Matching Waveform Envelopes for Earthquake Early Warning

Thesis by

Becky Roh

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2021

(Defended September 29, 2020)

© 2020

Becky Roh
ORCID: 0000-0002-3905-0086

ACKNOWLEDGEMENTS

Above all, I would like to thank God through Jesus Christ for His abundant grace that empowered me to work harder in the good times and to not lose joy in the difficult times. Pursuing a PhD has truly been a humbling experience where I faced countless challenges, and with each challenge, I received encouragement and support through many people in my life. No amount of words can possibly express my gratitude to all the people who made my thesis possible, but I will do my best to extend my appreciation in this acknowledgements section.

I would like to thank my advisor, Tom Heaton, for broadening my knowledge in engineering seismology and boosting my confidence. His constant enthusiasm for improving existing methods overflowed to me, which led me to seek, with excitement, different approaches in solving problems. Also, I thank Tom for keeping me on track. Every time I could not see the forest for the trees (in other words, be overwhelmed by details to the point it obscures the real importance), he would remind me to step back and see the bigger picture. I would also like to thank my thesis advisory committee: Domniki Asimaki, John Hall, Richard Allen, Sarah Minson and Zach Ross.

I would also like to thank the earthquake early warning (EEW) community for the helpful feedback on my research progress: Lucy Yin for explaining the concepts of earthquake early warning during my initial stages of research in the field, Men-Andrin Meier and Zach Ross for graciously sharing their impressive data collection, Jen Andrews for explaining in great detail how to improve my algorithm, Egill Hauksson for checking in on my research progress, Angie Chung for her quick, informative responses to my countless questions on the current ShakeAlert system, and Debi Kilb, Annemarie Baltay, and Sarah Minson for the enjoyable discussions we had at academic conferences. I would also like to thank those who did not let distance and time difference affect the exchange of advice; I thank Maren Bose and Masumi Yamada for their valuable advice on how to improve my algorithm for application to complicated earthquake sequences. I am aware that without funding capabilities, I would not have been able to pursue research in this field. Therefore, I gratefully acknowledge the Gordon and Betty Moore Foundation for the funding sources so that I could not only do research, but

also take classes in other disciplines and travel to places I have never been to – Seattle, Denver, Japan, etc. – to share my research with fellow scientists at academic conferences.

My bachelor's degree is in civil engineering; therefore my knowledge in earthquakes was initially limited to seismic loadings and wave propagation. I would like to thank the person that introduced me to the urgency of earthquake early warning and to the idea to pursue a PhD at Caltech: Lucy Jones. Through my internship with her at the USGS, I was inspired to pursue research in a field I had little experience in.

Along with academic support came emotional support, especially from my colleagues in the Caltech community. I would like to thank the members of my research group, Lucy Yin, Kenny Buyco, Anthony Massari, and Gokcan Karakus, for reminding me to stay positive in stressful situations. I was also blessed with an amazing group of friends, Tori, Erika, Kavya, and Ying Shi, that took the time to plan trips and activities for us to take a break and recharge. I would also like to thank Carolina Oseguera for making sure I was not stressed, especially with making travel arrangements for my conference trips. And I would like to thank my community at Holliston UMC for taking the time to check in on me and keeping me in their prayers.

Finally, I would like to acknowledge my family, in particular to my mom, dad, and sister, Michelle. They have seen me at my best and my worst, and they have never failed to show me the love and support I needed to keep going. I thank them for their patience and for being participants in all of the milestones in my life. My family has been cheering for me since day one, and without them, I would not have been able to start and complete my PhD.

ABSTRACT

Current earthquake early warning (EEW) algorithms are continuously optimized to strive for fast, accurate source parameter estimates for the rupturing earthquake (i.e. magnitude, location), which are then used to predict ground motions expected at a site. However, they may still struggle with challenging cases, such as offshore events and complex sequences. An envelope-based two-part search algorithm is developed to handle such cases. This algorithm matches different templates to the incoming observed ground motion envelopes to find the optimal earthquake source parameter estimates.

The algorithm consists of two methods. Method I is the standard grid search, and it uses Cua-Heaton ground motion envelopes as its templates; Method II is the extended catalog search, and its templates are waveform envelopes from past real and synthetic earthquakes. The grid search is intended for robustness and provides approximate average solutions, whereas the extended catalog search matches envelopes considering the station's specific site and path effects. In parallel execution, Methods I and II work together – either by confirming each other's solutions or accepting the solution with stronger fits – to provide the best parameter estimates based on waveform-based data.

The main advantage of the two-part search algorithm is its ability to find parameter estimates of reduced uncertainties using the P-wave data from a single station. Many algorithms wait until multiple stations are triggered to reduce tradeoffs between the magnitude and location. This waiting time, however, is detrimental in EEW, for it jeopardizes the warning time that can be issued to nearby regions expected to experience strong shaking. The use of a single station would virtually eliminate this waiting time, maximizing the warning time without the cost in accuracy of the estimates.

Because EEW is a race against time, further actions are taken for more rapid estimation of the earthquake source parameters. A Bayesian approach using prior information has the potential to reduce uncertainties that arise in the initial time points due to tradeoffs between the magnitude and location. This essentially increases the confidence of the initial parameter estimates, allowing alerts to be issued faster. A KD tree nearest neighbor search is also

introduced to reduce latency in the time it takes to find the best-fitting solutions. In comparison to an exhaustive, brute-force search, it cuts the searching time by only examining through a fraction of the total database.

An envelope-based algorithm examines the shape and relative frequency content and makes appropriate judgments, just as a human seismologist would; it also addresses the issue of data transmission latencies. Overall, this algorithm is able to interpret the complexity of earthquakes and assess the features they hold to ultimately communicate information of significant ground shaking to different regions.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 General Concept of Earthquake Early Warning (EEW)

Earthquakes endanger lives and properties for urban areas near major active faults on land and subduction zones offshore. Seismic history strongly indicates that California is well acquainted with earthquakes. It would be a great advantage to give communities an advance, confident warning before the damaging shaking arrives. An advance warning about a potentially damaging earthquake could reduce injuries, destruction of properties, and increase effectiveness of emergency response. Such alerts could help control elevators, issue go-around commands to aircrafts, save data, secure equipment in surgeries, stop trains, and give instructions to factories, construction sites, schools, hospitals, and shopping centers. It is not possible to predict future earthquakes, however, seismic waves can be detected after the earthquake ruptures. Of course, if earthquakes ruptured much more slowly, then the judgments made by *human* seismologists would suffice in broadcasting the information to the public. However, earthquakes occur much more quickly in reality, and the judgments by human seismologists would essentially provide no warning time. Fortunately, current technologies are automated, providing rapid detection of the seismic waves and identification of the earthquake source parameter estimates. Doing so, alerts can be sent to regions expected to experience strong ground shaking with maximized warning times. These warning times can range between a few seconds to minutes, depending on the user's distance to the epicenter.

1.2 EEW in the World

EEW is not a recent concept. The very first published plan for an EEW system is the most basic version in which it does not provide any warning time. J.D. Cooper suggested in the San Francisco Daily Chronicle in 1868 that a bell be rung when ground shaking exceeded a certain threshold. This idea was never implemented. Today, new technologies provide data in great speeds that allow warning time to be maximized as much as possible. Specifically, damaging S-waves from earthquakes travel at about 3.5 km/s, whereas the less damaging P-waves travel 5-8 km/s and data can travel from stations to processing centers at speeds up to

300,000 km/s without interference. Therefore, before strong shaking arrives, P-wave data can be processed and alerts can be issued, providing users warning times ranging from a few seconds to even minutes.

This valuable concept of EEW was realized and implemented by countries devastated by large earthquakes, such as Japan and Mexico. Other countries also recognized the value of EEW, like Italy, China, Switzerland, Turkey, Taiwan, and the West Coast of the United States. Each country uses methods that cater to its needs. Some use time picks with event associators, while some use amplitude-based methods. Some use a single-station approach, while some use a network-based approach of multiple stations. Some require a central processing network, while some already combines onsite processing and wireless communications with the sensor. The available data from triggered stations is used to estimate the earthquake source parameters, but some countries also use information that stations have not yet triggered to locate the earthquake. Hybrids of these different methods are also used.

The current EEW system for the West Coast of the United States is called ShakeAlert. This system produces both point source and line source solutions. The Earthquake Point-Source Integrated Code, or EPIC, is the algorithm that determines the parameter estimates for a point source. EPIC is a modified version of the Earthquake Alarm Systems, or ElarmS (Allen 2007). ElarmS is a network-based approach that uses picks; it requires at least four triggered stations before issuing an alert. It is currently the fastest and most accurate of the ShakeAlert algorithms, making it the basis for declaring alerts. The Finite-Fault Rupture Detector, or FinDer, provides the line source solutions (Bose 2012). The line source assumption of the rupture is especially valuable for identifying larger earthquakes ($M>6.5$), events of longer duration and longer fault lengths. Rather than depending on picks, FinDer matches spatial pattern of ground motion amplitudes. Because of the limited template sets, FinDer alone cannot generate alerts in the overall ShakeAlert system.

Originally, the ShakeAlert system comprised of three point-source algorithms: ElarmS, Onsite, and Virtual Seismologist. The previously mentioned EPIC is the resulting algorithm from modifying ElarmS and merging it with Onsite. Onsite is a single-station approach to EEW (Kanamori 2005). This algorithm may be faster than methods of a network-based approach because it reduces the waiting time for stations to trigger. However,

it may be less reliable. To address this issue of reliability, Onsite requires 3 seconds of data. The Virtual Seismologist is a Bayesian probabilistic approach that uses both waveform envelopes and prior information (Cua 2005). The multiple algorithms yield earthquake source parameters, which are then combined in the Solution Aggregator (SA) algorithm. The eqInfo2GM algorithm takes the information from the SA algorithm to predict ground motions. The final step is for the Decision Module (DM) algorithm to check if thresholds are exceeded, determining whether to issue alerts to users.

1.3 Statement of Problem

The concept of source-based EEW method can be presented in two questions:

1. Given available data, what are the most probable magnitude and location estimates?
2. Given the most probable magnitude and location estimates, what are the expected ground motions in specified regions?

Algorithms do exist in which ground motions are predicted directly from the available data, skipping the source parameter estimation (i.e. Japanese method by propagation of local undamped motion, or PLUM). This thesis assumes a source-based method and focuses on addressing the very first question, just as the previously mentioned ElarmS and FinDer do. Predicting expected ground shaking from the source parameter estimates is beyond the scope of this thesis, but generally speaking, ground motion prediction equations, or GMPEs, are commonly used to accomplish this conversion.

In developing the problem, a variety of EEW systems from different parts of the world have been assessed. One of the challenges the current EEW system faces is the detection and identification of offshore events and complicated earthquakes. Between 2014 and 2016, E2 missed 213 $M \geq 3$ earthquakes, in which the majority of them were offshore or in areas without dense station coverage (Chung et al. 2019). Locating offshore events is infamously known to be difficult due to poor azimuthal seismic ray-path coverage and sparse station spacing about the epicenter (Chung et al. 2019). The misidentification of complicated earthquakes is most likely due to invalid point-source characterization of the earthquake rupture by one of the two independent algorithms of the current system, ElarmS.

In this thesis, an envelope-based search algorithm is designed to address these challenging types of earthquakes. It consists of two methods that run in parallel (see Fig.

1.1). Originally, the algorithm was envisioned as a real-time implementation based on the previously mentioned Virtual Seismologist. Therefore, Method I of the algorithm is a standard grid search that matches Cua-Heaton ground motion envelopes to the incoming ground motion envelopes, using probabilistic measures. Additions were made to the original idea to enhance accuracy and rapidity of the solutions. Thus, Method II is an extended catalog search that matches envelopes of both real and synthetic past earthquakes. The solution is a magnitude and location, corresponding to the matched envelope, that best describe the incoming envelopes.

1.4 Objectives of Thesis

The main objective of this thesis is to develop an EEW algorithm intended for real-time implementation that has the ability to accurately describe the incoming earthquake with only the P-wave data from fewer than three stations. By accurately characterizing the earthquake with limited data, this search algorithm has the potential to detect and identify challenging events, especially those in regions of sparse station coverage, those offshore, and those spaced close together in time in complicated sequences.

This thesis is organized in nine chapters. Chapter 1 introduces the research problem and provides the general roadmap of the search algorithm developed in this thesis. Chapter 2 discusses the basic processing methodology applied to the waveform data. Chapter 3 describes Method I of the full search algorithm, which is the standard grid search using the Cua-Heaton ground motion envelopes (Cua 2005). It includes a test sweep of M>4.5 events in in Southern California to emphasize its intension, which is the robustness in identifying critical earthquakes. Chapter 4 describes Method II of the full search algorithm, which is the extended catalog search using envelopes from past real and synthetic earthquakes. It defines and validates the use of a spectral scaling model to extend the earthquake catalog, extending it to ensure sufficient coverage of earthquakes. Alongside a step-by-step description of the methods, Chapters 3 and 4 also include real application to past M>5 recent earthquakes, such as the 2020 Northern coast offshore event, 2020 Lone Pine sequence, and 2019 Ridgecrest sequence. Chapter 5 describes a way to optimize Method II, with respect to search time, with KD trees. Chapter 6 presents special cases where Method II can be modified for accurate parameter estimates. These special cases include the 2016 Kumamoto

sequence, 2010 El Mayor-Cucapah mainshock, and the recent 2019 Ridgecrest sequence. Chapter 7 proposes how the two independent methods can be combined into a single working algorithm. Chapters 3-7 describe how the methods find waveform-based solutions, whereas Chapter 8 introduces prior information that can be used to reduce uncertainties in the initial parameter estimates. Finally, Chapter 9 provides concluding remarks and future work.

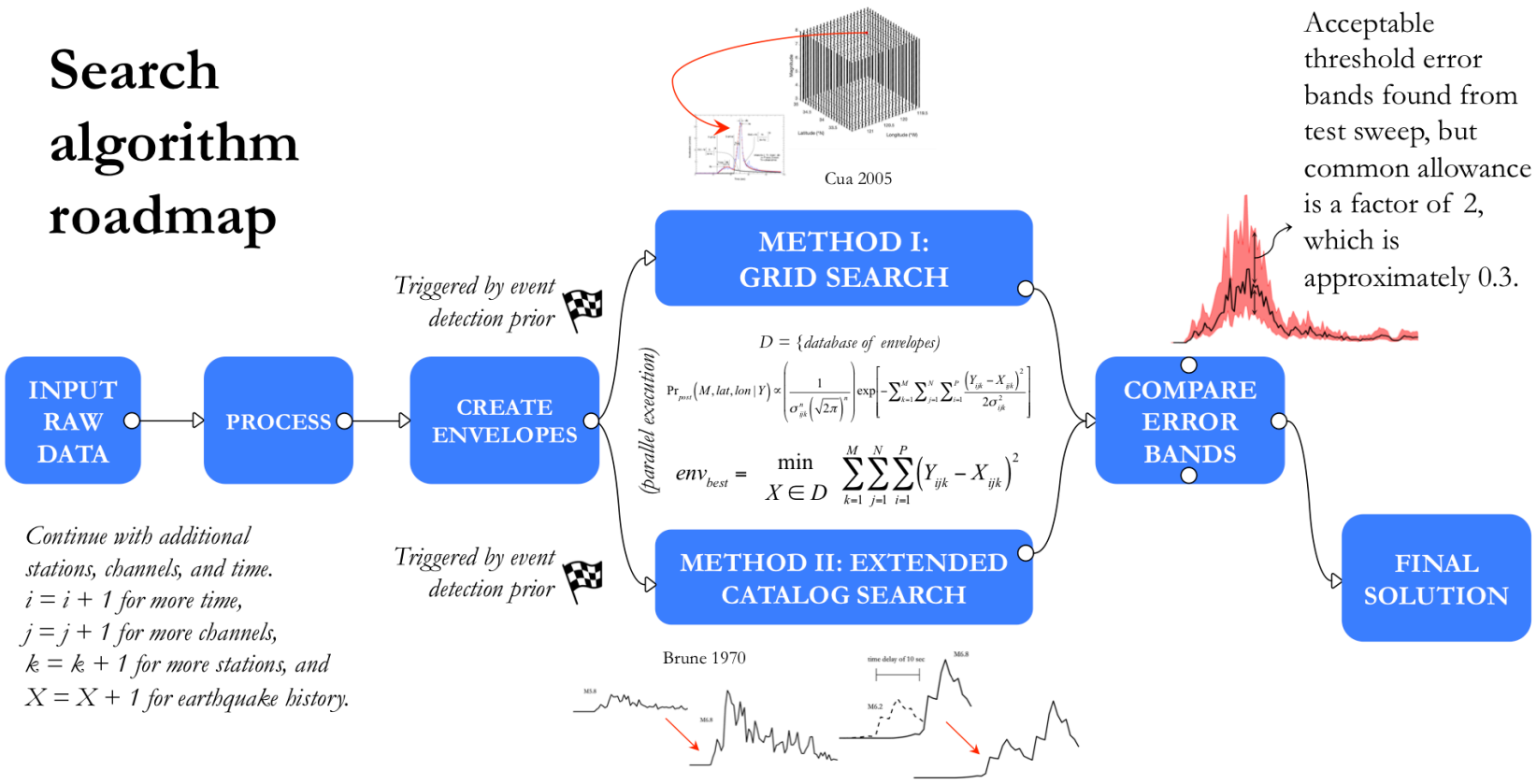


Figure 1.1. Search algorithm roadmap. This roadmap illustrates the parallel execution of the full envelope-based two-part search algorithm presented in this thesis: Method I, the grid search based on Cua-Heaton ground motion envelopes, and Method II, the extended catalog search of past real and synthetic earthquakes.

2 Data collecting and processing

The search algorithm is intentionally designed to detect and identify critically moderate to large earthquakes, particularly those that the current earthquake early warning (EEW) system finds challenging. Therefore, the records that are downloaded from different data centers reflect this goal.

2.1 Raw Data Collection

Strong motion datasets of earthquakes listed in Table 2.1 are downloaded from different data centers: the Northern California Earthquake Data Center (NCEDC), Southern California Seismic Network (SCSN), Kyoshin Network (K-NET), and California Strong Motion Instrumentation Program (CSMIP). The raw acceleration waveforms are downloaded 100 to 200 samples per second for three components: EW, NS, and UD. In this thesis, only the first three to seven seismic stations closest to the observed epicenter are considered in the computations. In reality, as time goes on after the earthquake ruptures, more stations are triggered. However, the computations in this thesis focus on the initial estimates for EEW-relevant purposes. The triggers are assumed to be real-time from P-wave arrivals. In other words, it is assumed that there is no latency in data retrieval.

For smaller earthquakes, broadband channels are commonly used to ensure small signals are visible. As seen in Table 2.1, the earthquakes chosen for this study are of moderate to large sizes. Records from broadband channels when ground motion amplitudes grow large may result in clipping issues. Clipping would result in unrealistic ground motions, which may impact the search algorithm's solutions. This is why only strong motion datasets are downloaded, and broadband records are not considered. Furthermore, the high dynamic range of current technologies, which are 24-bit digitizers, allows small amplitudes from even smaller earthquakes to be visible on the strong motion records. Once the raw acceleration records are collected, they are demeaned before further processing.

Table 2.1. Earthquake dataset used for analyses in this thesis.

Earthquake	Mag	# of records		Data center
		Used for observed	*Used to build catalog**	
2020 Northern coast offshore	M5.80	9	259	NCEDC
2020 Lone Pine	M5.80	12	12	SCSN
2019 Ridgecrest	M7.10	21	2583	SCSN
2016 Kumamoto	M7.00	30	60	K-NET
2012 Brawley	M5.41	24	360	SCSN
2010 El Mayor-Cucapah	M7.20	6	42	CSMIP
2010 – 2020 Southern California**	M>4.50	675	--	SCSN

* Before extension; before applying spectral scaling law to create synthetic earthquakes.
** 75 earthquake events used in mini test sweep for the grid search in Chapter 3.

2.2 Processing Methodology

Though only raw accelerations are initially downloaded, it is still important to fully represent the available frequency information of the incoming signals. Therefore, the raw acceleration records are processed and integrated to obtain the velocity and filtered displacement records as well. Doing so, it has the potential to reduce high uncertainties in distinguishing ground motions of a small earthquake from a larger one. The full use of the frequency information of the incoming signals is intended to reduce high uncertainties, especially for the more rapid single-station approach (Meier et al. 2015). The search algorithm's use of acceleration, velocity, and filtered displacement characterize high (3-10 Hz), medium (0.5-3 Hz), and low (<0.5 Hz) frequencies, respectively.

Raw acceleration waveforms are easily downloaded. To acquire velocity and filtered displacement waveforms, the raw acceleration waveforms are properly processed. They are filtered with a causal fourth-order Butterworth high-pass filter at a corner frequency of 0.075 Hz (Yamada et al. 2007). Then, a single integration provides the velocity, and a double integration provides the displacement. Another filter at a corner period of 3 seconds is applied to the displacement to reduce the influence of noisy microseisms on the small amplitude displacements and to remove long-period noise introduced by the initial processing of the strong motion data (Cua 2005). Ideally, the lower frequency energy would not be filtered out of the displacement record, especially because this frequency range helps discriminate between small and large earthquakes. However, Cua-Heaton envelopes are

created using this high-pass filtering. Therefore, for consistency in matching, this high-pass filtering is also used in this thesis. Throughout the thesis, the term "filtered" is attached to the term "displacement" to accurately represent this processing methodology. See Fig. 2.1 for a visualization of this processing methodology.

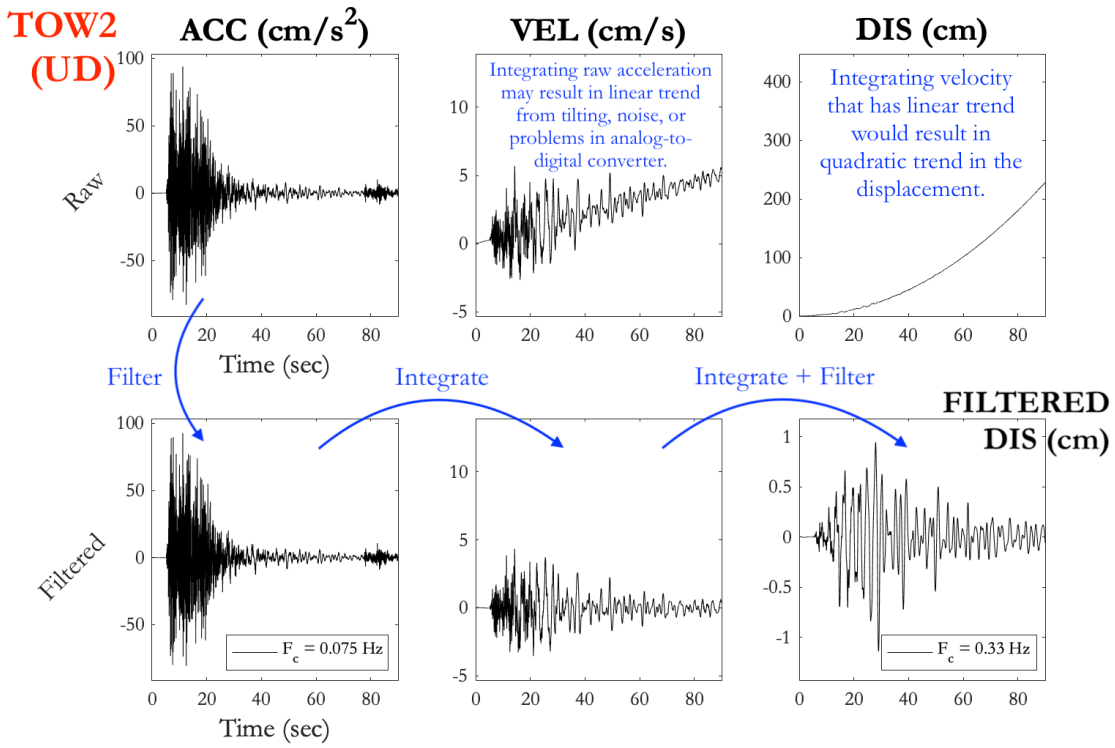


Figure 2.1. Processing waveforms from Station TOW2 in the UD direction. The raw acceleration is downloaded from HNZ channel. If this is integrated without any filters, linear and quadratic trends occur in the velocity and displacement (top row). However, a causal fourth-order Butterworth high-pass filter removes these trends in the velocity and filtered displacement (bottom row). For real-time implementation, the filters would be applied recursively in the time domain. Therefore, the processing time would be deemed negligible.

For simplicity, this processing methodology is uniformly applied to all the waveforms listed in Table 2.1. However, this processing methodology should be taken with caution, particularly for near-source records of large earthquakes. The removal of long period components may not be noticeable in the high-frequency acceleration, but it is obvious in the displacement. Specifically, for near-source records of large earthquakes, the displacement exhibits large static offsets. Because the search algorithm is consistent in matching filtered displacements, this matter can be regarded negligible. In cases like

nonlinear building response analysis, the removal of long period components may have a stronger negative impact.

2.3 Phase Determination for Offline Analyses

A triggered station indicates the start of the computations, and the trigger is based on the P-wave arrival. Therefore, another course of action using the data is to distinguish the arrival of the different phases. For this, a polarization analysis is used (Ross et al. 2014). A phase filter (Eqs. C), determined from the covariance matrix (Eq. 2.1), is multiplied to the three-component data to separate the P and S phases. The three-component data from Station SRT is shown in Fig. 2.2 to illustrate the application of the polarization analysis.

$$\sigma = \begin{bmatrix} Cov(N,N) & Cov(N,E) & Cov(N,Z) \\ Cov(E,N) & Cov(E,E) & Cov(E,Z) \\ Cov(Z,N) & Cov(Z,E) & Cov(Z,Z) \end{bmatrix} \tag{2.1}$$

where σ is the covariance matrix, and N, E, and Z refer to the data from the NS, EW, and UD components, respectively.

$$r = 1 - \left(\frac{\lambda_2+\lambda_3}{2\lambda_1}\right), 0 \leq r \leq 1 \tag{2.2}$$

where r is the rectilinearity, or degree of linear polarization, and λ_1, λ_2, and λ_3 are the eigenvalues corresponding to the covariance matrix, σ, of the data.

$$p = r\cos\varphi = r\,u_{11} \tag{2.3.1}$$

$$s = r(1-\cos\varphi) = r(1-u_{11}) \tag{2.3.2}$$

where p and s are the polarization filters, and u_{11} is the first components of the eigenvector, $\boldsymbol{u} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3})$, corresponding to the eigenvalues, $\lambda = (\lambda_1, \lambda_2, \lambda_3)$.

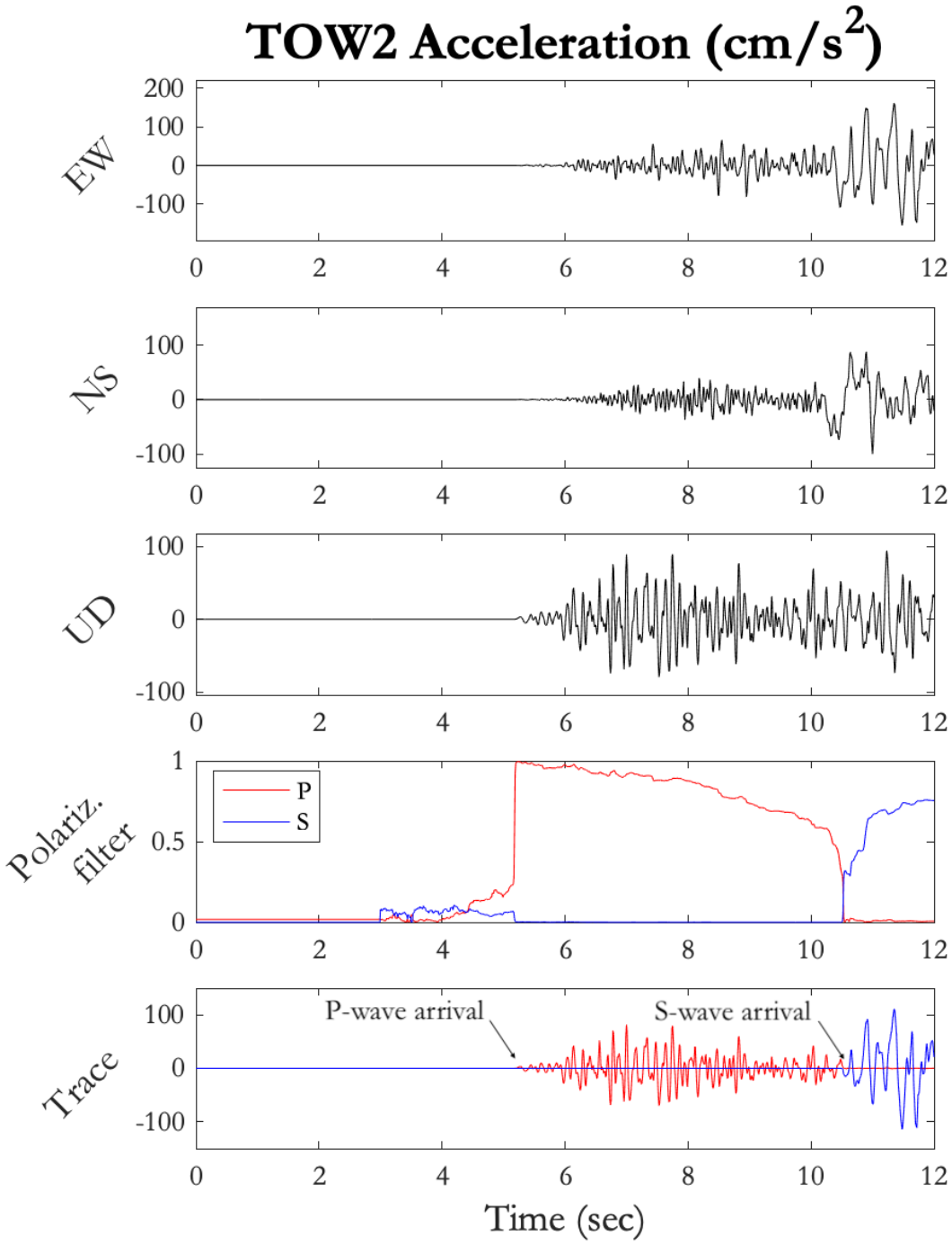


Figure 2.2. Application of polarization analysis to find P- and S-wave arrivals. The P-wave arrival indicates the start of the search algorithm.

2.4 Initiation of Algorithm by Prior for Real-Time Analyses

The analyses done using the earthquake records in Table 2.1 are offline. However, to initiate the search algorithm in real-time, an event detection prior is applied (see Chapter 8). This prior allows a fast detection of the incoming earthquake. It is a probabilistic approach in distinguishing the signal as either noise or an event.

2.5 Converting Full Waveforms to Envelopes

Finally, the input data of the algorithm is in the form of ground motion envelopes, instead of the full raw waveforms downloaded at 100 to 200 samples per second. An envelope of a signal is essentially a trace of the signal's absolute peaks, where the peaks are found using a sliding window (see Fig. 2.3). In this thesis, the size of the window is 1 second but, for other applications, can be modified by the user. Therefore, combining the acceleration, velocity, and filtered displacement envelopes in the three components, there are 9 samples per second retrieved from each triggered station. For comparison, the current system uses the full raw waveforms, with sample rates that vary from 100 to 200 Hz. Taking the acceleration and velocity in the three components, this means the number of data points acquired per second can grow as large as 1200.

The search algorithm depends on the rapid acquirement of real-time waveform data. Therefore, any large data latencies would wreak havoc in the system. One real example where data latencies increased is the 2019 Ridgecrest sequence. Near-source stations, CCC, LRL, WVP2, and CLC, experienced packet delays, most likely due to inefficient data compression and limited bandwidth (Chung et al. 2020). The efficiency of data compression fell sharply due to the large amplitudes of long duration. In fact, data at CLC was delayed by more than one minute. As seen in Chapters 3 and 4, these four stations are amongst the ones considered in the search algorithm. This means the search algorithm would have to wait for the delayed data, which would adversely increase the time it takes to find the parameter estimates. To clearly understand how envelopes may help reduce these detrimental data latencies, it is helpful to visualize how bandwidth and amount of data affect data transmission. Bandwidth refers to the volume of data that a network can transfer within a given time. The higher the bandwidth, the faster the data transfer. However, if the

bandwidth is fixed, the speed of the data transfer is dependent on the amount of data. For instance, large amounts of data, like 1200 samples, would take longer than smaller amounts, like 9 samples.

Decreasing the sample rate from 100 Hz to 1 Hz per channel would lessen the stress in the data collection process without removing valuable information. The waveform envelopes are sufficient for the algorithm to be able to make judgments on the incoming ground motions. It would observe the shape and relative frequency content and classify it with a certain magnitude and location, just as a human seismologist would. This is the main idea behind the Virtual Seismologist. Because the hope is that the use of envelopes will lower latencies, throughout this thesis, a zero latency of the data packets is assumed.

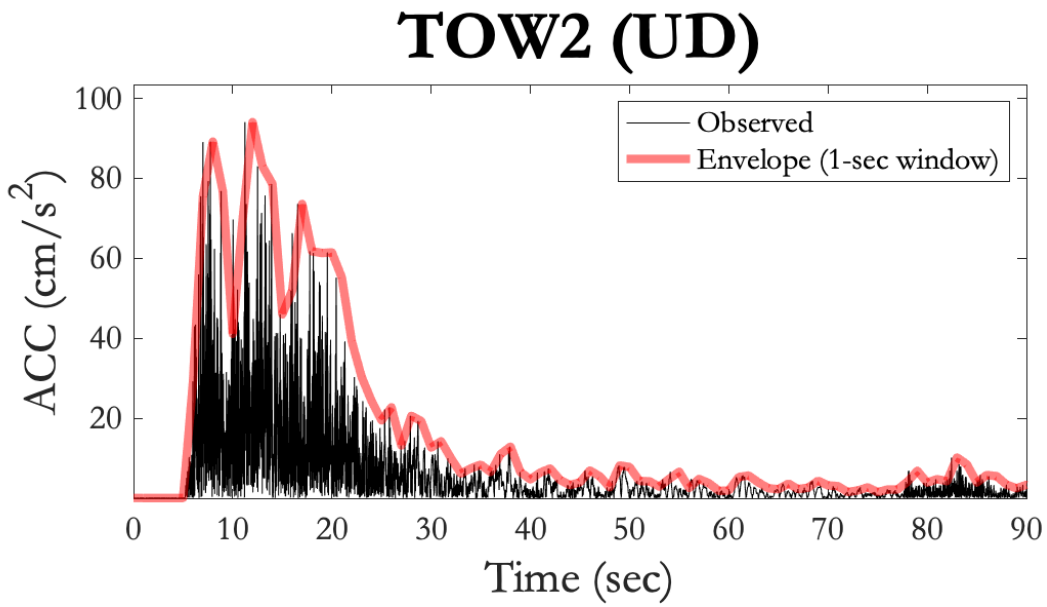


Figure 2.3. Envelope created using waveform data at Station TOW2 in UD direction. Envelope (in red) is converted from taking the maximum absolute amplitudes of raw full waveform data (in black) in 1-second windows. In other words, for every 100 samples of the raw data, the absolute peak is taken.

2.6 Summary

This chapter covers the initial components of the roadmap shown in Fig. 1.1 (shown in Chapter 1). More specifically, the chapter describes how the input data is processed and how ground motion envelopes are created. These ground motion envelopes drive the two-part envelope-based search algorithm. Therefore, it is critical for the initial components to not be prone to delays in data retrieval or prone to any processing errors.

3 Method I: Grid search

As mentioned in the previous chapter, this thesis presents a two-part search algorithm that provides earthquake source parameter estimates. The intent of earthquake early warning (EEW) is to provide solutions as quickly as possible without compromising accuracy. Therefore, the two parts of the algorithm run in parallel to enhance speed and accuracy, as they either provide confirmation to one another or provide replacements to the other. Method I of the two-part search algorithm is a standard grid search using the Cua-Heaton ground motion envelopes. The ultimate goal of the grid search is to achieve robustness in the earthquake source parameter estimates that describe the incoming observed ground motions. Therefore, a mini test sweep is done on recent 4.5<M<7 earthquakes in Southern California to show how well the grid search performs with various station densities and different waveform data. Additionally, specific critical earthquakes that are frequently missed or misidentified by the current EEW system are also included in this study. In this chapter, the grid search considers uniform prior information, meaning every magnitude and location is assumed to occur equally likely. With a uniform prior, the posterior probability is essentially the waveform-based likelihood. To see how prior information affects the grid search solutions, in terms of speed and uncertainty, refer to Chapter 8. Overall, the grid search is able to estimate the magnitude with less than 3 seconds of P-wave data from one station and the location with three stations. While this chapter only shows the performance of the grid search, Chapter 7 compares it to the performance of Method II of the two-part algorithm. The comparisons of the error bands resulting from the two different methods infers that the extended catalog search generally finds a better fit to the incoming ground motions, especially because it considers the specific conditions of the station and channel. However, the grid search does provide a form of confirmation, which would reduce the uncertainty of the overall results.

3.1 Introduction to the Grid Search Method

A standard grid search is a simple, straightforward approach to exhaustively scan through a set of potential parameters to find the best match that portrays the true observed data. It is the most basic method used in data inversion procedures, which is essentially the type of

problem EEW aims to solve for earthquake source parameters and expected ground shaking. A grid search thoroughly tests out all the various combinations of possible parameters, within the specified constraints, to find the best one that describes the input data. This calculation of the total combinations of the parameters clearly reveals, if existent, multiple optimal solutions. This way, the uniqueness of the optimal solution can be compared with the rest of the parameter space (i.e. when there are multiple best fitting parameters). Furthermore, this ability to search without requiring derivative information allows the grid search to solve nonlinear problems more easily, as it is used to achieve convergence by searching the parameter space by brute-force.

The history of the usage of the grid search method shows that though time-consuming, it almost always brings parameter values sufficiently close to the optimum values (Pederson 1997). Nevertheless, it is important to know the disadvantages of the grid search before applying it to the data. The main disadvantage is that the operation is computationally expensive. To address this disadvantage, the grid space can be modified to be smaller with coarser increments. However, this is dangerous, as grids need to cover sufficient space to ensure the optimal solution is not missed. Fortunately, the computational efficiency can still be saved by another advantageous feature of the grid search: its ability to run independently, or in parallel. Essentially, this feature allows a computationally large problem to be broken into smaller, more manageable ones without impacting timeliness.

Overall, the advantages of the grid search, which includes simplicity, directness, and robustness in obtaining optimal solutions, outweigh the disadvantages, which is mainly its computational inefficiency when grids grow large. In this chapter, a grid search attempts to describe the incoming ground motion envelopes with a magnitude and epicenter (i.e. latitude, longitude) by finding templates that matches the observed. The set of pre-determined templates are ground motion envelopes based on attenuation relationships developed by Cua. Throughout this thesis, they are referred to as Cua-Heaton envelopes. Ultimately, the outputs of the grid search method are the best-fitting earthquake source parameters to describe the incoming ground motions, which then can be transformed to EEW-relevant solutions, such as expected ground shaking levels.

3.2 Creating the Grid Space

3.2.1 Grids

There is no preceding insight in determining the total grid space properly before the earthquake ruptures. Therefore, the total grid space is created based on the information from the first-triggered station. Once it is created, however, it remains the same throughout the remainder of the calculations, which makes parallel execution of the grid search at various stations and channels possible. The total grid space is defined as the total possible combinations of the individual parameters in consideration, which are magnitude, latitude, and longitude. Shifts in time are also embedded within the grid search to account for early and late arrivals of the signal, which may occur due to differences in depth from different wave propagation paths.

The magnitude constraints are M3 to M7 at increments of 0.1. Modifying the increments to larger than 0.1 will make the grid search more computationally efficient but may do so at the cost of accuracy in the parameter estimates. On the other end, making the increments finer to less than 0.1 will only make the grid search more computationally expensive, without improving the optimal solution. Therefore, 0.1 is chosen as the increment, which is sufficient to capture the best-fitting magnitude estimates with an initial error of 0.53, using a single-station approach, which decreases to 0.28 as additional two stations are triggered. The minimum considered magnitude is M3, where the incoming signal can be clearly identified, and the maximum considered magnitude is M7, where point source characterization of the earthquake may not be valid. Because the Cua-Heaton ground motion envelopes are intended for earthquakes that are characterized as point source, the confidence in the results of the grid search for larger earthquakes may be lower. For such cases, the second method of this search algorithm, an extended catalog search, or the existing algorithm that characterizes the rupture as line source, FinDer, are better choices.

As previously mentioned, the first-triggered station provides insight into the determination of the total grid space. In particular, its location initializes the constraints of the latitude and longitude. Assuming the incoming ground motions are nearby this first-triggered station, that is less than 100 km away, the spatial constraint is a 2°x2° square at increments of 0.1°, which approximately maps to a 200x200 km^2 square at increments of 10

km, centered at the first-triggered station's location. This particular increment of 10 km is sufficient to capture the best-fitting location estimates with an overall error of approximately 5 km.

Together, if there are X possible magnitudes, Y latitudes, and Z longitudes, then the total grid space consists of $X \cdot Y \cdot Z$ grid points, where each grid point represents a single combination of a magnitude, latitude, and longitude (see Fig. 3.1).

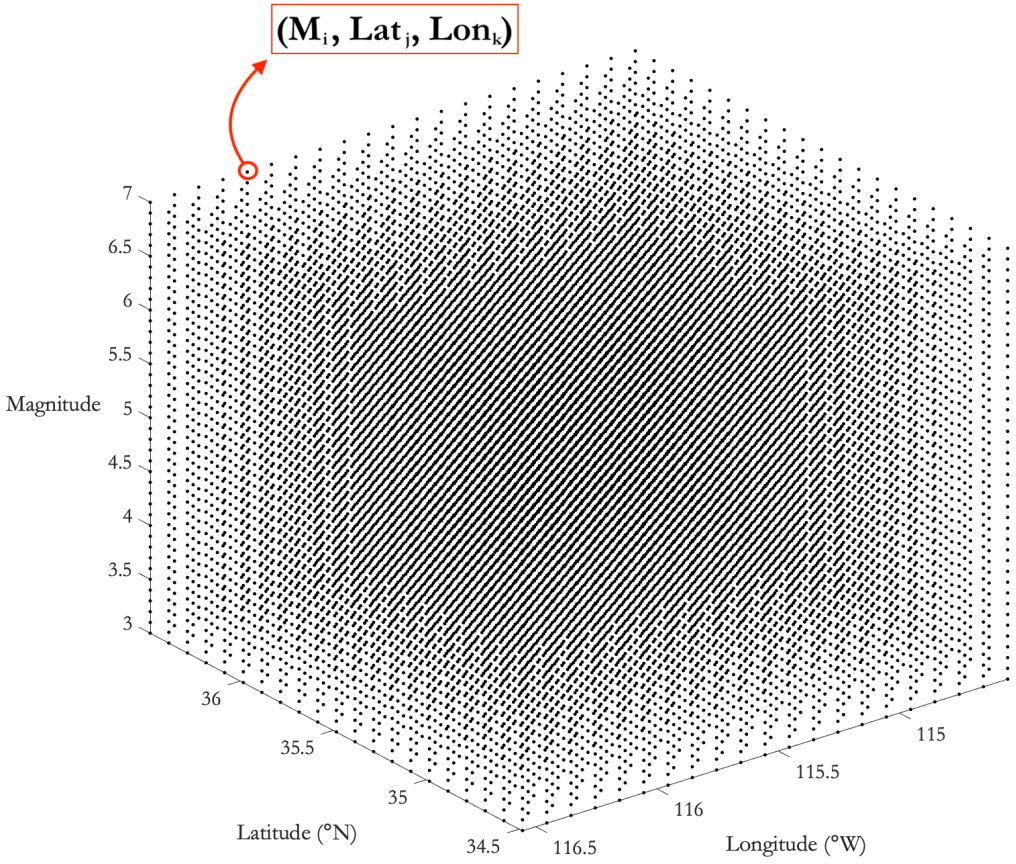


Figure 3.1. Visualization of the total grid space using the following constraints: M3 to M7 for magnitude and (35.48°N, 115.55°W) for the first-triggered station's location. The grid point (in red) refers to a combination of the i^{th} magnitude, j^{th} latitude, and k^{th} longitude.

3.2.2 Templates

Once the total grid space is determined, the templates can be created at each grid point. As seen in Fig. 3.1, each grid point specifies a magnitude, latitude, and longitude. However, the creation of the templates requires a re-parameterization of the latitude and longitude to epicentral distance, for Cua-Heaton envelopes are dependent on magnitude, epicentral distance, and site classification (Cua 2005). The Haversine formula achieves this appropriate re-parameterization; it finds the distance between two points on a sphere.

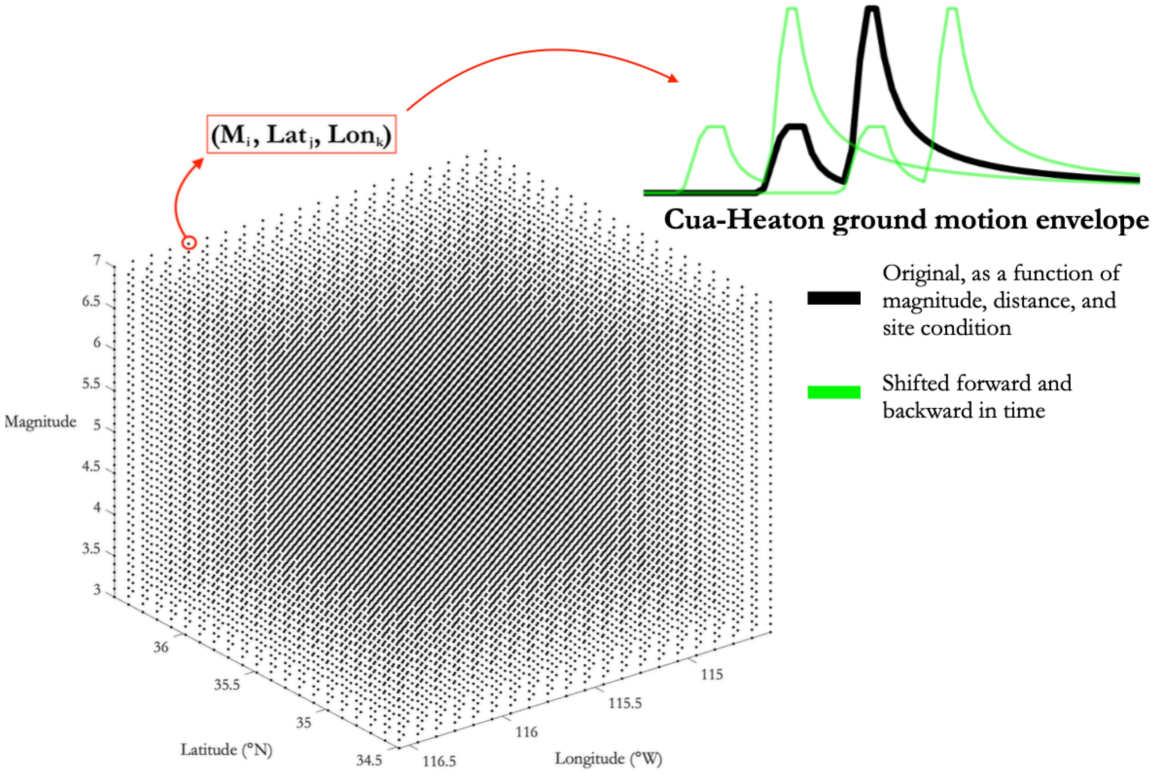


Figure 3.2 Templates created at each grid point of the total grid space. They are the Cua-Heaton ground motion envelopes (Cua 2005).

Before moving on with the specifics of the grid search method, it is important to understand how the templates are made (i.e. assumptions, special features). The Cua-Heaton ground motion envelopes are characterized by a total of only 11 parameters: the rise time, duration, constant amplitude, 2 coda decay (each for P-wave and S-wave component), and ambient noise. The amplitudes are dependent on the magnitude, epicentral distance, and site

classification (i.e. rock, soil). For the evolution of the envelopes with respect to time, a lookup table of travel times is used (Cua 2005). Initially, the envelopes for the P-wave, S-wave, and ambient noise are created individually, and under the assumption of random phase, the finalized envelopes used in the grid search are found by taking the square root of the sum of squared individual envelopes.

The Cua-Heaton ground motion envelopes have key features that make them more preferable than ground motion envelopes based on other attenuation relationships. First and foremost, they consider magnitude saturation dependent on distance. The saturation is most pronounced in the acceleration at close distances to large events. Also, soil and rock sites are treated differently. Soil sites exhibit stronger degree of saturation than rock sites. For the most accurate results, the grid search would generate and use envelopes for soil sites in regions of $V_{S30} < 464$ m/s and would use envelopes for rock sites in regions of $V_{S30} > 464$ m/s. However, for simplicity, this analysis only uses envelopes assuming rock sites.

The two-part search algorithm is envelope-based; both the input data and templates are in the form of envelopes. An envelope of a signal is essentially a trace of the signal's peaks, where the peaks are found using a sliding window (see Fig. 3.3). The user defines the size of the sliding window, but in this thesis, the size is set to 1 second. The algorithm emphasizes the use of envelopes to mimic the analysis human seismologists conduct as closely as possible. Particularly, the use of envelopes allows the algorithm to make judgments on the incoming earthquake by observing the shape and relative frequency content of the waveform data. The 1-second windows of the envelopes also have the potential in reducing latencies in real-time data collection, as high sampling frequencies of 100 to 200 Hz may add additional computational stress.

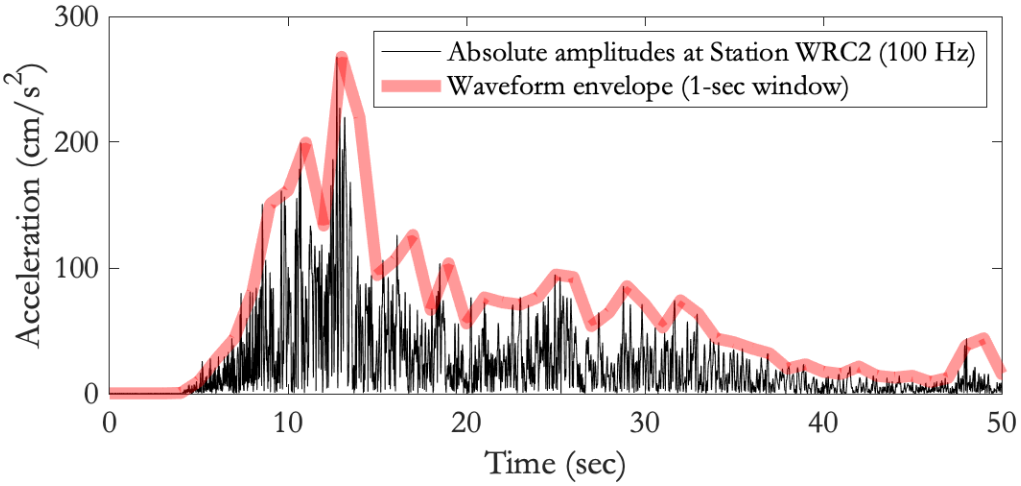


Figure 3.3. Envelope created by taking trace of signal's peaks in 1-second windows. Envelope (in red) converted from taking the maximum absolute amplitudes of raw full waveform data (in black) in 1-second windows.

3.3. Defining the "Goodness-of-Fit"

As previously mentioned, the grid search ultimately aims to find the best-fitting parameters, which are the magnitude, latitude, and longitude. To find these best-fitting parameters, a goodness-of-fit test is used. Generally, a goodness-of-fit test measures how well the observed data corresponds to the fitted (predicted) model by evaluating a particular function. Though the grid search and extended catalog search uses different templates as their respective predicted models, both methods of the two-part algorithm determines the goodness-of-fit score by maximizing the same function (see Eq. 3.1.3), which is the posterior probability assuming a normal distribution of the logarithmic residuals of the data. This is essentially the same as assuming a lognormal distribution of the absolute residuals of the data. Because normality is a strong assumption, it is important to check that it is not violated. That is, normality is only valid when the residuals, the difference of the logarithmic amplitudes of the observed and predicted Cua-Heaton envelopes, are normally distributed about 0.

$$Pr_{post}(M, lat, lon|Y) \propto Pr_{like}(Y|M, lat, lon) = \prod_{k=1}^{M} \prod_{j=1}^{N} \prod_{i=1}^{P} Pr_{like}\left(Y_{ijk}\middle|M, R\right) \quad (3.1.1)$$

$$Pr_{like}\left(Y_{ijk}\middle|M, lat, lon\right) \propto \exp\left[-\left(\frac{\left(Y_{ijk}-X_{ijk}\right)^2}{2{\sigma_{ijk}}^2}\right)\right] \quad (3.1.2)$$

$$Pr_{post}(M, lat, lon|Y) \propto \exp\left[-\sum_{k=1}^{M}\sum_{j=1}^{N}\sum_{i=1}^{P}\left(\frac{\left(Y_{ijk}-X_{ijk}\right)^2}{2{\sigma_{ijk}}^2}\right)\right] \quad (3.1.3)$$

where Y_{ijk} is the logarithmic (base 10) amplitude of the ground motion envelope observed at the j^{th} channel, k^{th} station, and i^{th} time point. X_{ijk} is the logarithmic (base 10) amplitude of the Cua-Heaton ground motion envelope, and σ_{ijk} is the uncertainty of the fit. The total posterior probability is taking the product of the individual probabilities (Eq. 3.1.2) for N channels, M stations, and P time points.

In this chapter, uniform prior is assumed. Therefore, the posterior probability is essentially just the waveform-based likelihood, as shown in Eq. 3.1.1 and Eq. 3.1.3. Independence is assumed to find the total posterior probability, as shown in the multiplication of the individual probabilities in Eq. 3.1.1. Independence is assumed in the following:

- Acceleration, velocity, and displacement amplitudes. For instance, velocity cannot be found based on information on position alone (Cua 2005).
- Time. For instance, amplitude at one time will not determine amplitude at a different time.
- Stations. Amplitudes at stations are causatively independent because the same earthquake causes them, but they can be considered stochastically independent because knowledge at one station does not imply knowledge in another (Cua 2005).

Therefore, for simplicity, independence is assumed. However, it is worthwhile to keep in mind that if large accelerations occur in one station, nearby stations may experience large amplitudes as well.

Working in the natural logarithmic form helps avoid computations with very small values. Therefore, maximizing the posterior probability over a set of Cua-Heaton ground motion envelopes can be transformed to a minimization problem. It is computationally simpler to minimize the sum of squared residuals (SSR), seen in Eq. 3.2, over a set of Cua-Heaton ground motion envelopes, which is simply the negative function within the

exponential in Eq. 3.1.3. This is valid, as logarithmic amplitude residuals follow a normal and independent distribution about a zero mean and constant variance (Cua 2005).

$$SSR(Y|M, lat, lon) = \sum_{k=1}^{M} \sum_{j=1}^{N} \sum_{i=1}^{P} \left(\frac{\left(Y_{ijk} - X_{ijk}\right)^2}{2\sigma_{ijk}{}^2} \right) \tag{3.2}$$

where Y_{ijk} is the logarithmic (base 10) amplitude of the ground motion envelope observed at the j^{th} channel, k^{th} station, and i^{th} time point. X_{ijk} is the logarithmic (base 10) amplitude of the Cua-Heaton ground motion envelope, and σ_{ijk} is the uncertainty of the fit.

3.4 Interpreting the Best Fits with Error Bands

Maximizing the posterior probability, or minimizing the SSR, finds the best-fitting Cua-Heaton ground motion envelope. Once this envelope is found, it is important to also include another value that quantifies how precisely this chosen envelope fits the incoming observed one. In the world of statistics, many terms exist to quantify how precise the predicted model fits the observed data, such as uncertainty and standard deviation. However, in this thesis, the term used to quantify the misfit of the Cua-Heaton envelope with respect to the true incoming envelope is error band. Error bands, defined in Eq. 3.3 and illustrated in Fig. 3.4, enclose the area about the best-fitting envelope in which the true observed data can be found; they portray the volatility around the best-fitting envelope. They vary based on the user-defined confidence band. For example, to satisfy a 95% confidence band, the error bands are adjusted to allow 95% of the true observed data.

$$\log X_{ijk} - \eta \leq \log Y_{ijk} \leq \log X_{ijk} + \eta \tag{3.3}$$

where Y_{ijk} is the logarithmic (base 10) amplitude of the ground motion envelope observed at the j^{th} channel, k^{th} station, and i^{th} time point. X_{ijk} is the logarithmic (base 10) amplitude of the Cua-Heaton ground motion envelope, and η is the error band that is adjusted accordingly to satisfy the confidence band.

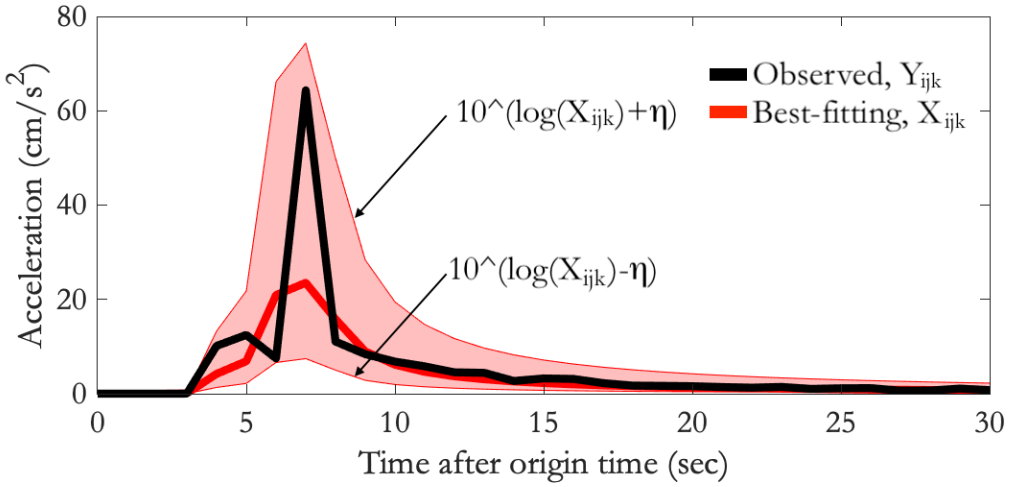


Figure 3.4. Error bands (transparent red shade) that give visual sense of level of confidence. It illustrates how much the chosen envelope (in red) must be scaled to best fit the observed envelope (in black).

3.5 Assessing Convergence: a Test Sweep on 5<M<7 Events

For simplicity, the first test sweep is done on only 5<M<7 events in Southern California. From year 2010 to 2020, 12 5<M<7 earthquakes are recorded in the ANSS catalog. M>7 events are excluded because they may require a finite fault characterization of the earthquake, instead of point source. Therefore, results using M>7 events would negatively implicate the grid search solutions. For practicality of mimicking real-time EEW analysis as closely as possible, only the first three to six triggered stations are considered. In EEW, it is preferable to avoid waiting for more stations because this would severely minimize, or even lose altogether, warning time for regions of expected strong shaking. Therefore, the initial hypothesis of the grid search is that accurate estimates can be achieved with at most three stations.

From the Southern California Seismic Network, the raw acceleration waveform time series from the nearest seismic stations at each event listed in Table 3.1 are downloaded. The raw acceleration includes three components of the ground motion: EW, NS, and UD directions. To avoid clipping issues, especially with 5<M<7 earthquakes, only the strong motion channels are downloaded. The raw acceleration is filtered with a 4^{th} order Butterworth high-pass at a corner frequency of 0.075 Hz (~ 13 s) and integrated, once for velocity and twice for displacement. However, for displacement, it is important to filter with a 3-second Butterworth high-pass to distinguish from microseisms that dictate the lower bound on the magnitude range (M<3.5). Therefore, it is denoted as "filtered displacement".

The acceleration, velocity, and filtered displacement time series are then converted to ground motion envelopes, as described in Fig. 3.3, ready to be used as input data for the grid search. The corresponding parameters to the maximum posterior probability (Eq. 3.1.3), or the minimized SSR (Eq. 3.2), are compared to the "true" parameters recorded in the ANSS catalog (see Table 3.1). To avoid bias in the solutions, this comparison refers to the *absolute* value of the difference between the grid search estimates and the "true" parameters in the ANSS catalog. Throughout this thesis, this comparison is denoted as "absolute error". The parameters in Table 3.1 are assumed to be true with negligible error.

Table 3.1. List of 5<M<7 events in Southern California in the years 2010 to 2020.

Event ID	Origin time	Magnitude	Latitude	Longitude
14607924	2020/04/04,23:25:07.190	M5.38	32.2662	115.2925
39462536	2020/06/04,01:32:11.140	M5.53	35.6148	117.4282
38457687	2019/07/06,03:47:53.420	M5.50	35.9012	117.7495
38443183	2019/07/04,17:33:49.000	M6.40	35.7053	117.5038
37374687	2016/06/10,08:04:38.700	M5.19	33.4315	116.4427
15481673	2014/03/29,04:09:42.170	M5.09	33.9325	117.9158
15200401	2012/08/26,20:57:58.220	M5.41	33.0185	115.5403
15199681	2012/08/26,19:31:23.040	M5.32	33.0172	115.5537
14937372	2011/02/18,17:47:35.770	M5.09	32.047	115.0622
10736069	2010/07/07,23:53:33.480	M5.42	33.4173	116.4747
14745580	2010/06/15,04:26:58.240	M5.71	32.705	115.9113
10589037	2010/04/08,16:44:25.010	M5.29	32.1647	115.2683

Running the grid search on the events listed in Table 3.1 reveals a few interesting observations. Overall, the performance of the grid search depends on the station coverage about the observed epicenter. To assess the performance of the grid search, the convergence of the parameter estimates is observed. Convergence in the parameter estimates refers to a single local maximum in the posterior probability. As seen in Fig. 3.7, the location estimate found using data from the first 2 seconds remains similar, even as additional data is acquired with time. This implies that convergence in the location estimate is not dependent on the amount of data retrieved from a station. Rather, it depends on the number of triggered stations and how they are distributed about the epicenter. As seen in Fig. 3.5, a *uniform* distribution of at least three stations is required for convergence in the location estimate. A uniform distribution of stations has station coverage at different directions about the epicenter, not on only one side of the epicenter. Uniformity is emphasized because it increases the chance of the epicenter being surrounded by stations (i.e. in network). When station coverage is sparse and distribution is non-uniform about the epicenter, the grid search believes multiple locations have similar likelihoods. Such region includes Baja California, where the first triggered stations are above the United States and Mexico border and far away from the observed epicenter (R > 50 km). Seen in Fig. 3.6, the stations look

like they are distributed in a line and to one side of the epicenter, which confuses the grid search in finding a single optimal location estimate. In California, interstation distances between seismic stations vary from region to region. They are less than 5 km in densely populated regions, like San Francisco and Los Angeles, but they are larger than 70 km in northeastern California. This study suggests that uniformity in interstation spacings is suggested for a robust grid search. As previously mentioned, the increment of the latitude and longitude grid space is 10 km. The grid search is able to robustly capture the epicentral locations with this increment. If each grid point represents a potential epicenter, uniform station coverage would mean there are at least three stations about each point, which is an interstation spacing of 10 to 20 km.

At least three stations are required for accurate location estimates. However, for the magnitude estimate, three stations are not required. Instead, a single station is able to find median absolute error of 0.53 units for the magnitude. Waiting for three stations would decrease this error to 0.28 units. Using a single station results in a median alert time of approximately 3 seconds, assuming no latencies. Alert time refers to the time between the origin and when alerts would be sent to users. Waiting for three stations increases the median alert time to 6 seconds. The median location error is approximately 5.25 km, whether the grid search uses P-wave data from a single station or three stations. The estimates are satisfactory as it is within a factor of 2, considering the grid increments are 10 km.

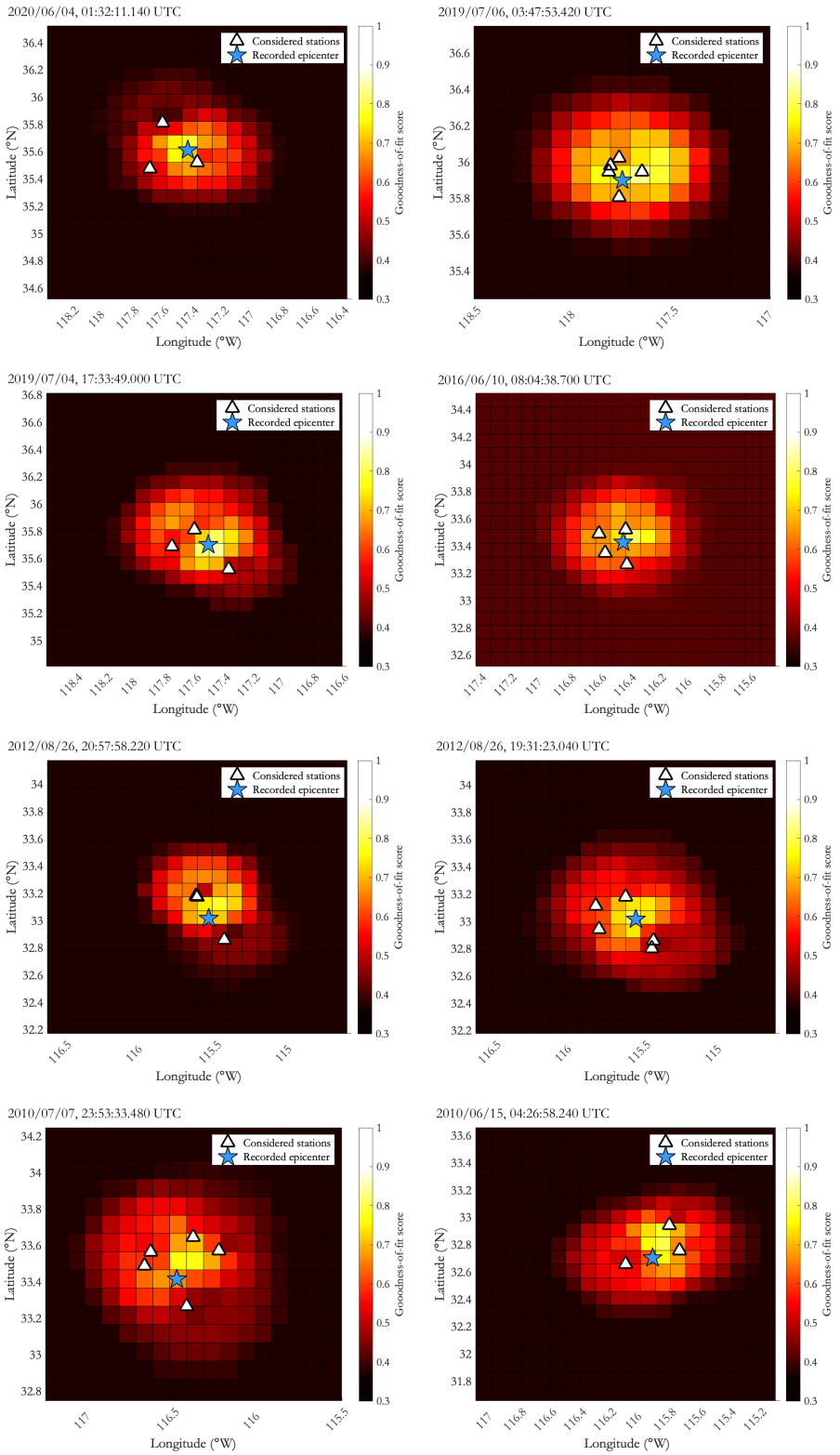


Figure 3.5. Convergence to single optimal location. The pattern seen here is a uniform station distribution (white triangles) about the observed epicenter (blue star).

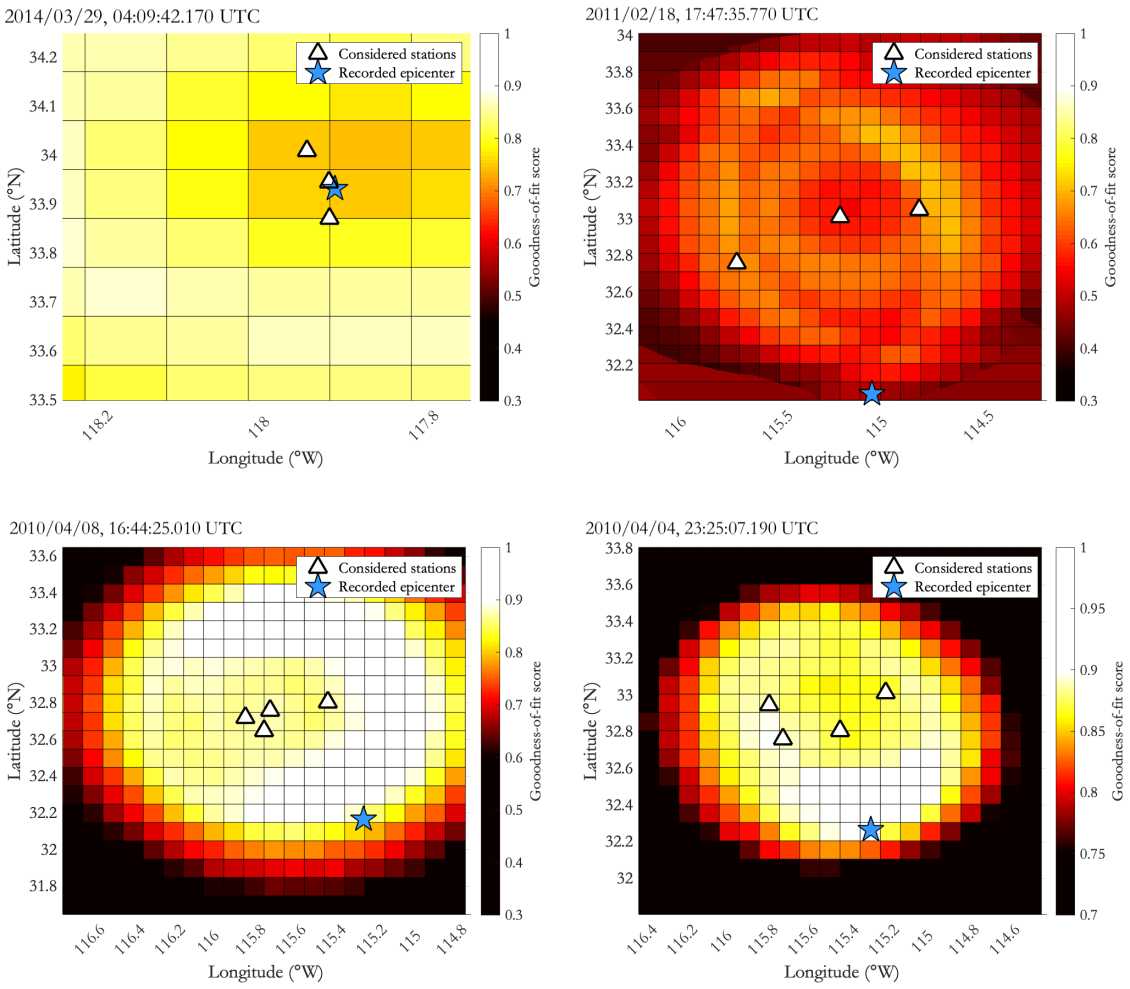
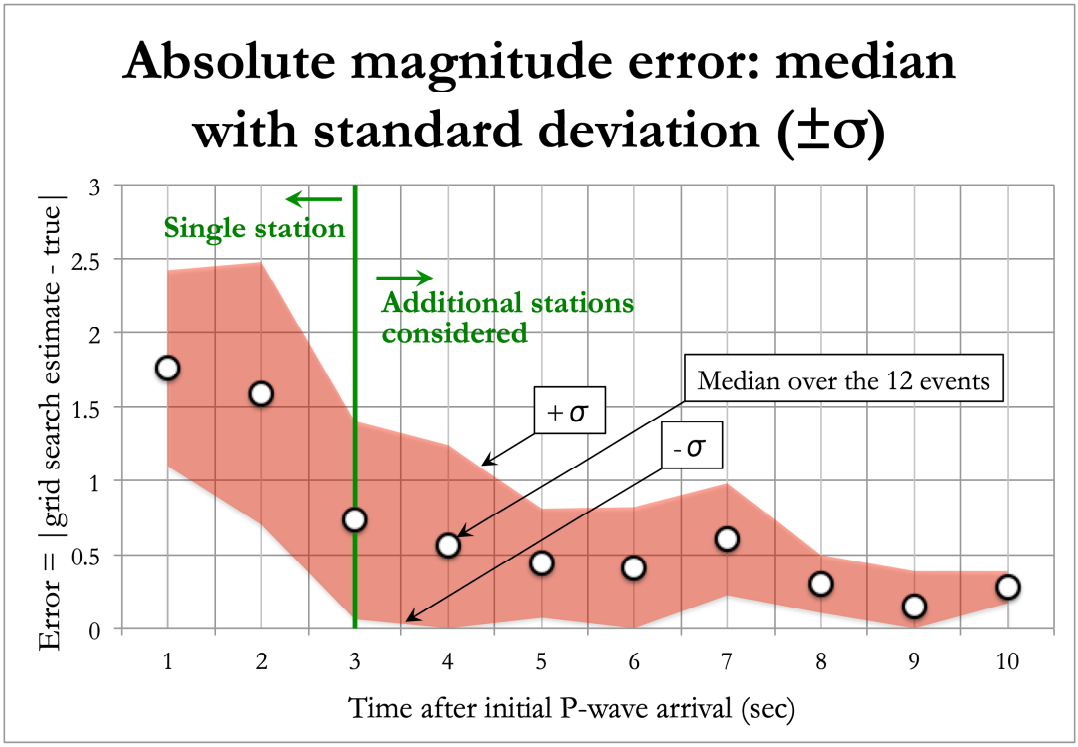


Figure 3.6. Multiple optimal location estimates (multiple local maxima in posterior probabilities). The pattern seen here is the linear station distribution (white triangles) on one side of the observed epicenter (blue star).



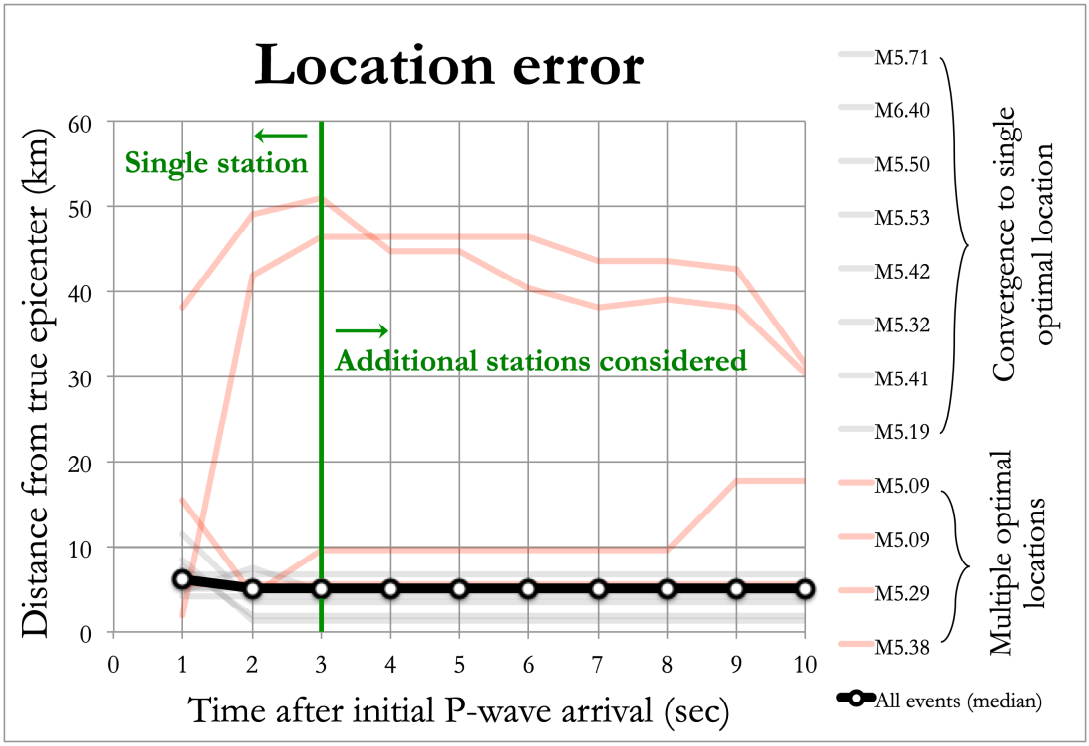


Figure 3.7. Parameter estimates from test sweep of 12 5<M<7 events. Cases consists of events located in regions of uniform station distribution (in gray) and events located out-of-network or in non-uniform station distribution (in red). The median of all 12 events is shown in black. The estimates found in the 3 seconds are from a single station (left of green vertical line). Afterwards, two to three stations are considered in the grid search.

From observing the trends of the median absolute magnitude and location errors with increasing time and triggered stations, a single criterion is not adequate to base a decision on. Instead, two criteria are suggested in making a decision to send an alert to users. The first criterion is to send an alert to regions *near* the estimated epicenter with the initial P-wave data from a single station. Seen in Fig. 3.6, even with multiple optimal locations with similar posterior probabilities, the epicenter is still found within reasonable range of 50 km of error. The alert would tell users within a radius of 50 km about the location estimate to expect strong shaking. A specific quantitative measurement of the parameter estimates is not necessary, especially due to saturation effects in near-source regions for a large (M>6.5) earthquake. In other words, the saturation encloses at approximately 0.5g for both a M6.5 and a M7.5 earthquake at close distances. The second criterion is to wait for three triggered stations to send alerts to regions that are *farther* away from the estimated epicenters, as in those regions, accuracy of the parameter estimates cost more.

3.6 Assessing Robustness: a Test Sweep on 4.5<M<7 Events

It is difficult to assess *robustness* in the parameter estimates with only 12 events. Therefore, the test sweep is expanded to include 75 4.5<M<7 single events in Southern California from 2010 to 2020. Again, raw acceleration waveforms from the three nearest triggered stations are downloaded. They are processed to find the velocity and filtered displacement data. Then, the data is converted to ground motion envelopes as inputs to the grid search. Considering the EW, NS, and UD components for acceleration, velocity, and filtered displacement envelopes, there are 9 envelopes each for 75 events, total of 675 envelopes. The grid search solutions are the best-fitting magnitude and location estimates, and they are updated every 1 second with additional data from more stations and time. Fig. 3.8 shows the different distributions of the absolute errors in the magnitude estimates for a single station, two stations, and three stations. An absolute error of 0 means the estimate equals the true recorded in the historic ANSS catalog. With additional triggered stations, the median and standard deviation of the absolute magnitude error decreases. The distribution becomes more left skewed with additional stations; this implies that more triggered stations result in magnitude estimates that are closer to the true ones recorded in the ANSS catalog.

To observe for robustness in the set of 75 earthquakes, the median and the standard deviation of the absolute errors are calculated (see Table 3.2). Using the P-wave data from a single station, the median and standard deviation of the absolute magnitude error are 0.53 and 0.40 units, respectively. The location error ranges from 0 to 16.76 km. The median is 5.25 km, and the standard deviation is 11.50 km. However, as stated before in the previous test sweep of 5<M<7 events, a single station does not guarantee convergence of the location estimate. With three stations, the median absolute error reduces to 0.28 units for the magnitude. The standard deviation decreases to 0.33 units. The location error remains similar, approximately 5.13 km, but convergence is guaranteed with three triggered stations. With the grid increment being 10 km, the location error is satisfactory as it is within a factor of 2.

For comparison to current EEW algorithms, the solutions for ElarmS are shown in Table 3.2. The most updated ElarmS algorithm has a median magnitude error of 0.3 units and location error of 2.3 km using four stations (Chung et al. 2019). By minimizing the required amount of triggered station from four to one, the grid search gains alert time. However, it costs accuracy of 0.23 units in the magnitude estimate error. Waiting for three stations, still fewer than the amount ElarmS waits for, gains accuracy of 0.02 units in the magnitude error. As previously mentioned, the grid search results in this chapter are based on waveform-based likelihood only. There is, however, potential for faster magnitude estimates with the application of prior information (see Chapter 8).

Table 3.2 (a). Comparison of solutions (median): Grid Search vs. ElarmS.

Using P-wave data from…	Magnitude absolute error	Location error	Alert time*
Grid Search			
One station	0.53	5.25 km	3.0 sec
Three stations	0.28	5.13 km	6.0 sec
ElarmS			
Four stations	0.30	2.3 km	6.7 sec

Table 3.2 (b). Comparison of solutions (standard deviation): Grid Search vs. ElarmS.

Using P-wave data from…	Magnitude absolute error	Location error	Alert time*
Grid Search			
One station	0.40	11.50 km	--
Three stations	0.33	7.40 km	--
ElarmS			
Four stations	0.20	16.7 km	--

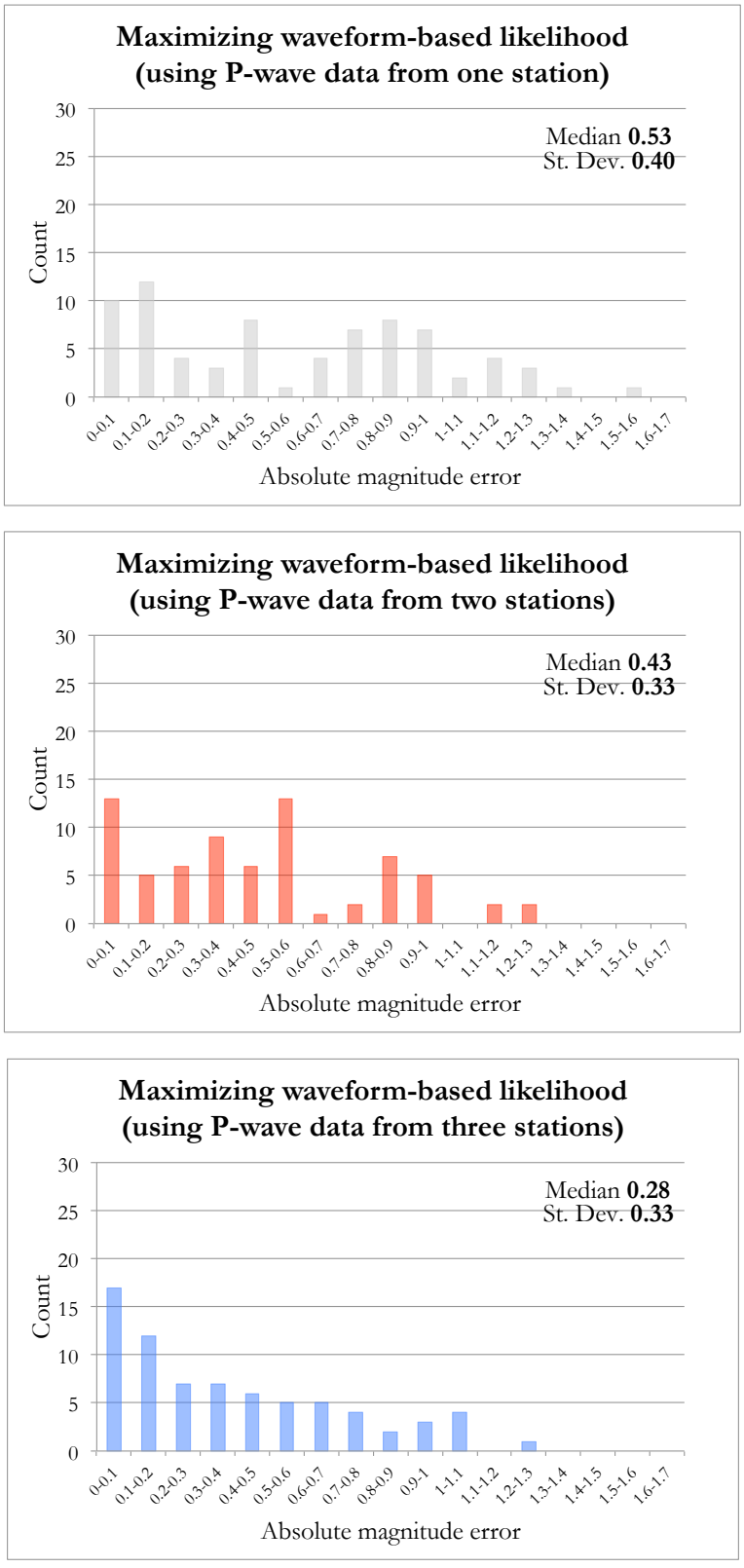


Figure 3.8. Absolute magnitude error using P-wave data from one station (in gray), two stations (in red), and three stations (in blue).

From observing the trends of the median absolute magnitude and location errors with increasing time and triggered stations, a single criterion is not adequate to base a decision on.

2 criteria are suggested for 4.5<M<7.

1) With P-wave data at a **single station**, send alerts to expect strong shaking to regions **near** estimated epicenter without specific quantitative magnitude estimates.
2) Wait for **three stations** to trigger before sending alerts to regions **farther** away, as accuracy costs more than speed of alerts.

3.7 Application to Past Real Earthquakes

The test sweeps portray robustness in the grid search parameter estimates. The following section shows the specific details of the calculations and analysis done in the grid search using recent real earthquakes. The particular types of earthquakes of interest are offshore events and foreshock-mainshock pairs in a sequence. The grid search is applied to the available envelopes from the 2020 Northern coast offshore event, the 2020 Lone Pine event, and the 2019 Ridgecrest mainshock.

3.7.1 2020 Northern coast offshore event

One of the challenges of EEW is detecting and identifying large offshore events. In fact, most recently between 2014 and 2016, E2 missed 213 M≥3 earthquakes, in which the majority of them were offshore or in areas without dense station coverage (Chung et al. 2019). Locating offshore earthquakes is infamously known to be difficult due to poor azimuthal seismic ray-path coverage and sparse station spacing around the epicenter (Chung et al. 2019). Offshore events are frequently missed, but if the system manages to detect them, it still takes a long time to issue the first alert due to the lack of stations between the epicenter and mainland and the requirement of ElarmS to have at least four triggered stations. This is where the previously mentioned advantage of the grid search comes into play, in which it is able to find magnitude estimates of absolute errors 0.28 to 0.53 using one to three stations, respectively. This advantage has the potential to shorten the time it takes to find parameter estimates, which would increase the warning time for nearby regions.

On March 09, 2020 at 02:59:08 UTC, a M5.8 offshore event occurred near Petrolia, CA. This event should not have been a surprise, as earthquake history shows at least ten M>5 earthquakes in the region in the past 20 years. In this type of event, time is of the essence, especially with the first P-wave arriving 14 seconds after the origin time (see Table 3.3). Waiting for more stations to trigger would jeopardize warning time for mainland regions closest to the epicenter, regions that would feel the strong shaking first. Therefore, the stations considered in this analysis are the first three to be triggered: Petrolia Fires Station (89101), Cooskie Peak (KCO), and Cape Town (KCT). As seen in Fig. 3.10, the grid search finds matches to the incoming observed acceleration, velocity, and filtered displacement envelopes. As previously mentioned, error bands are the allowance about the best-fitting envelope for it to represent the incoming ones to a user-defined confidence. To show range, confidence of 68%, 90%, and 95% are chosen, as shown in Fig. 3.9.

To mimic a real-time analysis as the current EEW system, the extended catalog search updates the fits and the corresponding magnitude estimates as data becomes more available with time (see Fig. 3.9). The grid search is able to estimate the magnitude as M>5 within the first 5 seconds after the initial P-wave arrival, which FinDer fails to do. Based on the trend of the error bands in Fig. 3.9, the confidence in the magnitude and location estimates increases with additional data with more time. It ultimately converges to M5.7 with time. The computations involve the first three stations. Therefore, a consistent comparison to real-time Finder is valid up until 20 seconds after the origin time, where the grid search magnitude estimate is M5.1, which is 1.2 units more accurate than FinDer's solution. After 20 seconds, the estimates in Fig. 3.9 are worst-case scenario, as more stations in the grid search computations would reduce the uncertainties. Even so, the grid search estimates the magnitude to be 0.3 units closer to the true magnitude and 5 seconds faster than FinDer's solutions. For the location, the initial grid search estimate is 11 km from the true epicenter. With time, specifically 30 seconds after the origin time, the location estimate converges to 2.87 km from the true epicenter. This estimate is 67 km closer than FinDer's location estimate.

In reality, the current EEW system misidentifies this M5.8 event, which leads to a large location error and failure to send an alert. FinDer simulation takes 35 seconds to lock in at a magnitude estimate, ultimately underestimating it to M5.4. The grid search manages to

avoid misidentifying this event and essentially eliminate the blind zone for onshore regions, as illustrated in Fig. 3.11.

Table 3.3. Triggered stations from the 2020 Northern coast offshore event with P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	**Latitude(°N)**	**Longitude (°W)**	**P-wave arrival (sec after origin time)**
CE.89101	40.3250	124.2877	14
NC.KCO	40.2567	124.2660	14
NC.KCT	40.4756	124.3375	14

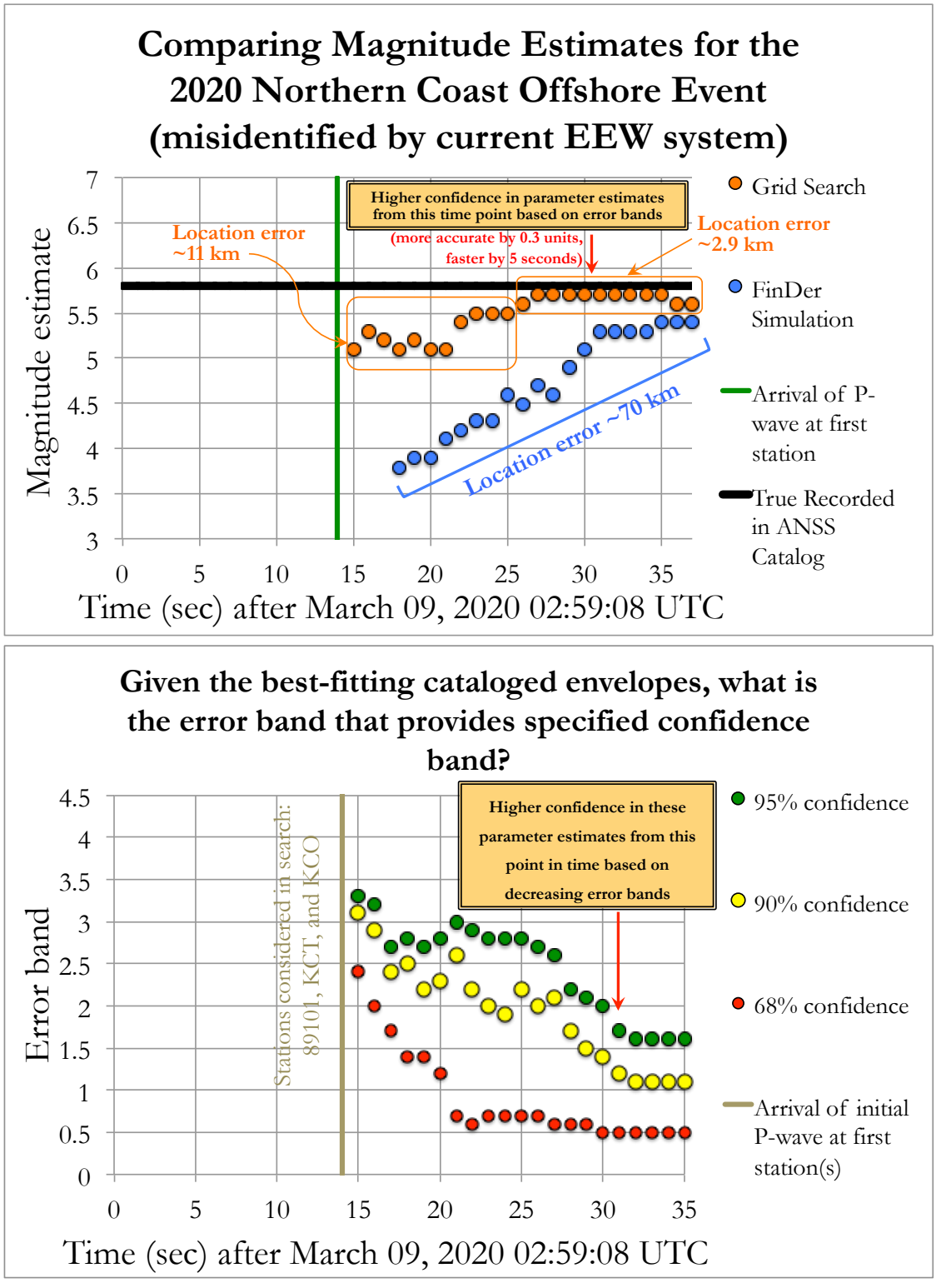
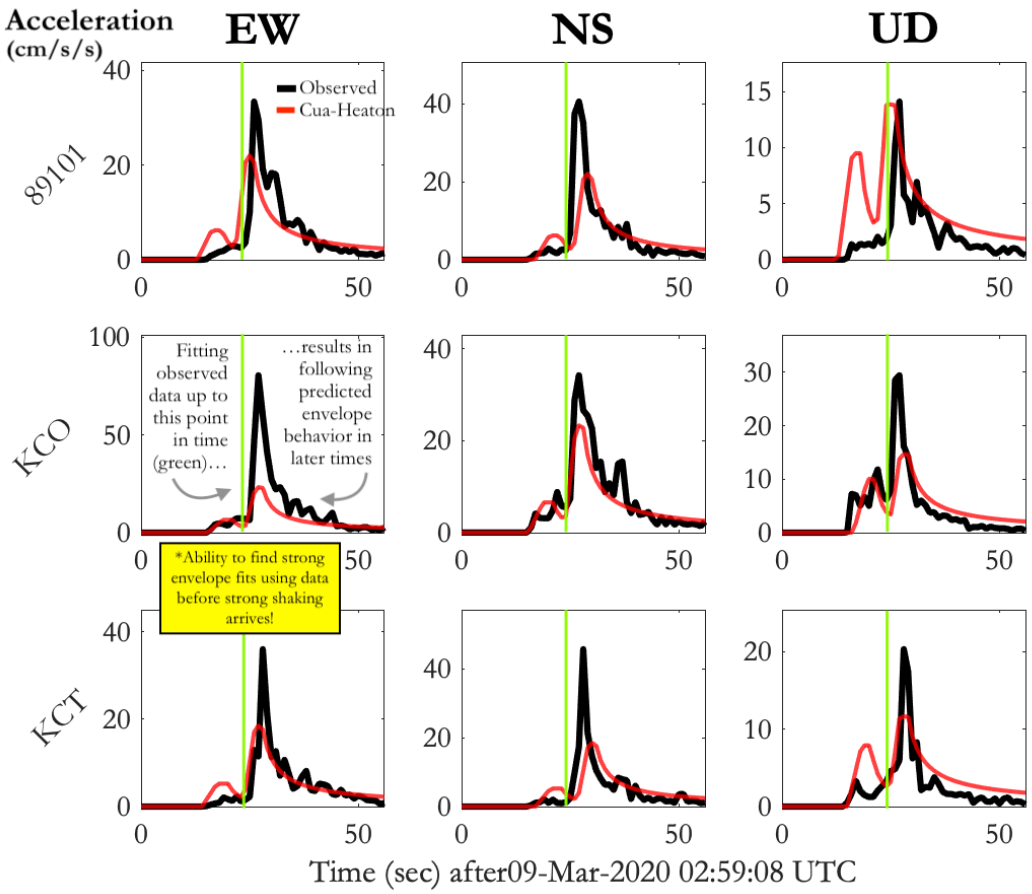
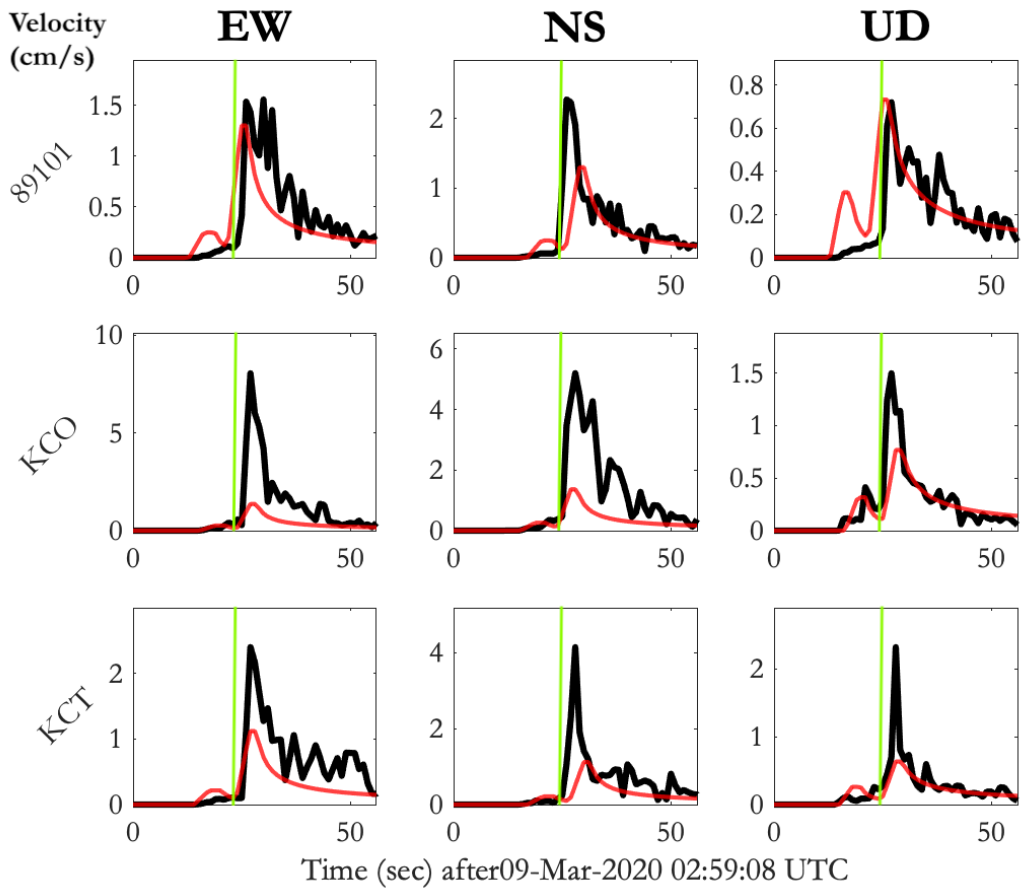


Figure 3.9. Grid search magnitude estimates for the 2020 Northern coast offshore event. Along with magnitude estimates, error bands needed for 95%, 90%, and 68% confidence bands are plotted.



(figure continues next page)



(figure continues next page)

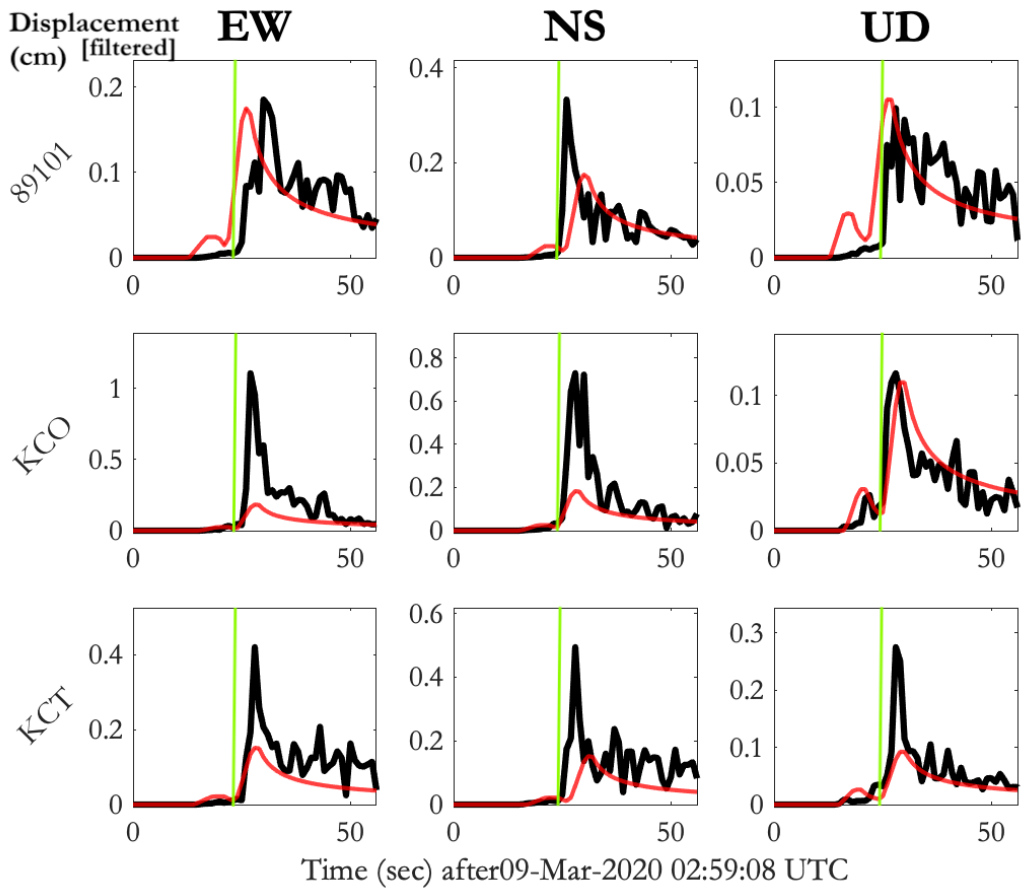


Figure 3.10. Comparing the best-fitting cataloged (in red) and incoming observed envelopes (in black) for the 2020 Northern coast offshore event. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and filtered displacement are also labeled accordingly. The data lying on the left side of the green vertical line is used to find the best-fitting Cua-Heaton envelopes. Using data before strong shaking arrives is still able to accurately predict amplitudes to come later in time.

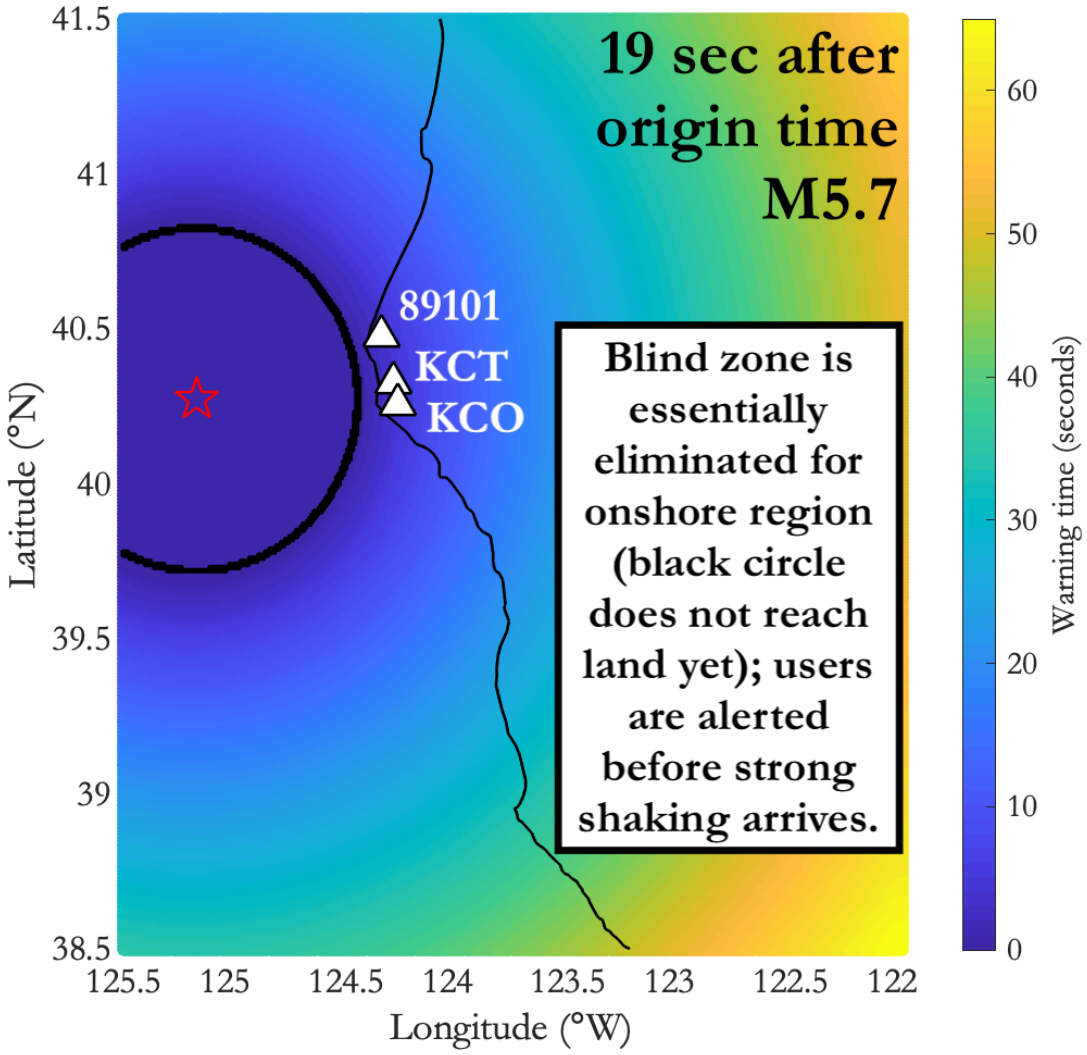


Figure 3.11. Estimated warning time in specified Northern coast region. 19 seconds after the origin time, the magnitude estimate approaches M5.7, and the location estimate converges to 15 km from the true epicenter. At this time point, alerts can be issued to users before strong shaking arrives onshore. Essentially, the blind zone is eliminated for the onshore region.

3.7.2 2019 Lone Pine foreshock-mainshock pair

Another type of earthquakes considered in this study is a foreshock-mainshock pair part of a sequence. The next earthquake to study is the 2020 Lone Pine mainshock.

Approximately 41 hours before the M5.8 mainshock, earthquake history shows a M4.62 foreshock that occurred 0.7875 km away. The computations involve waveform envelopes from the first four triggered stations listed in Table 3.4. The first station, Cottonwood Creek (CWC), is triggered just 2 seconds after the origin time, and the rest, Cerro Gordo (CGO), McCloud Flat (MWF), and Darwin (DAW) follow 4, 7, and 7 seconds after, respectively. Overall, the Cua-Heaton envelopes provide strong fits for acceleration and velocity, as shown in Fig. 3.13, but have difficulty fitting the long period components in the filtered displacement. The accuracy of these envelope fits is quantitatively shown through the trends in the error bands. Initially, the error bands remain similar until 18 seconds after the origin time where they decrease by 63%, 71%, and 68% for the 68%, 90%, and 95% confidence bands, respectively (see Fig. 3.12).

Initially, there is tradeoff between the magnitude and location estimates; the initial magnitude estimate is underestimated to M4.1 with the location error being about 8 km (see Fig. 3.12). Just 7 seconds after, which is faster than the current system's solutions, the magnitude estimate grows to M5.9 with the location error being approximately 0.89 km. Here, the grid search finds a location estimate 29 km closer to the true epicenter than the current system. Eventually, the magnitude estimates approach M6.1, which is 0.3 units from the true value. The confidence here is amplified, as the error bands get smaller. In real-time application, more stations would be considered in the computations. More stations would reduce uncertainties, farther decreasing the error bands. A consistent comparison to the current system solutions is valid up until 10 seconds after the origin time. Afterwards, worst-case scenario is shown in Fig. 3.12. Even so, obtaining an accurate estimate 18 seconds after the origin time is similar to the current EEW system performance.

In reality, the current EEW system detects the mainshock, but with tradeoffs between the location and magnitude estimates. It overestimates the magnitude to M6.0 and locates the epicenter with an error of 30 km. It also takes at least 20 seconds for estimates to converge and lock in. The grid search is able to recognize the incoming ground motions with similar time, 18 seconds after the origin time, with a magnitude estimate error of 0.3 and

location error of 0.89 km. Grid search performs better in accuracy but similar in speed in obtaining the parameter estimates. Because both grid search and the current system obtain the parameter estimates in similar speeds, the warning times and blind zones remain relatively the same.

The 2020 Lone Pine mainshock, along with the 2019 Ridgecrest mainshock, is intentionally studied for this thesis because the preceding foreshock provides valuable information. This chapter addresses calculations using only waveform-based data, but a foreshock-mainshock pair is a case where prior information from the foreshock has the potential to reduce tradeoffs in location and magnitude estimates for the mainshock. One such prior is the Epidemic Type Aftershock Sequence (ETAS) model, which is described in Chapter 8 (Felzer 2009).

Table 3.4. Triggered stations from the 2020 Lone Pine mainshock with P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	**Latitude(°N)**	**Longitude (°W)**	**P-wave arrival (sec after origin time)**
CI.CWC	36.4399	118.0802	2
CI.CGO	36.5504	117.8029	4
CI.WMF	36.1176	117.8549	7
CI.DAW	36.2715	117.5921	7

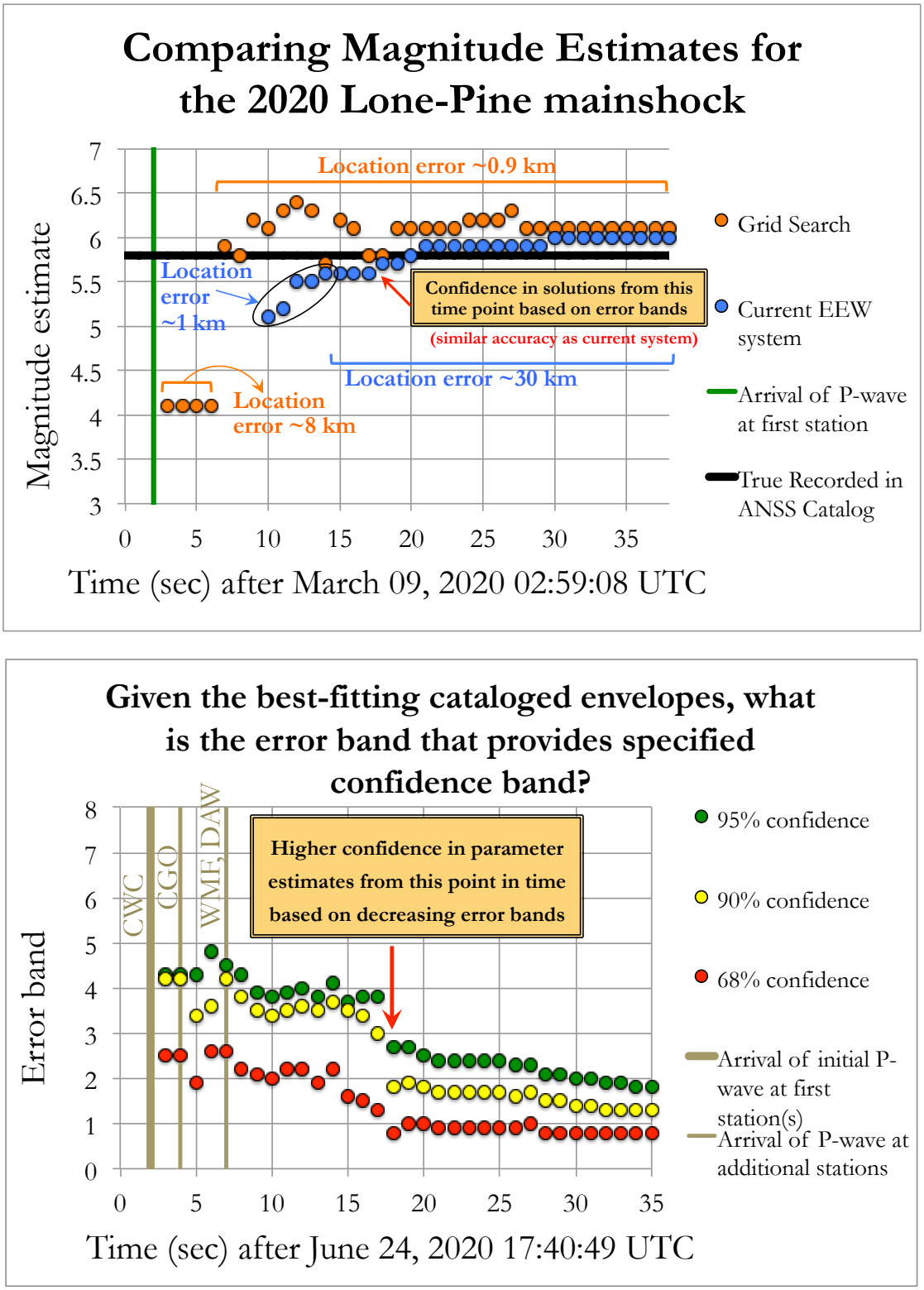
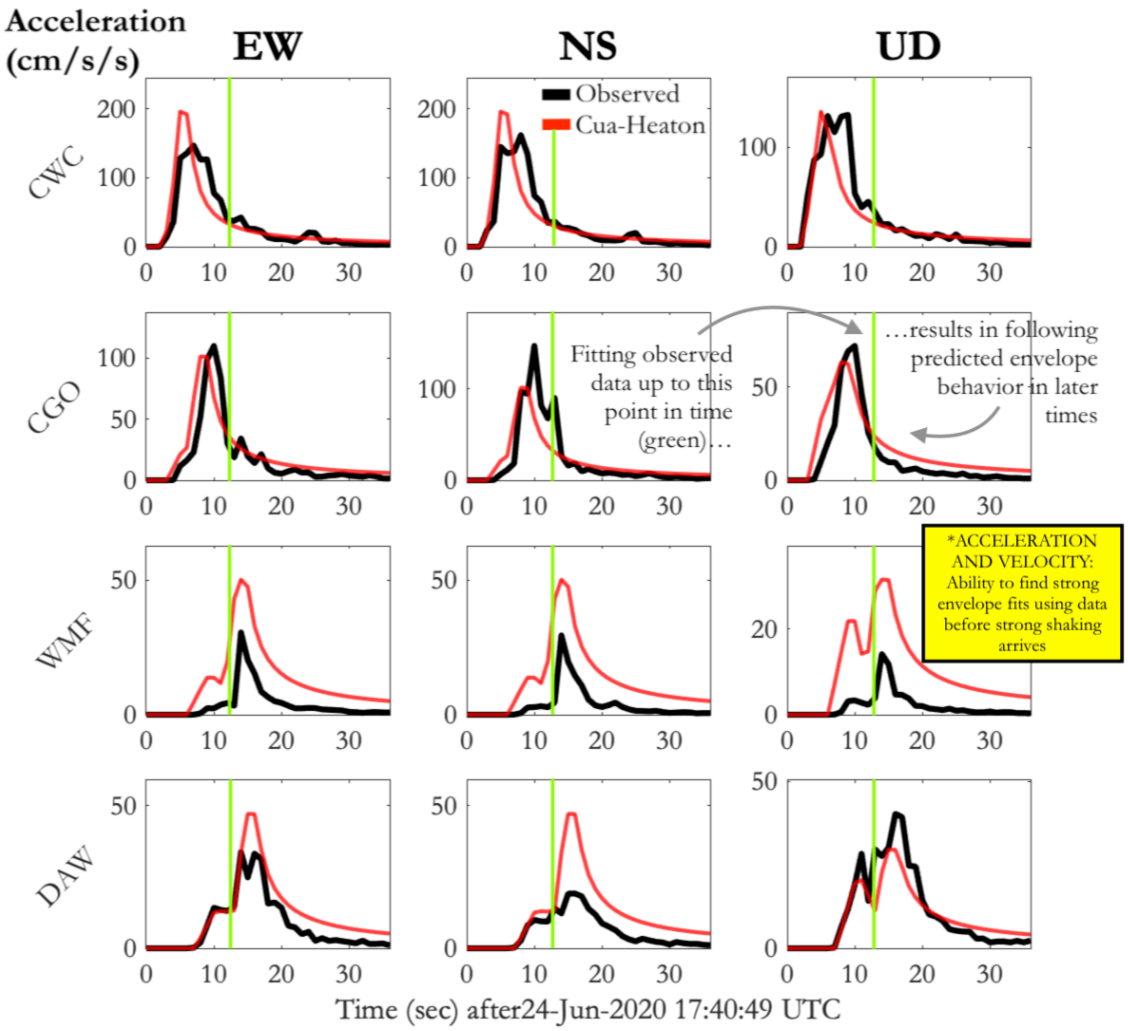
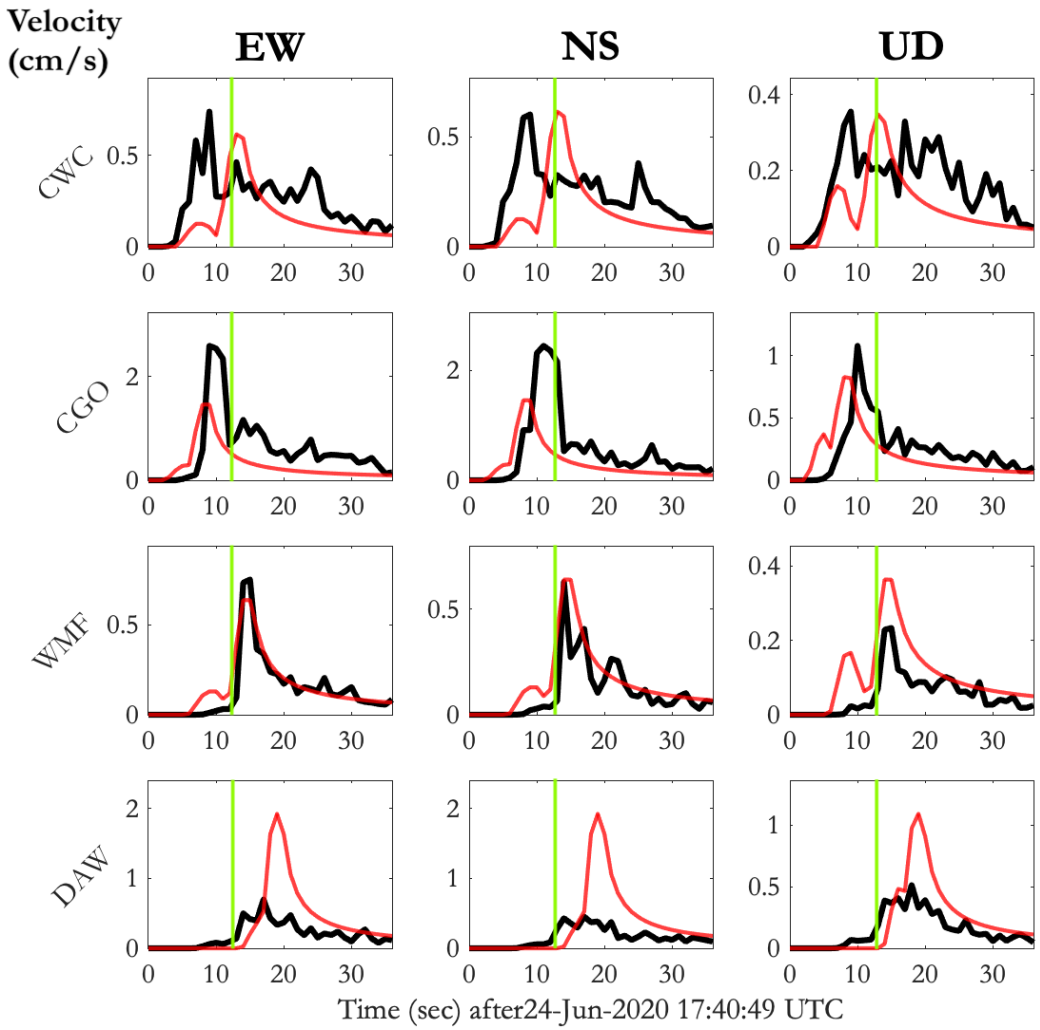


Figure 3.12. Grid search magnitude estimates for the 2020 Lone Pine mainshock. Along with magnitude estimates, error bands needed for 95%, 90%, and 68% confidence bands are plotted.



(figure continues next page)



(figure continues next page)

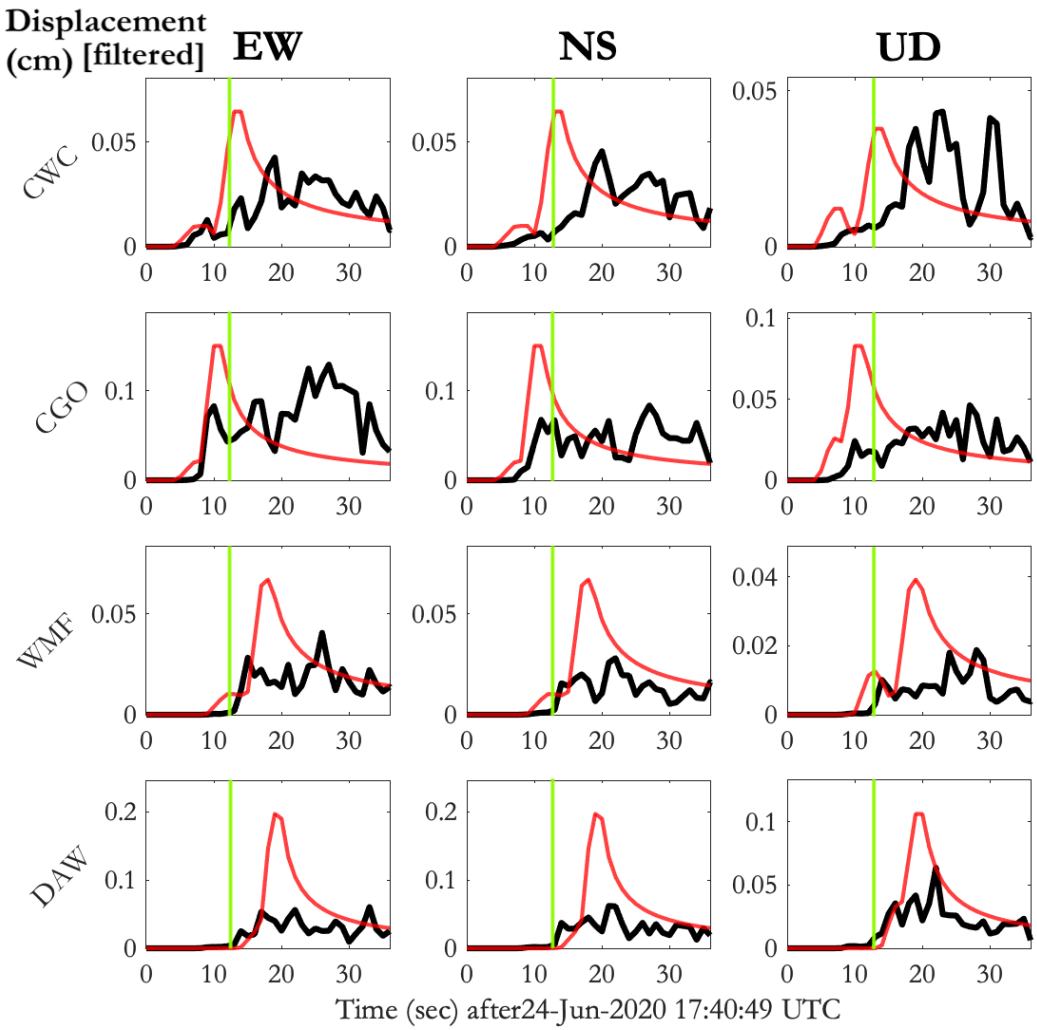


Figure 3.13. Comparing the best-fitting Cua-Heaton (in red) and incoming observed envelopes (in black) for the 2020 Lone Pine mainshock. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and filtered displacement are also labeled accordingly.

3.7.3 2019 Ridgecrest sequence

The 2020 Northern coast offshore event and 2020 Lone Pine mainshock are examples where point source characterization of the earthquake is valid. However, this may not be the case for larger earthquakes (M>6.5). The next example to study is the 2019 Ridgecrest sequence, where the M6.4 foreshock was followed a M7.1 mainshock. This particular foreshock-mainshock pair is spaced apart in time by nearly 34 hours and in space by 11 km. The mainshock ruptured bilaterally in the NW-SE direction for a cumulative length of ~65 km (Ross et al. 2019). Some discrepancy in the envelope fits is expected due to this long duration and rupture length. If the station-to-epicenter distance is less than the rupture length, a point source characterization may not be valid and a finite fault characterization is needed. However, for this particular chapter, point source is assumed as the Cua-Heaton envelopes were designed with this in mind (Cua 2005). An earthquake of this large magnitude is the upper limit of the grid search.

The computations involve waveform envelopes from the first seven triggered stations listed in Table 3.5. Overall, the Cua-Heaton envelopes provide strong fits for the acceleration, but have difficulty fitting the longer periods in the velocity and filtered displacement. The grid search struggles to find Cua-Heaton envelopes that capture the larger ground motions at the two stations closest to the fault, which are China Lake (CLC) and Christmas Canyon China Lake (CCC). The fits are most accurate for the remaining stations, which are Tower 2 (TOW2), Snort (SRT), Renegade Canyon (WRC2), Slate Mountain (SLA), and Laurel Mtn (LRL). For these five stations, point source assumption is valid as they are located at least 15 km from the epicenter. For reference, models suggest the rupture length is approximately 20 km.

Table 3.5. Triggered stations from the 2019 Ridgecrest mainshock with P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	Latitude(°N)	Longitude (°W)	P-wave arrival (sec after origin time)
CI.CLC	35.8157	117.5975	1
CI.TOW2	35.8086	117.7649	3
CI.SRT	35.6923	117.7505	3
CI.WRC2	35.9479	117.6504	4
CI.SLA	35.8909	117.2833	5
CI.LRL	35.4795	117.6821	5
CI.CCC	35.5249	117.3645	6

Initially, the grid search underestimates the magnitude to M3 with location error 11 km (see Fig. 3.14). A consistent comparison to the current system solutions is valid up until 7 seconds after the origin time. Afterwards, worst-case scenario is shown in Fig. 3.14 as only the first seven stations are considered in the analysis. Even with a limited amount of stations, 10 seconds after the origin time, the magnitude estimates grow to ~M6 with location error 2 km, which is essentially the same as the current system's results. The grid search finds the final magnitude to be M6.9, which is 0.6 units more accurate than the current system. In comparison, the current EEW system underestimated the final magnitude to M6.3. As reference, the current system solutions approach M6.4 as the final magnitude. It is only known in retrospect that the M7.1 Ridgecrest mainshock is one of the complicated sequences with multiple sources rupturing close in time and in space. The use of error bands creates an opportunity to recognize complicated sequences in real-time. As the rupture continues, the error bands help distinguish when envelopes of point source assumption do not suffice and when envelopes of finite fault characterization are required.

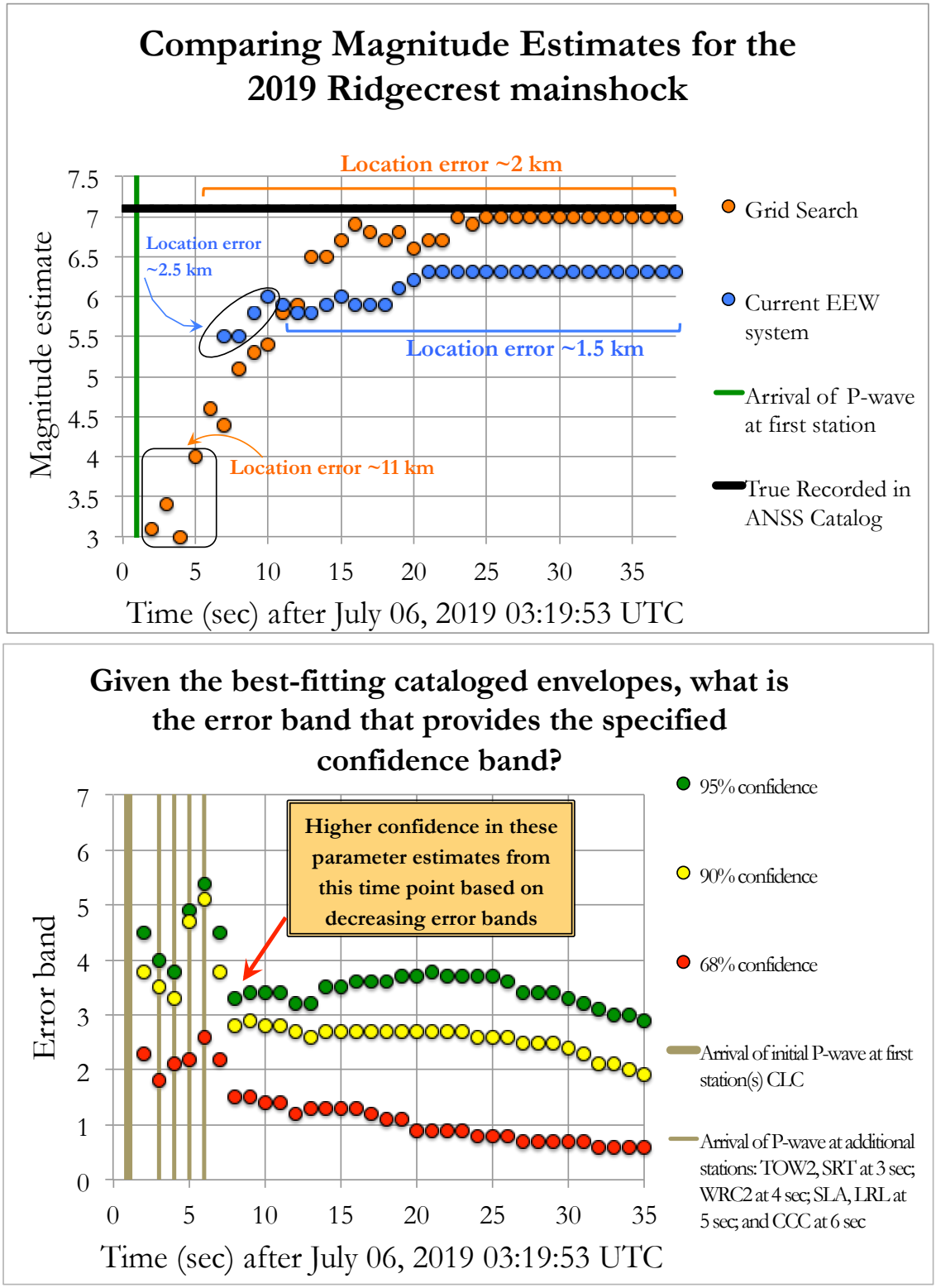


Figure 3.14. Grid search magnitude estimates for the 2019 Ridgecrest mainshock. Along with magnitude estimates, error bands needed for 95%, 90%, and 68% confidence bands are plotted.

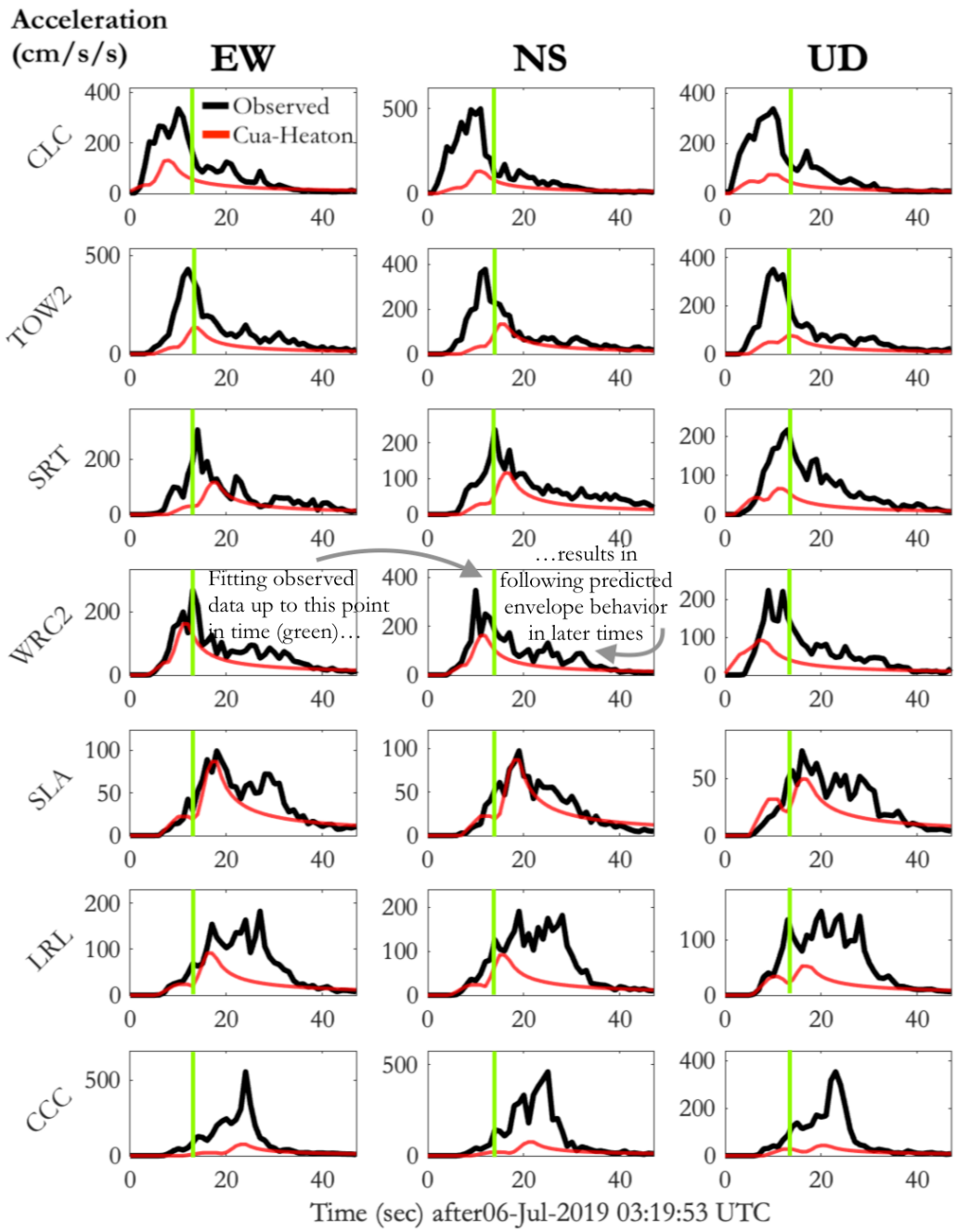


Figure 3.15. Comparing the best-fitting Cua-Heaton (in red) and incoming observed envelopes (in black) for the 2019 Ridgecrest mainshock. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration is shown only as fits for velocity and displacement have large uncertainties.

As previously mentioned, a large earthquake like the M7.1 Ridgecrest is the upper limit of the grid search. Grid search is intended for M<6.5 events (Cua 2005). Therefore, the grid search is expected to perform well using the waveform envelopes from the M6.4 Ridgecrest foreshock. By eye inspection, as the rupture continues in time, the grid search finds significantly more accurate envelope fits for the M6.4 foreshock than for the M7.1 mainshock. To quantitatively measure the fits, the error bands are plotted in Fig. 3.16. The smaller the error bands are, the stronger the fits are, as the envelopes do not have to be shifted as much to allow for 95% representation of the incoming envelopes. To acquire better fits for the observed envelopes of the M7.1 mainshock, additional templates are considered (see Chapter 6). These additional templates are based on multi-source models (Yamada 2007).

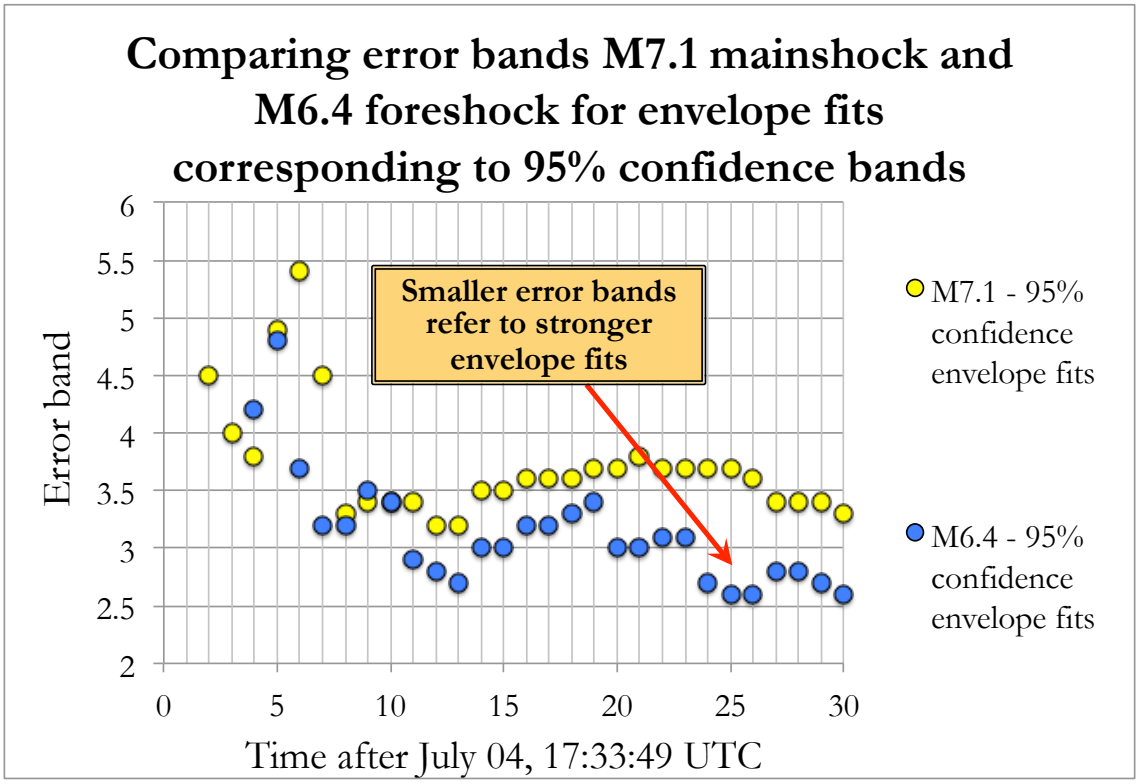


Figure. 3.16. Error bands to quantify envelope fits for the M7.1 (yellow) and M6.4 (blue) Ridgecrest events. These error bands quantify the differences between the envelope fits.

3.8 Further Magnitude Constraints Using Amplitude Ratios

One of the challenges in obtaining accurate parameter estimates quickly is the tradeoff between magnitude and location. Due to the tradeoffs, with only 1 to 2 seconds of data, the grid search may find multiple optimal parameter estimates. Therefore, constraining the magnitude may reduce the tradeoffs that occur in the initial part of the rupture. Cua uses the ratios between acceleration and displacement to estimate the magnitude using as little as 1 second of data (Cua 2005). The use of ratios stems from the idea that different frequency content shows different energies that are radiated, which can distinguish whether the incoming ground motion is one from a small earthquake or from a large one. Chapter 8 describes how Cua came up with the relationship that constrains magnitude estimates with available observed data, independent of epicentral distance. The relationship is applied to the same 4.5<M<7 earthquakes from the mini test sweep to observe how much waiting time can be saved. As seen in Fig. 3.17, the use of the ratios of ground motion amplitudes improves the waveform-based magnitude estimates from a median of 0.53 (standard deviation 0.40) to

0.14 (standard deviation of 0.21). This decrease increases the confidence in the magnitude estimates being more closely distributed about the true magnitudes.

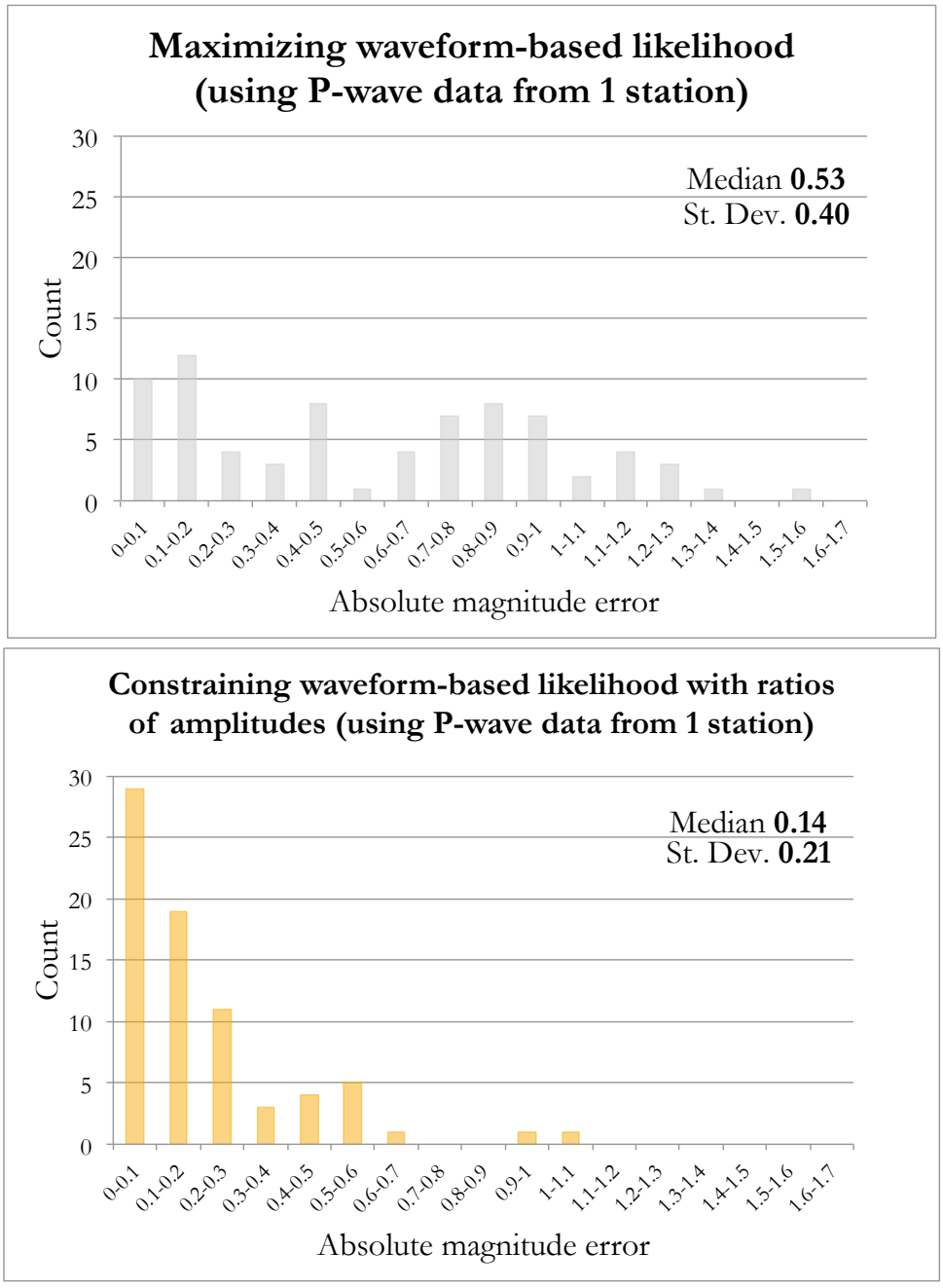


Figure 3.17. Maximizing waveform-based posterior probability (left) vs. maximizing waveform-based probability that is constrained by the ratios between ground motion amplitudes (right). Using the ratios as a constraint on the magnitude, the tradeoffs between the parameters are reduced, increasing the overall accuracy in the magnitude estimates.

3.9 Summary

The grid search is a simple, direct method in finding optimal parameters that describe the incoming observed ground motions from a rupturing earthquake. The calculations at each station and channel are independent and are done using the same grid space, allowing parallel execution. The parameters of interest are magnitude, latitude, and longitude. Taking all possible combinations of these parameters, the total grid space is defined as well as the corresponding Cua-Heaton ground motion envelopes. The posterior probability, which is based on the assumption of a normal distribution of the logarithmic residuals of the observed and predicted Cua-Heaton envelopes, is maximized for the best parameter estimates.

A test sweep on $4.5<M<7$ events reveals certain criteria that need to be satisfied for robust performance of the grid search. A uniform distribution of the seismic stations is required, and the suggested interstation spacing is 10 to 20 km. Therefore, the grid search is intended for regions of dense station coverage, like Los Angeles and San Francisco, but not for regions of sparse station density, like Baja California and northeastern California near Nevada. P-wave data from one station is adequate to find magnitude estimates that differ from the true values by 0.53 units. However, three stations are required to ensure the location estimates are of, in fact, global maximum probability. There may be multiple optimal location estimates when fewer than three stations are used. With three stations, location estimates differ from the true epicenters by 5 km and the magnitude estimates differ from the true values by 0.28 units. To reduce tradeoffs in the magnitude and location in the initial estimates, a constraint is applied using the ratios between the ground motion amplitudes (Cua 2005). Doing so, the standard deviation of the magnitude estimates found using P-wave data from one station decreases from 0.53 to 0.14, suggesting a higher confidence that the estimates are closely distributed about the true magnitudes.

The grid search uses Cua-Heaton ground motion envelopes, therefore, it is important to understand the assumptions that were made when Cua designed them. The strongest of the assumptions is the point source characterization of the earthquake. Therefore, for large earthquakes ($M>6.5$), the grid search may underestimate the ground motions as the Cua-Heaton envelopes may not have the ability to capture the large ground motions that are amplified along the direction of the longer fault rupture. For such cases, the

extended catalog search using templates of complex sequences, or a combination of multiple subevents, is more appropriate.

4 Method II: Extended Catalog Search

The second method of the two-part search algorithm is the extended catalog search. It is intended to run in parallel to the first method, which is the standard grid search (see Chapter 3). If both methods agree on a solution, it further increases confidence in the alerts. The extended catalog search is based on the general idea of image comparison and template matching. Given a waveform envelope from an incoming earthquake, the extended catalog search looks for the best match from a catalog consisting of real and synthetic earthquakes. In this chapter, the extended catalog search is applied using waveform data only. In Bayesian probability terms, this means the extended catalog search maximizes the normalized likelihood of the envelope amplitudes assuming uniform prior information, just as in Chapter 3. Uniform prior information implies this search algorithm assumes earthquakes of every magnitude and location occur equally likely. To see how prior information is applied to the extended catalog search, refer to Chapter 8. In this chapter, the extended catalog search is performed on the same three real earthquakes studied in Chapter 3. Once again, an earthquake is assumed to be point source. Chapter 6 addresses how the extended catalog search handles complicated sequences by considering additional templates for multi-source models. Complicated sequences refer to those that defy the conventional assumptions that earthquake ruptures are from individual faults (i.e. M7.8 Kaikoura multi-fault rupture).

4.1 The Usefulness of Catalog-Based Search Algorithms

A technological advancement that many people take special interest in today is the facial recognition system. There are many different techniques to create this system, but essentially, the main idea behind it is an image comparison, or template matching. The general strategy is to consider all possible positions (brute-force) of the templates from a database or catalog and find the one that best matches, or that shares the most similarities with, the target image. The challenge in template matching is guaranteeing an accurate match to the target image. Inaccuracies can have serious consequences in some cases. For instance, law enforcement agencies use facial recognition software to identify criminals. The credibility would be lost with false positive results and large error bands.

The extended catalog search is based on this general idea of template matching, where an incoming observed ground motion is compared to previously observed ground motions from an earthquake catalog. The goal is to find one that best describes the observed ground motion, or more specifically, to find the best parameter estimates (i.e. magnitude and location). To ensure accuracy, the catalog must cover a large enough range of earthquakes. A sufficient catalog consists of a huge variety of both epicentral locations and magnitudes. A past observation is that almost 50% of all earthquakes have foreshocks recorded in earthquake history, and this is the rationale behind the extended catalog search (Abercrombie and Mori 1996). This past observation strongly implies that an event already exists in the catalog that closely matches the incoming earthquake. Assuming the original catalog already consists of many epicentral locations, it is further extended with respect to earthquake magnitude only. This is to ensure representation of the larger earthquakes that are less frequent, as stated by the Gutenberg-Richter law.

The feature of the extended catalog search that sets it apart from other algorithms is its uniqueness of the solutions to the station and channel at hand. Therefore, the main advantage of this search algorithm is its inclusion of not only earthquake source effects, but also site and path effects. In other words, the search does not only consider distance for predicted ground motion amplitudes; direction is also considered. The focal mechanisms derived from the first ground motions show how the polarity of the first P-wave arrival varies between seismic stations at different directions from an earthquake (i.e. compressional has material displacing towards the station while dilatational has material displacing away from the station). This is different from the standard grid search from Chapter 3 where the solutions are more of an average result using pre-determined GMPEs. The unique matches from the extended catalog search especially help with reliability in estimates for single-station approach. It has the potential to reduce false alerts without waiting for multiple stations.

The term "extended" refers to the catalog being extended with respect to magnitude only. It is not extended with respect to epicentral distance or focal mechanism. The catalog of interest is Southern California, and based on Fig. 4.1(a), a variety of epicenters can be seen by eye inspection. In fact, based on 154,671 earthquakes in California, an event exists almost every 25 km in space for the region between the U.S.-Mexico border and San Francisco, as shown in Fig. 4.1(b). Therefore, the original earthquake catalog is assumed to have a

sufficient range of epicenters. To vary the focal mechanisms, another model is needed to further extend the catalog (Heaton 1979).

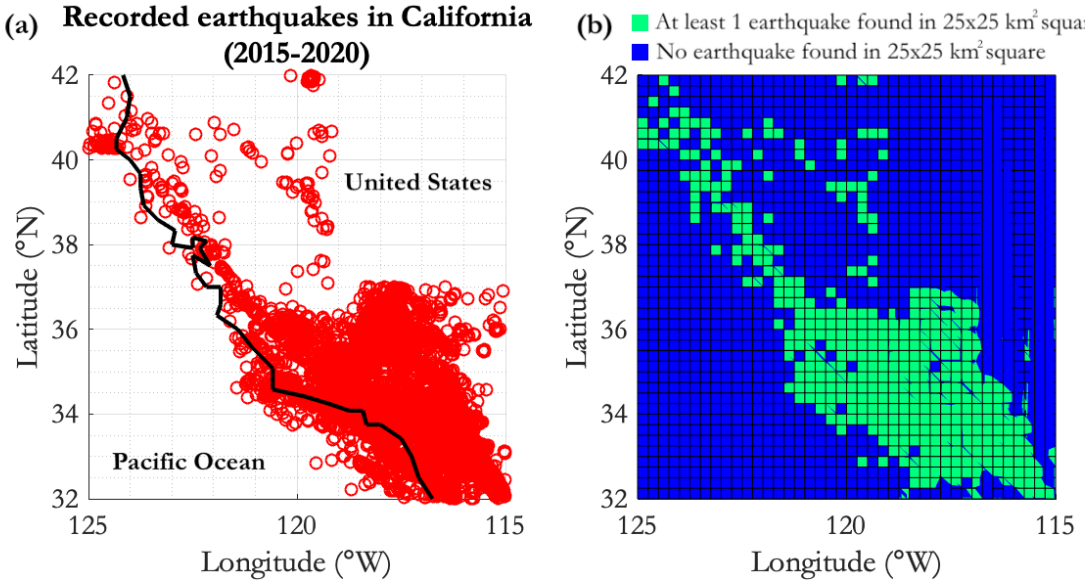


Figure 4.1. All recorded epicenters in California from 2015 to 2020. (a) Recorded clustered earthquakes in California. (b) Because at least one earthquake is found every 25 km in the onshore region between San Francisco and the U.S.-Mexico border, sufficient range of epicenters is assumed, and the catalog is not extended with respect to location.

4.2 Defining the "Goodness-of-Fit"

As previously mentioned, the goal of the extended catalog search is to find a close match to the ground motions of the incoming earthquake. A close match is defined as the one with the best score, denoted as the "goodness-of-fit". The measurement used to define the goodness-of-fit is the posterior probability of the parameters of interest (i.e. magnitude, epicentral location) given the incoming observed ground motions. The best estimates of the parameters are obtained by maximizing the posterior probability over a set of cataloged earthquakes. Bayes' theorem states the posterior probability is the normalized product of a prior and a likelihood function. In this chapter, the simplest case is assumed, that the prior is uniform. A uniform prior in the earthquake magnitude and location implies that earthquakes of all magnitudes and locations are equally likely. However, certain relationships prove this is not true. For instance, the Gutenberg-Richter law states smaller earthquakes occur more frequently than large ones (Gutenberg and Richter 1944). It is also commonly known earthquakes often cluster in time and space (Ogata 1998). Therefore, in a later chapter,

different prior information is applied to the likelihood to calculate a posterior probability that aims to provide accurate parameter estimates without jeopardizing warning time for regions expected to experience strong shaking. Using a uniform prior, the posterior probability is merely the normalized likelihood, which is calculated using only waveform information.

When an earthquake ruptures, the seismic stations receive waveform information at high sample frequencies, such as 100 and 200 samples per second. To decrease chances of latency in the early warning system that arises from heavy data collection, the input data used in the extended catalog search is waveform envelopes, instead of the full waveform time series, with amplitudes taken in 1-second windows. Using ground motion envelopes, the posterior probability to maximize for the best parameter estimates is defined in Eqs. 4.1. Eqs. 4.1 may look familiar, as it is the same function as the one mentioned in Chapter 3. The only difference is the type of templates used as the variable X_{ijk}. As previously mentioned in Chapter 3, the likelihood function models the logarithmic amplitude residuals as a normal random variable.

$$Pr_{post}(M, lat, lon|Y) \propto Pr_{like}(Y|M, lat, lon) = \prod_{k=1}^{M} \prod_{j=1}^{N} \prod_{i=1}^{P} Pr_{like}\left(Y_{ijk}\middle|M, R\right) \tag{4.1.1}$$

$$Pr_{like}\left(Y_{ijk}\middle|M, lat, lon\right) \propto \exp\left[-\left(\frac{\left(Y_{ijk}-X_{ijk}\right)^2}{2{\sigma_{ijk}}^2}\right)\right] \tag{4.1.2}$$

$$Pr_{post}(M, lat, lon|Y) \propto \exp\left[-\sum_{k=1}^{M}\sum_{j=1}^{N}\sum_{i=1}^{P}\left(\frac{\left(Y_{ijk}-X_{ijk}\right)^2}{2{\sigma_{ijk}}^2}\right)\right] \tag{4.1.3}$$

where Y_{ijk} is the logarithmic (base 10) amplitude of the ground motion envelope observed at the j^{th} channel, k^{th} station, and i^{th} time point. X_{ijk} is the logarithmic (base 10) amplitude of the cataloged earthquake, and σ_{ijk} is the uncertainty of the fit. The total posterior probability of the cataloged earthquake is taking the product of the individual probabilities (Eq. 4.1.2) for N channels, M stations, and P time points.

Working in the natural logarithmic form helps avoid computations with very small values. Therefore, maximizing the posterior probability over a set of cataloged earthquakes (defined in Eq. 4.1.1 and Eq. 4.1.3) can be transformed to a minimization problem. It is computationally simpler to minimize the sum of squared residuals (SSR), defined in Eq. 4.2, over the set of cataloged earthquakes. The SSR is simply the negative function within the exponential in Eq. 4.1.3. The minimization of the function defined in Eq. 4.2 over the set of cataloged earthquakes is justified as constant variance is assumed. This assumption has been

proven true; the logarithmic amplitude residuals follow a normal and independent distribution about zero mean and constant variance (Cua 2005).

$$SSR(Y|M, lat, lon) = \sum_{k=1}^{M} \sum_{j=1}^{N} \sum_{i=1}^{P} \left(\frac{\left(Y_{ijk} - X_{ijk}\right)^2}{2\sigma_{ijk}{}^2} \right) \tag{4.2}$$

where Y_{ijk} is the logarithmic (base 10) amplitude of the ground motion envelope observed at the j^{th} channel, k^{th} station, and i^{th} time point. X_{ijk} is the logarithmic (base 10) amplitude of the cataloged earthquake, and σ_{ijk} is the uncertainty of the fit.

Different methods in data fitting exist to meet different needs. As previously mentioned, the extended catalog search uses the minimization of the sum of squared residuals to find the best parameter estimates. Generally, a drawback of using the sum of squared residuals is its sensitivity to outliers. For example, a cataloged ground motion envelope may fit the arrival of the incoming observed envelope, but may not necessarily fit the coda. A standard sum of squared residuals using absolute amplitudes will heavily penalize the goodness-of-fit based on the large misfit in the coda, despite the small misfit in the arrival. The use of logarithmic amplitudes addresses this issue as it emphasizes both small and large amplitudes in evaluating the goodness-of-fit. Therefore, the use of logarithmic amplitudes makes the evaluation of the goodness-of-fit more robust. Also, logarithmic amplitudes follow the Gaussian distribution, which is the assumption made in defining the posterior probability in Eq. 4.1.3. The minimization of the sum of squared residuals resembles an L_2 norm. For more robustness, a hybrid of the L_1 and L_2 norms may be applied instead. Future work may include a test sweep using this hybrid, called the Huber norm.

4.3 Interpreting the Best Fits with Error Bands

It is not adequate to merely minimize the SSR to find the synthetic envelopes from the catalog that best describe the incoming ground motions. Another important solution to include is the quantification of how precisely the best-fitting envelopes fit. In the world of statistics, many terms exist to quantify how precise the model fits the observed data. Such terms are uncertainty, standard deviation, and margin of error. To avoid confusion, throughout this thesis, the terms to quantify the model's performance against the observed data are error bands that satisfy a specified confidence band. Error bands enclose the area about the best fit in which the true observed data can be found. They give a visual sense of

how well the observed envelope fits the best-fitting envelope. For example, a 95% confidence band ensures the error bands about the best fit contain 95% of the true observed data. The error band, or the area about the best fit, is defined in Eq. 4.3. A visual of the error band is given in Fig. 3.4 in Chapter 3.

$$\log X_{ijk} - \eta \leq \log Y_{ijk} \leq \log X_{ijk} + \eta \tag{4.3}$$

where Y_{ijk} is the logarithmic (base 10) amplitude of the ground motion envelope observed at the j^{th} channel, k^{th} station, and i^{th} time point. X_{ijk} is the logarithmic (base 10) amplitude of the cataloged earthquake, and η is the error band that is adjusted accordingly to satisfy the confidence band.

While the minimized SSR, or highest relative probability, chooses the envelopes that best match the incoming ground motions from the extended catalog, the error bands provide information on the rest of the envelopes that the whole catalog holds. For instance, the best-fitting envelope has the smallest SSR but may still have large error bands if the catalog itself does not contain envelopes that accurately represent the incoming observed ground motions.

Therefore, along with the SSR, the error bands are calculated. The initial confidence band for the error bands to satisfy is chosen as 68%, based on the Empirical Rule (68-95-99.7) rule. This rule states that for a normal distribution, about 68% of the observed data is within 1 standard deviation of the mean, 95% of the data is within 2 standard deviations, and 99.7% of the data is within 3 standard deviations. Because normal distribution is assumed for the logarithmic difference of the envelope fits (lognormal distribution for absolute difference), the mean for the best fit would approach a mean of 0. Therefore, one standard deviation would be the error band about the best-fitting envelope, which would contain at least 68% of the incoming ground motion amplitudes.

Once the error band about the envelope fit is found, it can be transformed into the uncertainty in ground shaking, which is the focus of EEW. The modified Mercalli intensity scale, or MMI, is the value used to characterize the ground shaking. Based on current GMPEs used, a generally acceptable threshold for the uncertainty in logarithmic PGA and PGV amplitudes is a factor of 2, which is approximately 0.3. As seen in Eq. 4.4 and Eq. 4.5, this value is converted to an uncertainty in MMI of 1.1, based on the relationships by Wald 1999 (Wu et al. 2007). However, for a more accurate acceptable threshold, a test sweep on a variety of events is required. Ultimately, the algorithm will use this acceptable threshold to

declare if the confidence is high enough to send parameter estimates to users. The scope of this thesis, however, finds solutions in terms of confidence bands of 68%, 90%, and 95%.

Algorithm A:

while $\frac{\sum_{k=1}^{M}\sum_{j=1}^{N}\sum_{i=1}^{P} Z_{ijk}}{\sum_{k=1}^{M}\sum_{j=1}^{N}\sum_{i=1}^{P} 1} < 68\%$

$$Z_{ijk} = \begin{cases} 1, if\ \log X_{ijk} - \eta \le \log Y_{ijk} \le \log X_{ijk} + \eta \\ 0, else \end{cases}$$

update η

end

$$MMI = 3.66 \log PGA - 1.66, V \le MMI \le VIII \quad (4.4)$$

$$MMI = 3.47 \log PGV - 2.35, V \le MMI \le VIII \quad (4.5)$$

where MMI is the modified Mercalli intensity scale, PGA is the peak ground acceleration, and PGV is the peak ground velocity.

4.4 Defining the Original Catalog

This study focuses on finding the best envelope fits to earthquakes that are particularly difficult to detect and identify. They include offshore events, like the 2020 Northern coast (Petrolia, CA), and sequences, like the 2019 Ridgecrest, 2020 Lone Pine, and 2012 Brawley swarm. Full waveforms (raw acceleration from strong-motion sensors) for these mainshocks are downloaded from the Southern California Seismic Network (SCSN), Northern California Earthquake Data Center (NCEDC), and Kyoshin Network (K-NET). Again, for earthquake early warning purposes, waveforms from the first few three to seven triggered seismic stations are considered. From the same stations, waveforms of previous M>3 earthquakes are also downloaded to initiate the creation of the catalog. Comparing the epicentral distributions in Fig. 4.1(a) and Fig. 4.2(a), there is already great variety in epicenters for M<3 earthquakes. Therefore, only M>3 earthquakes are downloaded for the purpose of scaling and creating synthetic waveforms for unavailable epicenters. For computational storage purposes, downloaded waveforms are within 100 km about the mainshock epicenter and date back less than 5 years before the mainshock. In the future, with ample storage space,

the original catalog can include additional waveforms from earthquakes from farther back in earthquake history (i.e. 10-20 years before the mainshock) as well as smaller magnitudes (i.e. M<3).

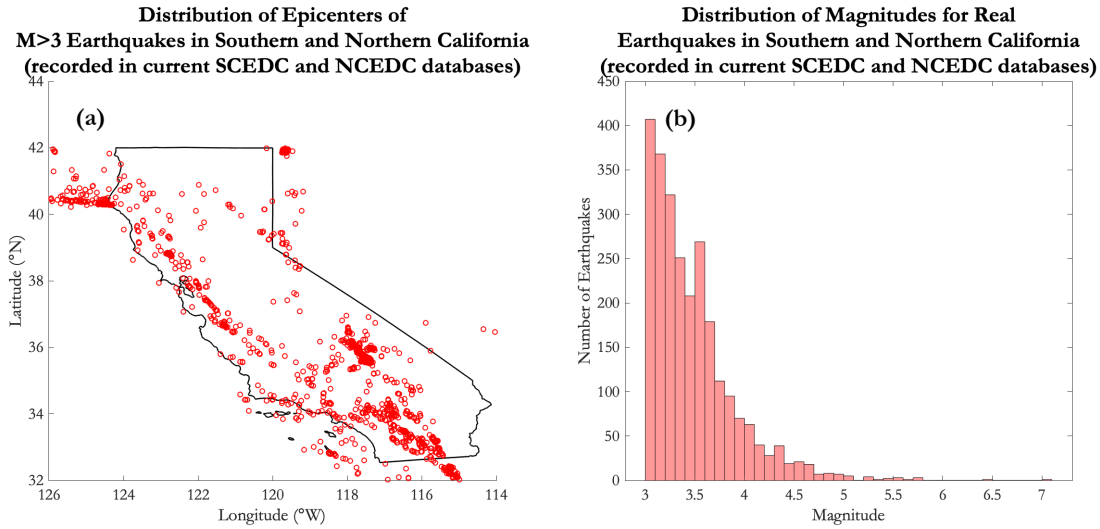


Figure 4.2. Distribution of epicenters and magnitudes in the current database for Southern and Northern California, looking back 5 years (years 2015 – 2020). Accuracy of extended catalog search relies on (a) variety of epicenters and (b) magnitudes. However, Gutenberg-Richter law says number of smaller earthquakes is greater than larger ones. The aim of the spectral scaling model is to create more (synthetic) earthquakes for M>5, in which records are scarce.

The collected raw acceleration time series are processed to find the corresponding velocity and displacement time series. The velocity and displacement are also included in the extended catalog search to ensure the waveform envelope fits cover a wide range of frequency bands, as acceleration covers only the high frequencies. Obtaining the velocity requires a single integration of the acceleration. However, integrating the raw acceleration sometimes leads to a linear trend due to tilting, response of the transducer to strong shaking, or issues with the analog-to-digital convert (Yamada et al. 2007). Therefore, before integration, raw accelerations are filtered using a causal fourth-order Butterworth high-pass filter with a corner frequency of 0.075 Hz. Raw velocity data from broadband sensors are not used to avoid issues of clipping, especially for earthquakes of large magnitudes (M>5). If the catalog includes M<3 earthquakes, raw velocity data from broadband sensors are downloaded instead of integrating the raw acceleration data. Displacement time series are acquired after another integration.

As mentioned earlier, the feature of the extended catalog search that sets it apart from other search algorithms its uniqueness to the seismic station at hand. Given an incoming ground motion from a single station and channel, the extended catalog search only compares it to waveforms recorded at the same station and channel. The search for a station-specific fit allows it to include the effects of not only the earthquake source (i.e. location, size), but also the site conditions. The specific station has unique local soil deposit, and the sediments near the ground surface at the site may have a large impact on the ground motion amplitudes, frequency, and duration. However, because most of the earthquakes in the catalog are of moderate sizes (M<7) where point source characterization is valid, one disadvantage is its inability to consider rupture propagation path effects. Chapter 6 addresses how this catalog can be further extended to include envelopes for complex sequences that consider rupture propagation. This chapter, however, only uses earthquakes of point source characterization.

4.5 Extending the Catalog

As previously mentioned, for this search algorithm to work adequately, diversity in the database is necessary. In other words, the searched database needs to sufficiently cover a wide range of potential earthquakes with respect to space and size. If similar data is not present in the searched database, then accuracy of the parameter estimates cannot be guaranteed, despite the high probability, or goodness-of-fit. It is assumed the original catalog consists of earthquakes with sufficient spatial coverage, that is, there are a variety of epicenters. However, due to the Gutenberg-Richter law, large earthquakes occur less frequently and are scarce in the database. To ensure the catalog also covers a wide range of magnitudes, a spectral scaling model is applied to raw ground motion records to generate synthetic ground motions. The impact of the spectral scaling model on the original catalog is shown in Fig. 4.3. While the original catalog of real earthquakes is scarce in magnitudes M>5, as shown in Fig. 4.3(a), the extended version of the catalog ensures those larger magnitudes are included, as shown in Fig. 4.3(b). Despite the extension, the shape of the histogram is maintained to ensure the Gutenberg-Richter law is still satisfied. With this, the

synthetics in the catalog still follow a realistic representation of earthquakes, with smaller earthquakes occurring more frequently than larger ones.

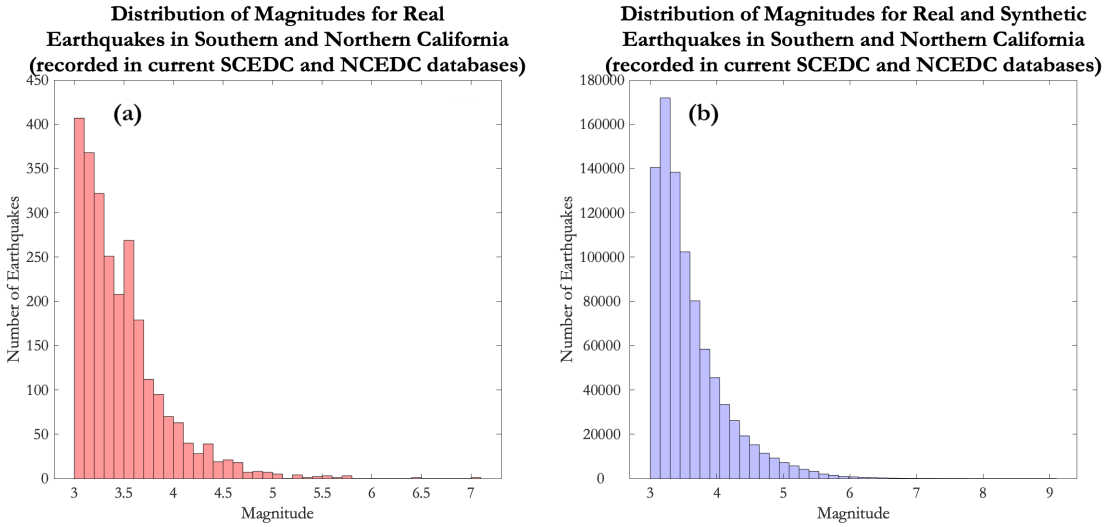


Figure 4.3. Extension of the catalog with respect to magnitude only. (a) Original catalog of real earthquakes. (b) Extended catalog includes larger magnitudes as well. The shape of the histogram is maintained to satisfy the Gutenberg-Richter law.

To extend the catalog, a transfer function is applied to the existing ground motions. The resulting synthetic ground motions would represent those from a different magnitude. This transfer function is a *simplified* version of the standard spectral scaling model (Brune 1970). The commonly used standard spectral scaling model is the ω^2-model. To simplify, a few assumptions are made.

First, this model assumes earthquakes are characterized as a point source. Also, the spectrum of an earthquake is assumed to take the form of Eq. 4.6, which is dependent on the frequency, corner frequency, and potency (Aki 1967).

$$|u_i(f)| \sim \frac{P_i}{1+\left(\frac{f}{f_{c_i}}\right)^2} \tag{4.6}$$

where $|u_i(f)|$ is the spectrum of i^{th} earthquake with spectral decay f^{-2}, f_{c_i} is the corner frequency that describes the rupture dimension, and P_i is the potency.

To show that earthquake spectra follow Eq. 4.6, the spectra of three real earthquakes are compared in Fig. 4.4. These earthquakes are from the Ridgecrest sequence: the M5.36, M6.4, and M7.1. The epicenters for these three earthquakes are similar (within 12 km of each other). Despite the different magnitudes, the shapes of the spectra are similar: constant for

lower frequencies and ω^2 drop-off trend for the higher frequencies. Other assumptions are constant stress drop independent of source size and constant rupture velocity. Together, these assumptions help create the simplified spectral scaling model needed to extend the catalog.

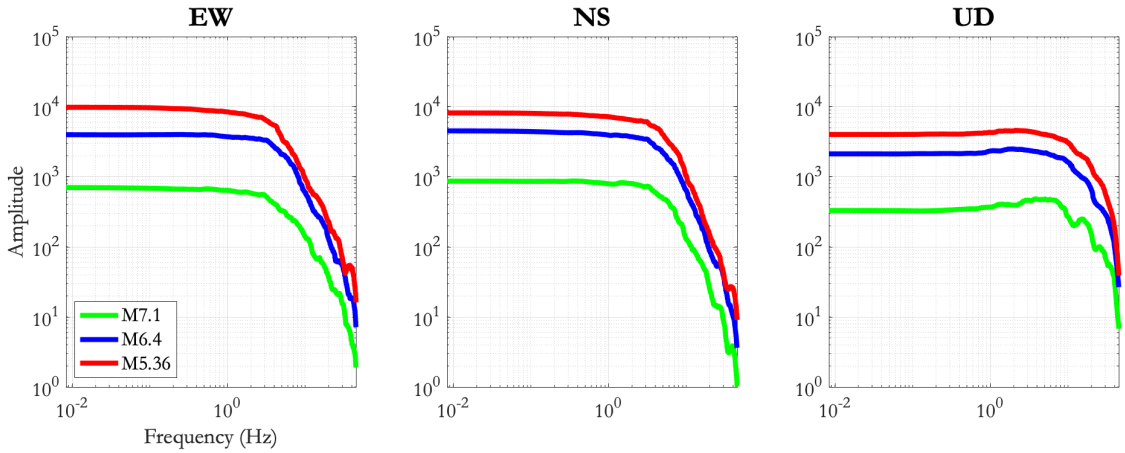


Figure 4.4. True spectra of available raw acceleration waveforms recorded at Station SRT (HNE, HNN, and HNZ). As the magnitude increases from M5.36 to M6.4 to M7.1, the spectrum scales up in amplitude but the shape remains similar, which is approximately constant for lower frequencies and ω^2 decay for higher frequencies.

Again, the key feature of the extended catalog search is its simplicity in producing a synthetic spectrum of an earthquake for which records are not available, such as scarce earthquakes of large magnitudes. Scaling a spectrum for which records are available produces this synthetic spectrum. Specifically, this scaling procedure refers to an application of a transfer function and is simple in that it reduces the depiction of the source spectrum to only one parameter, the magnitude (Heaton & Hartzell 1989). Because the transfer function is essentially a scaling factor of one earthquake to another, it can be written as a ratio of one spectrum to another, in which the spectra are in the forms defined in Eq. 4.6. The definition of moment magnitude given by Kanamori, as seen in Eq. 4.7, modifies this ratio into a function of only the corner frequency, magnitude, and frequency. Applying Eq. 4.7 to Eq. 4.6 and taking the ratio of one spectrum to another gives the transfer function, Eq. 4.8.

$$M = \frac{\log W - 4.8}{1.5} = \frac{\log(\sigma P) - 4.8}{1.5} = \frac{\log[2MPa \cdot P] - 4.8}{1.5} \approx \frac{\log P + 1.5}{1.5} \rightarrow P(M) = 10^{(\frac{3}{2})(M-1)} \qquad (4.7)$$

where M is the moment magnitude, W is the total work, σ is the effective stress, and P is the potency. The effective stress is assumed to be 2 MPa given the average crustal rigidity of 40 GPa.

$$S(f) = \frac{|u_j(f)|}{|u_i(f)|} \sim \left(\frac{P_j}{1+\left(\frac{f}{f_{c_j}}\right)^2} \right) \left(\frac{1+\left(\frac{f}{f_{c_i}}\right)^2}{P_i} \right) = 10^{\left(\frac{3}{2}\right)(M_j - M_i)} \left(\frac{{f_{c_i}}^2 + f^2}{{f_{c_j}}^2 + f^2} \right) \frac{{f_{c_j}}^2}{{f_{c_i}}^2} \tag{4.8}$$

where $S(f)$ is the transfer function and f is the frequency. The following variables correspond to earthquakes of same epicentral location but different magnitudes, M_i and M_j: $u_i(f)$ and $u_j(f)$ are spectra, P_i and P_j are potencies, and f_{c_i} and f_{c_j} are corner frequencies.

Eq. 4.8 is further simplified by applying Eq. 4.9, which is assuming constant stress drop of 2.7 MPa and constant rupture velocity of 2.8 km/s, and Eq. 4.7.

$$f_c \approx \frac{1}{T_c} = \frac{1}{\left(\frac{2L}{V_R}\right)} = \frac{V_R}{2L} = \frac{V_R}{2\left(\frac{\mu C P}{\Delta\sigma}\right)^{\frac{1}{3}}} = \left(\frac{{V_R}^3}{8C\mu}\frac{\Delta\sigma}{P}\right)^{\frac{1}{3}} = \left(\frac{{V_R}^3}{20\mu}\frac{\Delta\sigma}{P}\right)^{\frac{1}{3}} = \left(\frac{(2.8km/s)^3}{20(35GPa)}\frac{(2.7MPa)}{(10^{1.5M-1.5}m^3)}\right)^{\frac{1}{3}} \approx 10^{2.8-0.5M} s^{-1} \tag{4.9}$$

where f_c is the corner frequency, T_c is the duration, V_R is the rupture velocity (assumed to be 2.8 km/s), L is the rupture length, $\Delta\sigma$ is the stress drop (assumed to be 2.7 MPa), μ is the rigidity (assumed to be 35 GPa for upper crust), C is a constant describing the aspect ratio of the rupture dimensions (assumed to be 2.55 from Tom's notes on earthquake scaling), and P is the potency. Also assuming the width is the same as the length, the corner frequency can be written in terms of magnitude only.

This further simplification puts the transfer function, $S(f)$, in terms of magnitude and frequency (see Eq. 4.10). Because the transfer function in Eq. 4.8 is dependent on only the magnitude, scaling waveforms using the ω^2-model implies the epicentral location does not vary. In other words, applying the model to the original catalog only extends it with respect to earthquake magnitude.

$$S(f) = 10^{\left(\frac{3}{2}\right)(M_j - M_i)} \left(\frac{{f_{c_i}}^2 + f^2}{{f_{c_j}}^2 + f^2} \right) \frac{{f_{c_j}}^2}{{f_{c_i}}^2} \approx 10^{\left(\frac{1}{2}\right)(M_j - M_i)} \left(\frac{10^{5.6-M_i} + f^2}{10^{5.6-M_j} + f^2} \right) \tag{4.10}$$

where $S(f)$ is the transfer function and f is the frequency. The following variables correspond to earthquakes of same epicentral location but different magnitudes, M_i and M_j. f_{c_i} and f_{c_j} are the corresponding corner frequencies.

The spectra from the previously mentioned earthquakes of the Ridgecrest sequence provide an excellent comparison of the true observed with the synthetic. To create the synthetic of a M7.1 earthquake, the transfer function in Eq. 4.10 is applied to scale the true

spectra of the real M6.4 earthquake. Inverse Fourier transform of the scaled synthetic spectrum gives the ground motion in the time domain, shown in Fig. 4.5. Long period components observed in the true M7.1 earthquake seem to be well represented in the M6.4 true spectra. Similarly, to create the synthetic of a M6.4 earthquake, the transfer function is applied to scale the true spectra of the real M5.36 earthquake. However, long period components are not well represented in the true M5.36 earthquake, meaning they are also missing in the synthetic M6.4 ground motions. This is unclear in the synthetic acceleration (high frequency), but is clearer in the synthetic velocity. The envelope fits seem to match better at the initial part of the waveform than the coda. This discrepancy is not too significant for the purpose of EEW. However, to avoid missing long period components in synthetics, the original catalog is only extended up to +2 from the true recorded magnitudes. For instance, a waveform from a M3 earthquake scaled to one for a potential M7 earthquake will not resemble a true M7 due to missing long period components.

As shown in Figs. 4.5 and 4.6, the error bands are also calculated. They are denoted as σ, and they refer to the allowed tolerance about the synthetic envelope that contains at least 68% of the true amplitudes. The smaller the error bands, the better the fit is between the true and synthetic envelopes. Shown in Fig. 4.5, the spectra for the true M6.4 contains the long components present in the true M7.1. Therefore, scaling up the true M6.4 to produce a M7.1 synthetic leads to relatively accurate fits (majority have error bands less than factor of 2). However, the spectra for the true M5.36 are missing long period components present in the true M6.4. Specifically, this is seen in the discrepancy between the true and synthetic spectra in Fig. 4.6. A M6.4 synthetic produced by scaling up spectra that are missing long period components fails to capture the behavior of the true M6.4.

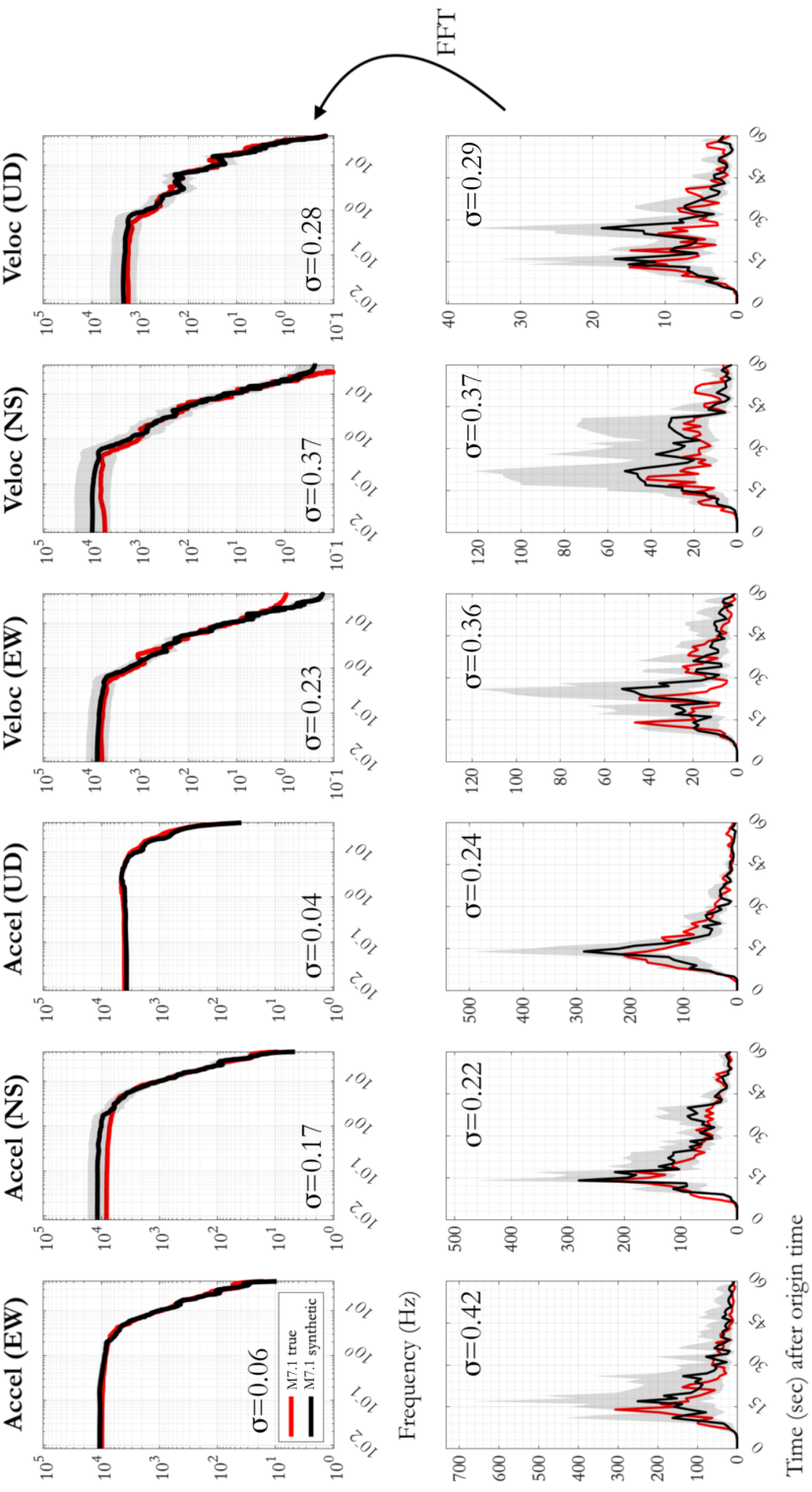


Figure 4.5. Synthetic spectra (in black) for a potential M7.1 earthquake generated using available waveforms (in red) at Station SRT (HNE, HNN, and HNZ). This is done by applying the transfer function to the raw waveforms that are available for M6.4 earthquake. The top row compares the true and synthetic spectra, and the bottom row compares the acceleration and velocity waveform envelopes. σ refers to the error band needed to represent at least 68% of the true observed envelopes.

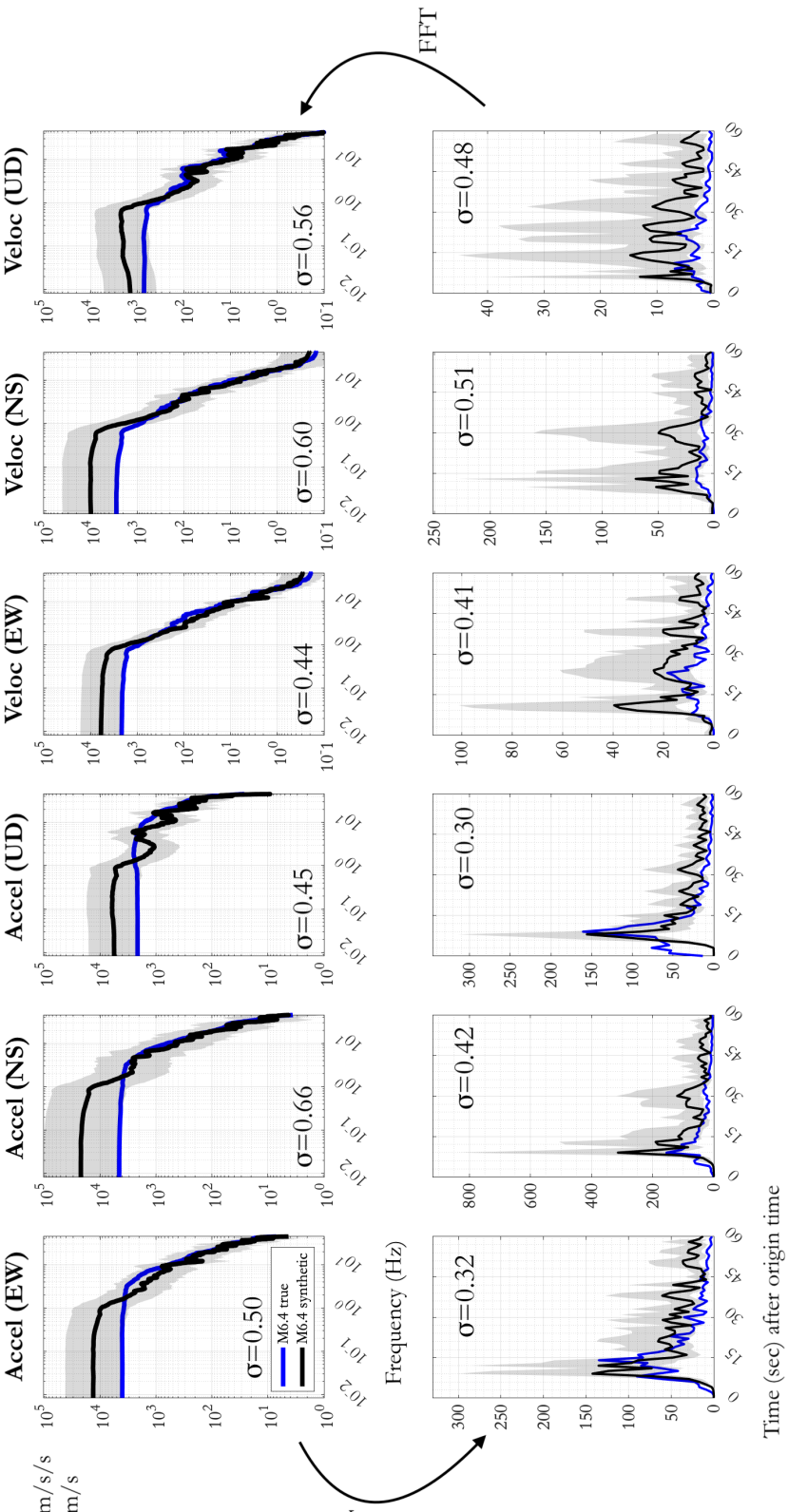


Figure 4.6. Synthetic spectra (in black) for a potential M6.4 earthquake are generated using available waveforms (in red) at Station SRT (HNE, HNN, and HNZ). This is done by applying the transfer function to the raw waveforms that are available for M5.36 earthquake. The top row compares the true and synthetic spectra, and the bottom row compares the acceleration and velocity waveform envelopes. σrefers to the error band needed to represent at least 68% of the true observed envelopes.

4.6 Application to Past Real Earthquakes

Applying the transfer function, defined in Eq. 4.8, scales the spectra of the earthquakes available in the original catalog to spectra of earthquakes of different magnitudes that are unavailable. Once more waveforms are generated to create a sufficient database to search through, they are transformed to ground motion envelopes of 1-second windows. Then, just as before, the posterior probability, defined in Eq. 4.1.1, is maximized to find the best earthquake source parameter estimates. For simpler computations, instead of maximizing the posterior probability, the SSR, defined in Eq. 4.2, is minimized. The following section looks at the application of the extended catalog search to real earthquakes, particularly those that are frequently missed or misidentified by the current EEW system. They are the same ones used in Chapter 3 to assess the performance of the grid search. The extended catalog search finds magnitude and location estimates as well as their corresponding error bands about the best-fitting envelopes. Again, the error bands provide a visual sense of the best-fitting envelopes at a given level of confidence; they show how much confidence can be given to the corresponding magnitude estimates. Because the extended catalog search is intended for real-time EEW application, expected warning times are also calculated by applying Eq. 4.11. In this equation, it is assumed the amount of time for data to travel between stations, processing centers, and users is relatively much smaller, essentially negligible, than the time it takes for parameter estimation. The warning time is the time it takes for strong shaking from the S-wave to arrive at a specified region.

$$T_{warn} = T_S - T_{param} - T_{transit} \approx T_S - T_{param} \tag{4.11}$$

where T_{warn} is the expected warning time, T_S is the S-wave arrival time at a given region (from a lookup table by Cua), T_{param} is the time it takes for algorithm to collect data and find parameter estimates of the earthquake source, and $T_{transit}$ is the time it takes for data to travel between stations, processing centers, and users.

4.6.1 2020 Northern coast offshore event

One of the challenges of EEW is detecting and identifying large offshore events, as mentioned in Chapter 1. In fact, most recently between 2014 and 2016, E2 missed 213 M≥3 earthquakes, in which the majority of them were offshore or in areas without dense station coverage (Chung et al. 2019). Locating offshore earthquakes is infamously known to be difficult due to poor azimuthal seismic ray-path coverage and sparse station spacing around the epicenter (Chung et al. 2019). Offshore events are frequently missed, but if the system manages to detect them, it still takes a long time to issue the first alert due to the lack of stations between the epicenter and mainland and the requirement of ElarmS to have at least four triggered stations. Therefore, one of the goals of the extended catalog search is to find parameter estimates corresponding to small error bands with fewer than four stations. The intension is to shorten the time it takes to find parameter estimates, which would increase the warning time for nearby regions.

On March 09, 2020 at 02:59:08 UTC, a M5.8 offshore event occurred near Petrolia, CA. This event should not have been a surprise, as earthquake history shows at least ten M>5 earthquakes in the region in the past 20 years. In this type of event, time is of the essence, especially with the first P-wave arriving 14 seconds after the origin time (see Table 4.1). Waiting for more stations to trigger would jeopardize warning time for mainland regions closest to the epicenter, regions that would feel the strong shaking first. Therefore, the stations considered in this analysis are the same first three to be triggered: 89101, KCO, and KCT. As seen in Fig. 4.8, the extended catalog search finds matches to the incoming observed acceleration, velocity, and displacement envelopes. From the minimization of the SSR, the chosen cataloged event is the M4.5 event from July 25, 2018 05:06:06 UTC. The waveforms from this cataloged event are eventually scaled to M5.8 to best fit the incoming observed envelopes. In comparison to the best-fitting Cua-Heaton envelopes found by the grid search, the shape of the envelopes is much more specific due to specific site and path effects. The fits found by the extended catalog search are stronger especially for the displacement.

To mimic a real-time analysis as the current EEW system, the extended catalog search updates the fits and the corresponding magnitude estimates as data becomes more available with time (see Fig. 4.7). The error bands are relatively large for the first 3 seconds;

therefore, the magnitude estimates found at the first 3 seconds after the origin time should be taken with a grain of salt. The initial magnitude is M5.3 and is of low confidence. However, 4 seconds after the origin time, the error bands decrease substantially, meaning the magnitude estimate of M5.7 has higher confidence. With time, the magnitude estimate eventually approaches the true M5.8. The location estimate is set constant throughout as the epicenter of the cataloged event, which is approximately 15 km from the true location of the observed M5.8 event.

In reality, the current EEW system misidentifies this M5.8 event, leading to a missed alert, and FinDer simulation takes 35 seconds to lock in at a magnitude estimate, ultimately underestimating it to M5.4. The extended catalog search manages to avoid these pitfalls: missed events, delay of accurate parameter estimates, and misidentification of the earthquake magnitude. Shown in Fig. 4.7, the extended catalog search estimates M5.8 at 22 seconds after the origin time, which is more accurate than the current system by 0.4 units and faster by approximately 13 seconds.

Fig. 4.9 shows the expected warning times different regions would receive using data available at different times after the origin time. For the levels of shaking, the resulting parameter estimates would have to be converted to MMI, using existing relationships such as those by Wald.

The solutions found by the extended catalog search 20 seconds after the origin have higher confidence than those found by the grid search. If the alerts are sent at this point in time, then the blind zone is virtually eliminated for the users in the onshore region.

Table 4.1. Triggered stations from the 2020 Northern coast offshore event with corresponding P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	Latitude(°N)	Longitude (°W)	P-wave arrival (sec after origin time)
CE.89101	40.3250	124.2877	14
NC.KCO	40.2567	124.2660	14
NC.KCT	40.4756	124.3375	14

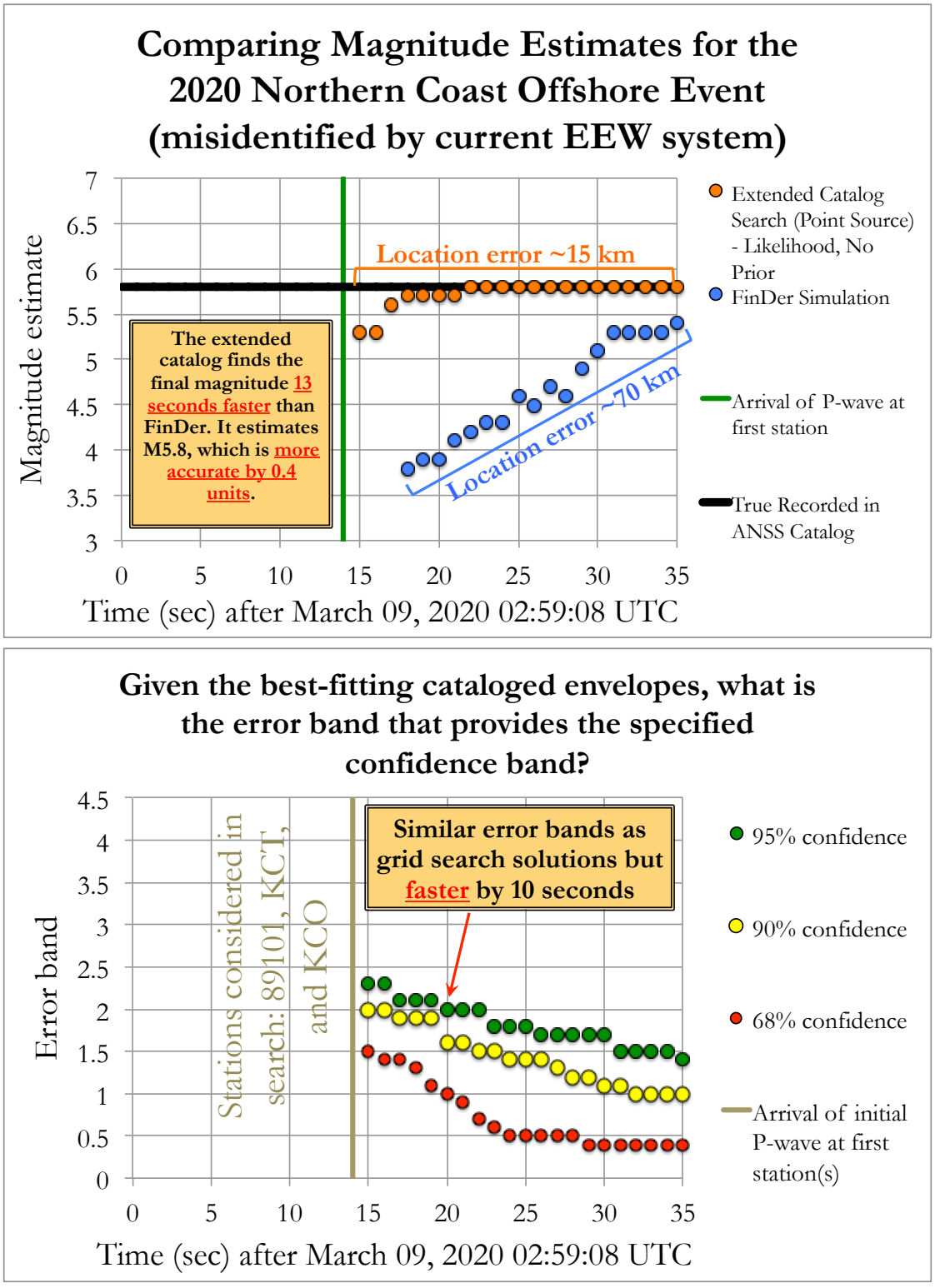
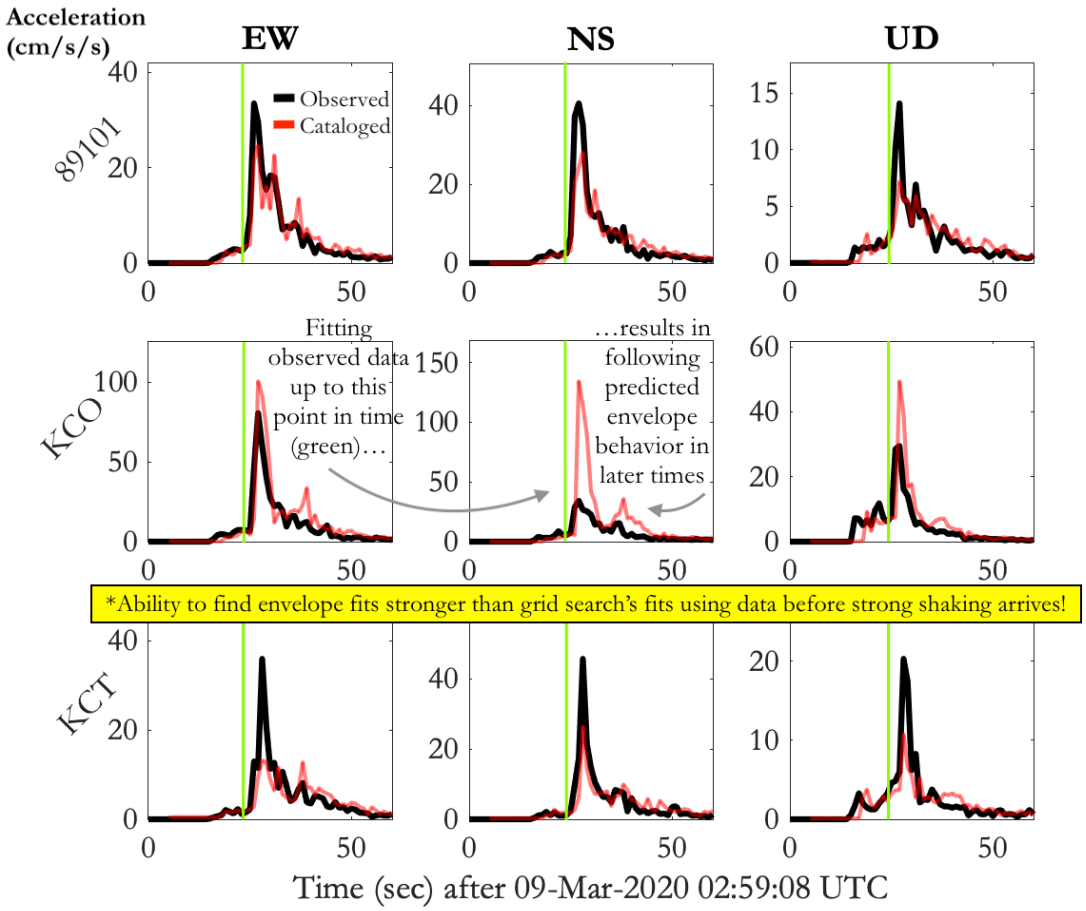
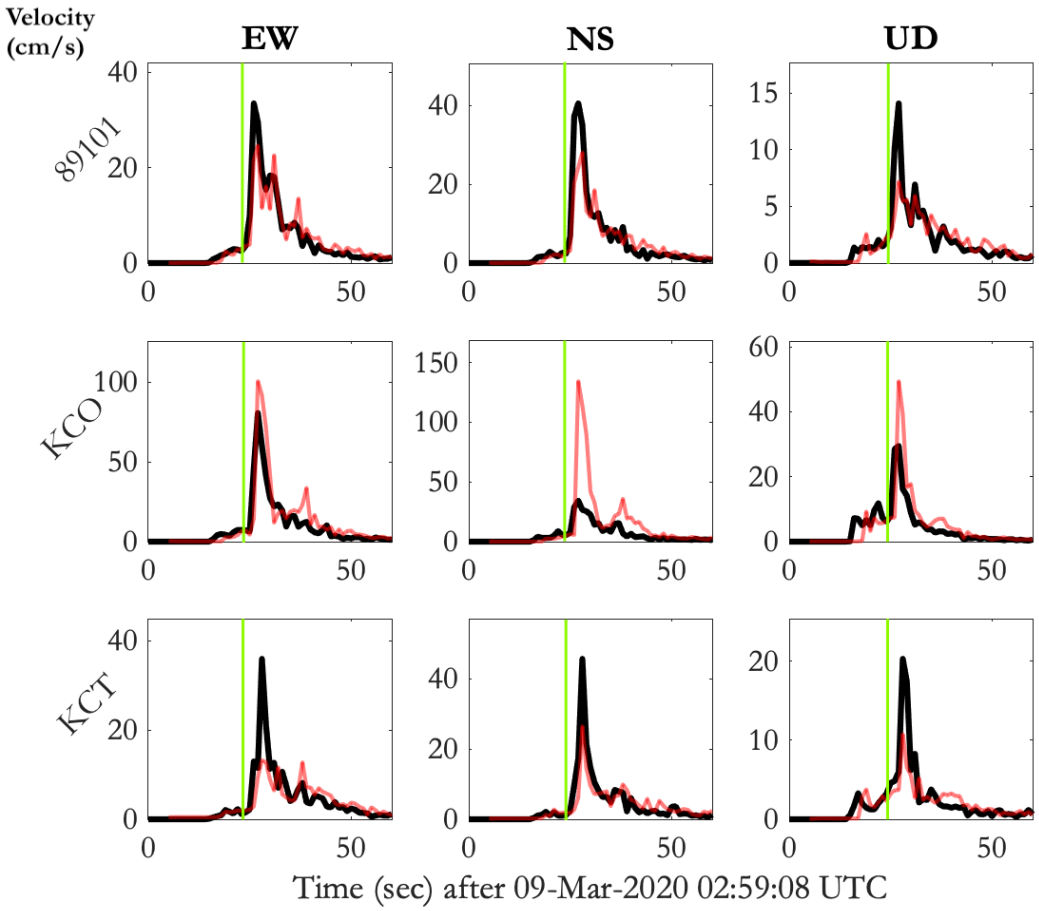


Figure 4.7. Extended catalog search magnitude estimates for the 2020 Northern coast offshore event. Along with magnitude estimates, error bands needed for 95%, 90%, and 68% confidence bands are plotted.



(figure continues next page)



(figure continued on next page)

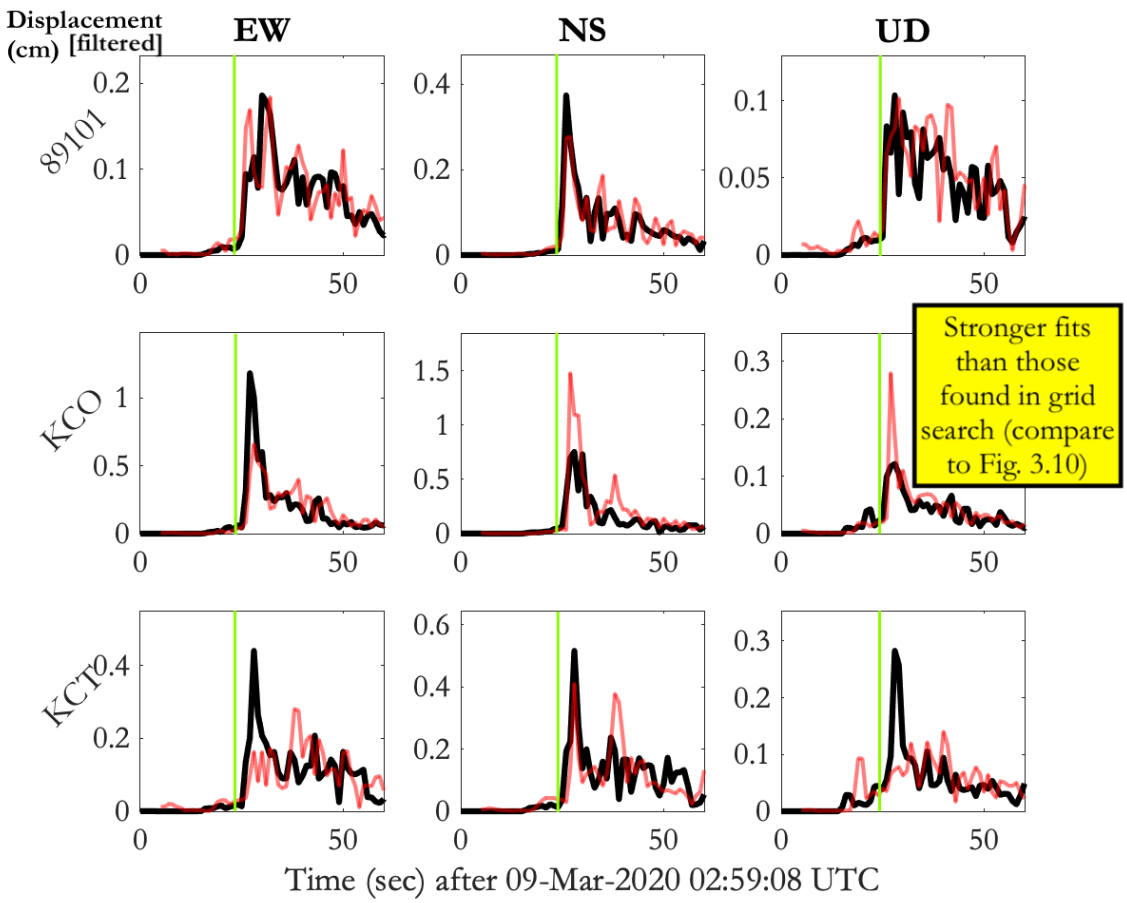


Figure 4.8. Comparing the best-fitting cataloged (in red) and incoming observed envelopes (in black) for the 2020 Northern coast offshore event. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and displacement are also labeled accordingly.

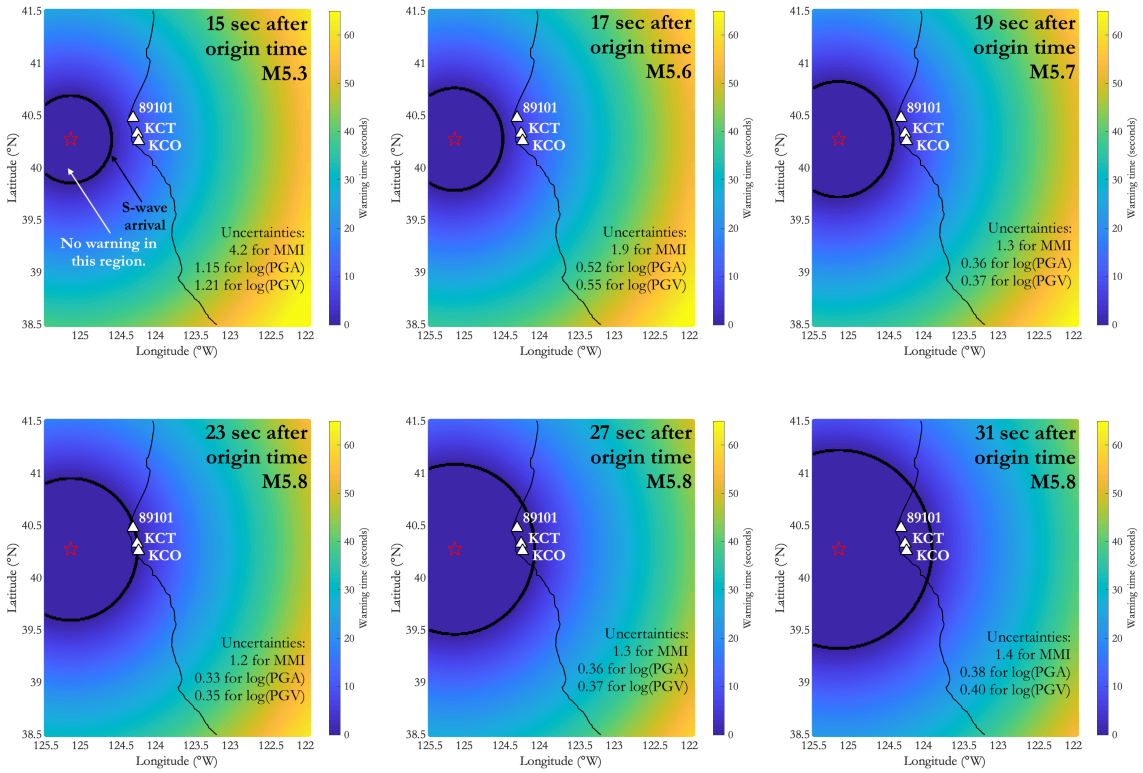


Figure 4.9. Warning times for regions near the epicenter using ground motions from the 2020 Northern coast offshore event. In each subfigure, the black circle represents the arrival of S-wave. The magnitude estimate corresponding to the time it takes to find it is written in black at the top right corner. The corresponding error bands (for MMI, logarithmic acceleration, and logarithmic velocity) are written at the bottom right corner. Stations are represented in white triangles, labeled accordingly. The warning times are represented in the colorbar on the right, with blue being the shortest and yellow being the longest warning time.

4.6.2 2020 Lone Pine foreshock-mainshock pair

For the extended catalog search to work properly in finding accurate parameter estimates with small error bands, the catalog itself must include waveforms resembling the incoming ground motions. A foreshock occurring close by in space and in time most likely satisfies this criterion. Foreshock-mainshock pairs hold a special advantage over other types of earthquakes because the extended catalog search does not necessarily have to look back over a large amount of years in earthquake history, saving computation space and efficiency. A specific foreshock-mainshock pair to look at is the 2020 Lone Pine sequence.

Approximately 41 hours before the M5.8 mainshock, earthquake history shows a M4.62 foreshock that occurred 0.7875 km away. In this study, the waveforms from the closest triggered stations, listed in Table 4.2, are observed. Initially, there is scatter in the error bands (see Fig. 4.10). Approximately 11 seconds after the origin time, the error bands start to decrease and remain relatively small. The corresponding magnitude estimates are M5.7. Obtaining this accurate estimate 11 seconds after the origin time is 7 seconds faster than the current EEW system. The extended catalog search constrains the location estimate at the epicenter of the foreshock, which has a location error of approximately 0.7875 km, while the current EEW system estimates the location with an error of ~30 km.

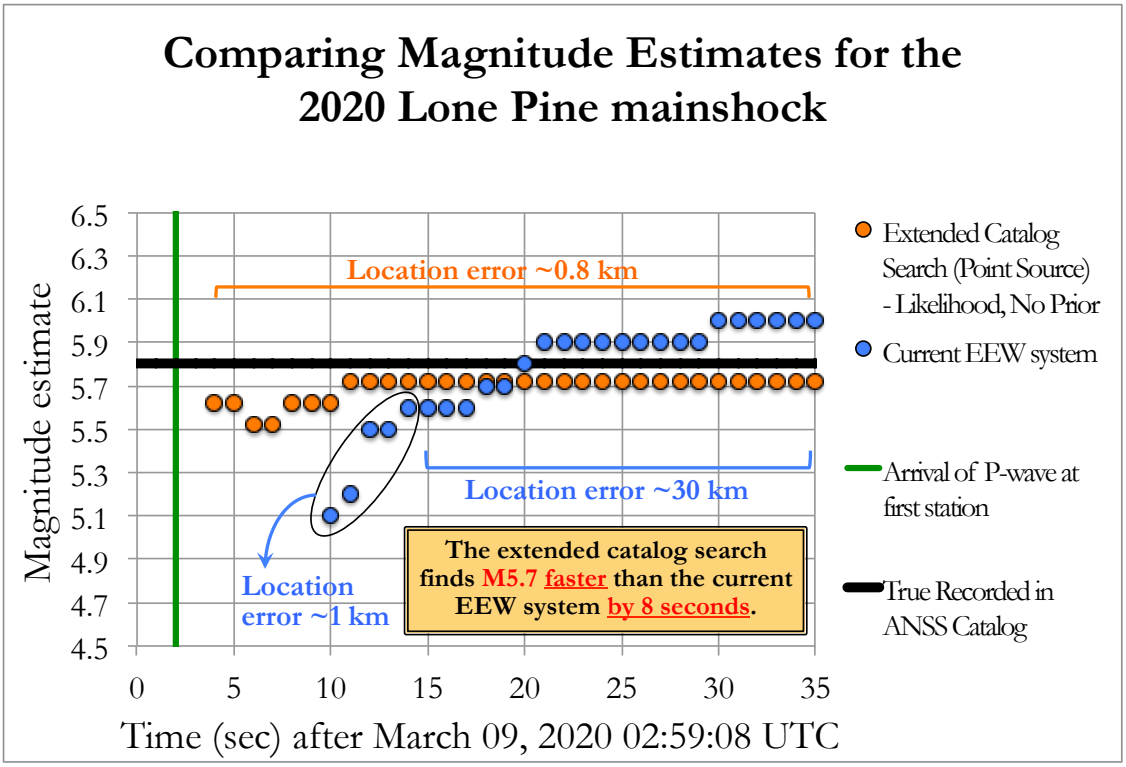
Seen in Fig. 4.11, the envelope fits are weakest in the first station, CWC, which is located approximately 9 km from the epicenter. As the station-to-epicenter distance increases, the error band generally tends to decrease. This behavior is connected to the assumption made at the very beginning of this chapter: the point source characterization of the earthquake. Point source characterization is only valid for stations located at greater distances than the fault length. If the fault length is greater than 9 km, the invalid characterization of the earthquake may explain the discrepancy in the envelope fits at the first station. Another possible explanation is the difference in focal mechanism, which is not considered in this study of the extended catalog search.

The extended catalog search is able to recognize the incoming ground motions rather quickly. The initial magnitude estimate is M5.6 with just 2 seconds of P-wave data. The location estimate is constrained at the foreshock's epicenter, which is about 0.8 km from the true mainshock epicenter. The magnitude estimate eventually grows to M5.7 at 11 seconds after the origin time. This is faster than the current system by 8 seconds. In reality, the

current EEW system detects the mainshock, but with tradeoffs between the location and magnitude estimates. It overestimates the magnitude to M6.0 and locates the epicenter with an error of 30 km. It also takes at least 20 seconds for estimates to converge and lock in. The performance of the extended catalog search is successful in both speed and accuracy in obtaining the parameter estimates.

Table 4.2. Triggered stations from the 2020 Lone Pine mainshock with corresponding P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	**Latitude(°N)**	**Longitude (°W)**	**P-wave arrival (sec after origin time)**
CI.CWC	36.4399	118.0802	2
CI.CGO	36.5504	117.8029	4
CI.WMF	36.1176	117.8549	7
CI.DAW	36.2715	117.5921	7



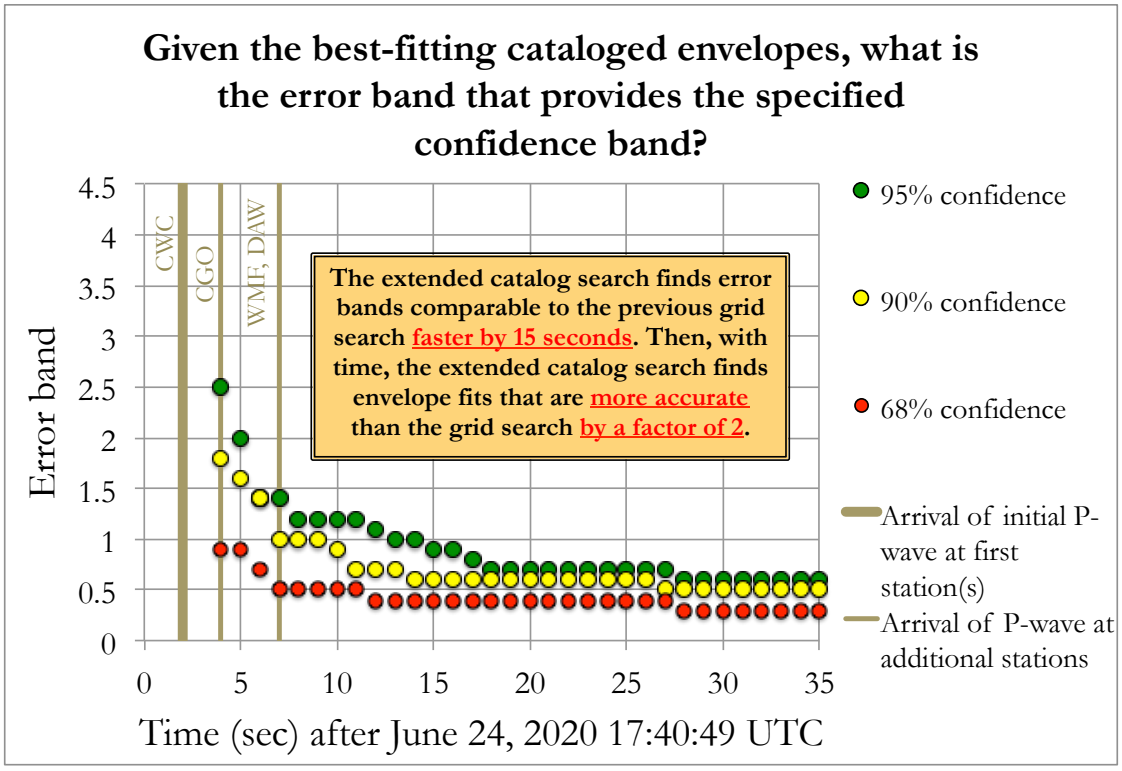
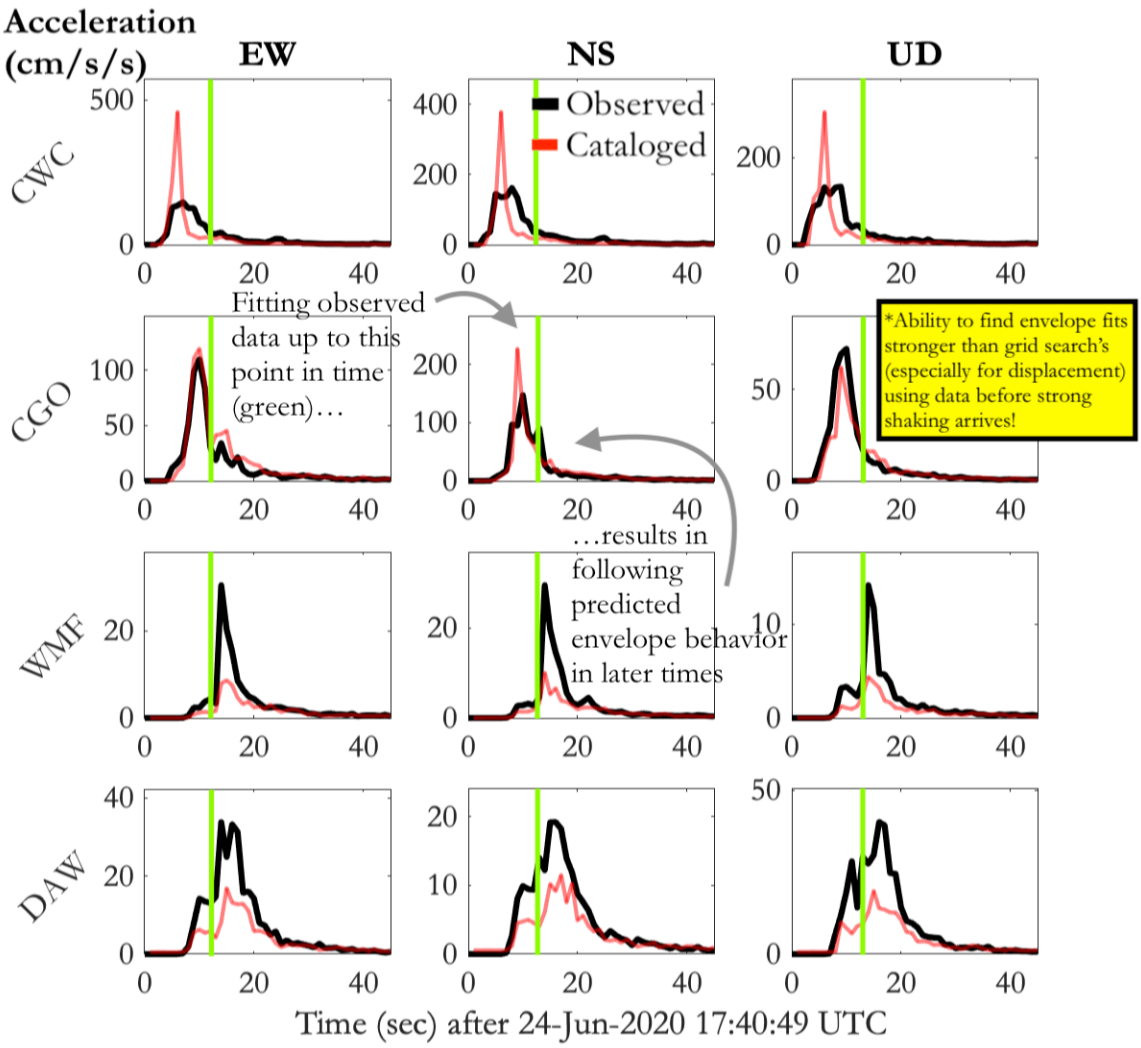
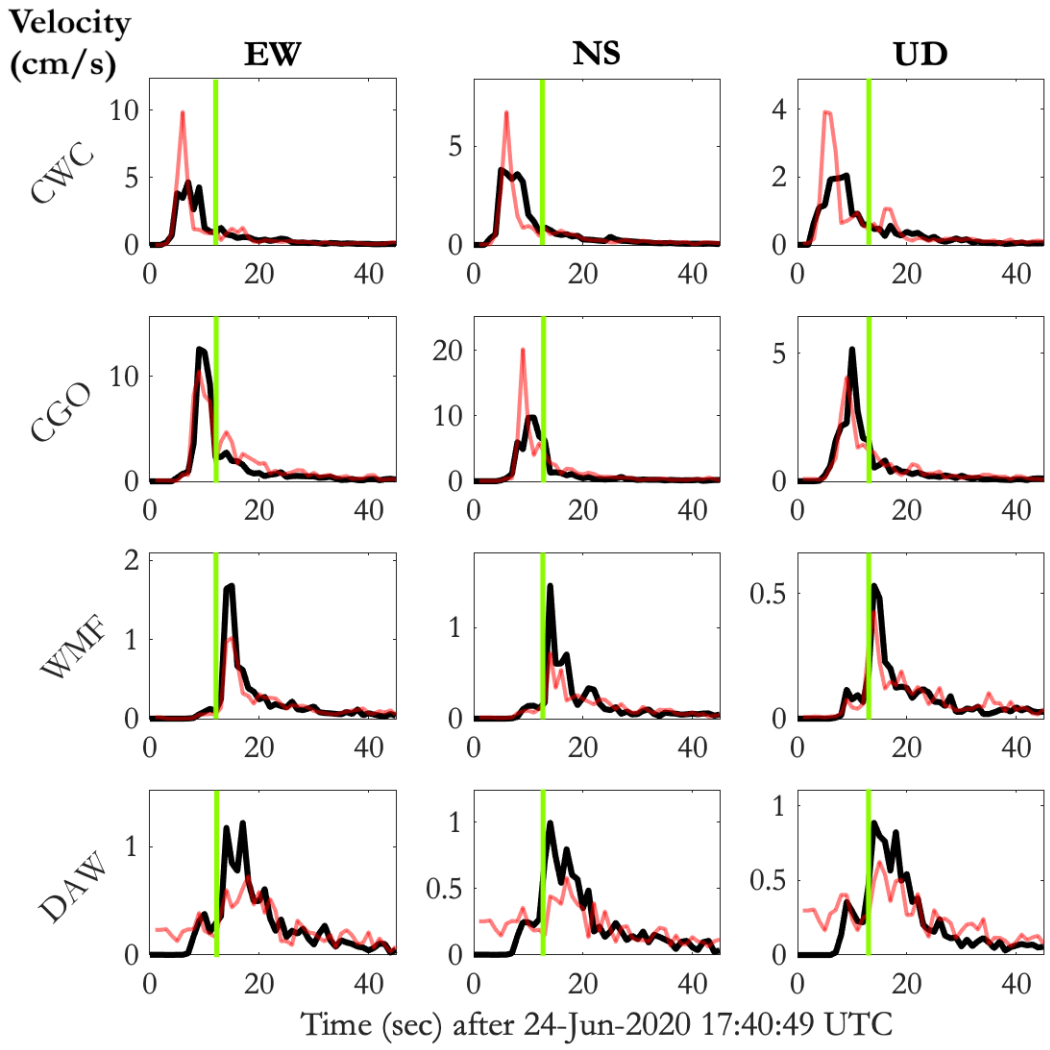


Figure 4.10. Extended catalog search magnitude estimates for the 2020 Lone Pine mainshock. Along with magnitude estimates, error bands needed for 95%, 90%, and 68% confidence bands are plotted.



(figure continues on next page)



(figure continues on next page)

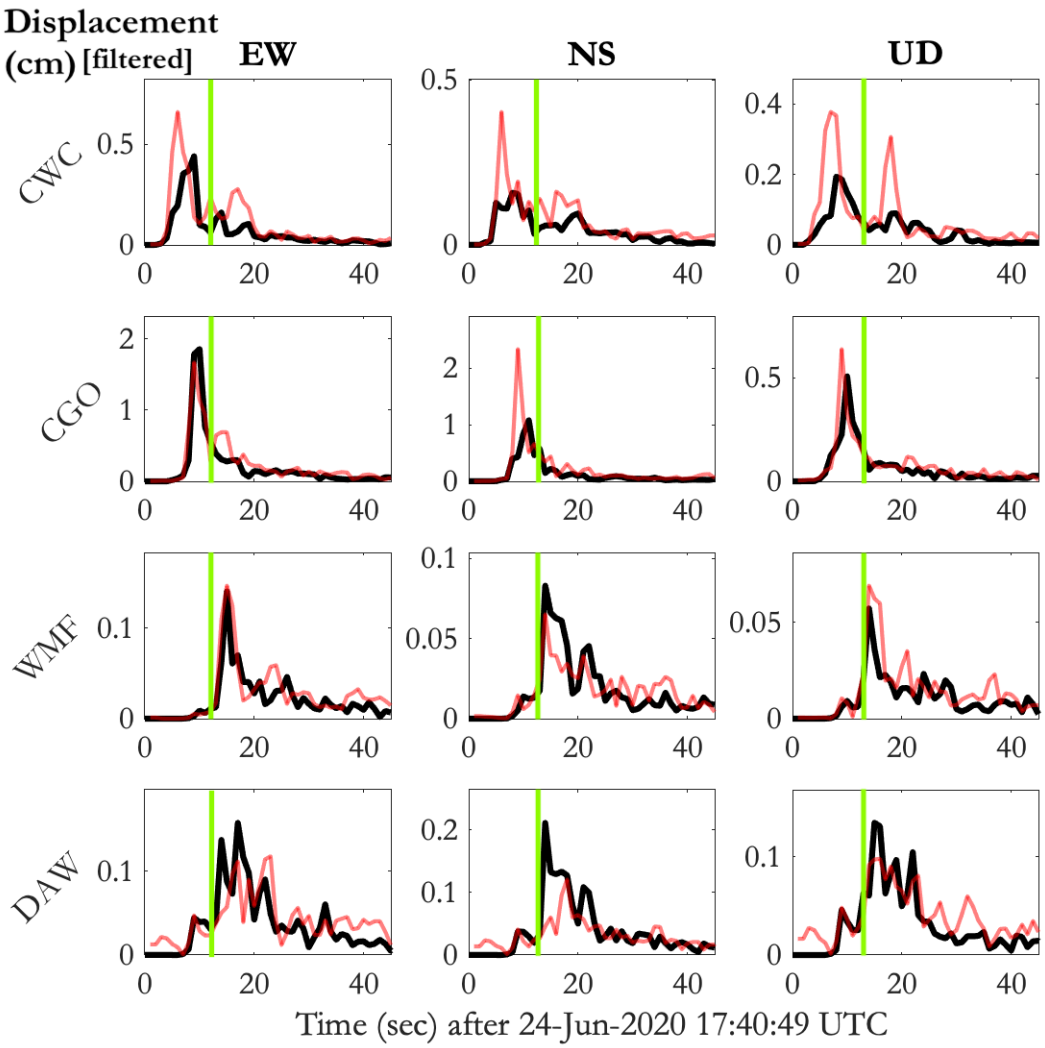


Figure 4.11. Comparing the best-fitting past cataloged (in red) and incoming observed envelopes (in black) for the 2020 Lone Pine mainshock. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and displacement are also labeled accordingly.

4.6.3 2019 Ridgecrest sequence

So far, the two examples of real earthquakes generally show point source characterization is valid. However, this may not be the case for larger earthquakes (M>6.5). The following example is the 2019 Ridgecrest sequence, where the M6.4 foreshock was followed a M7.1 mainshock. This particular foreshock-mainshock pair is spaced apart in time by nearly 34 hours and in space by 11 km. The mainshock ruptured bilaterally in the NW-SE direction for a cumulative length of ~65 km (Ross et al. 2019). Some discrepancy in the envelope fits is expected due to this long rupture length. If the station-to-epicenter distance is less than this rupture length, a point source characterization may not be valid and a finite fault characterization is needed. However, for this particular study in this chapter, point source is assumed. In Chapter 6, the use of additional templates considering complex sequences (i.e. combinations of subevents at different time delays) may reduce the error bands. The use of templates considering complex sequences especially helps explain discrepancies in the fits for stations close to the rupturing fault, such as stations China Lake (CLC) and Christmas Canyon China Lake (CCC), where the ground motions are amplified due to the propagation of multiple ruptures.

Using the waveforms from the stations listed in Table 4.3, the magnitude estimates are initially underestimated to M5.4. Then, they jump up to M6.4 approximately 8 seconds after the origin time. The magnitude estimates ultimately approach M6.9 with the location estimate constrained to 11 km from the true epicenter. In reality, the current EEW system underestimates the final magnitude to M6.3 at 21 seconds after the origin time. In comparison, as shown in Fig. 4.12, the extended catalog search estimates M6.4 faster than the current EEW system by 13 seconds. It is also able to identify the eventual growth of the magnitude to M7.0, which is 0.7 units more accurate than the current EEW system.

Table 4.3. Triggered stations from the 2019 Ridgecrest mainshock with corresponding P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	**Latitude(°N)**	**Longitude (°W)**	**P-wave arrival (sec after origin time)**
CI.CLC	35.8157	117.5975	1
CI.TOW2	35.8086	117.7649	3
CI.SRT	35.6923	117.7505	3
CI.WRC2	35.9479	117.6504	4
CI.SLA	35.8909	117.2833	5
CI.LRL	35.4795	117.6821	5
CI.CCC	35.5249	117.3645	6

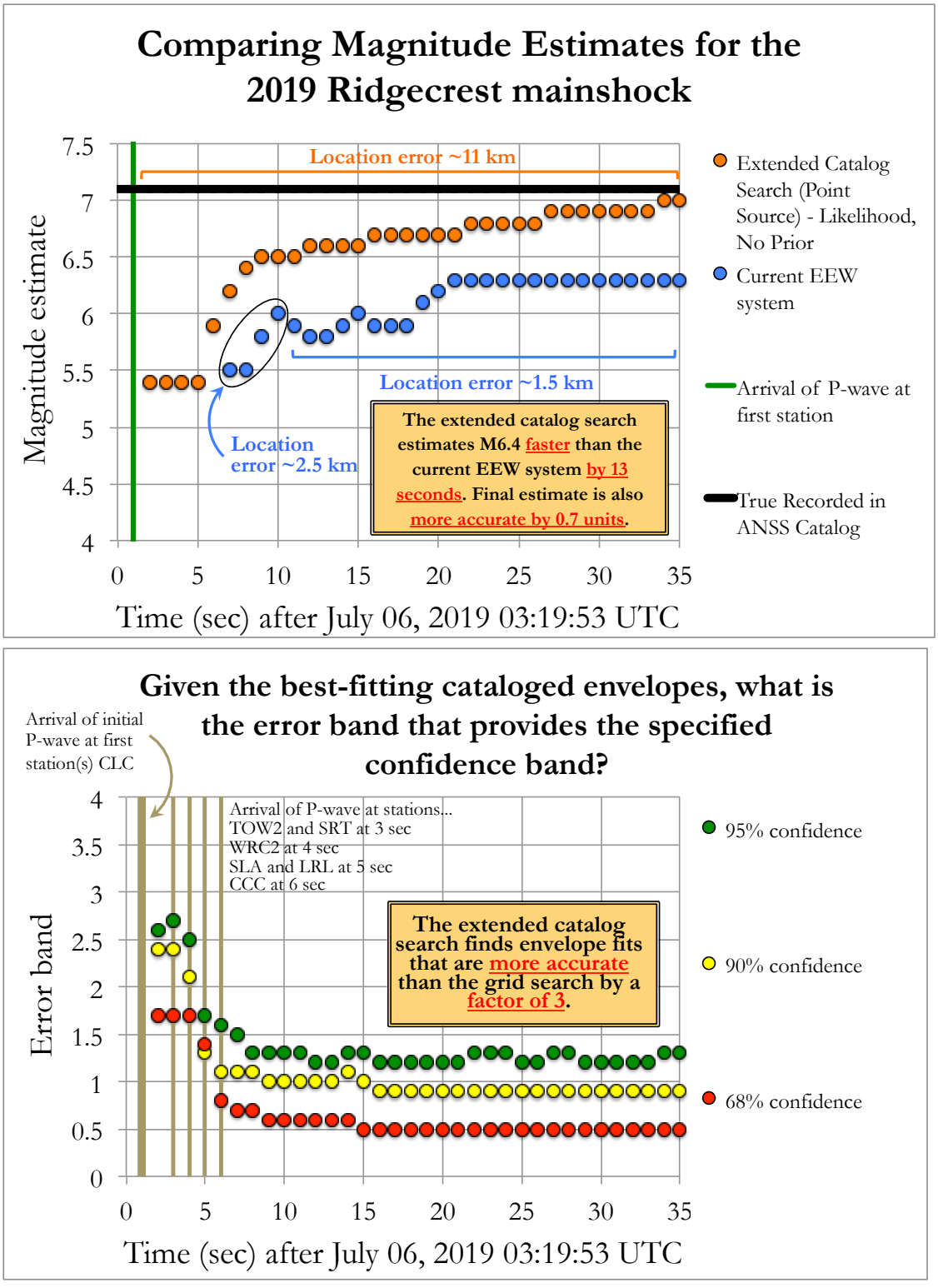
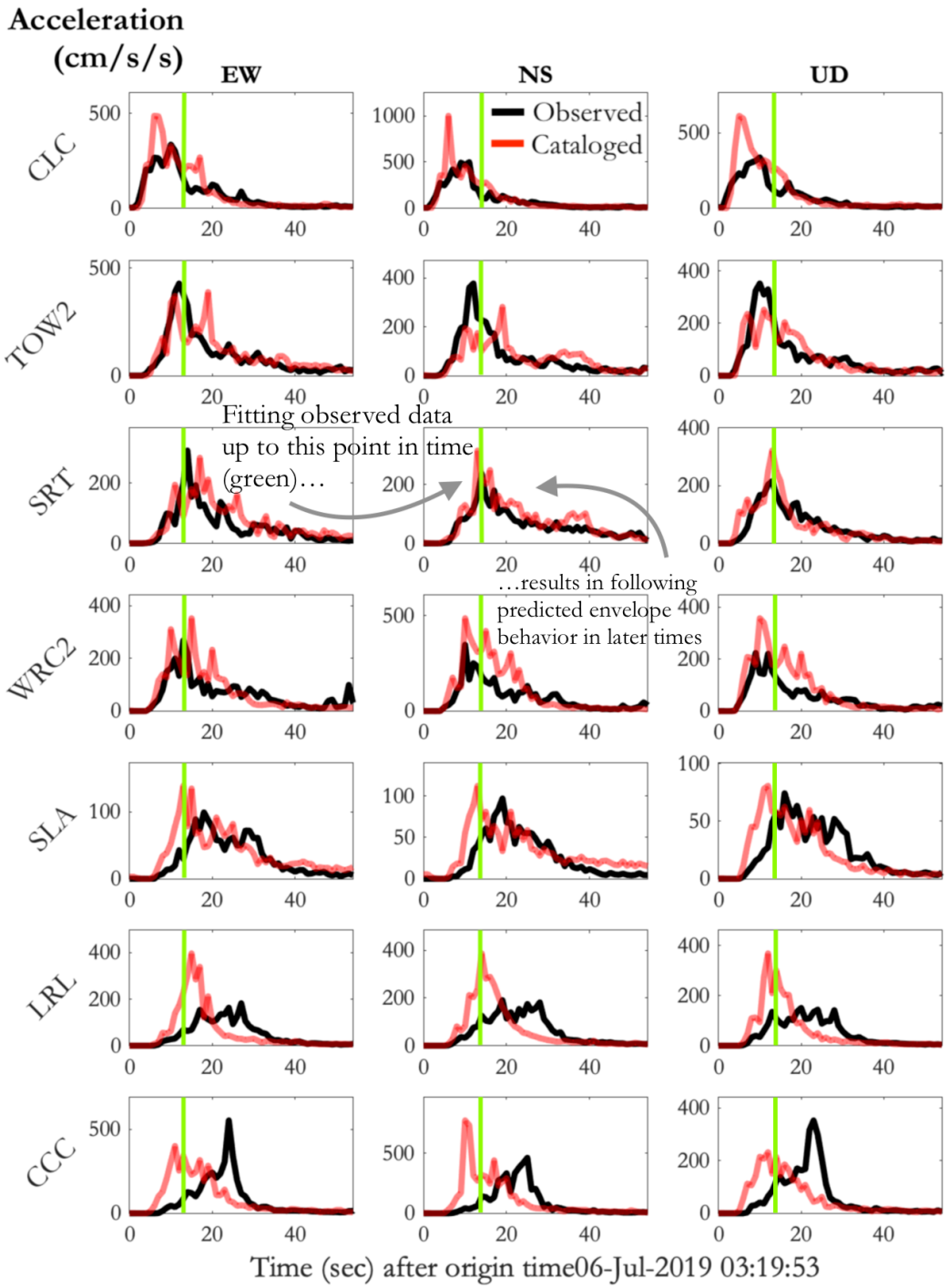
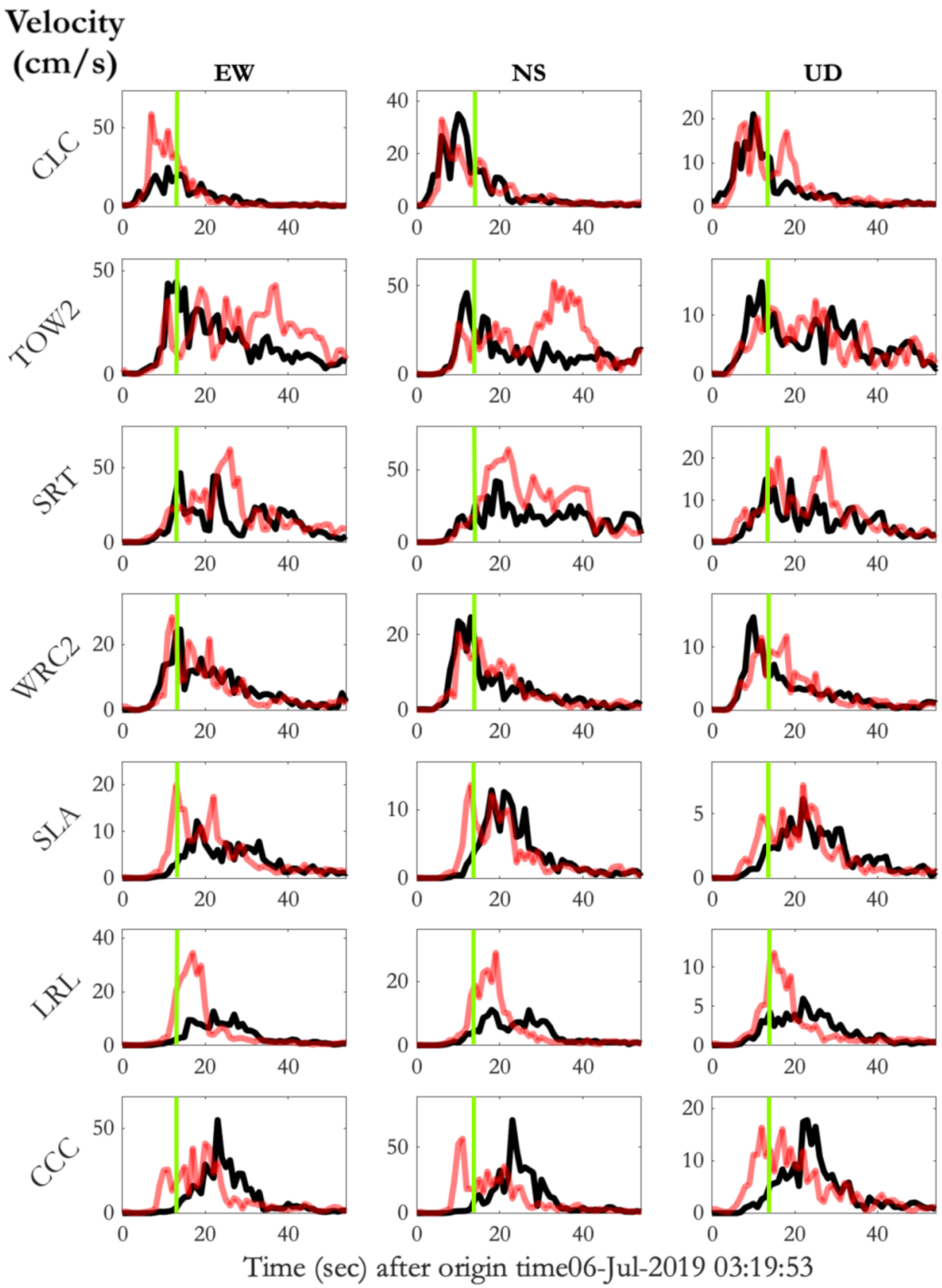


Figure 4.12. Extended catalog search magnitude estimates for the 2019 Ridgecrest mainshock. Along with magnitude estimates, error bands needed for 95%, 90%, and 68% confidence bands are plotted.



(figure continues on next page)



(figure continues on next page)

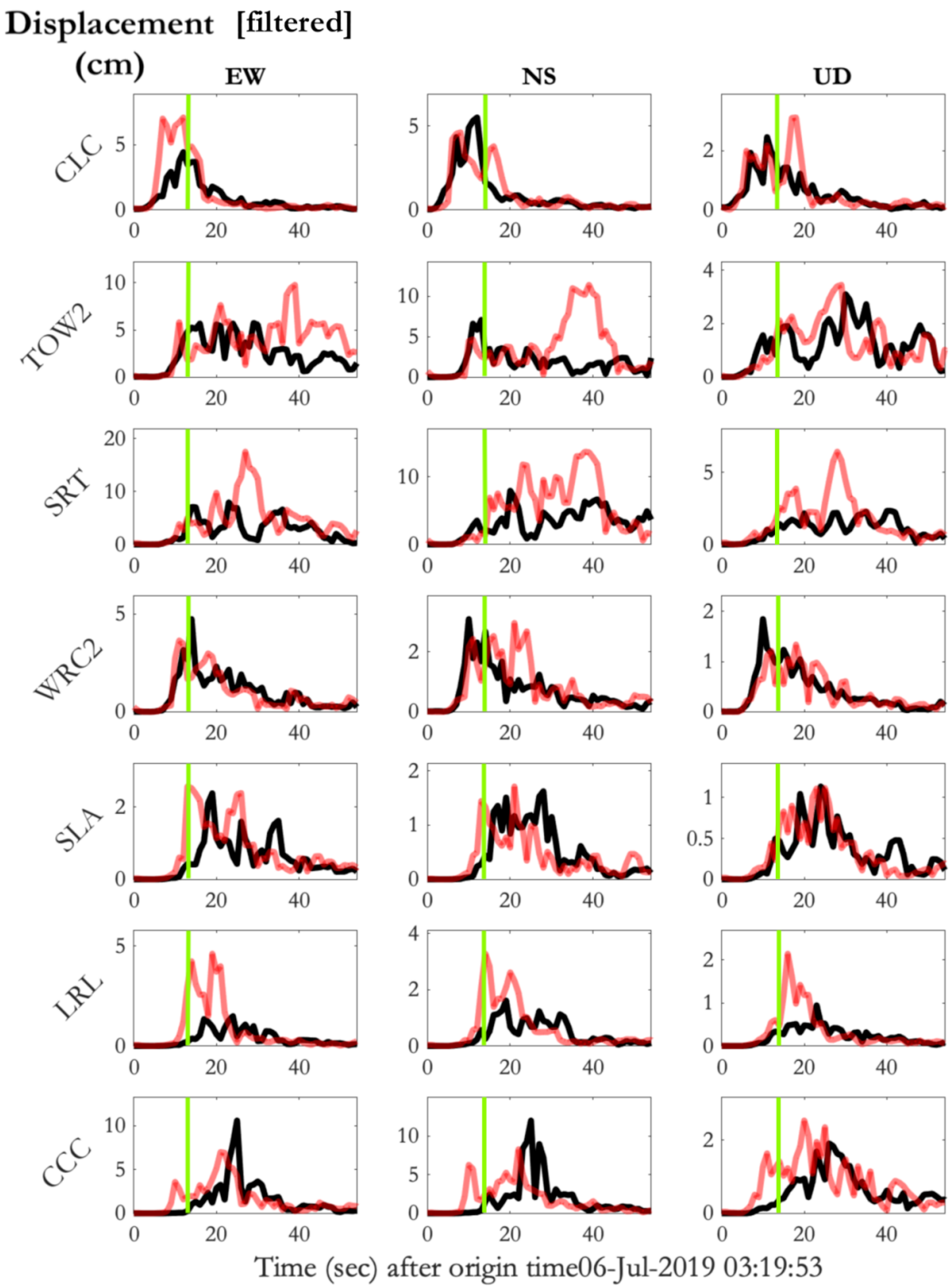


Figure 4.13 Comparing the best-fitting cataloged (in red) and incoming observed envelopes (in black) for the 2019 Ridgecrest mainshock. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and displacement are also labeled accordingly.

4.7 Summary

The simplest form of the extended catalog search is assuming uniform prior and considering only the waveform-based likelihood. In other words, maximizing the posterior probability is essentially the same as maximizing the likelihood. To ensure the extended catalog search finds envelopes from the past that resemble the incoming observed ones, a simplified spectral scaling model is applied. A commonly used standard model by Brune is used, and it characterizes the earthquake as a point source and assumes earthquake spectra obey certain simple similarity laws. This model allows synthetics to be produced from available waveforms of real earthquakes. It is important, however, for the extension of the catalog to still have a realistic, natural distribution of magnitudes that follows the Gutenberg-Richter law.

As time goes, catalogs are updated with additional earthquake data, meaning the distribution of earthquake parameters (i.e. magnitude, epicenters) continues to diversify. This chapter only considers past events from years 2015 to 2020. This constraint can be modified to allow waveforms from further back in time. Therefore, the distribution of epicenters in Fig. 4.2 would include more variations if the catalog considers more years, such as 2000 to 2020. This reduces the chance of cases where the catalog does not hold waveforms that resemble incoming observed ground motions.

Based on the applications to real earthquakes, the extended catalog search performs best for sequences where plenty events that are nearby in space and in time are available. When earthquakes occur in regions of low seismicity, the extended catalog must look further back in earthquake history to ensure a variety of waveforms is used. For instance, the extended catalog looks back a few hours for the 2020 Lone Pine and 2019 Ridgecrest sequences, but it looks back a few years for the 2020 Northern coast offshore event. For regions of very low to no seismicity, the final solution will take the grid search estimates. This is the value of having two methods in the search algorithm; they serve as checks to one another or as replacements if solutions to one method are unavailable.

Seen in the results, generally, the error bands of the parameter estimates are smaller for the extended catalog search compared to those for the grid search. This means the solutions found by the extended catalog search correspond to stronger envelope fits, so alerts based on their estimates would have higher confidence. Because the error bands are

relatively smaller, even for initial time points, the extended catalog search may not be significantly impacted by prior information. The grid search, however, may benefit more by the use of prior information, as shown in Chapter 8.

A real-time application of the extended catalog search requires an acceptable threshold for the error bands to satisfy before sending an alert to users. To avoid false alerts, the current EEW system waits until four stations are triggered, but to save time, the extended catalog search aims to find accurate parameter estimates with reduced uncertainties with only one to two stations. This benefits regions with sparse station coverage. To send accurate alerts with only one to two stations, the error bands would have to satisfy an acceptable threshold. However, a test sweep is required to calculate the acceptable threshold, which is beyond the scope of this thesis.

For real-time application, the extended catalogs will be pre-determined based on the following criteria. The basis is to look back 1 month in earthquake history, which works sufficiently for regions of high seismicity. For regions of moderate seismicity, the extended catalog is built looking farther back in time (i.e. 3 months, 1 years, 10 years, and so on). It is important to continuously update the catalog with time as new events occur.

So far, both extended catalog search and grid search find the best-fitting envelopes by brute-force. They exhaustively search through every single available envelope. However, exhaustive searches can be time-consuming and are practical when the size of the catalog is relatively small (< 1,000 records). Chapter 5 focuses on optimizing the search time for Method II, the extended catalog search. In the same three specific earthquakes that are studied throughout this thesis, between 90 and 3,000 envelopes per station are used. It may be necessary for the catalog to have >10,000 or even 100,000 records to ensure a variety of envelopes are represented in the search. Otherwise, an accurate match may not be guaranteed.

Once the database grows to > 1,000 records, a faster search method may be needed. A KD tree nearest neighbor search is one method that has the potential to search a fraction of the whole database without risking accuracy. It is based on the idea of Internet searching. Because this chapter uses relatively small catalogs, using a brute-force search is permitted as it takes the same amount of searching time as a KD tree nearest neighbor search.

5 Optimizing Method II with KD trees

Methods I and II described in Chapters 3 and 4, respectively, use brute-force search, where the posterior probability is calculated using every available waveform in the catalog. Brute-force refers to a search that scans through every ground motion envelope in the extended catalog to find the one that best describes the incoming observed envelope. However, this choice of search procedure is only practical when the size of the total catalog is relatively small. Brute-force search time increases exponentially as the database of earthquake increases, which is devastating in EEW applications. For instance, brute-force search may be sufficiently fast when looking back only 2 days in earthquake history, such as in the 2019 Ridgecrest sequence, but it may not be quick enough when looking back multiple years, such as in the 2020 Northern coast offshore event. Though search time may increase, looking back farther into earthquake history and having a larger database to work with ensures the extended catalog search works robustly. Therefore, another method of searching is introduced in this chapter: a KD tree nearest neighbor search (Bentley 1975). Brute-force search remains for Method I to ensure an optimal solution is not missed. However, Method II can be optimized using the suggested KD tree nearest neighbor search.

5.1 Introduction

The goal is to collect and store information in a way where the retrieval of relevant information can be done in a prompt manner, especially when databases grow large. To do so, the main focus is on the initial step of building a specific data structure that efficiently represents the data. The suggested data structure is tree-based. Once the structure is built, a search can be conducted to find the relevant information that satisfies the user's requests. Many Internet searching services, like Google Maps and Facebook's facial recognition, use tree-based structures to do fast searches. Tree-based structures allow searches to avoid exhaustive, brute-force observation of every component in a database. Instead, a fraction of the database is searched, and the time spent on searching large databases is reduced.

5.2 Re-structuring the Format of the Dataset

A KD tree is a re-organized storage of data points for the purpose of fast retrieval. The K refers to the total amount of dimensions the dataset is characterized by. We refer to the K dimensions to construct the tree. While a standard binary tree looks to only one component for all levels of the tree, a KD tree is constructed using K components that are cycled for each level of the tree. For instance, for a database of K-dimensional points comprising of $(x_1, x_2, \dots, x_K)$ coordinates, a tree is constructed by cycling $x_1, x_2, \dots, x_K, x_1, x_2, \dots, x_K, \dots$ for consecutive levels.

In application to Method II, the extended catalog search, there are multiple KD trees, each with a different value for "K" and each made to compare with observed data of different lengths. Each KD tree comprises of data points characterized by coordinates $(x_1, x_2, \dots, x_K)$. As seen in Table 5.1, For each of the different KD trees, eight coordinates remain constant, which are the event information and the station information. The remaining coordinates are the recorded amplitudes of the ground motion envelopes (for time points $t = 0,1,2, \dots$).

Table 5.1. Format of the dataset in preparation for KD tree construction.

Dimension	Description of featured dimension	Note
1	Origin time of cataloged earthquake	Remain constant for every KD tree constructed
2	Magnitude of cataloged earthquake	
3	Latitude of cataloged earthquake	
4	Longitude of cataloged earthquake	
5	Latitude of station	
6	Longitude of station	
7	Channel (HN* refers to strong motion, HH* refers to broadband)	
8	Location term (--,10,00,01)	
9	Amplitudes of ground motion envelope recorded at station specified by dimensions 5-8 for K-8 seconds	Depends on the length of the ground motion envelope
…		
K		

5.3 Constructing the KD Tree

It requires initial effort to construct a KD tree before searching it. The general procedure in constructing a KD tree is to recursively subdivide the total dataset into subsets with respect to the median until the final subset consists of only one data point. As previously mentioned, the median is calculated while cycling through the coordinates $x_1, x_2, \dots, x_K, x_1, x_2, \dots, x_K, \dots$ for each successive level. The median of the data is used for each recursive subdivision to ensure the construction of a well-balanced KD tree (Brown 2020). To clearly understand the construction of a KD tree, it is best to construct the simplest case, which is using a 2D dataset $(K = 2)$. A sample dataset of size 10 is defined in Table 5.2. The two dimensions of this sample dataset are the amplitudes of the envelope at the first and second time points, denoted as E_1 and E_2. The 10 data points, each represented by a node in the tree, are labeled form A to J.

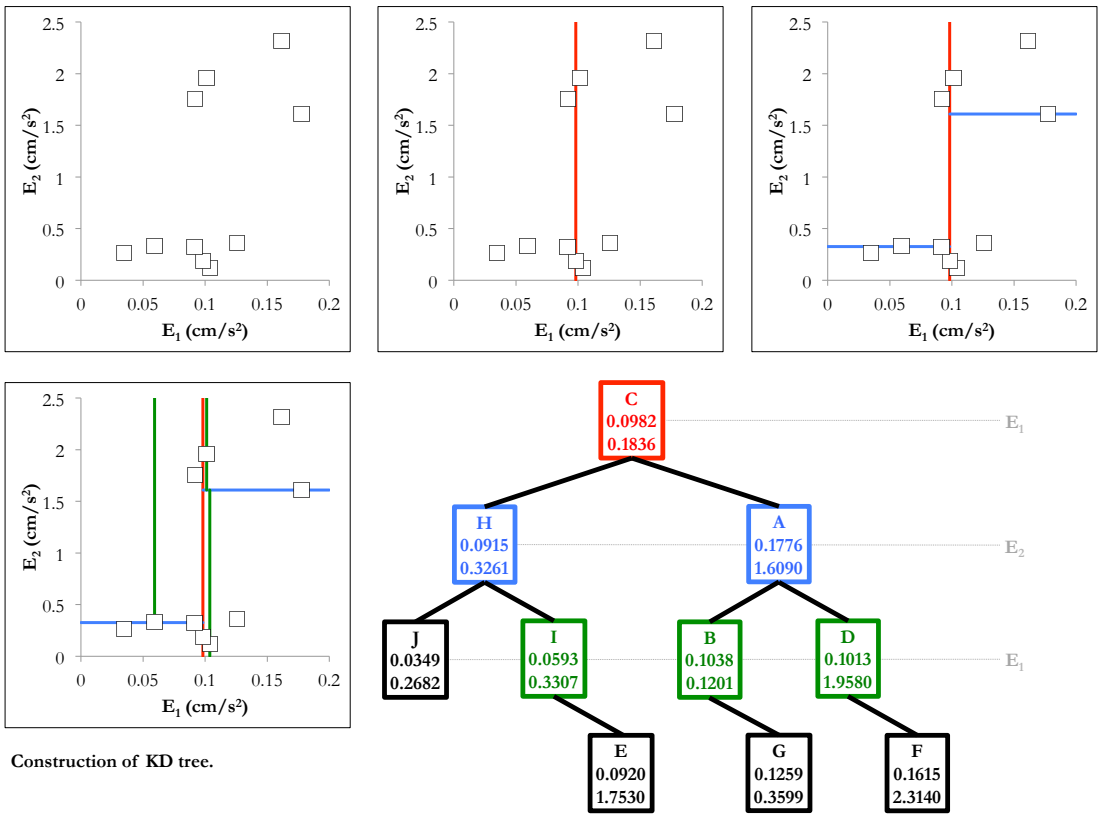
Table 5.2. Sample 2D dataset ($K = 2$).

	K = 2		**Parameters corresponding to specified data point**			
Data point/Node	E_1 (cm/s^2)	E_2 (cm/s^2)	Origin time (UTC)	Magnitude	Latitude (°N)	Longitude (°W)
A	0.1776	1.6090	2019/06/23 03:53:02.800	M5.60	40.2730	124.3000
B	0.1038	0.1201	2019/01/13 09:35:48.000	M4.10	40.3750	124.9930
C	0.0982	0.1836	2018/07/25 05:06:06.740	M4.50	40.3850	125.0520
D	0.1013	1.9580	2018/03/23 03:09:36.710	M4.70	40.4480	124.4910
E	0.0920	1.7530	2017/06/24 21:22:03.000	M4.00	40.2880	124.2990
F	0.1615	2.3140	2016/12/05 18:33:15.480	M4.35	40.2785	124.3860
G	0.1259	0.3599	2016/01/07 05:49:52.430	M4.31	40.2732	124.3395
H	0.0915	0.3261	2016/01/02 05:11:46.620	M4.44	40.3083	124.6872
I	0.0593	0.3307	2015/05/25 10:17:35.820	M4.34	40.6508	124.7120
J	0.0349	0.2682	2015/01/29 19:13:55.180	M4.25	40.3113	124.5892

A KD tree is represented in two ways. Both representations are necessary to clearly understand how the nearest neighbor search will be conducted, which is explained in detail later in this chapter. The first representation is a structure of nodes and branches. The second representation is a visualization of the points in partitioned spaces, or hyperplanes.

To construct a KD tree structure using the first representation with nodes and branches, the root node is defined. The data point assigned to the *root node* is found by taking the median of the total dataset with respect to one dimension, either E_1 and E_2. Choosing E_1, the data point C is assigned to the root node, as shown in red in Fig. 5.1, which is placed at the top of the tree. The rest of the data points are subdivided into subsets: to the left, if the E_1 coordinates are less than that of the root node's, and to the right, if the E_1 coordinates are greater than that of the root node's. The nodes for the second level of the tree, as shown in blue in Fig. 5.1, are found by taking the median of each subset, now with respect to the E_2. The nodes for the third level of the tree, as shown in green in Fig. 5.1, are found by taking the median of each subset with respect to E_1. At this point, the remaining subsets consist of only one data point, indicating the end of the process of subdivisions.

So far, the construction of a KD tree has been described only using nodes and branches. To construct a KD tree using partitioned spaces, the first step is to visualize how the data points are plotted in space (see boxes in Fig. 5.1). Just as before, the median is taken with respect to the E_1 coordinate. At this median, data point C, a line is drawn perpendicular to the axis corresponding to the E_1 coordinate, as shown in red in Fig. 5.1. This line splitting the space is defined, and referred to throughout this chapter, as the *hyperplane*. Next, the median is taken once more, but with respect to the E_2 coordinate, using the data points in the left side of the red hyperplane. This time, the hyperplane is drawn perpendicular to the axis corresponding to the E_2 coordinate, shown in blue in Fig. 5.1. The same is done with the subset right of the red hyperplane and with the rest of the remaining subsets, alternating the coordinate with each subdivision. Just as before, the construction is complete once the remaining subsets consist of only one data point.



Construction of KD tree.

Figure 5.1. Construction of a well-balanced KD tree where $K = 2$. Each node in this tree contains the coordinates (E_1, E_2).

5.4 Searching the KD Tree

Once the KD trees are constructed, a nearest neighbor search can finally be conducted. A nearest neighbor search, or similarity search, is a common technique in data mining and machine learning. An *exact* nearest neighbor search is applied, which is finding the single data point from the dataset that best fits the newly observed point, which is referred to throughout this chapter as the *query*. The data point that best fits the query point, denoted p^*, is the one that satisfies Eqs. 5.0.

$$p^* = \arg\min_{p \in S} d(q,p) \tag{5.0.1}$$

$$d(q,p) = \|q - p\|_2 \tag{5.0.2}$$

$$d(q,p) = d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{5.0.3}$$

where $d(q,p)$ is the distance function, L2 norm, of q, the query point, and p, the data point from the catalog database, S.

5.4.1 Steps

For the sake of clarity in explaining the steps of the nearest neighbor search, an arbitrary query point, $(0.0108,0.2957)$, is chosen. This query point is not part of the initial dataset the KD tree is originally constructed from and is a newly observed data point from a new earthquake with envelope amplitudes 0.0108 and 0.2957 cm/s^2.

The first step of the nearest neighbor search is to create a path from the root node to the leaf node nearest to the query. Leaf nodes are the nodes located at the very bottom of the tree. The partitioned space helps visualize where to locate this nearest leaf node, which is done by finding the partitioned space containing the query point. Once this particular partitioned space is found, the distance function $d(q,p)$ is calculated and initialized as the "best-fitting distance." Applying this concept to the previously defined 2D sample dataset, the nearest leaf node is node J, and the root node is node C. The distance found between data point J and query point is initialized as the "best-fitting distance."

All the visited nodes along the first path down the tree are to be examined. Therefore, the next step is to find the next nearest node amongst the visited nodes and calculate the distance $d(q,p)$. The next nearest node to examine is node H. The distance between data point H and query point is greater than the "best-fitting distance," which means no update is made.

The advantage of the KD tree nearest neighbor search is the *pruning* process of certain nodes. Pruning reduces the search space by eliminating nodes to examine. If the node is unvisited along the first path down the tree, the feature that decides if it is pruned or examined is its associated *hypersphere*. With the query point at the center, the hypersphere is drawn with its radius as the distance, $d(q,p)$. The most recently examined node was node H. Here, the hypersphere is drawn with a radius of 0.0862 cm/s^2, as shown in Fig. 5.2. The hyperplane corresponding to node I lies within this hypersphere, indicating node I is to be examined. The next node is node E, which lies outside of the hypersphere, meaning it is to be pruned from the tree.

The next node to examine is the root node because it was previously visited along the first path down the tree. The steps are repeated. The distance between data point C and query point is greater than the "best-fitting distance," which means no update is made. A

hypersphere associated with the data point C is drawn, as shown in Fig. 5.2. The hyperplanes corresponding to nodes A and B are found lying within this hypersphere. Therefore, the next nodes to examine are nodes A and B. The distances calculated at these nodes are greater than the "best-fitting distances," indicating no updates. The remaining nodes, D, G, and F, are pruned because the associated hyperplanes lie outside the hypersphere. The final best-fitting data point is data point J.

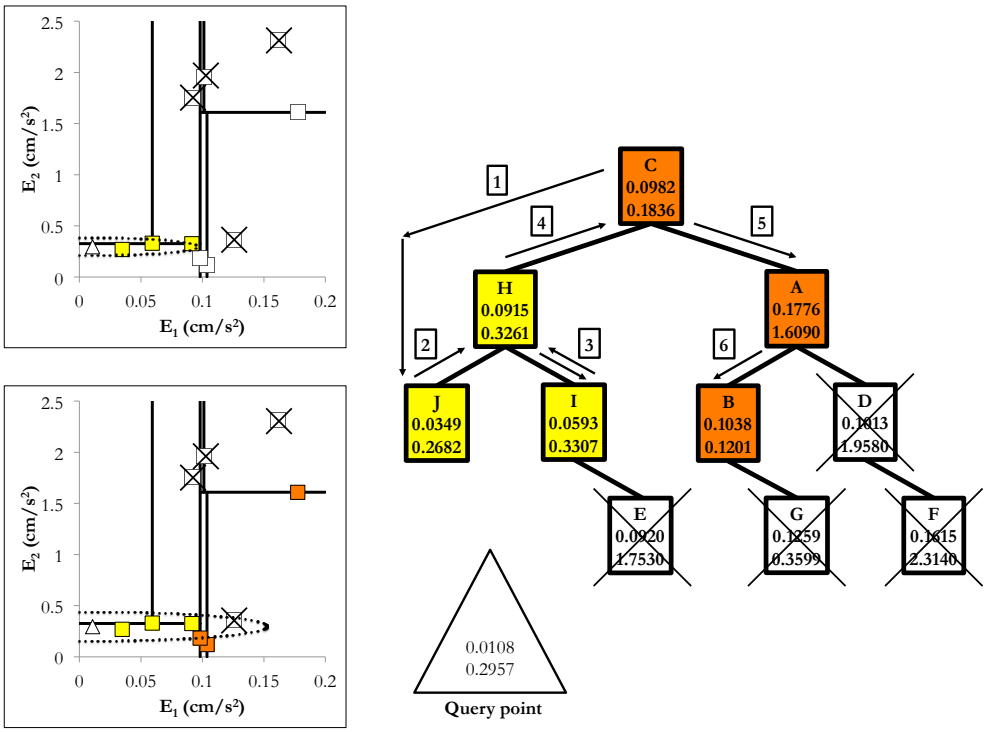


Figure 5.2. KD tree nearest neighbor search. The query point (triangle) is the newly observed data point. The data points organized in partitioned spaces (left) show which nodes are examined and pruned. Those lying within the dotted hypersphere are examined (yellow and orange nodes), while those lying outside it are pruned (i.e. nodes that are crossed out). The order in which the nodes are examined are numbered in the tree structure (right).

5.5 Advantages Compared to Brute-Force Search

The most important advantage of a KD tree nearest neighbor search is its speed. To demonstrate this advantage, the time complexity in constructing and searching a KD tree is shown below. The time complexity refers to the computational complexity that describes the amount of time to run an algorithm. It looks at the size of the dataset and the amount of operations. The first observation is the time complexity in constructing a KD tree. Also as previously mentioned, the main idea of constructing a KD tree is recursively calculating the median and splitting the dataset until the final subsets only consists of only one data point. Therefore, the time complexity looks to the number of times the dataset is divided. Given a dataset of size n with k dimensions, the total is kn values. Recursively, the dataset splits by 2, until only one data point, which is characterized by k values, remains (see Eqs. 5.1).

$$\frac{kn}{2^x} = k \tag{5.1.1}$$

$$n = 2^x \tag{5.1.2}$$

$$\log n = \log 2^x \tag{5.1.3}$$

$$\log n = x \log 2 \tag{5.1.4}$$

$$\frac{\log n}{\log 2} = x \tag{5.1.5}$$

$$\log_2 n = x \tag{5.1.6}$$

where k is the amount of dimensions the dataset of size n is characterized by and x is the amount of times the dataset is split by to construct the finalized KD tree.

Seen in Eq. 5.1, the dataset splits $\log_2 n$ times, meaning the depth, or amount of levels, of the finalized KD tree is $\log_2 n$. At each level of the KD tree, there is a computational cost. At the top of the KD tree, every data point is observed, which is a total of kn values (see Fig. 5.3). Traversing the tree, each time the data is split, half the amount of the data is observed (i.e. $\frac{kn}{2}, \frac{kn}{4}, \frac{kn}{8}, \ldots$). The total cost of constructing the tree is C (see Eq. 5.2).

$$C = kn + \left(\frac{kn}{2} + \frac{kn}{2}\right) + \left(\frac{kn}{4} + \frac{kn}{4} + \frac{kn}{4} + \frac{kn}{4}\right) + \cdots = \Theta(kn \log_2 n) \tag{5.2}$$

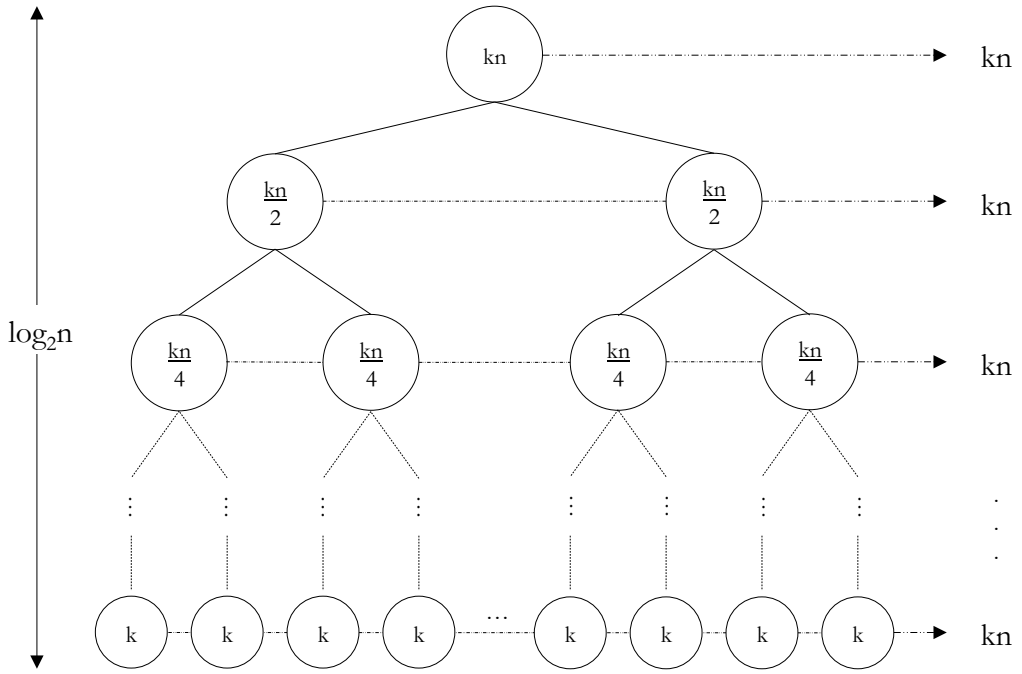


Figure 5.3. Understanding the time complexity of constructing a KD tree. It takes $\log_2 n$ operations to split the dataset into individual nodes. Therefore, the total depth of the tree is $\log_2 n$. At each level of the tree, there are kn operations. kn refers to the amount of values observed to compute the median. Together, the time complexity is $\Theta(kn \log_2 n)$.

The second observation is the time complexity in searching the finalized KD tree. As seen in Fig. 5.4, the worst-case scenario in searching it is if every single node has to be visited, meaning the time complexity is $\Theta(n)$. However, if the finalized KD tree is built properly and well-balanced, then the time complexity for the best-case scenario is $\Theta(\log_2 n)$. This scenario is if only one node in each level of the tree is visited. The bulk of the run time comes from the construction of the KD tree, which is $\Theta(kn \log_2 n)$. Therefore, pre-construction and storage of the KD tree is suggested. The spatial complexity to store a KD tree is linear, which is $\Theta(n)$.

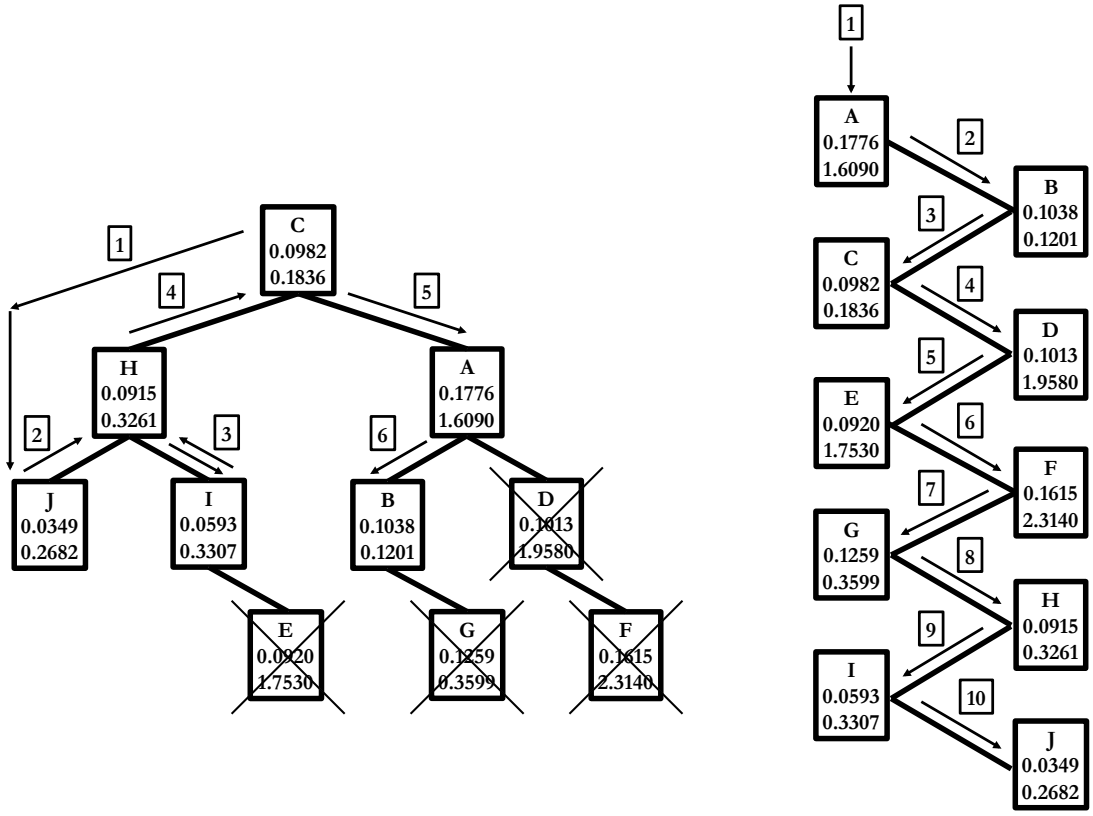


Figure 5.4. Comparing brute-force search and KD tree nearest neighbor search, using the same sample 2D dataset and query point from Fig. 5.2. The KD tree nearest neighbor search (left) is faster because it searches 60% of the total dataset, while the brute-force search (right) goes through 100%.

With the previous sample database of size 10 points, it is not obvious to observe how much faster the KD tree nearest neighbor search is in comparison to a brute-force search. The KD tree nearest neighbor search is notably more valuable with increasing database sizes. In application to the study in this thesis, the database size refers to the amount of events taken from earthquake history. Fig. 5.5(a) compares results for varying databases sizes, from 100 to 100,000 points. With a database size of 10,000, the amount of visited nodes decreases to 50% of the total database. Converting the amount of visited nodes to searching time in MATLAB, the brute-force search time increases exponentially. As the database size approaches 100,000, the search time differs by approximately 78%. The search time using KD trees does not differ significantly from brute-force search time when the database size is less than 1,000. This is the primary reason behind the use of brute-force search for the previous case studies shown in Chapters 3 and 4.

The results in Fig. 5.5(b) are found assuming the dimension, or K, is 38. This refers to KD trees consisted of 30 second long envelopes. For EEW, there is no need for such long envelopes, so this refers to a worst-case scenario. When K is smaller, the impact of the KD tree search is much more obvious, meaning the search time would differ by more than 78%. The curse of dimensionality from an increasing K does not apply here because a threshold is set on K to ensure that the tree-based search performs faster than brute-force.

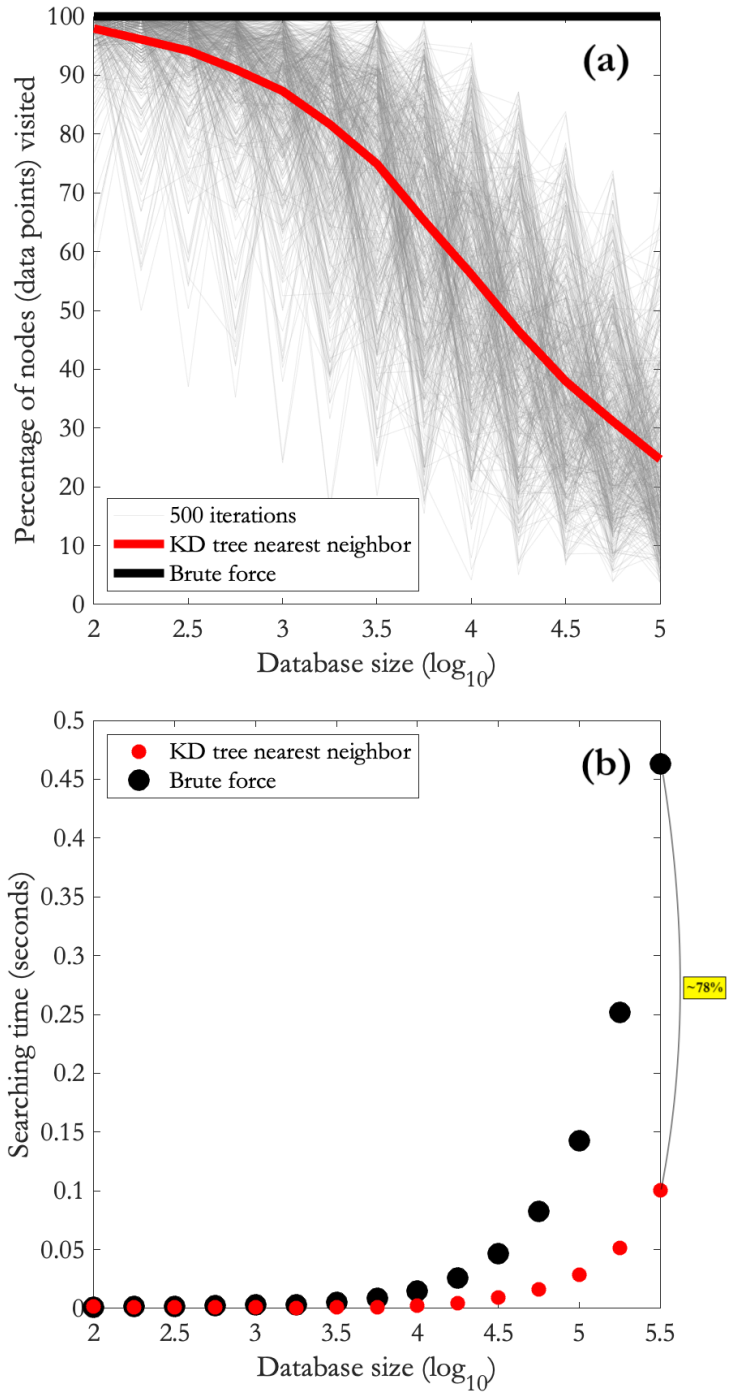


Figure 5.5. (a) Comparing the amount of visited nodes that KD tree nearest neighbor and brute-force search observes. As database size increases, the KD tree nearest neighbor search scans through a fraction of the total dataset to find the best-fitting data point. Comparisons are based on the mean (red) of 500 iterations using random datasets (gray). (b) Comparing the search times (MATLAB) of the KD tree nearest neighbor and brute-force search. In this case, the value of $K = 38$ is constant, which means the dataset is characterized by 30 seconds of ground motion envelopes.

5.6 Conditions

Many papers describe and analyze the performance of the KD tree search using a constant value for K. However, in application to Method II, K is not constant. K depends on the length of the observed waveform of the incoming earthquake, which increases with additional data with increasing time. For instance, if the extended catalog search calls for ground motion envelopes for 30 seconds of data, then $K = 38$, where the additional 8 dimensions characterize the station and event information. If the envelope of the incoming earthquake consists of 60 seconds of data, then the extended catalog search uses a database of $K = 68$.

The curse of dimensionality refers to the invalidation of the results plotted in Fig. 5.5 due to a varying K. This is seen in Fig. 5.6; the computational searching time increases as the value of K increases. The computational searching time essentially increases as the amount of visited nodes increases. One observation is when $K > 308$, which is a tree consisting of approximately 3 minute long ground motion envelopes, the KD tree search takes as long as the exhaustive brute-force search. Fortunately, in EEW, 5 minutes of ground motion amplitudes is unnecessary.

Therefore, the value for K is constrained to save computational efforts in constructing the tree structures. As seen in Fig. 5.6, for the KD tree nearest neighbor search to have faster speed than the exhaustive brute-force search by at least 50%, the value for K must be less than 83. The trees to construct consist of ground motion envelopes of less than 75 seconds, which is hardly an issue for EEW-relevant application.

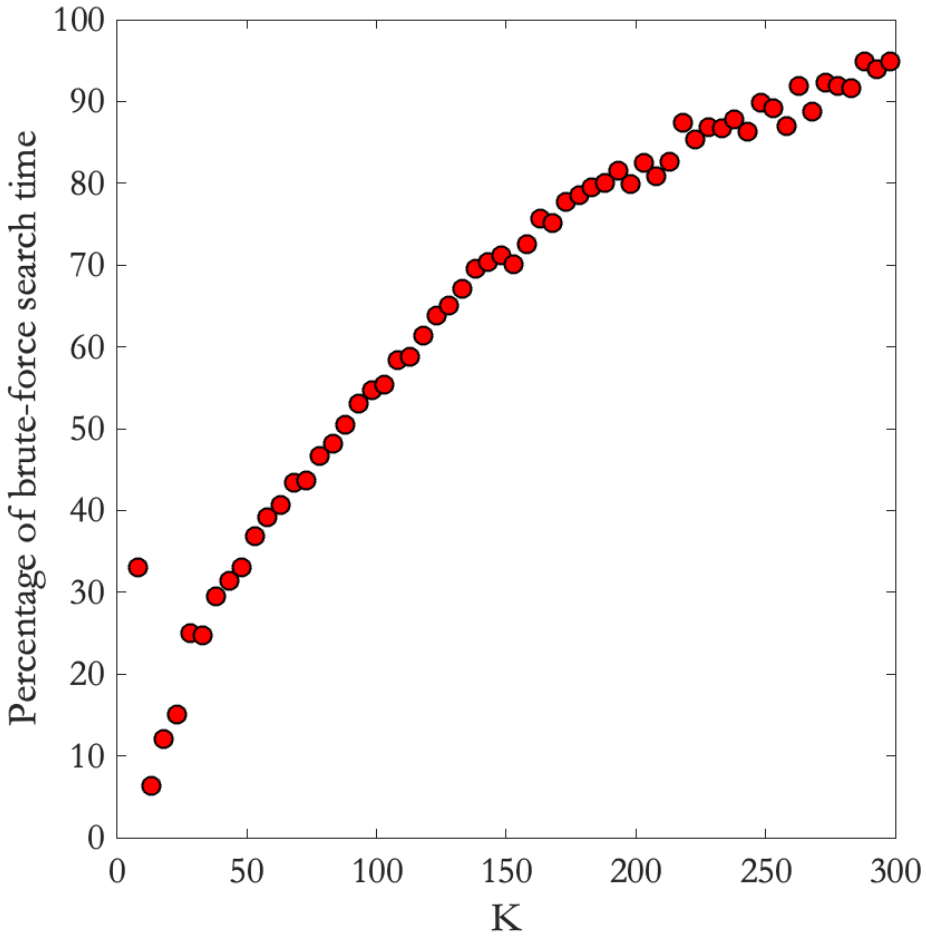


Figure 5.6. KD tree nearest neighbor search in comparison to brute-force search when database size is approximately 560,000.

5.7 Application to Current SCSN Catalog

To show the impact of the KD tree nearest neighbor search on the searching time, the number of visited nodes is observed. This number is compared to the number of nodes the brute-force search visits, which is every single one.

Tables C-E list the number of waveform envelopes from the current SCSN catalog from years 2000 to 2020. Table 5.3 considers stations within 100 km about the epicenter for M>5 events, Table 5.4 considers stations within 50 km about the epicenter for M>4 events, and Table 5.5 considers stations within 10 km about the epicenter for M>3 events.

Seen in Table 5.3, as the number of waveform envelopes increases, the impact of the KD tree nearest neighbor search remains similar. However, as the number of dimensions (K) increases, the amount of visited nodes increases from about 5% to 42% but remains less than the amount the brute-force search visits (100%).

As seen in Fig. 5.5, the value of the KD tree nearest neighbor search is amplified when the database size increases and the number of dimensions (K) remains constant. However, a real-time EEW application involves a varying K. Particularly, the curse of dimensionality is prevalent as the number of waveform envelopes increases to > 5,000,000 and the number of dimensions increases to > 39. An increasing K for large database sizes requires every single node to be visited, in which case is the same as the behavior of a brute-force search.

Table 5.3. KD tree performance with stations within 100 km about epicenter for M>5 events.

Looking back *x* amount of years	# of earthquakes	# of waveform envelopes in total database after applying spectral scaling model* (*N*)	KD tree nearest neighbor search time as percentage of brute-force's, given the dimensionality (*K*)				
			12	24	39	54	69
5 years	9	289,980	6%	20%	33%	34%	42%
10 years	22	539,820	5%	16%	31%	36%	41%
15 years	32	779,940	5%	17%	30%	37%	42%
20 years	36	843,480	6%	17%	30%	38%	42%

Table 5.4. KD tree performance with stations within 50 km about epicenter for M>4 events.

Looking back *x* amount of years	# of earthquakes	# of waveform envelopes in total database after applying spectral scaling model (*N*)	KD tree nearest neighbor search time as percentage of brute-force's, given the dimensionality (*K*)				
			12	24	39	54	69
5 years	142	1,744,560	5%	17%	30%	36%	42%
10 years	329	3,475,080	5%	13%	31%	45%	51%
15 years	433	5,064,300	5%	12%	31%	77%	81%
20 years	522	6,653,880	6%	12%	100%	100%	100%

Table 5.5. KD tree performance with stations within 10 km about epicenter for M>3 events.

Looking back *x* amount of years	# of earthquakes	# of waveform envelopes in total database after applying spectral scaling model (*N*)	KD tree nearest neighbor search time as percentage of brute-force's, given the dimensionality (*K*)				
			12	24	39	54	69
5 years	1,702	4,687,200	5%	12%	31%	70%	75%
10 years	3,714	8,249,760	6%	14%	100%	100%	100%
15 years	4,626	10,807,920	7%	31%	100%	100%	100%
20 years	5,462	13,580,460	6%	78%	100%	100%	100%

5.8 Summary

In previous chapters, brute-force search is used to conduct studies on the 2020 Northern coast offshore, 2020 Lone Pine, and 2019 Ridgecrest events. These cases involve relatively small catalogs of sizes approximately 90 to 3,000 envelopes per station. It may be necessary for the extended catalogs to have greater than 10,000 records to ensure sufficient representation of a variety of earthquake magnitudes and locations to guarantee an accurate match. In such cases, a KD tree nearest neighbor search is recommended to potentially reduce search time.

It is computationally most effective to pre-determine the KD trees for storage. A separate KD tree exists with respect to the length of ground motion envelopes. For instance, matching 15 second envelopes would search a 23-D tree and matching 30 second envelopes would search a 38-D tree. For catalogs that consist of less than 560,000 waveform envelopes of lengths less than 3 minutes, the KD tree search scans through a fraction (< 50%) of the total catalog to find the best fit. If the catalog size increases to 3,000,000 waveform envelopes, for the KD tree search to scan through < 50% of the total catalog, the length of the ground motion envelopes must be less than 1 minute. The KD tree search performs similarly to the brute-force search when the database size grows to 5,000,000 and the length of ground motion envelopes grows to 30 seconds $(K = 38)$. The dependence of the number of dimensions (K) and database size on the computational searching time illustrates the curse of dimensionality. With time, more earthquakes will occur, meaning the catalog will continue to grow. To run the KD tree nearest neighbor search in real-time, it is important to continuously update it.

6 Complex Earthquakes

As mentioned in previous Chapter 4, a simplified spectral scaling model extends the original catalog of waveform envelopes to include a variety of earthquake magnitudes. A strong assumption made by the model is that the earthquake is characterized as a point source, not a finite fault. However, for larger earthquakes (M>6.5), the rupture length is longer, and characterizing the earthquake as a point source would make it difficult to predict large shaking, particularly in regions close to the fault but not necessarily close to the epicenter. This chapter addresses how the extended catalog search is modified for complex earthquakes, where point source characterization may not provide the best envelope fits to the incoming ground motions. Complex earthquakes refer to a cascade of multiple ruptures, also denoted as subevents, that generates large ground motions. Regions like Japan and New Zealand have such earthquakes of high complexity, where waves from one end of the fault pile up with waves from another end, causing amplification in the direction of the rupture propagation. Real complex earthquakes to study are the 2016 Kumamoto, 2016 Kaikoura, 2010 El Mayor-Cucapah, and 2019 Ridgecrest. Though these earthquakes are commonly branded as single events, multiple peaks and arrivals are detected in the ground motions, indicating the potential for multiple sources. Therefore, to find appropriate waveform envelope fits and accurate parameter estimates, the extended catalog search includes additional templates of complex earthquakes. This chapter addresses the addition of these new templates for complex earthquakes using waveforms of past real earthquakes.

6.1 Point Source vs. Finite Fault Characterization

In case studies of real earthquakes presented in Chapter 4, the extended catalog search finds envelope fits that accurately describe the incoming ground motions. As previously mentioned, EEW aims to find parameter estimates as soon as possible, which means the waveforms considered in these case studies are from triggered stations near the epicenter. Though the stations are located close, their epicentral distances are still greater than the fault rupture length, making the point source approximation of the earthquake valid. However, for earthquakes of larger magnitudes, a finite fault approximation may be necessary to provide more accurate parameter estimates. Fortunately, a real-time Finite Fault Rupture

Detector (FinDer) algorithm already exists for such cases (Bose 2012). This algorithm uses image recognition techniques to detect fault ruptures, assuming a line source. Therefore, it is sensible to have FinDer running in parallel to the extended catalog search and grid search as a form of confirmation.

The effectiveness of using the early frequency content of the waveforms to estimate the final magnitude estimate is questioned for larger earthquakes. Larger earthquakes are more complex, where multiple sources rupture closely in time and in space. They have longer rupture duration and length. Method II in Chapter 4 uses envelopes that represent single medium-sized events. Therefore, additional templates are introduced to represent complex earthquakes.

6.2 Additional Templates

Additional templates for complex earthquakes are added to the extended catalog. The inclusion of these templates that represent multiple sources may help avoid missed or false alerts for expected large ground motions in regions that are close to the fault. When earthquake history is sufficient with respect to epicentral locations, the combination of individual waveform envelopes (see Eq. 6.1) from past real earthquakes at different time delays is used as templates for complex earthquakes. When waveforms of past real earthquakes are scarce or do not exist, the combination of individual envelopes based on Cua-Heaton GMPEs at different time delays is used instead.

$$E_{complex}(t) = \sqrt{\sum_{i=1}^{n}(E_i(t-\delta))^2} \tag{6.1}$$

where $E_{complex}(t)$ is the estimated envelope of the total complex earthquake, $E_i(t-\delta)$ is the envelope at the i^{th} source with time delay of δ seconds. $E_i(t-\delta)$ can be a waveform envelope from a past real earthquake or based on the Cua-Heaton GMPEs.

6.3 Application to Real Complex Earthquakes

6.3.1 2016 Kumamoto sequence

According to the ANSS catalog, the 2016 Kumamoto sequence started with a M6.2 foreshock on April 14, 2016 at 12:26:35 UTC. Approximately 6 km away, the M7.0 mainshock soon followed on April 15, 2016 at 16:25:06 UTC. Assuming 36 hours is ample time for stations to collect true parameter information on the foreshock, the extended catalog search uses templates generated from the waveforms of this M6.2 foreshock. In previous case studies where the point source approximation is valid, a simple application of the simplified spectral scaling model to original, raw waveforms creates sufficient amount of envelopes that have the potential to accurately describe the incoming ground motions. It is mentioned in Chapter 4 that this search is specific to the station and channel, making the search include not only source (i.e. magnitude, location) effects but also path (i.e. depth) and site (i.e. rock, soil) effects. However, the directivity effects are not taken into consideration. Therefore, for a large, complicated earthquake like the Kumamoto mainshock, it is sensible for the templates to consider directivity effects that happen due to the rupture propagation, especially because 5 seconds after the origin time, the main rupture starts to propagate towards the northeast, along the Futagawa fault (Yagi et al. 2016).

In this analysis, the stations are chosen based on their distances to the mainshock epicenter and to the Futagawa fault. Because EEW aims to find parameter estimates before strong shaking arrives, waveforms from stations close to the epicenter ($R<50$ km) are selected for the extended catalog search. These particular stations are triggered, meaning the P-wave arrives, within 6 seconds of the origin time of the mainshock. As shown in Table 6.1, only the first ten triggered stations are considered in this analysis, meaning the magnitude estimates are found using 90 observed envelopes (i.e. acceleration, velocity, and displacement for EW, NS, and UD components). For real-time application, the amount of considered stations will continue to increase with time and may provide more accurate magnitude estimates with reduced error bands. However, this analysis emphasizes the accuracy of the magnitude estimates using only the initial parts of the incoming waveform envelopes (<10 seconds after the origin time).

Table 6.1. Triggered stations from the 2016 Kumamoto event with P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	Latitude(°N)	Longitude (°E)	P-wave arrival (sec after origin time)
KMM006	32.79	130.78	4
KMM008	32.69	130.66	5
KMM005	32.88	130.88	5
KMM011	32.62	130.87	6
KMM009	32.69	130.99	6
KMM010	32.61	130.49	7
KMM002	33.02	130.68	7
KMM012	32.51	130.60	7
KMM007	32.83	131.12	8
KMM004	32.93	131.12	8

The key feature of the extended catalog search is the accuracy of its initial magnitude estimates. As shown in Fig. 6.1, FinDer produces initial estimates 4 seconds after the origin time, but heavily underestimates the magnitude to M3.2. It finally approaches the true M7.0, taking 15 seconds after the origin time. The jump from M5.8 to M6.5 at 7 seconds after the origin time is consistent with the theorized directivity effect: the rupture propagation northeast along the Futagawa fault occurring 5 seconds after the origin time. With the initial rupture of the mainshock arriving at the first station 2 seconds after the origin time, the delayed subevent from the rupture propagation would arrive at the first station approximately 7 seconds after the origin time. Similar to the solutions of FinDer, the extended catalog search finds the magnitude estimate to approach M6.9, an underestimation to the true recorded M7.0. This is assuming the magnitude of the foreshock, M6.2, is accurate. On the other hand, the extended catalog search immediately recognizes the incoming ground motions to resemble those of a M6.6 event. This is approximately 4 seconds faster than FinDer's solutions.

Fig. 6.2 shows the best-fitting cataloged envelopes and how they compare to the incoming observed ones. The stations with the worst envelope fits, relative to the stations considered in this analysis, are KMM004, KMM007, KMM005, and KMM006. The differences in the waveform envelopes are amplified in the velocity and displacement, where longer period components are represented. Seen in Fig. 6.3, the stations' locations with

respect to the ruptured fault explains the differences in the cataloged envelopes and the incoming observed ones. These stations are located near the fault and/or in the direction of the rupture, where higher complexity may exist due to amplified ground motions. The generation of templates using two subevents to represent the M7.0 mainshock may not be adequate to capture the complexity, meaning there may be more than two. However, the use of two subevents to represent the M7.0 mainshock works for the rest of the stations. These stations are either located farther away from the fault or in the opposite direction of the rupture, where ground motions are less likely to amplify and rupture propagation is not as complex.

In this particular mainshock, the point source and finite fault characterizations do not have a huge impact on magnitude estimates. Generally, the difference in magnitude estimates is only 0.1. However, the differences are seen more clearly in the error bands. With the complex earthquake templates (ones that are generated using two subevents), the error bands are reduced, meaning the waveform envelopes fit better than the ones used in point source approximation (using single event). The definition of the error bands is provided in Chapter 3.

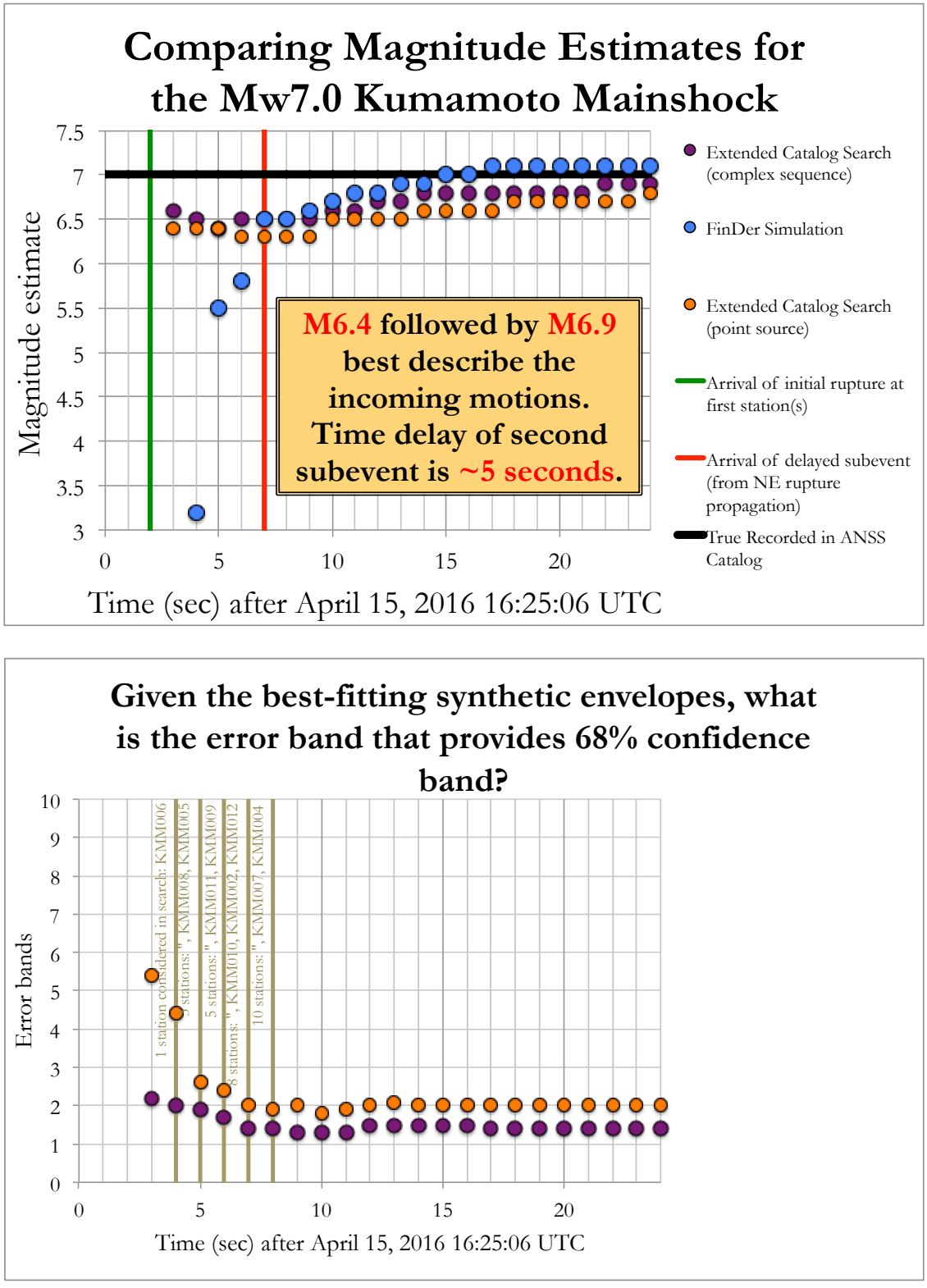
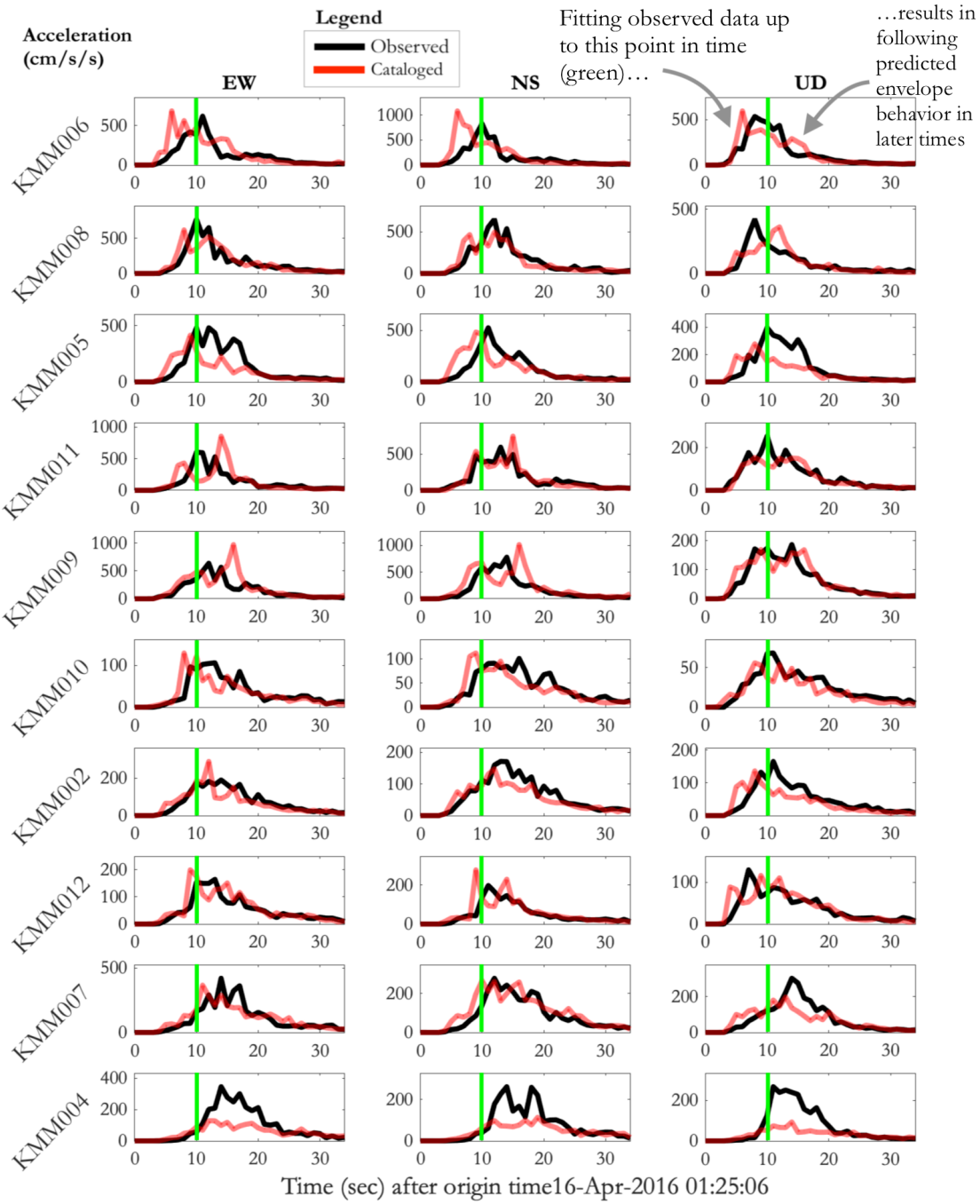
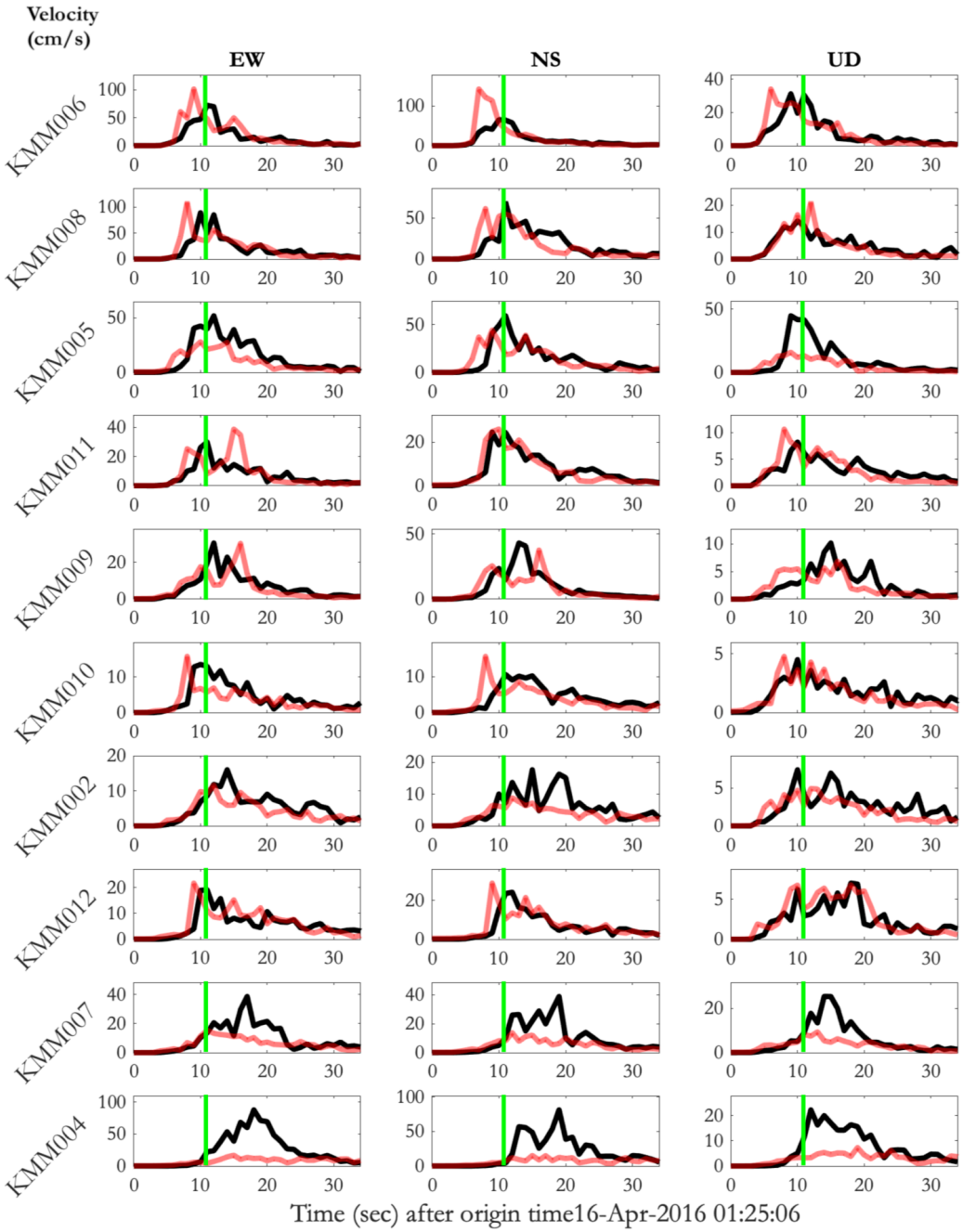


Figure 6.1. Complex earthquake catalog search magnitude estimates for the 2016 Kumamoto. Along with magnitude estimates, error bands for 68% confidence band compare envelope fits between point source and complex sequence cases.



(figure continues on next page)



(figure continues on next page)

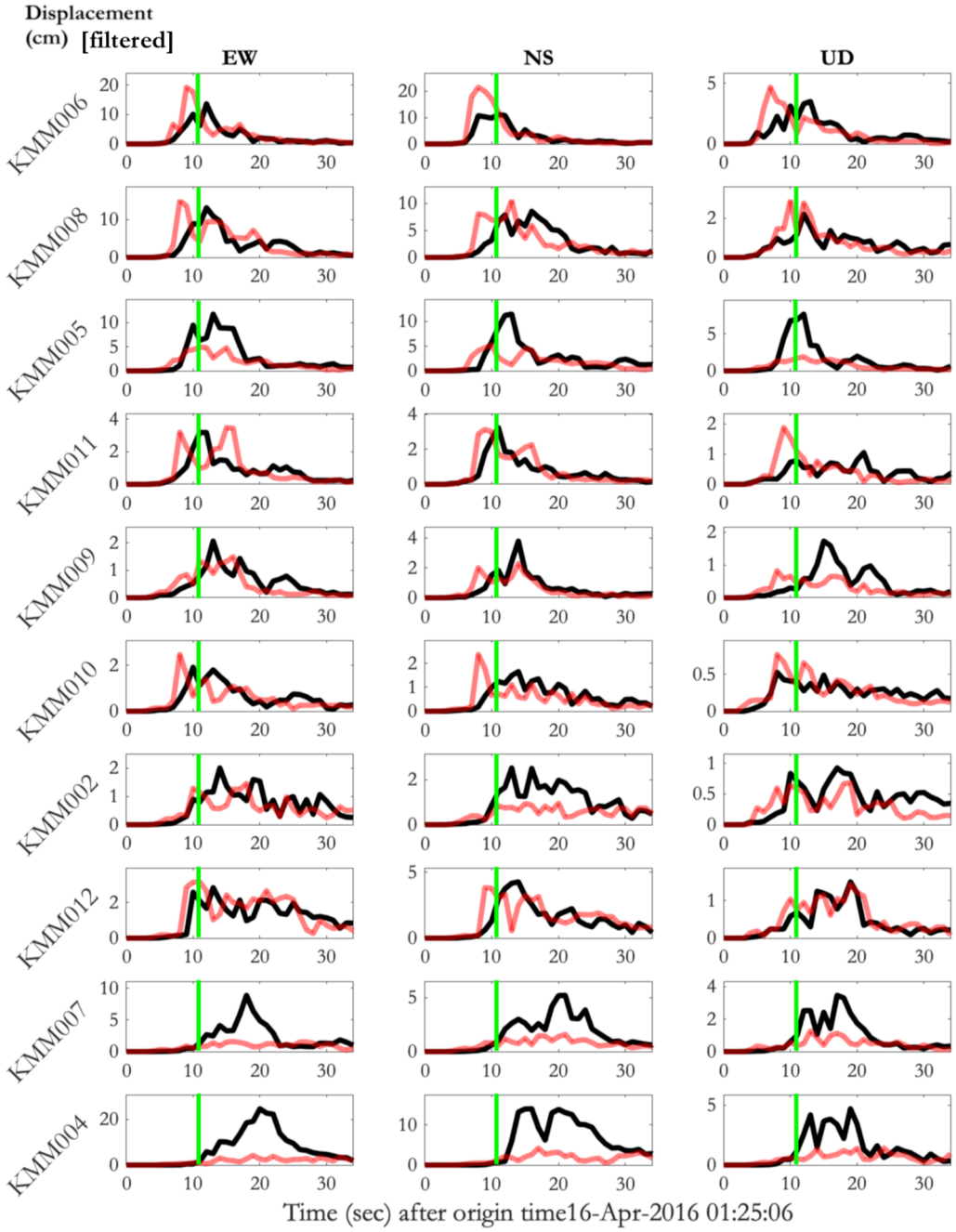


Figure 6.2. Comparing the best-fitting complex earthquake cataloged (in red) and incoming observed envelopes (in black) for the 2016 Kumamoto mainshock. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and displacement are also labeled accordingly.

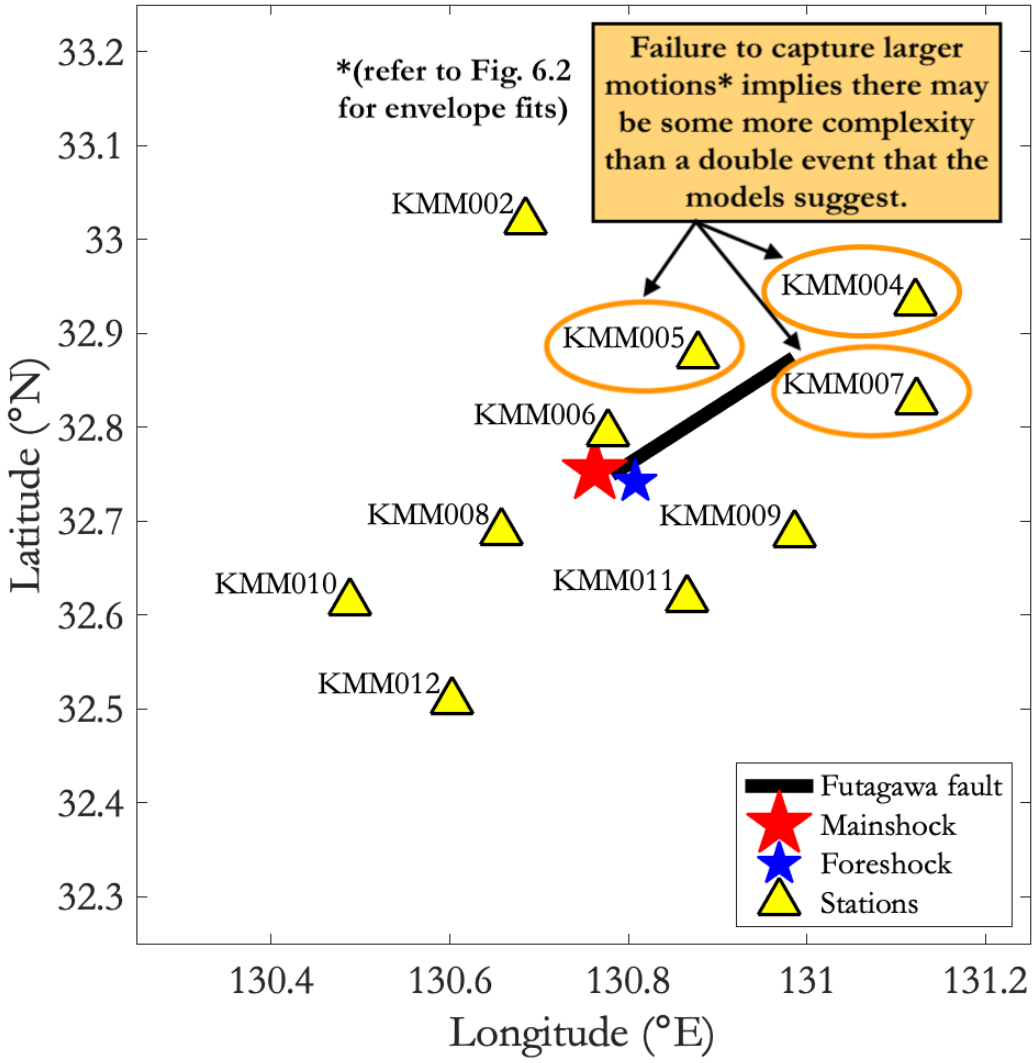


Figure 6.3. Station locations with respect to the epicenter of the mainshock, epicenter of past event, and ruptured fault for the 2016 Kumamoto mainshock. This particular rupture propagation is the northeast direction.

6.3.2 2010 El Mayor-Cucapah

Another complex earthquake is the 2010 El Mayor-Cucapah mainshock that occurred in the Baja California, Mexico region on April 04, 2010 at 22:40:52 UTC. The uniqueness of the complexity in this mainshock is the closeness in time between the rupturing events and a long rupture length (~120 km). It is believed that the mainshock started with a ~M6 normal faulting, rupturing bilaterally from the epicenter in the northwest and southeast directions, and followed by the main M7.2 event only 15 seconds later (Hauksson et al. 2010). Therefore, the templates used in this analysis will take waveform envelopes from two past events at different time delays.

The data for this analysis is downloaded from the California Strong Motion Instrumentation Program (CSMIP). To ensure the extended catalog search consists of a variety of waveform envelopes, earthquake history includes years back to 2002. This particular earthquake is located out of network, or in a region of sparse station coverage. Therefore, the accessibility of stations is limited. As shown in Table 6.2, the two closest stations to the epicenter of the mainshock for which ample waveforms are available are Calexico Fire Station, denoted NP 5053, and Holtville, Post Office, denoted NP 5055. These two stations are the closest to the US-Mexico border. The aim is to use only these two stations to find accurate parameter estimates. The finite fault approximation should reduce the error bands, for judgment of envelope fits, and error in parameter estimates, for comparing to the true values, that the point source characterization produces.

Table 6.2. Triggered stations from the 2010 El Mayor-Cucapah event with P-wave arrivals. Maximization of posterior probability considers data from only these stations.

Station	Latitude(°N)	Longitude (°W)	P-wave arrival (sec after origin time)
NP 5055	32.811	115.379	6
NP 5053	32.670	115.493	9

The best-fitting cataloged envelopes are the ones that are scaled and combined at different time delays using the waveforms from the past M5.70 event from February 22, 2002 19:32:41 UTC. Initially, the magnitude estimates are underestimated. As seen in Fig. 6.4, by 10 seconds after the origin time, the magnitude estimate approaches M6, which is consistent with the theorized first ~M6 subevent (Haukkson et al. 2010). 21 seconds after

the origin time, the jump in magnitude estimate from M6 to M6.7 implies the arrival of a second subevent, which is consistent with the theorized second M7.2 subevent (Haukkson et al. 2010). Also seen in Fig. 6.4, the error bands remain the same for point source (envelopes using single event) and finite fault assumption (envelopes using two subevents at different time delays) until 21 seconds after the origin time. At that point in time, the error band is significantly reduced, further satisfying the implication of the arrival of the second rupture at the stations. Eventually, the magnitude estimate approaches the true recorded M7.2.

The comparison of the error bands in Fig. 6.4 indicates when to choose point source or finite fault approximation. The one that produces smaller error bands has more accurate envelope fits, therefore, more confidence in the resulting parameter estimates. As shown in Fig. 6.4, the arrival of the second subevent is when the error bands begin to diverge, which is indicated by the vertical red line. The increase in the error bands at 21 seconds after the origin time implies the envelopes that previously used to fit the incoming ground motions ceases to fit and envelopes for a complex sequence are needed. This shows both goodness-of-fit score (i.e. maximization of posterior probability, minimization of sum of squared residuals) and error bands are important in accurately describing the incoming ground motions. A point source assumption for the whole, complex rupture may fail to alert regions for stronger shaking to come.

The best-fitting envelopes corresponding to relatively smaller error bands are illustrated in Fig. 6.5. Fig. 6.4 and Fig. 6.5 suggest the 2010 El Mayor-Cucapah mainshock is a double event, with a M6.0 followed by a M7.2. The time delay of the second subevent is approximately 16 seconds. The additional templates considered in this study use waveforms from the M5.70 event from February 22, 2002 19:32:41 UTC. Of course, the search can include more complexity, such as a triple event, using other nearby events plotted in Fig. 6.6. However, this study constrains the model to a double event for consistent comparison to the solutions from Haukkson et al. 2010.

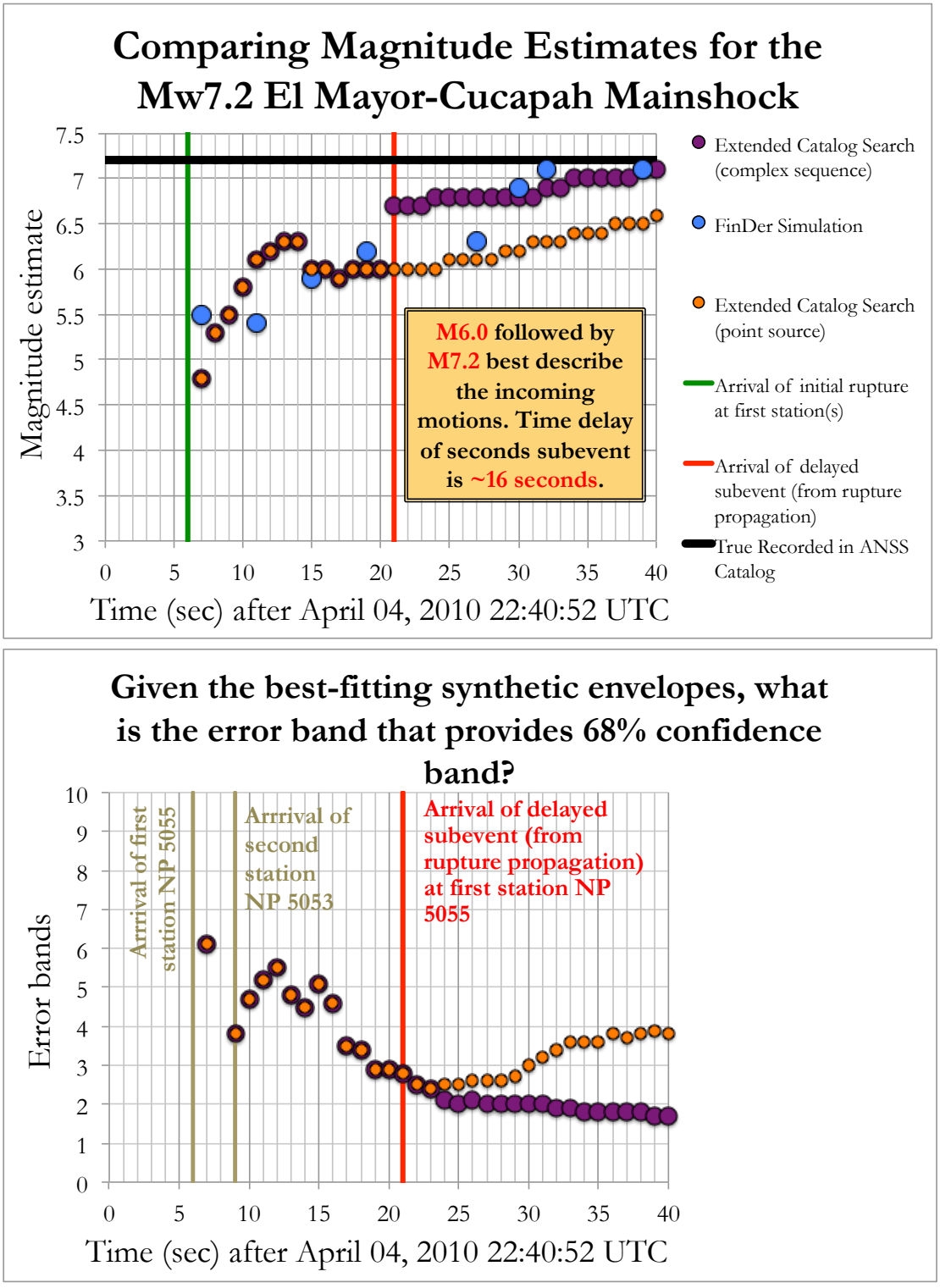
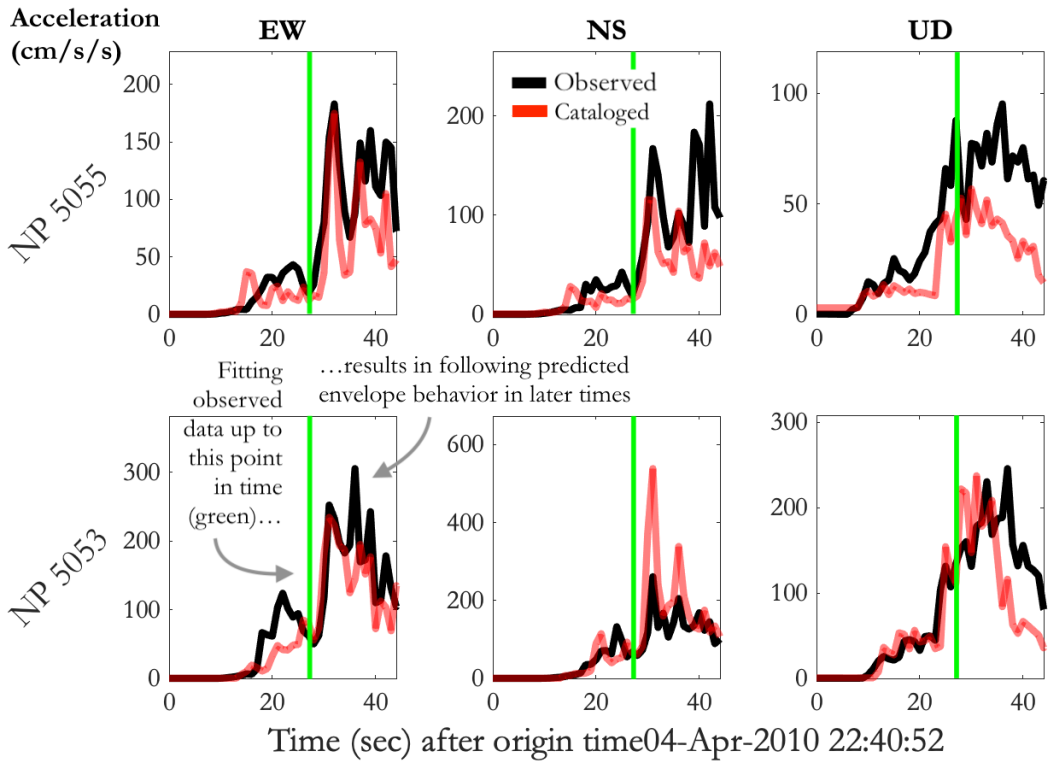
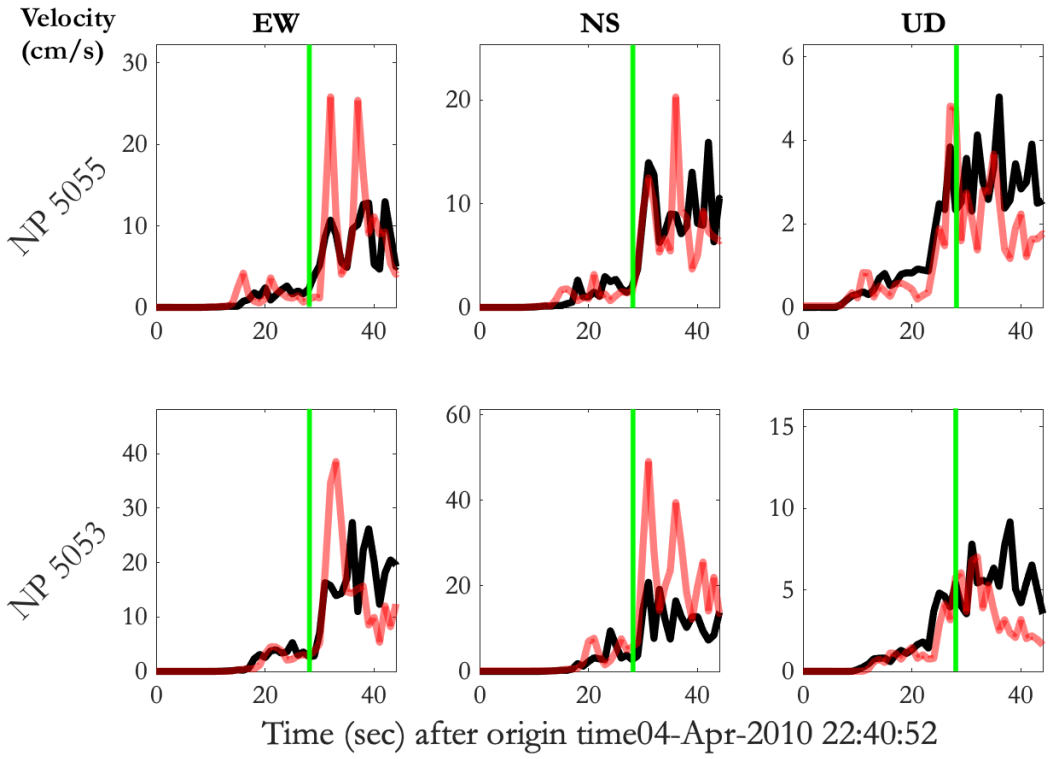


Figure 6.4. Complex earthquake catalog search magnitude estimates for the 2010 El Mayor-Cucapah mainshock. Error bands illustrate the arrival of the second subevent, indicating the need for complex earthquake templates.





(figure continues on next page)

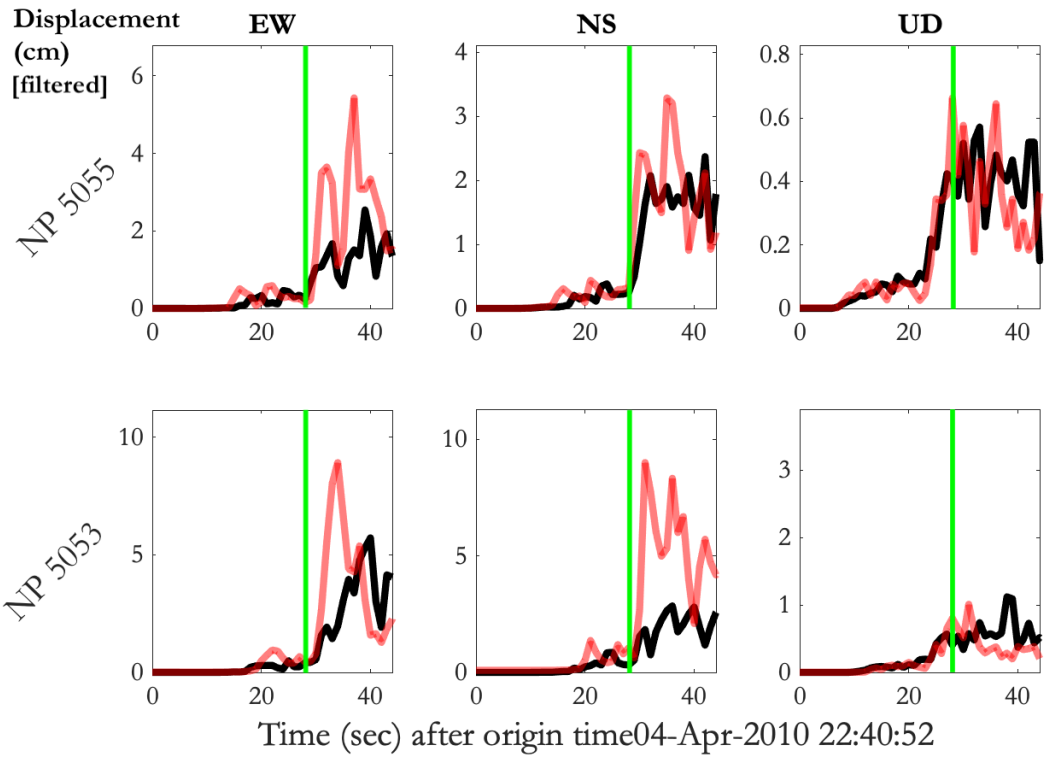


Figure 6.5. Comparing the best-fitting complex earthquake cataloged envelopes (in red) and the incoming observed envelopes (in black) for the 2010 El Mayor-Cucapah mainshock. Each row represents a station (labeled in the y-axis), and each column represents a component (labeled at the top). Acceleration, velocity, and displacement are also labeled accordingly.

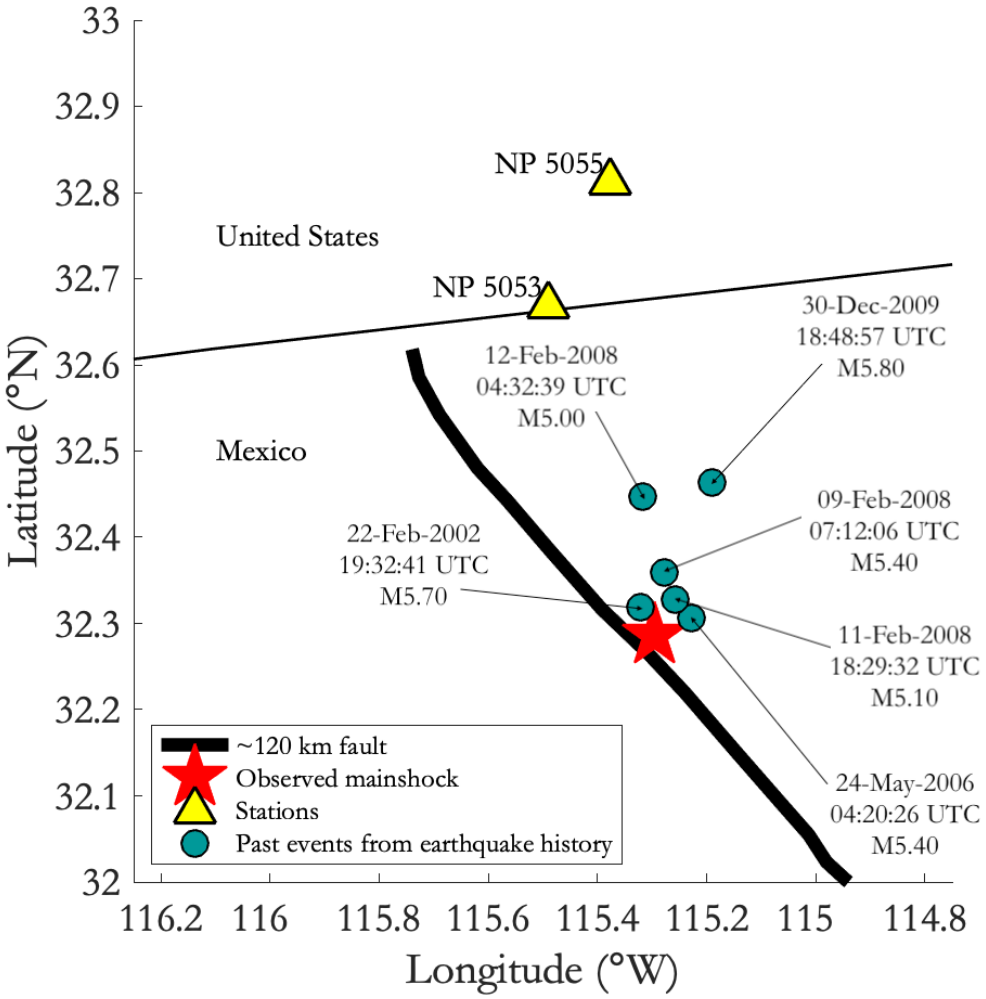


Figure 6.6. Station locations with respect to the epicenter of the mainshock, epicenter of past events, and ruptured fault for the 2010 El Mayor-Cucapah mainshock. This particular rupture propagation is bilateral in the northwest and southeast direction.

6.3.3 2016 Kaikoura

The M7.8 Kaikoura earthquake is also known for its high complexity, as the rupture propagates northwards across multiple faults. Therefore, modeling the earthquake as a point source may not be appropriate in this case as well. In this analysis, the stations to consider are WTMC (R~8 km), CULC, (R~30 km), and HSES (R~30 km). Additional stations that are near the northern end of the fault are KIKS (R~60 km), KEKS (R>100 km), and WDFS (R>100 km). These additional stations experience strong shaking, even though they are located far from the epicenter. The earthquake history at these stations, however, is not sufficient to ensure accuracy in waveform envelope fits and parameter estimates. In other words, the GeoNet catalog does not have available waveforms for the regions near the ruptured faults. Therefore, the templates cannot be generated in the same way as before, where existing waveforms from the past would be scaled to cover enough earthquake magnitudes. This particular earthquake is a case where the extended catalog search has insufficient database to produce accurate parameter estimates. An algorithm that is suited well for this case is FinDer, which is recommended to have running in parallel with the extended catalog search as a form of confirmation.

Another method that may have the potential to solve this problem is a multi-source model using Cua-Heaton ground motion envelopes (Yamada 2007). Here, the templates are created using the attenuation relationships developed by Cua, the ones used in the Virtual Seismologist (VS) method. However, instead of characterizing the earthquake as a point source model as the VS method does, the multi-source model creates the templates by combining the Cua-Heaton ground motion envelopes at different time delays, like in Eq. 6.1. In her thesis, Yamada divides the fault surface into "sub-sources", with each sub-source representing a single point source. Templates for complex earthquakes are simply combinations of Cua-Heaton waveform envelopes at each sub-source. The use of Cua-Heaton envelopes is at a disadvantage in comparison to the extended catalog search because they do not consider the path effects at the specific station and channel.

6.3.4 2019 Ridgecrest sequence

As mentioned in Chapter 4, the 2019 Ridgecrest mainshock also exhibits complex behavior. However, majority of the stations that are considered in the extended catalog search have epicentral distances greater than the ruptured portion of the fault. Therefore, assuming point source produces error bands similar to those considering higher complexity (see Fig. 6.7). Here, the templates for complexity are the combinations of two single events at different time delays.

Examining each individual station and channel shows that the two stations that have the poorest envelope fits relative to the other stations are CLC and CCC. CCC, in particular, is located in the direction of the rupture, where amplification of the ground motions is likely to occur. It is theorized that the Ridgecrest mainshock consists of four subevents (Ross et al. 2019). Therefore, the templates that would provide more accurate envelope fits are combinations of four single events, not two, spaced at different time delays. As shown in Fig. 6.8 in purple, these additional envelopes that represent higher complexity, larger motions that occur later in time are captured. Envelopes assuming point source, shown in orange in Fig. 6.8, fail to do so. Table 6.3 quantitatively exemplifies this trend. The ability to capture the larger motions means it has enhanced ability to alert regions of stronger shaking to come.

As shown in Fig. 6.9, the stations located closest to the fault are CLC and CCC. This explains the large amplifications that occur later in time in which finite fault approximation is needed. For the remaining stations farther away from the fault, point source approximation produces relatively small, satisfactory error bands. The results suggest the 2019 Ridgecrest mainshock is a complex event, with a M7.1 followed by a M6.9 at a delay of 5 seconds, followed by a M6.4 at a delay of an additional 10 seconds, and followed by a M6.0 at a delay of an additional 6 seconds. These envelope fits improve by a factor 2.24 and 1.51 for stations CLC and CCC, respectively. As shown in Table 6.3, for majority of the envelope fits, the error band and sum of squared residuals (SSR) are reduced when the templates for complex sequences are used (purple). Point source characterization (orange) produces poorer envelope fits (larger error band and SSR).

The additional templates considered in this study use waveforms from the M6.4 foreshock. Of course, the search can include even more complexity. However, this study

constrains the model to a double event for consistent comparison to the solutions from Ross et al. 2019.

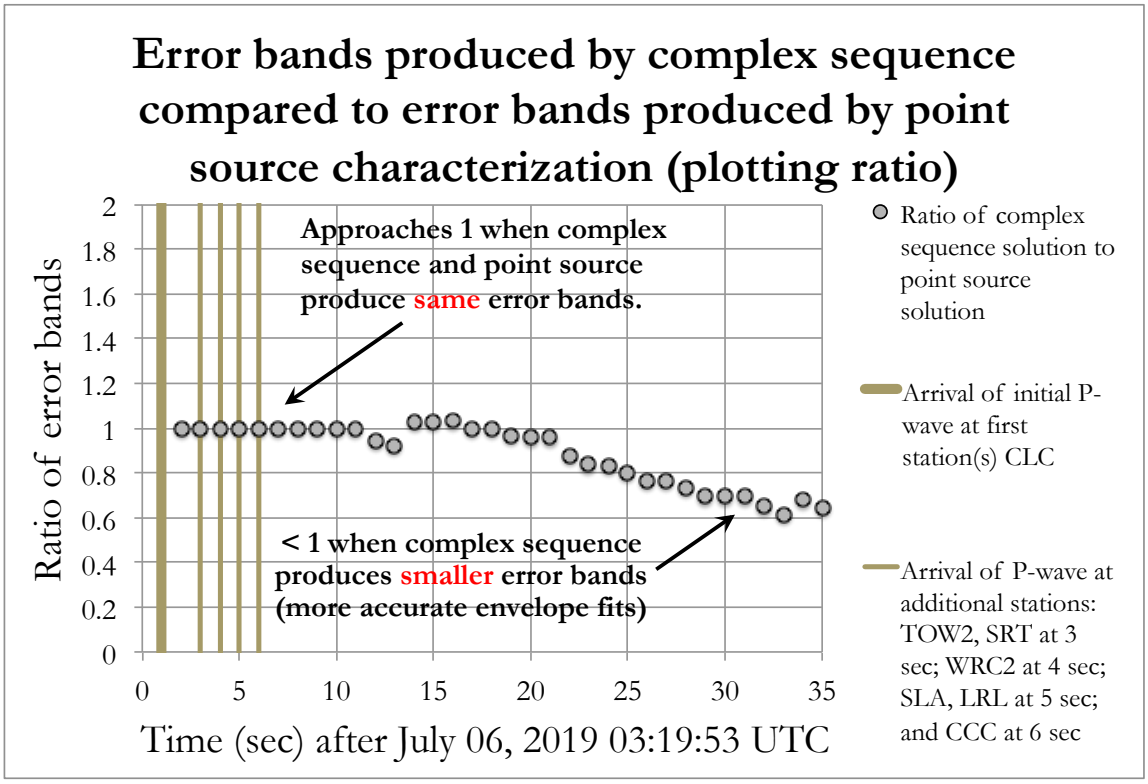


Figure 6.7. Comparing error bands for point source characterization vs. assumption of double events occurring in the 2019 Ridgecrest mainshock. Initially, the differences are essentially nonexistent.

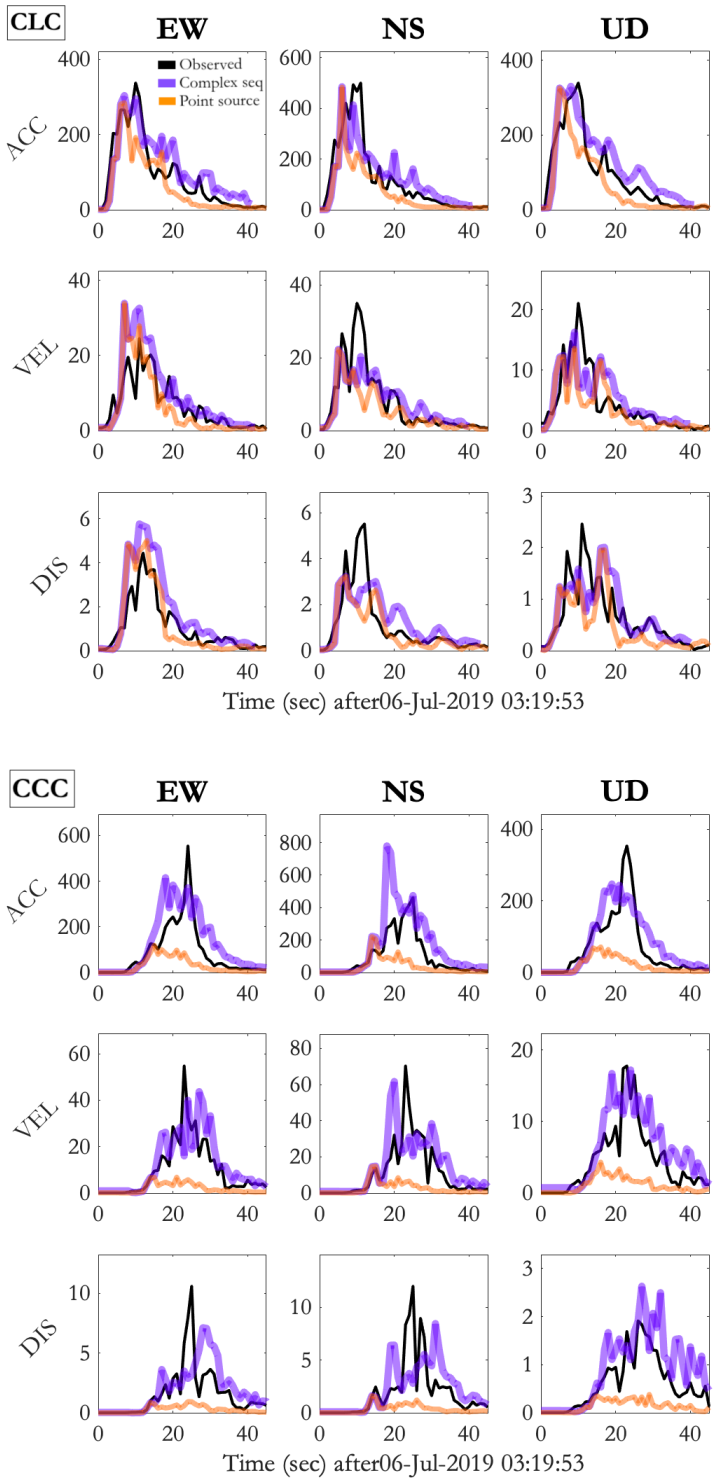


Figure 6.8. Comparing envelope fits for point source characterization vs. complex sequence assumption for the 2019 Ridgecrest mainshock at stations near the fault, CLC and CCC. Templates using multiple sources capture the large ground motions that occur later in time (between 20 and 40 seconds after the origin time). The improvements in envelope fits are quantifiably described in Table 6.3.

Table 6.3. Comparing the error bands and sum of squared residuals at stations CLC and CCC.

ACC (cm/s/s)		CLC			CCC		
		EW	NS	UD	EW	NS	UD
Error band	Point	0.50	0.50	0.50	0.60	0.70	0.70
	Complex	0.30	0.30	0.30	0.60	0.60	0.50
Sum of squared residuals (SSR)	Point	6.44	8.33	6.59	19.73	23.38	27.54
	Complex	3.66	5.45	4.18	14.27	10.58	13.57

VEL (cm/s)		CLC			CCC		
		EW	NS	UD	EW	NS	UD
Error band	Point	0.50	0.40	0.40	0.90	1.00	0.90
	Complex	0.30	0.30	0.30	0.50	0.50	0.40
Sum of squared residuals (SSR)	Point	7.25	3.19	6.23	28.41	28.84	23.25
	Complex	2.35	2.65	2.92	24.01	18.48	26.77

DIS (cm)		CLC			CCC		
		EW	NS	UD	EW	NS	UD
Error band	Point	0.50	0.40	0.40	0.80	1.00	0.90
	Complex	0.40	0.40	0.20	0.60	0.50	0.40
Sum of squared residuals (SSR)	Point	7.13	3.92	8.52	24.89	28.40	22.60
	Complex	3.49	2.93	1.55	12.79	20.46	22.34

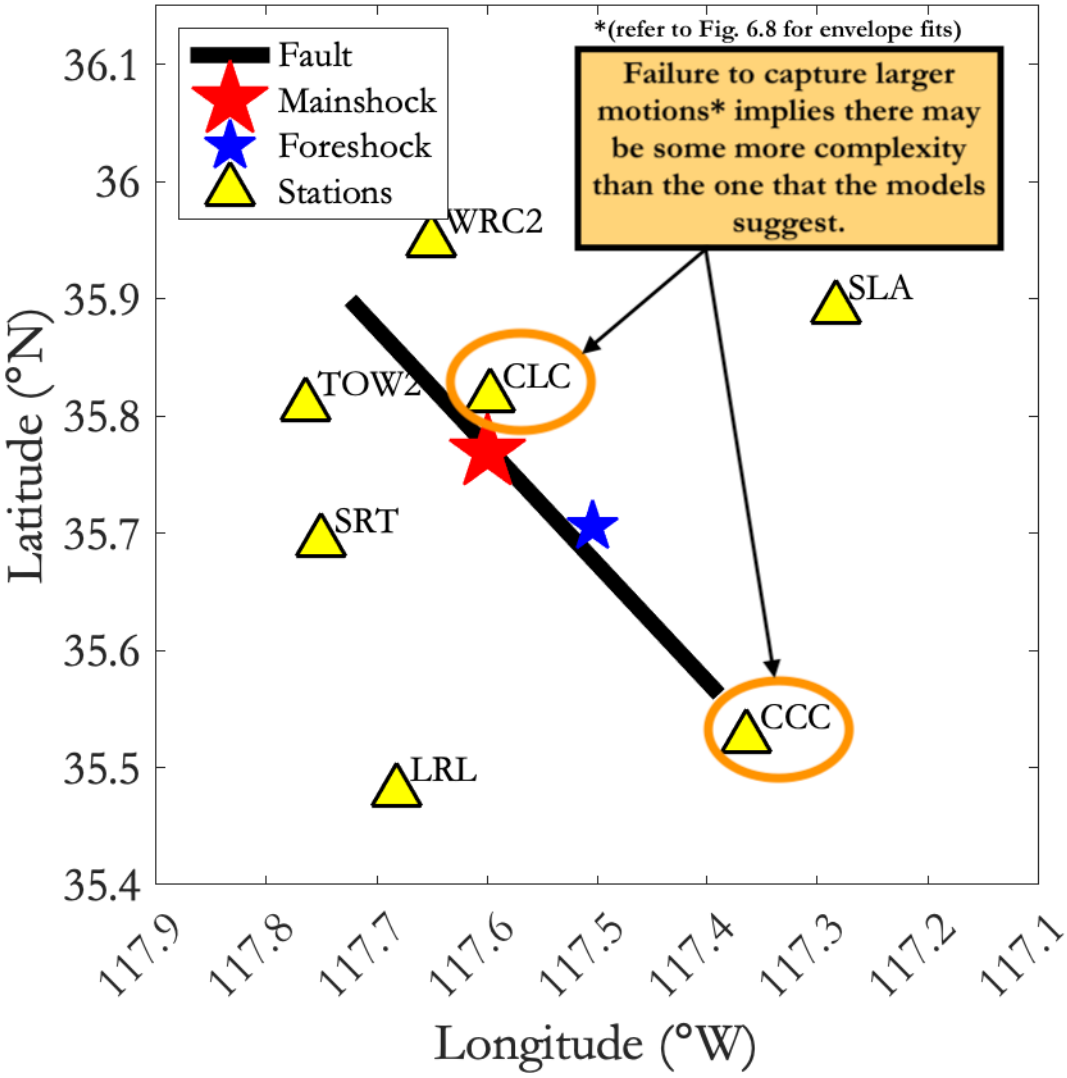


Figure 6.9. Station locations with respect to the epicenter of the mainshock, epicenter of past event, and ruptured fault for the 2019 Ridgecrest mainshock and foreshock. This particular rupture propagation is bilateral in the northwest and southeast direction.

6.4 Summary

This chapter addresses real cases where point source characterization may not be appropriate. For larger earthquakes (M>6.5), finite fault characterization would capture larger ground motions that occur near the rupturing fault.

For real-time application of Method II, the extended catalog search, the catalogs would be pre-determined. All the computations that use catalogs assume the earthquake database is accessible. The basis in creating the extended catalogs is looking back 1 month in history, which would work sufficiently for sequences and regions of high seismicity. For regions of lower seismicity, however, 1 month of earthquake history may be insufficient, and the extended catalog will be constructed looking farther back in time. The catalogs must be updated with time as new events are added.

Case studies of the 2016 Kumamoto, 2010 El Mayor-Cucapah, and 2019 Ridgecrest show that sufficient earthquake history is needed for accurate parameter estimates with small error bands. Therefore, if seismicity is not high in the region, looking back only 1 month in earthquake history may not be adequate. One such case is the 2016 Kaikoura earthquake. This is a special case in which both Method I and Method II of the search algorithm do not provide accurate envelope fits. Here, the extended catalog search may look back more than 10 years in earthquake history and still fail to find envelopes that fit the incoming ground motions well. The grid search is also unable to capture the high complexity of the rupture. The best algorithm to use for the 2016 Kaikoura earthquake is FinDer. Therefore, it is suggested to have FinDer running in parallel as a form of confirmation. A case of moderate seismicity is the 2010 El Mayor-Cucapah. Here, the extended catalog search needs to look back 10 years in earthquake history for envelope fits that produce acceptable error bands. The extended catalog search performs best in cases of high seismicity, such as foreshock-mainshock sequences. For the 2016 Kumamoto and 2019 Ridgecrest earthquakes, available waveforms of the foreshock resemble those from the mainshock.

To address cases where seismicity is low, the pre-determined catalog is to be built based on the following criteria:

- Because smaller earthquakes (M<5) are more frequent by the Gutenberg-Richter law, waveform envelopes from 3 months in earthquake history are used to build the

original catalog. For larger earthquakes (M>5), waveform envelopes from year 2000 are used to build the catalog. The catalog is continuously updated with time.

- Spatial coverage is also taken into consideration when building the original catalog. For regions of high seismicity (spaces filled with red circles in Fig. 6.10), the catalog considers waveform envelopes from earthquakes where the stations are located within 100 km from the epicenter. But in regions where seismicity is low (empty space in Fig. 6.10), this threshold is extended to 200 km.

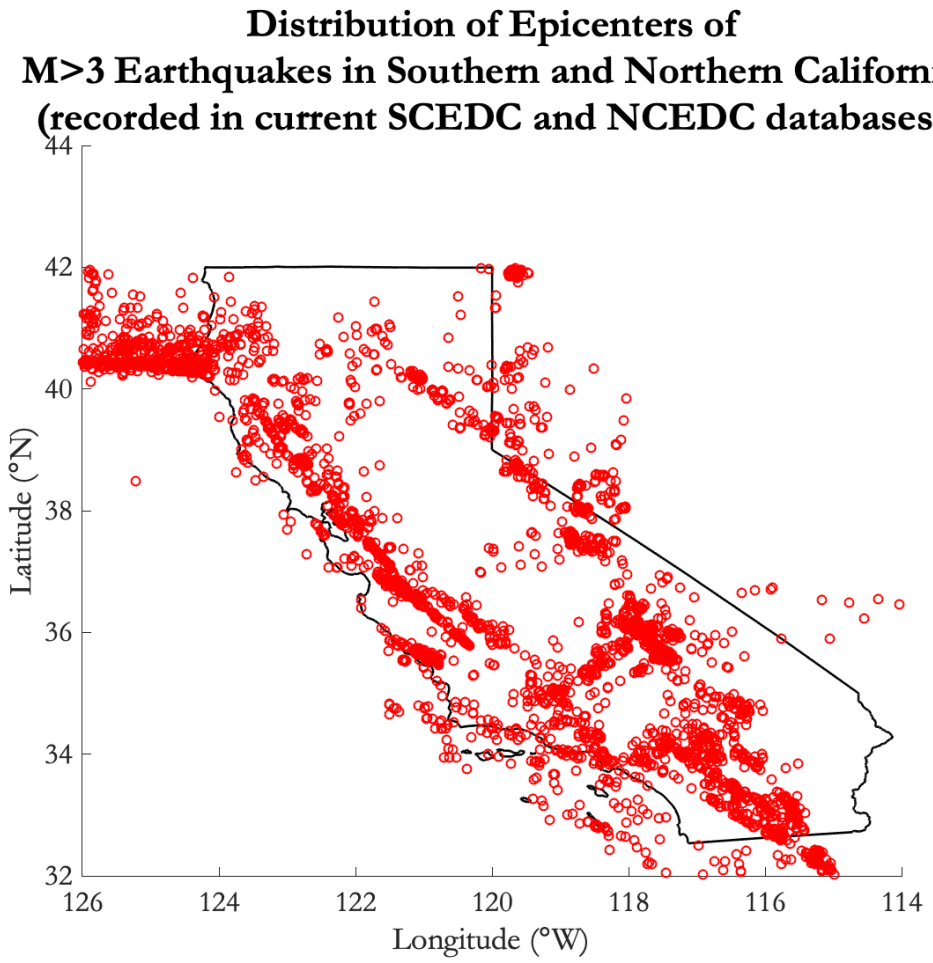


Figure 6.10. Distribution of M>3 earthquakes in the California region for years 2000 – 2020.

7 Parallel execution of Methods I and II

As previously mentioned, Methods I and II are independent. Separately, they find parameter estimates and error bands that capture the preciseness in the solutions. The total algorithm combines the two parts that work together to provide accurate parameter estimates as quickly as possible. To do so, error bands were defined in Chapter 3 and emphasized throughout this thesis. The final solution takes one of two cases. One case is when both Methods I and II have similar fits and provide similar parameter estimates. The user can have even higher confidence in these results as one method confirms the solutions of the other. If the error bands are similar, then the user takes both solutions, either by average or individually. The alternative case is when the methods provide different solutions, but one method has significantly smaller error bands, indicating stronger envelope fits.

7.1 Application to past real earthquakes

For a consistent comparison, this chapter refers to the same real earthquakes from Chapters 3 and 4.

7.1.1 2020 Northern coast offshore event

Seen in Fig. 7.1, the extended catalog search estimates M5.7, an error of 0.1 units, 9 seconds faster than the grid search. In fact, the cataloged envelopes from the extended catalog search fit more accurately by 28% than the Cua-Heaton envelopes from the grid search. However, 30 seconds after the origin time, envelopes from both methods have similar confidence in fitting the incoming observed ones (see Fig. 7.2).

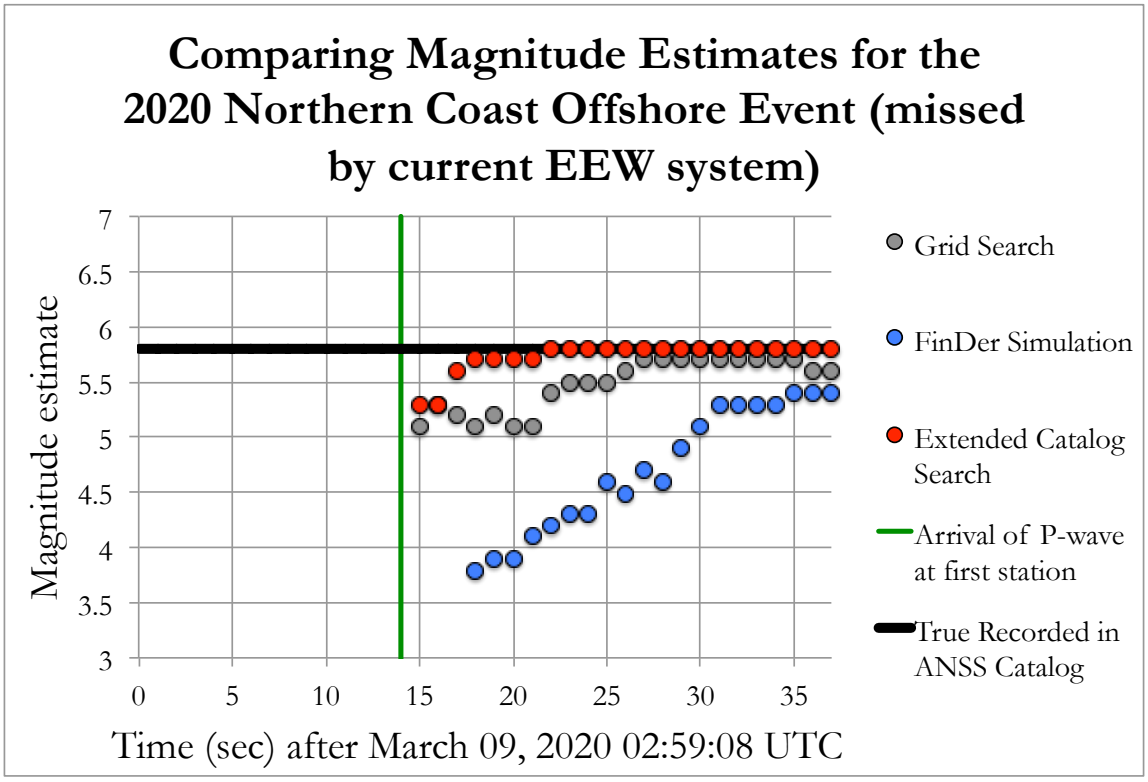


Figure 7.1. Comparing Methods I and II magnitude estimates for the 2020 Northern coast offshore event.

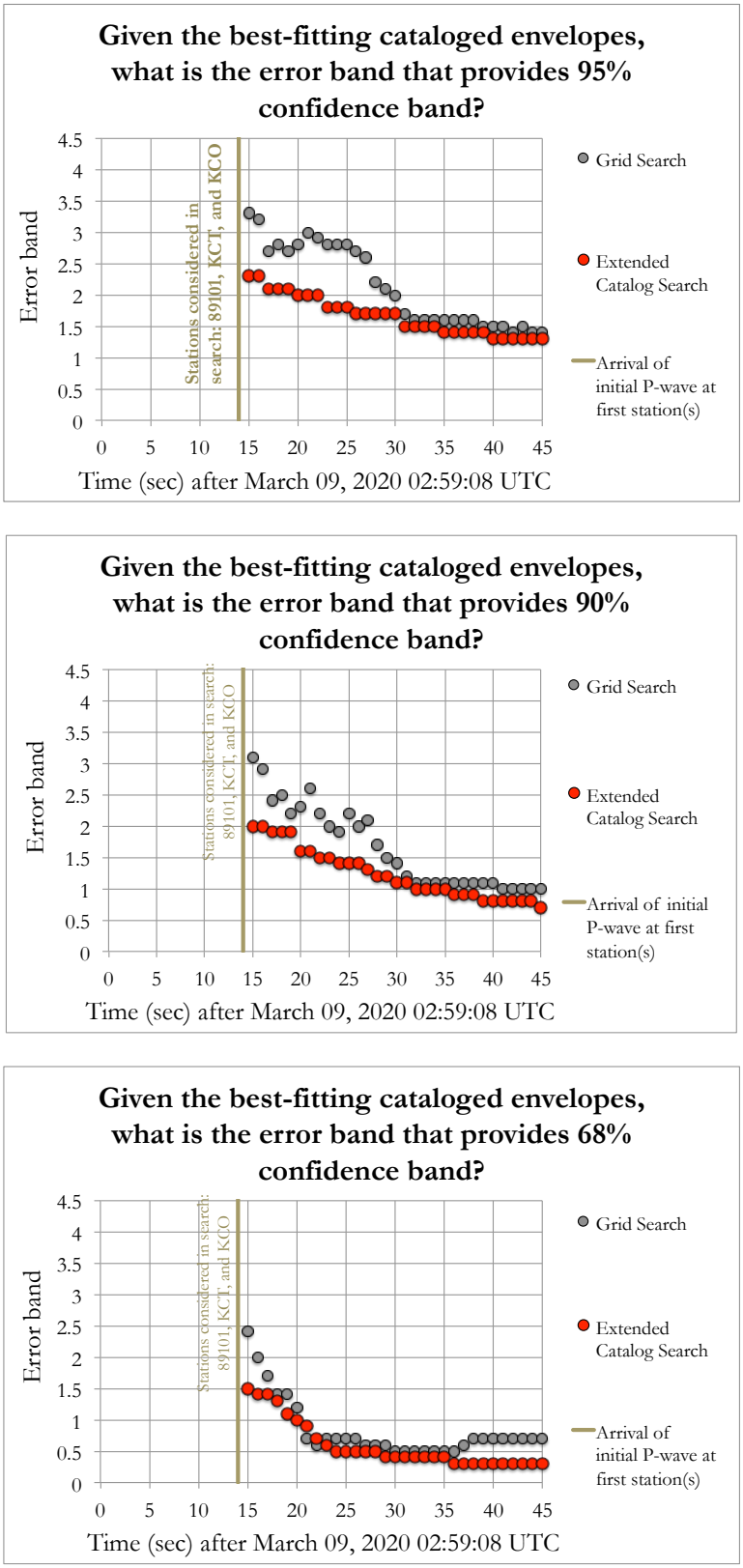


Figure 7.2. Comparing Methods I and II error bands for the 2020 Northern coast offshore event.

7.1.2 2020 Lone Pine foreshock-mainshock pair

Seen in Fig. 7.3, the extended catalog search instantly recognizes the incoming earthquake as M5.6 using the first 2 seconds of data. This estimate grows to M5.7 11 seconds after the origin time. On the other hand, the grid search takes 6 seconds of data to estimate M5.9. As previously mentioned, the two-part algorithm performs well with earthquakes that are part of a sequence because of the similar envelopes the foreshock provides. Comparing Figs. B and D, the gaps between the error bands are larger in this earthquake than those in the previous 2020 Northern coast offshore event, emphasizing the lowered confidence in the grid search estimates for the 2020 Lone Pine mainshock. Instead, the cataloged envelopes from the foreshock provide more accurate fits than the Cua-Heaton envelopes from the grid search not just for the initial time points but for the whole rupture.

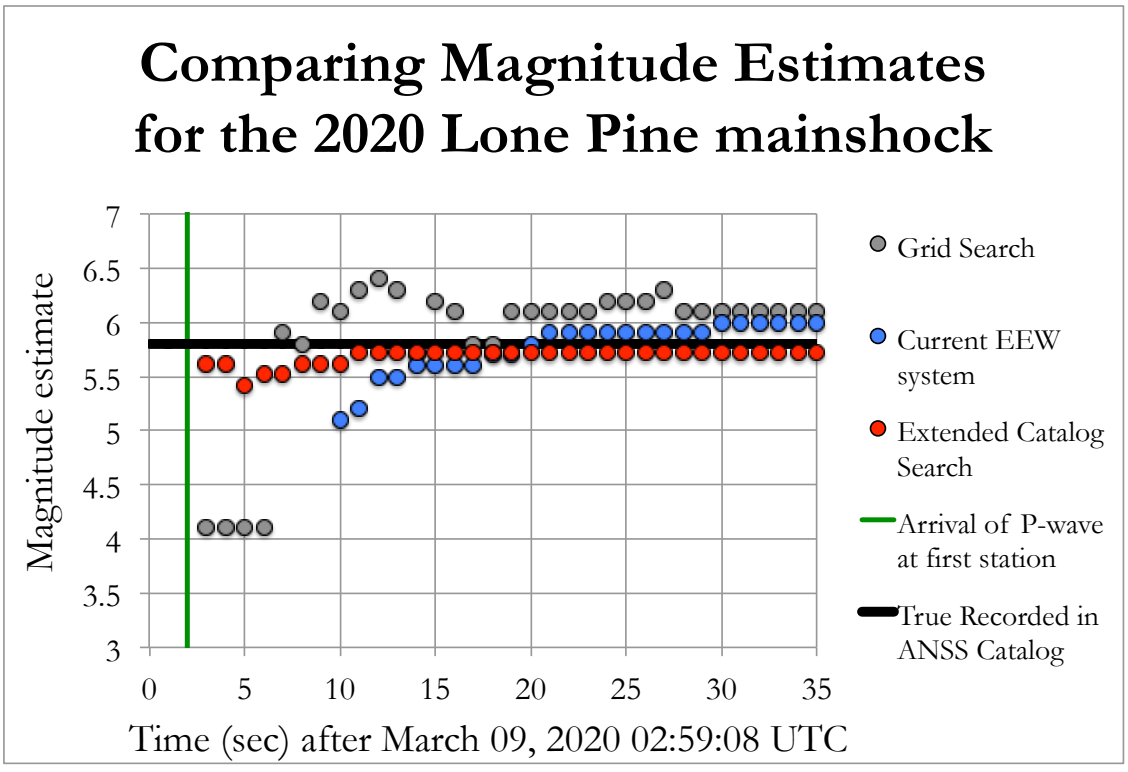


Figure 7.3. Comparing Methods I and II magnitude estimates for the 2020 Lone Pine event.

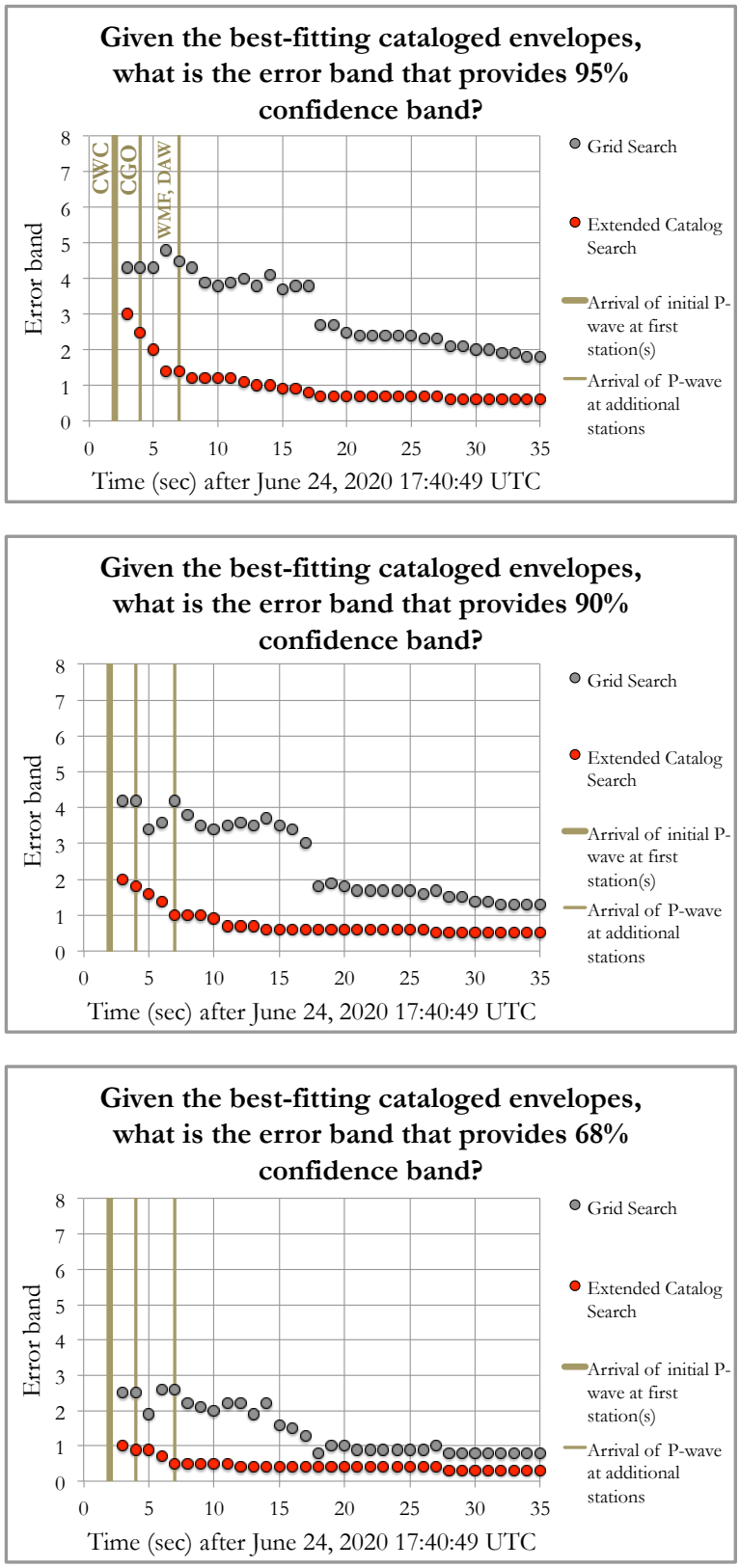


Figure 7.4. Comparing Methods I and II error bands for the 2020 Lone Pine event.

7.1.3 2019 Ridgecrest sequence

Seen in Fig. 7.5, the extended catalog search estimates the incoming earthquake as M5.4 using the first 4 seconds of data. This estimate grows to M6.5 10 seconds after the origin time and to M6.9 27 seconds after the origin time. On the other hand, the grid search estimates resemble those of the extended catalog search 13 seconds after the origin time. However, seen in Fig. F, the comparison of error bands shows the cataloged envelopes are stronger matches to the incoming envelopes than the Cua-Heaton envelopes by 63%.

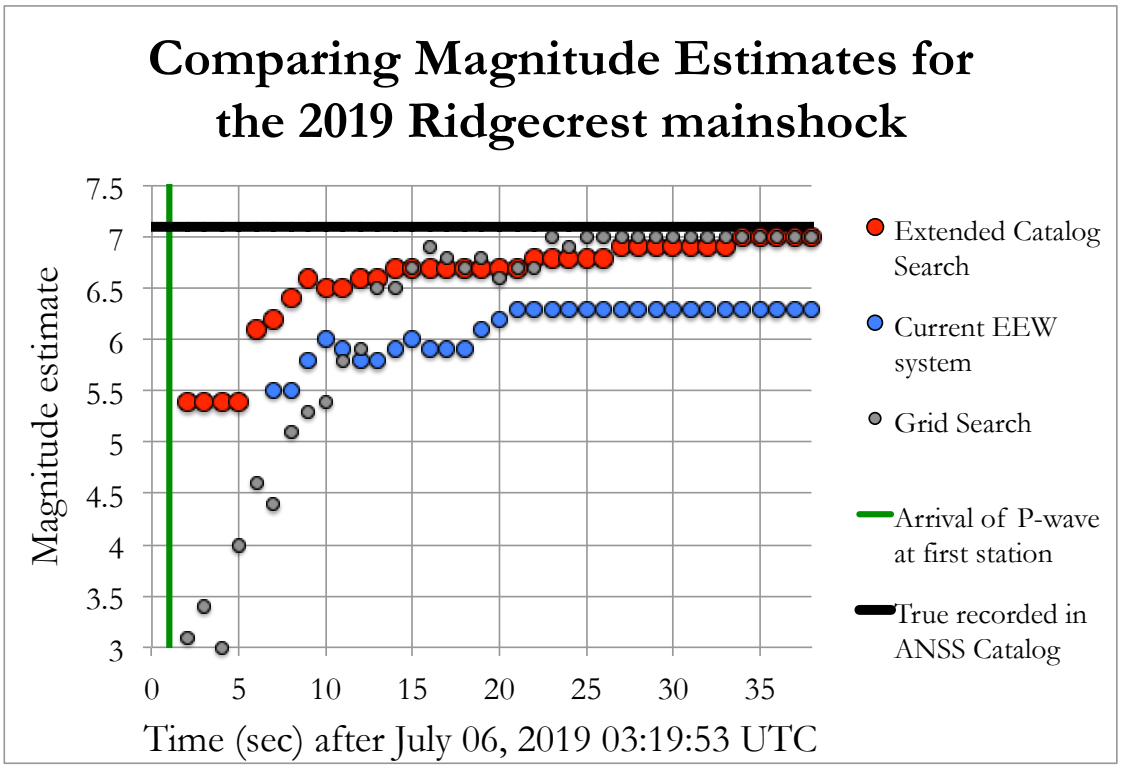


Figure 7.5. Comparing Methods I and II magnitude estimates for the 2019 Ridgecrest mainshock.

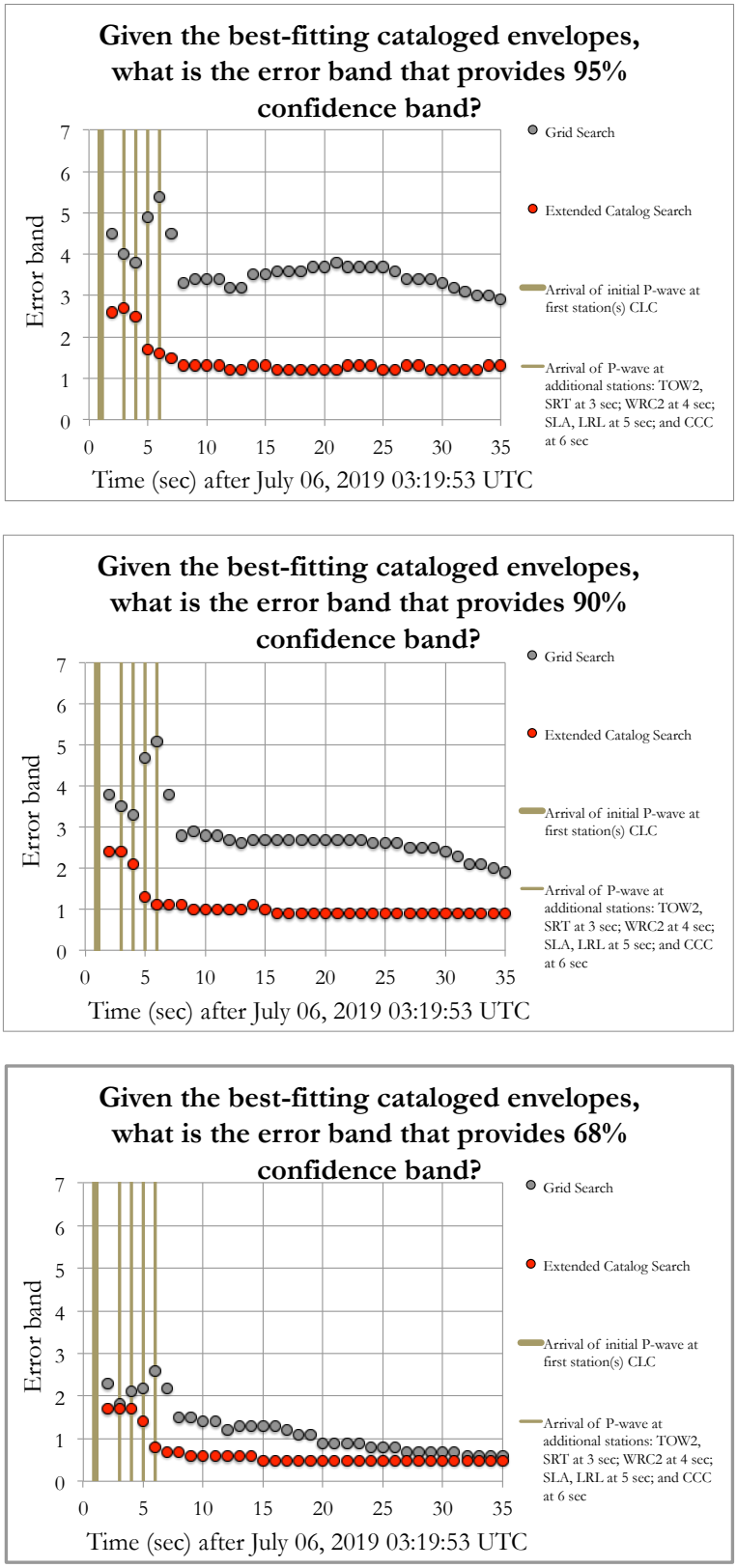


Figure 7.6. Comparing Methods I and II error bands for the 2019 Ridgecrest mainshock.

7.2 Summary

While the grid search is generally a robust method for different types of earthquakes, the extended catalog search provides more accurate envelope fits, most likely due to the specific consideration of the site and path effects at the specified stations. In other words, if the earthquake history at the specified stations contains ground motion envelopes that resemble those of the incoming observed earthquake, the extended catalog search is strongly recommended over the grid search. The error bands quantify how well the cataloged envelopes resemble those of the incoming observed earthquake. A test sweep on a variety of earthquakes is required to calculate the error bands that would be considered acceptable. This numerical analysis is beyond the scope of this thesis, therefore, a common acceptable threshold is used, which is a factor of 2.

Method I and Method II run in parallel to provide accurate parameter estimates. When both methods agree in error bands and magnitude estimates, such as in the 2020 Northern coast offshore event, the user may have high confidence in the solutions. If one method has stronger envelope fits than the other, such as in the Lone Pine and Ridgecrest mainshocks, the algorithm would choose the method of enhanced performance.

8 Prior Information

Current automated decision making modules in earthquake early warning (EEW) do not incorporate previous experience and judgment. However, previous experience and judgment may have the potential in providing faster, yet still reliable, earthquake source parameter estimates. The Virtual Seismologist (VS) method developed by Cua and Heaton incorporates prior information by taking a Bayesian approach. The idea of Bayesian approach is to imitate the analysis human seismologists would make, which is to combine previous and current data to make a well-informed judgment. This chapter describes the different types of prior information that target different parameter estimates, particularly the magnitude and location. As previously mentioned, the grid search and extended catalog search require waveform information to find parameter estimates. Unfortunately, waiting for sufficient waveform information decreases the warning time, compromising the regions close to the earthquake source that would experience strong shaking. Finding estimates using both prior and waveform information would make the algorithm much faster without jeopardizing the confidence of the initial estimates.

8.1 Introduction

In previous chapters, the posterior probability to maximize is essentially the waveform-based likelihood. A uniform prior is assumed, meaning every magnitude and location is equally likely. However, looking at past earthquake seismicity, this is not true as earthquakes cluster in time and in space. Different types of prior information are applied to the waveform-based likelihood. The purpose is to reduce uncertainties in the initial estimates, especially for a single-station approach. With time, as more data becomes available, the waveform-based likelihood has dominating influence over the solutions.

8.2 Seismicity Prior for Faster Event Detection

To maximize available warning time, the algorithm must decide whether available data is from an earthquake or noise as early as possible. The seismicity prior gives information specifically for two cases:

1. If a generally noisy station is in a region of low to no seismicity, then it is more likely for incoming data to be from noise.
2. If a generally quiet station is in a region of an earthquake sequence, then it is more likely for incoming data to be from an earthquake.

Before seismic data is used to identify the earthquake parameter estimates, it is initially processed to detect whether the incoming signal is from an earthquake or from noise. Finding estimates for non-earthquake data would result in false alerts. False alerts can bring negative impacts economically, such as financial losses resulting from emergency shutdowns in nuclear power plants, and even psychologically, such as extreme fear and flight reactions. Therefore, many EEW systems require triggers from multiple seismic stations to avoid false alerts. For instance, ElarmS, the current network-based algorithm in the ShakeAlert system, uses a minimum of four stations to alert (Chung et al. 2019). OnSite, a single-station approach in ShakeAlert, uses two stations but waits for the arrival of 3 seconds of data (Bose et al. 2012). Unfortunately, waiting for multiple stations may jeopardize the warning time for regions near the earthquake source, which is why it is valuable to have a method that detects an earthquake using 1 to 2 seconds of data from the closest one to two stations. Earthquake history provides important information unique to the station that has the potential to detect the incoming ground motion as an earthquake. A faster detection leads to a faster identification of the earthquake source parameters using the previously mentioned grid search and/or extended catalog search. A probabilistic approach is taken to detect if the incoming ground motion is an earthquake or ambient noise.

The prior information in distinguishing the incoming signal as an earthquake or ambient noise depends on the past seismicity at the specified station. Therefore, it is assumed the EEW system has access to this station-specific catalog of past earthquakes. It is also assumed the catalog is continuously updated with new earthquakes as time passes. Table 8.1 illustrates the format of this catalog. Eqs. 8.1-8.4 demonstrate the required calculations to transform the catalog of past earthquakes is to a probability, a probability that can distinguish whether

to initiate performing the waveform-based algorithms, such as the grid search and extended catalog search.

Table 8.1 Format of catalog to extract prior information from.

Past earthquake from catalog	Corresponding information				
Origin time (UTC)	Magnitude	Longitude	Latitude	Station-to-epicenter distance	Waveform envelopes

* For a single station and channel, the catalog lists the recorded earthquakes in a given time span. For each cataloged earthquake, there is a recorded magnitude, epicenter location (in terms longitude and latitude), epicentral distance from the station, and waveform envelopes.

$$Pr_E(pga) = \frac{R_E(>pga_{trigger})}{R_E(>pga_{trigger})+R_N(>pga_{trigger})} \tag{8.1}$$

where $Pr_E(pga)$ is the probability that the incoming ground motion is an earthquake, $R_E(> pga_{trigger})$ is the expected rate of earthquakes with $pga > pga_{trigger}$ (modeled using ETAS), $R_N(> pga_{trigger})$ is the rate at which noise exceeds $pga > pga_{trigger}$ (modeled using a lognormal distribution to account for envelopes that consider absolute value of amplitudes).

$$R_N(> pga_{trigger}) = \frac{1}{pga_{trigger}\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln pga_{trigger}-X}{\sigma}\right)^2\right] \tag{8.2}$$

where $R_N(> pga_{trigger})$ is the rate at which noise exceeds $pga > pga_{trigger}$, X and σ are the average and standard deviation, respectively, of the amplitudes of past noise recorded at the specified station and channel. If the station is known to have a history of noise exceeding $pga_{trigger}$, then this rate of noise would notify the system to be particularly cautious in saying the incoming signal is from an earthquake. $pga_{trigger} > 0$.

$$R_E(> pga_{trigger}) = \mu(> pga_{trigger}) + \sum_{i=1}^{N} R_i(> pga_{trigger}) \tag{8.3}$$

where $R_E(> pga_{trigger})$ is the expected rate of earthquakes with $pga > pga_{trigger}$ (modeled using ETAS), $\mu(> pga_{trigger})$ is the long-term background activity, and $\sum_{i=1}^{N} R_i(> pga_{trigger})$ is the short-term observed seismicity.

$$R_i\left(> pga_{trigger}\right) = \frac{K \cdot 10^{\alpha(M_i - M_{min})}}{(t - t_i + c)^p r^n} = \frac{0.008 \cdot 10^{1(M_i - M_{min})}}{(t - t_i + 0.095)^{1.34} r^{1.37}} \quad (8.4)$$

where $R_i\left(> pga_{trigger}\right)$ is the individual seismicity rate computed using the i^{th} earthquake with corresponding M_i recorded in the catalog, M_{min} is the minimum threshold of the magnitude of the forecast earthquakes, $t - t_i$ is the time (in days) between the forecast earthquakes and the i^{th} recorded earthquake, and K, α, c, p, and n, are constants in the ETAS model that are dependent on the region. The following values used for the constants in the equation above are based on the Southern California region (Felzer 2009).

A detailed description on the application of Eqs. 8.1-8.4 to real earthquakes may provide a clearer understanding. The following sections describe the steps and result of the seismicity prior for the real M5.8 Northern coast offshore event and the 2019 Ridgecrest sequence.

8.2.1 2020 Northern coast offshore event

The current EEW system has difficulty in detecting and identifying offshore events, particularly those near the Northern coast. A specific case is a M5.8 offshore event that occurred on March 09, 2020, at 02:59:08 UTC. This analysis looks at the data from earthquake history of Station KCO, one of the closest stations to the observed epicenter, to calculate the probability of the incoming signal is one from an earthquake. If so, the algorithm will relay this information to decide whether to start performing the grid search and extended catalog search.

Table 8.2 lists the past earthquakes recorded at Station KCO in the region specified in Fig. 8.1. From each recorded earthquake, the amplitude associated with the P-wave arrival is extracted as well as the amplitudes of the pre-signal noise. A histogram is created using the amplitudes of the pre-signal noise, and a lognormal distribution is fitted. Varying the parameters of a lognormal distribution, the best-fitting model is found. This distribution is used to model the expected rate, $R_N\left(> pga_{trigger}\right)$. As seen in Fig. 8.2, as the amplitude increases with the arrival of a signal, the chance of it being noise tends to zero. A wider distribution represents a noisier station.

Once the proper noise model is found, the ETAS model is used to calculate the rate of earthquakes, denoted $R_E\left(> pga_{trigger}\right)$. This is thoroughly explained in the next section, Section 8.3 Location prior using ETAS model. Together, the seismicity prior,

denoted $Pr_E(pga)$, provides the probability of the incoming signal being one from an earthquake (see Fig. 8.3). A relatively high probability means the search algorithm can immediately start finding the best parameter estimates, without waiting for multiple stations to trigger to issue an alert. This can potentially save seconds of warning time.

Table 8.2. Earthquake history at Station KCO Channels HNE, HNN, and HNZ.

Origin time (UTC)	Magnitude	Epicenter Longitude	Epicenter Latitude	Station-to-epicenter distance
22-Feb-2020 19:14:27	4.3300	124.6736	40.2895	34.7837
19-Dec-2019 15:30:12	4.0600	124.3633	40.2758	8.5271
23-Jun-2019 03:53:02	5.6000	124.3000	40.2730	3.4064
12-Apr-2019 14:06:27	4.6000	126.8690	40.4110	221.3647
29-Mar-2019 04:55:25	4.1900	124.4655	40.4288	25.5409
24-Mar-2019 11:32:02	4.3500	125.1371	40.4190	76.0292
24-Feb-2019 21:05:12	4.0800	125.0716	40.3881	69.8634
03-Feb-2019 23:43:21	4.1900	124.4965	40.3070	20.3432
03-Feb-2019 22:18:12	4.4800	124.4741	40.2893	18.0291
02-Feb-2019 10:52:21	4.3000	124.4946	40.3006	20.0037
15-Jan-2019 20:20:20	4.0800	124.4723	40.3216	18.9329
13-Jan-2019 09:35:48	4.1000	124.9930	40.3750	63.0464
25-Jul-2018 05:06:06	4.5000	125.0520	40.3850	68.1674
07-May-2018 23:51:04	4.4700	125.3246	40.6770	101.0432
23-Mar-2018 03:09:36	4.7000	124.4910	40.4480	28.5717
22-Mar-2018 16:24:49	4.4200	124.3891	40.7513	55.9857
09-Mar-2018 06:01:28	4.4800	124.5423	40.2918	23.7711
03-Feb-2018 03:12:46	4.3300	125.4455	40.8010	116.6554
25-Jan-2018 17:24:33	5.0000	126.3863	40.4296	180.7753
25-Jan-2018 16:39:43	5.8000	126.3034	40.4541	174.0806
07-Jan-2018 19:01:00	4.5000	125.2450	40.3990	84.5110
29-Jul-2017 00:02:38	5.1000	125.1920	40.7640	96.5178
24-Jun-2017 21:22:03	4.0000	124.2990	40.2880	4.4648
08-Dec-2016 16:32:46	4.7000	126.3622	40.4270	178.7166
08-Dec-2016 14:49:45	6.6000	126.1937	40.4535	164.8531
05-Dec-2016 18:33:15	4.3500	124.3860	40.2785	10.4694
27-Oct-2016 06:37:23	4.1100	124.5465	40.3462	25.7924

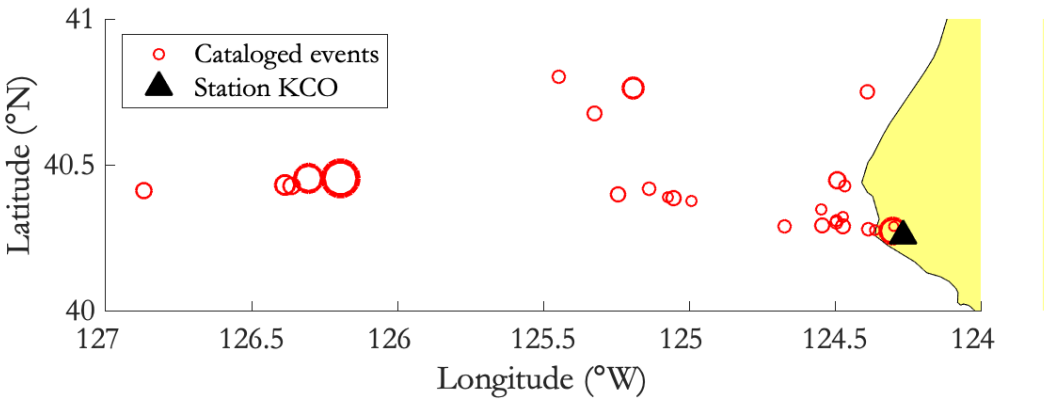


Figure 8.1. Earthquake history of M>4 earthquakes from years 2015-2020 in region constrained (40°N, 127°W) to (41°N, 124°W). These events are extracted from the NCEDC catalog.

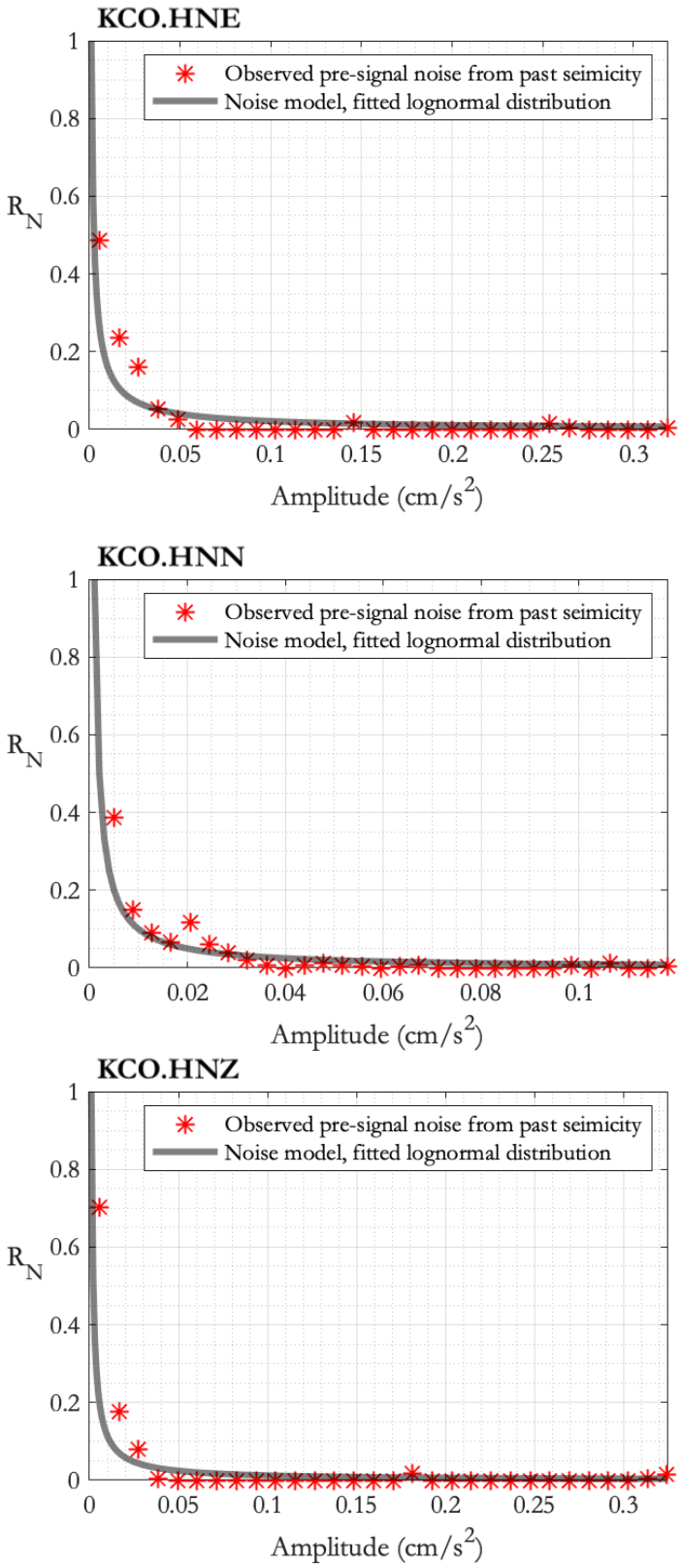


Figure 8.2. Noise model based on the pre-signal noise data from the catalog in Table 8.2.

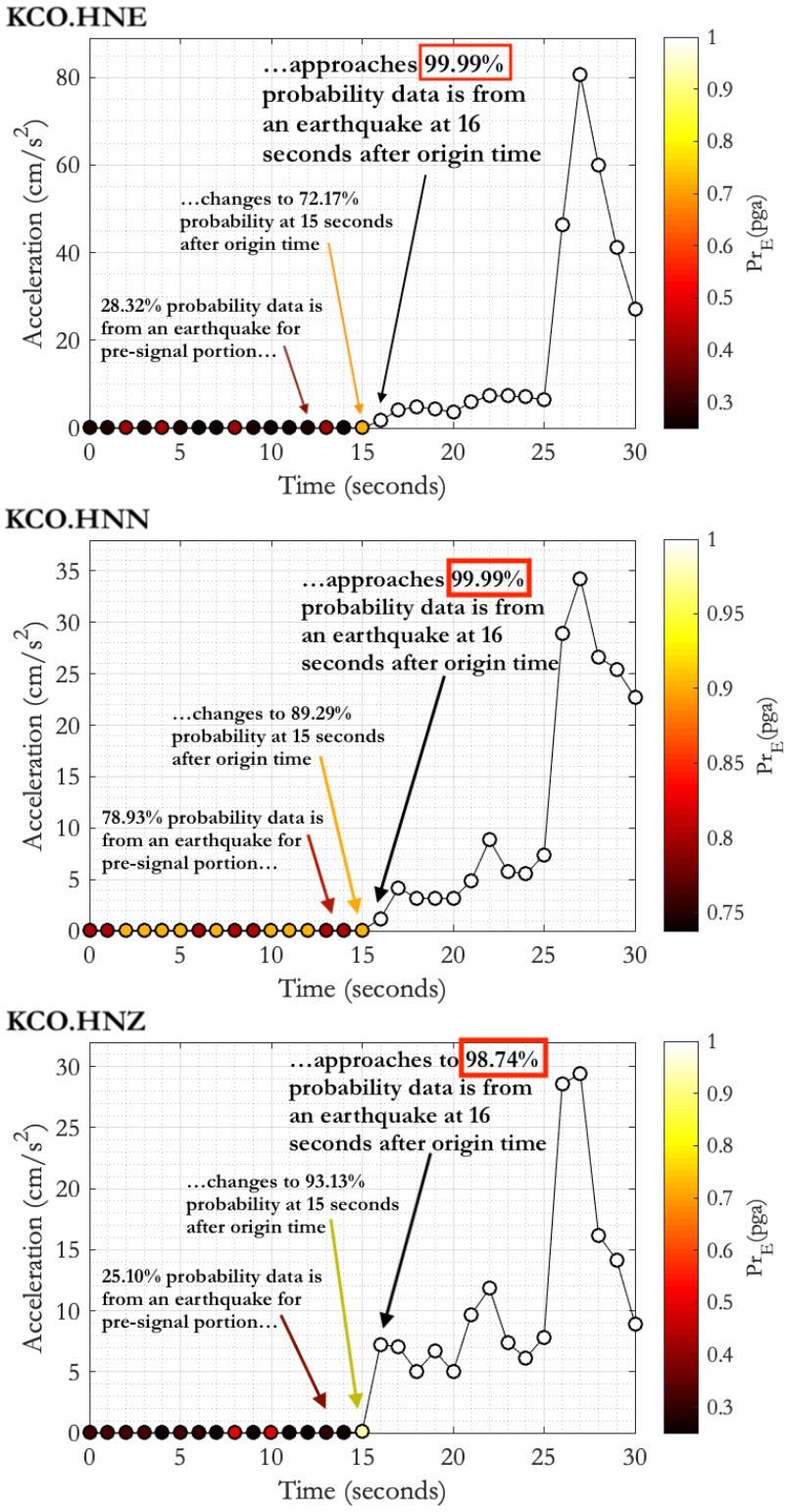


Figure 8.3. Application of event detection prior to the 2020 Northern coast offshore event. With each amplitude of the ground motion envelopes, there is a corresponding probability, $Pr(> pga_{trigger})$, which is the probability that the incoming ground motion is an earthquake.

8.2.2 2019 Ridgecrest sequence

Similar to the format of Table 8.2, the catalog for the Ridgecrest sequence consists of recorded earthquakes unique to a station. This complex sequence consists of 124 M>3 earthquakes recorded at a station that are within the specified region in Fig. 8.4. The time span of this catalog is 1.4 days before the observed mainshock.

As before, taking the pre-signal noise data from the recorded earthquakes listed in Table 8.2, the expected rate of noise, $R_N(> pga_{trigger})$, is modeled after a lognormal distribution. Seen in Fig. 8.5, as the amplitude increases, the chance of it being noise tends to zero. Again, a wider distribution represents a noisier station. The proper depiction of noise by the lognormal model allows the calculations to be more cautious when labeling a signal an earthquake. Together, the rate of noise and rate of earthquakes are used to calculate the seismicity prior. A probability is assigned to each amplitude.

For a consistent comparison to the performance of ElarmS, the first four triggered stations are considered: CLC, TOW2, SRT, and WRC2. ElarmS waits until four stations are triggered before declaring an event. This is detrimental in terms of EEW as the P-wave arrives at the fourth station 5 seconds after the origin time. The corresponding blind zone has a radius of approximately 14 km. As shown in Fig. 8.6, the seismicity prior finds that there is between 94.55% and 99.03% probability that an event has occurred after only 1 second after the origin time. A high probability almost immediately after the rupture results in a blind zone that has a radius of less than 3 km. In a region of high seismicity, especially during a sequence, the seismicity prior can reduce uncertainties in a single-station approach.

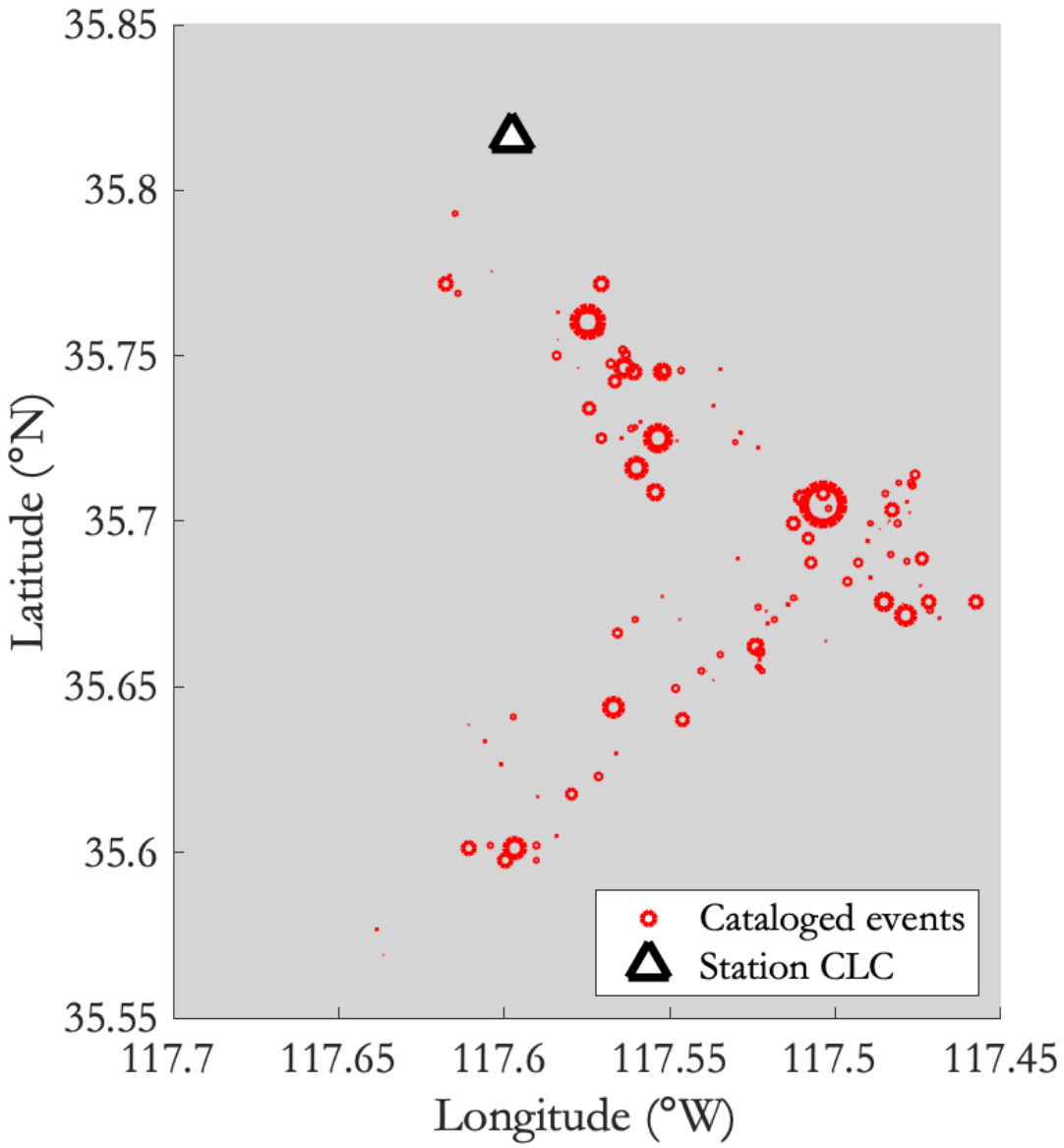


Figure 8.4. Earthquake history at Station CLC Channels HNE, HNN, and HNZ. The size of the circles refers to the earthquake magnitude. The larger it is, the larger the recorded magnitude.

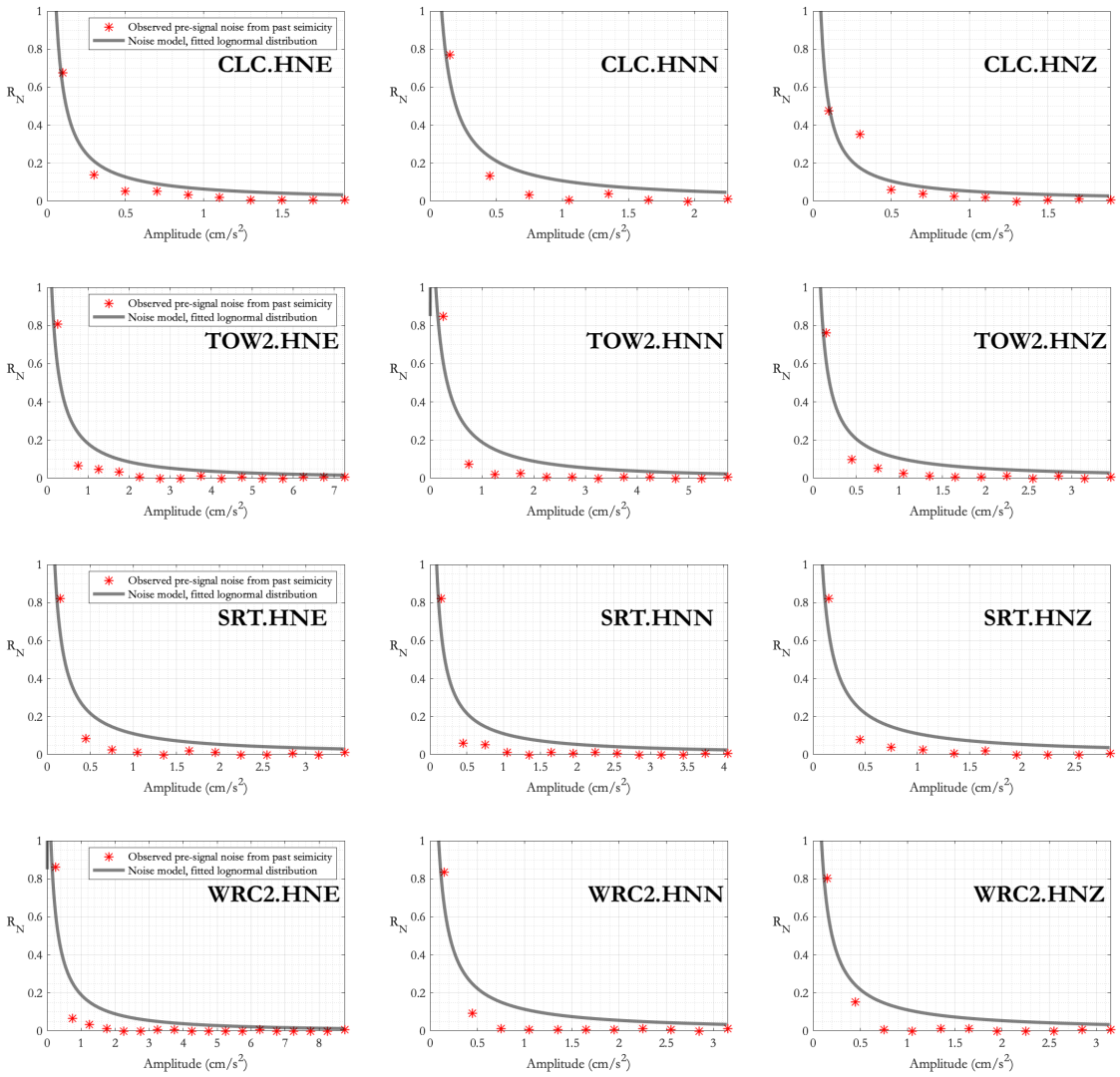


Figure 8.5. Noise model based on the pre-signal noise data from the catalog in Fig. 8.4.

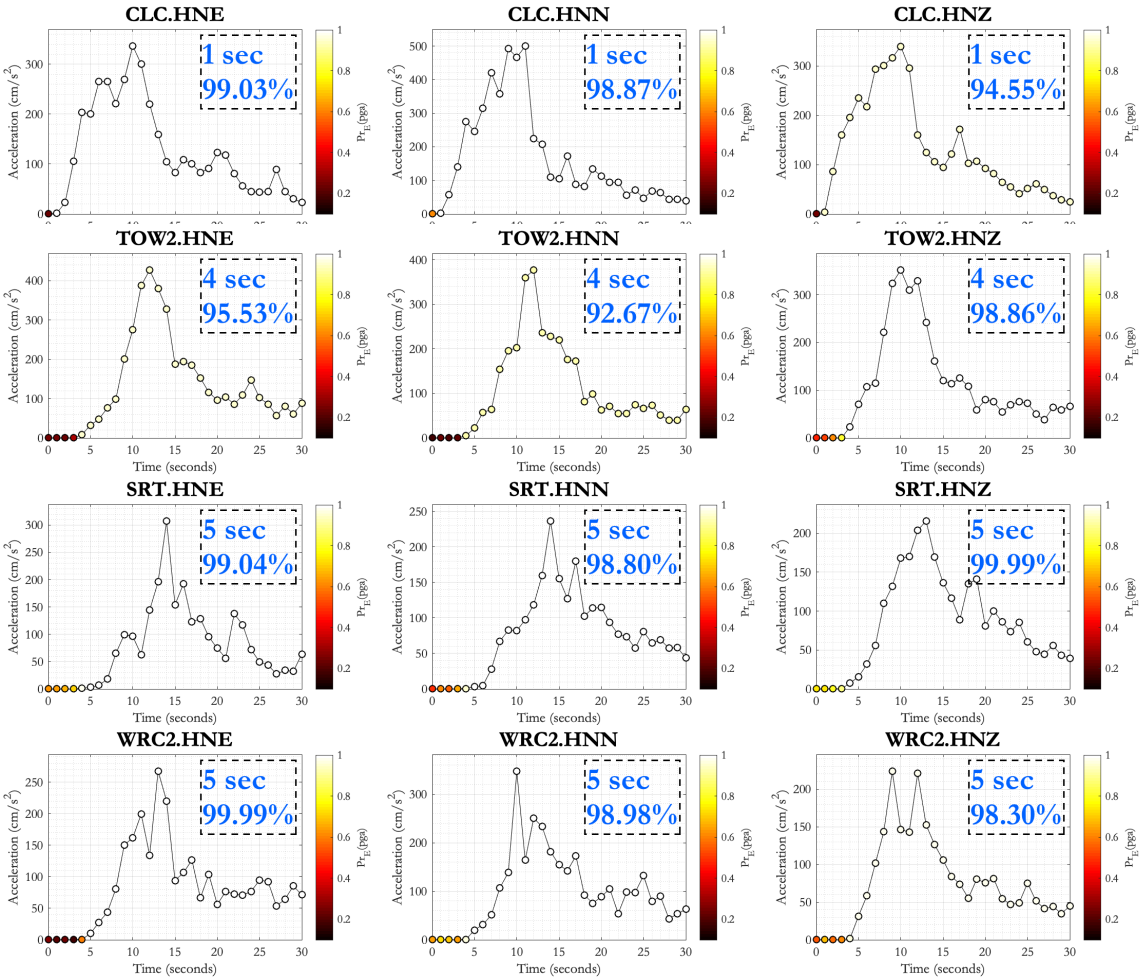


Figure 8.6. Application of seismicity prior to the 2019 Ridgecrest mainshock. With each amplitude of the ground motion envelopes, there is a corresponding probability, $Pr(> pga_{trigger})$, which is the probability that the incoming ground motion is an earthquake.

8.3 Location prior using ETAS model

Once the probability exceeds the threshold to announce the incoming signal is from an earthquake and not noise, the waveform-based algorithm can start analyzing the data to find parameter estimates. Another prior is the epidemic-type aftershock sequence (ETAS) model. This prior is in the form of an earthquake occurrence probability and is calculated using an ETAS model, a model that generates aftershocks stochastically using established empirical relationships for the distribution of aftershocks as a function of magnitudes, times, and locations (Ogata 1998). Defined in Eqs. 8.5-8.6, this earthquake occurrence probability, for magnitudes $M > M_{min}$, at location (lat, lon) at T days after a specified time point is modeled as a nonhomogeneous Poisson process with respect to time.

$$Pr_{ETAS}(lat, lon) = 1 - \exp\left(\int_0^T \lambda(t, lat, lon)dt\right) \tag{8.5}$$

where $\lambda(t, lat, lon)$ is the rate of earthquakes at current time and at location (lat, lon).

$$\lambda(t, lat, lon) = \sum_{j=1}^{N} \lambda_j\,(t, lat, lon) \tag{8.6}$$

where N is the amount of earthquakes of magnitudes $M > M_{min}$ observed in past seismicity. While the full ETAS model uses the long-term background seismicity as well, this analysis focuses only on the short-term past observed seismicity.

As previously mentioned, to calculate the rate of the short-term past observed seismicity, it is assumed the EEW system has access to the catalog of past seismicity that is continuously updated with time. Seen in Eq. 8.7, the individual rate of the short-term past seismicity for the j^{th} earthquake recorded in the catalog is modeled after the following:

1. Omori's law (Utsu 1961): the frequency of aftershock decays hyperbolically with time after a strong earthquake.
2. Gutenberg-Richter's law (Gutenberg & Richter 1944): the magnitude-frequency distribution of earthquakes (there are 10 times more earthquakes of M_i than earthquakes of magnitude $M_i + 1$).
3. Felzer and Brodskey relation (Felzer & Brodskey 2006): for short times after the mainshock, the decay of aftershocks in space is modeled after a single inverse power law.

$$\lambda_j(t, lon, lat) = \lambda_j(t, R) = \frac{K}{(\Delta_j + c)^p R^n} 10^{\alpha(M_j - M_{min})} \quad (8.7)$$

where Δ_j is the difference from the current time to the origin time of the j^{th} earthquake recorded in the catalog, R is the distance from (lon, lat) to the epicenter of the j^{th} earthquake, M_j is the magnitude of the j^{th} earthquake, and M_{min} is the minimum magnitude of the forecast earthquakes. The chosen values for the constants, K, c, p, n, and α, are based on the ETAS model for Southern California (Felzer 2009).

Earthquake history, especially recent seismicity, provides important information, even before any waveform information is collected. Since seismic activity clusters in time and space, the EEW system can use this prior information to deduce recent seismicity may in fact turn into a foreshock of a larger earthquake (Reasenberg & Jones 1989). Fig. 8.7 illustrates the ETAS model using real earthquakes from years 2016 to 2020 in the Northern coast offshore region, and Fig. 8.8 illustrates it for the Ridgecrest region. The prior information is calculated assuming the EEW system has access to the catalog of seismicity.

8.3.1 2020 Northern coast offshore event

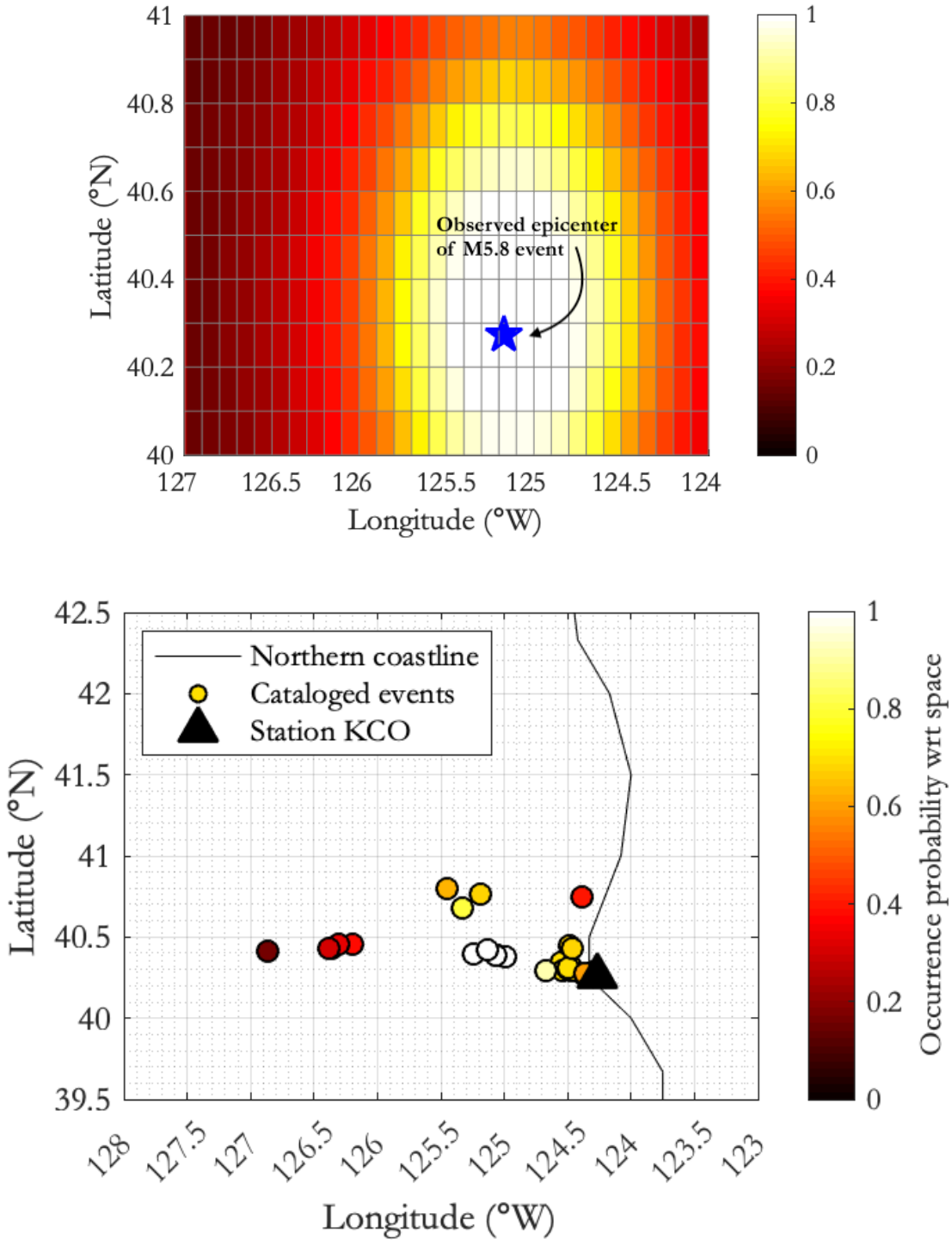


Figure 8.7. Occurrence probability by ETAS model of estimated location for the 2020 Northern coast offshore event, using prior information (no waveforms involved).

8.3.2 2019 Ridgecrest sequence

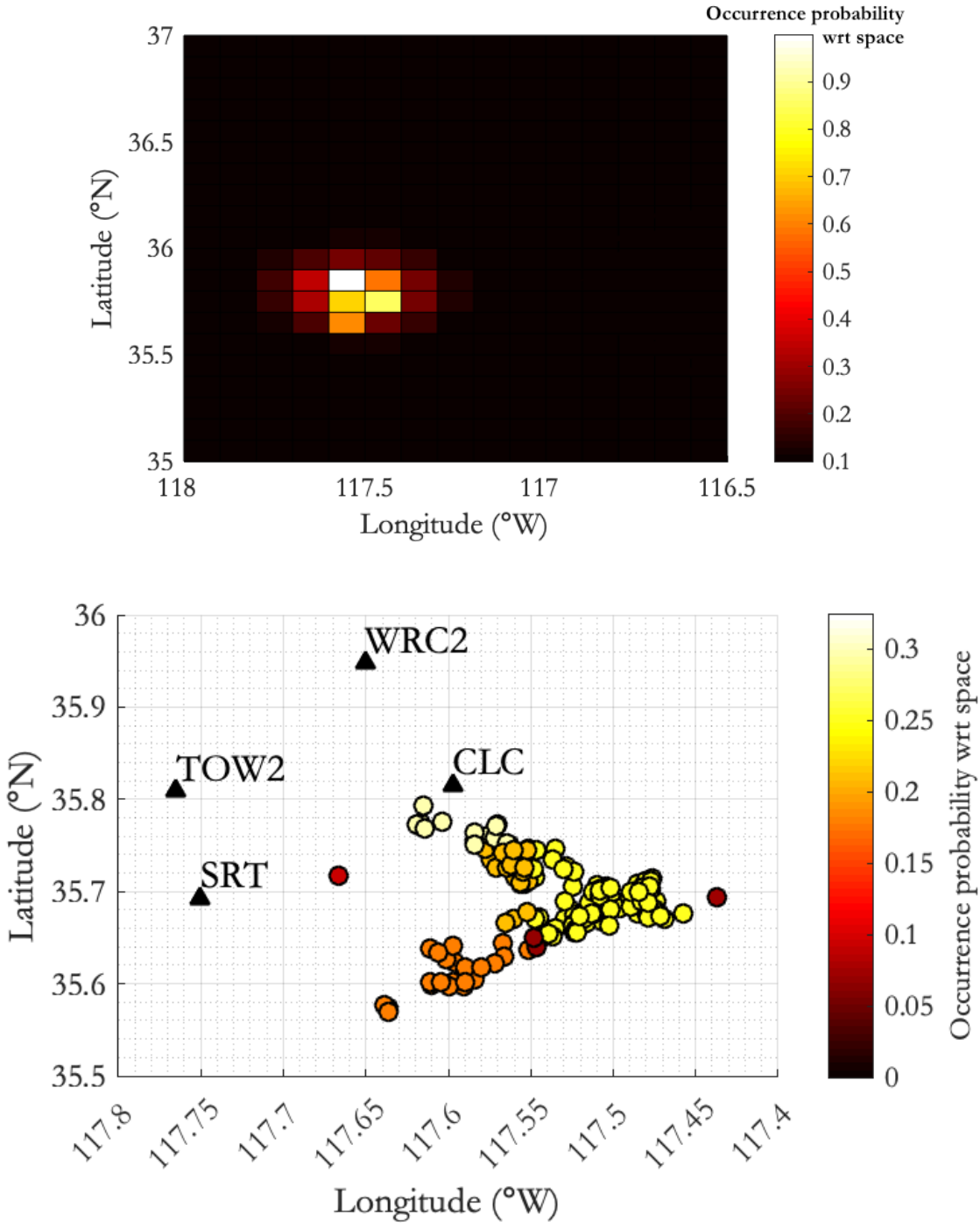


Figure. 8.8. Occurrence probability by ETAS model of estimated location for the 2019 Ridgecrest mainshock, using prior information (no waveforms involved). The catalog here is based on 1.4 days of earthquake history before the observed M7.1 mainshock.

8.4 Magnitude Estimate Using Amplitude Ratios

One valuable aspect written in Cua's thesis is an approach that finds the magnitude estimate without any attenuation relationships that are used in calculating the likelihood. For instance, attenuation relationships do not help avoid trade-offs between magnitude and location when data is only available from one station. However, incoming ground motion at this single station is still valuable because the ratios between acceleration and displacement reveal information about the frequency content. The different frequency content indicates different energies that are radiated, which then distinguish the incoming ground motion as one from a small earthquake or from a large one.

Cua takes this idea and applies a linear discriminant analysis on real earthquake records to best separate them into groups of different earthquake magnitude ranges. For instance, the first group consists of magnitudes less than 3, while another consists of magnitudes between 3 and 4 and so on. The process of a linear discriminant analysis also aims to maximally cluster within the group. The result from this analysis is a relationship that constrains magnitude estimates from available observed data (seen in Eq. 8.8-8.9). However, Eqs. 8.8-8.9 only show how the best magnitude estimate is found. Eq. 8.10 allows Eqs. 8.8-8.9 to be expressed in a probabilistic sense.

$$Z = \frac{PA^{0.36}}{PD^{0.93}} = 0.36 \log PA - 0.93 \log PD \tag{8.8}$$

where PA and PD are the peak acceleration and displacements in a specified time window.

$$M_{LDA} = \begin{cases} -1.627Z + 8.94 \; for \; P - wave \; amplitudes \\ -1.459Z + 8.05 \; for \; S - wave \; amplitudes \end{cases} \tag{8.9}$$

where M_{LDA} is the magnitude estimate found using relationships based on linear discriminant analysis on amplitude observations. The uncertainties are 0.45 and 0.41 for P-wave and S-wave amplitudes, respectively.

$$Pr_{Mag}(M_{arb}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(M_{arb}-M_{LDA}))^2}{2\sigma^2}\right) \quad (8.10)$$

where $P_r(M_{arb})$ is the likelihood of observing an arbitrary magnitude, M_{arb}, and M_{LDA} is the magnitude estimate found using the amplitude ratios in Eqs. 8.8-8.9. The uncertainty, σ, depends on the phase, which can be found using a standard STA/LTA analysis or the polarization analysis (Ross et al. YEAR). As previously stated, it is 0.45 for P-wave and 0.41 for S-wave. Because Eq. 8.8 and Eq. 8.9 consider waveform information, Eq. 8.10 should not be mistakenly considered prior information. It is merely an additional constraint on the parameter estimates to reduce tradeoffs.

8.4.1 2020 Northern coast offshore event

Table 8.3. Magnitude estimates using P-wave amplitudes for the 2020 Northern coast offshore event.

Network	Site Name	Component	Magnitude estimates
NC	KCO	EW	5.6
		NS	5.5
		UD	5.9
CE	89101	EW	6.4
		NS	5.9
		UD	6.3
NC	KCT	EW	6.0
		NS	6.1
		UD	6.3
BK	PETL	EW	5.9
		NS	5.9
		UD	6.2

8.3.2 2019 Ridgecrest sequence

Table 8.4. Magnitude estimates using P-wave amplitudes for the 2019 Ridgecrest mainshock.

Network	Site Name	Component	Magnitude estimates
CI	CLC	EW	6.7
		NS	7.2
		UD	6.8
CI	TOW2	EW	6.7
		NS	6.6
		UD	6.8
CI	SRT	EW	7.0
		NS	7.1
		UD	6.8
CI	WRC2	EW	6.3
		NS	7.0
		UD	6.9
CI	SLA	EW	7.2
		NS	7.1
		UD	7.1
CI	LRL	EW	6.7
		NS	6.8
		UD	7.0
CI	CCC	EW	6.9
		NS	7.2
		UD	6.8

8.5 Bayes' Theorem: Applying Prior to Likelihood

So far, the priors have been defined in terms of probability: (i) Eq. 8.1 distinguishes a trigger as one from an earthquake versus noise, (ii) Eq. 8.5 estimates the epicenter location based on previous observed seismicity, and (iii) Eq. 8.10 estimates the earthquake magnitude based on incoming ground motion amplitudes. To apply these priors to waveform-based likelihoods, Bayes' theorem is used. These priors have the potential to reduce uncertainty in the initial waveform-based likelihoods (i.e. estimates found in the first few seconds). For instance, waveform-based likelihoods may give a large range of possible parameter combinations to describe the incoming ground motion, but prior information may constrain them. Once more data is available with more time, the emphasis of the priors decays while the waveform-based likelihoods take over.

Bayes' theorem is defined as a normalized product of a prior probability density function (PDF) and a likelihood function, as seen in Eq. 8.11. Eqs. 8.12-8.14 show how to apply the general case of Bayes' theorem, seen in Eq. 8.11, to a single station and channel.

$$Pr_{post}(A|B) = \frac{Pr_{like}(B|A) \times Pr_{prior}(A)}{Pr_{norm}(B)} \propto Pr_{like}(B|A) \times Pr_{prior}(A) \tag{8.11}$$

where $Pr_{post}(A|B)$ is the posterior, $Pr_{like}(B|A)$ is the likelihood, $Pr_{prior}(A)$ is the prior, and $Pr_{norm}(B)$ is the normalizing constant. $Pr_{like}(B|A)$ is the likelihood calculated using waveform-information. The priors defined earlier in this chapter (Eqs. 8.1, 8.5, and 8.10) would be used as $Pr_{prior}(A)$ in Eq. 8.11.

$$Pr_{post}(M, lat, lon|Y) \propto Pr_{like}(Y|M, lat, lon) \times Pr_{prior}(M, lat, lon) \tag{8.12}$$

$$Pr_{like}(Y|M, lat, lon) = \prod_{i=1}^{n} Pr_{like}(Y_i|M, lat, lon) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(Y_i - X(M,lat,lon)_i)^2}{2\sigma^2}\right) \tag{8.13}$$

$$Pr_{prior}(M, lat, lon) = Pr_{Mag}(M) \times Pr_{ETAS}(lat, lon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(M_{arb} - M_{LDA}))^2}{2\sigma^2}\right) \times \left[1 - \exp\left(\sum_{j=1}^{N} \frac{K}{(\Delta_j + c)^p R(lat,lon)^n} 10^{\alpha(M_j - M_{min})}\right)\right] \tag{8.14}$$

Maximizing the posterior probability finds the magnitude and location estimates that best describe the incoming ground motions. However, to avoid computations with very small numbers, Eqs. 8.12-8.14 are rewritten in logarithmic form. This transforms the process of maximizing the posterior probability to minimizing the negative functions within the exponential. Doing so, the following estimates are found with respect to time after the mainshock ruptures for the 2020 Northern coast offshore event, 2019 Ridgecrest mainshock, and the 2012 Brawley mainshock. As seen in the figures below, the use of priors has the potential to provide relatively more accurate magnitude estimates well before the waveform-based likelihood can. For the Northern coast offshore event, using prior information would give 8 more seconds of warning time. For the Ridgecrest mainshock, using prior information would give 4 more seconds of warning time. For the Brawley mainshock, the waveform-based likelihood is sufficient but the prior information gives the

results more confidence. For all of these events, as time passes, the waveform-based likelihood has more weight in the posterior probability, and the results with prior and without prior converge.

8.5.1 2020 Northern coast offshore event

Prior information is not *always* useful. As seen in Fig. 8.10, prior information makes virtually no impact on the location and magnitude estimates for the extended catalog search. For this particular method, the waveform-based likelihood is sufficient. However, for the grid search, the prior information reduces the tradeoffs between the location and magnitude, leading to more accurate magnitude estimates in the initial time points (see Fig. 8.9). With the prior information reducing tradeoffs in the first few seconds estimates are available, it is assumed the first 2 seconds of P-wave data can be used to issue expected appropriate warning times (see Fig. 8.11). The first station is triggered 14 seconds after the origin time due to the location of the offshore event being approximately 60 km from the coast. Shown in this case, prior information is most valuable for regions of sparse station coverage. This way, uncertainties can be reduced without waiting for more stations to be triggered.

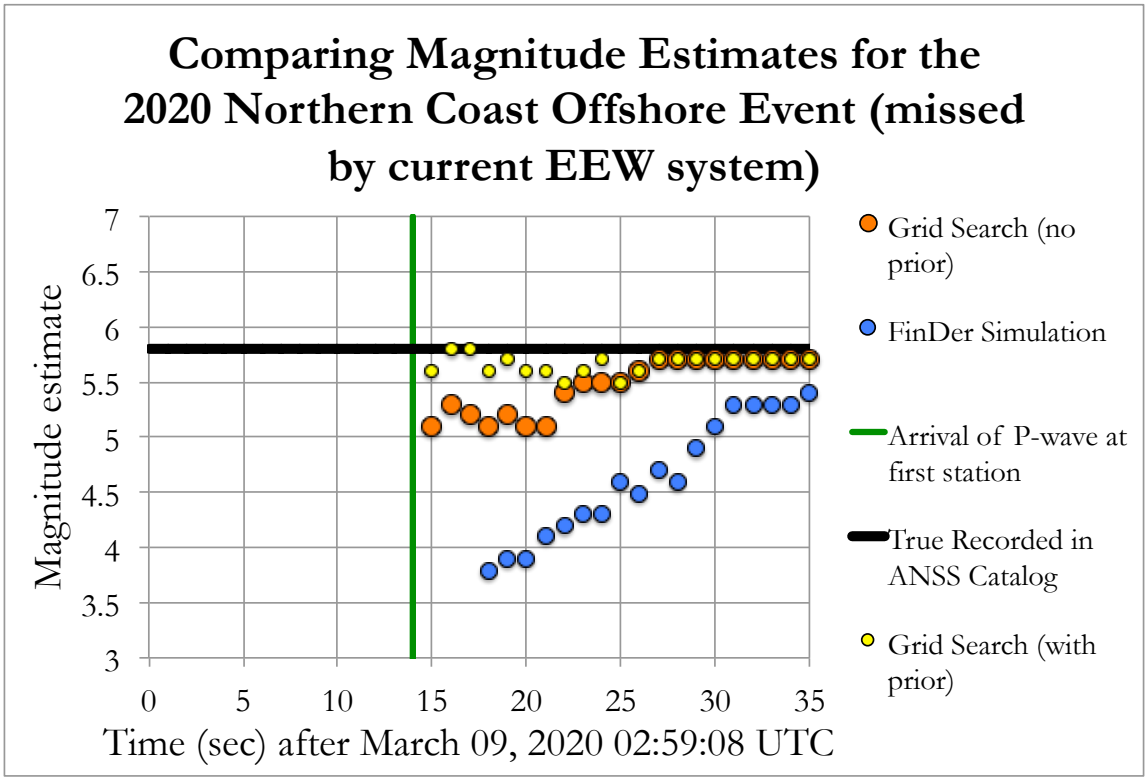


Figure. 8.9. Comparing magnitude estimates using waveform-based likelihood vs. using prior information for the 2020 Northern coast offshore event.

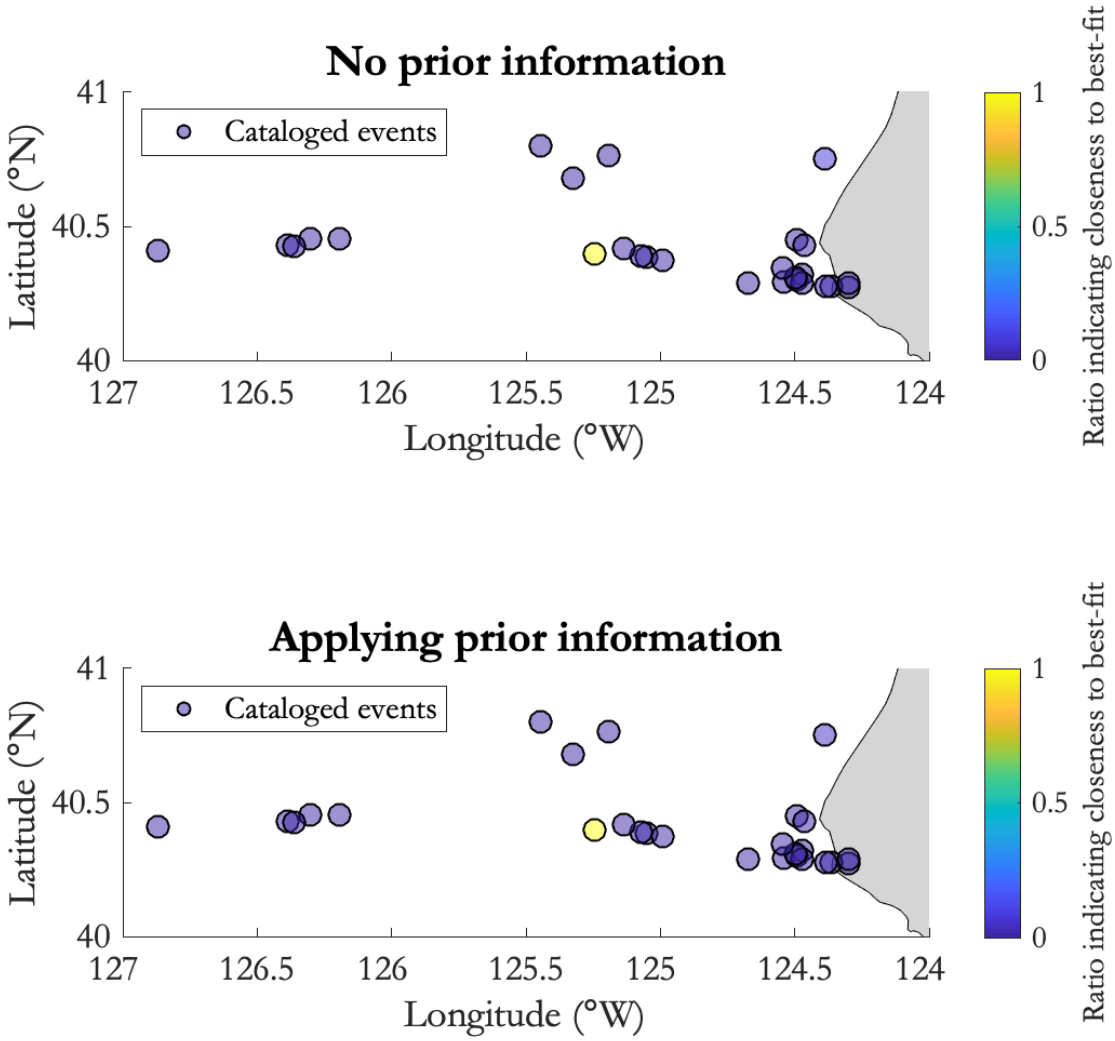


Figure 8.10. Applying prior information made little impact on location estimate in the extended catalog search for the 2020 Northern coast offshore event.

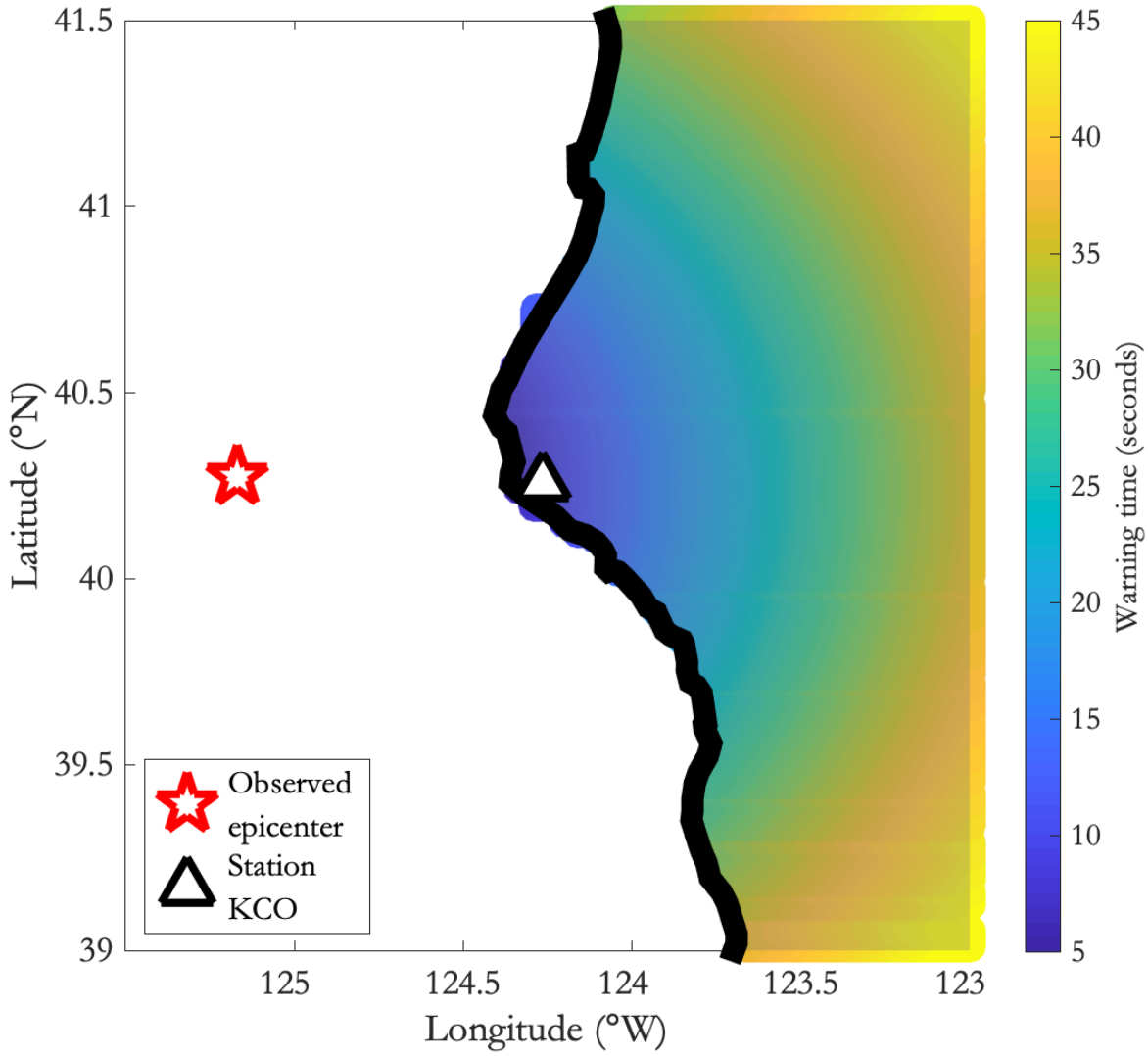


Figure 8.11. Warning times based on first parameter estimates at Station KCO, which occurs 16 seconds after origin time.

8.5.2 2019 Ridgecrest sequence

Prior information does not significantly change the initial parameter estimates for the extended catalog search. However, prior information makes a larger, more obvious impact on the grid search solutions. As seen in Fig. 8.13, prior information gives valuable insight on the location estimates. The magnitude estimates based on waveform-based likelihood are initially underestimated. The prior information provides some constraint on the location, providing magnitude estimates similar to the current EEW system, just 4 seconds earlier. It reduces the tradeoffs between the location and magnitude, leading to more accurate magnitude estimates in the initial time points (see Fig. 8.12). With the prior information reducing tradeoffs in the first few seconds estimates are available, it is assumed the first 2 seconds of P-wave data can be used to issue expected appropriate warning times (see Fig. 8.14).

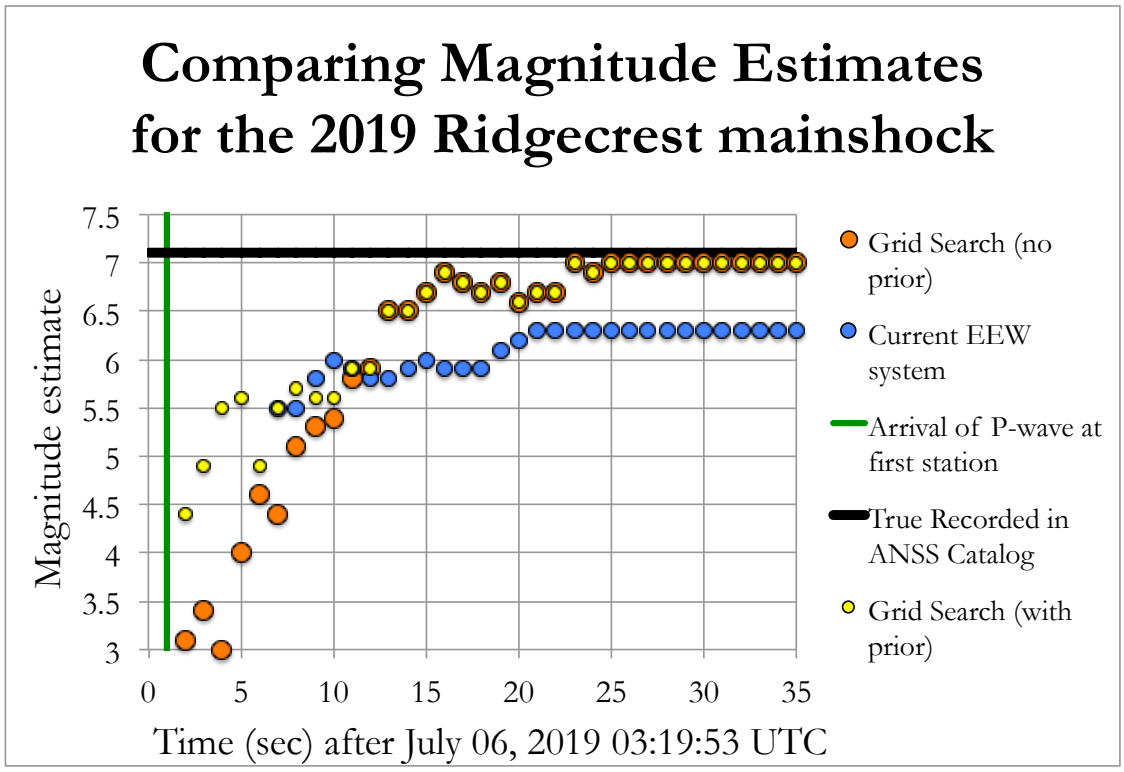


Figure 8.12. Comparing magnitude estimates using waveform-based likelihood vs. using prior information for the 2019 Ridgecrest mainshock.

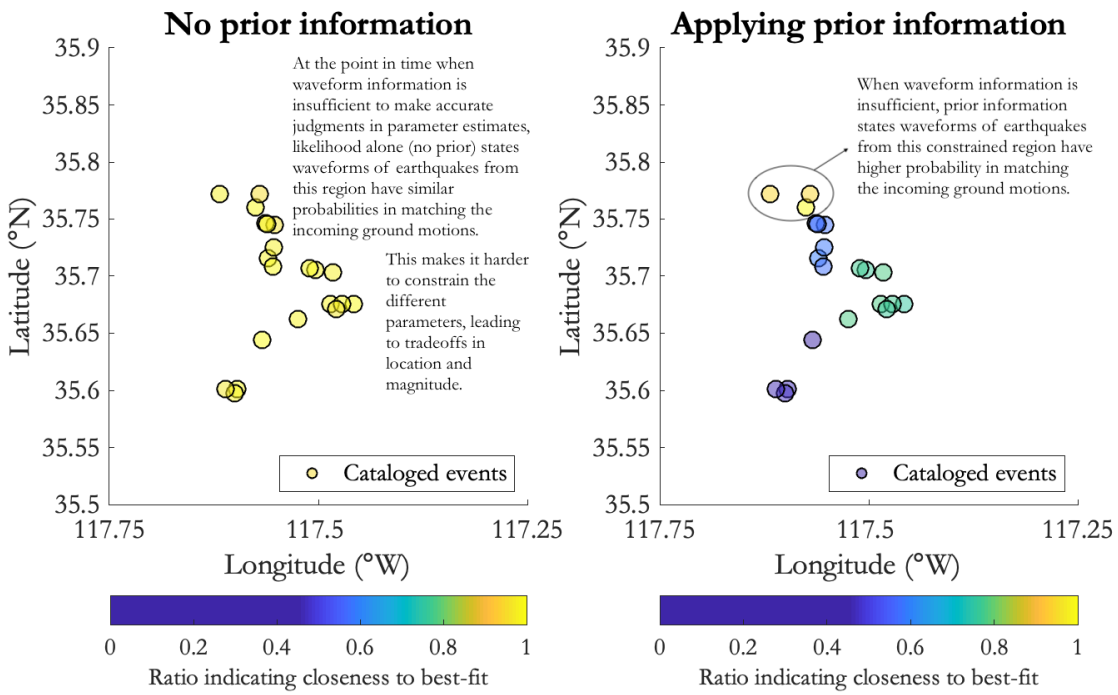


Figure 8.13. Valuable insight to the location estimate from prior information for the 2019 Ridgecrest mainshock.

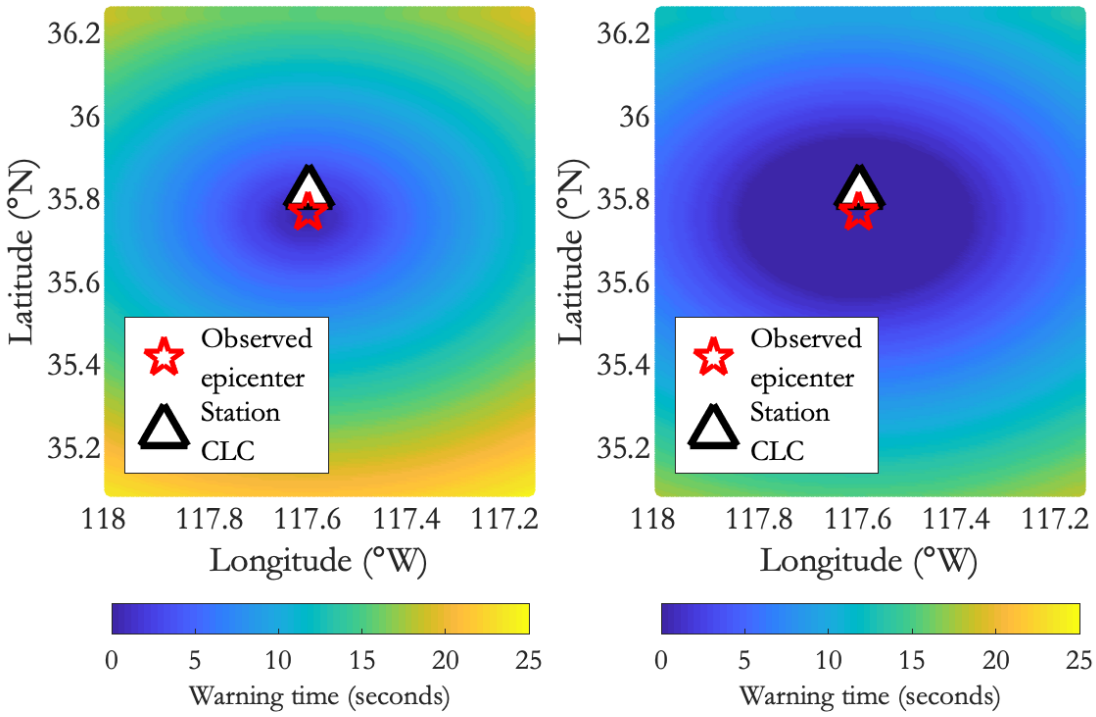


Figure 8.14. Warning times based on first parameter estimates at Station CLC, which occurs 2 seconds after origin time. (left) No prior information is used, and (right) prior information is applied.

8.6 Summary

Previous chapters find parameter estimates based on observed waveforms and a uniform prior. Applying additional prior information and constraints in a Bayesian probabilistic approach may reduce the tradeoffs between magnitude and location in the initial time points of the earthquake rupture, tradeoffs that occur due to insufficient amount of waveform data. However, as additional data is acquired with time, the impact of the prior information diminishes and waveforms have dominating influence over the final solutions. As seen with the offshore event, prior information is most valuable for regions of sparse station coverage and regions undergoing an earthquake sequence. This way, uncertainties can be reduced without waiting for more stations to be triggered and with data immediately available from an occurring sequence. Furthermore, the seismicity prior rapidly provides the arrival of a signal, indicating when to initiate the search algorithm.

9 Concluding Remarks and Future Work

9.1 Concluding Remarks

As stated in the introduction, the two-part search algorithm is developed to address the challenges the current EEW system faces. One of the challenges is identifying offshore events. The application of the algorithm to past real earthquakes shows the extended catalog search has the ability to find parameter estimates that approach the true observed parameters with the P-wave data at a single station. This is an advantage compared to the algorithms that wait for multiple triggered stations before issuing an alert. It is also advantageous for events with sparse station coverage, as only a single station is necessary. Another challenge is identifying complex earthquakes. The extended catalog search also is able to address this challenge by further extending the catalog to additional templates that consider multiple ruptures spaced closely together in time. Doing so, the P-wave data at a single station is generally sufficient to provide accurate parameter estimates. The unique feature of the extended catalog search is its consideration of the specific site and path effects observed at the single station and channel. Because these effects are factored into the goodness-of-fit, the extended catalog search overall finds stronger envelope fits than the grid search. The concluding remark in regards to the full two-part search algorithm is the extended catalog search solutions are accepted initially, and the grid search is used to confirm. However, if earthquake history is insufficient, such as an inadequate amount of epicenters, then the grid search solutions would be accepted over the extended catalog search.

The test sweep of the grid search method on a variety of waveform envelopes shows it is generally robust for $4.5<M<7$ events that are in network and surrounded by at least three stations in different directions. For the best chance of epicenters meeting this special criterion, a uniform distribution of stations is recommended with an interstation spacing between 10 to 20 km. Though the grid search struggles with tradeoffs in the initial time points, with time, it converges to the parameter estimates found by the extended catalog search. The use of prior information reduces the uncertainties in the initial estimates. Together, the two methods run in parallel, providing the best parameter estimates as quickly

as possible that are updated with additional data and with time. The test sweep results suggest to send an alert based on 2 different criteria:

1. P-wave data from a single station is used to alert regions near the epicenter, within 50 km, to expect strong shaking.
2. Wait for three triggered stations for enhanced accuracy in parameter estimates to alert regions farther away.

9.2 Future work

This thesis lays the groundwork for the two-part search algorithm. Future work would include a test sweep on a variety of datasets, including events and non-events, to find an acceptable threshold in uncertainties that would be allowed for issuing alerts. This thesis only refers to error bands of the best-fitting envelopes in reference to the incoming observed envelopes and assumes that those within a factor of 2 are acceptable fits. This thesis also attempts to mimic real-time analyses as closely as possible, calculating the probability using data that would be available at the specified time points (i.e. consider stations after their P-wave arrivals). However, it assumes the algorithm will have the stations' earthquake histories in the appropriate re-parameterized format at hand. Therefore, future work would include a real-time streaming of the earthquake database for building the extended catalog and the KD trees. It is suggested in Chapter 5 that KD trees be pre-determined and stored to save construction time. For KD tree nearest neighbor to be implemented practically for EEW purposes, it cannot be constructed in real-time. For EEW, it is practical to only search. Therefore, future work would include construction and storage of the KD trees for easy access, each one based on the length of the ground motion envelopes. Finally, this thesis covers event detection using priors and rapid estimation of the earthquake source. Future work would include the next step in finding the expected ground motion shaking from the earthquake source parameter estimates. It would be of great value to compare how different the expected ground motion shaking would be with respect to the uncertainties of the extended catalog search and grid search solutions. With the future work, the envelope-based two-part search algorithm presented in this thesis has high potential to improve the current EEW system.

Bibliography

Abercrombie, R. E. and Mori, J. (1996). Occurrence patterns of foreshocks to large earthquakes in the western United States. *Nature,* 381: 303-307.

Aki, K. (1967). Scaling Law of Seismic Spectrum. *Journal of Geophysical Research,* 72(4): 1217-1231.

Bentley, J. L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM,* 18(9): 509-517.

Bose, M., Heaton, T. H., and Hauksson, E. (2012). Real-time Finite Fault Rupture Detector (FinDer) for large earthquakes. *Geophysical Journal International,* 191: 803-812.

Bose, M. Heaton, T., and Hauksson, E. (2012). Rapid Estimation of Earthquake Source and Ground-Motion Parameters for Earthquake Early Warning Using Data from a Single Three-Component Broadband or Strong-Motion Sensor. *Bulletin of the Seismological Society of America,* 102(2): 738-750.

Bose, M. R., Allen, R. M., Brown, H., Cua, G., Fischer, M., Hauksson, E., Heaton, T., Hellweg, M., Liukis, M., Neuhauser, D., Maechling, P., Solanki, K., Vinci, M., Henson, I., Khainovski, O., Kuyuk, S., Carpio, M., Meier, M.-A., and Jordan, T. (2014). CISN ShakeAlert: An earthquake early warning demonstration system for California, in Early Warning for Geological Disasters, F. Wenzel and J. Zschau (Editors), Springer-Verlag, Heidelberg, Germany, 40-69, doi: 10.1007/978-3-642-12233-0_3.

Brown, R. A. (2015). Building a Balanced k-d Tree in O(kn log n) Time. *Journal of Computer Graphics Techniques,* 4(1): 50-68.

Brune, J. N. (1970). Tectonic Stress and the Spectra of Seismic Shear Waves from Earthquakes. *Journal of Geophysical Research,* 75(26): 4997-5009.

Chung, A. I., Henson, I., and Allen, R. M. (2019). Optimizing Earthquake Early Warning Performance: ElarmS-3. *Seismological Research Letters,* 90(2A): 727-743.

Chung, A. I., Meier, M. A., Andrews, J., Bose, M., Crowell, B. W., McGuire, J. J., and Smith, D. E. (2020). ShakeAlert Earthquake Early Warning System Performance during the 2019 Ridgecrest Earthquake Sequence. *Bulletin of the Seismological Society of America,* 110(4): 1904-1923.

Cua, G. (2005). *Creating the Virtual Seismologist: developments in ground motion characterization and seismic early warning.* PhD thesis, California Institute of Technology.

Cua, G., Fischer, M., Heaton, T., and Wiemer, S. (2009). Real-time Performance of the Virtual Seismologist Earthquake Early Warning Algorithm in Southern California. *Seismological Research Letters,* 80(5): 740-747.

Felzer, K. R. and Brodsky, E. E. (2006). Evidence for dynamic aftershock triggering from earthquake densities. *Nature,* 441.

Felzer, K. (2009). Simulated Aftershock Sequences for an M7.8 Earthquake on the Southern San Andreas Fault. *Seismological Research Letters,* 80(1): 21-25.

Gutenberg, R. and Richter, C. F. (1944). Frequency of earthquakes in California. *Bulletin of Seismological America,* 34: 185-188.

Hauksson, E., Stock, J., Hutton, K., Yang, W., Vidal-Villegas, A., and Kanamori, H. (2010). The 2010 Mw 7.2 El Mayor-Cucapah Earthquake Sequence, Baja California, Mexico and Southernmost California, USA: Active Seismotectonics along the Mexican Pacific Margin. *Pure and Applied Geophysics,* 168: 1255-1277.

Heaton, T. H. (1979). *Generalized ray models of strong ground motion.* PhD thesis, California Institute of Technology.

Heaton, T. H. (1985). A Model for a Seismic Computerized Alert Network. *Science,* 228(4702): 987-990.

Heaton, T. H. and Hartzell, S. H. (1989). Estimation of Strong Ground Motions from Hypothetical Earthquakes on the Cascadia Subduction Zone, Pacific Northwest. *Pageoph,* 129(1/2): 131-201.

Meier, M.-A., Heaton, T., and Clinton, J. (2015). The Gutenberg Algorithm: Evolutionary Bayesian Magnitude Estimates for Earthquake Early Warning with a Filter Bank. *Bulletin of the Seismological Society of America,* 105(5): 2774-2786.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.,* 50(2): 379-402.

Ogata, Y. (1988). Statistical models for earthquake occurrence and residual analysis for point processes. *Journal of the American Statistical Association,* 83(401): 9-27.

Pederson, J. S. (1997). Analysis of small-angle scattering data from colloids and polymer solutions: modeling and least-squares fitting. *Advances in Colloid and Interface Science,* 70: 171-210.

Reasenberg, P. A. and Jones, L. M. (Year). Earthquake Hazard After a Mainshock in California. *Science,* 243(4895): 1173-1176.

Ross, Z. E. and Ben-Zion, Y. (2014). Automatic picking of direct *P, S* seismic phases and fault zone head waves. *Geophysical Journal International,* 199: 369-381.

Ross, Z. E., Idini, B., Jia, Z., Stephenson, O. L., Zhong, M., Wang, X., Zhan, Z., Simons, M., Fielding, E. J., Yun, S., Hauksson, E., Moore, A. W., Liu, Z., and Jung, J. (2019). Hierarchical interlocked orthogonal faulting in the 2019 Ridgecrest earthquake sequence. *Science,* 366: 346-351.

Utsu, T. (1961). A statistical study on the occurrence of aftershocks. *Geophysics Magazine,* 30: 521-605.

Wu, Y. M., Kanamori, H., Allen, R. M., and Hauksson, E. (2007). Determination of earthquake early warning parameters, τ_C and P_d, for southern California. *Geophysical Journal International,* 170: 711-717.

Yagi, Y., Okuwaki, R., Enescu, B., Kasahara, A., Miyakawa, A., and Otsubo, M. (2016). Rupture process of the 2016 Kumamoto earthquake in relation to the thermal structure around Aso volcano. *Earth, Planets and Space,* 68(118): 1-6.

Yamada, M., Heaton, T., and Beck, J. (2007). Real-Time Estimation of Fault Rupture Extent Using Near-Source versus Far-Source Classification. *Bulletin of the Seismological Society of America,* 97(6): 1890-1910.

Yamada, M. and Heaton, T. (2008). Real-Time Estimation of Fault Rupture Extent Using Envelopes of Acceleration. *Bulletin of the Seismological Society of America,* 98(2): 607-619.

Yamada, M. (2007). *Early Warning for Earthquakes with Large Rupture Dimension*. PhD thesis, California Institute of Technology.