2020-08-04

A Method for Reconstructing Historical Destructive Earthquakes Using Bayesian Inference

Hayden J. Ringer
Brigham Young University

A Method for Reconstructing Historical Destructive Earthquakes Using Bayesian Inference

Hayden J. Ringer

A thesis 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
Emily J. Evans
Blake H. Barker
Ron A. Harris

Department of Mathematics

Brigham Young University

ABSTRACT

A Method for Reconstructing Historical Destructive Earthquakes Using Bayesian Inference

Hayden J. Ringer
Department of Mathematics, BYU
Master of Science

Seismic hazard analysis is concerned with estimating risk to human populations due to earthquakes and the other natural disasters that they cause. In many parts of the world, earthquake-generated tsunamis are especially dangerous. Assessing the risk for seismic disasters relies on historical data that indicate which fault zones are capable of supporting significant earthquakes. Due to the nature of geologic time scales, the era of seismological data collection with modern instruments has captured only a part of the Earth's seismic hot zones. However, non-instrumental records, such as anecdotal accounts in newspapers, personal journals, or oral tradition, provide limited information on earthquakes that occurred before the modern era.

Here, we introduce a method for reconstructing the source earthquakes of historical tsunamis based on anecdotal accounts. We frame the reconstruction task as a Bayesian inference problem by making a probabilistic interpretation of the anecdotal records. Utilizing robust models for simulating earthquakes and tsunamis provided by the software package *GeoClaw*, we implement a Metropolis-Hastings sampler for the posterior distribution on source earthquake parameters. In this work, we present our analysis of the 1852 Banda Arc earthquake and tsunami as a case study for the method.

Our method is implemented as a Python package, which we call `tsunamibayes`. It is available, open-source, on GitHub: `https://github.com/jwp37/tsunamibayes`.

Keywords: Bayesian statistics, Markov chain Monte Carlo, inverse problems, earthquakes, tsunamis, seismic hazard analysis

Acknowledgements

I want to thank my advisor, Dr. Jared Whitehead, who has mentored me for the better part of four years, going back to my time as an undergraduate in the ACME program. He has patiently put up with my idiosyncrasies and foibles, and has helped me navigate the ups-and-downs of graduate student life. I am deeply thankful for his guidance and support.

I would like to thank my many collaborators on this project. Dr. Ron Harris's passion for protecting human life is inspiring. Dr. Justin Krometis has been a great resource for me in all of my questions. Dr. Nathan Glatt-Holtz has offered irreplaceable expertise on difficult mathematical issues. Thanks to all of the following collaborators: Claire Ashcraft, Ashley Avery, Bryce Berrett, Camille Carter, Garret Carver, Marja Crawford, Bethany Fry, Joshua Fullwood, Spencer Giddens, McKay Harward, Ryan Hilton, Reagan Howell, Cody Kesler, Hunter Klein, Josh Lapicola, Kameron Lightheart, Martha Morrise, Carly Ringer, Adam Robertson, Timothy Sitze, Jake Voorhees, Mingyan Zhao, and Yajing Zhao. Chapters 1, 4, and 5 of this thesis include material which comes from a in-progress manuscript, authored by a number of individuals listed above.

I want to thank Dr. Tyler Jarvis and Dr. Emily Evans, who have taught me since I was an ACME student. They have been gracious towards me far beyond what I have deserved. I want to thank Dr. Jeff Humpherys for taking a chance on me as somewhat-wayward undergraduate. Thank you to Dr. Blake Barker, who sits on my thesis committee alongside Drs. Whitehead, Evans, and Harris. I also want to thank other BYU faculty and staff for their kindness over the last decade, especially Fred Lewis, Lonette Stoddard, Dr. Leonard Bakker, Dr. Benjamin Webb, and Dr. Vianey Villamizar.

Thank you to my parents, Rick and Andrea. They have supported me in every struggle and cheered for me in every success. I want to thank my grandparents, John, Carol, Robert, and Joyce, who I love dearly. Thank you to my sisters Jenna and Carly, as well as my brother-in-law Mike, for their love and friendship. Thank you to Claire, my niece (and best buddy!), who brings so much joy to my life.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Indonesia is one of the most tectonically active and densely populated places on Earth. It is surrounded by subduction zones that accommodate the convergence of three of Earth's largest plates. Some of the largest earthquakes, tsunamis and volcanic eruptions known in world history happened in Indonesia [1, 2]. Since these events, population and urbanization has increased exponentially in areas formerly destroyed by past geophysical hazards. Recurrence of some of these large events during the past two decades have claimed a quarter million lives [2].

Most casualties from natural disasters in Indonesia are caused by tsunamis, which, over the past 400 years, occur on average every 3 years [3]. Many potential tsunami source areas, such as the eastern Sunda [4] and Banda [5] subduction zones have no recorded mega-thrust earthquakes [6]. However, some historical accounts of earthquakes and tsunamis in Indonesia provide enough detail about wave arrival times and wave heights from multiple locations to test if mega-thrust events have happened in apparently quiet regions, and assess the likely consequences of these events reoccurring.

Reliance on modern instrumental records of earthquake events to determine seismic risk severely biases hazard assessments, as the relevant temporal scales are hundreds or thousands of years on a given fault zone. To improve risk estimates, it is imperative to draw from historical records of damaging earthquakes, which reach beyond the fifty to seventy year horizon provided by modern instrumental records. To this end, there has been substantial effort invested in the quantification of the characteristics of pre-instrumental earthquakes and tsunamis; see e.g. [4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 2, 18, 19, 20]. As noted in these references, the historical and prehistorical data sources are sparse in details and laced with high levels of uncertainty. To improve the usage of these imprecise data sources, we develop a systematic framework that provides estimates in concert with associated uncertainties on earthquake parameters. We then apply a Bayesian statistical inversion approach already

leveraged in a variety of disciplines in the physical, social and engineering sciences, (see [21, 22, 23] as well as [24, 25, 26, 27]), to reconstruct large seismic events from historical accounts of the resulting tsunamis. Our focus here is on an initial case study concerning the reconstruction of the 1852 Banda Arc earthquake and tsunami in Indonesia detailed in the recently translated Wichmann catalog of earthquakes [2, 28] and from contemporary newspaper accounts [29]. To formulate a Bayesian posterior distribution that estimates various parameters of the 1852 seismic and tsunami event, we develop a 'forward model' that associates seismic parameters specifying earthquake location and magnitude with shoreline observations that include wave arrival times, maximal wave height at the shoreline, and coastal inundation. The forward model utilizes the Geoclaw software package [30, 31, 32, 33] to numerically integrate the shallow water equations for predicting the evolution of the tsunami initiated by seafloor deformation due to the earthquake itself.

Chapter 2. Background

2.1 Bayesian Inference

Bayesian probability is a natural setting for this inference problem. Under the Bayesian, or *epistemic*, interpretation of probability, we can model uncertainty regarding the location, intensity, and geometry of the source earthquake as random variables. In contrast to a frequentist interpretation of probability, these random variables need not represent a *truly random* process; correspondingly we do not interpret subjective probabilities as the *chance* of an event occurring, but as the *credence* of an event occurring.

The tools of Bayesian inference provide a robust framework for quantifying epistemic uncertainty. Here we recite the standard definitions of measure-theoretic probability, as well as Bayes' Theorem for random variables. It is assumed that the reader has at least some familiarity with these topics. While we follow the presentation in [22], the reader may see [34] for an extended treatment of the topic.

Given a set Ω, a σ-algebra $\mathcal{F} \subseteq 2^\Omega$, and a countably-additive function μ on $\mathcal{F}$, the triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*, the elements of $\mathcal{F}$ are called *measurable sets*, and μ is called a *measure*. A standard example of a measure space is $(\mathbb{R}^n, \mathcal{M}, \mu)$, where μ is Lebesgue measure and $\mathcal{M}$ are the Lebesgue-measurable sets in $\mathbb{R}^n$. Within the context of probability theory, it is usually sufficient to consider the smaller σ-algebra of Borel sets, denoted $\mathcal{B}(\mathbb{R}^n)$.

A measure space with $\mu(\Omega) = 1$ is also called a *probability space*. The measure on a probability space is called a *probability measure*, and is usually denoted with the letter P. In a probability space, elements of $\mathcal{F}$ are called *events*, and the function P assigns to each event E a value $P(E) \in [0, 1]$, which is the *probability* of E occurring. When Ω is a countable (possibly finite) set, the space is called a *discrete probability space*.

A function $f : \Omega_1 \to \Omega_2$ between measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ is called *measurable* if the pre-image of every measurable subset of Ω_2 is a measurable subset of Ω_1. A measurable function $X : \Omega \to \mathbb{R}^n$ from a probability space $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu)$ is called a *random variable*. A random variable with countable (perhaps finite) range is called a *discrete random variable*. Otherwise, it is called a *continuous random variable*. A collection $X_1, \ldots, X_N$ of random variables defined on the same probability space form the product, or *joint* random variable $X_1 \times \cdots \times X_N$.

Random variables are the workhorses of probability theory. Each random variable induces a probability measure μ_X on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$:

$$\mu_X(A) = P(X^{-1}(A)), \quad A \in \mathcal{B}(\mathbb{R}^n).$$

This probability measure is called the *distribution* of X. The distribution has a very reasonable interpretation: $\mu_X(A)$ is the probability that $X(\omega) \in A$. This information is also contained within the *cumulative distribution function* (CDF) of X:

$$F_X(\boldsymbol{x}) = \mu_X((-\infty, x_1] \times \cdots \times (-\infty, x_n]), \qquad \boldsymbol{x} = [x_1, \ldots, x_n]^\mathsf{T}.$$

$F_X(\boldsymbol{x})$ is the probability that $X(\omega)$ is entry-wise less-than or equal-to $\boldsymbol{x}$. When the distribution of a random variable is absolutely continuous with respect to Lebesgue measure, there exists a nonnegative function f_X, called the *probability density function* (pdf) of X, such that

$$\mu_X(A) = \int_A f_X(x)dx.$$

The pdf of X, when it exists, is extremely useful when working with random variables.

For continuous random variables $X : \Omega \to \mathbb{R}^n$, we will abuse notation and write $p(x)$ to mean the evaluation of the probability density function (pdf) of X at $x \in \mathbb{R}^n$ (when it exists). Given another continuous random variable $Y : \Omega \to \mathbb{R}^m$, we will write $p(x, y)$ to mean the evaluation of the joint pdf of X and Y at $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$. We will also write $p(x|y)$ to mean the evaluation of the conditional pdf for $X|Y = y$. We will write $x_1, x_2, \ldots$ to indicate *realizations* of the random variable X.

For discrete random variables, the analog of the pdf is the *probability mass function*, or pmf. The pmf p_X of X gives the probability of single points in the range of X: $p_X(x) = P(X(\omega) = x)$. We will follow the abuse of notation above and use p to denote the pmf/joint pmf/conditional pmf when context allows.

The heart of Bayesian inference is, of course, *Bayes' Theorem*, which is simply a relation between the conditional densities/mass functions $p(x|y)$ and $p(y|x)$:

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

where the integral is defined appropriately, i.e. it reduces to a summation in the discrete case.

Bayes' Theorem allows for the "inversion" of conditional probabilities. This is particularly useful when X represents some unknown quantity of interest and Y represents data that is generated in some known way, conditioned on a given value of X. In this context, Bayes' Theorem updates the knowledge about X in light of known data $Y = y$. Here, $p(x)$ is

called (the density of) the *prior distribution*, as it represents the knowledge about X *before* considering the data. The conditional density $p(y|x)$ is called the *likelihood*, and represents the known connection between the data Y and the unknown quantity X. Finally, $p(x|y)$ is called (the density of) the *posterior distribution*, which represents the knowledge about X after incorporating the data. Said concisely: Bayesian inference is the computation (or estimation) of posterior distributions given observations of the data, assumptions about the likelihood, and any prior knowledge of the unknown variable X.

As an example, consider the age A and height H of a child. For children of a given age, height can be said to follow some fixed distribution, say a normal distribution with age-dependent mean μ_a and fixed variance σ^2. Consider a child whose exact age is unknown, but is known to be between 5 and 10 years old. This corresponds to a uniform prior distribution for A on $[5, 10]$. Suppose the height of the child is measured to be h. The likelihood $p(h|a)$ representing the probability the child is a certain height h, given that they are a years old, would then be:

$$p(h|a) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(h-\mu_a)^2}{2\sigma^2}\right).$$

Noting that $p(h) = \int_5^{10} p(h|a)p(a)da$ is a normalizing constant, the posterior pdf is:

$$p(a|h) \propto p(h|a)p(a) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(h-\mu_a)^2}{2\sigma^2}\right) \cdot \frac{1}{5}.$$

If μ_a is linear in a, then the posterior distribution is also a normal distribution. (In particular, if $\mu_a = ma + b$, then the posterior distribution has mean $\frac{h-b}{m}$ and variance $\frac{\sigma^2}{m^2}$). If μ_a is not linear in a, then the posterior distribution may not belong to a known family of distributions, but it may still be estimated by various methods, including the sampling techniques discussed below.

2.2 Sampling Methods

A number of issues appear when naively applying Bayes' Theorem to more complicated problems. If the unknown quantity X is high-dimensional (taking values in $\mathbb{R}^n$), then the normalizing constant in the denominator requires a high-dimensional integral that may be impossible to compute exactly and effectively impossible to estimate numerically. Furthermore, the computation of expectations or higher-order moments of the posterior (something of interest in many applications) will also require computing intractable high-dimensional integrals. In short, although Bayes' Theorem may specify the theoretical existence of the posterior distribution, it may not be feasible to apply Bayes' Theorem directly to obtain useful and meaningful interpretations of the posterior itself.

Despite these challenges, suppose that it is possible to generate samples from the posterior distribution. Given samples $x_1, x_2, \ldots, x_N$, the expectation $E[X]$ of the posterior distribution can be estimated as $\frac{1}{N}\sum x_i$. Similarly, higher-order statistics of X, and statistics of functions of X, could be estimated using standard estimators from inferential statistics. With a sufficiently large number of samples, desired information about X could be estimated to high accuracy.

A robust family of sampling methods fall under the umbrella of *Markov chain Monte Carlo* (MCMC) methods. These methods work by simulating a Markov chain which has the desired sampling distribution, i.e. the posterior, as its stationary, or steady-state, distribution. Thus, a large number of samples from this Markov chain will adequately represent the desired posterior distribution and its relevant statistics. We introduce here one such method, the *Metropolis-Hastings algorithm*, originally developed by Nicholas Metropolis [35], and extended by W.K. Hastings [36].

A Markov chain is a sequence of random variables $X_1, X_2, \ldots$ satisfying the *Markov property*:

$$P(X_{n+1} \in A | X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} \in A | X_n = x_n) \quad \forall A \in \mathcal{B}(\mathbb{R}^n)$$

In other words, the future states of the chain depend on the past states only though the present state. This is often called being *memoryless*. A Markov chain where $P(X_{n+1} \in A|X_n = x_n) = P(X_n \in A|X_{n-1} = x_{n-1})$ for all n is called *time-homogeneous*: the transition probabilities do not vary with the time index n.

It is well-known that time-homogeneous finite-state Markov chains have stationary distributions given by a particular eigenvector $\boldsymbol{v}$ of their transition matrix P: $P\boldsymbol{v} = \boldsymbol{v}$. Identifying the stationary distribution of a continuous-state Markov chain is more challenging. However, the derivation of the Metropolis-Hastings algorithm only requires a simplified condition, called *detailed balance*. Although we omit the full derivation, the interested reader may consult pp. 92-96 of [22].

The key is to consider the *probability transition kernel* for a continuous-state Markov chain (the analogue of the transition matrix) to be of the form:

$$P(X_{n+1} \in A|X_n = x) = \int_A K(x, y)dy + r(x)\chi_A(x).$$

Here $K(x, y)$ is the probability density of a transition from the state x to the state y, and $r(x)$ is the probability of remaining at the state x. The *balance equation* states that $\pi(x)$ is the probability density of the stationary distribution of the above Markov chain if:

$$\int_{\mathbb{R}^n} \pi(x)K(x, y)dy = \int_{\mathbb{R}^n} \pi(y)K(y, x)dx.$$

The *detailed* balance equation, which clearly implies the above, is:

$$\pi(x)K(x, y) = \pi(y)K(y, x).$$

This can be read as saying: "the steady-state frequency of x, times the frequency of transitions from x to y, is equal to the steady-state frequency of y, times the frequency of transitions from y to x."

In order to construct a Markov chain with a specified stationary distribution, it is enough to construct a probability transition kernel so that detailed balance is satisfied. As a matter of fact, it is possible to *modify* a given Markov chain to have a desired stationary distribution. Let $\pi(x)$ be the desired stationary distribution pdf, and suppose $q(x, y)$ is the density function for a Markov probability transition kernel. If we can find a "correction function" $\alpha(x, y)$ such that:

$$\pi(y)\alpha(y, x)q(y, x) = \pi(x)\alpha(x, y)q(x, y),$$

then detailed balance holds for the Markov chain with transition kernel:

$$P(X_{n+1} \in A | X_n = x) = \int_A \alpha(x, y)q(x, y)dy + (1 - \alpha(x, y))\chi_A(x).$$

The Metropolis-Hastings correction function is to set:

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right).$$

In this context, $q(x, y)$ is called the *proposal kernel*, and $\alpha(x, y)$ is called the *Metropolis-Hastings acceptance function*. The algorithm is to use q to "propose" a new state in the chain, and then to accept that new state with probability α, otherwise remaining at the state x. If $q(x, \cdot)$ is easy to sample from, then the only work is to evaluate the desired stationary distribution pdf. The full algorithm is as follows:

Algorithm 1: Metropolis-Hastings

```
Initialize: x_0
for i = 0, ... do
    Propose u ~ q(x_i, ·)
    Set α = min(1, π(u)q(u,x_i) / π(x_i)q(x_i,u))
    Draw p ~ U(0, 1)
    if p < α then
        Set x_{i+1} = u
    else
        Set x_{i+1} = x_i
```

To draw from the posterior distribution of Bayes' Theorem, one only needs to be able to evaluate the un-normalized density $p(y|x)p(x)$, since the unknown normalization constant cancels out in the definition of α.

While the Metropolis-Hastings algorithm will function with many proposal kernels, a good choice of proposal kernel can significantly speed up convergence. A common choice is the random walk proposal kernel, which is a multivariate Gaussian centered at the current state x_i, with a fixed covariance matrix.

The Metropolis-Hastings algorithm is not, strictly-speaking, parallelizable. However, multiple chains can be simulated simultaneously, drawing samples at a faster rate. This, however, does not necessarily speed up the convergence to the stationary distribution, as the Markov chains may not be initialized within the "bulk" of the distribution. The period of time before the Markov chain begins to display convergence is called burn-in. There is, however, a trick to boot-strapping multiple chains to speed up burn-in. Periodically, the chains can be paused and re-initialized by resampling from the current states of all chains, proportional to the value of the sampling distribution pdf at each state [23]. This means that chains in poorly-performing regions of the parameter space can be "jumped" to regions that are more important in the stationary distribution. Initializing multiple chains in different parts of the parameter space can accelerate burn-in, and therefore convergence.

2.3 Inverse Problems

Inverse problems are ubiquitous in mathematics, pure and applied. As the name says, an inverse problem is the inverse of a direct problem. For example, solving a polynomial equation is the inverse of the direct problem of evaluating a polynomial function. Inverse problems are often more challenging than the corresponding direct problem.

An important kind of inverse problem arises when the direct problem is physical in nature. Often, the direct problem is given as an initial/boundary value problem for a partial differential equation. For example, solving the heat equation is the direct problem for mod-

eling the flow of thermal energy (or, more broadly, a diffusive phenomenon). Given an initial condition and boundary data, a solution to the heat equation specifies the temperature at future times. The inverse problem for the heat equation IBVP is: given some future temperature state, determine the initial temperature state. This is a very challenging problem, as the very nature of the heat equation means that information about the initial condition is lost as time moves forward.

Statistical methods for inverse problems treat the unknown parameters/initial condition as a random variable. Bayesian statistical methods for inverse problems attempt to compute a posterior distribution for the parameters/initial condition. These methods often require the ability to solve the direct, or forward problem for a variety of parameters/initial conditions. This means they often rely on robust numerical methods for PDEs. In essence, the statistical method computes a sequence of solutions to the forward problem that, in a probabilistic sense, converge to the parameter values that likely produced the observed data. However, since the direct problem is "merely" a model for the real-world data-generating process, care must be taken to insure that the *forward model* is robust. When implemented properly, the Bayesian approach is a powerful way to provide a solution to an inverse problem that captures the relevant uncertainties.

Chapter 3. Overview of Method

There are three key components to our method, as a Bayesian inverse approach:

- The *likelihood*
- The *forward model*
- The *prior distribution*

The likelihood is an interpretation of the anecdotal accounts (such as those found in the Wichmann catalog) as a family of probability distributions. We source three kinds of

observations from the accounts: on-shore arrival time of the tsunami wave, maximal on-shore wave height, and coastal inundation distance. Each recorded observation is cast as a probability distribution, and those distributions together form the likelihood function. Chapter 4 is focused on the likelihood.

The forward model is the map from source earthquake parameters (location, size, geometry) to observations. The forward model includes the formation of the tsunami from the source earthquake (via seafloor deformation) and the propagation of the tsunami waves to the observation locations. This is accomplished by using the *GeoClaw* software package. GeoClaw both computes the seafloor deformation and simulates the tsunami via a finite-volume solver for the shallow water equations [33]. Chapter 5 contains our discussion of the forward model.

The prior distribution encodes knowledge about earthquakes in general, as well as information about the subduction zone in question. There are physical constraints on earthquake location, size, and geometry, and these are all enforced via the prior distribution. Here we also define our approach to parameterizing the space of possible earthquakes, giving attention to reducing the dimensionality of the parameter space while maintaining a physically-reasonable representation. We make special use of the USGS Slab2 3D model of the Banda Arc subduction zone [37]. Chapter 6 discusses the prior and parameterization in detail.

Having specified the Bayesian inverse problem, we use random walk Metropolis-Hastings to sample from the posterior distribution. Given that each sample requires a computationally-expensive call to the forward model, we make use of multiple Markov chains combined using periodic resampling. All of this is implemented in Python as a package that we call `tsunamibayes`. An overview of the software implementation is provided in chapter 7.

Throughout the text, we hold close to the 1852 Banda Arc case study. While we anticipate that a number of alterations will be made as we study other events, the 1852 Banda Arc event has been our focus, and is useful for illustration purposes. Chapter 8 contains a summary of our results when applying our method to the 1852 Banda Arc event.

Chapter 4. Construction of the Likelihood

4.1 Overview of historical account and potential observations

Observations are selected from the historical accounts in the Wichmann catalog [38, 28] based on two key criteria. First, the account has to provide an identifiable location (latitude-longitude) that can be incorporated into the modeling. In other words, the details provided in the historical account must be sufficiently accurate to yield a precise location via modern-day maps and information. Second, the account has to be sufficiently detailed that some level of confidence can be placed on the observable in question. Note that drawing from a catalog of this kind introduces unavoidable ambiguities that do not apply to modern instrumental data. For example, we specify the wave height based on passages of the form "[t]he water rose to the roofs of the storehouses and homes," as described in more detail below.

Thirteen different observations for the 1852 Banda Arc tsunami meet these criteria spread across nine locations, which are shown in Figure 4.1. These include three types of observations:

(i) *Arrival time.* The arrival of the first significant wave after the shaking stopped. We assume that the arrival time refers to the first wave, not the maximal one.

(ii) *Maximum wave height.* This is the most frequent observable, and is identified at every location.

(iii) *Inundation length.* This refers to the distance inland that the wave traveled onshore, and is actually interpreted for our purposes as a deterministic function of the wave height. This essentially places a double amount of weight on those locations that have observations of both wave height and inundation.

Based on the text of each account, a probability distribution is developed describing the probability that each observation took a given value. These distributions, which are assumed

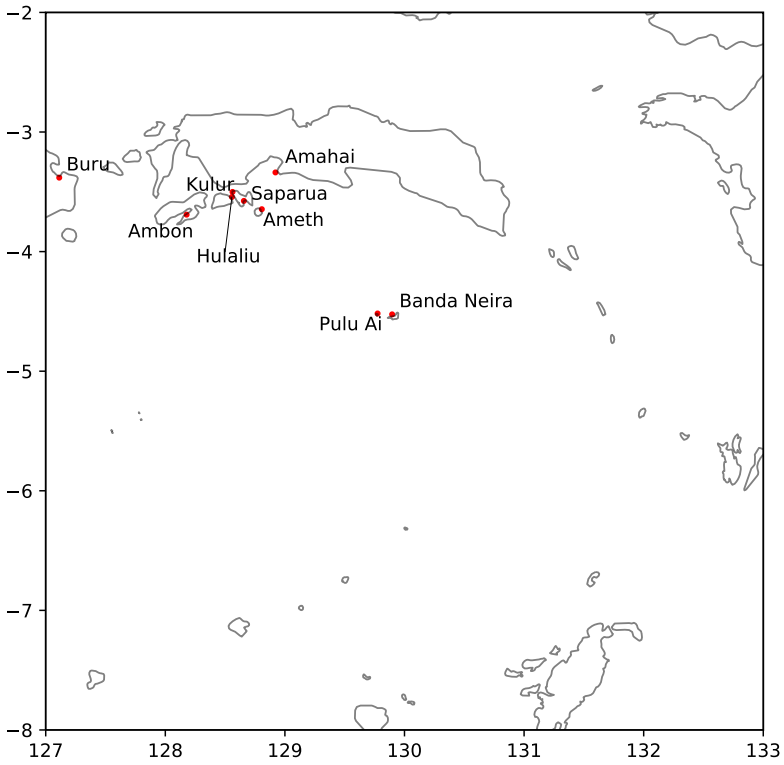

Figure 4.1: The nine observation locations from the Wichmann catalog for the 1852 Banda Arc earthquake and tsunami.

to be independent, are shown in 4.2. Rather than explain the reasoning behind all thirteen of these likelihood distributions for each of the nine locations, we only provide a detailed discussion of the likelihood for a single location: Banda Neira.

4.2 Banda Neira: a sample likelihood distribution

From page 242 in the Wichmann catalog: "Barely had the ground been calm for a quarter of an hour when the flood wave crashed in...The water rose to the roofs of the storehouses and homes...[the wave] reached the base of the hill on which Fort Belgica is built on Banda Neira". Ideally we would expect the wave height observation to be near the boat dock on Banda Neira which is just east of Fort Nassau. For the available bathymetry data we seek

a location near shore that will maintain a sizable wave for a reasonably initiated tsunami. With this in mind, we select $-4.524°$ latitude and $129.8965°$ longitude.

Using 15 minutes as the anticipated arrival time of the wave at Banda Neira is too simplistic for these circumstances. In particular it is noted in other locations that the shaking lasted for at least 5 minutes, but the modified Okada model used in Geoclaw here assumes an instantaneous rupture. Hence we build into the likelihood, a skew toward longer times with a mean of 15 minutes. This is done with a skew-normal distribution with a mean of 15 minutes, standard deviation of 5 minutes, and skew parameter 2.

Assuming a standard construction for the time period for the homes (and storehouses) we can assume the water rose at least 4 meters above standard flood levels as most buildings of the time were built on stilts and had steep vaulted roofs. Based on the regular storm activity in the region we can expect that with high tide, and normal season storm surge, the standard flood level is also approximately 2 meters in this region. This leads us to select a normally distributed likelihood for wave height with a mean of $6.5m$ and standard deviation of $1.5m$, allowing for reasonable likelihood for wave heights in the range from $3m$ to $9m$.

To quantify the wave reaching the base of the hill, we measured the distance from 20 randomly selected points along the beach to the edge of said hill in ARCGIS. The mean of these measurements was 185 meters, with a standard deviation of roughly 65 meters. Thus we choose a normal distribution with those parameters.

4.3 Overview of all likelihoods

The likelihood distributions for the other 8 locations are constructed in a very similar manner to that described above for Banda Neira. The total likelihood of a given event is then computed as the product of these individual observational likelihoods (we rely heavily on the assumption that each observable is independent of the others). The assumption of independence of the different observations is certainly questionable, but there is also no reason to suppose that a more complicated construction of the total likelihood is preferable, i.e. we

have chosen to take the most simplified approach without making additional unjustifiable assumptions about the structure of the likelihood.

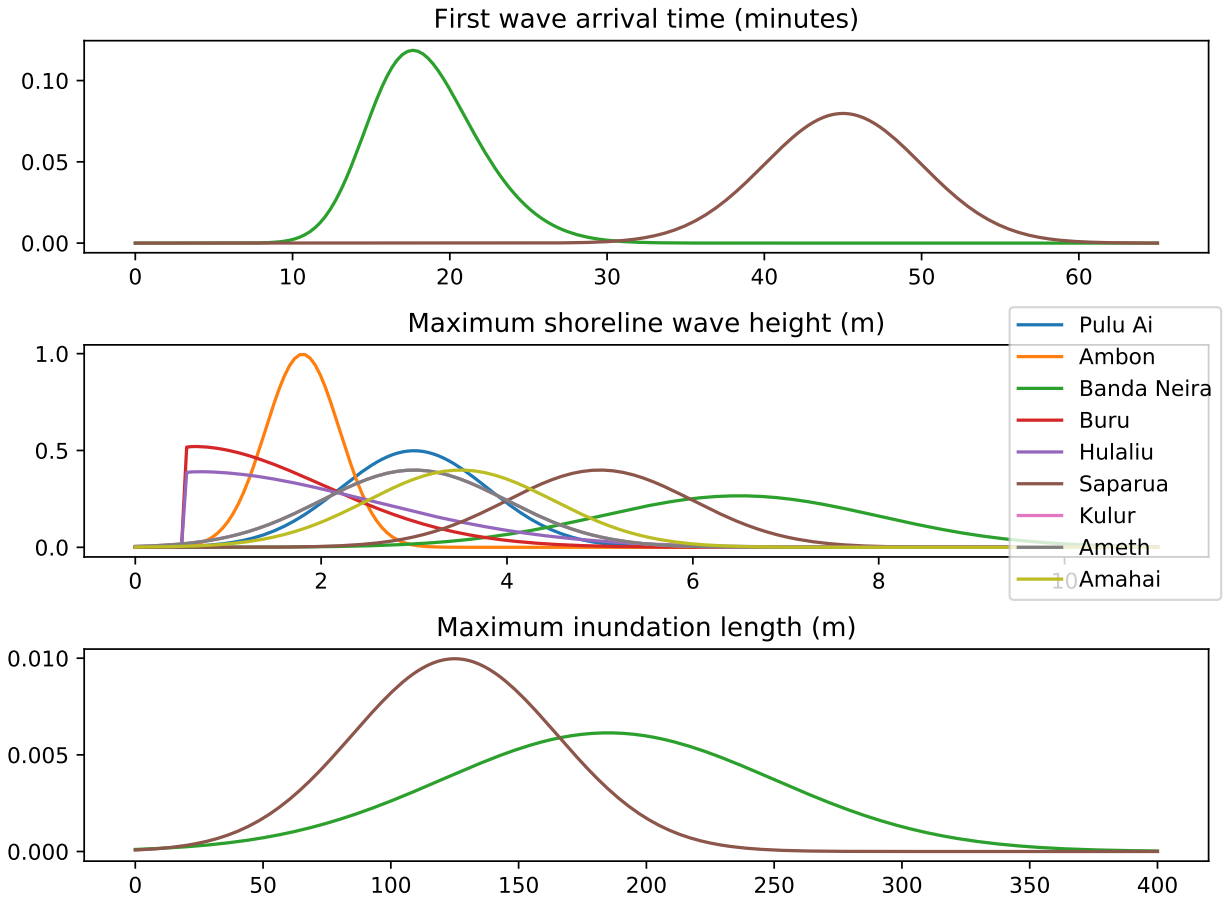

Figure 4.2: 1852 Banda Arc tsunami likelihood densities for the 13 observations at 9 locations. Each likelihood density represents an interpretation of the Wichmann catalog description

Chapter 5. Forward Model

To solve the inverse problem of inferring earthquake parameters from tsunami observations, the forward problem (computing tsunami observations from earthquake parameters) must first be specified. A tsunami is produced when an earthquake causes a sudden and significant deformation in the shape of the seafloor. The seafloor deformation produces a displacement of water, which propagates as a long wavelength tsunami wave. Because of this it is necessary

to model both the seafloor displacement from the earthquake, and the resultant propagation of the tsunami wave in the open ocean and eventual run up on the shoreline. Since our observations include the arrival time of the tsunami relative to the seismic waves of the earthquake (physically observed shaking), we will need to model the tsunami dynamically.

The software package *GeoClaw* [30, 31, 32, 33], which is part of *ClawPack* (see `http://clawpack.org`), provides a suite of tools that cover each of these modelling problems. To compute seafloor deformation, GeoClaw uses a Green's function solution to a particular elastics problem that looks for surface deformation in a half-space for an instantaneous rectangular disturbance [39, 40]. This widely accepted model for surface displacement of a specified earthquake is known as the Okada model. To compute the tsunami propagation, GeoClaw uses an adaptive mesh-refinement finite-volume solver for the shallow water equations. Here we provide a brief overview of both components of this forward model.

5.1 The Okada model for seafloor deformation

The Okada model computes seafloor deformation as an idealized elastic dislocation problem. The model assumes that the Earth is made of a homogeneous isotropic elastic material, with infinite extent and a flat surface (hence the half-space problem) [39]. For rupture zones which are small (relative to the radius of the earth) these assumptions are quite reasonable, and the Okada model has been shown to be a very useful approximation for seafloor deformation in the context of tsunami modeling.

Elastic mechanics is concerned with determining *displacement fields* for elastic solids. A displacement field is a vector-valued function $\boldsymbol{u}(x, y, z)$ specifying how the solid has deformed relative to some reference configuration. Linear elasticity theory for static problems in isotropic media is centered on the *Navier-Cauchy equilibrium equation*:

$$(\lambda + \mu)\nabla(\nabla \cdot \boldsymbol{u}) + \mu\nabla^2\boldsymbol{u} + \boldsymbol{F} = 0. \tag{5.1}$$

Here $\boldsymbol{F}$ is the total body force per unit volume. λ and μ are constants referred to as *Lamé's coefficients*. In geodynamics, we often take $\lambda = \mu$, which is a valid assumption for the Earth's crust [41]. μ is also known as the *shear modulus*.

The Okada model computes the displacement field at the surface $z = 0$ produced by a discontinuous displacement on a rectangular patch within the elastic material. The equations of the Okada model are closed-form expressions for a Green's function solution integrated over the rectangular source [39]. The model requires nine independent parameters, which describe the location, size, and orientation of the rectangle, as well as the direction and size of the displacement. Expressed in geographic coordinates, these parameters are:

- *latitude* and *longitude* of the rectangle center
- *depth* of the rectangle center
- *length* and *width* of the rectangle
- *strike*, the angle the top edge of the rectangle makes with due north
- *dip*, the angle the rectangle makes with the free surface
- *slip*, the amount of displacement (parallel to the rectangular surface)
- *rake*, the angle of the slip direction, relative to the strike direction

See Figure 5.1 for a diagram of the Okada parameters.

The computed deformation of the flat seafloor is added to the actual bathymetry to produce the post-earthquake bathymetry. This deformed seafloor is used to produce the initial condition for the GeoClaw finite volume solution to the shallow water equations.

In cases where the earthquake cannot be reasonably modeled as a single rectangular source, multiple rectangular patches can be used, and the resulting deformation summed. This is the case in the 1852 Banda Sea earthquake, as the curved arc of the subduction zone rules out a single rectangular source. In the following chapter, we discuss our approach for generating rectangular patches for earthquakes on the Banda Arc.

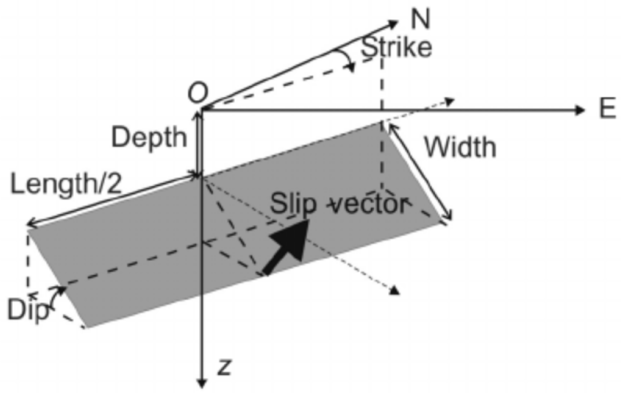

Figure 5.1: Okada model for a rectangular source. Figure from [42].

5.2 Forward propagation of the tsunami through Geoclaw

The propagation of the tsunami wave is computed via the nonlinear shallow water equations supplemented with the appropriate initial and boundary conditions dictated by the specified Okada parameters and bathymetry of the region. We simulate the tsunami generated by each Monte Carlo sample using the Geoclaw software package, [30, 31, 32, 33] which employs an adaptively-generated mesh for a finite volume based scheme. For bathymetry (sea-floor topography) we use the 1-arcminute etopo datasets available from the open access NOAA database[1] referred to hereafter as NOAA bathymetry, and for the coastline near each observational point we utilize higher resolution Digital Elevation Models (DEM) from the Consortiom for Spatial Information (CGIAR-CSI) [2] referred to below as DEM coastlines. These higher resolution topographical files yield a 3-arcsecond resolution on land, but give no additional information on the sub-surface bathymetry. The extent of each of these files is provided in Table 5.1.

In addition to these DEM coastline datasets and the NOAA bathymetry, we also took advantage of detailed sounding maps available at `http://inarisk.bnpb.go.id`. To convert this data into digitally accessible information, contours were taken from images exported from the website and then traced and interpolated in ArcGIS to produce approximate depths in

[1] `https://www.ngdc.noaa.gov/mgg/global/global.html`
[2] `http://srtm.csi.cgiar.org/srtmdata/`

Observation location	latitude extent	longitude extent
Banda Neira & Pulau Ai	$[-4.6, -4.467]$	$[129.6, 129.983]$
Ambon, Saparua, Haruku, & Nusa Laut	$[-3.881, -3.411]$	$[127.844, 128.909]$
Pulau Buru	$[-3.381, -3.271]$	$[127.041, 127.213]$
Amahai	$[-3.414, -3.269]$	$[128.866, 128.999]$

Table 5.1. Specification of the extent of the DEM coastline files used near each of the historically observed accounts.

the same regions as specified in Table 5.1. For example, the bathymetric readings based on this data are shown in Figure 5.2 for the bay of Amahai. The upper left panel in Figure 5.2 depicts the bathymetry data that is gleaned from `http://inarisk.bnpb.go.id` and digitized by interpolating across contours of constant depth in ArcGIS. The upper right panel of Figure 5.2 depicts the bathymetry/topography from the NOAA bathymetry dataset. Using the built in interpolative methods in Geoclaw's topotools package (`topotools.interp_unstructured` with the cubic interpolant, and a proximity radius of 1000), we interpolate the coastline and coarse bathymetry from the NOAA dataset to match the bathymetric contours from the upper right panel to produce the lower left panel. This lower left panel does not accurately capture any of the topographical features of the coastline and suffers significantly from interpolant error onshore as there are no bathymetric readings there. The actual shoreline and onshore topography is then overlaid from the DEM coastlines on top of the bottom left panel of Figure 5.2 to create the final product which is seen in the bottom right panel of the same Figure. This retains the improved bathymetric contours, and yields an accurate coastline and near-shore topographical profile.

This same process is repeated for Palau Buru, and the coastline near the islands of Ambon, Saparua, Haruku, and Nasu Luat. The resultant final bathymetric files are not shown here, but similar results hold. Finally, all of these high resolution bathymetric files are used by Geoclaw when the wave approaches these locations onshore.

For the region near Banda Neira and Palau Ai, the bathymetric data was still quite rough, particularly for the narrow channels between Banda Neira, Banda Api, and Lonthor. We obtained a set of soundings for this region from local government officials. Using the same

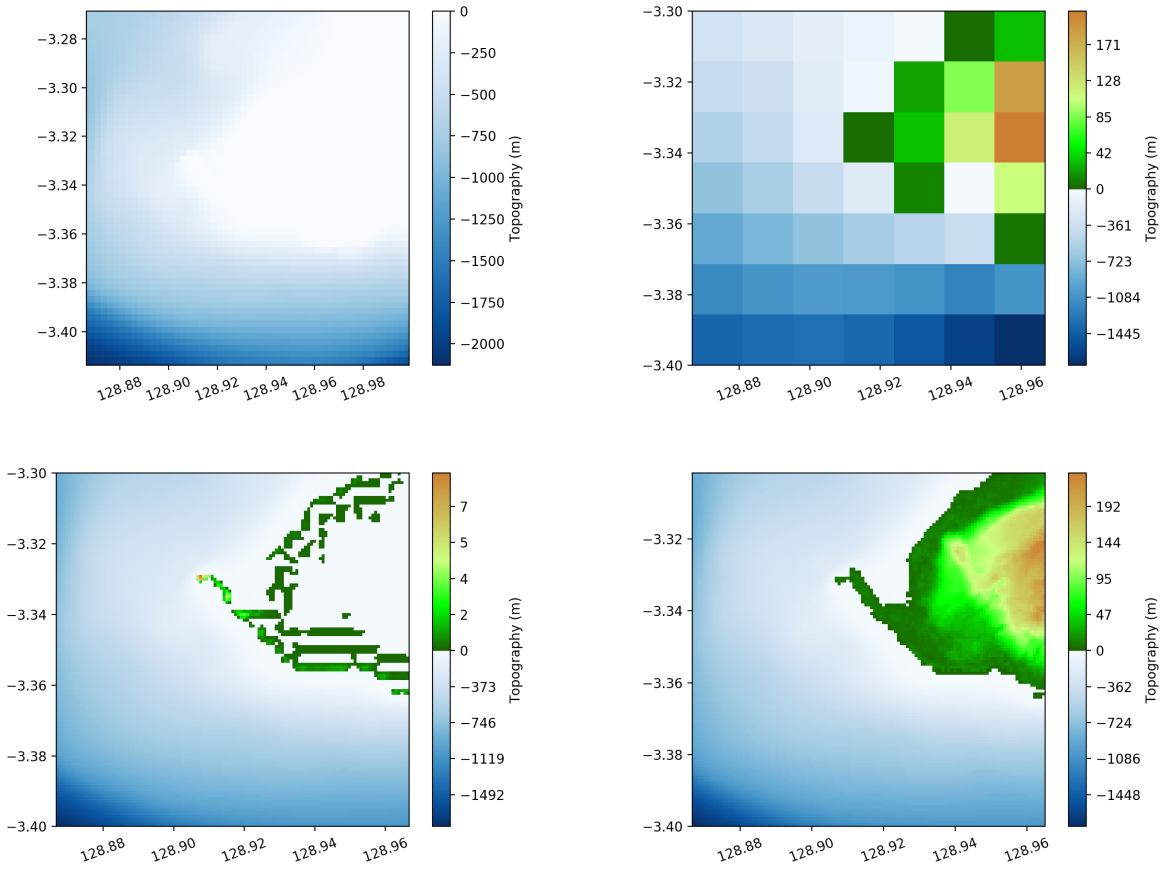

Figure 5.2: Combining all of the bathymetric and topographical sources into a single file for the bay near Amahai. The upper left figure demonstrates the bathymetry drawn from the level curves exported from `http://inarisk.bnpb.go.id`. The upper right figure shows the level of resolution for the NOAA bathymetry data. The lower left figure shows the interpolation of these two data sets (omitting the interior of the coast, i.e. all grid points from the NOAA bathymetry that are *not* below sea level, or border a grid below sea level). The lower right figure is the final product, combining the improved bathymetric data with the DEM coastline dataset.

approach as described above for the bay of Amahai, these discrete soundings are interpolated for the entire region surrounding the Banda islands (except that a linear interpolant is used instead of cubic due to the sparsity of the measurements) and overlayed with the DEM coastlines. The resultant bathymetry files for Palau Ai and the Banda islands are significantly improved and are critical for the inference as discussed below.

For the forward simulations of the tsunami wave, we employ an adjoint-based adaptive mesh strategy [43]. This entails solving a linearized adjoint equation backward in time with sources centered at each gauge location The solution of the adjoint equation produces waves that propagate backward in time from the desired observation locations to indicate what part of the forward wave will eventually influence the tsunami at those locations (see [43] for details). To initialize the adjoint solver, we place a smoothed Gaussian perturbation $h(x, y)$ to the wave height at each gauge location given by:

$$h(x, y) = \sum_k \exp(-r_k^2/150), \tag{5.2}$$

where r_k is the distance from the point (x, y) to the gauge location (x_k, y_k). The solution of the linearized adjoint problem guides the choice of refinement regions of the fully nonlinear forward model, indicating where the wave that will reach the observed locations will be at specific times. The benefit of using this approach as noted in [43] is that only those parts of the wave that will reach the desired locations are refined, i.e. the mesh refinement is restricted to those parts of the domain (in both space and time) that will most influence the final wave at the desired location. In addition, for the application at hand, we only need to run the backward adjoint solver once, and then the generated output can be used for every sample so long as the gauge locations are not changed. This saves a substantial amount of computational cost, allowing us to use a much finer mesh near the observational locations than a standard adaptive mesh would have allowed.

We use an adaptive mesh with 6 levels, starting with 6 arcminute resolution in the open water with no motion, and then going through $2\times$, $2\times$, $2\times$, $3\times$, and $5\times$ grid refinements

Observation location	latitude extent	longitude extent
Banda Neira	$[-4.576, -4.49]$	$[129.86, 129.95]$
Pulau Ai	$[-4.525, -4.515]$	$[129.76, 129.785]$
Ambon	$[-3.8, -3.66]$	$[127.98, 18.2]$
Hulaliu	$[-3.515, -3.478]$	$[128.53, 128.577$
Pulau Buru	$[-3.39, -3.27]$	$[127.05, 127.28]$
Saparua (near port)	$[-3.592, -3.572]$	$[128.65, 128.7]$
Saparua (main bay)	$[-3.626, -3.592]$	$[128.66, 128.717]$
Nusa Laut	$[-3.653, -3.644]$	$[128.804, 128.82]$
Amahai	$[-3.352, -3.328]$	$[128.9, 128.927]$
channel between Haruku & Saparua	$[-3.54, -3.515]$	$[128.53, 128.563]$
channel between Haruku & Saparua	$[-3.594, -3.54]$	$[128.552, 128.6]$

Table 5.2. Specification of the statically refined regions labeled according to the historically observed data. To maintain computational tractability, some choices were necessarily made regarding which regions in the computational domain were needed at the highest mesh refinement level. For instance, the narrow channel between the islands of Haruku and Palau Saparua is captured via two distinct refined grids to avoid having to much spatial refinement unnecessarily placed over land.

to those regions where the adjoint indicates the wave will be, resulting in the finest grid of 3 arcseconds which matches the fine resolution of the DEM coastline files. This means that the mesh levels are given by 6 arcminute, 3 arcminute, 1.5 arcminute, 45 arcsecond, 15 arcsecond, and 3 arcsecond resolution respectively. In addition to this dynamic adaptation of the mesh, we statically fix regions near each gauge at the highest mesh resolution (3 arcseconds) for the entirety of the simulation, thus accurately capturing the wave characteristics near the observed locations. These regions are explicitly specified in Table 5.2. Implementation of such a highly refined grid for the region in question required some minor modification of the default list lengths in the fortran code as described in the code repository. The backward adjoint solver is run on a 15 arcsecond grid and the output files are saved every 5 minutes to ensure adequate spatial and temporal resolution for the dynamic grid refinement. Geoclaw interpolates these output files temporally to determine the wave location throughout the entire simulation.

All other settings in Geoclaw are set to their default values. An adaptive time step is adjusted according to the Courant-Friedrichs-Lewy (CFL) condition with a desired CFL of 0.75. The spatial discretization in Geoclaw is a second order scheme with the MC limiter

[44] employed to avoid the development of un-physical shocks. All simulations are run for a physical time window of 1.5 hours to ensure that the wave has reached all of the relevant locations (for this event the longest historically recorded time between the earthquake and the arrival of the wave was approximately 40-45 minutes as shown in Figure 4.2). Each simulation of Geoclaw generates wave heights and arrival times at the locations shown in Figure 4.1.

Chapter 6. Parameters and Priors

6.1 Parameter Selection

To make efficient use of Bayesian methods, it is necessary to consider the dimensionality of the parameter space. As the number of parameters to be estimated increases, so does the difficulty of the sampling problem. This 'curse of dimensionality' appears in this setting because Bayesian inference boils down to the computation of high dimensional integrals. It is known that random walk MCMC methods converge arbitrarily slowly as the dimension of the parameter space increases [45].

A zeroth order approach is to consider the 9-dimensional parameter space for the Okada model. However, as discussed in the previous chapter, it is unreasonable to model the source earthquake as a single rectangular rupture. Naively, an N-subfault rupture zone would require a $9N$-dimensional parameter space, which produces an intractable sampling problem for any useful value of N (for even the simplest fault around the Banda Arc, we would expect $N \geq 3$).

To reduce the dimensionality of the parameter space, we make a distinction between the *(forward) model parameters* and the *sample parameters*. The model parameters are the direct inputs to the forward model: in this case, the $9N$ Okada parameters for an N-subfault rupture. The sample parameters are some sufficient lower-dimensional set, from which the model parameters can be computed. The sample parameters define the MCMC search space.

For reducing the number of sample parameters, a good starting point is to consider model parameters that may be assumed to take constant values. Among the nine Okada parameters, the rake angle can be reasonably fixed to 90°. This corresponds to pure thrust motion, which acts perpendicular to the strike of the fault. While strike-slip motion is certainly present in real megathrust earthquakes, thrust motion is the primary driver of seafloor deformation, and thus tsunami formation. Within the Okada model, rake angles other than $\theta = 90°$ are roughly equivalent to a reduction in slip by a factor of $\sin(\theta)$. It is therefore difficult to infer both the rake angle and the slip distance.

Another avenue is to seek model parameters that can be determined from other model parameters in the context of prior information. In the case of the 1852 Banda Arc Earthquake, a detailed model of the subduction zone geometry is available from the USGS *Slab2* dataset [37]. The Slab2 data for the Banda Arc is depicted in Figure 6.1. Depth, dip angle, and strike angle can be determined from latitude and longitude.

We are left with five of the Okada parameters: latitude, longitude, length, width, and slip. These could be chosen as the sample parameter space. However, a problem arises in choosing the triple of (length, width, slip) as sample parameters, due to their relationship with earthquake magnitude. The *scalar seismic moment* M_0 of an earthquake of length L, width W, and average slip S is defined as

$$M_0 = \mu L W S \tag{6.1}$$

where μ is the shear modulus of the rock, with dimensions of force per unit area. The scalar seismic moment was introduced by H. Kanamori in his definition of *moment magnitude* M_w [46]. Moment magnitude is an improvement over the classical Richter magnitude scale, and is now the standard magnitude scale used by the U.S. Geological Survey [47]. Moment magnitude is defined as

$$M_w = \frac{2}{3}(\log_{10} M_0 - 9.05). \tag{6.2}$$

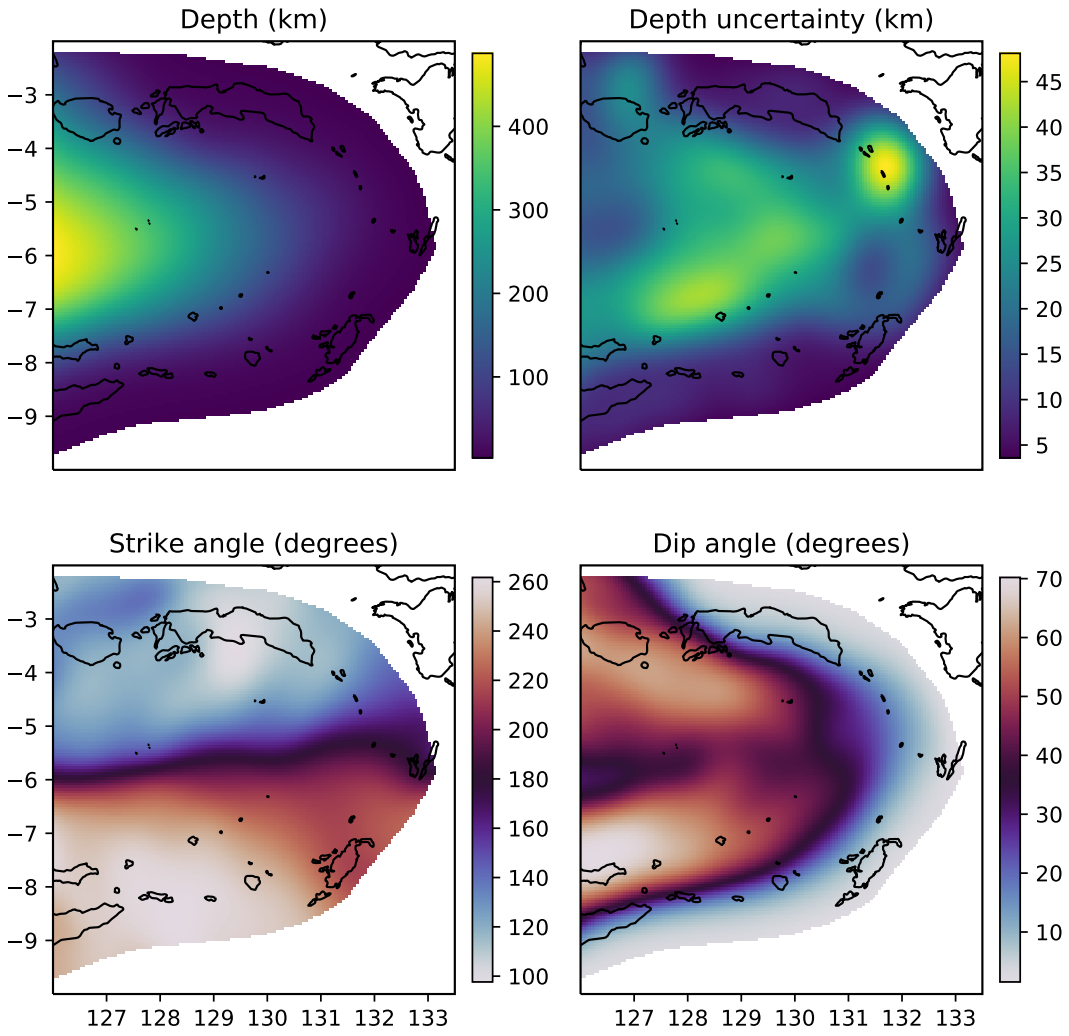

Figure 6.1: Slab2 depth, depth uncertainty, strike angle, and dip angle maps for the Banda Arc subduction interface.

It is observed that the empirical frequency of earthquakes of a given magnitude follows an exponential distribution [48]. In order to ensure that magnitude follows an exponential prior, we remove slip from the sample parameters and replace it with moment magnitude. Given values of magnitude, length, and width, slip can be back-calculated via Equations 6.1 and 6.2.

Equations 6.1 and 6.2 also highlight a challenge when using a random walk proposal kernel with these parameters. Since magnitude grows with the logarithm of length and width, any fixed choice of variance for length and width in the Gaussian proposal kernel will be inappropriate for all but a limited range of magnitudes. Therefore, we introduce magnitude-normalized substitutes for length and width as sample parameters. Using the Wells-Coppersmith dataset [49] (augmented with additional collected data), we computed linear least squares fits for $\log L$ and $\log W$ against magnitude. These fits are displayed in Figure 6.2. Our magnitude-normalized substitutes are $\Delta \log L$ and $\Delta \log W$, the "residuals" compared to the linear best fit. In other words: given values for M_w, $\Delta \log L$, and $\Delta \log W$, length and width are computed as:

$$\log L = aM_w + b + \Delta \log L$$

$$\log W = cM_w + d + \Delta \log W$$

where a, b, c, d are the coefficients of the linear best fits.

To the five sample parameters (latitude, longitude, magnitude, $\Delta \log L$, $\Delta \log W$), we add a sixth parameter: *depth offset*. The Slab2 data includes estimates of uncertainty in the subduction interface depth (see Figure 6.1). Depth offset accounts for this uncertainty by allowing for earthquakes that are situated somewhat deeper or shallower than is specified in the Slab2 depth map.

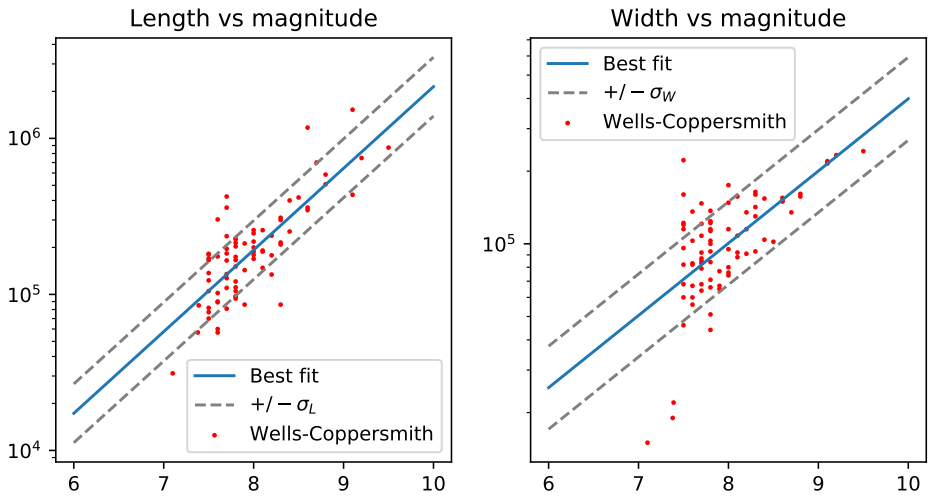

Figure 6.2: Wells-Coppersmith data and linear best fits for $\log L$ and $\log W$ against M_w.

6.2 Computing Subfault Model Parameters

As discussed before, it is necessary to model the earthquake as a collection of rectangular subfaults that conform to the subduction interface geometry. Here we describe our approach for "decompressing" the six sample parameters into the Okada parameters for N rectangular subfaults.

The basic approach is to "break" a single rectangular rupture zone into an $m \times n$ grid of identical subrectangles, which are then placed to conform to the interface geometry. Each of these subrectangles has length L/m and width W/n, where L and W are the length and width of the full rupture zone.

To fit this grid to the fault, it is easiest to use odd values for m and n. By experimentation, we determined to use $m = 11$ and $n = 3$ for the 1852 Banda Arc Earthquake. We place a single point at the latitude and longitude of the centroid of the full rupture zone. Using the Slab2 map of strike angle, we move in opposite directions, staying parallel to strike. Every L/m kilometers, we place another point. This continues until m points are placed. For each point, we then move in opposite directions, perpendicular to the strike angle, placing points every W/n kilometers, until all mn points have been placed. These points

are the latitude/longitude coordinates for the centers of the subrectangles. This procedure is displayed in Figure 6.3.

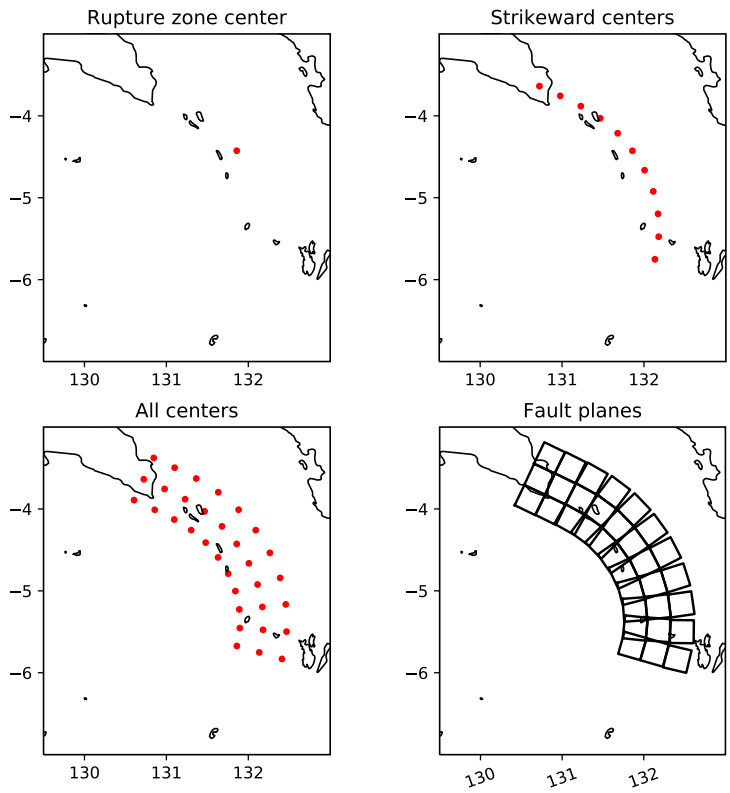

Figure 6.3: Placing subrectangles contoured to interface geometry. First, a point is placed at the center of the rupture zone. Points are then placed forwards and backwards following the strike angle (essentially following level curves of depth). Additional points are placed up-dip and down-dip. Using Slab2 depth, dip, and strike data, Okada parameters for rectangles centered at each point are computed.

Having specified the latitude, longtitude, length, and width for each subrectangle, the remaining Okada parameters are determined as follows. Each subrectangle is given the same slip value as determined by Equations 6.1 and 6.2. The strike and dip angles are determined by the Slab2 strike and dip maps. The depth is determined by the Slab2 depth map, plus the value of the depth offset sample parameter. As discussed above, all subrectangles are assigned a rake angle of 90°.

6.3 Prior Distributions

Selection of appropriate prior distributions is a key step in good Bayesian inference. An over-specified prior can overwhelm the data, and an under-specified prior may allow for parameter values that are non-physical. Here we discuss our choice of prior distributions for latitude, longitude, magnitude, $\Delta \log L$, $\Delta \log W$, and depth offset for the 1852 Banda Arc event.

Prior constraints on earthquake latitude and longitude are derived from the subduction interface geometry. Large earthquakes can only be supported in a certain range of depth: too deep, and the crust is too plastic to store the strain energy of a large earthquake [50], too shallow, and the rupture interface would extend above the surface. We take the approach that, a priori, depth is the primary constraint on earthquake location. Since the Slab2 dataset gives a depth map for the entire Banda Arc, any probability distribution on depth produces an implied distribution on latitude and longitude. Based on the augmented Wells-Coppersmith dataset, we chose a truncated normal distribution for depth. This distribution is supported on $[2.5, 50]$ kilometers, with a mean of 30km and a standard deviation of 5km. Evalutating the pdf of this distribution at each latitude/longtitude coordinate, via the Slab2 depth map, gives a non-negative continuous function. Although this function does not integrate to unity, the normalizing constant cancels out in the evaluation of the Metropolis-Hastings acceptance parameter α. The unnormalized logpdf of the latitude/longitude prior is displayed in Figure 6.4.

As discussed above, earthquake magnitude is observed to approximately follow an exponential distribution. It is known that the exponential scaling cannot continue indefinitely in the large magnitude regime, and a number of approaches have been used to address this (see [48]). We take a simple approach of right-truncating the exponential distribution at magnitude 9.5. A consensus estimate for the parameter of the exponential distribution is $\lambda = .5$ [48].

Since $\Delta \log L$ and $\Delta \log W$ are magnitude-normalized length and width, defined as residuals against a linear best-fit, we chose Gaussian prior distributions with mean zero. The

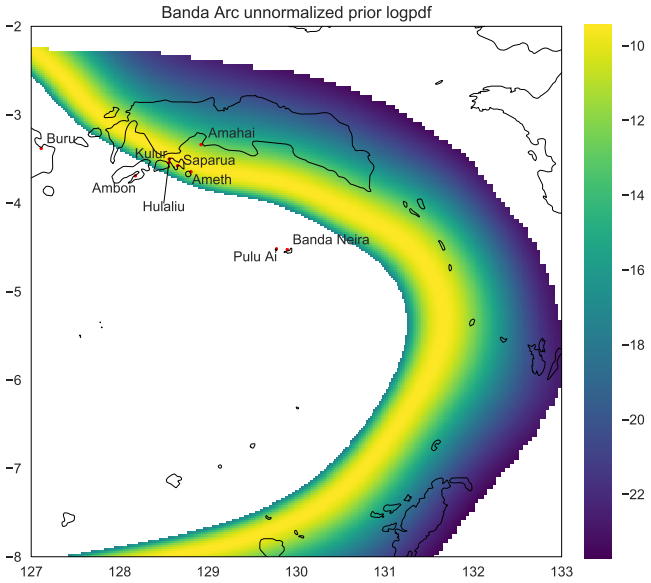

Figure 6.4: Unnormalized logpdf of the latitude/longitude prior distribution

standard deviations for these distributions are determined from the sample variances for the residuals in the augmented Wells-Coppersmith dataset against the linear fit. These values are $\sigma_{\Delta \log L} = 0.188$ and $\sigma_{\Delta \log W} = 0.172$.

The prior for depth offset was chosen based on the Slab2 depth uncertainty data. The average reported uncertainty is roughly 5km, so a mean-zero normal distribution with standard deviation 5 was selected.

Table 6.1. Prior distributions for the 1852 Banda Arc earthquake

Parameter name(s)	Kind	Distribution Parameters
Latitude & longitude	pre-image of truncated normal via depth	• $\mu = 30$km • $\sigma = 5$km • $(a, b) = (2.5\text{km}, 50\text{km})$
Magnitude	truncated exponential	• $\lambda = .5$ • $(a, b) = (6.5, 9.5)$
$\Delta \log L$	normal	• $\mu = 0$ • $\sigma = .188$
$\Delta \log W$	normal	• $\mu = 0$ • $\sigma = .172$
depth offset	normal	• $\mu = 0$ • $\sigma = 5$km

Chapter 7. Software Implementation

A key product of our research is the development of a Python package which implements our method. Called `tsunamibayes`, this package is designed to be modular and flexible. Since each hypothetical scenario may have a unique interpretation as a Bayesian inference problem (different parameters/priors, modified/generalized forward model, additional types of observations), the core code of the module does not assume particular features, but rather provides a suite of tools that can be recombined or modified to suit the needs of the user.

The central part of the module is a collection of base classes that implement an abstract interface between the general components of a Metropolis-Hastings sampler. These classes are summarized in Table 7.1. Each base class contains several unimplemented methods that are called within the core Metropolis-Hastings loop. These methods are intended to be implemented in scenario-specific inherited classes. For example, the class `BaseScenario` has an unimplemented method `BaseScenario.propose()`. Although to this point we have only used a random walk proposal kernel, `tsunamibayes` does not assume that random walk will always be used, and thus leaves the proposal kernel to be defined by the user.

The modularity of the software is on display when defining the choice of sample parameters and specifying the relationship to the forward model parameters. `BaseScenario` acts as a "central processing unit", and allows the user to easily swap in alternative parameterizations of the sample space while using the same forward model parameters. This is important when considering scenarios other than the 1852 Banda Arc event, which have different prior constraints on earthquake parameters and thus may need to search over different sample parameter spaces.

The package is open-source, available on GitHub: `https://github.com/jpw37/tsunamibayes`.

Table 7.1. `tsunamibayes` abstract base classes

Class	`BaseScenario`	`BaseForwardModel`	`BasePrior`	`BaseFault`
Description	Implements Metropolis-Hastings, file I/O	Computes forward model and evaluates likelihood	Evaluates prior distribution	Manages geographic data for fault zones, computes Okada parameters
Core methods	• `sample()`: Metropolis-Hastings algorithm • `propose()`*: Proposal kernel • `proposal_logpdf()`*: logpdf of proposal kernel • `map_to_model_params()`*: Map from sample parameters to forward model parameters	• `run()`*: Compute forward model for given model parameters • `llh()`*: Evaluate log-likelihood for given forward model output	• `logpdf()`*: Evaluates prior logpdf • `rvs()`*: Random sample from prior (optional)	• `subfault_split()`: Compute split into subfaults conforming to fault geometry • `depth_map()`*: Depth of fault interface at given lat/lon coordinates • `strike_map()`*: Strike angle of fault interface • `dip_map()`*: Dip angle of fault interface

*methods to be defined in inherited classes

Chapter 8. Results

8.1 Summary

Our analysis of the 1852 Banda Arc Earthquake was conducted over a period of four months, from April to July 2020. Using the compute resources available through BYU's Office of Research Computing, we ran fourteen (14) parallel chains. These chains were initialized at locations around the Banda Arc, and then run for a burn-in period of 6000 samples. After the burn-in, the chains were resampled from the last states. The chains were run until each produced 12000 samples, for a total of 168000. At 6000 samples, the chains were again re-initialized with a resampling step.

From prior analysis, we anticipated that the posterior mass would be concentrated in a region near 4.5°S, 131.5°E. This was indeed the case. However, a separate mode of the posterior was located during burn-in, in a wider region to the south around 5.5°S, 131.5°E. However, the posterior pdf values for samples in this other region were on the order of 100 times lower in comparison to samples in the northern region. Interestingly, chains in the south would not cross into the north, and vice-versa. We discovered that this was due to major differences in rupture zone geometries: northern samples tended to be longer and narrower for their given magnitude, while chains in the south tended to be wider and shorter. This means that while the two groups of chains appeared superficially close in latitude/longitude/magnitude, they were actually in well-separated regions of the full parameter space. After resampling at the end of burn-in, all fourteen chains were located in the north.

Figures 8.1 and 8.2 summarize the posterior distribution. We see some stark differences compared to the prior distributions. The posterior is notably concentrated in the small region near 4.5°S, 131.5°E, which is situated in a shallow part of the subduction interface. Also notable is the marginal posterior for magnitude: despite a prior that heavily preferred lower magnitudes, the posterior is still concentrated around earthquakes of magnitude 8.8. More subtle inference is seen in magnitude-normalized length and width. The posterior favors

rupture zones that are relatively narrow. This can be explained as a wave-height/arrival-time tradeoff. In order for an earthquake to produce the observed wave heights in Banda Neira, for instance, it needs to be quite a large event. However, larger earthquakes, all else being equal, have rupture zones that are closer to Banda Neira, thus reducing the arrival time of the wave. Therefore, a large but narrower rupture zone captures the balance in wave-height/arrival-time.

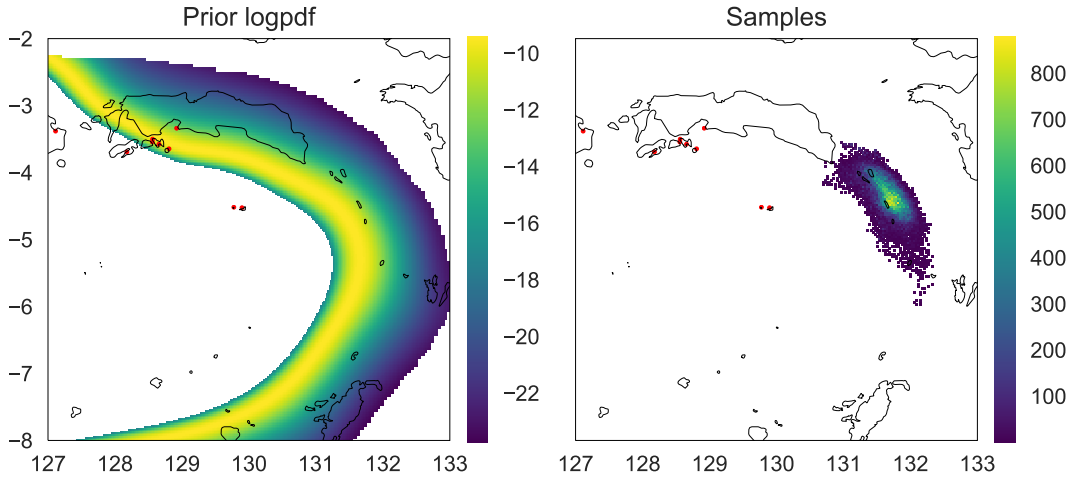

Figure 8.1: Sample epicenters from the posterior compared with the prior in latitude/longitude. The posterior is concentrated in a small region in the northeast.

Given that the posterior is 6-dimensional, it can be difficult to visualize. Figure 8.3 displays the approximate conditional expectation for magnitude, depth offset, $\Delta \log L$, and $\Delta \log W$, conditioned on latitude and longitude. Several trends are apparent. The farther outside the arc, the higher the expected magnitude. This is not surprising, as higher magnitudes would be required to produce large enough waves at that distance. Furthermore, the farther outside the arc, the greater the value of depth offset. This appears to counteract the shallowing of the fault interface towards the outside of the arc, ultimately producing earthquakes at constant depth among accepted samples. The closer the center of the rupture is to the coast of Seram, the shorter the rupture. This is likely due to the rupture extending

underneath Seram Island, which leads to a smaller tsunami (as only some of the rupture occurs beneath the ocean). Thus, a shorter rupture zone increases the slip (and thus wave height), counteracting the influence of Seram Island.

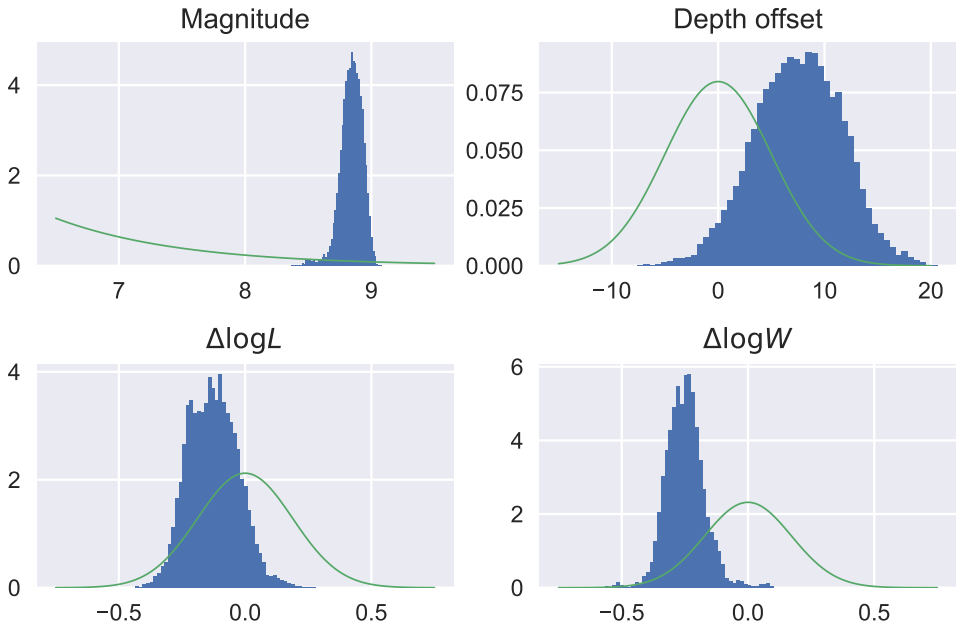

Figure 8.2: Magnitude, depth offset, $\Delta \log L$, and $\Delta \log W$ posterior histograms, compared to the associated prior distribution densities (plotted in green).

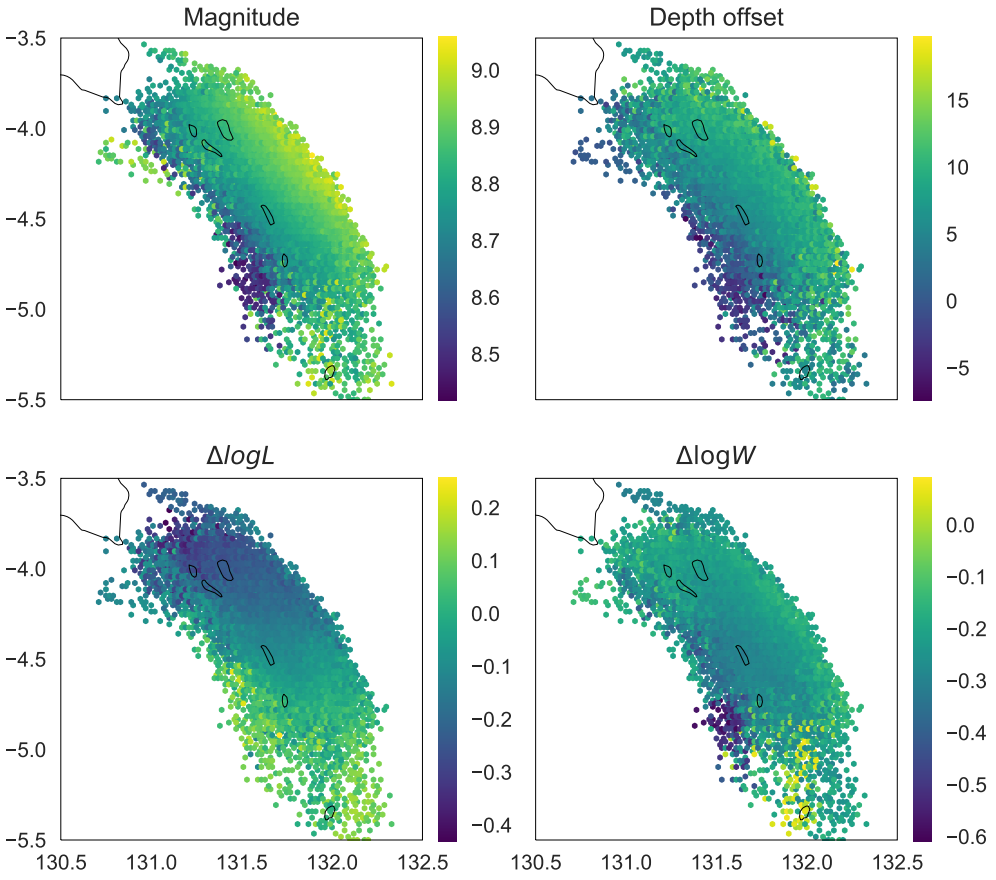

Figure 8.3: Posterior conditional expectation of magnitude, depth offset, $\Delta \log L$, and $\Delta \log W$ conditioned on latitude/longitude.

8.2 Mixing and Convergence

It is important to give attention to the quality of the mixing of the Markov chains, as well as to quantify, as much as possible, the degree to which the chains have converged to the true posterior. There are a number of approaches to analyze this. For mixing, autocorrelation analysis of time-series data is a useful measure. A well-mixed chain will have low autocorrelation for sufficiently long lag intervals. Figure 8.4 displays the chain-by-chain autocorrelation for the sample log-likelihood. It can be seen that a number of chains still displayed nontrivial autocorrelation even across longer lag intervals. Figure 8.5 shows a

single chain as a time series, and it becomes more clear that the proposal kernel is perhaps too conservative in $\Delta \log L$, $\Delta \log W$, and depth offset. Increasing the variance in the proposal kernel for these parameters could improve mixing, and thus the rate of convergence.

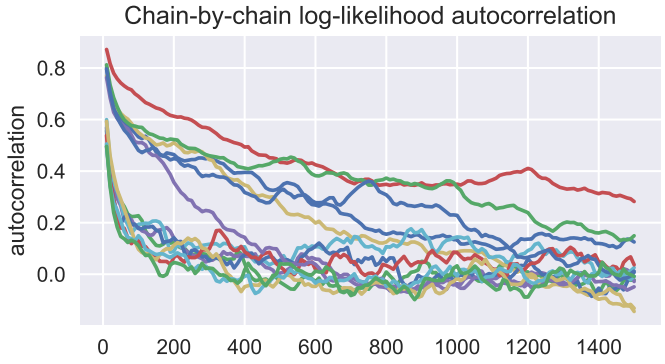

Figure 8.4: Autocorrelation for the 14 chains' log-likelihood values

Although tuning of the proposal kernel may help accelerate convergence, it is still possible that our chains have indeed converged. A standard metric for convergence when multiple chains can be compared is the *Gelman-Rubin diagnostic* R [51]. R is a measurement of the extent to which the chains have converged to the same distribution. When R is close to 1, and generally less than 1.1 or 1.2, the chains are all mixing around the same distribution, presumably the posterior. Figure 8.6 shows a rolling plot of R for each sample parameter. The scores all fall below 1.1 by 8000 samples. This is good evidence that our chains have converged to the posterior. It is notable that $\Delta \log L$, $\Delta \log W$, and depth offset take longer to cross the threshold of 1.1, providing further evidence that mixing in those parameters could be improved.

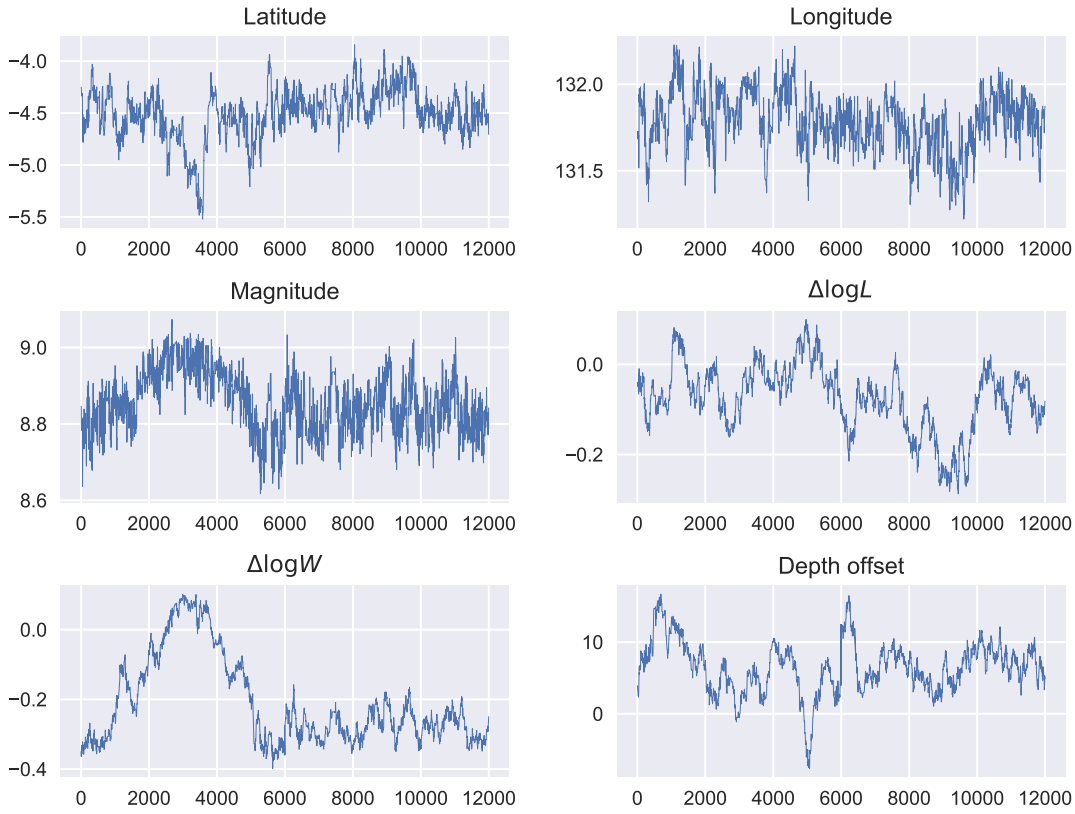

Figure 8.5: Time series for single chain. Depth offset, $\Delta \log L$, and $\Delta \log W$ display nontrivial autocorrelation.

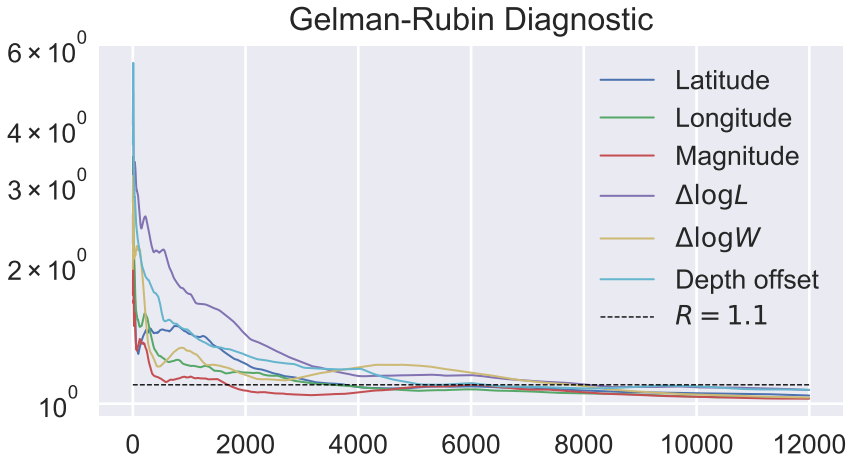

Figure 8.6: Cumulative Gelman-Rubin diagnostic scores for each sample parameter. At around 8000 samples, all sample parameters had scores under 1.1, indicating convergence.

8.3 Forward Model and Output

It is important to remember that our samples represent the posterior distribution for our particular choice of parameterization of the latent space of possible earthquakes. Figure 8.7 shows the mean seafloor deformation produced by the Okada model, among the 168000 samples. This demonstrates that the rupture zone is, on average, quite large.

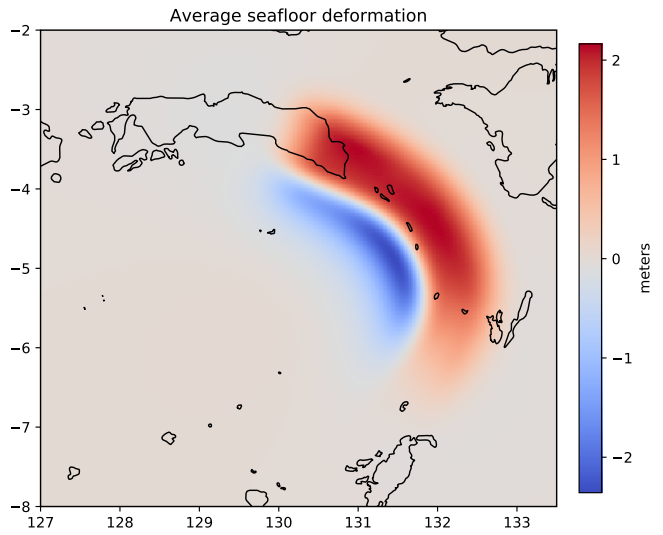

Figure 8.7: Average computed seafloor deformation for posterior samples.

It is also worth considering the implied observation distributions. This is known as the *posterior predictive distribution*. Figure 8.8 shows the model output for the posterior samples, compared against the likelihood densities for each observation. Banda Neira and Saparua provided the largest contribution to the likelihood, and we see that the posterior samples broadly matched our interpretation of the observations there. The posterior samples at Kulur and Ameth stand out as different from the likelihood, with waves smaller than our interpretation of the accounts. However, this is acceptable, given that these accounts were not specific, and we assigned wide distributions to them. Overall, the posterior distribution is consistent with the observations recorded in our sources.

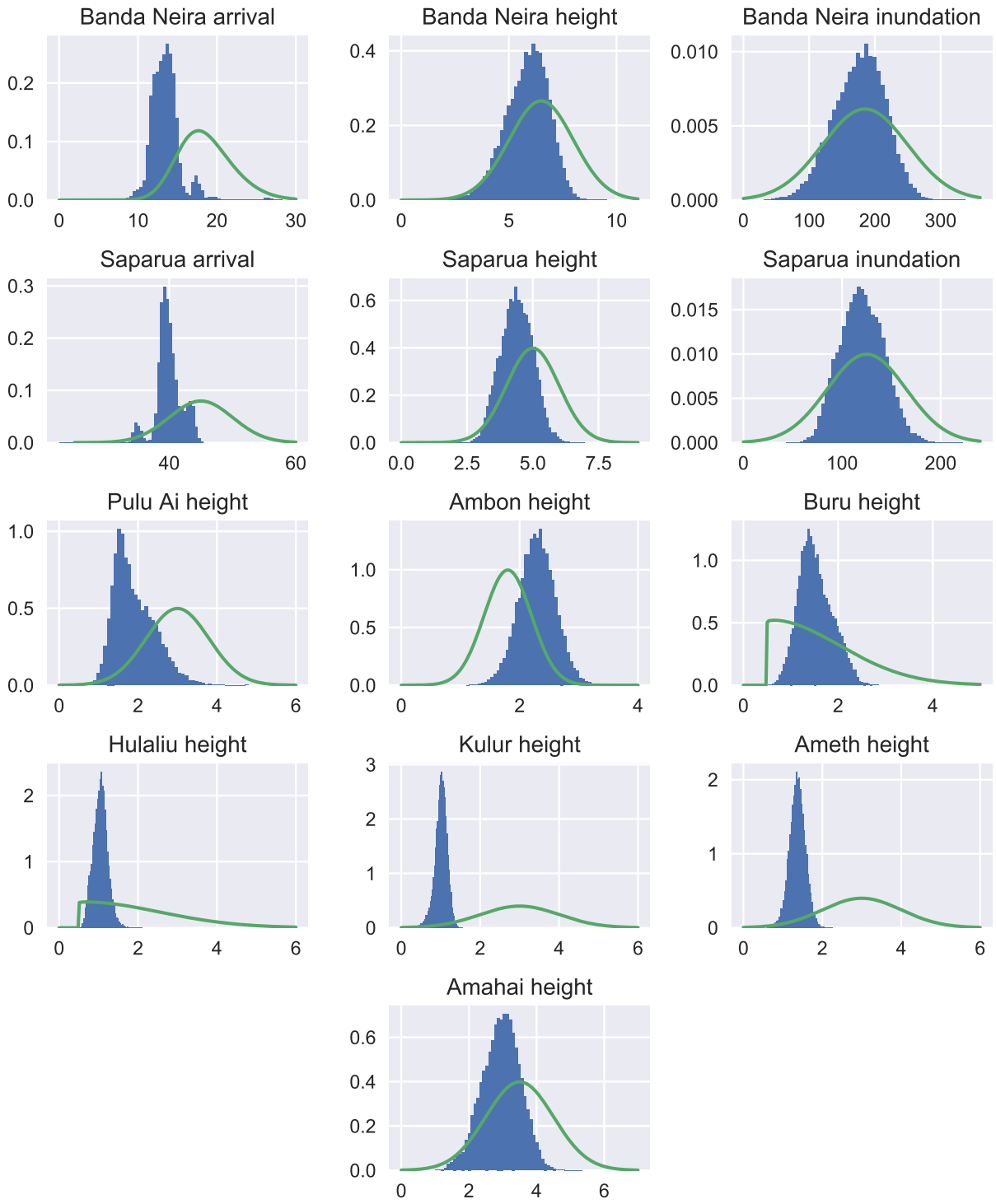

Figure 8.8: Model output compared to likelihood densities (plotted in green). Arrival times are in minutes, wave height and inundation length are in meters.

8.4 Claim & Corroborating Evidence

The implied claim of our posterior distribution is this: the 1852 Banda Arc tsunami was caused by a magnitude ∼8.8 megathrust earthquake centered near 4.5°S, 131.5°E. During analysis, we discovered an item of corroborating evidence for this claim, in the form of the Slab2 depth uncertainty data. The Slab2 model of the subduction zone is based on seismic data that can be used to infer the interface geometry. The more earthquakes that have occurred on a particular segment of a fault, the more certain we can be of the geometry. Regions of uncertainty correspond to "seismic gaps": fault segments that have been relatively silent during the modern period of instrumental data. A seismic gap may represent a location where hundreds of years of stress has accumulated, which eventually results in a large earthquake when the fault slips and the stress is released [52]. While not all seismic gaps turn out to be dangerous [53], they are still important to consider as possible sources for an event such as the 1852 Banda Arc earthquake.

Both the Slab2 depth uncertainty map, and the underlying seismic dataset, demonstrate the presence of a seismic gap in the region where our posterior distribution is concentrated (see Figures 8.9 and 8.10). This can be viewed as evidence that supports the results of our analysis.

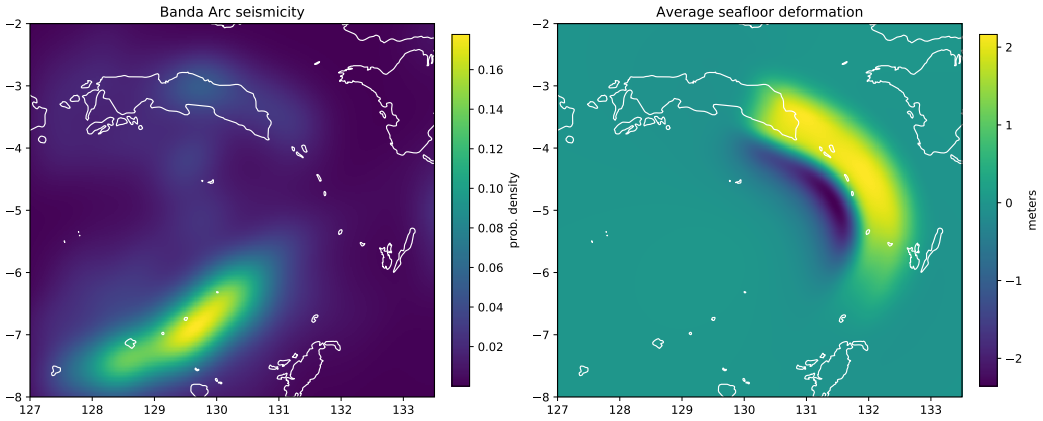

Figure 8.9: Seismicity of the Banda Arc compared to the average posterior seafloor deformation. The plot on the left displays the frequency of earthquakes in the Banda Arc by location. The average posterior earthquake is located within a seismic gap. Seismicity data was smoothed using a Gaussian KDE.

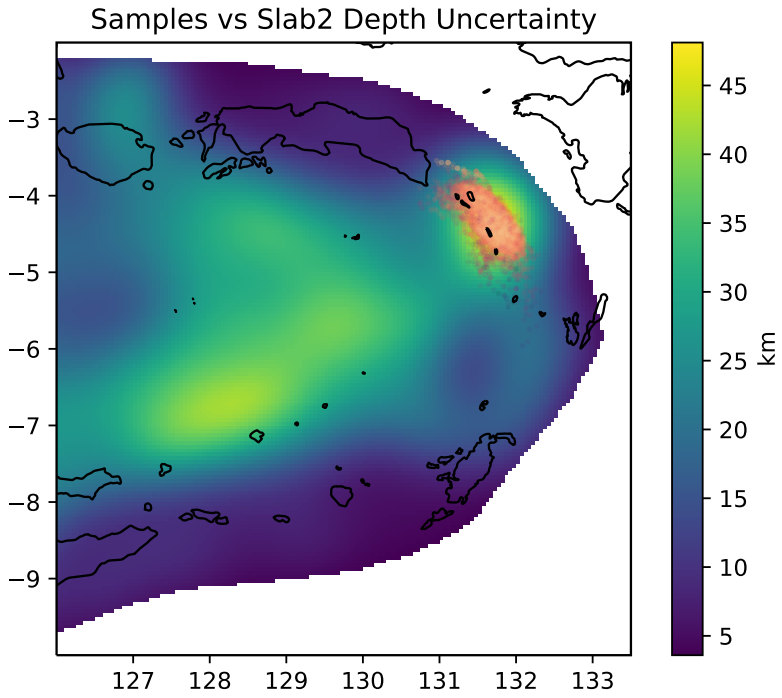

Figure 8.10: Posterior samples overlaid on Slab2 depth uncertainty map. Earthquake epicenters match with the Slab2 region of high uncertainty, which is derived from a seismic gap there.

Chapter 9. Conclusion

Probabilistic interpretation of anecdotal accounts appears to be a promising approach, inasmuch as it opens the door to using the tools of Bayesian inference. We have demonstrated that it is feasible to express the problem of historical earthquake reconstruction in the Bayesian framework and that the resulting problem can be made computationally tractable.

A priority for future work is to thoroughly validate our method. While we are encouraged by the results for the 1852 Banda Arc tsunami, we intend to run similar analyses for modern tsunamis for which the ground truth is known. Reconstructing events such as the 2004 Sumatra-Andaman tsunami and the 2011 Tōhoku tsunami would be a good test for our method.

We intend to apply our method to a number of other Wichmann-catalog recorded events, including the 1820 Makassar tsunami. Other events pose new challenges for our method. For example, the 1820 Makassar tsunami does not have a definite fault of origin. Expanding our method to handle the case of multiple candidate source faults will be important.

Other challenges are computational in nature. Despite having access to large-scale computing power, our method still required roughly four months to reach just over 150,000 samples. We have so far been limited to slowly-converging random walk Metropolis-Hastings. There are a number of more sophisticated MCMC sampling methods in existence, such as Hamiltonian Monte Carlo. It will be non-trivial to adapt our method to make use of higher-order samplers, as most require evaluation of the gradient of the forward model, something which is not available in this setting.

Perhaps the most important ongoing work will be to make the results of our analysis useful for seismic hazard mitigation. To this end, we plan on modeling a number of scenarios that represent what would happen if the 1852 Banda Arc earthquake were to reoccur. The ultimate justification for our work is in trying to better inform efforts to protect people who live in harm's way.

Bibliography

[1] K. McCue. Seismic hazard mapping in Australia, the Southwest Pacific and Southeast Asia. *Annals of Geophysics*, 42:1191–1198, 1999.

[2] Ron Harris and Jonathan Major. Waves of destruction in the East Indies: the Wichmann catalogue of earthquakes and tsunami in the Indonesian region from 1538 to 1877. *Geological Society, London, Special Publications*, 441:SP441–2, 2016.

[3] Hamzah Latief, Nanang Puspito, and Fumihiko Imamura. Tsunami Catalog and Zones in Indonesia. *Journal of Natural Disaster Science*, 22:25–43, 01 2000.

[4] KR Newcomb and WR McCann. Seismic history and seismotectonics of the Sunda Arc. *Journal of Geophysical Research: Solid Earth*, 92(B1):421–439, 1987.

[5] Ron Harris. The Nature of the Banda Arc-Continent Collision in the Timor Region. In D. Brown and P.D. Ryan, editors, *Arc-Continent Collision*, pages 163–211. Springer Berlin Heidelberg, Berlin Heidelberg, 2011.

[6] Emile A Okal and Dominique Reymond. The mechanism of great Banda Sea earthquake of 1 February 1938: applying the method of preliminary determination of focal mechanism to a historical event. *Earth and Planetary Science Letters*, 216(1-2):1–15, 2003.

[7] Kerry Sieh, Danny H Natawidjaja, Aron J Meltzner, Chuan-Chou Shen, Hai Cheng, Kuei-Shu Li, Bambang W Suwargadi, John Galetzka, Belle Philibosian, and R Lawrence Edwards. Earthquake supercycles inferred from sea-level changes recorded in the corals of west Sumatra. *Science*, 322(5908):1674–1678, 2008.

[8] Aron J Meltzner, Kerry Sieh, Hong-Wei Chiang, Chuan-Chou Shen, Bambang W Suwargadi, Danny H Natawidjaja, Belle E Philibosian, Richard W Briggs, and John Galetzka. Coral evidence for earthquake recurrence and an AD 1390–1455 cluster at the south end of the 2004 Aceh–Andaman rupture. *Journal of Geophysical Research: Solid Earth*, 115(B10), 2010.

[9] Aron J Meltzner, Kerry Sieh, Hong-Wei Chiang, Chuan-Chou Shen, Bambang W Suwargadi, Danny H Natawidjaja, Belle Philibosian, and Richard W Briggs. Persistent termini of 2004-and 2005-like ruptures of the Sunda megathrust. *Journal of Geophysical Research: Solid Earth*, 117(B4), 2012.

[10] Aron J Meltzner, Kerry Sieh, Hong-Wei Chiang, Chung-Che Wu, Louisa LH Tsang, Chuan-Chou Shen, Emma M Hill, Bambang W Suwargadi, Danny H Natawidjaja, Belle Philibosian, et al. Time-varying interseismic strain rates and similar seismic ruptures on the Nias–Simeulue patch of the Sunda megathrust. *Quaternary Science Reviews*, 122:258–281, 2015.

[11] Kruawun Jankaew, Brian F Atwater, Yuki Sawai, Montri Choowong, Thasinee Charoentitirat, Maria E Martin, and Amy Prendergast. Medieval forewarning of the 2004 Indian Ocean tsunami in Thailand. *Nature*, 455(7217):1228, 2008.

[12] Katrin Monecke, Willi Finger, David Klarer, Widjo Kongko, Brian G McAdoo, Andrew L Moore, and Sam U Sudrajat. A 1,000-year sediment record of tsunami recurrence in northern Sumatra. *Nature*, 455(7217):1232, 2008.

[13] Stein Bondevik. Earth science: The sands of tsunami time. *Nature*, 455(7217):1183, 2008.

[14] E Bryant, Grant Walsh, and Dallas Abbott. Cosmogenic mega-tsunami in the Australia region: are they supported by Aboriginal and Maori legends? *Geological Society, London, Special Publications*, 273(1):203–214, 2007.

[15] Barbara Dix Grimes. Mapping Buru: The Politics of Territory and Settlement on an Eastern Indonesian Island. *Sharing the Earth, Dividing the Land: Land and territory in the Austronesian world*, pages 135–155, 2006.

[16] Anthony Reid. Two hitherto unknown Indonesian tsunamis of the seventeenth century: Probabilities and context. *Journal of Southeast Asian Studies*, 47(1):88–108, 2016.

[17] Z. Y. C. Liu and R. A. Harris. Discovery of possible mega-thrust earthquake along the Seram Trough from records of 1629 tsunami in eastern Indonesian region. *Natural Hazards*, 72:1311–1328, 2014.

[18] TszMan L Fisher and Ron A Harris. Reconstruction of 1852 Banda Arc megathrust earthquake and tsunami. *Natural Hazards*, 83(1):667–689, 2016.

[19] Jonathan Griffin, Ngoc Nguyen, Phil Cummins, and Athanasius Cipta. Historical Earthquakes of the Eastern Sunda Arc: Source Mechanisms and Intensity-Based Testing of Indonesia's National Seismic Hazard Assessment. *Bulletin of the Seismological Society of America*, 109, 12 2018.

[20] Stacey Servito Martin, Linlin Li, Emile A. Okal, Julie Morin, Alexander E. G. Tetteroo, Adam D. Switzer, and Kerry E. Sieh. Reassessment of the 1907 Sumatra "Tsunami Earthquake" Based on Macroseismic, Seismological, and Tsunami Observations, and Modeling. *Pure and Applied Geophysics*, 2019.

[21] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, 2005.

[22] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer Science & Business Media, 2005.

[23] Masoumeh Dashti and Andrew M Stuart. The Bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 311–428, 2017.

[24] Alberto Malinverno. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3):675–688, 2002.

[25] Jun'ichi Fukuda and Kaj M Johnson. A fully Bayesian inversion for spatial distribution of fault slip with objective smoothing. *Bulletin of the Seismological Society of America*, 98(3):1128–1146, 2008.

[26] Ihab Sraj, Kyle T Mandli, Omar M Knio, Clint N Dawson, and Ibrahim Hoteit. Uncertainty quantification and inference of Manning's friction coefficients using DART buoy data during the Tōhoku tsunami. *Ocean Modelling*, 83:82–97, 2014.

[27] Ihab Sraj, Kyle T Mandli, Omar M Knio, Clint N Dawson, and Ibrahim Hoteit. Quantifying uncertainties in fault slip distribution during the Tōhoku tsunami using polynomial chaos. *Ocean Dynamics*, 67(12):1535–1551, 2017.

[28] A Wichmann. The earthquakes of the Indian archipelago from 1858 to 1877. *Verhandl. Koninkl. Akad. van Wetenschappen, 2nd sec*, 22(5):1–209, 1922.

[29] Jacob Swart. Verhandelingen en Berigten Betrekkelijk het Zeewegen en de Zeevaartkunde (English: Treatises and Reports Related to the Seaways and Nautical Sciences). 13:257–274, 1853.

[30] Randall J LeVeque and David L George. High-resolution finite volume methods for the shallow water equations with bathymetry and dry states. In *Advanced Numerical Models for Simulating Tsunami Waves and Runup*, pages 43–73. World Scientific, 2008.

[31] Randall J. LeVeque, David L. George, and Marsha J. Berger. Tsunami modelling with adaptively refined finite volume methods. *Acta Numerica*, 20:211–289, 2011.

[32] Frank I González, Randall J LeVeque, Paul Chamberlain, Bryant Hirai, Jonathan Varkovitzky, and David L George. Validation of the GeoClaw Model. In *NTHMP MMS Tsunami Inundation Model Validation Workshop. GeoClaw Tsunami Modeling Group*, 2011.

[33] Marsha J Berger, David L George, Randall J LeVeque, and Kyle T Mandli. The GeoClaw software for depth-averaged flows with adaptive refinement. *Advances in Water Resources*, 34(9):1195–1206, 2011.

[34] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[35] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[36] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[37] Gavin P. Hayes, Ginevra L. Moore, Daniel E. Portner, Mike Hearne, Hanna Flamme, Maria Furtney, and Gregory M. Smoczyk. Slab2, a comprehensive subduction zone geometry model. *Science*, 362(6410):58–61, 2018. Slab2 GitHub: `https://github.com/usgs/slab2`, data download: `https://www.sciencebase.gov/catalog/item/5aa1b00ee4b0b1c392e86467`.

[38] A Wichmann. The earthquakes of the Indian archipelago to 1857. *Verhandl. Koninkl. Akad. van Wetenschappen, 2nd sec*, 20(4):1–193, 1918.

[39] Yoshimitsu Okada. Surface deformation due to shear and tensile faults in a half-space. *Bulletin of the Seismological Society of America*, 75(4):1135–1154, 1985.

[40] Yoshimitsu Okada. Internal deformation due to shear and tensile faults in a half-space. *Bulletin of the Seismological Society of America*, 82(2):1018–1040, 1992.

[41] A. Udias, Raúl Madariaga, and Elisa Buforn. *Source Mechanisms of Earthquakes: Theory and Practice*. Cambridge University Press, 01 2013.

[42] Gilda Currenti, Ciro Negro, Luigi Fortuna, and G. Ganci. Integrated inversion of ground deformation and magnetic data at Etna volcano using a genetic algorithm technique. *Annals of Geophysics*, 50, 01 2009.

[43] B Davis and R LeVeque. Adjoint methods for guiding adaptive mesh refinement in tsunami modeling. *Pure & Applied Geophysics*, 173(12), 2016.

[44] Randall J LeVeque. *Finite Volume Methods for Hyperbolic Problems*, volume 31. Cambridge University Press, 2002.

[45] Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Annals of Applied Probability*, 24(6):2455–2490, 12 2014.

[46] Thomas C. Hanks and Hiroo Kanamori. A moment magnitude scale. *Journal of Geophysical Research: Solid Earth*, 84(B5):2348–2350, 1979.

[47] William Bakun. *USGS Earthquake Magnitude Working Group*, 2002 (accessed July 19, 2020).

[48] Yan Y. Kagan. Seismic moment distribution revisited: I. Statistical results. *Geophysical Journal International*, 148(3):520–541, 03 2002.

[49] Donald L Wells and Kevin J Coppersmith. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America*, 84(4):974–1002, 1994.

[50] Valentí Sallarès and César R. Ranero. Upper-plate rigidity determines depth-varying rupture behaviour of megathrust earthquakes. *Nature*, 576(7785):96–101, 2019.

[51] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 11 1992.

[52] W. R. McCann, S. P. Nishenko, L. R. Sykes, and J. Krause. Seismic gaps and plate tectonics: Seismic potential for major boundaries. *Pure and Applied Geophysics*, 117(6):1082–1147, 1979.

[53] Yan Y. Kagan and David D. Jackson. Seismic gap hypothesis: Ten years after. *Journal of Geophysical Research: Solid Earth*, 96(B13):21419–21431, 1991.