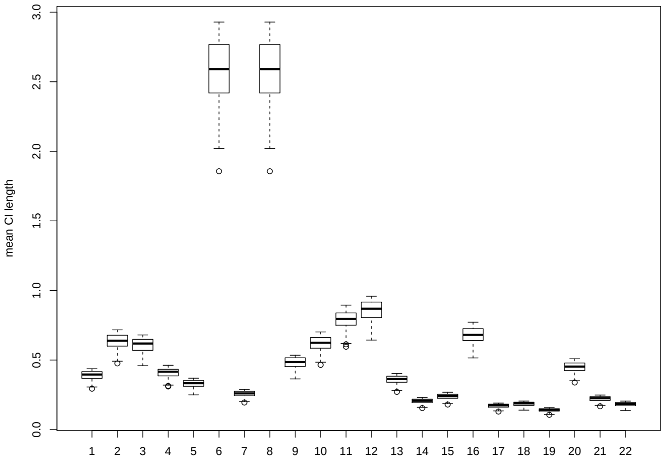
Therefore, it can be concluded that the application of statistical emulation for each location separately, without taking into consideration any spatial correlations between the gauges, and using an automatic process that can make this process simultaneous, can result in sufficiently reasonable predictions. This process can be improved further by a better choice of registration process. The generation of emulators for all the locations investigated can help in the performance of sensitivity and uncertainty analysis in the whole area.

5.6 Limiting Event

In order to identify and visualise risk exposures to tsunami inundation for the whole area around the Cascadia Subduction Zone, an investigation of a “limiting” event case is performed. A



(a) Boxplots for the emulators predictions' RMSE at the two gauges with the largest RMSE and mean CI length and also another 20 randomly selected gauges at Victoria area



(b) Boxplots for the emulators predictions' mean CI length at the two gauges with the largest RMSE and mean CI length and also another 20 randomly selected gauges at Victoria area

Figure 5.40: Statistics for the emulators predictions at the 22 gauges at Victoria area. The two corresponds to the largest RMSE and mean CI length and the other 20 are randomly selected.

"limiting" event is an extreme, almost implausible in real life event, which is explored in order to obtain the maximum possible water inundation. This is an efficient way to obtain the worst case scenario of tsunami disaster in just a single run of the tsunami model. The predictions for this extreme event can help eliminating hazard for the next tsunami events in the area by creating efficient defences and prepare the population accordingly.

For this event, the trench height z_t is considered to take values up to 15m and also it is assumed to be three times the height of the maximum middle uplift, z_{max}, which therefore takes

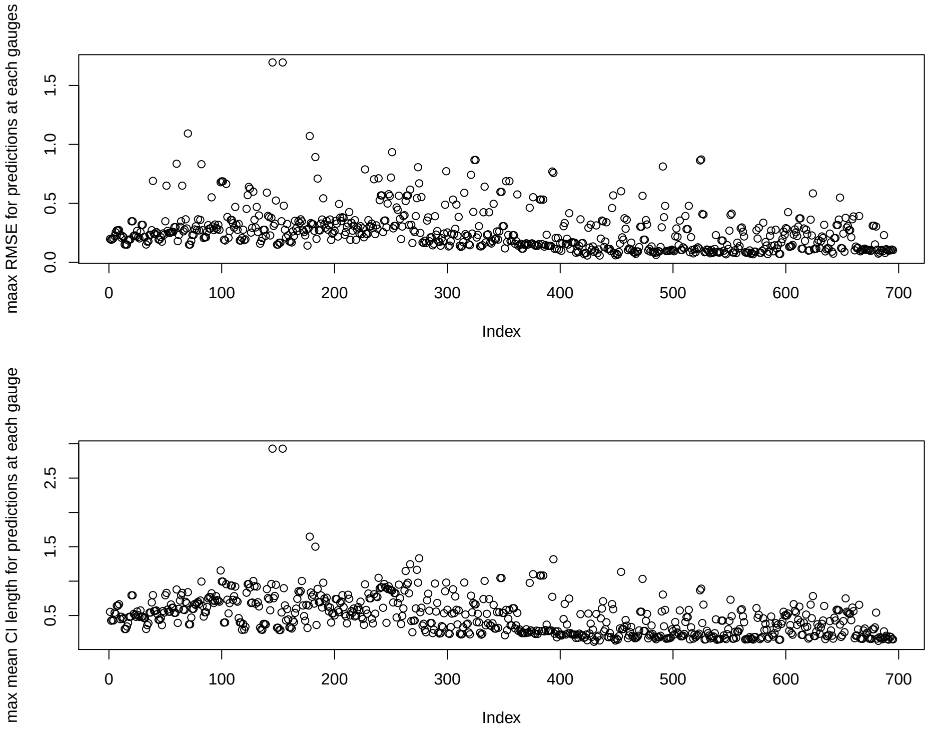


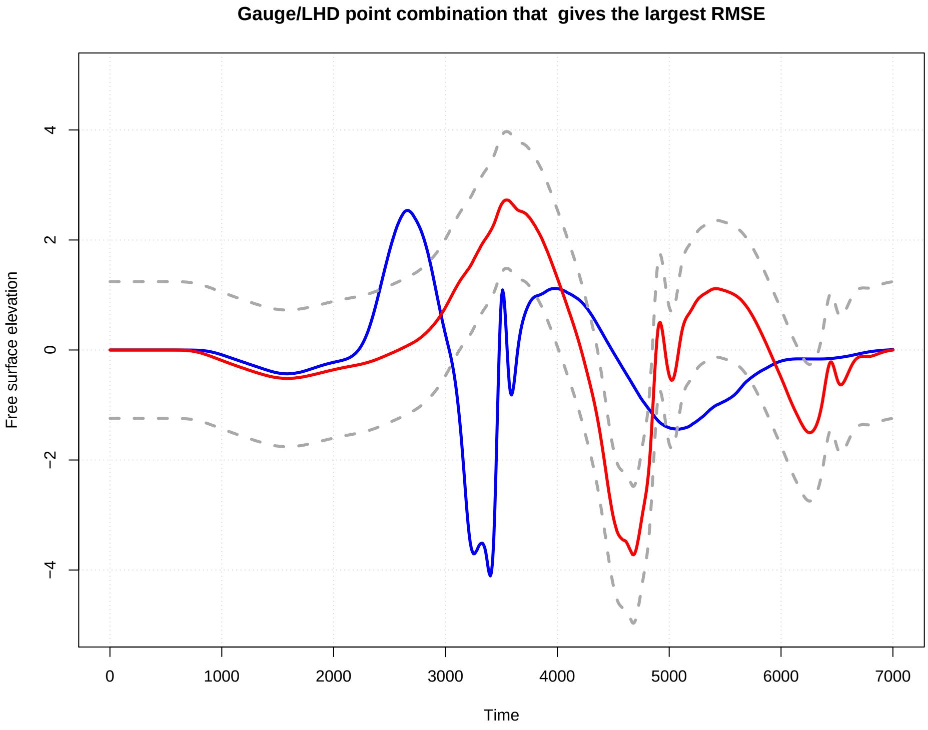
Figure 5.41: Maximum values of the RMSE and mean CI length for all the 695 gauges where statistical emulation has been applied.

amplitudes up to 5m. Unlikely to the previous analysis with the 40 different deformation cases, the height of the uplift is regarded to be the same along the whole length of the deformation area. Additionally, for this event the propagation is assumed to be simultaneous, with all the polygons across the margin to move together for 150 seconds. This kind of deformation is implausible. The shape of the deformation investigated for this "limiting" event is presented in Figure 5.44.

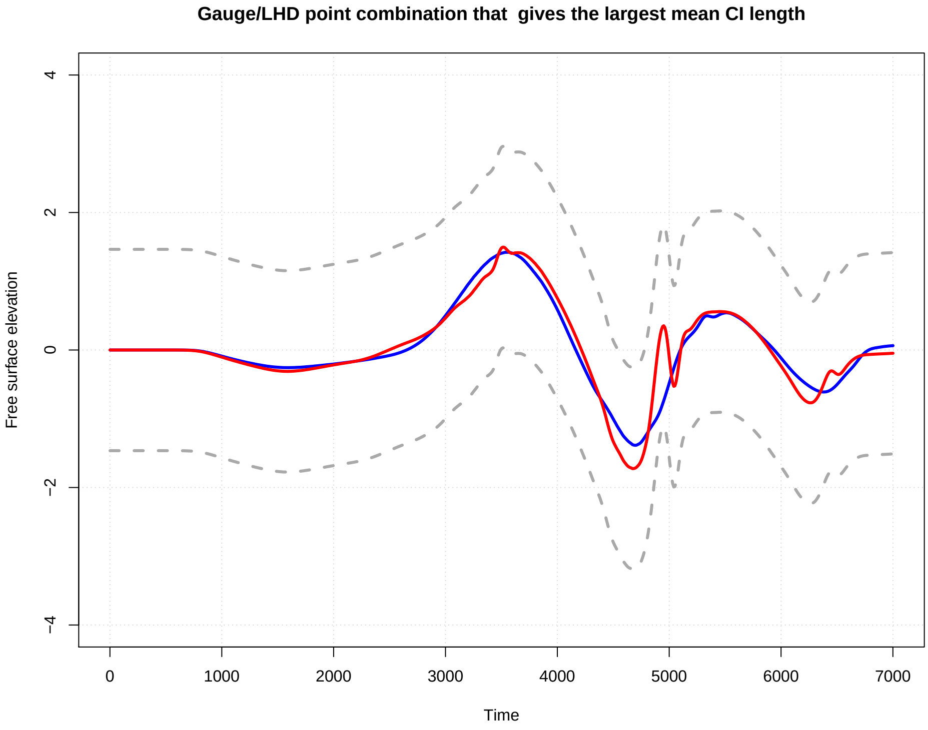
VOLNA evaluations are obtained for this limiting event bathymetry deformation for a long time interval, up to 10,000 seconds, at a large number of gauges covering the whole computational domain. These gauges are located at the barycenters of all the triangles in the computational mesh, excluding the ones located more than $10m$ above the sea level. The whole domain is covered by an unstructured triangular mesh which consists of 2,392,352 triangles. Using these evaluations, inundation maps are produced and visualised using Google Earth.

5.6.1 Inundation maps

The wave elevation time series evaluations resulting from the "limiting" event case are used for the generation of the worst case scenario inundation map for the Cascadia area. For the inundation map only the triangles located at the coast are considered. At the process of the selection

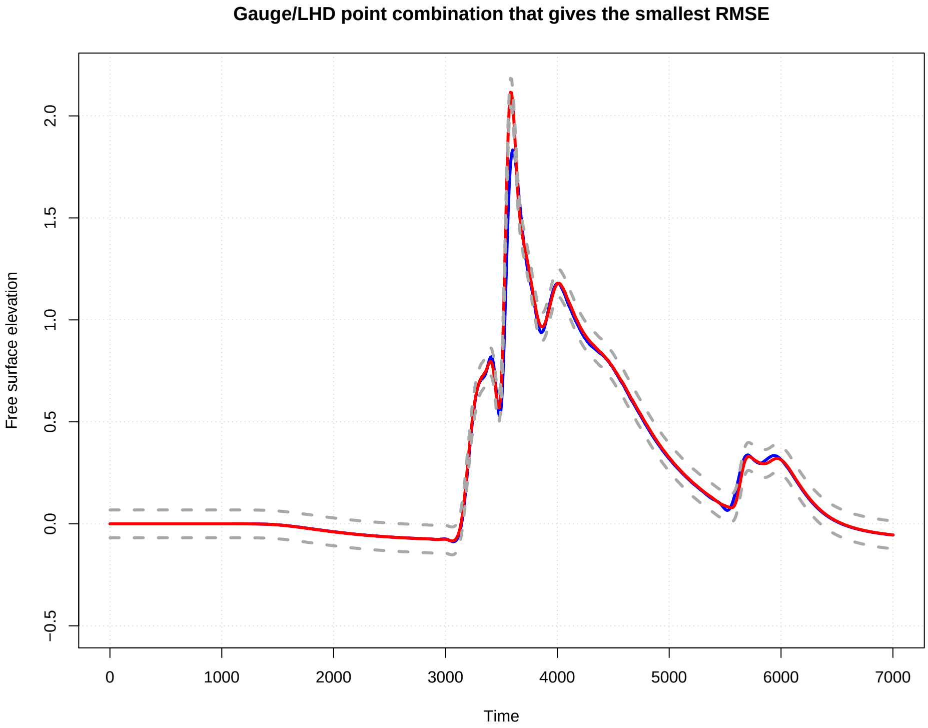


(a) Largest RMSE

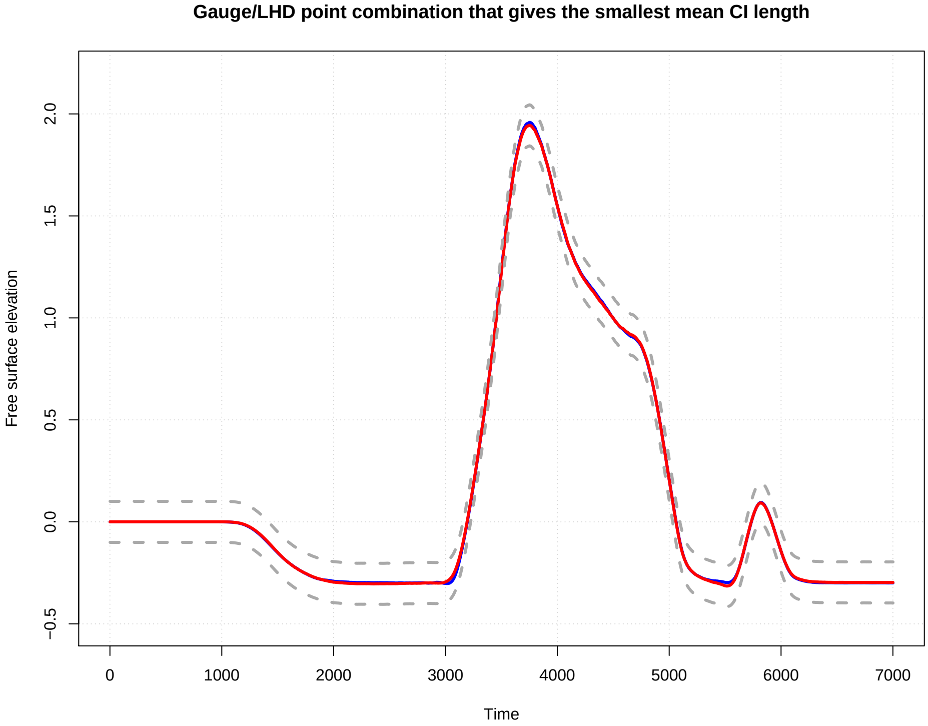


(b) Largest mean CI lengths

Figure 5.42: LOO diagnostics of gauge and LHD input point combinations results in the worst emulators predictions out of the 695 emulations.



(a) Smallest RMSE



(b) Smallest mean CI lengths

Figure 5.43: LOO diagnostics of gauge and LHD input point combinations results in the best emulators predictions out of the 695 emulations.

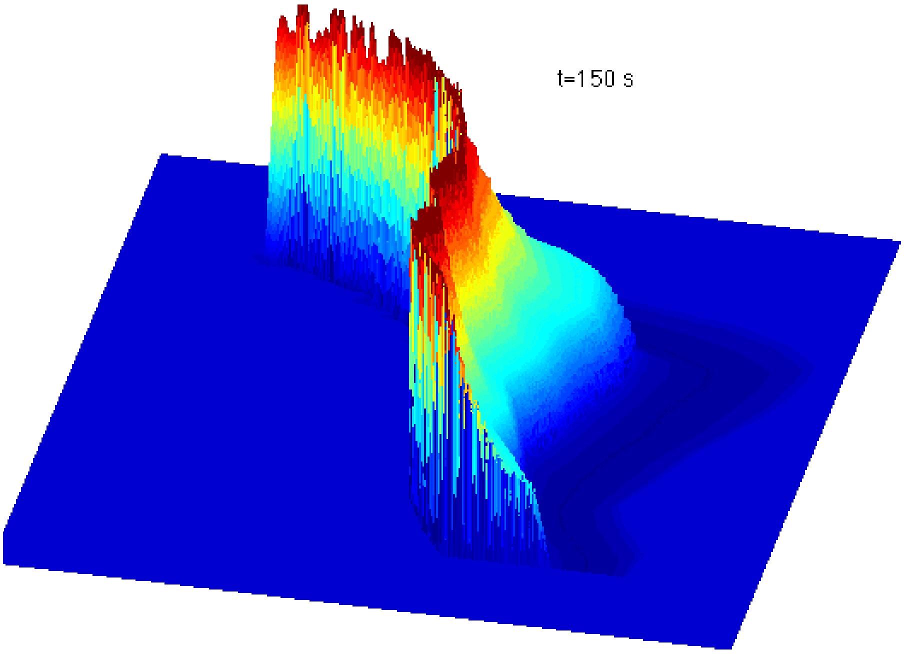


Figure 5.44: Bathymetry deformation for the case of the “limiting” event investigated.

of these triangles, the initial bathymetry of each the triangles barycenters is considered. In the case that this bathymetry is greater than zero, then the particular triangle is regarded as land triangle and it is included in the analysis. 833,087 triangles satisfy this condition and therefore considered as land initially. From these triangles only the ones that undergo water inundation are selected, which are only 49,955.

The number of unique nodes that compose these triangles is 36,316. However, some of these nodes belong also to triangles that are not in the 49,955 ones. Specifically, 45,175 triangles share nodes with the ones included in the analysis. The reason that these triangles are not included in the coast and inundated triangles, even the fact that they are neighbours of ones that included, is investigated. Some of the 45,175 triangles are not considered as land (25,113), as their initial bathymetry is below the water level. Looking at this in more detail, 0.5% of them have initial bathymetry greater than -0.001m, 2.5% greater than -0.01m and almost 30% greater than -0.1m. Also, around 69%, 81% and 87% of the 45157 triangles have barycenteres with initial bathymetry greater than -1m, -2m and -3m, respectively. Therefore, the reason that this large number of triangles is not considered as land is the too demanding criterion that defines coastal points only the ones that have initial bathymetry greater or equal to 0m from the sea level. However, due to the fact that these triangles are neighbours of the ones included, it means that they are very close to the coast, or even at the coast in case of a steep topography.

Since only the barycenters are tested for the initial bathymetry, there might be cases of triangles covering both land and water areas, which are located in deep sloping bathymetry location and therefore have below water surface initial bathymetry at the barycenter, even the fact that they are partly in land. These triangles should have been included in the inundation map. By not including the specific cases of triangles, gaps are generated around the coastline in the inundation map. In order to avoid these unnecessary gaps in the inundation map, the triangles which are neighbours of the 49955 ones, have initial bathymetry greater or equal to -10m and present water inundation, are included in the analysis. By including these triangles, a sufficient idea of the maximum possible inundation in the area can be obtained and also an impressive inundation map can be generated.

The range of the maximum water elevation at all the triangles is divided into five levels that approximately represent actual building floors, and each level is plotted with different colour. Google Earth is used to visualise the water inundations and it is presented in Figures 5.45, 5.46.

From Figures 5.45 and 5.46a it can be concluded that at Victoria area, where the largest city in the investigated area is located, high inundations are obtained. The major inundation at the Victoria area is taking place in the central business district around the harbour with a tsunami wave elevation estimated up to $7.5\mathrm{m}$. For the Grays Harbor area, which is presented in Figure 5.46d, very high tsunami wave elevation predictions are observed, with even more than $10\mathrm{m}$ amplitudes. Similarly, at Columbia River estuary, deep inundation of the populated coast can be obtained in Figure 5.46c with the inundation to be extended for more than $30\mathrm{km}$ upriver.

5.7 Results and Discussion

Subduction zone earthquakes can be accompanied by significant seabed deformations, uplift and subsidence, resulting in devastating tsunami events. In this chapter, the effect of the co-seismic seabed deformation on the resulting tsunami wave elevation is investigated, focused on the Cascadia Subduction Zone. Different event scenarios are examined by considering different earthquake source characteristic combinations. The source deformation shape is realistically represented using quadratic curves to describe the bathymetry uplifted and subside parts and also polygons are used to realistically describe the geometry of the deformation area. For each of these different scenarios, which are selected using an Experimental Design method, the advanced tsunami computer model VOLNA is evaluated, resulting to the wave elevation time series at many gauges in this area, particularly focused in the most significant areas of Victoria and Grays Harbor. Some assumptions are used for the seabed displacement representation, such as that the subsidence is always 15% of the uplifted height and also that each block is propagated for 150 seconds. An investigation for the future could be a sensitivity analysis of

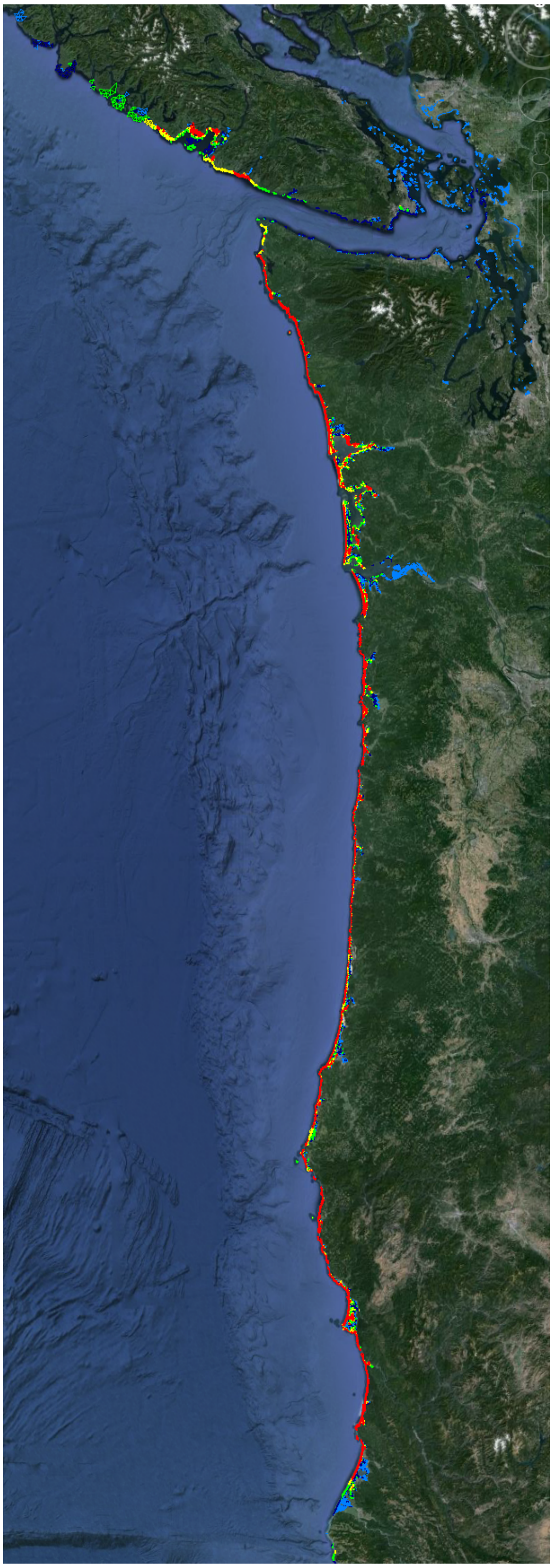


Figure 5.45: Water inundation at the coast for the whole domain. The colours represent different levels of inundation, where light blue: [0m,2.5m], dark blue: [2.5m,5m], green: [5m,7.5m], yellow: [7.5m,10m], red:>10m.

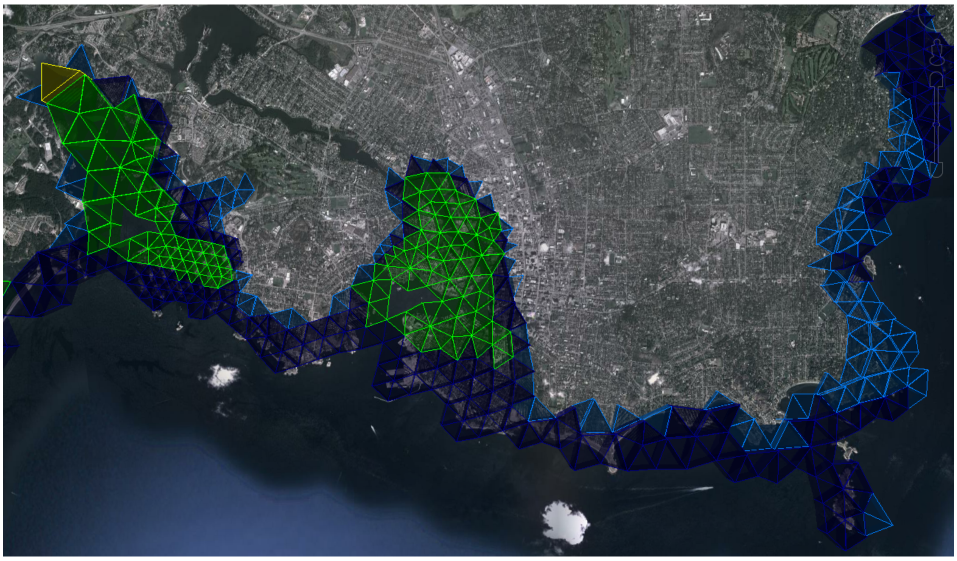
the wave elevation on changes in these parameters that are now considered as fixed.

A dense unstructured triangular mesh is used as input together with the bathymetry deformation as changes with time. The computational demanding VOLNA evaluations have become feasible with the beneficial use of the Emerald supercomputer. Using these evaluations, statistical emulation has been applied in order to generate an accurate statistical surrogate model that can make feasible computational demanding analyses such as sensitivity analyses and uncertainty quantifications. Also, the resulting emulator can be used for fast predictions in emergency cases and also for any event case, which is particularly useful in case of an actual event.

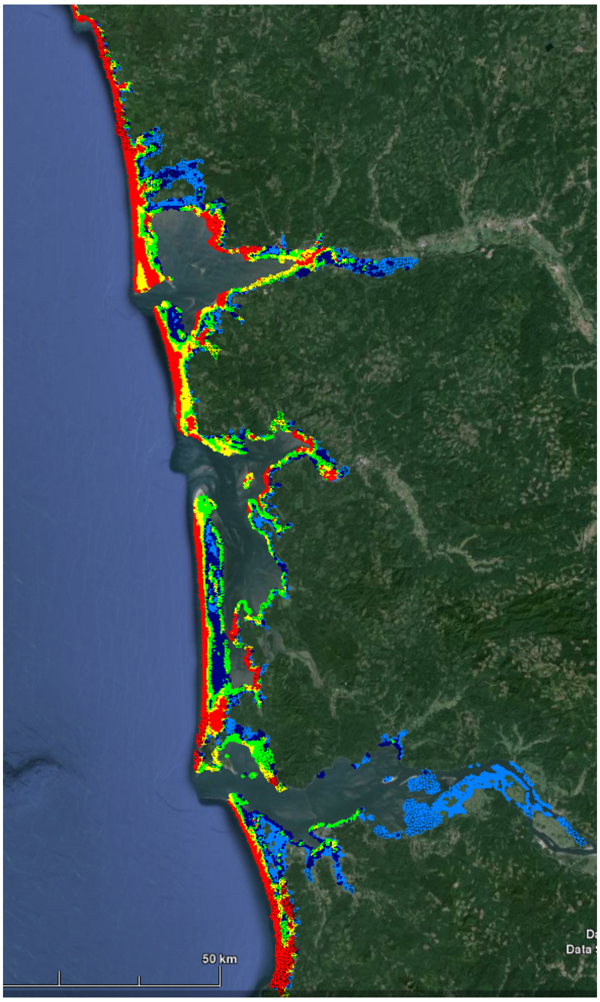
Tsunami warnings are announced by the local Warning Centres. These warnings have to be fast, but also accurate. Many precomputed event scenarios are available and at the time of the event, considering all the information accessible for the earthquake source. The Warning Centres usually find which of the precomputed scenarios is the “closest” to the actual event just happened and the warnings announced are based on this particular scenario. Unfortunately, this is not the best way to conclude to the warnings, since in many cases, the real event case with the precomputed scenario case are not that “close”. This means that there are large uncertainties in predictions, even nowadays, after so many years of research on tsunami and earthquake events. The use of the emulator in the particular case could be extremely beneficial, as the warnings could be fast and based on the actual event source characteristics and furthermore provide uncertainties attached to the predictions. As it is proved in the analysis in this chapter, the emulator’s predictions can be highly accurate. Therefore, the use of a surrogate model in place of the actual computer model is a big advantage and almost necessary in these cases.

For the generation of the statistical emulator, landmark registration as well as Functional Principal Component analysis have been applied, resulting in impressively accurate predictions. Furthermore, the resulting emulator is used for sensitivity analysis and uncertainty quantification, highlighting the necessity of the use of the emulator in place of the expensive-to-run computer model to make these analyses feasible. Moreoever, an initial attempt for the simultaneous generation of emulators at a large number of locations is presented in this chapter. Unfortunately, due to the bad choice of landmarks for the registration of the data, predictions are obtained only for 695 out of 3023 locations investigated. For the 695, the emulators have be validated using LOO diagnostics and it is concluded that the emulator predictions are very satisfactory for most of the cases. The improvement in the registration process is an analysis that is going to be completed in the future.

Finally, at the end of this chapter, an investigated for a “limiting” case event is performed. Specifically, an extreme event that is far from realistic is generated. This event considers very high uplift at the trench, up to $15m$, for the whole length of the deformation, with the middle maximum uplift to be three time less and therefore up to $5m$. We investigated such an implausible event in order to obtain the maximum plausible coastal inundation in this area, which can help in preparation for the next events and as a result to prevent large disasters. This is an efficient way since just a single run of the computer model is necessary to obtain the worst case scenario. Inundation maps are produced for this event and are visualised using Google Earth, presenting which locations are more vulnerable in tsunami wave inundations in case of an earthquake event.

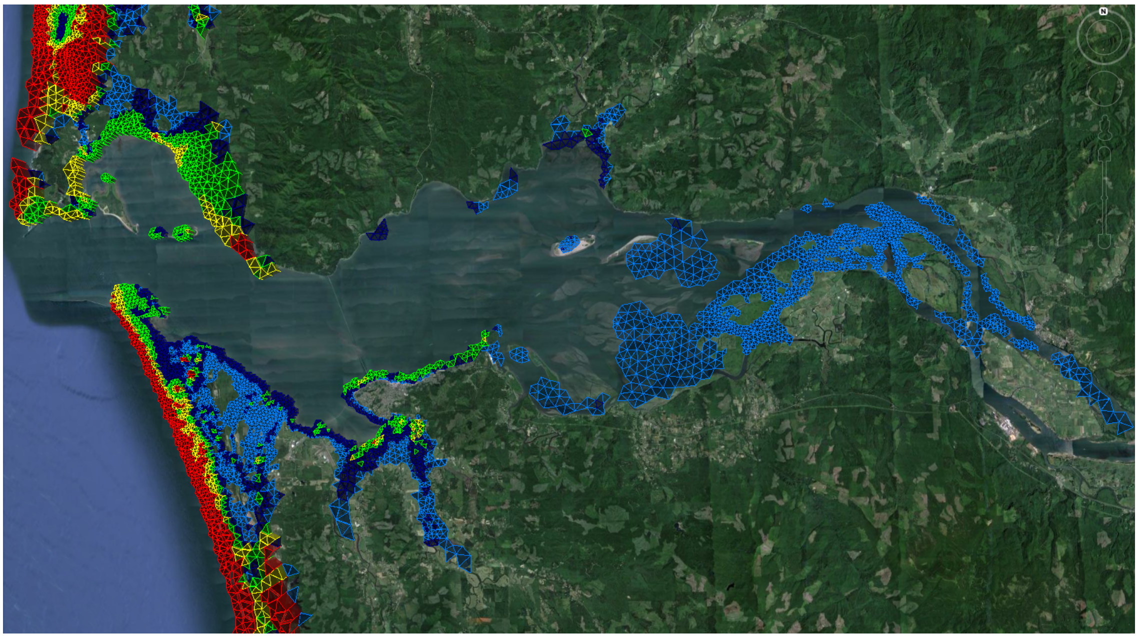


(a) Victoria area

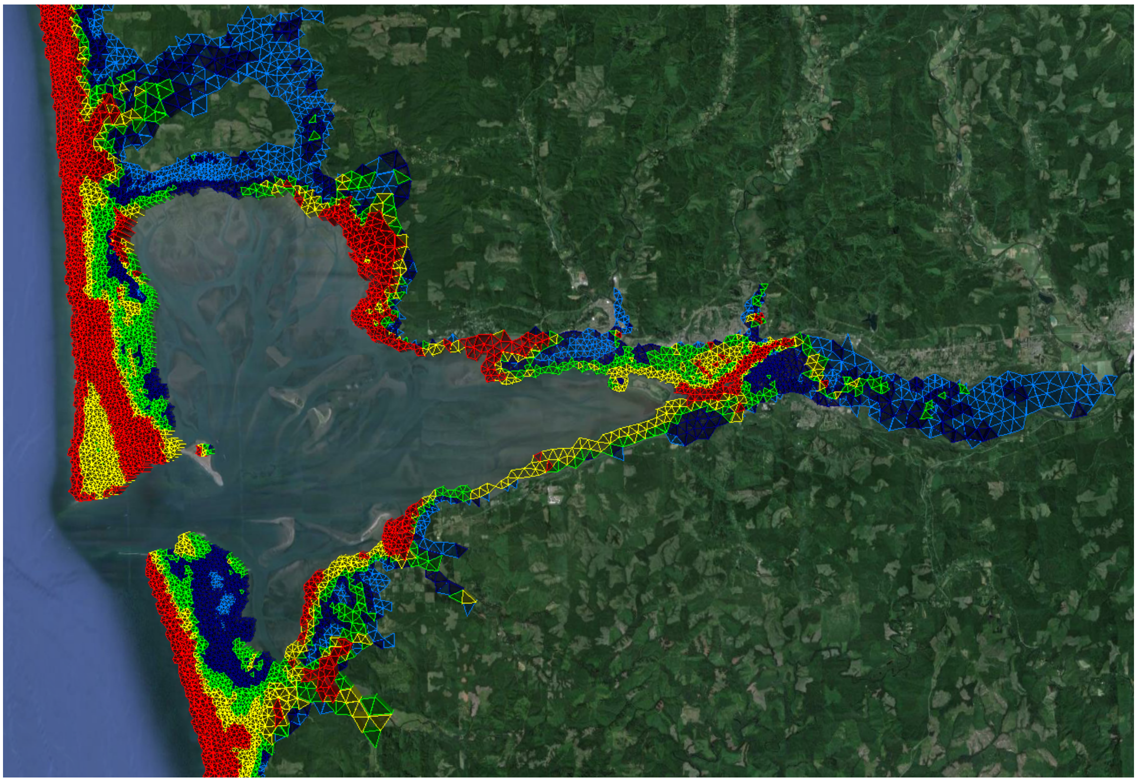


(b) South Washington area

Figure 5.46: Water inundation at coast focusing on specific areas. The colours represent different levels of inundation, where light blue: [0m,2.5m], dark blue: [2.5m,5m], green: [5m,7.5m], yellow: [7.5m,10m], red:>10m.



(c) Columbia river area



(d) Grays Harbor area

Figure 5.46: Water inundation at coast focusing on specific areas. The colours represent different levels of inundation, where light blue: [0m,2.5m], dark blue: [2.5m,5m], green: [5m,7.5m], yellow: [7.5m,10m], red:>10m.

Chapter 6

Conclusions and Further Work

6.1 Statistical Emulation of a landslide-generated tsunami model

A proof-of-concept case study for the statistical analysis of an analytical landslide-generated tsunami model was performed in the first part of the thesis. The main strategy followed was to build a statistical emulator, specifically an Outer Product Emulator, that is a fast accurate representation of the analytical model and subsequently use it for fast predictions, quantification of uncertainty and sensitivity analysis. The Bayesian framework enables the creation of an emulator by combining prior knowledge of the computer model properties with few carefully chosen model evaluations. These model evaluations resulted from the combinations of three inputs, which are features of the landslide. Specifically, these inputs are the initial location, speed and shape of the landslide, and they selected using “maximin” Latin Hypercube Design. The resulting emulator was validated using Leave-One-Out diagnostics that showed impressively accurate predictions. Following that, the emulator used for sensitivity analysis and uncertainty quantification, with the time required for these analyses very short compared to the time would have been required if the expensive-to-run computer model was used instead. The results in this part imply that approximation of the computer model with an accurate statistical emulator ends up to very positive consequences. Therefore, the application of statistical emulation for tsunami events investigations is recommended as it gives the capability for more advanced examination of the current tsunami models available, due to the flexibility to obtain as many model evaluations as necessary. The generation of improved tsunami models could result in more accurate predictions and this is definitely necessary for the case of tsunami events where human lives are in danger.

6.2 Sensitivity of tsunami models to different representations of the coseismic source deformation

In the second part of the thesis, the earthquake source deformation is extensively examined. The motivation for this analysis was that even nowadays, after so many years of scientific research in the area of earthquake-generated tsunamis, the description of the earthquake source is highly uncertain. So far, the most common approach in literature to obtain the source features is by inverse modelling, that is applied after the occurrence of the events. In this approach, the wave amplitude measurements observed at some gauges after an event are used to obtain the source characteristics that give the most close predictions possible to the actual wave elevations. A more recent approach to describe the source is with blocks. Nevertheless, strong assumptions are implied by this type of description, such as fixed elevation at the trench of the deformation.

In the research at this chapter of the thesis, a novel realistic seabed displacement is generated for this investigation, that is flexible and uses quadratic curves to represent the surface of the deformation. Simple bathymetry with slope is used to represent the seabed and its deformation is given as input to the advanced accurate tsunami model, VOLNA. Following that, VOLNA is evaluated to obtain the resulting wave elevation at a set of gauges. In order to assess the effect of the bathymetry deformation on the wave elevation, our advanced source description has been compared with four simpler bathymetry deformation representations: one block, four blocks, sixteen blocks and “piano keys”. Specifically, VOLNA model simulations performed for the four simpler cases, and the wave elevations obtained are compared with the ones obtained from the more advanced representation. This gives an insight whether or not the bathymetry representation in the case of an earthquake affects significantly the model output. It is concluded that the description of the source is particular important for earthquake-generated tsunami events. This is due to the fact that large differences on the amplitude of the waves are obtained from different source descriptions. Therefore, a simplistic unrealistic representation of the source could impose large biases and uncertainties in the predictions.

6.3 Cascadia study using VOLNA evaluations

In the last part, earthquake-generated tsunami events on the Cascadia Subduction Zone were examined. The analysis concentrated both on the generation phase of the tsunami, but also in the propagation and inundation. The earthquake source was modelled using an advanced representation which tried to mimic as close as possible reality, based on the existing information available about the different aspects of the area investigated. Following that, statistical emulation was applied to the resulting tsunami wave time series. Finally, sensitivity analysis and uncertainty quantification were performed using the resulting statistical emulator.

For the description of the earthquake source, quadratic curves are used to describe the subsidence and uplift surfaces and polygons to describe the geometry of the source. The specific bathymetry displacement representation is similar to the one investigated in Chapter 4. It is very flexible, avoiding unnecessary simplifications and assumptions. However, two assumptions were made for the source description in order to reduce the input parameters to just three and make the investigation tractable. Specifically, the subsidence was assumed to be 15% of the uplift and also the propagation time for each block was assumed fixed and equal to 150 seconds. A further analysis where these assumptions are removed and instead both the amplitude of the subsidence and the blocks' propagation time are considered as inputs parameters could be performed in the future. However, this will increase the computational burden of the analysis due to additional parameters.

This advanced bathymetry deformation together with real bathymetric and topographic data was used as inputs to the VOLNA tsunami model, which was evaluated to obtain the tsunami wave elevation time series at a large number of gauges covering the whole area, especially focusing on the Victoria and Grays Harbor areas. The approach for this problem was to consider different hypothetical events, in order to capture many possible earthquake event cases that can possibly happen in the future. These forty cases were generated by different combinations of three source characteristics, which are the maximum uplift height, the ratio of the trench to the maximum uplift heights and also the duration of the deformation propagation. The inputs combinations are selected using experimental design methods in order to cover efficiently the input computational domain.

The resulting wave elevation time series were used for statistical emulation and specifically for Outer Product Emulation. Landmark registration and Functional Principal Component analysis have been applied in the emulation process, resulting to a very accurate statistical representation of the computer model. Using Leave-One-Out validation diagnostics and estimating the mean Credible Interval length and RMSE of the predictions, it is concluded that emulator predictions are improved by the use of registration and Functional Principal Component analysis. The statistical emulator can be used for fast prediction in crucial situations, such as real-time predictions at Tsunami Warning Centres, and also for analyses that require a large number of model evaluations, such as uncertainty and sensitivity analyses. The use of emulators at Warning Centres would be very beneficial, since they can be evaluated in a short time at any combination of source characteristics. Hence, the Warning Centres could provide fast predictions based on the actual event source characteristics, instead of the "closer" precomputed event case. Additional beneficiaries are insurance companies and coastal planning agencies.

From the sensitivity analysis performed using the statistical emulator, it is concluded that the wave elevation is independent of the rupture propagation duration. On the other hand, it is strongly correlated to the maximum uplift height. Additionally, the wave amplitude is slightly positively correlated to the ratio of the trench to the maximum heights. The sensitivity analysis was also extended to the dependency of the model output to pairs of inputs. In total, 31696 model evaluations have been used for this demanding analysis and the total time required was impressively very short.

Furthermore, the statistical emulator was used for uncertainty analysis, in order to understand how the uncertainty in inputs affects the total uncertainty in the predictions. Due to the fact that in tsunami events investigations the description of the source and its characteristics are usually highly uncertain, the three inputs were assigned distributions based on the literature and expert knowledge. The distributions were used to obtain 2000 random samples for each of the source characteristics. These samples were combined to form 2000 different event case scenarios, with the positive correlation of the maximum uplift to the ratio of the trench to maximum heights to be taken into consideration. Using the emulator, the maximum wave elevation of the time series for each of these events was obtained and conclusions were drawn for the distribution of the maximum elevation and the uncertainty in the predictions. As for the case of the sensitivity analysis, the time required for the analysis was very short.

Therefore, the large benefit of using emulators in place of the computer models is highlighted, since both of the sensitivity and uncertainty analyses would not be feasible to be performed in such a high resolution using the expensive computer model. However, the best way to perform uncertainty quantification would be to consider all the possible sources of uncertainty, such as uncertainty in the data, model and prior informations, and combine them together to obtain the overall uncertainty in predictions, instead of focusing just to the uncertainty due to uncertain inputs. This is a challenging effort.

Additionally, an automatic process for simultaneous application of statistical emulation at 3023 gauges covering the whole area was presented in this last chapter. For the emulation, the Landmark Registration method was used in which landmarks were automatically selected for all the locations. However, this process resulted in error for a large number of locations and this did not let the emulation process to be completed. Specifically, statistical emulators successfully generated for only 695 gauges. Functional Principal Component analysis was employed on the registered data for the selection of the output regression functions. The parameters required were estimated for all the locations simultaneously. This analysis was not computationally demanding, due to the fact that it was automated and it was running in parallel for all the locations at Emerald supercomputer. The conclusion from this investigation is that the emulator's performance was very satisfactory in almost all the 695 cases investigated. Nevertheless, the process could be improved further in the registration step in order to avoid errors due to the selection of landmarks. This can be achieved by a better choice of the landmarks locations or by using another registration method, such as the continuous registration that does not require landmarks specification.

The final part of the analysis in this chapter is the investigation of a "limiting" case event. This particular event case is almost impossible to happen in real life, as it assumes very high rupture elevation at the trench along the whole margin. However, the reason of this analysis was to obtain the maximum possible coastal inundation, and this is definitely an efficient way since just a single run of the tsunami model is required. Inundation maps are obtained from this analysis, which are visualised using Google Earth, and could be used by insurance and planning agencies.

Bibliography

Anastasiou, K. and Chan, C. (1997). Solution of the 2d shallow water equations using the finite volume method on unstructured triangular meshes. *International Journal for Numerical Methods in Fluids* **24,** 1225–1245.

Atwater, B. F., Satoko, M.-R., Kenji, S., Yoshinobu, T., Kazue, U., and Pararas-Carayannis, G. (2006). The orphan tsunami of 1700. *Science of Tsunami Hazards* **24,** 49.

Atwater, F. B., Nelson, R. A., Clague, J. J., Carver, A. G., Yamaguchi, K. D., Bobrowsky, T. P., Bourgeois, J., Darienzo, E. M., Grant, C. W., Hemphill-Haley, E., Kelsey, M. H., Jacoby, C. G., Nishenko, P. S., Palmer, P. S., Peterson, D. C., and Reinhart, M. A. (1995). Summary of Coastal Gecologic Eevidence for Psat Great Eearthquakes at the Cascadia Subduction Zone. *Earthquake Spectra* **11,** 1–18.

Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., and Perthame, B. (2004). A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing* **25,** 2050–2065.

Babuška, I., Nobile, F., and Tempone, R. (2007). Reliability of computational science. *Numerical methods for partial differential equations* **23,** 753–784.

Bardet, J. P., Synolakis, C. E., and Davies, H. L. (2003). Landslide tsunamis: recent findings and research directions. *Pure Appl. Geophys.* **160,** 1793–1809.

Bastos, L. (2010). *Validating Gaussian Process Models in Computer Experiments*. PhD thesis, University of Sheffield.

Bastos, L. S. and O'Hagan, A. (2009). Diagnostics for gaussian process emulators. *Technometrics* **51,** 425–438.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007). A framework for validation of computer models. *Technometrics* **49,** 138–154.

Beck, J. and Guillas, S. (2014). Sequential design with mutual information for computer experiments (MICE): Emulation of a tsunami model.

Bowman, V. and Woods, D. C. (2012). Thin-plate spline emulation of multivariate simulators with application to atmospheric dispersion.

Byrne, D. E., Davis, D. M., and Sykes, L. R. (1988). Loci and maximum size of thrust earthquakes and the mechanics of the shallow region of subduction zones. *Tectonics* **7,** 833–857.

Chang, W. (2014). Supplement of probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties. *Supplement of Geosci. Model Dev.* **7,** 1933–1943.

Chang, W., Applegate, P. J., Haran, M., and Keller, K. (2014). Probabilistic calibration of a greenland ice sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties. *Geoscientific Model Development Discussions* **7,** 1905–1931.

Chui, C. K. (2014). *An introduction to wavelets*, volume 1. Academic Press.

Clague, J. J. (1997). Evidence for large earthquakes at the cascadia subduction zone. *Reviews of Geophysics* **35,** 439–460.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* **86,** 953–963.

Dominey-Howes, D., Dunbar, P., Varner, J., and Papathoma-Kohle, M. (2010). Estimating probable maximum loss from the Cascadia tsunami. *Nat Hazards* **53,** 43–61.

Dutykh, D. and Dias, F. (2010). Influence of sedimentary layering on tsunami generation. *Computer Methods in Applied Mechanics and Engineering* **199,** 1268–1275.

Dutykh, D., Dias, F., and Kervella, Y. (2006). Linear theory of wave generation by a moving bottom. *C. R. Acad. Sci. Paris, Ser. I* **343,** 499–504.

Dutykh, D., Mitsotakis, D., Gardeil, X., and Dias, F. (2013). On the use of finite fault solution for tsunami generation problems. *Theor. Comput. Fluid Dyn.* **27,** 177–199.

Dutykh, D., Poncet, R., and Dias, F. (2011). The VOLNA code for the numerical modelling of tsunami waves: generation, propagation and inundation. *Eur. J. Mech. B-Fluid* **30,** 598–615.

Enet, F., Grilli, S. T., and Asce, M. (2005). Tsunami landslide generation: Modelling and experiments. *Ocean Waves Measurement and Analysis, Fifth International Symposium WAVES 2005* .

Enet, F., Grilli, S. T., and Asce, M. (2007). Experimental study of tsunami generation by three-dimensional rigid underwater landslides. *J. Waterw. Port Coast. Ocean Eng* .

Fricker, T. E., Oakley, J. E., and Urban, N. M. (2013). Multivariate gaussian process emulators with nonseparable covariance structures. *Technometrics* **55,** 47–56.

Garcia-Navarro, P. and Vazquez-Cendon, M. E. (2000). On numerical treatment of the source terms in the shallow water equations. *Computers & Fluids* **29,** 951–979.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics* **18,** 681–695.

Gesch, D. B., Oimoen, M. J., and Evans, G. A. (2014). Accuracy assessment of the U.S. Geological Survey National Elevation Dataset, and comparison with other large-area elevation datasets - SRTM and ASTER. Technical report, U.S. Geological Survey Open-File Report.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association* **97,** 590–600.

Goldfinger, C. (2011). Submarine paleoseismology based on turbidite records. *Annual review of marine science* **3,** 35–66.

Goldfinger, C., Nelson, C. H., and Johnson, J. E. (2003). Holocene earthquake records from the Cascadia subduction zone and northern San Andreas Fault based on precise dating of offshore turbidites. *Annual Review of Earth and Planetary Sciences* **31,** 555–577.

Goldfinger, C., Nelson, C. H., Johnson, J. E., Morey, A. E., Gutierrez-Pastor, J., Karabanov, E., Eriksson, A. T., Gracia, E., Dunhill, G., Patton, J., Enkin, R., Dallimore, A., and Vallier, T. (2012). Turbidite event history: Methos and implications for Holocene paleoseismicity of the Cascadia Subduction Zone. *U.S. Geol. Surv. Professional Paper* **1661,**.

Goldstein, M. and Rougier, J. (2009). Reified bayesian modelling and inference for physical systems. *Journal of statistical planning and inference* **139,** 1221–1239.

Grezio, A., Marzocchi, W., Sandri, L., and Gasparini, P. (2010). A Bayesian procedure for Probabilistic Tsunami Hazard Assessment. *Nat. Hazards* **53,** 159–174.

Grilli, S. and Watts, P. (2005). Tsunami generation by submarine mass failure part I: Modeling, experimental validation and sensitivity analyses. *J. Waterw. Port Coast. Ocean Eng, ASCE* **131,** 283–297.

Grilli, T. S., Ioualalen, M., Asavanant, J., Shi, F., Kirby, T. J., and Watts, P. (2007). Source Cconstraints and Model Simmolation of the December 26, 2004, Indian Odean Ttsunami. *J. Waterway, Port, Coastal, Ocean Engineering* **133,** 414–428.

Guilbert, J., Vergoz, J., Schissele, E., Roueff, A., and Cansi, Y. (2005). Use of hydroacoustic and seismic arrays to observe rupture propagation and source extent of the Mw= 9.0 Sumatra earthquake. *Geophysical research letters* **32,**.

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *Applied Statistics* **38,** 1–50.

Hodges, J. S. (2013). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. CRC Press.

Hyndman, R. D. and Wang, K. (1995). The rupture zone of cascadia great earthquakes from current deformation and the thermal regime. *Journal of Geophysical Research: Solid Earth (1978-2012)* **100,** 22133–22154.

Imamura, F. and Hashi, K. (2002). Re-examination of the source mechanism of the 1998 Papua New Guinea Earthquake and Tsunami. *Pure Appl. Geophys.* **160,** 2071–2086.

Joseph, V. R. and Hung, Y. (2008). Orthogonal-Maximin Latin Hypercube Designs. *Stat. Sinica* **18,** 171–186.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *J. Roy. Stat. Soc. B Met* **63,** 425–464.

Kervella, Y., Dutykh, D., and Dias, F. (2007). Comparison between three-dimensional linear and nonlinear tsunami generation models. *Theoretical and computational fluid dynamics* **21,** 245–269.

Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103,** 1155–1165.

Kulkarni, R., Wong, I., Zachariasen, J., Goldfinger, C., and Lawrence, M. (2013). Statistical Analyses of Great Eearthquake Recurrence along the Cascadia Subduction Zone. *Bulletin of the Seismological Society of America* **103,** 3205–3221.

Lay, T., Ammon, C. J., Kanamori, H., Xue, L., and Kim, M. J. (2011). Possible large near-trench slip during the 2011 M (w) 9. 0 off the Pacific coast of Tohoku Earthquake. *Earth, planets and space* **63,** 687–692.

Lay, T., Kanamori, H., Ammon, C. J., Nettles, M., Ward, S. N., Aster, R. C., Beck, S. L., Bilek, S. L., Brudzinski, M. R., Butler, R., et al. (2005). The great Sumatra-Andaman earthquake of 26 december 2004. *Science* **308,** 1127–1133.

Lay, T., Yamazaki, Y., Ammon, J. C., Cheung, F. K., and Kanamori, H. (2011). The 2011 Mw 9.0 off the Pacific Coast of Tohoku Earthquake: Comparison of deep-water tsunami signals with finite-fault rupture model predictions. *Earth Planets Space* **63,** 797–801.

Leonard, L. J., Currie, C. A., Mazzotti, S., and Hyndman, R. D. (2010). Rupture area and displacement of past cascadia great earthquakes from coastal coseismic subsidence. *Geological Society of America Bulletin* **122,** 2079–2096.

Liu, P., Woo, S., and Cho, Y. (1998). Computer programs for tsunami propagation and inundation. *Cornell University, Ithaca, NY 14850, USA* .

Liu, P. L.-F., Lynett, P., and Synolakis, C. E. (2003). Analytical solutions for forced long waves on a sloping beach. *J. Fluid Mech.* **478,** 101–109.

Liu, P. L.-F., Wu, T.-R., Raichlen, F., Synolakis, C. E., and Borrero, J. C. (2005). Runup and rundown generated by three-dimensional sliding masses. *J. Fluid Mech.* **536,** 107–144.

Lorito, S., Selva, J., Basili, R., Romano, F., Tiberti, M. M., and Piatanesi, A. (2014). Probabilistic Hazard for Seismically-Induced Tsunamis: Accuracy and Feasibility of Inundation Maps (accepted). *Geophysical Journal International* .

Lynett, P. and Liu, P. (2005). A numerical study of the run-up generated by three-dimensional landslides. *J. Geophys. Res.* **110,**.

Masterlark, T. and Hughes, K. L. (2008). Next generation of deformation models for the 2004 M9 Sumatra-Andaman earthquake. *Geophysical Research Letters* **35,**.

McKay, M., Beckman, R., and Conover, W. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21,** 239–245.

Murty, T. S. (1979). Submarine slide-generated water waves in Kitimat Inlet, British Columbia. *J. Geophys. Res.* **84,** 7777–7779.

Murty, T. S. (2003). Tsunami wave height dependence on landslide volume. *Pure Appl. Geophys.* **160,** 2147–2153.

Nason, G. (2010). *Wavelet methods in statistics with R.* Springer Science & Business Media.

Nelson, A. R., Sawai, Y., Jennings, A. E., Bradley, L.-A., Gerson, L., Sherrod, B. L., Sabean, J., and Horton, B. P. (2008). Great-earthquake paleogeodesy and tsunamis of the past 2000 years at Alsea Bay, central Oregon coast, USA. *Quaternary Science Reviews* **27,** 747–768.

NRC (2012). *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification.* Natl. Acad. Press, Washington DC.

Oakley, J. E. and O'Hagan, A. (2002). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. Roy. Stat. Soc. B Met* **66,** 751–769.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Syst. Safe.* **91,** 1290 – 1300.

Okada, Y. (1992). Internal deformation due to shear and tensile faults in a half-space. *Bulletin of the Seismological Society of America* **82,** 1018–1040.

Panizzo, A. and De Girolamo, P. (2005). Forecasting impulse waves generated by subaerial landslides. *J. Geophys. Res.* **110,**.

Poncet, R., Campbell, C., Dias, F., Locat, J., and Mosher, D. (2010). A study of the tsunami effects of two landslides in the st. lawrence estuary. In *Submarine Mass Movements and Their Consequences*, pages 755–764. Springer.

Power, W., Downes, G., and Stirling, M. (2007). Estimation of Tsunami Hazard in New Zealand due to South American Earthquakes. *Pure Appl. Geophys.* **164,** 547564.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Function Data Analysis with R and MATLAB.* Springer.

Ramsay, J. O. (2006). *Functional data analysis.* Wiley Online Library.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* MIT Press.

Rogers, G. C. (1988). An assessment of the megathrust potential of the Cascadia subduction zone. *Can. J. Earth Sco.* **25,** 844–852.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *J. Comput. Graph. Stat.* **17,** 827–843.

Rougier, J. (2011). A representation theorem for stochastic processes with separable covariance functions and its implications for emulation. *under review* .

Rougier, J., Guillas, S., Maute, A., and Richmond, A. D. (2009). Expert knowledge and multivariate emulation: The thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Technometrics* **51,** 414–424.

Ruff, L. J. (2003). Some aspects of energy balance and tsunami generation by earthquakes and landslides. *Pure Appl. Geophys.* **160,** 2155–2176.

Saito, T., Ito, Y., Inazu, D., and Hino, R. (2011). Tsunami source of the 2011 Tohoku-Oki earthquake, Japan: Inversion analysis based on dispersive tsunami simulations. *Geophysical Research Letters* **38,**.

Sammarco, P. and Renzi, E. (2008). Landslide tsunamis propagating along a plane beach. *J. Fluid Mech.* **598,** 107–119.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer Series in Statistics.

Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation*, pages 130–143. Winter Simulation Conference.

Sarri, A., Guillas, S., and Dias, F. (2012). Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification. *Natural Hazards and Earth System Sciences* **12,** 2003–2018.

Satake, K. and Atwater, B. F. (2007). Long-term perspectives on giant earthquakes and tsunamis at subduction zones. *Annual Review of Earth and Planetary Sciences* **35,** 349–374.

Satake, K., Fujii, Y., Harada, T., and Namegaya, Y. (2013). Time and space distribution of coseismic slip of the 2011 Tohoku earthquake as inferred from tsunami waveform data. *Bulletin of the seismological society of America* **103,** 1473–1492.

Satake, K., Shimazaki, K., Tsuji, Y., and Ueda, K. (1996). Time and size of a giant earthquake in Cascadia inferred from Japanese tsunami records of January 1700. *Nature* **379,** 246–249.

Shuto, N. (1991). Numerical simulation of tsunamis - its present and near future. *Natural Hazards* **4,** 171–191.

Simons, M., Minson, S. E., Sladen, A., Ortega, F., Jiang, J., Owen, S. E., Meng, L., Ampuero, J.-P., Wei, S., Chu, R., et al. (2011). The 2011 magnitude 9.0 Tohoku-Oki earthquake: Mosaicking the megathrust from seconds to centuries. *Science* **332,** 1421–1425.

Spiller, T. E., Bayarri, J. M., James, O. B., Calder, S. E., Patra, K. A., Pitaman, B. E., and Wolpert, L. R. (2014). Automating Emulator Construction for Geophysical Hazard Maps. *SIAM/ASA J. Uncertainty Quantification* **2,** 126–152.

Sraj, I., K., M. T., Knio, M. O., Dawson, N. C., and Hoteit, I. (2014). Uncertainty Quantification and Inference of Mawning's Friction Coefficients using DART Buoy Data during the Tohoku tsunami. *Ocean Modelling* .

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stephanakis, T. (2013). *Tsunami Amplification Phenomena*. PhD thesis, Ecole Normale Superieure De Cachan and University College Dublin.

Suzuki, W., Aoi, S., Sekiguchi, H., and Kunugi, T. (2011). Rupture process of the 2011 Tohoku-Oki mega-thrust earthquake (M9. 0) inverted from strong-motion data. *Geophysical Research Letters* **38,**.

Synolakis, C., Bernard, E., Titov, V., Kânoğlu, U., and González, F. (2007). Standards, criteria, and procedures for NOAA evaluation of tsunami numerical models: Seattle, Washington, NOAA/Pacific Marine Environmental Laboratory. Technical report, Technical Memorandum OAR PMEL-135.

Synolakis, C. E., Bardet, J.-P., Borrero, J. C., Davies, H. L., Okal, E. A., Silver, E. A., Sweet, S., and Tappin, D. R. (2002). The slump origin of the 1998 Papua New Guinea Tsunami. *P. Roy. Soc. Lond. A Mat.* **458,** 763–789.

Synolakis, E. C., Bernard, N. E., Titov, V. V., Kanoglu, U., and Gonzalez, I. F. (2008). Validation and Verification of Ttsunami Nnumerical Modes. *Pure appl. geophys.* **165,** 2197–2228.

Tanioka, Y. and Seno, T. (2001). Sediment effect on tsunami generation of the 1896 Sanriku tsunami earthquake. *Geophysical Research Letters* **28,** 3389–3392.

Tappin, D. R., Watts, P., and Grilli, S. T. (2008). The Papua New Guinea tsunami of 17 July 1998: anatomy of a catastrophic event. *Nat. Hazard Earth Sys.* **8,** 243–266.

Tinti, S., Zaniboni, F., Pagnoni, G., and Manucci, A. (2008). Stromboli Island (Italy): Scenarios of tsunamis generated by submarine landslides. *Pure Appl. Geophys.* **165,** 2143–2167.

Titov, V., Rabinovich, B. A., Mofjeld, O. H., Thomson, E. R., and Gonzalez, I. F. (2005). The Global Reach of the 26 December 2004 Sumatra Ttsunami. *Science* **309,** 2045–2048.

Titov, V. V. (2009). Tsunami forecasting. *The sea* **15,** 371–400.

Titov, V. V., Gonzalez, F. I., Bernard, E., Eble, M. C., Mofjeld, H. O., Newman, J. C., and Venturato, A. J. (2005). Real-time tsunami forecasting: Challenges and solutions. In *Developing Tsunami-Resilient Communities*, pages 41–58. Springer.

Titov, V. V. and Synolakis, C. E. (1998). Numerical modeling of tidal wave runup. *Journal of Waterway, Port, Coastal, and Ocean Engineering* **124,** 157–171.

Urban, N. M. and Fricker, T. E. (2010). A comparison of latin hypercube and grid ensemble designs for the multivariate emulation of an earth system model. *Comput. Geosci.* **36,** 746–755.

Wahba, G. and Craven, P. (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31,** 377–404.

Watts, P. (1998). Wavemaker curves for tsunamis generated by underwater landslides. *J. Waterw. Port Coast. Ocean Eng, ASCE* **124,** 127–137.

Watts, P. (2000). Tsunami features of solid block underwater landslides. *J. Waterw. Port Coast. Ocean Eng, ASCE* **126,** 144–152.

Watts, P. (2004). Probabilistic predictions of landslide tsunamis off Southern California. *Mar. Geol.* **203,** 281 – 301.

Watts, P. and Grilli, S. (2003). Underwater landslide shape, motion, deformation and tsunami generation. *Proc. 13th Offshore and Polar Engineering Conf., ISOPE03, Honolulu, Hawaii* **3,** 364–371.

Watts, P., Grilli, S. T., Asce, M., and Fryer, G. J. (2005). Tsunami generation by Submarine Mass Failure II: Predictive equations and case studies. *J. Waterw. Port. Coastal. Ocean Eng* **131,** 283–297.

Watts, P., Grilli, S. T., Kirby, J. T., Fryer, G. J., and Tappin, D. R. (2003). Landslide tsunami case studies using a Boussinesq model and a fully nonlinear tsunami generation model. *Nat. Hazard Earth Sys.* **3,** 391–402.

Wells, D. L. and Coppersmith, K. J. (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America* **84,** 974–1002.

Wiegel, R. L. (1955). Laboratory studies of gravity waves generated by the movement of a submerged body. *Trans. AGU* **36,** 759–774.

Wood, N. and Goof, J. (2004). Vulnerability of port and harbor communities to earthquakes and tsunami hazards: the use of GIS in community hazard planning. *Coast Manage* **32,** 243–269.

Worldwide, A. (2013). Study of impact and the insurance and economic cost of a major earthquake in British Columbia and Ontario/Quebec. *Commissioned by the Insurance Bureau of Canada* .

Wright, S. G. and Rathje, E. M. (2003). Triggering mechanisms of slope instability and their relationship to earthquakes and tsunamis. *Pure Appl. Geophy.* **160,** 1865–1877.

Yeats, S. R., Sieh, K., and Allen, R. C. (1997). *The Geology or Earthquakes*. Oxford University Press.

Yue, H. and Lay, T. (2013). Source rupture models for the mw 9.0 2011 Tohoku earthquake from joint inversions of high-rate geodetic and seismic data. *Bulletin of the Seismological Society of America* **103,** 1242–1255.