Developing a Scenario to Validate the Tsunamibayes Methodology

Carol Herrera

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Jared P. Whitehead, Chair
Tyler Jarvis
Ron Harris

Department of Mathematics

Brigham Young University

ABSTRACT

Developing a Scenario to Validate the Tsunamibayes Methodology

Carol Herrera
Department of Mathematics, BYU
Master of Science

Modern quantitative measurements of earthquakes and tsunamis gives us a small glimpse into understanding seismic activity between tectonic plates. We can potentially widen this scope by using non-instrumental data from past earthquakes and tsunamis. Our research team has implemented a method to reconstruct historical earthquakes by developing a Python package called `tsunamibayes`. We have tested this package using the 1852 Banda Sea Earthquake and the 1820 Sulawesi Earthquake. Now we intend to validate our method using a recent earthquake, the 2004 Boxing Day Earthquake, which has instrumental seismic data to which we can compare our results. This paper covers the data collection, sample parameter selection, adjoint computation, and a novel way to interpolate topography and bathymetry files, for the validation process.

Keywords: Earthquake, Tsunami, ArcGIS, Sumatra, Bayesian Inference

Acknowledgements

I want to first thank God and my parents. I would like to thank my advisor, Dr. Jared Whitehead, who was an example to me of humility, professionalism, and composure throughout this project.

Thank you to the many people who have collaborated on this project. Thank you to Hayden Ringer, whose work was the basis of this research group and who was an inspiration to me. Isaac Sorensen has been a great resource to me with all of my questions. I value his patience and care for the work we have done together. Thanks to the ArcGIS employees of the BYU library: Teresa Gomez, Thomas Britt, and Matthew Rowley; who helped source the bathymetry files for this project. I want to also thank the following collaborators: Claire Ashcraft, Marja Crawford, and Taylor Paskett.

Thank you to all of my teachers who believed in me and helped me reach this point that I didn't think was possible. Specifically, thank you to Dr. Jarvis, who has been a mentor for me since my ACME years. Thank you to my graduate professors: Drs. Bakker and Barker; who pushed me and gave me confidence when I started my master's program, and to Dr. Ron Harris, who sits on my committee. I also want to thank a few other BYU faculty and staff for their kindness, specifically: Dr. Dorff, Dr. Kempton, Dr. Corey, and Lonette Stoddard.

Thank you to my husband Nicholas Glenn, who has been my biggest cheerleader during my master's program, who is helping me be a better person, and for editing this paper for grammar, syntax, and structure. I would like to thank all of the people who supported me to become the person I am today. I especially want to thank my friends Nathan Brown, who reminds me of the joy in life, and Matthew Gong, who believed in me first.

Contents

Contents iv

List of Tables v

List of Figures vi

1 Introduction 1

2 Historical Background 2

3 Data Collection 4

3.1 Observational Data 4

3.2 Using Bayesian Inference 8

3.3 Initial Parameter Selection 8

3.4 GeoClaw and the Adjoint 11

3.5 Bathymetry 13

4 Methodology 17

5 Conclusion 18

A The Overlay Model Python Script 19

Bibliography 25

List of Tables

3.1 Prior Distribution for the 2004 Sumatra Earthquake 12

List of Figures

2.1 Indonesian Plate Boundaries and Seismic Gaps 3

3.1 Wave Height Observation Example 5

3.2 2004 Telwatta Tsunami Train Wreck 6

3.3 Probability Distributions from Empirical Data 7

3.4 Observation Locations 14

3.5 Trials to Overlay Topography and Bathymetry Files 16

3.6 Overlay Model 17

4.1 The `tsunamibayes` Model 18

Chapter 1. Introduction

The goal of our research group is to provide the information needed to yield a hazard assessment for future seismic events and identify areas along fault lines that have not had seismic activity for decades. To attain this goal, we need seismic data, but modern methods of measuring earthquakes only goes back about 50 years. Using modern instrumental data alone to estimate the probabilities of future earthquakes is grossly inadequate [1]. So, how can we find the origin of historical earthquakes when they occurred before having the use of modern seismic instruments?

Our research group aimed to answer this question by developing the `tsunamibayes` package and with it, we modeled earthquakes that have occurred in the 1800's using historical data, Bayesian Inference, MCMC, and the GeoClaw software package. We have used `tsunamibayes` to investigate massive earthquakes and tsunamis that struck the Banda Neira islands in 1852 and Sulawesi in 1820. From this, we were able to find highly probabilistic results for the earthquake's location, magnitude, and rupture zone. Though, with some confidence, the results of this event are plausible, we are unable to validate our results by direct comparison to actual seismic information, meaning there is no 'ground truth' for this problem to validate our method. Instead to increase confidence in the 1852 event and future scenarios, I have begun to test our approach on the 2004 Indian Ocean Boxing Day Earthquake and Tsunami, and compare the results to modern seismic data for this event.

To compare the results from these historical earthquakes to a modern event, it is necessary to approach the 2004 Boxing Day event as if there is no access to modern instrumental data. That is to say that any observations used must be similar to the observations used for the 1852 and 1820 event. Therefore, the data needs to be strictly accounts from eyewitnesses, such as: stories from news articles, blogs, or interviews. The only exception being photos and videos, because this evidence is comparable to that of an eyewitness account. Also, any earthquake characteristics likely to apply to the area in question must be drawn from previous earthquakes along the same fault, which means there cannot be any seismic information from after 2004. This allows for an adequate comparison with the historical recreation from our first experiment, and will inform us of the veracity and reliability of our previous work. In this paper, I present my efforts to collect and prepare the data and files needed, so that our model can collect samples for the Boxing Day 2004 event.

Chapter 2. Historical Background

The Indonesian archipelago is the focal point of our study due to the tectonic activity caused by the subduction of the Indo-Australian plate into the Eurasian plate. (Fig.2.1) This causes a range of natural disasters, including: volcanic eruptions, earthquakes, landslides, and tsunamis; all over Indonesia. With the world's 4^{th} largest population at 279 million and a growth rate of 1% yearly, approximately 60% of Indonesia's population lives in low-lying coastal areas which puts them at high risk for tsunamis around the coast and inundation further inland [2][3]. Earthquakes with a magnitude of 7 or higher occur almost yearly and generate tsunamis 88% of the time [4].

For instance, off the west coast of Sumatra, there are historical records of many severe earthquakes. From these records it is observed that between the geographic regions of these events are seismic gaps (Fig.2.1). "A seismic gap is a segment of an active fault known to produce significant earthquakes that has not slipped in an unusually long time, compared with other segments along the same structure.... Any large and long standing gap is, therefore, considered to be the fault segment most likely to suffer future earthquakes" [5].

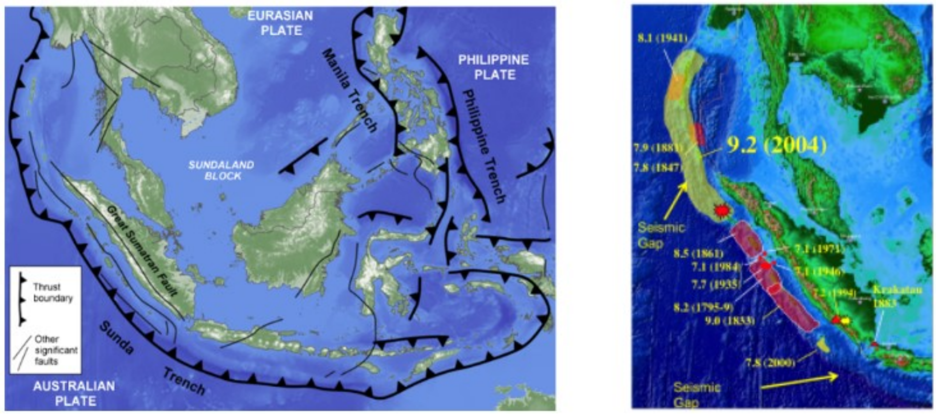

Figure 2.1: (Left) The tectonic plates, faults, and plate boundaries. (Right) The red regions are recorded earthquake occurrences. The yellow regions are the seismic gaps.

To gain more information about these seismic gaps, beyond the modern instrumentation on recent earthquakes, we have studied earthquakes from the 1800s. The Wichmann catalog, which is a collection of eyewitness accounts of Indonesian earthquakes from 1538 to 1877, gives us a starting point with which to use the `tsunamibayes` method to attempt to map out these earthquakes. This allows us to estimate the source of these earthquakes with our model. The two earthquakes we chose to study are the 1852 Banda Neira and 1820 Sulawesi events. We believe we have an accurate model of these earthquakes, but due to the lack of instrumentation data, we cannot verify our results. Therefore, we have chosen to validate these results by retesting the model on the 2004 Boxing Day Earthquake and Tsunami.

The reason we chose to use the 2004 Boxing Day Earthquake and Tsunami to retest our model is that due to its devastating impact, this seismic event was well publicized. On December 26, 2004 at 7:58 AM local time, a magnitude 9.1 undersea megathrust earthquake struck off the northern tip of Sumatra, one of the Sunda Islands of western Indonesia [6]. This seismic event was caused by the Indian Plate slipping and subducting under the Burma Plate [7]. It was the third-largest earthquake since the early 1900s and lasted up to ten minutes. The resulting tsunami killed more than 230,000 people across 14 countries. The worst affected areas were Indonesia, Malaysia, Thailand, Myanmar, Bangladesh, India, Sri

Lanka, and the Maldives, and the effects could be seen as far as East Africa, North America, Antarctica, and Australia [6]. As one of the most destructive and devastating earthquakes in recent recorded history, this event received a lot of coverage which has provided several opportunities for our group to collect data similar to the 1852 and 1820 events.

Chapter 3. Data Collection

There are several steps necessary to take before beginning to test our model on the 2004 Boxing Day scenario. The anecdotal accounts need to be collected and then converted into measurable data. The initial parameters need to be chosen to input in the model. Next, the topography and bathymetry data, or files that measure land elevation and depth in large bodies of water, need to be prepared for each of our observation locations. Finally, it is necessary to identify where refinement in the model is needed [8].

3.1 Observational Data

I started by collecting empirical data which comes in the form of anecdotal accounts. For example, an eyewitness recounted: "All both of us could do was hang on for dear life to the leaves of the tall palm trees as the sea water surged through" [9]. This shows that the wave is approximately as tall as the palm trees native to the area (Fig.3.1). For this observation, I then research the average height of the local palm trees for this particular location and derive a probability distribution of the possible wave height based on this information.

Figure 3.1: This picture shows us the maximum wave height of the tsunami crashing on shore at Ao Nang, Thailand. We would estimate the height of the wave to be 80% of the height of the palm tree in the photo.

As discussed above, I needed data that comes from eyewitness accounts similar to the data we obtained for the 1852 and 1820 scenario data. To stay consistent, the data needs to fulfill two criteria. First, the data needs to refer to an identifiable location. For example, a news article, could be interviewing a local eyewitness, so either the article or witness will need to mention the city name, the beach, a port, a hotel, etc. I would then determine the latitude and longitude by measuring the closest shoreline to that location. One of the observations was at Chittagong Port in Chittagong, Bangladesh by the Karnaphuli River that flows into the Indian Ocean. I could not place the coordinates by the river, because it is too far inland from the ocean. I instead picked a spot on the oceanic shoreline closest to the port. Secondly, the data needs to contain sufficient detail in order approximate one of the following:

(i) Wave height: Height of the first significant wave that strikes the shore after the earthquake.

(ii) Inundation length: How far inland the wave traveled from the shoreline.

Figure 3.2: (Left) Photo evidence of train in Telwatta that was pushed off of the tracks due to the 2004 tsunami. (Center) Depiction of how far the train track is from the shore in comparison to how far the wave inundated based on how far the train moved. (Right) A probability distribution showing the possible values of the inundation length based on this observation.

(iii) Wave arrival time: The local time that the first significant wave struck the shore after the earthquake.

Based on the information provided by each account, I fit a probability distribution to assign each observation the likelihood that it has a certain value. For example, one of the observations we obtained stated: “At Telwatta, the track cuts through thick palm groves and the sea, 200 meters away, is barely visible” [10]. The article says the sea (or shore rather) was 200 meters away. Using Google maps, the area measures that the track was 254 meters from shore and the train was 412 meters from the shore. As a result, I decided to center the distribution at 400 meters (Fig.3.2). A Wikipedia article on the train wreck due to the tsunami states, “[the train] was carried 100 meters” [11]. Thus the distribution was modified to skew right, as it appears that the wave inundated much further than where the train was pushed to.

After collecting 12 observations, I went through a similar process of analyzing each account and determining what the distribution is based on the information given in the account (Fig. 3.3)

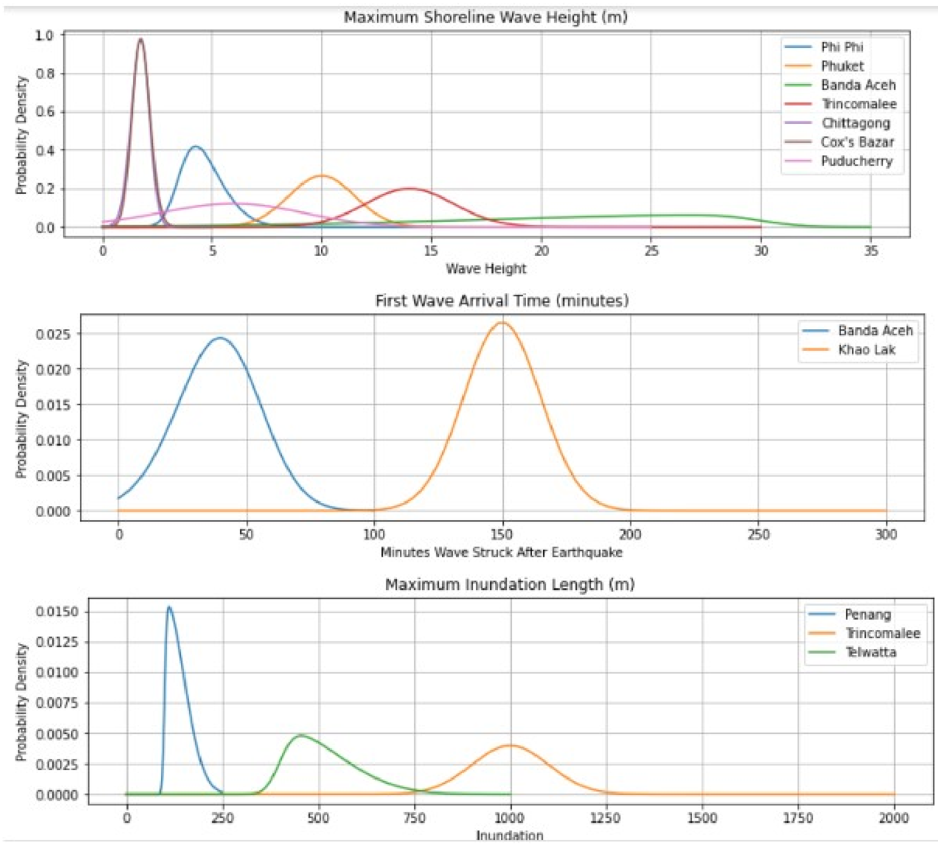

Figure 3.3: Probability distributions for each observation.

3.2 Using Bayesian Inference

Given the empirical nature of the collected data, we used Bayesian Inference to model the inherent uncertainties. As determining the earthquake from tsunami observations is an inverse problem, Bayes' Theorem provides a natural inversion of conditional probabilities and is expressed as:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$$

where $p(y|x)$ is the likelihood, $p(x)$ is the prior distribution, and $p(x|y)$ is the posterior. In relation to our problem, we let x be the earthquake parameters and y be the tsunami observations, then $p(x)$ is the probability of a set of earthquake parameters, $p(y|x)$ is the probability of a certain tsunami observations for a specific earthquake x, and $p(x|y)$ gives the probability of a specific earthquake given the tsunami observations.

To give more context, the likelihood describes measured data and observational uncertainties from the anecdotal accounts using probability distributions. The prior is a collection of distributions of different parameters drawn from information of past earthquakes that occurred on the same fault. This helps define what a "plausible" earthquake in the observed region would be like. The posterior will provide an estimation of the characteristics of the originating earthquake.

3.3 Initial Parameter Selection

The `tsunamibayes` package developed by previous members of our group, simulates thousands of possible earthquakes and measures the resulting tsunami wave in various locations. Instead of sampling random combinations of earthquake parameters, the program identifies samples via Markov Chain Monte Carlo (MCMC) which will take input samples and converge on an estimation of the posterior distribution.

To begin with, `tsunamibayes` uses the Okada model as a basis for earthquake parameters. These parameters are:

- Strike Angle - Specifies orientation of the fault and is measured clockwise in relation to North.
- Dip Angle - Measures the angle of the fault in relation to the horizontal surface.
- Rake Angle - Angle in relation to the strike (i.e. if the angle is parallel to the strike, the rake is $0^{\underline{o}}$, or if the angle is perpendicular to the strike, the rake is $90^{\underline{o}}$).
- Hypocenter/Depth - Measures the depth of the source of the earthquake.
- Epicenter - Point on the earth's surface vertically above the hypocenter; prescribed by latitude and longitude points.
- Slip - Pertaining to the two rock bodies in the fault, the distance one body moves in relation to the other during the rupture.
- Length - Distance the fault measures along the surface that ruptured during the earthquake.
- Width - Distance the fault measures along the dip until maximum depth is reached.

Given sufficient fault and bathymetry data (from previously obtained instrumental records), approximations of the depth, strike, and dip parameters are derived from the latitudinal/-longitudinal position of the epicenter. Using these derived relationships, the five parameters left to fully describe the earthquake are latitude, longitude, rupture length, rupture width, and rupture slip. However these parameters are not independent of one another. Length, width, and slip are all directly related to the earthquake magnitude. The scalar seismic moment M_0 of an earthquake of length L, width W, and average slip S is defined as:

$$M_0 = \mu LWS \tag{3.1}$$

where μ is the (roughly) constant shear modulus of the rock, or Earth's crust, with dimensions of force per unit area. Using this definition, the moment magnitude M_w is defined as

$$M_w = \frac{2}{3}(\log M_0 - 9.05). \tag{3.2}$$

To capture this relationship, the magnitude, M_w, is chosen as one of the sampling parameters. Because smaller earthquakes are exponentially more likely to occur than large magnitude earthquakes, the magnitude is described as an exponential prior distribution. Sampling with magnitude removes slip from the sample parameters, because the slip can then be calculated from the magnitude, length, and width. Since the magnitude grows exponentially with length and width then sampling from magnitude, length, and width concurrently is problematic. Hence we search over magnitude, and logarithmically scaled length and width. Even then, the length and width of earthquakes is highly correlated with the magnitude, so searching over these parameters independently is not a good idea. Instead, the logarithm of length and width is fitted to the magnitude using modern seismic records, and then search over residual errors or deviations from this linear fit. In the end, the sample parameters are reduced to magnitude, Δ log L, and Δ log W with the latter two defined as

$$\log L = aM_w + b + \Delta \log L \tag{3.3}$$

$$\log W = cM_w + d + \Delta \log W \tag{3.4}$$

where a, b, c, and d are the coefficients of the obtained from the modern data, and M_w is the magnitude.

As mentioned above, the values for the dip, depth, strike, and rake are derived from the dataset for each latitude and longitude point in the region of interest using modern instrument data that maps out the major subduction zones. However, this mapping to the geometry of the fault is inherently uncertain with a substantial amount of potential error. Rather than rely solely on the predetermined mapped values, the model searches over off-sets for each of these geometric parameters, introducing: dip-offset, depth-offset, strike-offset, and rake-offset as sample parameters. These parameters are normal probability distributions centered at 0, with a user-defined standard deviation on each.

We found from the 1820 event that dip off-set, strike off-set and rake off-set were neglible in the result, so the earthquake parameters for the 2004 event are modified to be a vector that includes: magnitude, Δ log L, Δ log W, latitude, longitude, and depth-offset. Making probability distributions of these parameters from past earthquakes will form our prior. Some of these parameters that are more general were reused from the 1852 event, but other parameters that are more specific like latitude and longitude need to be updated to fit the fault in the 2004 scenario.

Appropriate distributions must then be chosen for each of these parameters. Latitude and longitude were derived from the subduction interface geometry and magnitude approximately follows a exponential distribution. The parameters Δ log L and Δ log W were given a Gaussian prior distribution with mean zero because they are magnitude-normalized and defined as residuals against a linear best-fit, and depth offset is given a normal distribution centered at zero [1]. The sigmas were found based on the data using residuals of linear best-fit. The sample parameters vector is then mapped to the Okada earthquake parameters as described above, before being fed into GeoClaw to generate and propagate the resultant tsunami.

3.4 GeoClaw and the Adjoint

The forward model in our setting is GeoClaw, an open-source software package that simulates wave propagation by seafloor deformation. The code implements the shallow water equations which is a high resolution finite-volume problem solver on rectangular grids with adaptive mesh refinement. For a given earthquake, GeoClaw will simulate the resultant tsunami and output the wave height, arrival time, etc. that are needed to compare with the observational probabilities described above.

GeoClaw has a dynamically adaptive mesh that makes use of a linearized adjoint equation. The forward model will take Okada earthquake parameters to generate an earthquake with consequent seafloor deformation and these parameters are also used as initial conditions for the shallow water equations that are solved forward in time. The linearized adjoint for the shallow water equations can be solved backwards in time to pre-determine locations where grid refinement is needed [12]. To do this, each observation location is initialized

Table 3.1: Prior Distribution for the 2004 Sumatra Earthquake

Parameter name(s)	Kind	Distribution Parameters
Latitude & Longitude	pre-image of truncated normal via depth	• $\mu = 31.886$ • $\sigma = 7.163$ • (a, b) = (2.5 km, 35 km)
Magnitude	truncated exponential	• $\lambda = 0.5$ • (a, b) = (6.5, 9.5)
Δ log L	normal	• $\mu = 0$ • $\sigma = 0.188$
Δ log W	normal	• $\mu = 0$ • $\sigma = 0.172$
Depth Offset	normal	• $\mu = 0$ • $\sigma = 7.163$

with a Gaussian perturbation in sea surface height (imitating a Dirac-delta function), then the adjoint equation starts at these points and evolves the wave backward from the final time, t_f to initial time t_0. Once adjoint solution is calculated, then a simulation runs where the forward solver propagates the tsunami forward in time and simultaneously the adjoint solution runs backward in time. For any point in time, $t_i \in [t_0, t_f]$, when the forward and backward waves are coincident, is where these spatial locations are refined in the forward solution.

Though calculating adjoint is computationally expensive, it is only computed it once, then it runs simultaneously with each of the simulations done by the forward model. This adjoint computation gives a method to refine the grid dynamically to capture only those parts of the tsunami that will directly impact the observation locations.

3.5 Bathymetry

From the anecdotal accounts, we have obtained 12 observations that cover nine locations (Fig: 3.4). Bathymetry data needs to be collected from all of these locations as well as the entire region that the tsunami evolves over. The region the tsunami evolves over is set to be just west of Puducherry, India at 79° longitude, just north of Chittagong, Bangladesh at 23° latitude, just east of Penang, Malaysia at 101° longitude, and as far south as the equator which reaches the lower edge of the subduction zone where the earthquake originated from.

Figure 3.4: Observation Locations from left to right: Puducherry (India), Telwatta (Sri Lanka), Trincomalee (Sri Lanka), Chittagong (Bangladesh), Cox's Bazar (Bangladesh), Banda Aceh (Indonesia), Khao Lak (Thailand), Phi Phi (Thailand), and Penang (Malaysia).

For all of the locations, topography was found readily available on Earth Explorer [13] with a resolution of about 1 arcsecond or 30 square meters. I pulled bathymetry data from two websites: GEBCO [14] and Indonesian Geospatial Portal [15]. I relied on GEBCO as our default option since it was difficult to find on Google bathymetry files for India, Sri Lanka and Bangladesh with a resolution higher than 5 arcseconds or 150 square meters. However, the Indonesian geospatial site had files with finer resolution at 3 arcseconds or 90 square meters, but only for the Indonesian coastline.

To interpolate both the topography and bathymetry files, I used ArcGIS Pro. The "Mosaic to New Raster" tool in ArcGIS merges two files into a new raster, or matrix of pixels. Using this tool produced a blurry result around the coastline of each of the observation locations (Fig: 3.5: Top Left). Because the shallow water equations in GeoClaw depend sensitively on the bathymetry and coastline, it is necessary to get a smoother transition between resolutions along the coastline. Alternatively, making a land and ocean mask, which is a stencil made in ArcGIS of the land or the ocean elevation, could resolve the grainy resolution along the coastline. Overlaying these two masks showed that because the resolutions differ between the topography and the bathymetry files, the resulting file left a gap at the coastline between both of the stencils (Fig: 3.5: Top Right). To have the raster for land and for ocean fit together, a model was customized to take in files of different resolutions and then output a file with a finer resolution along the coastline (Fig: 3.5: Bottom Left).

The Overlay Model in ArcGIS takes three inputs: the bathymetry file, and the same topography file twice (Fig. 3.6). The first tool used on the Input Topography file is "Reclassify" which uses binary classification that assigns the data with lower values a 1 and the data with higher values a 2. The "Raster to Polygon" tool will take pixels of the same value and aggregate them into a single shape. Since the file only has pixels with a value of 1 or 2, this tool is essentially making a land polygon and an ocean polygon. Next the two "Select" tools will isolate the ocean or land polygon to manipulate them later in the model as masks.

Before moving on to the next tool of the Overlay Model, Input Bathymetry file needs to be resampled. The "Resample" tool will take a pixel from an input file and split it into several pixels of the desired size. This helps modify the bathymetry file to have the same resolution as the topography file. Because the Input Topography file was altered to make two masks, a copy a this same file is used as input into the next tool of the model. The "Extract by Mask" tool crops the data from the input file within shape of the input mask. Thus the land and ocean masks extract land and ocean data from the topography and bathymetry files. The "Mosaic to New Raster" tool takes the cropped ocean and land data and merges them into a .tif file and saves it into an input path. (Fig. 3.6)

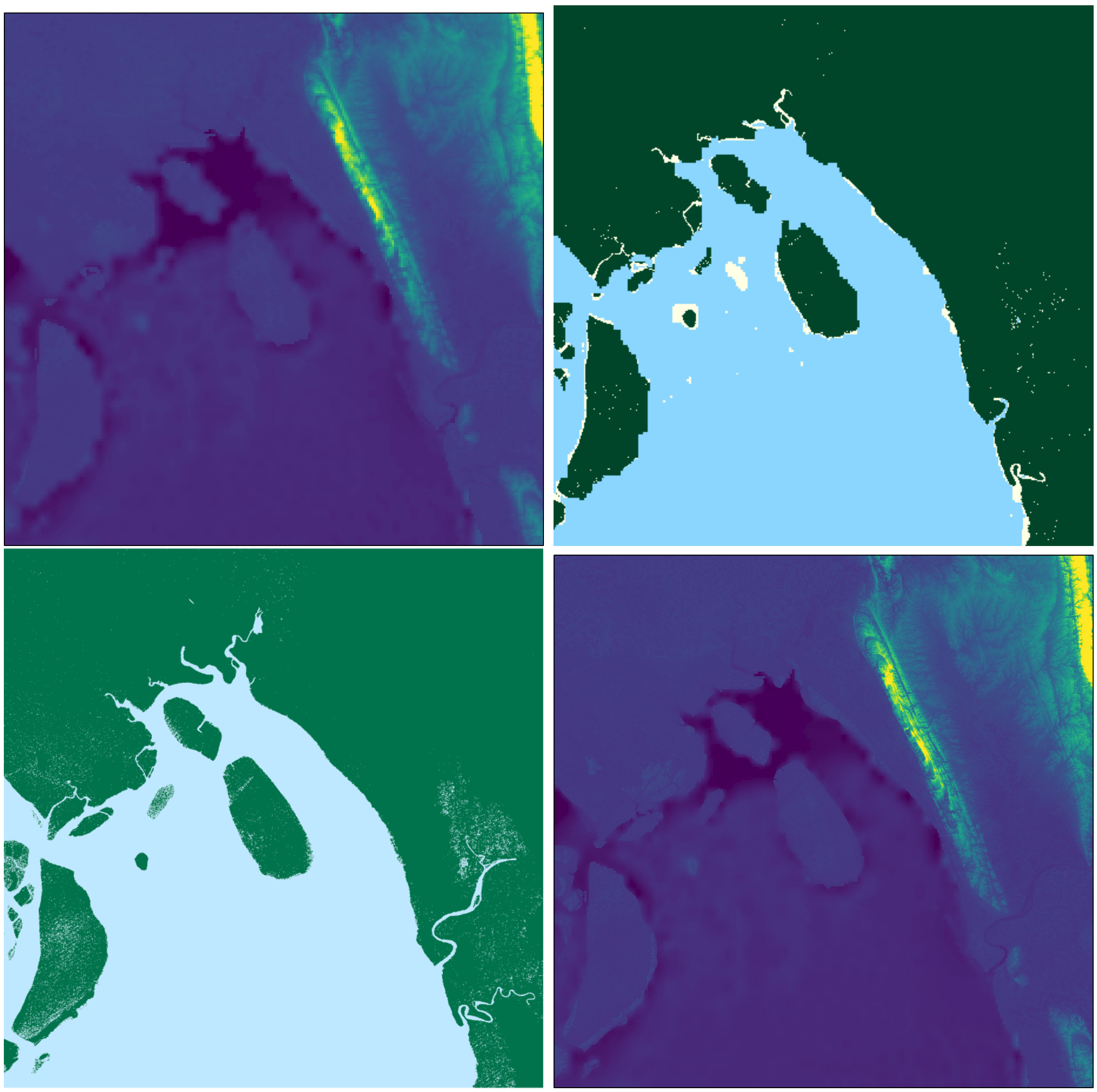

Figure 3.5: (Top Left) Results of “Mosaic to New Raster” tool in ArcGIS Pro. (Top Right) Land and ocean masks with gap at the coastline. (Bottom Left) Land and ocean masks with a more defined coastline. (Bottom Right) Resulting file from a custom ArcGIS model.

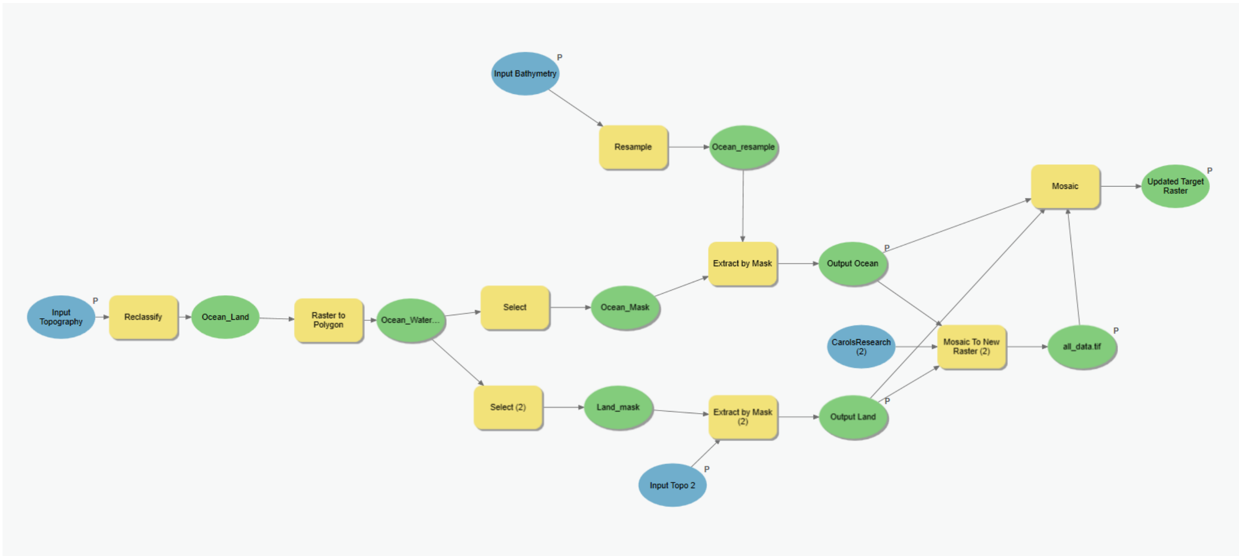

Figure 3.6: This is the Overlay Model. The blue ovals indicate initial inputs. The yellow squares indicate tools. The green ovals indicate outputs and inputs.

Chapter 4. Methodology

Here I will briefly touch on the `tsunamibayes` model which starts with an initial set of sample parameters. Since there are several characteristics that describe an earthquake rupture, it is necessary to narrow the parameter space down using the constraints that the prior places. The prior is then used on the initial sample to expand it into the Okada parameters which are required for the forward model. GeoClaw can only simulate rectangular rupture zones, so for faults that are curved, it is necessary to make small rectangular regions that cover the fault plane. Each of these rectangular patches can be defined by the Okada parameters. This subfault model is created using a fault class constructed in the `tsunamibayes` package. The prior is used again on the expanded parameters to verify that they are realistic. If so, then the prior is nonzero, otherwise, the parameters are rejected.

With a nonzero prior and the input topography and bathymetry files, a forward tsunami is simulated in GeoClaw and then a comparison can be made with the output model to the constructed likelihood. Once the likelihood of the model output is determined, then the MCMC algorithm either accepts or rejects these output parameters. Accepted parameters become the new proposed sample parameters. With the new sample parameters, the model runs again and continues to do so until it converges on a set of parameters that become the posterior. It is here where the accuracy of the results given by the model can be evaluated with actual seismic data of the event.

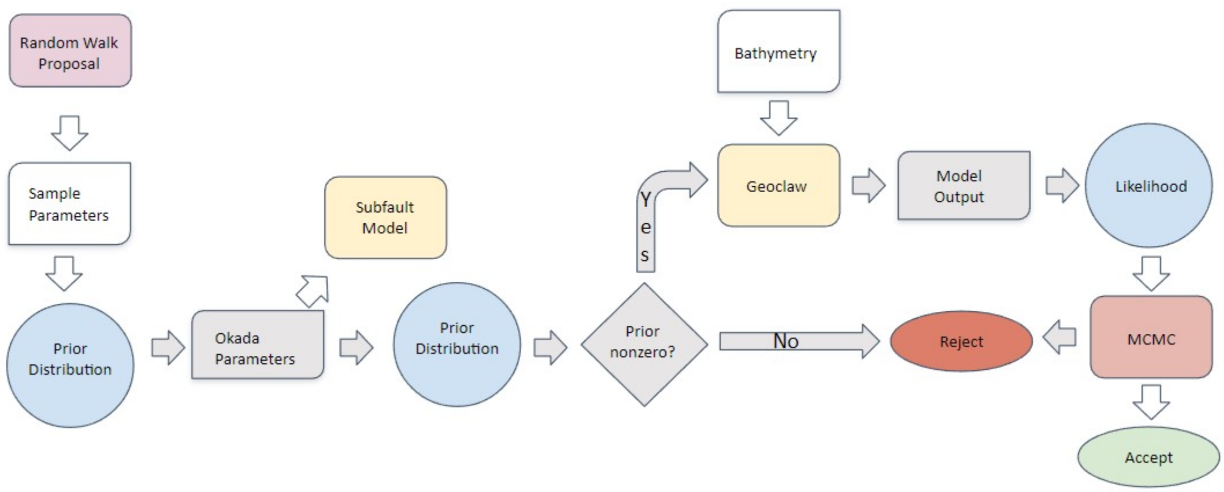

Figure 4.1

Chapter 5. Conclusion

Up to this point, I have created the likelihood from the anecdotal data, selected the initial sample parameters, calculated the adjoint, and interpolated the topography and bathymetry files. Everything I have done besides preparing the topography and bathymetry files has been a close rendition of the 1852 and 1820 scenarios. Previously, it has taken several weeks to collect and prepare the topography and bathymetry files. I expect that after finding the Indonesia Geospatial Portal that data collection will be faster and that the Overlay Model provides more accurate results.

Future work to be done on the 2004 Sumatra scenario is to set the threshold of the MCMC algorithm, run the chains on the supercomputer, and debug. If the results are accurate to the characteristics of the 2004 Sumatra Earthquake and Tsunami, then probabilistic interpretation of anecdotal accounts is a reliable method for modeling earthquakes and tsunamis that occurred before the use of modern seismic instruments. We can then aim to fine tune our model and make it available for seismic hazard mitigation. Ultimately, we hope this research can protect people that live in tsunami prone areas.

Appendix A. The Overlay Model Python Script

Here we have the Overlay Model as a Python script. This code uses the package `arcpy` which is available through Esri licensing in ArcGIS Pro. This code is available on GitHub at: `https://github.com/jpw37/tsunamibayes/blob/sumatra_2004/the_overlay_model.py`
The code is as follows:

```
def Model(Input_Topography=,#input topography filename as string
          Input_Bathymetry=,#input filepath of bathymetry file
          Input_Topo_2=, #input topography filename as string
          Output_Ocean=,#filepath to save your Output_Ocean file
          Output_Land=#filepath to save your Output_Land file
          ):

    # To allow overwriting outputs change overwriteOutput option to True.
    arcpy.env.overwriteOutput = False


    # Check out any necessary licenses.
```

```
arcpy.CheckOutExtension("3D")
arcpy.CheckOutExtension("spatial")

project_path="W:\\tsunamidata-selected\\ArcGIS\\Projects\\project_folder"

# Process: Resample (Resample) (management)
Ocean_resample = project_path+"\\my_gis_project.gdb\\Ocean_resample"
oCS = """GEOGCS['GCS_WGS_1984',
DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563]],
PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]]"""
with arcpy.EnvManager(outputCoordinateSystem=oCS,
                      snapRaster=Input_Topo_2):
    arcpy.management.Resample(in_raster=Input_Bathymetry,
                              out_raster=Ocean_resample,
                              cell_size=
                              "1.66666666666667E-03 1.66666666666667E-03"
                              resampling_type="NEAREST")
    Ocean_resample = arcpy.Raster(Ocean_resample)

# Process: Reclassify (Reclassify) (sa)
Ocean_Land = project_path+"\\my_gis_project.gdb\\Ocean_Land"
Reclassify = Ocean_Land
Ocean_Land = arcpy.sa.Reclassify(in_raster=Input_Topography,
                                 reclass_field="Value",
                                 remap = "-42 2 1;2 539 2",
                                 missing_values="DATA")
Ocean_Land.save(Reclassify)
```

```
# Process: Raster to Polygon (Raster to Polygon) (conversion)
Ocean_Water_Mask =project_path+"\\my_gis_project.gdb\\Ocean_Water_Mask"
#put "MULTIPLE_OUTER_PART", on the same line as create_multipart_features
with arcpy.EnvManager(outputMFlag="Disabled", outputZFlag="Disabled"):
    arcpy.conversion.RasterToPolygon(in_raster=Ocean_Land,
                                     out_polygon_features=Ocean_Water_Mas
                                     simplify="SIMPLIFY",
                                     raster_field="Value",
                                     create_multipart_features=
                                     "MULTIPLE_OUTER_PART", #(see above)
                                     max_vertices_per_feature=None)

# Process: Select (Select) (analysis)
Ocean_Mask = project_path+"\\my_gis_project.gdb\\Ocean_Mask"
arcpy.analysis.Select(in_features=Ocean_Water_Mask,
                      out_feature_class=Ocean_Mask,
                      where_clause="gridcode = 1")

# Process: Extract by Mask (Extract by Mask) (sa)
Extract_by_Mask = Output_Ocean
#cellSize=r"W:\tsunamidata-selected\ArcGIS\Projects\project_folder
#\n07_e098_1arc_v3.tif"
with arcpy.EnvManager(cellSize=,#(See comment above)
                      outputCoordinateSystem=oCS,
                      snapRaster=Input_Topo_2
```

```
                    ):
    Output_Ocean = arcpy.sa.ExtractByMask(in_raster=Ocean_resample,
                                          in_mask_data=Ocean_Mask)
    Output_Ocean.save(Extract_by_Mask)


# Process: Select (2) (Select) (analysis)
Land_mask = project_path+"\\my_gis_project.gdb\\Land_mask"
arcpy.analysis.Select(in_features=Ocean_Water_Mask,
                      out_feature_class=Land_mask,
                      where_clause="gridcode = 2")

# Process: Extract by Mask (2) (Extract by Mask) (sa)
Extract_by_Mask_2_ = Output_Land
Output_Land = arcpy.sa.ExtractByMask(in_raster=Input_Topo_2,
                                     in_mask_data=Land_mask)
Output_Land.save(Extract_by_Mask_2_)


# Process: Mosaic To New Raster (2) (Mosaic To New Raster) (management)
all_data_tif = arcpy.management.MosaicToNewRaster(
    input_rasters=[Output_Ocean, Output_Land],
    output_location=project_path,
    raster_dataset_name_with_extension="all_data.tif",
    coordinate_system_for_the_raster=oCS,
    pixel_type="8_BIT_UNSIGNED",
    cellsize=None,
```

```
        number_of_bands=1,
        mosaic_method="LAST",
        mosaic_colormap_mode="FIRST")[0]
    all_data_tif = arcpy.Raster(all_data_tif)

    # Process: Mosaic (Mosaic) (management)
    Updated_Target_Raster = arcpy.management.Mosaic(
        inputs=[Output_Ocean, Output_Land],
        target=all_data_tif,
        mosaic_type="LAST",
        colormap="FIRST",
        background_value=None,
        nodata_value=None,
        onebit_to_eightbit="NONE",
        mosaicking_tolerance=0,
        MatchingMethod="NONE")[0]
    Updated_Target_Raster = arcpy.Raster(Updated_Target_Raster)

    return all_data_tif, Updated_Target_Raster

if __name__ == '__main__':
    # Global Environment settings
    #scratchWorkspace=r"W:\tsunamidata-selected\ArcGIS\Projects
    #\project_folder\my_gis_project.gdb",
    #workspace=r"W:\tsunamidata-selected\ArcGIS\Projects\project_folder
    #\my_gis_project.gdb"
    with arcpy.EnvManager(scratchWorkspace=, #(see comment above)
```

```
                              workspace= #(see comment above)
                              ):
Model(*argv[1:])
```

Bibliography

[1] Hayden J. Ringer. A Method for Reconstructing Historical Destructive Earthquakes Using Bayesian Inference. Master's thesis, Brigham Young University, 2020.

[2] Jared Whitehead, editor. *Bayesian Inference of Seismic Events from Anecdotal Historical Accounts in Indonesia*, 2020.

[3] Indonesia population(live). Worldometer. https://www.worldometers.info/world-population/indonesia-population/.

[4] List of Earthquakes in Indonesia. Wikimedia Foundation. www.en.wikipedia.org/wiki/List_of_earthquakes_in_Indonesia.

[5] Seismic Gap. Wikimedia Foundation. https://en.wikipedia.org/wiki/Seismic_gap.

[6] Boxing Day Tsunami: How the Disaster Unfolded 10 Years Ago. ABC News, December 2014. https://www.abc.net.au/news/2014-12-24/boxing-day-tsunami-how-the-disaster-unfolded/5977568.

[7] 2004 Indian Ocean Earthquake and Tsunami. Wikimedia Foundation. www.en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake_and_tsunami.

[8] *Guiding AMR with Adjoint Flagging*. https://www.clawpack.org/adjoint.html.

[9] Brian O'Connell. Eyewitness: Irish Tsunami Survivors Recall Day the Wave Hit. *The Irish Times*, December 2014. www.irishtimes.com/news/world/asia-pacific/eyewitness-irish-tsunami-survivors-recall-day-the-wave-hit-1.2048633.

[10] Jonathan Steele. One Train, More than 1,700 Dead. *The Guardian*, December 2004. https://www.theguardian.com/world/2004/dec/29/tsunami2004.srilanka.

[11] 2004 Sri Lanka Tsunami Train Wreck. Wikimedia Foundation. https://en.wikipedia.org/wiki/2004_Sri_Lanka_tsunami_train_wreck.

[12] B. Davis and R. LeVeque. Adjoint Methods for Guiding Adaptive Mesh Refinement in Tsunami Modeling. *Pure & Applied Geophysics*, 173(12), 2016.

[13] Earth Explorer. https://earthexplorer.usgs.gov/.

[14] GEBCO. https://download.gebco.net/.

[15] Indonesia Geospatial Portal. https://tanahair.indonesia.go.id/demnas/#/.