# Reducing Latencies in Earthquake Early Warning

Thesis by

Lucy Yin

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy


CALIFORNIA INSTITUTE OF
TECHNOLOGY
Pasadena, California

2018
Defended August 3, 2017

Lucy Yin
ORCID: 0000-0002-0652-9330

To my father and Wentao

Most importantly, to the memory of my mother

ACKNOWLEDGEMENTS

The memorable years at Caltech have forever changed my life in numerous ways. The pursuit of my PhD degree was only possible with the support and assistance from the Caltech faculty, scientists, friends, and family. There are too many people to thank.

First and foremost, I would like to acknowledge my advisor Professor Tom Heaton. I am very grateful for his help throughout the time at Caltech. Working with him has been a great privilege and also very enjoyable. His inspiring ideas always guide me to think outside of box and be a creative innovator. What I learned from him is far beyond the knowledge of engineering seismology. In addition to his academic guidance, he has supported me during the emotionally tough times and gave me moral support. He is not only my academic advisor at Caltech; he is also my advisor in life.

I want to express my sincere gratitude to my thesis advising committee, Prof Domniki Asimaki, Prof Jean-Paul Ampuero, Prof Yisong Yue, and Dr Morgan Page, for the time they dedicated to my thesis and research progress. I learned so much from them, from seismology to applied mathematics, to statistics, and to computer science. I would also like to thank Prof Jim Mori and Prof Masumi Yamada for their hospitality during my visit to Kyoto University. Not only did I enjoy participating in the English seismology seminars, but also I will never forget the memories of Gion Matsuri, hanabi, and mushi atsui weather. I am grateful to have worked with all the researchers and scientists from the EEW group, especially my collaborators Dr Men-Andrin Meier and Dr Jennifer Andrews. I thank them for their insights during our discussions and the effort and time put into our manuscripts.

I am fortunate to have joined Caltech MCE with a group of wonderful people. We worked together for classes, fought through qualification exams, and camped in the wildness. They are my big family away from home. Special thanks to my MCE mentors, Dr. Gokcan

Karakus, Dr. Stephen Wu, Dr Ming Hei Cheng, and Dr Ramses Mourhatch, who taught me the survival strategies through PhD.

My heartfelt gratitude goes to my parents, Dr. John Jiahong Yin and Mrs. Ellen Guirong Tang, who support me with their unconditional love. Ever since I can remember, my father always inspired me to pursue my dreams in the academic path. And for my mom, I hope that she could see my accomplishment from heaven, and keep watching over me as my guardian angel. Lastly, I would like to thank Dr. Wentao Huang for always being on my side during the journey at Caltech. I am forever grateful for his love, patience, understanding and encouragement helping me going through the ups and downs. I look forward to the next adventurous chapter of our lives together away from Caltech.

I would also like to acknowledge the financial support from the Natural Science and Engineering Research Council of Canada (NSERC) through their postgraduate doctoral and master's scholarships (PGS), and the Gordon and Betty Moore Foundation.

ABSTRACT

Existing Earthquake Early Warning (EEW) algorithms use waveform analysis for earthquake detections, estimation of source parameters (i.e., magnitude and hypocenter location), and prediction of peak ground motions at sites near the source. The latency of warning delivery due to data collection significantly restricts the usefulness of the system, especially for users in the vicinity of the earthquake source, as the warning may not arrive before the strong shaking. This presentation discusses several methods to reduce the warning latency, while maintaining reliability and robustness, so that the warning time can be maximized for users to take appropriate actions to reduce causalities and economic losses.

Firstly, we incorporated the seismicity forecast information from Epidemic-Type Aftershock Sequence (ETAS) model into EEW as prior information, under the Bayesian probabilistic inference framework. Similar to human's decision-making process, the Bayesian approach updates the probability of the estimations as more information becomes available. This allows us to reduce the required time for reliable earthquake signal detection from at least 3 seconds to 0.5 second. Furthermore, the initial error of hypocenter location estimation is reduced by 58%. The performance of the algorithm is further improved during aftershock sequences and swarm earthquakes.

Secondly, we introduce the use of multidimensional (KD tree) data structure to organize seismic database, so that the querying time can be reduced for the nearest neighbor search during earthquake source parameter estimation. The processing time of KD tree is approximately 15% of the processing time of linear exhaustive search, which allows the potential use of large seismic databases in real-time.

EEW is an interdisciplinary subject that involves collaboration among different scientific and engineering communities. Only by optimizing the warning time, such a unified system could be successful in taking protective actions before, during, and after earthquake natural disasters.

PUBLISHED CONTENT AND CONTRIBUTIONS

Yin, L., M. Meier, and T. Heaton (2017), “Making earthquake early warning faster and more accurate using ETAS seismicity models as a Bayesian prior”, 16th world Conference on Earthquake Engineering proceeding, N 2850

Yin, L., J. Andrews, and T. Heaton, “Rapid Earthquake Discrimination for Earthquake Early Warning: A Bayesian Probabilistic Approach using Three-Component Single Station Waveforms and Seismicity Forecast”, The Bulletin of the Seismological Society of America (under review)

Yin, L., J. Andrews, and T. Heaton, “Reducing process delays for real-time earthquake parameter estimation – an application of KD tree to large databases for Earthquake Early Warning ”, Computers & Geosciences Journal (under review)

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS AND/OR TABLES

*Chapter 1*

# Introduction

## 1.1 Motivation

An earthquake is a natural disaster that develops over a very short time frame; the time interval between the initial of rupture to the end of damaging ground motion arriving at a site could be from the order of seconds to a minute. However, the aftermath damage that an earthquake brings could be permanent and significant. Scientists, the government and the private sector have put in tons of effort in mitigating earthquake losses. Although earthquake prediction is a challenging task, the development of Earthquake Early Warning (EEW) systems has progressed rapidly over the past few decades (Allen, Gasparini, et al. 2009).

The advancement of Earthquake Early Warning systems has been driven by the growth of information technology and the increase of awareness of seismic hazard. The goal of Earthquake Early Warning is to provide alerts to the community about the incoming ground shaking and take appropriate actions to save lives and reduce losses. Strauss and Allen 2016 have estimated that EEW could decrease the number of injuries during an earthquake by more than 50%, and reduce millions of dollars in economic savings from fire damage, semiconductor plant danger, and train collisions, with statistically three lives rescued annually. The obvious benefits of the application have brought the attention of researchers worldwide to develop and implement EEW systems.

The success of an EEW system is often measured by the accuracy and time of the delivered alerts. Although the existing EEW algorithms can provide reliable and accurate information

in the final updates (unfortunately, sometimes come after the arrival of the strongest shaking), the uncertainties in the earliest alerts could be largely due to the lack of available ground motion data (Bose, Allen, et al. 2014). In principal, there is a trade-off between accuracy and time: as more data is collected from the observation of the on-going earthquake with the progress of time, the analysis can produce more accurate estimations. However, warnings would be delayed if significant time was necessary for data collection. The earliest alerts are the most critical outcomes of the system because the strongest shaking is generally experienced near the earthquake hypocenter where the propagated seismic waves arrive earliest. To overcome the challenge of latency and accurate predictions, in this thesis we propose several methodologies to maximize warning time (for earliest alerts) while guaranteeing a robust accuracy level of the messages.

The latency for warning times in EEW, $\Delta t_{latency}$, is defined as:

$$\Delta t_{latency} = \Delta t_{data} + \Delta t_{est} + \Delta t_{trans} \qquad [1.1]$$

where $\Delta t_{data}$ is the time necessary to collect sufficient ground motion stream data, $\Delta t_{est}$ is the time needed to estimate parameters about the earthquake (such as magnitude, location or predicted ground motion), and $\Delta t_{trans}$ is the time required to transmit the alert information to the community. Since $\Delta t_{trans}$ highly depends on the hardware device and the allowable bandwidth of information transmission, it is out of the research scope for seismologists. In fact, $\Delta t_{trans}$ can be minimized to the order of fraction of seconds because of the rapid advancement in electronic information flow. Relatively, $\Delta t_{data}$ and $\Delta t_{est}$ contribute to the majority of the latency concern. The time needed for data collection and estimation might require seismic wave arrival at multiple stations before issuing the first alert (Kuyuk and Allen 2014), and some could range from 3sec to over 10 sec depending on the algorithm and station distribution (Wu and Kanamori 2005) (Satriano, et al. 2011). In this study, I aim to shorten $\Delta t_{data}$ and $\Delta t_{est}$ through different techniques. To

minimize $\Delta t_{data}$, I propose to incorporate prior information data, so that additional data can be collected simultaneously. As data from various independent sources are obtained in parallel, required observations can be gathered with less time. Chapters 2 through 4 of the thesis focus on the formulation and collection of prior information from earthquake forecasting; and then apply the Bayesian probabilistic approach to combine the seismic knowledge from different sources under an ensemble model to provide final predictions. To reduce $\Delta t_{est}$, I present a data structure organization method, multidimensional binary search (KD Tree) to efficiently query desired estimations in order in Chapter 5.

From hardware deployment to algorithm development, from decision making to public education, EEW involves contributions among different scientific and engineering communities. As an earthquake engineer, my goal is to reduce the damage brought by disasters to the minimum through the efforts to develop intelligent algorithms. Only by improving the accuracy and speed of the future alert, can the system reach its full potential in performance.

## 1.2 Background on Earthquake Early Warning System (EEW)

### *1.2.1 EEW concept and development*

With the development of information technology, Earthquake Early Warning (EEW) systems are able to analyze ground motions in real-time and provide alerts before the onset of the destructive wave at specific facilities. An earthquake nucleates at a point under the surface of the earth, and excites many types of ground motion waves, including P-wave, S-wave, and surface wave. The P-wave is the fastest wave with less destructive power, while the S-wave and the surface wave are slower in traveling speed with substantially large destructive power. EEW is based on the principle that the damaging earthquake ground motions propagate slower than electronic information, so warnings can be successfully

delivered immediately after detecting the first earthquake signals at a seismic station. The speed of the more damaging S-waves from earthquakes is about 3.5km/s, whereas electrically transmitted signals from the seismic network sensors travel at about $3.0x10^5$km/s. As the seismic waves propagate, the seismometers observe more streams of the real-time ground motion data. As a result, the real-time analysis of the earthquake parameters becomes more precise. The characterized information on the event is then delivered to the users of EEW. The warning time, as defined by the duration between the EEW alert received by the user and the arrival of strong shaking at the user's site, needs to be sufficient to respond with appropriate actions. In general, the warning time increases as the latency time decreases. In addition to the scientific effort to reduce the warning latency, many inevitable geological factors that could dominate the latency include the distance between the site and the hypocenter, depth of the earthquake source, and soil properties, etc.

The application of EEW alerts can reduce hazard risks through public awareness, automated decisions, and emergency responses. Public awareness increases the safety of society, with the personal preparation of "drop, cover, and hold on" to avoid minor injuries resulting from falling objects, especially in schools and public areas with a vulnerable and dense population (Horiuchi 2009) (Fujinawa and Noda 2013). Automated decisions include interrupting hazardous nuclear or chemical processes in manufacturing systems; this prevents secondary or cascading unsafe failures to protect personnel (Wu, Beck and Heaton 2012) (Ionescu, et al. 2007) (Wu, Beck and Heaton 2013). Lastly, EEW alerts can provide alerts to the rescuing groups for faster emergency response, including hospitals, police, and fireman, etc. Simply by providing alerts in advance, various groups and government can better facilitate resources and assign rescuing responsibilities, particularly during aftershock sequences when telecommunication is unstable. Nevertheless, errors in the system (including false alert and missed events) could potentially lead to serious

consequences in the societal adoption of EEW. It is critical for scientists and engineers to collaborate in developing robust and reliable EEW systems that provide timely alerts with guaranteed accuracy.

### *1.2.2 Overview of earthquake early warning systems around the world*

Although the concept of earthquake early warning has been around for awhile (since (Cooper 1868)), the implementation of the systems has been achieved only over the past few decades with the development of necessary instruments, computational power, and network communications. EEW systems have been in operation in several regions around the world (Normile 2004). Figure 1.1 shows a seismic hazard map and countries where EEW is in operation or being tested (Allen, Gasparini, et al. 2009).

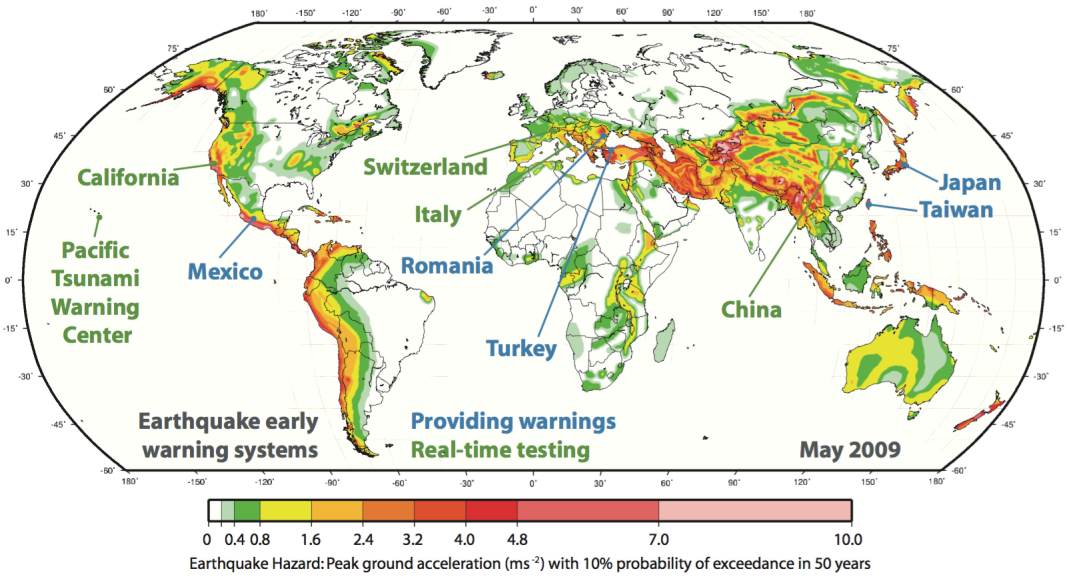



Figure 1.1 Seismic hazard map and countries where EEW is in operation or being tested under development by May 2009 (Allen, Gasparini, et al. 2009)

Japan has implemented one of the first applications of EEW. In the late 1960s, the Japan Railway (JR) has started monitoring ground shaking by deploying seismometers near their Bullet Train, also named as Shikansen, train tracks. The power of the Bullet Train is automatically shut off if the ground shaking intensity achieves a threshold of 40 gals. (Nakamura and Tucker 1988). The system was upgraded to the Urgent Earthquake Detection and Alarm System (UrEDAS) in the 1980s. Additional seismometers are deployed along the costal lines to provide more warning time for the trains (Nakamura 1984). The system worked well during the Niigata Chuestu earthquake in 2004, applying brakes to the Bullet Train within 3 sec after the detection of p-wave arrival (Nakamura, et al. 2006). Currently, the Japan Meteorological Agency is able to broadcast national wide public warnings of on-going earthquakes by television, cellphone, and other means of telecommunications (Doi 2003).

The Seismic Alert System (SAS) for Mexico City was established after the 1985 Michoaca earthquake (J. Espinosa-Aranda, et al. 1995). The SAS was the first public warning system in the world. Seismometers are deployed along the coastal line, located about 300km southwest of Mexico City, to detect subduction earthquake. The system is effective because the subduction zone is a few hundreds of kilometers away from Mexico City, so warnings can be provided to the city about 60 sec prior to the arrival of the damaging seismic waves (Lee and Espinosa-Aranda 2003) (J. Espinosa-Aranda, et al. 1996). Due to the high population density and soft soil properties of Mexico City, the SAS system provides very useful information for the departments in charge of emergency services.

Taiwan has an earthquake early warning system created by the Taiwan Central Weather Burea (CWB) (Wu, Shin and Tsai 1998). The system takes the waveform information from real-time seismometers, and determines EEW parameters, including the predominant period ($\tau_c$) and peak amplitude displacement in the initial 3 sec of the P-wave ($P_d$), to

estimate the earthquake magnitude (Wu, et al. 2006). Studies have shown that the system could provide 20 s of warning time to Taipei if the 1999 Chi-chi earthquake reoccurs (Wu and Kanamori 2005).

The California Integrated Seismic Network (CISN) research group has developed the CISN ShakeAlert System in California. The system combines estimations and uncertainties from three independent algorithms, $\tau_c$-$P_d$ Onsite Algorithm, Virtual Seismologist, and Elarms, and a Decision Module calculates the probability of earthquake source parameters: magnitude and hypocenter location. A user display delivers the warning messages by display the shaking information on a map in real-time. The current system is under development in the beta-testing phase in California, and plans to deliver to the public in the near future. With the collaborations of universities, government agencies, and private sectors, the group is also planning to expand the demonstration system to the entire west coast of the United States. Figure 1.2 shows the frame of ShakeAlert system.

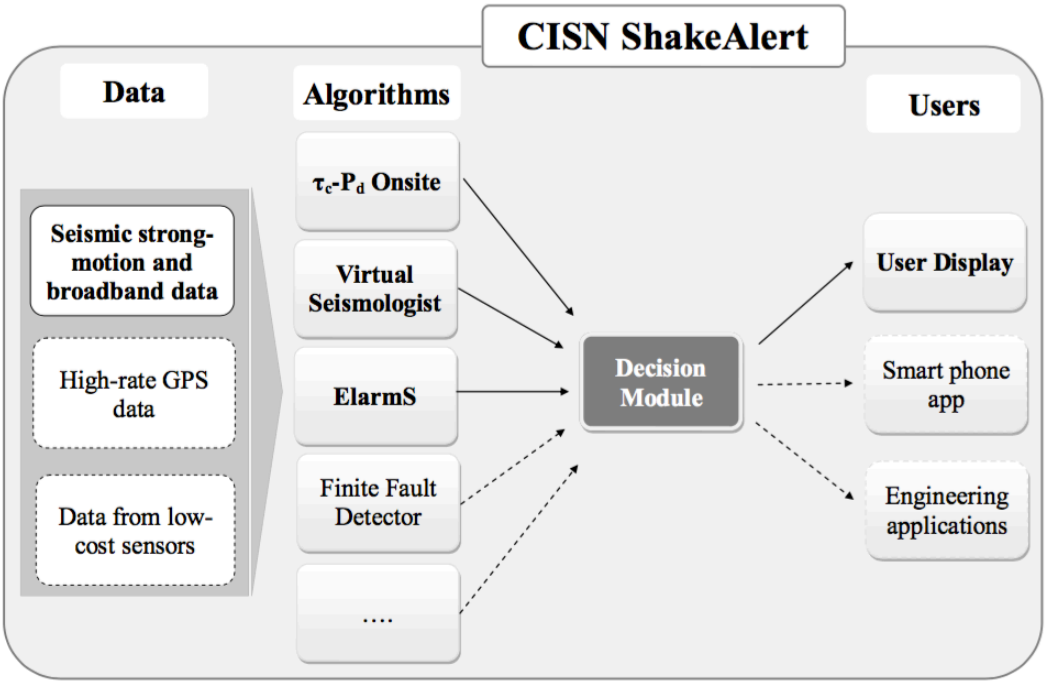



Figure 1.2 Framework of CISN ShakeAlert. The bold modules with solid lines are the existing system that is currently running. The dashed lines show components under development. (Bose, Allen, et al. 2014)

EEW has been proven to be beneficial in both reducing causalities and minimizing economic losses through many major earthquake events, especially during the M9 Japan Tohoku earthquake in March 2011. An increase of interest in earthquake early warning has been expressed around the world, particularly in seismic active regions. Southern Italy, Istanbul, China, Chile, Bucharest, and many other countries are trying to develop earthquake early warning systems to mitigate seismic damages.

### *1.2.3 Bayes' Theorem for EEW*

Bayesian Probabilistic approach to Earthquake Early Warning was first introduced in the Virtual Seismologist method (Cua 2005). The application of Bayes' theorem the probability of source characterization at any given time of the on-going earthquake, is a combination results from prior information and contributions from the available ground motion observations. The main differentiation of the Bayesian approach from other EEW algorithms is the exploitation of knowledge from previous experience or judgments that are not generally incorporated in automated decision-making process. A flow chart of Bayesian approach for EEW is shown in Figure 1.3. This methodology mimics the human ability to process many types of information simultaneously, combining the analyzed results to make a final decision at the end, and updating the decision over time as additional information is collected.

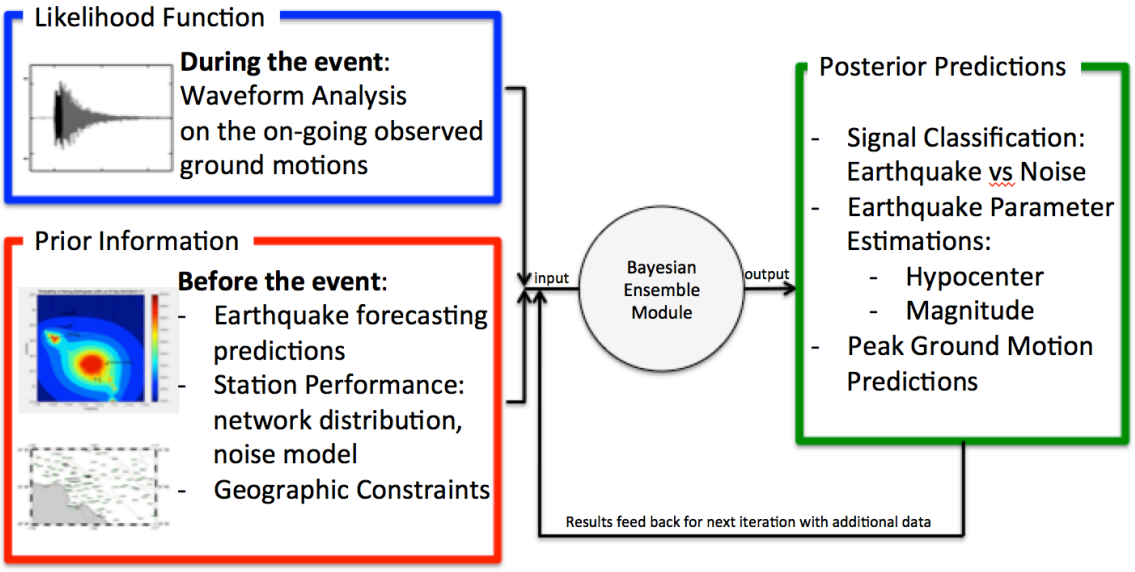



Figure 1.3 Flow Chart of Bayesian framework in EEW

The inspiration of such a decision making process in earthquake early warning comes from human beings' judgments on weather. For example, if a person needs to decide on bringing an umbrella outside, one would look at the weather forecast information and check if there are water drops outside the window. During this process, the person's brain is performing a Bayesian calculation: the weather forecast information provides a prior information of how probable the region is going to rain in general, and the actual observation water drop is a likelihood collection of the rain probability. The goal of my thesis is to apply a similar approach of this elegant concept to earthquake early warning, so information from multiple heterogeneous sources can be processed in parallel to make fast and reliable decisions.

The prior information in earthquake early warning systems includes recent seismicity rates, distribution of seismometer network, health status of the stations, and regional seismic hazard risk, etc. For example, earthquakes tend to be active near geologic faults, so for long-term predictions, the probability of earthquake occurrence is higher near the recognized faults. Also, earthquakes tend to cluster in space and time, forming a foreshock-mainshock-aftershock sequence, so recent seismic activities are good indications of near future seismicity. In Chapter 2 of this thesis, I present a detailed formulation of an earthquake forecasting model.

The general Bayes' theorem for EEW can be expressed as a product of the prior probability density function, $P(A)$, multiplied by the likelihood probability density function, $P(B|A)$, and normalized by the evidence function, $P(B)$, shown as the follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad [1.2]$$

$$\propto P(B|A)P(A)$$

where $A$ is the parameter we are interested in estimate (in EEW, this includes signal discrimination of earthquake source vs. noise source, regression estimation of source parameters such as magnitude and hypocenter location, prediction of peak ground motion intensities at users' sites, etc.), and $B$ is the given incoming observations (in EEW, this includes on-going ground motion observation data, such as time-series data of waveforms, GPS displacement, etc.). $P(A|B)$ is the posterior probability density function, meaning the probability of A given the observation data B. The evidence function $P(B)$ is the probability of the observation data that is independent of the estimating parameter $A$. This term is normalization constant, which does not affect the estimation of the parameter $A$, so the posterior function is proportional to the product of the likelihood function and the prior function.

Depending on the prediction task, the Bayesian framework can be applied to different estimating parameter and the predictive term $A$ is replaced with the parameter of interest. In the earthquake detection problem, Eq [1.2] becomes:

$$P(Y = Eq|S(t)) \propto P(S(t)|Y = Eq)P(Y = Eq) \quad [1.3]$$

where $Y = Eq$ is estimating the signal source being an earthquake event, and $S(t)$ is the avalaible ground motion observation at time $t$.

Similarly, in estimating the hypocenter location, the Eq [1.2] becomes:

$$P(Lat, Lon|S(t)) \propto P(S(t)|Lat, Lon)P(Lat, Lon) \quad [1.4]$$

where $(Lat, Lon)$ is estimating the coordinate location of the earthquake source, and $S(t)$ is the avalaible ground motion observation at time $t$.

Although the framework for different tasks is similar, the detailed constructions for the predictive functions vary dramatically. In Chapter 3 and 4 of this thesis, I present detailed approaches in answering both of the questions.

## 1.3 Research goal and Thesis plan

In order to construct an early warning system with faster initial alerts while maintaining the accuracy of the predictive information, we introduce methodologies from the seismology domain knowledge and computer science techniques to incorporate additional useful predictive information and efficiently organize large seismic database, respectively. The objectives of this thesis include:

- Reduce latency time for first early warning alert
- Improve accuracy of the earthquake parameter estimations
- Minimize process delays of large databases

This thesis is outlined as follows: Chapter 1 gives an overview of the research and developments of earthquake early warning systems. Chapter 2 discusses earthquake forecast models that could be incorporated into earthquake early warning. Chapters 3 through 4 provide applications of earthquake forecasting information to improve on the performance of earthquake early warning predictions; Chapter 3 focuses on the rapid earthquake detection algorithm, and Chapter 4 focuses on improvements of the hypocenter location estimation. Chapter 5 presents a method to reduce process delays of big data search for real-time seismology. Chapter 6 concludes the thesis with final remarks and suggestions of future work.

*Chapter 2*

# Earthquake Forecasting Methods

## 2.1 Background on Foreshock-Mainshock-Aftershock Sequences

Many seismologists have observed the temporal and spatial clustering properties of earthquakes (Kagan and Jackson 1991) (M. Bouchon, et al. 2013) (Gerstenberger, et al. 2005). This clustering pattern is often referred as foreshock – mainshock – aftershock sequences. The term "aftershock" is often defined as a series of smaller earthquakes following a "mainshock", which is an earthquake with larger magnitude. "Foreshock" is the smaller earthquakes that occurred prior to the mainshock in time. Of course, sometimes the predefined mainshock could produce aftershocks for years and the aftershocks produced may be larger than the mainshock (Lomnitz 1966), while other times, not all premonitory events are observed prior to large earthquakes (Abercrombie and Mori 1996). In addition, another type of seismic activity with location clustering pattern is the swarm earthquake (Shearer 2012), which occurs repeatitively over time at the same location.

Over the years, scientists are still trying to find explanations for the occurrences of earthquake sequences. While some researchers argue that triggering of near field earthquakes is due to the sudden change of dynamic and static stresses (Gomberg, et al. 2001), others believe that aftershock sequences are driven by physical mechanisms such as fluid flow, magnetic, or creep events (Hainzl and Ogata 2005) (Lohman and McGuire 2007). There are also theories that explain that foreshock occurrences are due to the interplate or heterogeneity of the Earth's crust (M. Bouchon, et al. 2013) (Mogi 1963).

Although the science behind earthquake formulation is complex and controversial, most scientists agree on the clustering properties of earthquake occurrences. No matter what the true explanation is behind the science of earthquake sequences, one conclusion is indisputable: the recent seismicity is a good indication of seismic activities of the near future.

## 2.2 Earthquake Forecasting and Earthquake Early Warning

As proposed in the Bayesian approach to earthquake early warning system, prior information can be incorporated to provide faster and more accurate warnings. Earthquake early warning, earthquake forecasting, and seismic hazard maps all provide a forecast of future earthquake occurrences, evaluated for different time frames. Figure 2.1 shows the relative time frame for the three earthquake information products.

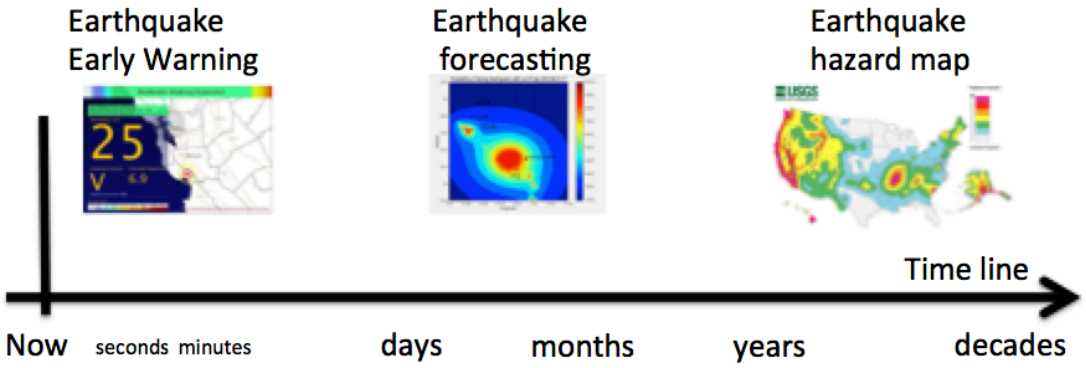



Figure 2.1 Time frame of various earthquake information products

Earthquake early warning, the focus of this thesis, provides earthquake information of the next few seconds to minutes. Even though the “heads-up” time is short, the intention is to

make automated decisions and take immediate action to avoid losses from the disaster, as shown in Chapter 1. The conventional concept of EEW is to send out warnings after detecting an initial seismic wave, so the analysis sorely depends on the observation of the on-going earthquake and not previously observed seismicity.

Earthquake forecasting tends to predict regional seismicity activities in the near future based on the recent seismicity. In this model, the recent change in seismicity is the major influence of model predictions. Scientists often use the forecasting models to predict aftershock patterns of a particular seismic sequence. However, they can be applied for any region or time of interest in general. This model is often created for the prediction range of the next few hours to months.

Lastly, seismic hazard maps are intended to provide insight to the general public and guidance in development. The input of this model is based on the long-term historical seismic occurrence that has lasted for years. The information provided from hazard maps is essential in creating and updating seismic designs provisions of building codes and facilitate government on urban planning. In general, the seismic hazard maps forecast the regional hazard level for the next few years to decades.

Up to now, the three earthquake information products provide independent information and were created separately for different audiences. However, it is not difficult to make the connections between them: the long-term predictions (forecasting and hazard maps) can be useful inputs for the short-term predictions. As mentioned in Chapter 1, the forecasting information can be applied as the prior information under the Bayesian framework, and the waveform analysis serves as the likelihood function. For the conventional waveform analysis of earthquake early warning, a minimum of time-series data is required to be collected before any decisions are made (e.g. 3 sec for Onsite, (Bose, Hauksson, et al., A Trigger Criterion for Improved Real-Time Performance of Onsite Earthquake Early

Warning in Southern California 2009)), and this process is repeated for every earthquake event. However, in the cases when we are expecting high seismicity, such as during aftershock sequences or swarm earthquakes, it is unnecessary to redundantly wait until the end of the data collection process to send out the alert because the new trigger is probably due to another aftershock earthquake in the sequence.

In such cases, the alerts can arrive much faster to the users near the source to mitigate potential dangers from the disaster. Table 2.1 shows the decision-making scenarios under Bayesian inference, where immediate decisions can be made when consistent predictions from waveform analysis and seismic forecast are observed. The earthquake forecasting models can provide the expected seismicity information necessary in the early warning system. Of course, the large earthquakes do not always occur when the expected seismicity is high; waiting is still required to collect additional data in these cases.

| | High earthquake probability from waveform analysis | Low earthquake probability from waveform analysis |
|---|---|---|
| High earthquake probability from seismic forecast | Send alert immediately | Wait for additional waveform analysis |
| Low earthquake probability from seismic forecast | Wait for additional waveform analysis | No alert immediately |

Table 2.1 EEW decision-making scenarios under Bayesian framework

Since EEW system aims to provide information to all earthquakes causing ground motions that could be dangerous, alerts should be issued faster for all earthquakes during the entire sequence including aftershocks, and not only emphasize the system performance during a large magnitude mainshock. During aftershocks, the repetitive ground shaking continuously deteriorates already weakened infrastructure components. Additional natural disasters, such as landslides and tsunami, can also be triggered from aftershocks as a consequence. The seismic damage can be even more significant if the aftershocks occur close to a populated urban area. The benefits of a rapid and reliable EEW system during the aftershocks of a large earthquake are equally (or more, in some cases) important than the mainshock, as rescue and repair personals are continuously working in then already damaged and fragile epicentral region (Bakun, et al. 1994). For example, over 200 aftershocks occurred after the single mainshock during the Northridge earthquake sequence. There is also a chance that what seemed like a recent mainshock turns out to be foreshock activity of another large event (Reasenberg and Jones, Earthquake Hazard After a Mainshock in California 1989), like the 1992 M6.5 Big Bear Earthquake occurring three hours after the M7.3 Landers Earthquake. If the prior information can assist in sending out faster alerts for all the aftershock events, then system performance would be improved for over 99% of all events.

## 2.3 General Epidemic-Type Aftershock Sequence (ETAS) model

Epidemic-Type Aftershock Sequence (ETAS) model simulates the entire sequences based on statistical relationships of earthquakes (Vere-Hones 1966) (Y. Ogata 1988) (Kagan and Knopoff 1981). The aftershocks are generated based on well-established empirical stochastic models derived from seismicity observations. Most importantly, in addition to the direct aftershocks produced by a mainshock, the generated aftershocks could produce aftershocks of its own, forming an epidemic-type effect, which differentiates this approach

from other aftershock simulation methods. These secondary aftershocks are true observations in the real earthquake sequences (Felzer, Becker, et al. 2002). This statistical method quantitatively describes the clustering property in earthquake sequence processes and the generated earthquakes that have the probability of generating secondary earthquakes. The construction of the model suggests that the distribution of aftershocks follows the Omori's Law in time (Utsu 1961), Gutenberg-Richter relationship in magnitude (Gutenberg and Richter 1944) and mainshock-aftershock distance relationships.

Taking the ETAS simulation created by Felzer (K. Felzer, Stochastic ETAS aftershock simulator 2007) as an example, the magnitude and location of the aftershocks are sampled from the distributions, and the primary aftershocks are fed back into the model to produce the secondary aftershocks; this process repeats. The generated aftershocks that match the time period and region of interested are selected to create a report of aftershock catalog. Every run of the simulation will produce different results due to the randomness of the sampling procedure. The maximum likelihood estimation (MLE) of aftershock locations can be calculated by running the simulation hundreds or thousands of times and taking the average results of all the simulations.

The simulation results of the Northridge aftershock for a 24-hour period of January 18 -19, 1994 calculated by Felzer's ETAS model, are as follows. Note here that each simulation result produces different sequences, show in Figure 2.2. There is always a small probability that a simulated aftershock is large enough that it initiates an unexpected sequence, so the seismicity clusters are slightly different in every simulation.

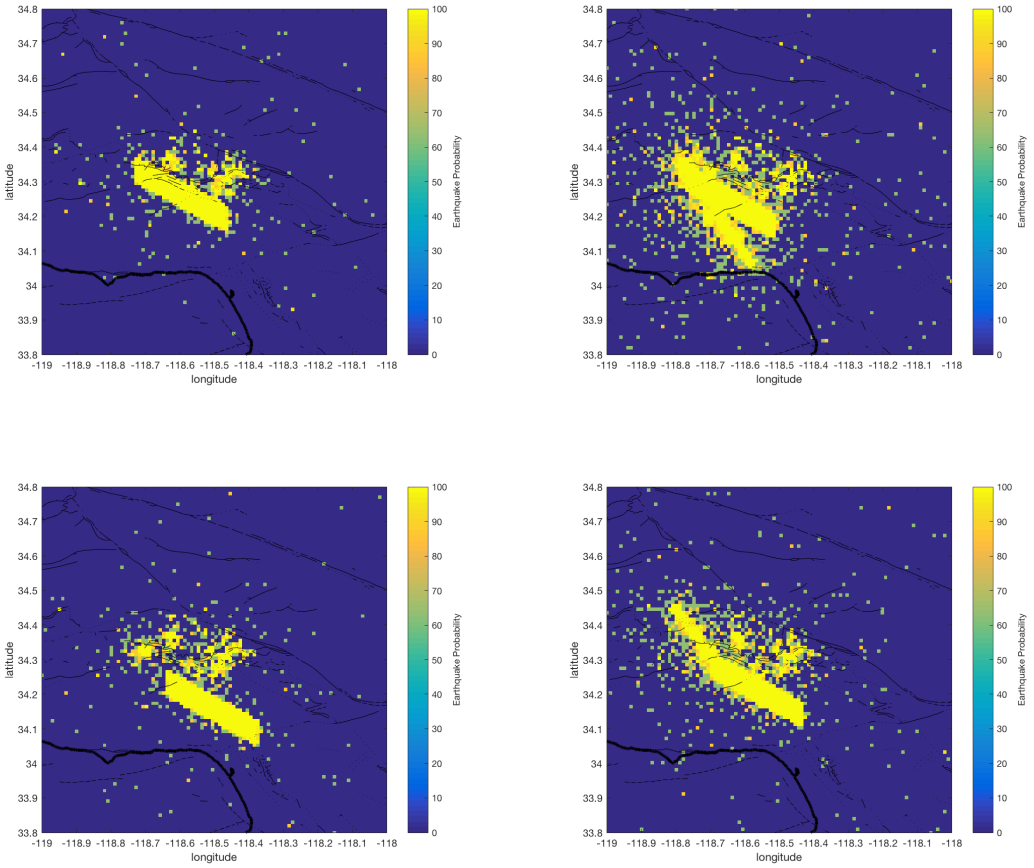



Figure 2.2 Four simulation results of the Northridge aftershock for a 24-hour period of January 18 -19, 1994 calculated by Felzer ETAS model

Figure 2.3 and Figure 2.4 are the average results of the Felzer ETAS model simulation after 50 and 500 runs, respectively. Although the average of 500 runs has more smoothing boundaries showing transition of change in seismic rates due to the averaging effect, the chance of producing outliers is much higher with more runs. Note here in the 500-run scenario that the diagonal lines show that the aftershocks might trigger additional seismicity on the fault lines that propagated outwards.

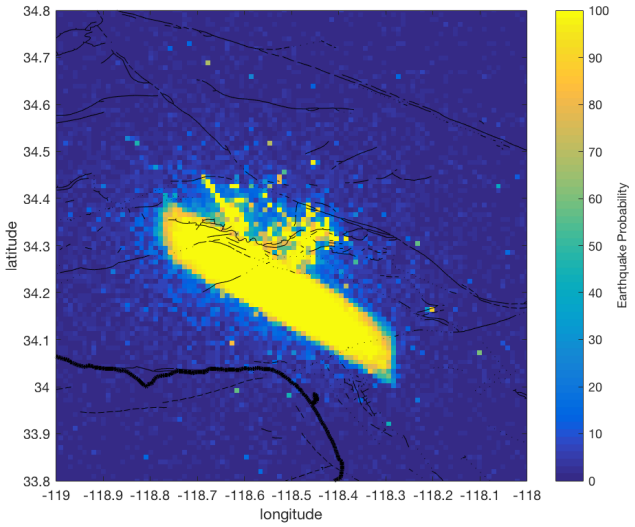



Figure 2.3 Average result of the Felzer ETAS model simulation after 50 runs

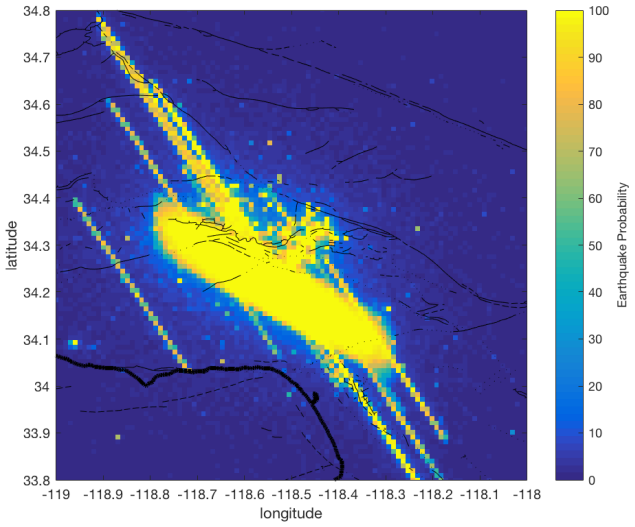



Figure 2.4 Average result of the Felzer ETAS model simulation after 500 runs

We can run the simulation repetitively and then get an average result for the earthquake-forecasting map. However, the computational delay introduced is not tolerable for real-time seismicity application of Earthquake Early Warning system. For example, a 50-run takes about 1 min on Matlab platform; the 500-run takes about 5 minutes. Started with the fundamental concept of ETAS simulation by (K. Felzer, Stochastic ETAS aftershock simulator 2007), I created an ETAS forecasting model that produces a MLE of earthquake forecasting map with a single run of negligible computational delay time about a single second. The real-time ETAS model can be incorporated into EEW upon the instantaneous requirements.

## 2.4 Modified Epidemic-Type Aftershock Sequence (ETAS) model

The forecast earthquake probability calculated using an ETAS seismicity model is based on the premise that the location of future earthquakes is significantly influenced by the accumulation of previously observed earthquakes. The concept of the ETAS seismicity model has been well established in the earthquake-forecasting field and the forecasting results have been validated through many earthquake sequences (Y. Ogata 1998) (K. Felzer 2009). The future earthquake occurrence process is modeled as a nonhomogeneous Poisson process in time; the probability of one or more earthquakes occurring above $M_{min}$ at location $(lat, lon)$ within the time range $\Delta t$ is:

$$prob_{ETAS}(lat, lon) = 1 - exp\left(\int_{t}^{t+\Delta t} \lambda(t, lat, lon)dt\right) \qquad [2.1]$$

where $\lambda(t, lat, lon)$ is the forecast rate of earthquake at current time $t$ and location (*lat, lon)*. It is composed of the long-term background seismicity $\mu(lat, lon)$ and the short-term observed seismicity.

$$\lambda(t, lat, lon) = \mu(lat, lon) + \sum_j \lambda_j(t, lat, lon) \quad [2.2]$$

I model earthquake sequences following Omori's Law in time (Utsu 1961), Gutenberg-Richter's relationship in magnitude (Gutenberg and Richter 1944) and Felzer and Brodsky's relation (Felzer and Brodsky 2006) in space. The short-term seismicity rate caused by each of the historical earthquakes in the catalogue is first calculated as a function of a distance from the hypocenter source, $\lambda_j(t, r)$, and then mapped to latitude and longitude, $\lambda_j(t, lat, lon)$, using a numerical transformation based on the distance-to-location mapping on the earth surface. The formulation for the seismicity rate by $j$th earthquake at the current time $t$ and distance $r$ km is:

$$\lambda_j(t, r) = \frac{K_o 10^{\alpha(M_j - M_{min})}}{(t - t_j + c)^p r^n} \quad [2.3]$$

where $K_o = 0.008, \alpha = 1, c = 0.095, p = 1.34, n = 1.37$ are ETAS model parameters of California obtained from (K. Felzer 2009) and $lat_j, lon_j, M_j$ are source parameters of the $j$th earthquake from the observed seismicity catalog. $M_{min}$ is the minimum magnitude of the forecast earthquakes. In the application of this proposed method to EEW, I assume that the EEW system has the access to the seismicity catalog record that continuously updates with time. As time passes, all the newly occurred events should automatically concluded in the catalog for the forecasting of future events.

In order to validate the accuracy of the ETAS predictions, Figure 2.5 to Figure 2.11 are examples of the earthquake probability forecasting maps produced from the modified

ETAS model and the true observation of seismicity for 1) Chino Hills earthquake sequence on 29 July 2008, 2) Northridge earthquake sequence on 17 April 1994, 3) Cucapah El Mayor Sequence on 9 April 2010, and 4) a seismic dormant region during 13 May 2015, respectively. The forecasting results not only match the location estimations of various seismic activation sequences, but also predict well during the seismic quiescence period. The size of the red circle scale, with observed magnitude of the earthquake records.

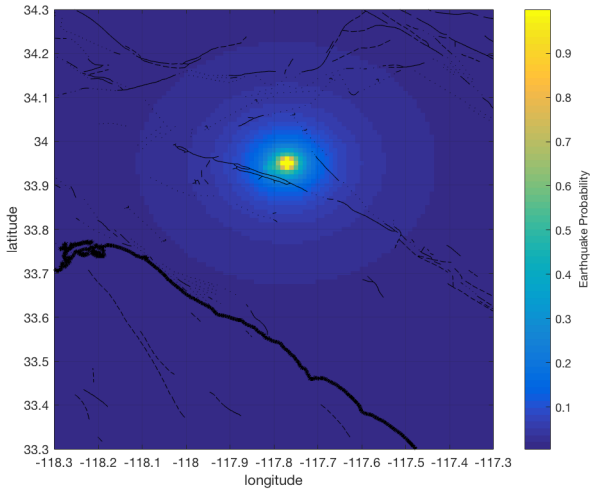



Figure 2.5 Modified ETAS forecast map for Chino Hills earthquake sequence on 29 July 2008

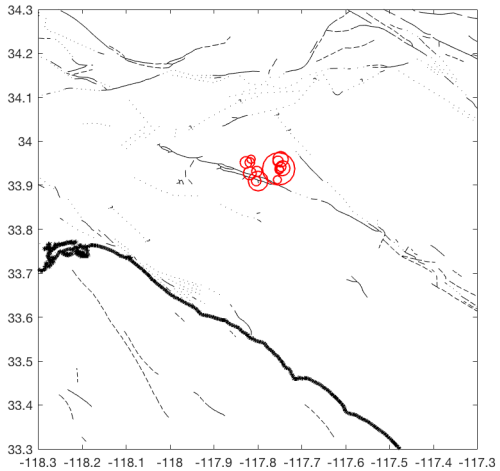



Figure 2.6 Observed seismicity of Chino Hills earthquake sequence on 29 July 2008. The size of the red circle scale with observed magnitude of the earthquake records.

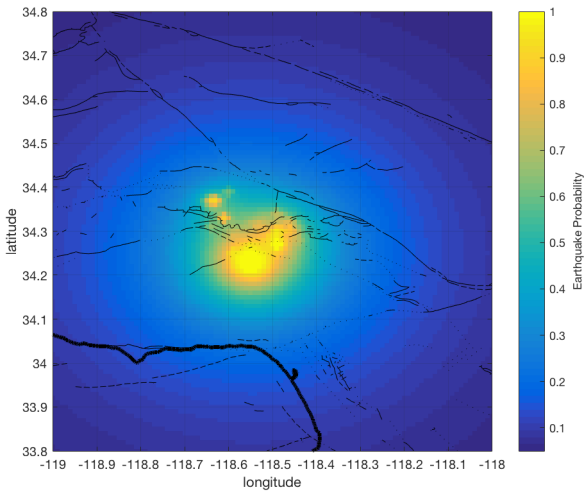



Figure 2.7 Modified ETAS forecast map for Northridge earthquake sequence on 17 April 1994

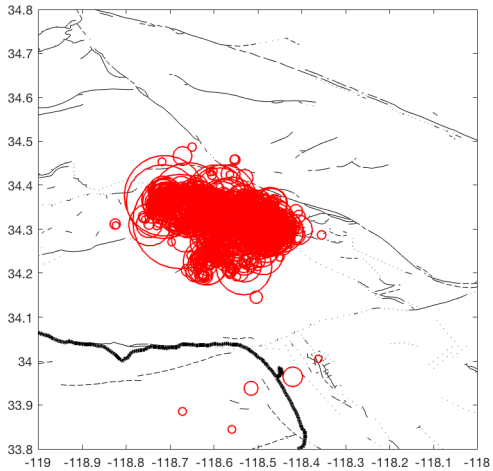



Figure 2.8 Observed seismicity of Northridge earthquake sequence on 17 April 1994. The size of the red circle scale with observed magnitude of the earthquake records.

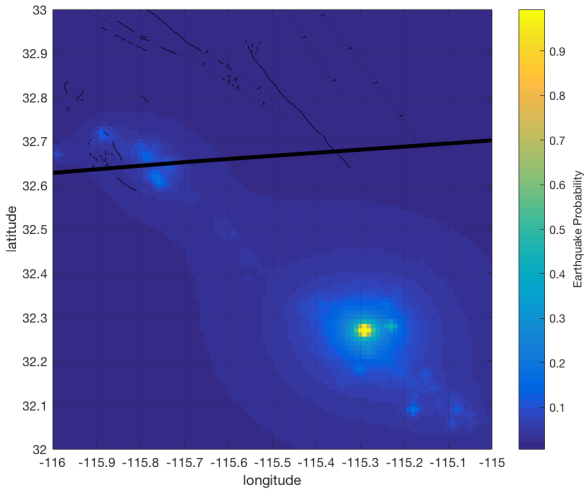



Figure 2.9 Modified ETAS forecast map for Cucapah El Mayor Sequence on 9 April 2010

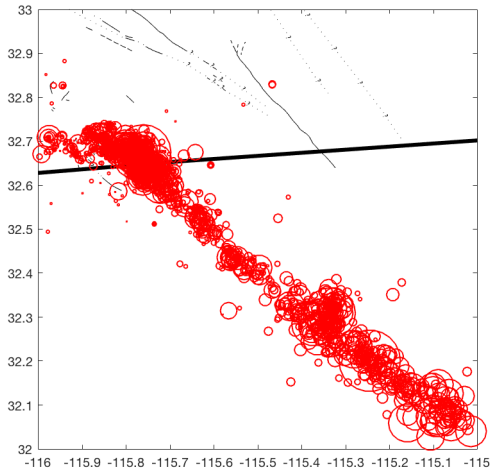



Figure 2.10 Observed seismicity of Cucapah El Mayor Sequence on 9 April 2010. The size of the red circle scale with observed magnitude of the earthquake records

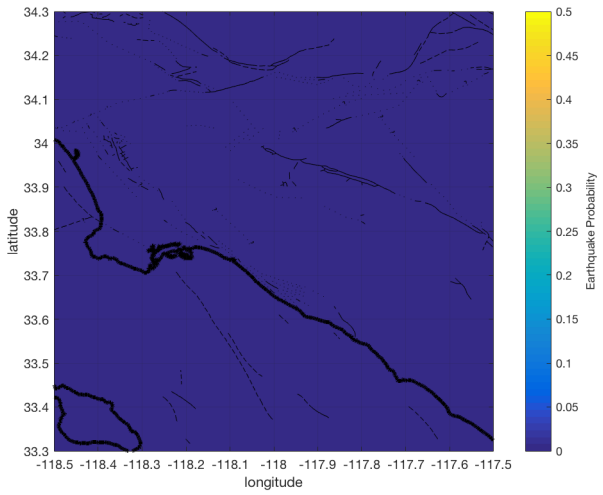



Figure 2.11 Modified ETAS forecast map for a seismic dormant region during 13 May 2015

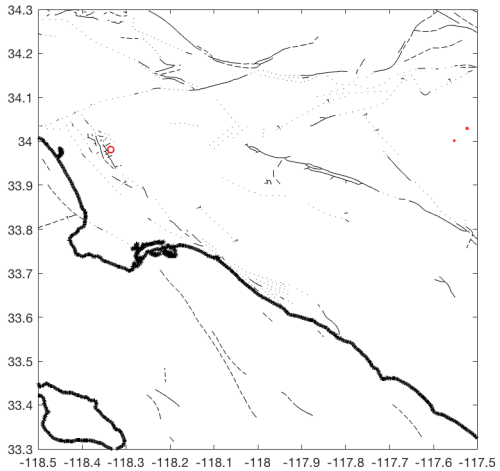



Figure 2.12 Observed seismicity of a seismic dormant region during 13 May 2015. The size of the red circle scale with observed magnitude of the earthquake records.

## 2.5 Summary

This chapter discusses the earthquake clustering properties in time and space, forming foreshock-mainshock-aftershock sequences. Furthermore, I presented methods to forecast near future earthquakes, especially the Epidemic-type Aftershock Sequence (ETAS) model. ETAS model predicts future seismicity based on statistical models of aftershock relationships. The seismic forecasting information brings significant insights to early warning systems. Despite the size of the earthquakes, most of the seismicity activities we observe are aftershocks of a sequence. Therefore, the previously observed earthquakes are good indications of near future events. I implemented an ETAS forecast model that provides real-time solution while maintaining the accuracy.

In the following two chapters I will present methodologies to apply the information provided by the modified real-time ETAS model into Bayesian approach to EEW to improve on the accuracy and speed of the earliest alerts.

*Chapter 3*

# 3. ETAS Prior application one: Rapid Earthquake Discrimination

## 3.1 Introduction

Due to the rapid advancement of digital seismic networks, Earthquake Early Warning (EEW) systems are currently able to analyze the real-time ground motion information and have the potential to provide warnings to potential users before strong shaking begins (Heaton 1985) (Allen and Kanamori 2003). We desire these EEW systems to provide both reliable and fast alerts, however, the goals of accuracy and speed are often in conflict with each other. Since the arrival of the destructive S-wave follows closely after the arrival of the P-wave in the epicentral region, processing delays must be minimized if we hope to provide warnings of the potentially damaging S-waves near an earthquake's epicenter.

A popular strategy for EEW is to identify the P-wave at a station and then warn of an impending S-wave. Unfortunately, systems reliant on these short windows of data are also commonly triggered by teleseisms and non-earthquake sources. The incorrect identification of the earthquake signals in EEW may cause false alarms or large uncertainties in source parameters. The negative impacts of the 'cry-wolf' syndrome can be critical in the societal adoption of EEW (Kuyuk et al., 2015), so speed may be sacrificed for improved accuracy in current systems. The first task to perform promptly after observing a shaking at a seismic station is to automatically make a decision on whether or not the shaking is caused

by an earthquake source, and different criteria have been imposed to filter out the non-earthquake triggers: the single-station Onsite algorithm collects and analyzes a fixed window of 3s before declaring an event (Bose, Hauksson, et al., A Trigger Criterion for Improved Real-Time Performance of Onsite Earthquake Early Warning in Southern California 2009); network-based algorithms require a minimum number of triggered stations for warning confirmation (e.g. Elarms-2 requires 4 stations for California (Kuyuk and Allen 2014), and Presto requires 3-5 stations for Southern Italy (Satriano, et al. 2011)). These methods can introduce a significant delay, especially in regions with low station density.

In this chapter, three predictive models are presented to identify earthquake source signals. First, a waveform analysis model uses a logistic regression method to predict the probability of incoming signals being generated by earthquake or non-earthquake sources. Then, two Bayesian models are presented that employ earthquake forecasting results (from Chapter 2) in addition to the waveform analysis model. One model uses the peak ETAS probability of the region as the Bayesian prior, and the other uses a derived earthquake probability from the ETAS model and noise distribution as the Bayesian prior.

## 3.2 Method and Data

We firstly collected local earthquake and non-earthquake strong-motion waveform data to train the model parameters in the waveform analysis. We also utilized earthquake catalog information for the ETAS forecasting analysis. The proposed model is then validated through different methods to demonstrate its reliability and robustness: 1) the performance of the proposed model is evaluated at every 0.5 sec since the triggered time up to 3.0 sec to estimate the speed-accuracy trade-off; 2) the leave-one-out cross validation test is

performed to demonstrate the robustness of the model in future predictions; 3) the proposed method is compared with the existing $\tau_c$-$P_d$ method to assess speed and accuracy gains; and 4) we demonstrate the application of the method in several test cases: an earthquake mainshock, an aftershock, an ambient noise false trigger, and a teleseismic event.

### *3.2.1 Data*

We collected three component strong-motion waveforms from local crustal earthquake and non-earthquake records in the southern California region to train the prediction model to identify earthquake signals. The non-earthquake records include ambient noise signals and teleseismic events that were detected by STA-LTA-type triggering at single seismic stations. All the strong-motion traces, 2,481 three-component records in total, are downloaded from the Southern California Earthquake Data Center. The station trigger times are provided by the Onsite algorithm (Kanamori 2005) (Y. Wu, et al. 2007) and are calculated using the modified characteristic function developed by R. Allen, 1978.

An important goal of EEW is to identify earthquakes that cause a significant level of ground shaking. Ground motion intensity depends on many factors including magnitude, hypocenter distance, local site conditions, details of source radiation, and wave propagation. We consider only records with observed Peak Ground Acceleration (PGA) greater than 2cm/s$^2$ (equivalent to Modified Mercalli Intensities > II) in the seismic network of Southern California during 2010 to 2015 (Wald, et al. 1999). With this threshold, our database consists of a total of 1,128 earthquake records. Ground motions with PGA less than 2cm/s$^2$ are not felt by humans and are unlikely to damage buildings (Cheng, et al. 2014). Figure 3.1 shows the distribution of the MMI shaking intensities of the earthquake records in our database. Mid- and large- size earthquakes contribute to a significant fraction of the records, since larger magnitude events cause MMI II shaking to greater distances. The data set includes records from the M7.2 El Mayor-Cucapah (4 April

2010), the M5.4 La Habra (28 March 2014) and the M5.4 Borrego Springs (7 July 2010) earthquakes. The majority of the records in the study created weak to light shaking. Although these records are minor concerns for the purpose of large earthquakes or human sensitivity, it is necessary to include them for a complete description of the statistical population of observations of an EEW system, since low PGA values are more often recorded due to the natural distribution of earthquake occurrence and ground motion attenuation with distance. A better identification of the low PGA earthquake records improves the overall performance of the earthquake detection.

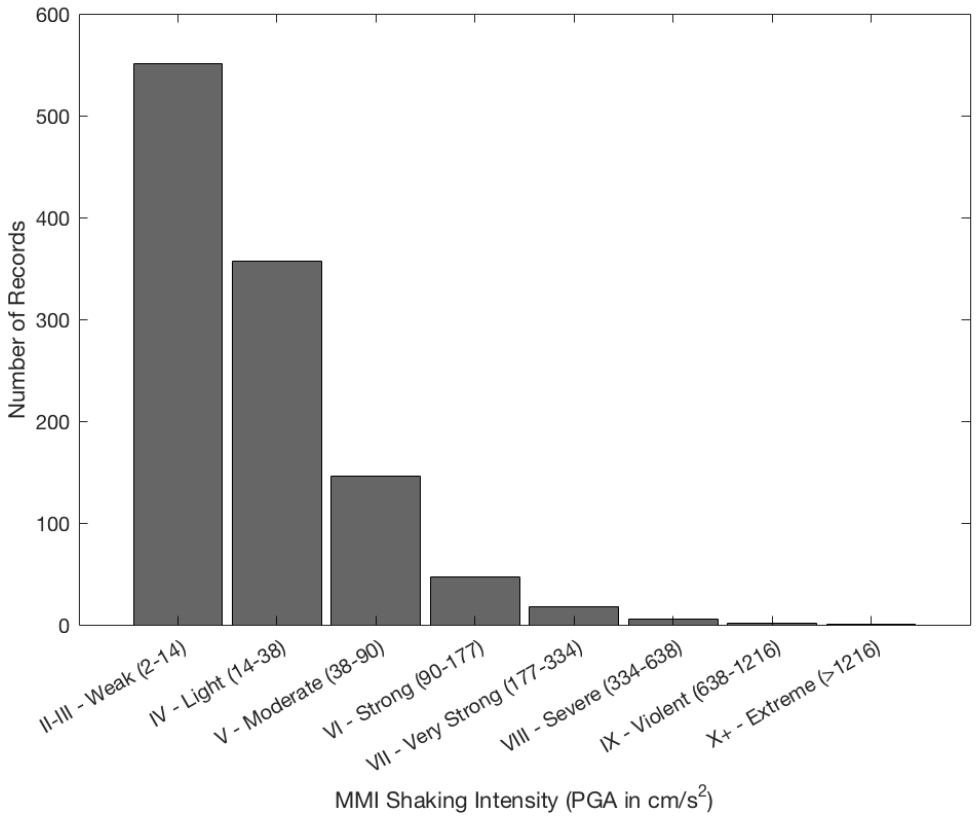



Figure 3.1 MMI shaking intensity distributions of the 1,128 earthquake records collected for the study

The data set of non-earthquake records consists of, 1000 noise and 353 teleseismic records. The noise signals include calibration pulses, jumps in electric current, glitches induced by machinery, ambient noise, etc. Since the total number of false triggers is on the scale of millions per year, the noise records were uniformly sampled from the top 100 noisiest stations in the CI network during 2015 (as observed by the Onsite algorithm STA-LTA triggering) to capture the general characteristics of noise disturbances most likely to mistakenly trigger the seismic network. The teleseism data set comprises records from 14 teleseismic events that triggered the Southern California seismic stations between 2008 and 2015. Table 3.1 shows the list of teleseismic events obtained in this study.

| Time | Region | Latitude | Longitude | Depth (km) | Magnitude |
|---|---|---|---|---|---|
| 2008-02-21 | Nevada,USA | 41.15 | -114.87 | 6.7 | 6.0 |
| 2010-02-27 | Offshore Bio-Bio, Chile | -35.9 | -72.73 | 35 | 8.8 |
| 2010-08-18 | Mariana Islands | 12.2 | 141.51 | 10 | 6.3 |
| 2011-03-11 | Tohoku, Japan | 38.30 | 142.37 | 30 | 9.0 |
| 2012-04-12 | Gulf of California | 28.79 | -113.14 | 10.3 | 6.9 |
| 2012-08-14 | Sea of Okhotsk | 49.78 | 145.13 | 625.9 | 7.7 |
| 2012-12-14 | Offshore Baja California | 31.09 | -119.66 | 13 | 6.3 |
| 2013-02-06 | Solomon Islands | -10.80 | 165.114 | 24 | 8.0 |
| 2013-05-24 | Sea of Okhotsk | 54.89 | 153.22 | 598.1 | 8.3 |
| 2014-03-05 | Vanuate | -14.42 | 169.54 | 648 | 6.3 |
| 2014-04-01 | Iquique, Chile | -19.6 | 70.77 | 25 | 8.2 |
| 2014-06-23 | Raoul Island, New Zealand | -29.98 | 177.73 | 20 | 6.9 |
| 2014-06-29 | New Mexico, USA | 32.582 | -109.17 | 6.4 | 5.3 |
| 2014-06-29 | Samoa Islands | -14.9 | -175.4 | 10 | 6.5 |
| 2015-05-30 | Chichi-shima, | 27.84 | 140.5 | 664 | 7.8 |

| | | | | | |
|---|---|---|---|---|---|
| | Japan | | | | |

Table 3.1 The teleseism events obtained in this study. Strong-motion sensors in Southern California record all these events.

### *3.2.2 Data Processing and Feature Extraction*

For each baseline-corrected record, the acceleration and velocity in the vertical and horizontal directions are processed. The acceleration records are directly obtained after removal of the trend and bias of the raw data; the velocity records are obtained by integrating the acceleration data in the time domain, and then applying a fourth-order causal Butterworth high pass filter with a corner frequency of 0.075Hz. This filter is applied recursively in the time domain, so the processed time is negligible. The horizontal records are calculated using the square root of the sum of the squares of the two horizontal components.

We extract the peak values of each ground motion in every half-second window from 0.5s to 3.0 sec after the triggered time for the training of model parameters. Figure 3.2 shows the distribution of the extracted ground motion amplitude features for noise, teleseismic, and earthquake data. We took the logarithm of the model features because the ground motion amplitudes span several orders of magnitude (Bose, Heaton, and Hauksson 2012). The distributions show clear differences between the earthquake and non-earthquake (noise and teleseismic) groups, although there are overlaps between the group distributions. The amplitudes of the high-frequency motions decay faster with distance (Hanks and McGuire 1981), so acceleration and velocity quantities are intuitively selected as indications of local earthquakes. Displacement records are excluded in the feature selection because the double integration required to obtain the displacement record from the acceleration data recorded from the strong motion sensors can lead to waveform artifacts (significant long-period trends are amplified during multiple integrations, DC shifts are obscured, etc.). Various

sophisticated Bayesian model selection methods can be also applied to extract the useful features; this is beyond the scope of this study.

These features of the *i*th record at the *k*th half-second time window after the triggered time are combined into a vector $X_{i,k} = [1, log_{10}(Za_{i,k}), log_{10}(Ha_{i,k}), log_{10}(Zv_{i,k}), log_{10}(Hv_{i,k})]$, where $H$ and $Z$ denote horizontal and vertical component, and $A$ and $V$ denote acceleration and velocity, respectively. We also label *i*th record $Yi = 1$ or $Yi = -1$ for earthquake and non-earthquake records, respectively. Note both noise and teleseismic records are considered as non-earthquake records.

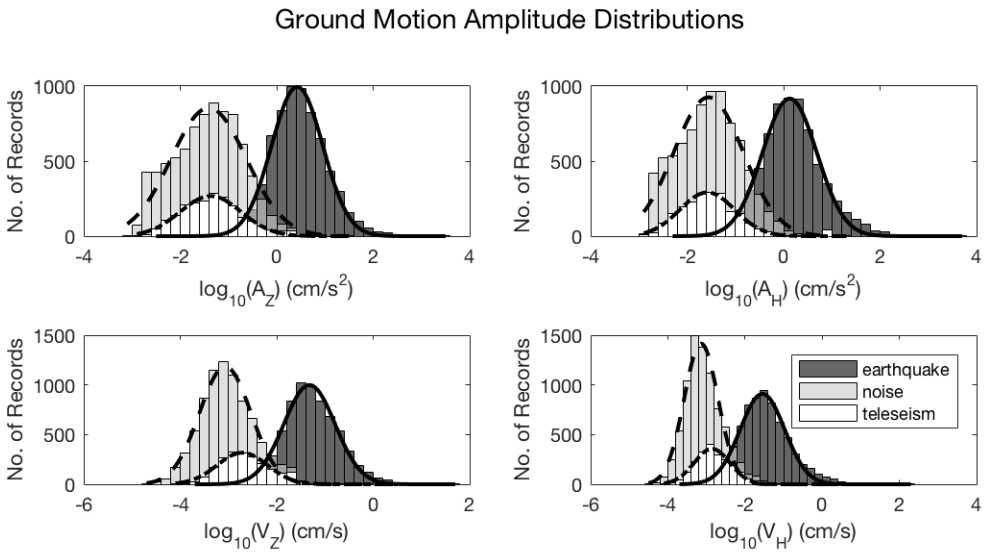



Figure 3.2 Maximum ground motion amplitude distributions collected for every half-second window within the initial 3.0s after the trigger time of all 2,481 three-component records used for this study. The labeled earthquake data are earthquake records with PGA greater than 2cm/s$^2$; noise data are false triggers

including calibration pulses, jumps in electric current, glitches induced by machinery, and ambient noise; the teleseism data include 353 records from 14 teleseismic events. The lines are the fitted Gaussian distributions to earthquake (solid), noise (dash) and teleseism (dot dash) data. The notations are A=acceleration, V=velocity, Z=vertical, H=horizontal.

## 3.3 Waveform Analysis

In waveform analysis, the goal is to predict the probability of the observed signal being caused by an earthquake source given only the available waveform information, $prob(Y_i|X_{i,t_1:t_n})$. We defined the classification result for station $i$ as $Y_i = 1$ as an earthquake record and $Y_i = -1$ as a non-earthquake record. $X_{i,t_1:t_n}$ is the waveform input of station $i$ recorded during time $t_1$to $t_n$. In general, $t_1$is the p-wave arrival time at the station; this is when the model starts to record the ground motion data for the predictions.

By assuming that the observed data $X_{i,t_1:t_n}$follows an independent and identically distributed random variable, the Bayesian equation can be written as:

$$prob(Y_i = 1|X_{i,t_1:t_n}) \propto \prod_{k=1}^{n} prob(Y_i = 1|X_{i,t_k}) \qquad [3.1]$$

A standard approach in binary classification is to define the predictive probability applying the logistic sigmoid function $\phi(t) = 1/(1 + e^{-t})$ to a linear function $t = f(x)$ (Yamada, Heaton and Beck 2007). The sigmoid function is a real-valued, differentiable, non-negative, and monotonically increasing function. Since the sigmoid function transforms linear inputs to a nonlinear output that is bounded between 0 and 1, it can be mathematically interpreted as probability. The predictive probability as a function of the observed ground-motion amplitudes is constructed using the sigmoid function:

$$prob(Y_i = 1|X_{i,t_k}) = \phi(X_i) = \frac{1}{1 + e^{-f(X_{i,t_k})}} \quad [3.2]$$

where

$$f(X_{i,t_k}) = c_0 x_{i0,t_k} + c_1 x_{i1,t_k} + \cdots + c_m x_{im,t_k} = \sum_{j=0}^{m} \theta_j x_{ij,t_k} = \theta \cdot X_{i,t_k}^T \quad [3.3]$$

$x_{ij,t_k}$ is the $j$th measurement of log of the ground motion during the $k$th half-second time window after triggered time at the $i$th station, $m$ is the total number of measurements, and $\theta_j$ is the $j$th model parameter. Let $X_{i,t_k} = [x_{10,t_k}, x_{i1,t_k}, x_{i2,t_k}, \dots, x_{im,t_k}]$, and $\theta = [c_0, c_1, \dots, c_m]$. The model parameters are determined from the training data set described earlier. In our study, we focus on four measurements of ground motion: vertical acceleration, horizontal acceleration, vertical velocity, and horizontal velocity. The best combination of features is chosen for $X_{i,t_k}$are based on the performance of model selection, details in the following section. According to this convention, as $f(X_{i,t_k})$ deviates further from 0 in the positive direction; the signal is more likely to be cause by an EEW-relevant earthquake source. The predicted probability of Eq[3.2] approaches one indicates that the event is very likely to be caused by an earthquake source; it also implies that the probability of detecting a non-earthquake source approaches zero, and vice versa in the opposite direction as $f(X_i)$ deviates from 0 to the negative direction.

Figure 3.3 shows an example of the chosen input features in the vertical acceleration at every half-second (red circle), where predictions are delivered at every half-second interval during the first 3 seconds. The predictions are updated based on the newly arrived waveform information

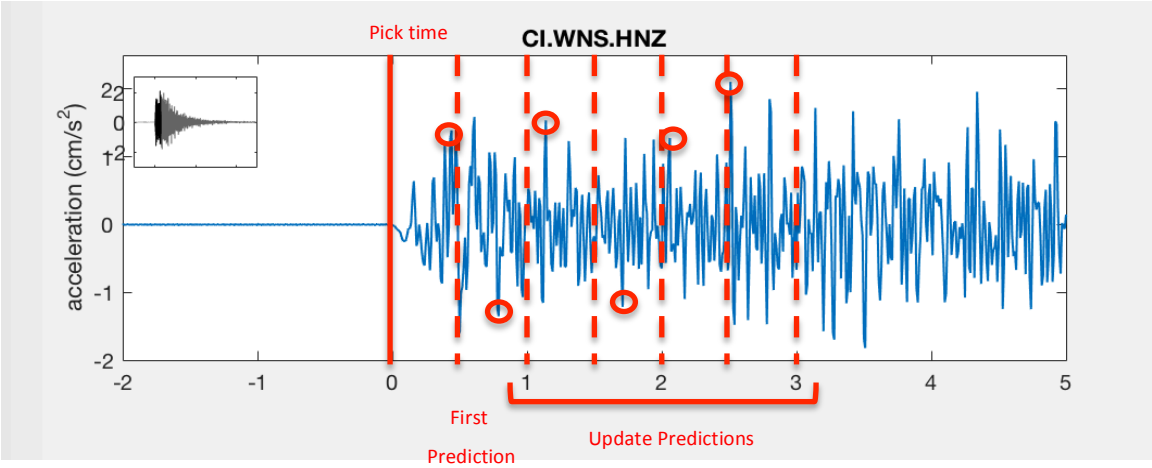



Figure 3.3 First and updated predictions at the initial 3 sec of the p-wave arrival

### *3.3.1 Determination of the Model Parameters*

Although the framework of the model is determined, the appropriate model parameters $\theta$ in the predictive formula from Eq[3.3] need to be specified to be useful to make predictions. To focus attention on the parameters of the likelihood function, we apply the Maximum Likelihood Estimation (MLE) method to determine the coefficients of the logistic regression that classifies earthquake and non-earthquake data. Classification methods such as Fisher's linear discriminant analysis (LDA) and the Least Squares estimates are alternative approaches to obtain the model coefficients. However, unlike the MLE method, these classification models do not provide a probabilistic interpretation to its predictive classes, which makes it challenging to measure the degree of uncertainties.

The MLE method can be interpreted as searching for an estimation of $\theta$ that best fit of the training data we collected. Assuming that all $D_{ik}$ are sampled independently and identically

from the distribution, the optimal model parameters $\hat{\theta}$ conditioned on the data $D_{mn} = \{(X_{i,t_k}, Y_{i,t_k}): i = 1 \dots m, k = 1 \dots n\}$ can be expressed as:

$$\hat{\theta} = argmax\ prob(D_{mn}|\theta) = argmax \prod_{i=1}^{m} \prod_{k=1}^{n} prob(D_{ik}|\theta) \quad [3.4]$$

where

$$prob(D_{ik}|\theta) = \frac{1}{1 + e^{-Y_{if}(X_{i,t_k}|\theta)}} \quad [3.5]$$

where $m = 2481$ is the total number of waveform records in the training, including earthquake, noise, and teleseism; $n = 6$ is the number of half-second windows in the initial 3.0 sec after triggered time.

### *3.3.2 Model Selection*

Applying the MLE method, we determined the model parameter coefficients for all 15 combinations of the four ground motion features using the training dataset. Table 3.2 demonstrates the model parameters and performance of all the candidate models. We focus on two performance measures for the model selection given the following definitions:

- True Positives (TP): true predicted earthquake data
- True Negatives (TN): true predicted non-earthquake data
- False Positives (FP): false predicted earthquake data, also referred to as false alerts
- False Negatives (FN): false predicted non-earthquake data, also referred to as missed events

First, we emphasize the initial precision rate of the predictions, defined as:

$$Precision\ (\%) = \frac{TP}{TP + FP} \quad [3.6]$$

at 0.5 sec after the trigger. A higher precision rate indicates a lower false alerts rate. This avoids modifications or cancelations of events that could potentially confuse the system and users. Secondly, we evaluate the final accuracy rate, defined as:

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \quad [3.7]$$

at the end of 3.0 sec. This measure is the representation of the final and overall performance of the predictions after all the prediction updates. As indicated in Table 3.2, model 1 satisfies both of the requirements, which demonstrates constancy in the highest accuracy and precision in both initial and final predictions.

| Model | Model Parameters | | | | | Initial Prediction - 0.5 s after TT | | Final Prediction – 3.0 s after TT | |
|---|---|---|---|---|---|---|---|---|---|
| | $c_0$ | $\log_{10}$(Za) | $\log_{10}$(Ha) | $\log_{10}$(Zv) | $\log_{10}$(Hv) | Accuracy (%) | Precision (%) | Accuracy (%) | Precision (%) |
| 1 | 6.884 | 4.8665 | -2.2965 | 0.2497 | 2.5895 | 89.24 | 90.88 | 97.70 | 96.85 |
| 2 | 8.0876 | - | 1.8542 | 3.3586 | -0.086 | 87.26 | 90.28 | 97.38 | 96.34 |
| 3 | 6.7442 | 2.7436 | - | 1.6677 | 0.9254 | 88.27 | 89.74 | 97.46 | 96.34 |
| 4 | 10.183 | - | - | 3.1113 | 1.645 | 88.19 | 91.30 | 96.37 | 94.51 |
| 5 | 6.7889 | 5.0952 | -2.4736 | - | 2.7924 | 89.24 | 90.73 | 97.70 | 96.85 |
| 6 | 7.1637 | - | 1.8846 | - | 2.6002 | 80.94 | 86.43 | 96.01 | 93.94 |
| 7 | 5.4392 | 3.516 | - | - | 1.7861 | 86.38 | 87.76 | 97.26 | 95.85 |
| 8 | 9.3083 | - | - | - | 4.1202 | 82.22 | 89.08 | 95.41 | 93.33 |
| 9 | 5.9532 | 3.1927 | -0.2574 | 2.2888 | - | 88.03 | 88.58 | 97.30 | 96.33 |
| 10 | 8.1628 | - | 1.8177 | 3.3123 | - | 87.22 | 90.19 | 97.42 | 96.42 |
| 11 | 6.0643 | 2.9228 | - | 2.3139 | - | 88.07 | 88.73 | 97.26 | 96.17 |
| 12 | 9.1132 | - | - | 4.4027 | - | 87.71 | 88.57 | 95.49 | 93.27 |
| 13 | 1.7111 | 5.3014 | -0.3945 | - | - | 85.33 | 84.85 | 96.49 | 95.38 |
| 14 | 2.3296 | - | 3.7818 | - | - | 76.99 | 79.85 | 93.23 | 91.59 |
| 15 | 1.8063 | 4.9229 | - | - | - | 86.17 | 85.12 | 96.41 | 95.06 |

Table 3.2 Coefficient parameters calculated using the MLE method, as well as accuracy and precision measures for all candidate models.

The model chosen for Eq [3.2] is:

$$prob(Y_i = 1|X_{i,t_k}) = \phi(X_i) = \frac{1}{1 + e^{-f(X_{i,t_k})}} \quad [3.8]$$

where

$$f(X_{i,t_k}) = 6.884 + 4.8665 * \log10(\mathrm{Za}) - 2.2965 * \log10(\mathrm{Ha}) + 0.2497 * \log10(\mathrm{Zv}) + 2.5895 \log10(\mathrm{Hv}) \quad [3.9]$$

### *3.3.4 Model Performance*

Through the model selection process, we chose model 1, by the combining of all 4 features, based on the performance measures. In order to demonstrate the time-accuracy of the model we performed, we evaluate the likelihood and posterior predictions at every time increments (0.5s window collected ended at 0.5s, 1.0s, 1.5s, 2.0s, 2.5s, 3.0s after the pick time at the station) on the entire dataset.

| Available Data | Predicted class | True Classes | | Precision | Accuracy |
|---|---|---|---|---|---|
| | | Earthquake | Non-Earthquake | | |
| 0.5s | Earthquake | 957 | 96 | 90.9% | 89.2% |
| | Non-Earthquake | 171 | 1257 | | |
| 1.0s | Earthquake | 1035 | 61 | 95.9% | 93.7% |
| | Non-Earthquake | 93 | 1292 | | |
| 1.5s | Earthquake | 1070 | 46 | 95.9% | 95.8% |
| | Non-Earthquake | 58 | 1307 | | |
| 2.0s | Earthquake | 1094 | 41 | 96.4% | 96.9% |
| | Non-Earthquake | 34 | 1312 | | |
| 2.5s | Earthquake | 1105 | 40 | 96.5% | 97.4% |
| | Non-Earthquake | 23 | 1313 | | |
| 3.0s | Earthquake | 1107 | 36 | 96.8% | 97.7% |
| | Non-Earthquake | 21 | 1317 | | |

Table 3.3 Waveform Analysis mode performance at time increments: 0.5s, 1.0s, 1.5s, 2.0s, 2.5s, 3.0s after the triggered time at the station

Table 3.3 shows the confusion matrix for the classification of earthquake versus non-earthquake records based waveform analysis. The decision boundary is set at 50%, and infers if the data is classified as an earthquake event if the predictive probability reaches above 50%; otherwise it is classified as a non-earthquake event. A summary of the results will be presented at a later section for comparison.

## 3.4 Bayesian Approach

(Cua 2005) and (Cua and Heaton, The Virtual Seismologist (VS) method: A Bayesian approach to earthquake early warning 2007) proposed that EEW could be made faster and more reliable by employing prior information in a Bayesian framework to estimate likely data interpretations. They suggested that seismicity information could be involved. In this paper, we show how this can be accomplished in the existing system. We propose a Bayesian probabilistic approach to rapidly identify earthquake source signals as quickly as 0.5 sec after the detection of a P-wave at a single station, and update the results every 0.5 sec up to 3.0 sec. This method analyzes both the waveform and the seismicity forecast information in parallel, and then combines the probabilistic results through a Bayesian framework. The idea is simple: triggers at a seismic station are more likely to have been caused by local earthquakes when 1) strong tremors are observed in the high frequency components of the ground motion and 2) recent seismic activities have been recorded in the proximity of the station.

Most existing earthquake detection algorithms focus only on waveform information, as explained in Chapter 3.3; that is, what is the likelihood an earthquake would produce the real-time waveform just recorded? However, the short time window for data collection in rapid earthquake signal identification can lead to high uncertainties. Also, such waveform

analysis ignores the fact that seismic risks vary consistently with time and location. Adding the seismic forecast information distinguishes the proposed method from any other current EEW detection/classification algorithm. As shown in Chapter 2.1, many studies have shown that seismic activity clusters in time and space, such as foreshock-mainshock-aftershock sequences and swarms earthquakes.

We apply a real-time Epidemic-Type Aftershock Sequences (ETAS) statistical model to forecast near-future seismicity rate as a function of location. This forecast is based on the spatial and temporal clustering properties of the recent earthquakes. For large earthquakes in California, roughly 40% of mainshocks have recorded foreshocks (Abercrombie and Mori 1996), and the forecast results demonstrate promising performance during seismicity sequences, such as all aftershocks and mainshocks following foreshocks. In these cases, the earthquake detection algorithm becomes extremely fast. Of course, not all strong earthquakes are preceded by foreshocks. For the cases without foreshock activity, the ETAS prior is non-informative on the solution due to the probabilistic formulation; the system proceeds just the way it does without any prior information. As sufficient waveform information is available with time, the posterior prediction is dominated by the observation. Combining the heterogeneous data sources using a Bayesian framework thereby improves rapidity and reduces uncertainty to detect of earthquake sources.

Using Bayesian framework, the algorithm aims to provide the probability that a station has been triggered by EEW-relevant earthquake source. Given the observed ground motion at $i$th station immediately following detecting an event, the Bayes' theorem can be expressed as:

$$prob(Y_i = 1|X_{i,t_1:t_n}) \propto prob(X_{i,t_1:t_n}|Y_i = 1)prob(Y_i = 1) \quad [3.10]$$

where $Y_i$ is the classification result at $i$th station, $X_{i,t1:tn} = [X_{i,t_1}, \dots, X_{i,t_n}]$ is a vector of the logs of the maximum ground-motion amplitudes observed at $i$th station from time $t_1$ to $t_n$

after the triggered time, the detailed definition is explained in the 3.2.1 Data section. The posterior probability, $prob(Y_i = 1|X_{i,t_1:t_n})$, is the predictive probability of the observed signal being caused by an earthquake source given the available ground motions. The likelihood function, $prob(X_{i,t_1:t_n}|Y_i = 1)$, describes the predictive probability that the trigger at the $i$th station is due to an earthquake source based on the characteristic similarity of the historical data, also referred to as the training set. The prior information, $prob(Y_i = 1)$, describes the relative probability in earthquake occurrence that may be helpful to identify EEW-relevant earthquake triggers.

By assuming that the observed data $X_{i,t_1:t_n}$ follows independent and identically distributed random variable, the Bayesian equation can be written as:

$$prob(Y_i = 1|X_{i,t1:tn}) \propto \prod_{k=1}^{n} prob(X_{i,tk}|Y_i = 1)prob(Y_i = 1) \quad [3.11]$$

### *3.4.1 Bayesian approach with a Simple Prior*

#### *Likelihood Function*

The definition of the likelihood function is given the signal source (e.g. $Y_i = 1$ as earthquake source or $Y_i = -1$ as non-earthquake source), that is the probability of recording the available waveform information $prob(X_{i,tk}|Y_i)$. In other words, the predictions are made sorely based on the observed ground motions from the waveforms, and the waveform analysis model trained in the previous section can be directly applied.

The likelihood function is described as:

$$prob\left(X_{i,t_k}|Y_i = 1\right) = \phi(X_i) = \frac{1}{1 + e^{-f\left(X_{i,t_k}\right)}} \quad [3.12]$$

where

$$f\left(X_{i,t_k}\right) = 6.884 + 4.8665 * \log10(\mathrm{Za}) - 2.2965 * \log10(\mathrm{Ha}) + 0.2497 * \log10(\mathrm{Zv}) + 2.5895\ \log10(\mathrm{Hv}) \quad [3.13]$$

***Prior Information***

Prior information represents a hypothesis statement regarding to our best knowledge about earthquake identification before examining the waveform data from the on-going rupture. The Bayesian prior, $prob(Y_i = 1)$, provides a relative probability of earthquake occurrence observed in the vicinity of the station. A uniform prior implies that any station in the network is equally likely to observe an earthquake signal versus a noise signal at any given time. The assumption of a uniform prior simplifies the calculation, but it is an overly biased representation of the seismic state. For example, large earthquakes are typically followed by aftershocks in the immediate spatial and temporal vicinity; as a result of the sudden increase of seismic activity during an aftershock sequence, the chance of having triggered signal due to earthquake is much higher than a triggered signal from noise.

The short-term earthquake forecast model aims to quantify the probability of earthquake occurrence. We define the Bayesian prior information as the maximum probability of the ETAS forecast model in the surrounding region of the triggered station:

$$prob(Y_i = 1) = \max prob_{ETAS}(lat, lon) \quad \text{with } |lat - lat_i| \leq 0.5\ , |lon - lon_i| \leq 0.5 \quad [3.14]$$

$prob_{ETAS}(lat, lon)$ is the earthquake probability resulted from an ETAS seismicity forecast model and $(lat_i, lon_i)$ is the location of the triggered station. With the assumption that earthquake source should be in the proximity of the earliest triggered stations, Eq[3.14] constrains the maximum forecast earthquake probability within the 0.5degree proximity (approx. 50km) of the station. Since observed seismicity activities tend to correlate highly with the results from the forecasting models, the Bayesian prior uses the ETAS probability as an indication of possible earthquake occurrence. A trigger at the $i^{th}$ seismic station is more likely to be created by an earthquake source when the ETAS forecast earthquake probability is large; this occurs during seismically active periods such as during an aftershock sequence. The method to calculate ETAS probability is presented in Chapter 2.3.

#### *3.4.1.1 Model Performance*

In order to demonstrate the time-accuracy of the Bayesian model, we compare the likelihood and posterior predictions at every time increments (0.5s window collected ended at 0.5s, 1.0s, 1.5s, 2.0s, 2.5s, 3.0s after the pick time at the station) on the entire dataset. The likelihood prediction is the results from waveform analysis model presented in Chapter 3.3 and the posterior prediction is the result from the Bayesian model described in the previous section.

Table 3.4 shows the confusion matrix for the classification of earthquake versus non-earthquake records based on the Bayesian approach under likelihood function and posterior function. The decision boundary is set at 50%, which is if the data is classified as an earthquake event if the predictive probability reaches above 50%; otherwise it is classified as a non-earthquake event. For the 0.5 sec analysis, the number of predicted earthquakes is reduced from 1053 (957 TP and 96 FP) in the likelihood prediction to 738 (727 TP and 11

FP) in the posterior probability, because the posterior function is more conservative in making predictions as it evaluates both the waveform and the prior seismicity information. As a result, the false alarm rate at the earliest prediction of 0.5 sec is significantly reduced in posterior prediction from 96 to 11. The seismicity prior model catches the dynamic change in the spatial-temporal clustering phenomenon in seismicity occurrences. For the same type of signals, the model tends to provide relatively higher probability results during a seismically active period, and predict a lower probability during seismically dormant period. For example, during a seismically active period such as an aftershock sequence, the relatively higher probability in prediction allows a quicker convergence to earthquake prediction for earlier alert delivery. On the other hand, when no seismic activity have been observed in the recent past, the system would take more time to identify the event to guarantee the level of accuracy in the prediction. As time progresses to 3.0sec, the posterior prediction also shows a decrease in the missed events, because the likelihood prediction dominates the solution as more available waveform data is collected.

| Available Data | Predicted class | True Classes | | | | Precision | | Accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| | | Likelihood $prob(X_i\|Y_i)$ | | Posterior $prob(Y_i\|X_i)$ | | Likelihood | Posterior | Likelihood | Posterior |
| | | Earthquake | Non-Earthquake | Earthquake | Non-Earthquake | | | | |
| 0.5s | Earthquake | 957 | 96 | 727 | 11 | 90.9% | 98.5% | 89.2% | 83.4% |
| | Non-Earthquake | 171 | 1257 | 401 | 1342 | | | | |
| 1.0s | Earthquake | 1035 | 61 | 939 | 14 | 95.9% | 98.5% | 93.7% | 91.8% |
| | Non-Earthquake | 93 | 1292 | 189 | 1339 | | | | |
| 1.5s | Earthquake | 1070 | 46 | 1010 | 18 | 95.9% | 98.2% | 95.8% | 94.5% |
| | Non-Earthquake | 58 | 1307 | 118 | 1335 | | | | |
| 2.0s | Earthquake | 1094 | 41 | 1049 | 23 | 96.4% | 97.8% | 96.9% | 95.8% |
| | Non-Earthquake | 34 | 1312 | 79 | 1330 | | | | |
| 2.5s | Earthquake | 1105 | 40 | 1076 | 25 | 96.5% | 97.7% | 97.4% | 96.9% |
| | Non-Earthquake | 23 | 1313 | 52 | 1328 | | | | |
| 3.0s | Earthquake | 1107 | 36 | 1089 | 27 | 96.8% | 97.5% | 97.7% | 97.3% |
| | Non-Earthquake | 21 | 1317 | 39 | 1326 | | | | |

Table 3.4 Model performance of the Bayesain model with a simple prior at time increments: 0.5s, 1.0s, 1.5s, 2.0s, 2.5s, 3.0s after the triggered time at the station

The objective of likelihood function is to minimize the loss function of the sum of missed and false alerts by analyzing the incoming waveform. However, the goal of the Bayesian posterior prediction prioritizes the minimization of false alerts since false alerts could confuse the decision making process of the entire seismic network while the initial missed alerts can be successfully identified (including with alternative existing algorithms) with more time and data. Similarly in Chapter 3.3.2 model selection, we focus on the precision and accuracy measures. Also shown in Table 3.4, the posterior prediction consistently provided a high precision rate, meaning low false alarm rate in the predictions. Although adjusting the decision boundary could reduce false alarms, it would be subjective as to how to adjust the value to achieve optimized results. The Bayesian approach is able to reduce the number of false alarms with the additional prior information. The initial accuracy is lower in posterior prediction than likelihood prediction because the initial high uncertainty in some of the earthquake records requires more time for discrimination; after the final update in the predictions, the accuracy rate at 3.0 s reaches to 97.3%.

The Likelihood prediction uses peak amplitudes recorded from the waveforms; the posterior prediction combines the Likelihood probability with ETAS forecast probability. The posterior probability shows 1) consistent high precision percentage-an indication of low false alarm rate, and 2) high final accuracy percentage-an indication of low missed alarm rate.

As shown in Figure 3.4, the predicted probability for all PGA ranges increases with time because as more data become available the uncertainties in the prediction are reduced. Also, the records with larger PGA range are predicted with a higher probability at an earlier

time. This indicates that the discriminant function performs well for the large events. To optimize the speed-reliability trade-off, we would recommend utilizing the classification starting at 0.5s after the trigger.

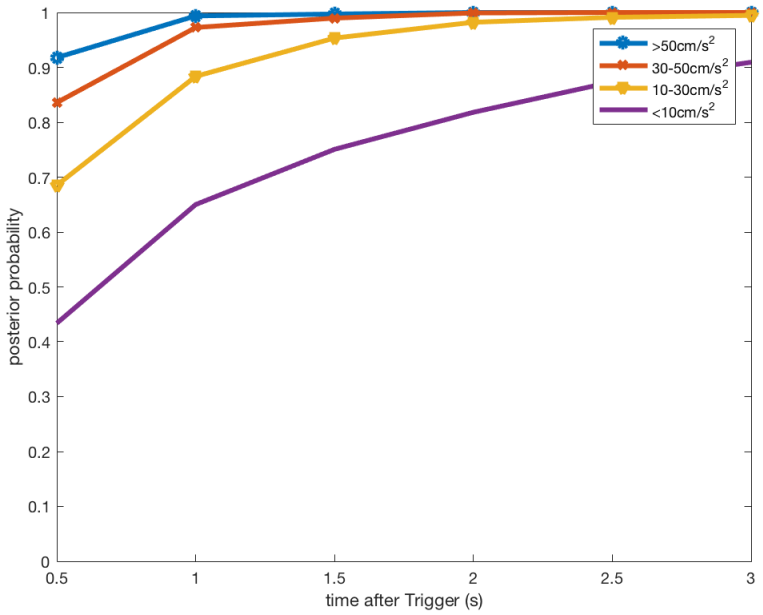



Figure 3.4 Average posterior prediction probabilities for earthquake records with various PGA range

### *3.4.2 Bayesian approach with a Modified Prior*

How to appropriately define the prior information in a Bayesian framework has been a challenging issue. Although using the maximum ETAS probability has significantly improved the initial precision of the predictions (that is the false alarm rate from the rapid predictions is low, and thus reliability of early alerts is improved), it is a tuning parameter that I have chosen based on the empirical analysis. I started to formulate a new prior information function from the fundamental.

Under Bayesian framework, the prior information in this particular question of identifying earthquake signals is expressed as $prob(Y_i = 1)$. Based on the observed triggering information at the seismometer stations, the probability that a trigger is due to an earthquake should be the fraction of expected earthquake trigger out of the total number of the expected triggers observed at the station (including earthquake and non-earthquake triggers), the equation can be formulated as:

$$prob(Y_i = 1) = \frac{\lambda_{Eq}}{\lambda_{Eq} + \lambda_{nonEq}} \qquad [3.15]$$

where $\lambda_{Eq}$ is the expected rate of earthquake triggers and $\lambda_{nonEq}$ is the expected rate of non-earthquake triggers.

In general, the prior information is defined as the knowledge collected before seeing any of the observation data. However, in the problem of EEW, the predictive model is activated only if a trigger has been observed at a seismometer station. A trigger is detected when the ratio of short-term average to long-term average (STA-LTA) of the real-time incoming waveforms has exceeded a threshold. Quantitatively, this implies that sudden ground shaking amplitude, $X_i$, must be higher than the ambient noise level in the ground shaking or

an amplitude threshold ($amp$). With the triggering amplitude information, we can further express Eq[3.15] as follows:

$$prob(Y_i = 1) = prob(Y_i = 1 | X_i > amp) = \frac{\lambda_{Eq_{> amp}}}{\lambda_{Eq_{> amp}} + \lambda_{nonEq_{> amp}}} \quad [3.16]$$

$\lambda_{Eq_{> amp}}$ is the expected rate of earthquake triggers that can create an amplitude greater than the observed value; $\lambda_{nonEq_{> amp}}$ is the expected number of non-earthquake triggers that can create an amplitude greater than the observed value.

First, let's focus on the calculation of $\lambda_{nonEq_{> amp}}$.

$$\lambda_{nonEq_{> amp}} = prob(amp_{nonEq} > amp_{obs}) * \lambda_{nonEq} \quad [3.17]$$

where $prob(amp_{nonEq} > amp_{obs})$ is the percentage or fraction of recording non-earthquake ground motion amplitudes greater or more extreme than the observed ground motion amplitude, and $\lambda_{nonEq}$ is the expected total rate of non-earthquake triggers. As observed in the waveform amplitude information, the vertical velocity waveform amplitude is chosen for the ground motion amplitude here because the initial p-wave motion is more evidently shown in the vertical channel, and velocity waveform is often used for STA-LTA triggers for stability. $\lambda_{nonEq}$ can be approximated by the average non-earthquake triggering rate observed at the station. Assuming that ambient noise, traffic, and regular surrounding activities, cause most of the non-earthquake triggers, the rate of non-earthquake triggers should be relatively stable per geographical locations. For example, one can keep track of station specific triggers over one month or six months period of time, and use the daily false triggers as the $\lambda_{nonEq}$. The station specific parameter is particularly important because the number of triggers varies significantly across the seismic network.

For instance, stations in the urban area, such as Downtown Los Angeles, could observe over 100 triggers daily, whereas stations in suburbs, such as near the Mojave Desert, sometimes observe less than a single trigger annually. Information on how likely a trigger is due to noise can directly be obtained from the location of the stations. Since it is challenging to obtain records of the number of false triggers with the current EEW system, I chose an estimation of $\lambda_{nonEq} = 10$ per day across all the stations in the network for simplicity. For future investigation, it is important to keep track of station specific information for the entire seismic network.

To calculate $prob(amp_{nonEq} > amp_{obs})$, the distribution of the noise amplitude needs to be statistically analyzed using p-value concept. Figure 3.5 is a histogram of the vertical log(PGV) at the 0.5 sec after trigger from the 1000 noise data randomly sampled from 100 most noisiest stations across the network during 2015. The line of best fit shown on top of the histogram shows that the data follows a normal distribution. The mean of the distribution is -2.8815, and the standard deviation is 0.5128.

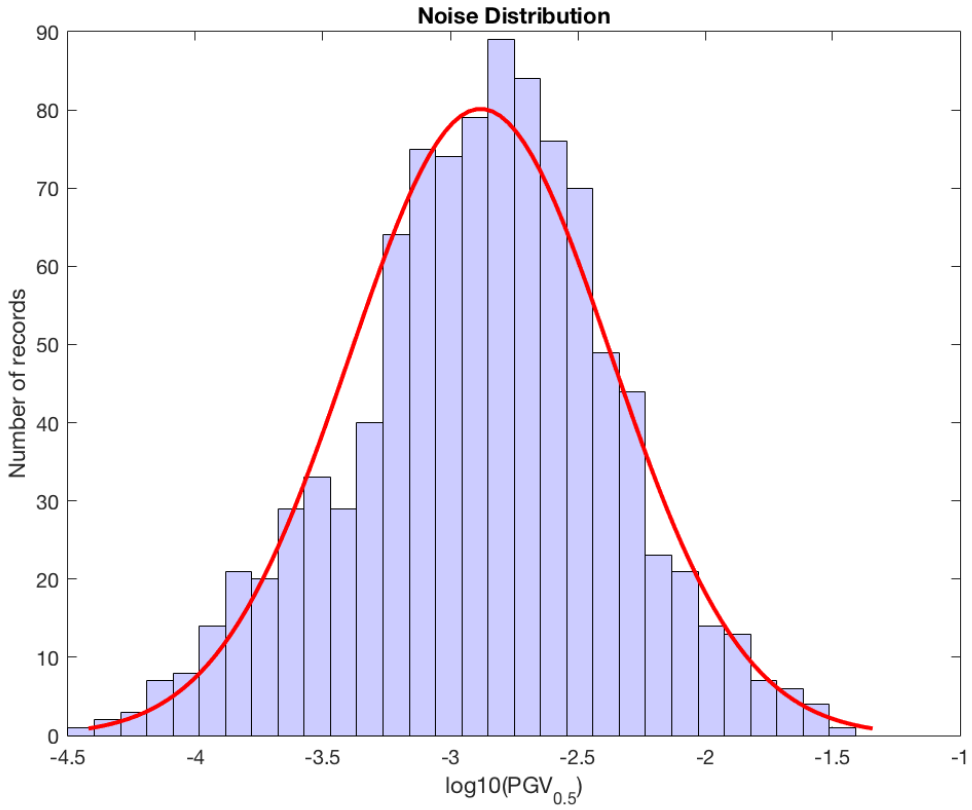



Figure 3.5 Histogram of the vertical $\log_{10}$(PGV) at the 0.5s after trigger from the 1000 noise data randomly sampled from 100 most noisiest stations across the network during 2015

***p-value***

Assuming that the $\log_{10}(PGV_{0.5})$ of noise triggers follows the normal distribution above, every newly recorded $\log_{10}(PGV_{0.5})$ will be evaluated to see how likely the data belongs to the noise distribution. Intuitively, the recorded $\log_{10}(PGV_{0.5})$ is very likely to a noise trigger if it is close to the mean the distribution since most of the noise observations are recorded around the center region in a normal distribution. On

the other hand, the observation does not very likely belong to the distribution if the $\log_{10}(PGV_{0.5})$ recorded is much higher or lower from the mean. Mathematically, p-value is a terminology to quantify the statistical significance of an observation under null hypothesis. The definition of p-value is the probability of finding the observed data (or more extreme data) assuming the null hypothesis is true; graphically, it is computed by summing the area under the normal curve from the observed data point to the infinities (extremes), as shown in Figure 3.6 for the example of noise identification. As an example, if the observed $log_{10}(PGV_{0.5})$ falls around the center of the normal distribution near the mean, the p-value approaches 100% implies that the triggered signal most likely belongs to the noise distribution. On the other hand, an observation that falls near the tails of the curve indicates low p-value. Even in the cases with very low p-value, the evidence can only suggest very low probability of being noise, but cannot reject the signal being caused by alternative sources (e.g. low p-value can not suggest the observation indicates earthquake signal). Since $prob(amp_{nonEq} > amp_{obs})$ serves a similar purpose, the calculated p-value is used to approximate $prob(amp_{nonEq} > amp_{obs})$:

$$prob(amp_{nonEq} > amp_{obs}) \approx \text{p-value}$$

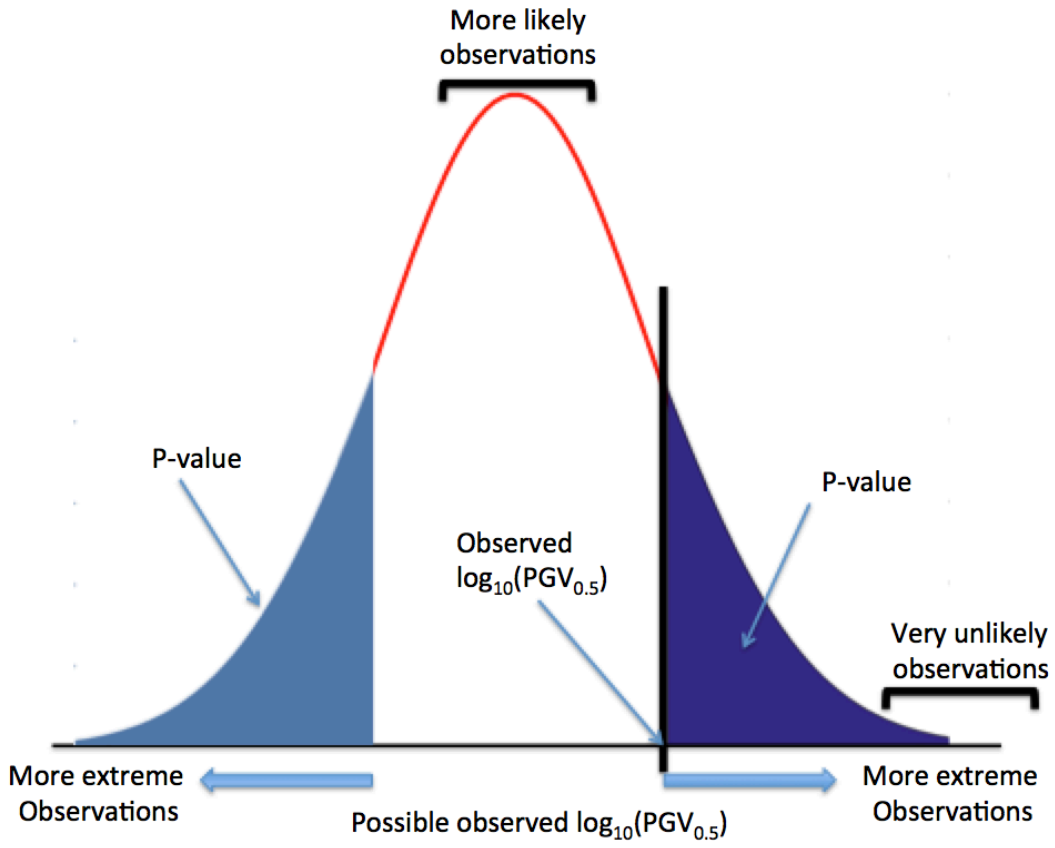



Figure 3.6 Visual explanation of two-tailed p-value computation under Normal Distribution

In summary, the expected rate of non-Earthquake triggers, $\lambda_{nonEq_{>amp}}$, is calculated based on the probability of having such amplitude noise trigger and the total number of false triggers observed at the specific station. This takes into consideration both the specific triggering information and the particular geological location data.

Now, let's focus on the calculation of $\lambda_{Eq_{>amp}}$, which is derived from ground motion prediction equations (GMPE) and seismicity forecasting models presented in Chapter 2.

The idea is clear: the goal is to calculate the total rate of seismicity that could produce the waveform amplitude recorded at the seismic station. Firstly, taking the vertical $\log_{10}(PGV_{0.5})$, a question is raised: what kind of earthquake source could create such observed waveform amplitude? The GMPE by (Cua and Heaton, 2007) can help answer the question. Mathematically, this GMPE is a function that takes into account of soil properties, hypocenter distance, and magnitude, type of wave phase (p-wave or s-wave), type of waveform (velocity or acceleration), and produces peak ground motion. By assuming that the observed $PGV_{0.5}$ is the peak ground velocity of the p-wave (a reasonable assumption because earliest triggers in EEW tend to cause by near-site earthquake sources), the magnitude of the earthquake source can be calculated for any given location by inverse the GMPE model. Next, the region surrounding the station is discretized into grids of $(lat, lon)$, the distance between each of the point on the grid to the triggered station can be calculated, calculates the minimum magnitude. As a sanity check, due to the wave-attenuation property, the calculated magnitude increases with distance for the same observed PGV. Then, the location of $(lat, lon)$ and the calculated magnitude needs to into the ETAS model in eq[2.2] to predict the expected rate of seismicity at each of the discretized grid. The ETAS seismicity forecast model presented in Chapter 2 calculates the all of the expected rate of earthquakes with a minimum magnitude and above at every point on the earth surface. Since the ETAS are produced in real-time, the regional seismicity rate can be directly pulled out with minimum computational effort (creating no delay in the EEW alerts), the calculation of $\lambda_{Eq_{>amp}}(lat, lon, M_{min})$ becomes simple. Lastly, the total seismicity rate is calculated by summing the discretized seismicity rate in the vicinity of the triggered station, mathematically it follows $\lambda_{Eq_{>amp}} = \sum_{lat,lon} \lambda_{Eq_{>amp}}(lat, lon, M_{min})$. For simplicity, the region of $[-0.5, +0.5]$ latitude degree by [-0.5, +0.5] longitude degree area centered at the triggered seismic station is considered as the region of interest in this study instead of applying a circular region with radius r km. For practical implementations,

the region to be considered should depend on the station density of the network due to the propagation of seismic waves, the earthquake source tend to locate very close to the first triggered station in the network. For a perfect seismic network with no malfunctioning stations, the concept of voronoi cell can be considered.

The concept follows the flow chart in Figure 3.7.

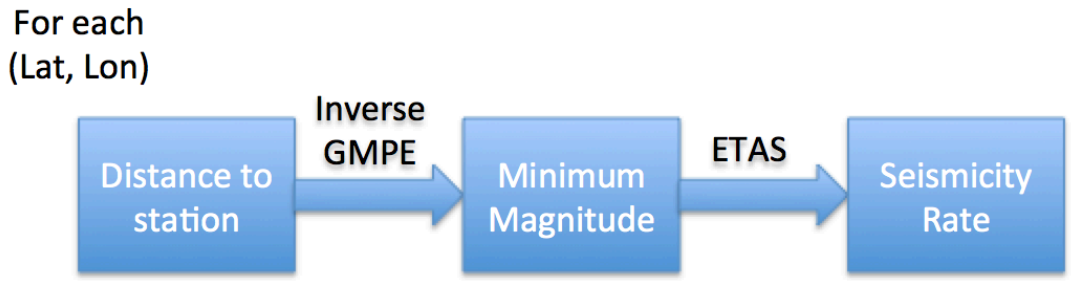



Figure 3.7 Seismicity rate calculation in a flow chart

To visually demonstrate the calculation of $\lambda_{Eq_{>amp}}$, consider the M4.3 La Habra earthquake as an example, as shown in Figure 3.8 to Figure 3.10. On March 29, 2014, 4:11:14 am a M4.3 La Habra earthquake triggered station CI.FUL along with other the near by local seismic stations. The recorded $\log_{10}(PGV_{0.5})$ is -1.1217 (about $PGV_{0.5}$ = 0.61 cm/s). Figure 3.8 shows the source to station distance map; the calculation is based on the great-circle distance between the discretized location and the station location. Figure 3.9 shows the minimum magnitude intensity that could produce such amplitude; this calculation is done by applying the inverse GMPE by (Cua and Heaton, 2007). Figure 3.10 shows the seismicity rate expected at each location from the ETAS model. Since the forecasted seismicity varies significantly depending on the previously observed seismic

activities, the same waveform amplitude would reflect dramatically different seismicity rate recorded at a different time or location. For this particular event, since three foreshocks with M2-M3 were observed near the station less than 24hr prior, the expected seismicity rate is relatively higher in the northwest region about 5km away from the station.

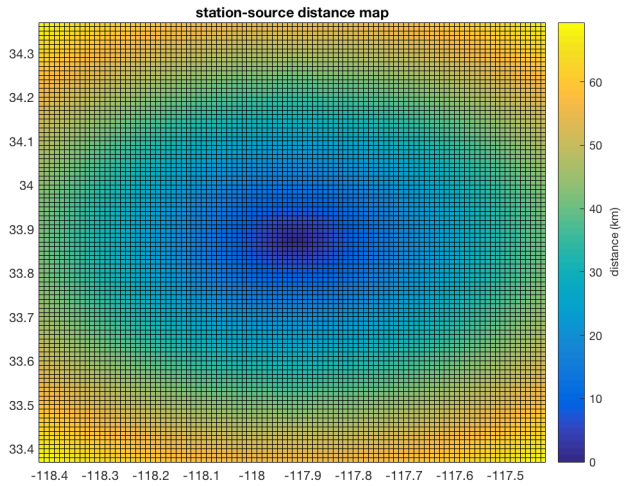



Figure 3.8 Source to station Distance Calculation based on the initial p-wave amplitude observed at the station

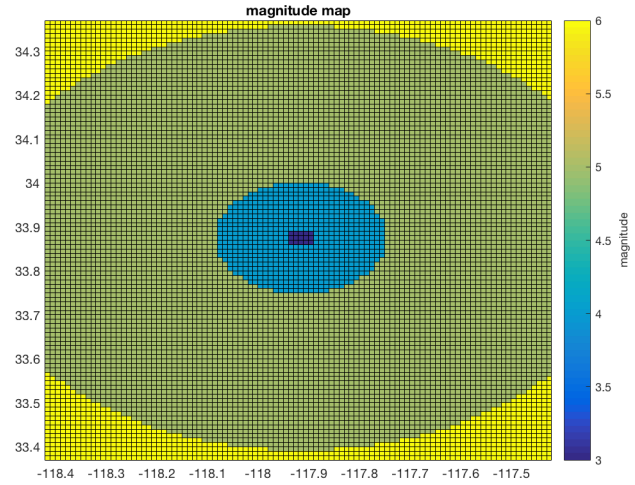



Figure 3.9 Magnitude calculation at the sources based on the initial p-wave amplitude observed at the station

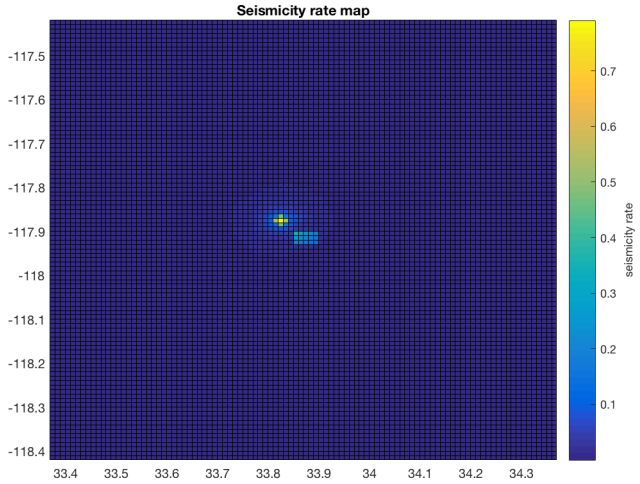



Figure 3.10 Seismicity rate calculation obtained from the ETAS model

After obtaining the seismicity rate map from Figure 3.10, the total seismicity rate $\lambda_{Eq_{>amp}}$ is calculated by summing all seismicity rates in the region. In this particular trigger at CI.FUL for the M4.3 La Habra earthquake, $\lambda_{Eq_{>amp}} = 18.6$.

### *3.4.2.1 Model Performance*

In order to demonstrate the time-accuracy of the Bayesian model, we compare the likelihood and posterior predictions at every time increments (0.5s window collected ended at 0.5s, 1.0s, 1.5s, 2.0s, 2.5s, 3.0s after the pick time at the station) on the entire dataset. The likelihood prediction is the result of the waveform analysis model presented in Chapter 4.3, and the posterior prediction is the result of the Bayesian model described in the previous section.

Table 3.5 shows the confusion matrix for the classification of earthquake versus non-earthquake records based on the Bayesian approach under likelihood function and posterior function. The decision boundary is set at 50%, meaning that the data is classified as an earthquake event if the predictive probability reaches above 50%; otherwise it is classified as a non-earthquake event.

| Available Data | Predicted class | True Classes | | | | Precision | | Accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| | | Likelihood $prob(X_i\|Y_i)$ | | Posterior $prob(Y_i\|X_i)$ | | Likelihood | Posterior | Likelihood | Posterior |
| | | Earthquake | Non-Earthquake | Earthquake | Non-Earthquake | | | | |
| 0.5s | Earthquake | 957 | 96 | 843 | 17 | 90.9% | 98.0% | 89.2% | 87.8% |
| | Non-Earthquake | 171 | 1257 | 285 | 1336 | | | | |
| 1.0s | Earthquake | 1035 | 61 | 944 | 20 | 95.9% | 97.9% | 93.7% | 91.8% |
| | Non-Earthquake | 93 | 1292 | 184 | 1333 | | | | |
| 1.5s | Earthquake | 1070 | 46 | 1013 | 24 | 95.9% | 97.7% | 95.8% | 94.4% |
| | Non-Earthquake | 58 | 1307 | 115 | 1329 | | | | |
| 2.0s | Earthquake | 1094 | 41 | 1052 | 28 | 96.4% | 97.4% | 96.9% | 95.8% |
| | Non-Earthquake | 34 | 1312 | 76 | 1325 | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2.5s | Earthquake | 1105 | 40 | 1081 | 25 | 96.5% | 97.5% | 97.4% | 97.0% |
| | Non-Earthquake | 23 | 1313 | 47 | 1325 | | | | |
| 3.0s | Earthquake | 1107 | 36 | 1095 | 28 | 96.8% | 97.5% | 97.7% | 97.5% |
| | Non-Earthquake | 21 | 1317 | 33 | 1325 | | | | |

Table 3.5 Model performance of the Bayesian model with a modified prior at time increments: 0.5 s, 1.0 s, 1.5 s, 2.0 s, 2.5s, 3.0 s after the triggered time at the station

## 3.5 Comparison Results of Waveform Analysis vs. Bayesian models

Following the definitions of TP (True Earthquake predictions), FP (False earthquake predictions), FN (missed earthquake predictions), and TN (True non-earthquake predictions), the results from the three proposed models at 0.5sec, 1.5sec, and 3.0sec are presented plot in Figure 3.11 to Figure 3.13, respectively. The results of the accuracy and precision metrics as functions of time are plotted in Figure 3.14 and Figure 3.15.

The predictive results for all three models vary the most at 0.5 sec, while the results are almost indifferent with no preference at 3.0 sec. This shows that the Bayesian approach with prior information has a much stronger influential initially when the waveform is not sufficient. In rapid detections of EEW, the algorithm aims to minimize the accuracy ratio, which is the rate of False Earthquake predictions and the rate of missed Earthquake predictions. As shown in Figure 3.11, waveform analysis has the highest number of false prediction, while the Bayesian framework with the simple prior has the highest number of missed predictions. The Bayesian framework with the modified prior compromises the benefits of moderate false and missed predictions. As time progresses, the waveform analysis component dominates the results because the model is able to observe more available on-going data. Therefore, the prediction results are similar at the 3.0sec analysis.

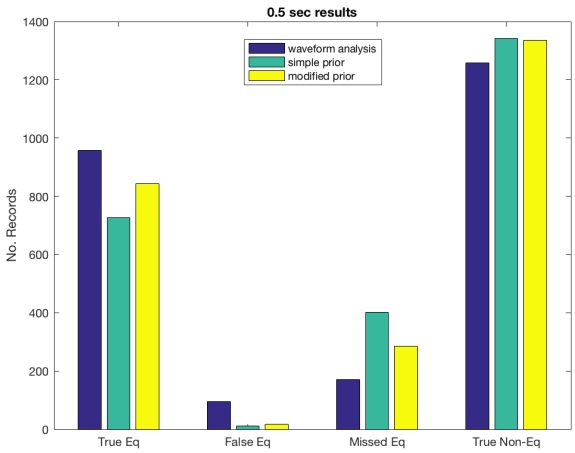



Figure 3.11 Comparison of the predictive results from waveform analysis, Bayesian model with simple prior, and Bayesian model with modified prior at 0.5sec after station trigger.

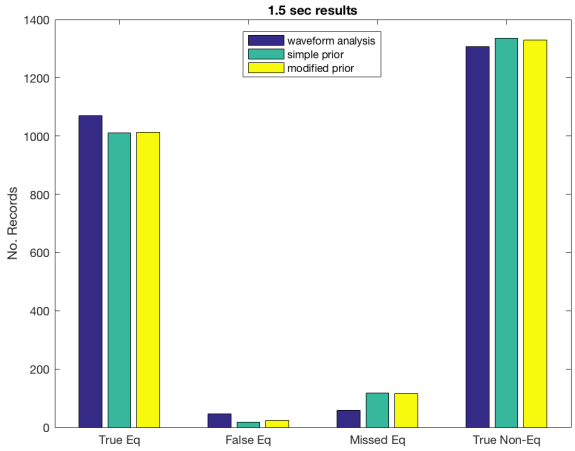



Figure 3.12 Comparison of the predictive results from waveform analysis, Bayesian model with simple prior, and Bayesian model with modified prior at 1.5sec after station trigger.

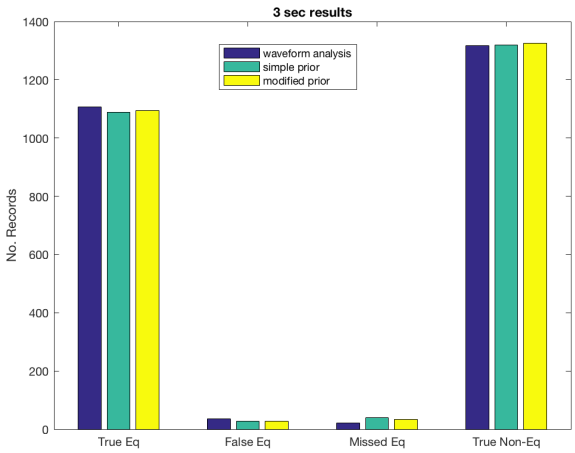



Figure 3.13 Comparison of the predictive results from waveform analysis, Bayesian model with simple prior, and Bayesian model with modified prior at 3.0 sec after station trigger.

In addition to the analysis of the raw categorical records, the precision and accuracy metrics from Eq [3.6] and Eq [3.7] are also considered. In reality, an earthquake is a localized rare event. Therefore, if all stations in the network are considered, the number of non-earthquake stations will dominate in the total number of triggers. The accuracy measure is sensitive to this unbalanced data records. As shown in Figure 3.14, accuracy measures for all three models improve with time. The waveform analysis demonstrates better accuracy over all 6 predictions because waveform analysis has much lower missed predictions. However, accuracy (the sum of all tree predictions) is not the only goal of performance measure. In fact, in the challenge of rapid earthquake identification, low false alarm rate is more important than missed alarms (false alarm leads to confusion in the system, while temporary missed alarm might be adjusted with more in-going data), so precision analysis also needs to be emphasized.

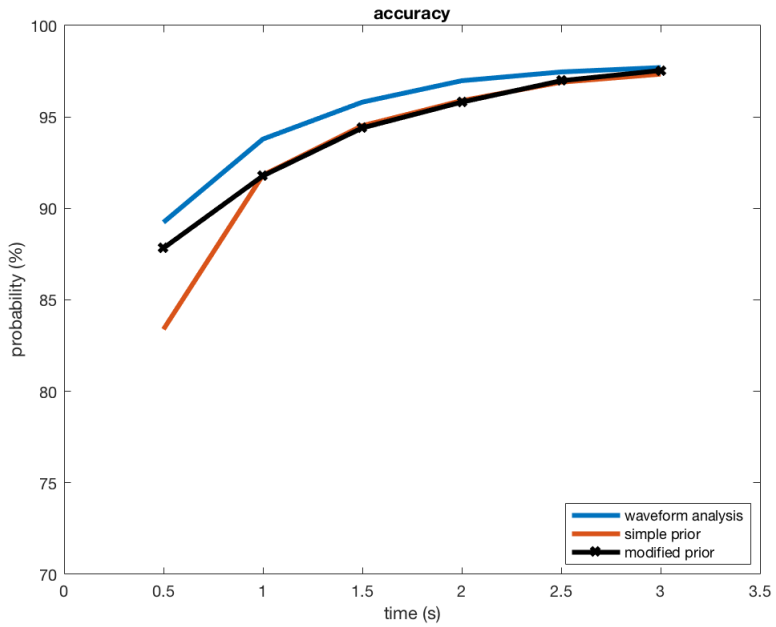



Figure 3.14 Accuracy rate (%) for waveform analysis, and Bayesian model with simple prior, Bayesian model with modified prior as a function of time.

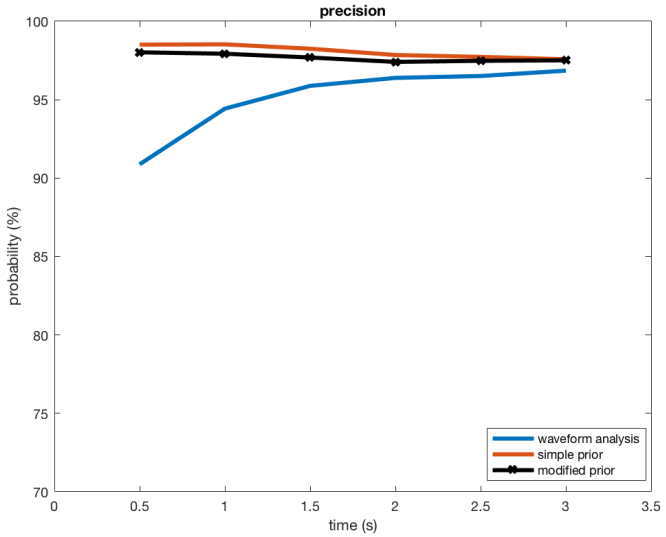



Figure 3.15 Precision rate (%) for waveform analysis, Bayesian model with simple prior, and Bayesian model with modified prior as a function of time.

In the precision analysis of Figure 3.15, both of the Bayesian models outperform over the waveform analysis over 5% in the first alert. The reason is that the rates of false predictions are extremely low for the Bayesian models. As more data comes in, the precision rates of all three models converge.

All three models demonstrate high accurate predictions. By compromising the precision and accuracy requirements for the EEW purposes, the Bayesian model with the modified prior is recommended. The predictions from this model are further validated through cross-validation for robustness and comparing with existing algorithms. To summarize the signal discrimination process (classifying earthquake versus noise) for real-time implementation for EEW, the steps in the proposed model follow the flow chart in Figure 3.16.

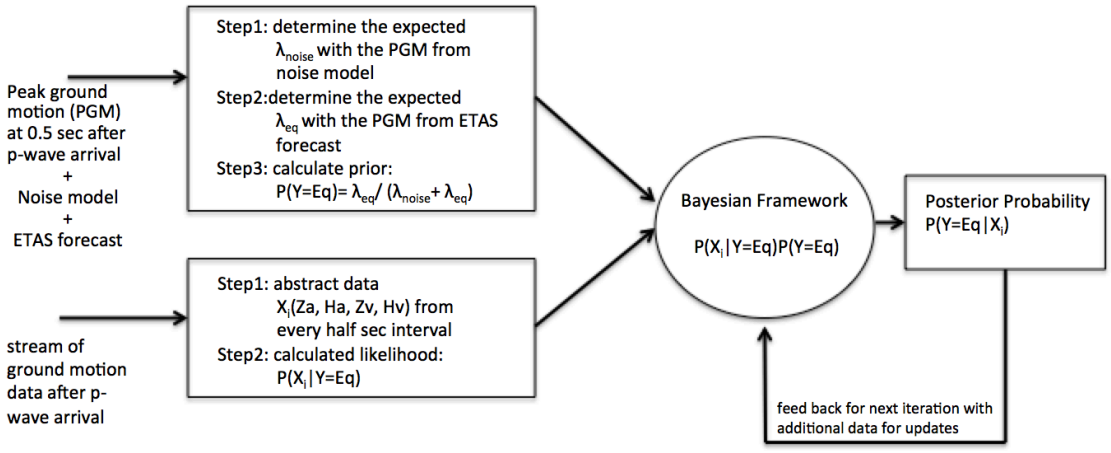



Figure 3.16 Flow chart of the proposed signal discrimination process for real-time implementation

## 3.6 Cross-validation Results

To examine the robustness of the model for future performance, we also performed the leave-one-out cross validation method on the data set for predictions at every time step. The leave-one-out method divides the data into two subsets, using all the data except one to train the model, and then validates the prediction on the left out data point to evaluate the performance of the model. The process is repeated until all the data has been assessed as the validation data, which in our case is 2,481 data points. The idea of the method is to use the validation set as 'new' data to test the performance of the trained model; the mis-classification rate on the validation data set is a good indication of how well the model will predict in the future.

The cross-validation performance is similar to the training performance. Note that Table 3.6 Cross-validation confusion matrix shows the prediction result on the validation set only.

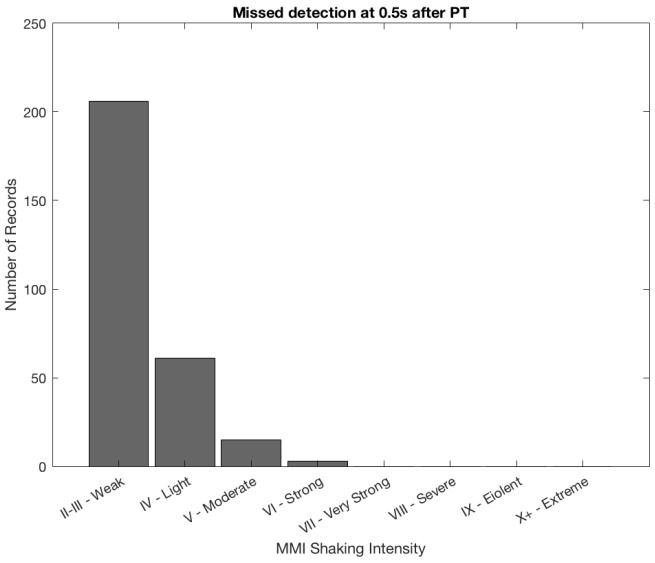
The precision rate of earthquake detection is above 98%, which is similar in the training evaluation in Table 3.5. This confirms the robustness of the model performance for future earthquake data.

| | Predicted class | True Class | | | |
|---|---|---|---|---|---|
| | | Earthquake | Non-Earthquake | Precision | Accuracy |
| 0.5 s | Earthquake | 843 | 18 | 97.91% | 87.79% |
| | None Earthquake | 285 | 1335 | | |
| 1.0 s | Earthquake | 943 | 20 | 97.82% | 91.74% |
| | Non-Earthquake | 185 | 1333 | | |
| 1.5 s | Earthquake | 1013 | 24 | 97.69% | 94.4% |
| | Non-Earthquake | 115 | 1329 | | |
| 2.0 s | Earthquake | 1052 | 28 | 97.41% | 95.81% |
| | Non-Earthquake | 76 | 1325 | | |
| 2.5 s | Earthquake | 1081 | 29 | 97.39% | 96.94% |
| | Non-Earthquake | 47 | 1324 | | |
| 3.0 s | Earthquake | 1095 | 28 | 97.51% | 97.54% |
| | Non-Earthquake | 33 | 1325 | | |

Table 3.6 Cross-validation confusion matrix

Figure 3.17 shows the MMI shaking intensity of the missed earthquake events at 0.5s and 3.0s after the triggered time. At 0.5 s after the triggered time, 80% of the 401 missed events are within the weak shaking intensity range; at 3.0s after the triggered time, all the 39 missed events are within the weak shaking intensity (less than 14cm/$s^2$). The consequence of the remaining missing events is negligible for the purpose of EEW. For a rapid discriminant algorithm that is activated promptly after the trigger time at a single station, the priority is to minimize the false alarm rate that could lead to confusion in the entire system, where as the temporary missed events can be identified slightly later using more available data. The cross-validation result is consistent with the training result in Table 2; it provides us with more confidence in the reliability of the algorithm in making predictions in the future.

a)




b)

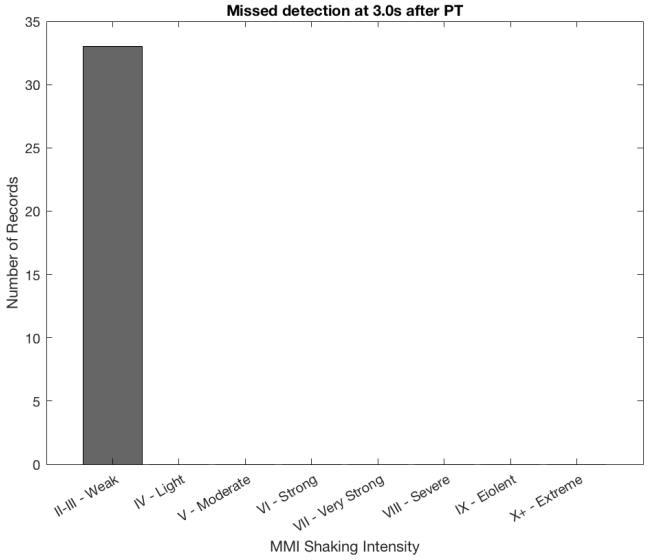



Figure 3.17 MMI Shaking Intensity for the missed earthquake events at a) 0.5 s and b) 3.0 s after triggered time

## 3.7 Comparison to the $\tau_c - P_d$ trigger criterion

The result of the new method is also compared with a $\tau_c - P_d$ trigger criterion applied with 3.0 s data. $\tau_c$ and $P_d$ are well-established EEW parameters observed during initial ground motion, which are used to predict the final magnitude of an ongoing earthquake (Kanamori 2005). The period parameter $\tau_c$ is defined as $\tau_c = 2\pi/\sqrt{r}$ where $r = \left[\int_0^{\tau_0} \dot{u}^2(t)dt\right]/\left[\int_0^{\tau_0} u^2(t)dt\right]$, $u(t)$ is the vertical displacement, and $\tau_o$ is the time window used. $\tau_o$ is set at 3 s in the current Onsite algorithm implemented in California. The $\tau_c - P_d$ trigger criterion was proposed by (Bose, Heaton and Hauksson 2012) to reduce the number of false alerts from non-earthquake signals. The $\tau_c - P_d$ criterion quantizes the quality of each trigger by assigning a parameter Q to the pair value of $\tau_c - P_d$ determined from the initial 3.0 s of P-wave data. The parameter Q is assigned with one of three values: 0, 0.5, and 1. Q=0.5 and Q=1 mean that the detected events are considered moderate to good quality data that are relevant to EEW, and Q=0 means that the trigger is due to a non-earthquake source. Figure 3.18 shows the $\tau_c - P_d$ plot of all earthquake and non-earthquake (noise and teleseism) data in our data set using a 3.0 s window following the P-wave trigger, Q=1 is the region between the 2 solid lines, and Q=0.5 is the region between the solid line and the dashed line. All the points outside of the dashed lines are assigned to be Q=0.

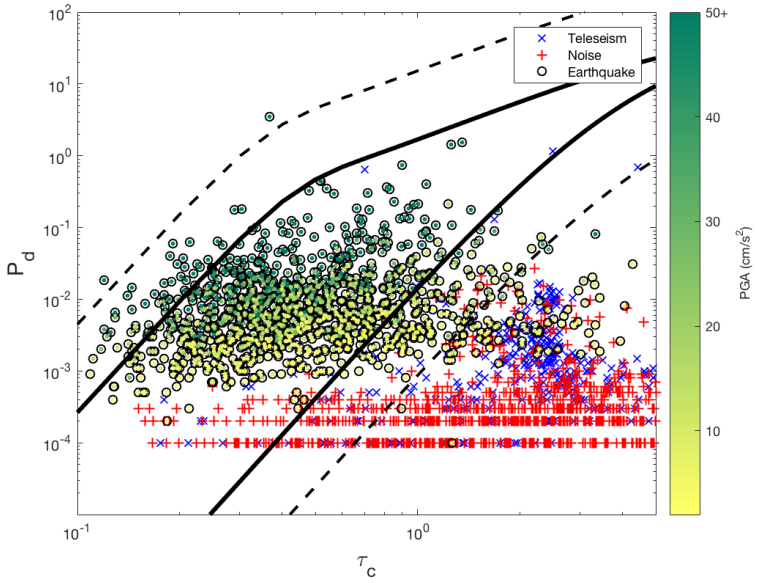



Figure 3.18 $\tau_c - P_d$ plot of all earthquake and non-earthquake (noise and teleseism) data in our data set using a 3.0 s window following the P-wave trigger for measurement. The solid line and dashed line are the decision boundaries of the parameter Q=1 and Q=0.5, respectively. The color intensity of the earthquake data represents the PGA observed

The predictive result of the new method at 0.5 s and 3.0s is compared to the result of the $\tau_c - P_d$ trigger criterion with the decision boundary at Q=1 and Q=0.5. The confusion matrix of the predicted result is shown in Table 3.7.

| | Predicted Class | True Class | | | | |
|---|---|---|---|---|---|---|
| | | Earthquake | Non-Earthquake | Total Predictions | Precision | Accuracy |
| 0.5s | Earthquake | 843 | 17 | 860 | 98.0% | 87.82% |
| | Non-Earthquake | 285 | 1336 | 1621 | | |
| 3.0s | Earthquake | 1095 | 28 | 1123 | 97.5% | 97.4% |
| | Non- | 33 | 1325 | 1358 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Earthquake | | | | | |
| Tc-Pd 3.0s Q=1 | Earthquake | 793 | 52 | 845 | 93.85% | 84.40% |
| | Non-earthquake | 335 | 1301 | 1636 | | |
| Tc-Pd 3.0s Q=0.5 | Earthquake | 955 | 206 | 1161 | 82.26% | 84.72% |
| | Non-earthquake | 173 | 1147 | 1320 | | |
| | Total Observations | 1128 | 1353 | | | |

Table 3.7 Comparison results of the proposed method with $\tau_c - P_d$ method.

Our proposed method has a higher accuracy rate compared to the $\tau_c - P_d$ method. This might be the result of several factors: 1) calculation of $\tau_c$ and $P_d$ uses the integration of displacement and velocity amplitudes, where long-period trends could be significantly amplified during multiple integration from the acceleration record, while the proposed method directly uses peak amplitudes to reduce processing artifacts; 2) $\tau_c - P_d$ uses the vertical component of ground motion data only, where the proposed method uses all three-component data and catalog data; 3) the proposed method is empirical and directly data-driven, which better characterizes the actual recorded observations. Most importantly, the proposed method provides confident results at 0.5 s after the trigger time, whereas $\tau_c - P_d$ currently for a fixed-window of data collection of 3 s. Of course, it can be modified for different time windows, but it requires further testification and calibration.

## 3.8 Examples

The application of the algorithm is demonstrated for four events, including ambient noise false triggers on 24 March 2015, a M3.7 Chino Hills aftershock on 29 July 2009, a M4.7

Los Angeles mainshock on 18 May 2008, and a M7.8 Japan teleseismic earthquake on 29 May 2015. All four events were excluded from the training data set, so the performance of these examples reflects the true behavior of the prediction of new data in the future. For each event, the analyses of the two near-field stations are presented.

***24 March 2015 – ambient noise false triggers in Southern California***

False picks are triggering the seismic networks hundreds to thousands times daily, especially in the urban areas. As an example, two false triggers at CI.CFS and CI.NEN stations occurred near Lancaster and San Bernardino on 24 March 2015, respectively. Although a sudden increase in acceleration has been observed at both stations, the PGA ground motions are less than 2cm/s$^2$, as shown in Figure 3.19. The benefits of the Bayesian approach are clearly demonstrated through these examples. If waveform analysis is the only model used for predictions, the large triggering amplitudes indicating high earthquake probability (over 60% being earthquake in both cases) could lead to incorrect alerts delivered. Although the conventional EEW method might correctly predict that the trigger is due to a non-earthquake source (less than 2% being earthquake in both cases), the decision is made after a fixed time window of 3.0 s after triggering. Alternatively, the posterior probability from the proposed method of Bayesian analysis makes the decision that the signal is non-earthquake immediately at 0.5 s after trigger, which optimize both the leading time for alerts and accuracy in the prediction. This is because 1) the forecasted earthquake probability in the prior information is low since no seismic activities were observed in the recent past and 2) the relatively high amplitude of ground shaking diminishes after less than a second which dissimilar to the patterns and characteristics of an ongoing earthquake. The two false triggers could be caused by busy traffic, collisions, or trains passing by.

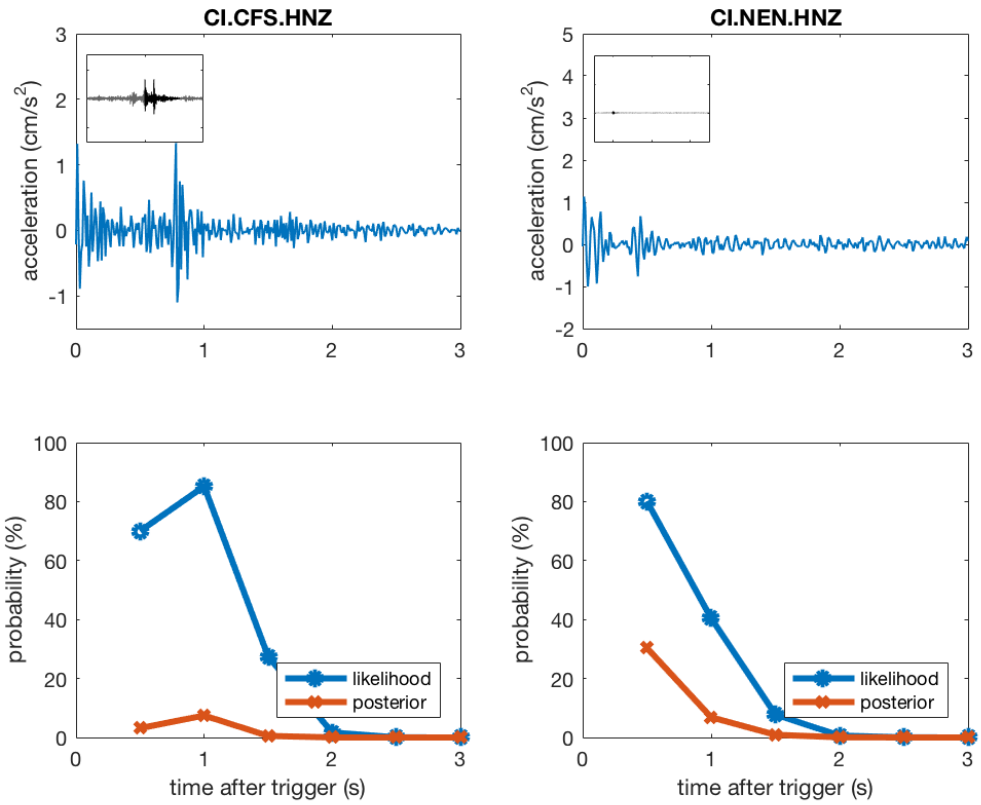



Figure 3.19 initial 3.0 sec vertical acceleration waveform and prediction results for stations CI.CFS and CI.NEN during ambient noise false triggers on 24 March 2015

### *29 July 2008 - M3.7 Chino Hills aftershock*

A M3.7 Chino Hills aftershock occurred on 29 July 2008, about 11 hours following the M5.5 mainshock. Immediately after the Chino Hills mainshock, the ETAS prior probability calculated at the stations located near the mainshock increased significantly due to the high-expected seismic activity during the aftershock sequence. Figure 3.20 indicates the location of the hypocenter and triggered stations. As shown in Figure 3.21, although the

amplitudes of P-wave reduced after the first trigger resulting in lower probability in the likelihood prediction, the posterior probability remains high with the assistance of the high prior seismicity probability. The PGA is recorded in the later phase after the arrival of S-wave.

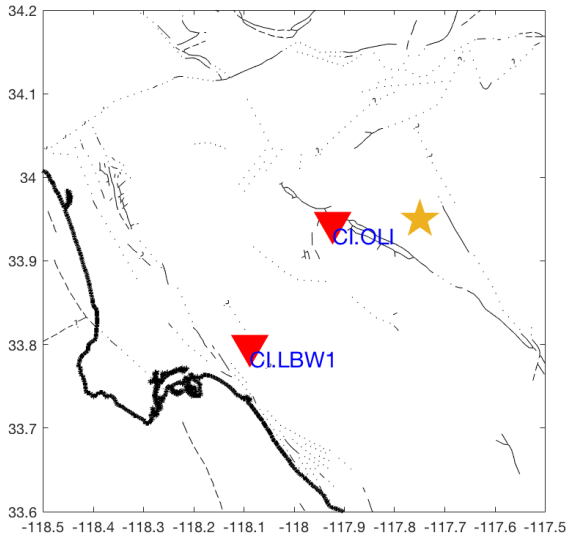



Figure 3.20 Map of M3.7 Chino Hill Aftershock: hypocenter in the yellow star and locations of CI.OLI and CI.LBW1 in red triangles

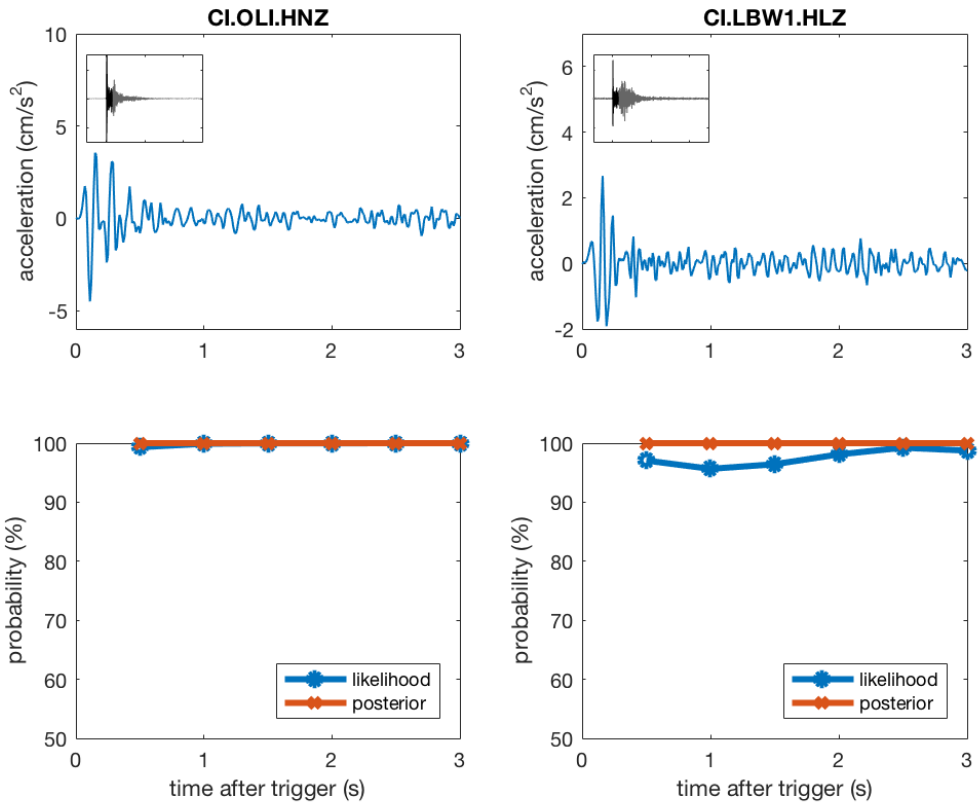



Figure 3.21 initial 3.0 sec vertical acceleration waveform and prediction results for stations CI.OLI and CI.LBW1 during M3.7 Chino Hill aftershock on 27 July 2008

### *18 May 2009 - M4.7 Los Angeles earthquake*

The M4.7 Los Angeles earthquake occurred on 18 May 2009. No recent seismicity was observed in the region prior to the event, so the prior probability of this event shows low probability of earthquake occurrence. Figure 3.22 indicates the location of the hypocenter and triggered stations. In Figure 3.23, the waveforms collected at the stations closest to the hypocenter, CI.WNS and CI.MIS, show strong on-going earthquake characteristics. Although the prior probability is low, the convincing indication of earthquake event from the likelihood probability dominates the posterior probability. This example demonstrates that if sufficient evidence from the arrival data shows earthquake characteristics, the effect of prior probability becomes negligible.

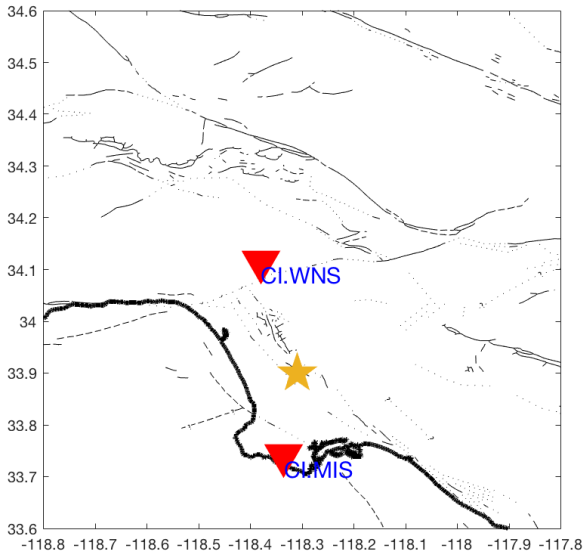



Figure 3.22 Map of M 4.7 Los Angeles earthquake: hypocenter in the yellow star and locations of CI.WNS and CI.MIS in red triangles

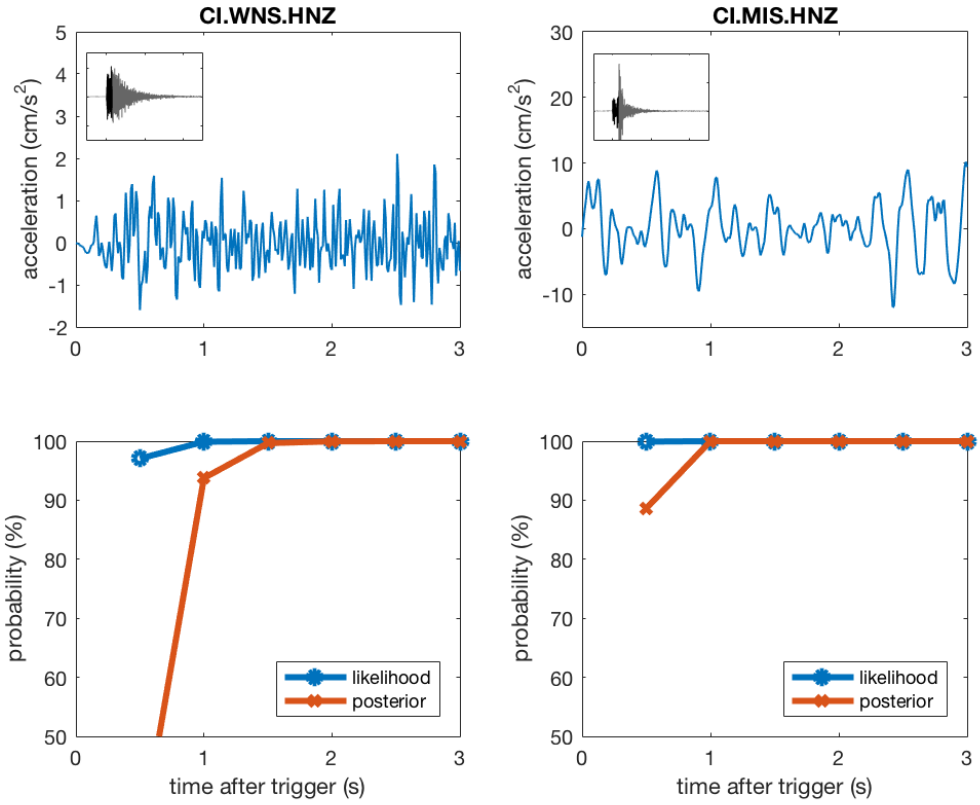



Figure 3.23 initial 3.0 sec vertical acceleration waveform and prediction results for stations CI.WNS and CI.MIS during M4.7 Los Angeles earthquake on 18 May 2009.

***30 May 2015 - M7.8 Japan teleseismic earthquake***

Teleseismic events are large earthquakes that occur at far distances, but are still observed by the local network. It is expected that the sensors will be sensitive (and triggered) by these signals, but the shaking does not affect the local region, and thus they should be classified as not EEW-relevant earthquake signals. The high frequency feature inputs in the proposed method aim to successfully discriminate teleseismic events from local earthquakes. As shown in Figure 3.24, CI.SMR and CI.SMW stations were triggered by the teleseismic event, but the amplitudes in acceleration records are small enough that they do not disturb the local community. The initial likelihood probability is relatively high due to the high uncertainty in the short data collected. Since no local seismicity was observed in the recent past, the ETAS prior probability was low, resulting in low posterior probability as well. The triggers from this teleseismic event would be ignored under the prediction of the proposed model.

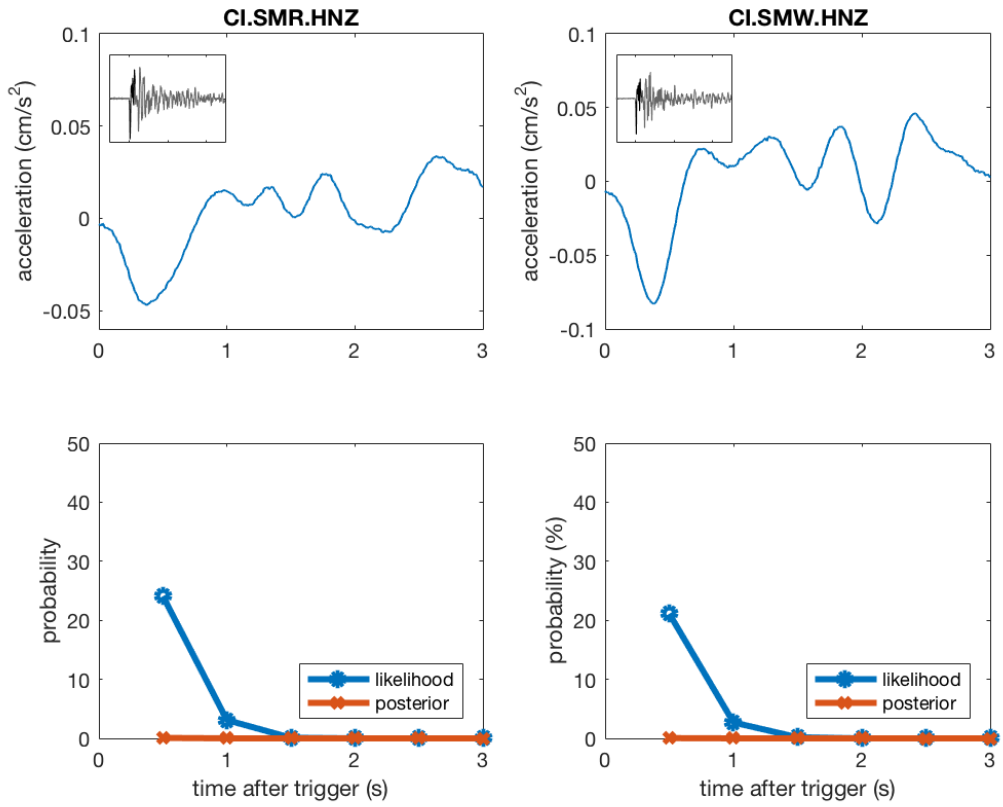



Figure 3.24 Initial 3.0 sec vertical acceleration waveform and prediction results for stations CI.SMR and CI.SMW during M7.8 Japan Teleseismic earthquake on 30 May 2015.

## 3.9 Discussion and Conclusion

We present a Bayesian probabilistic approach for single station rapid discrimination of EEW-relevant earthquake signals. The precision rate of earthquake classification at 0.5 sec after the P-wave achieves 98.5%. Waveform data from past earthquakes, noise, and teleseismic records are used in training the likelihood function. The prior information is constructed from the catalog information based on an ETAS seismicity forecast probability at the time of station trigger. Rapid earthquake detection becomes feasible due to the advancement in information technology, so that both waveform analysis and earthquake forecasting can be carried out in real-time.

The use of three component velocity and acceleration data extensively explores all the recorded information from the seismic sensors. Both accelerometer and broadband sensor data can be used with one-step integration or differentiation. The absolute amplitude is directly used to avoid computational errors accumulated through data processing.

The use of the seismicity Bayesian prior information differentiates the proposed method from other conventional earthquake discrimination methods that analyze waveforms only. It mimics human behavior in the decision-making process during an ongoing earthquake, especially the expectation that earthquake occurrence rises after a moderate earthquake. The forecasting process automatically applies to all the earthquake events in the catalog, with no need to label each individual event as foreshock, mainshock, aftershock, or aftershock of aftershock.

The practical benefits of the proposed method are revealed especially during aftershock and swarm sequences. During aftershock sequences, the repetitive ground shaking continuously deteriorates the already weakened infrastructure components. Seismic damage can be even more significant if the aftershocks occur close to a populated urban area, so the demand for

a fast EEW system becomes more significant during this time period to the rescue teams and residents. In addition, roughly 40% of earthquakes are preceded by smaller events known as foreshock-mainshock sequences (Abercrombie & Mori, 1996), and in such cases, the forecasting information can contribute to rapid identification of the large mainshock in our method. In all of the above cases, the proposed method can provide optimally fast alerts to users near the triggered station.

Unfortunately, it is recognized that the seismicity-based prior does not always give the 'correct' prediction, because not all mainshocks are preceded by foreshocks. If no previous seismic activity is recorded, the algorithm behaves similarly to the existing methods. As long as the catalog contains at least one prior observed seismic activity, the Bayes prior becomes important and influential when not enough waveform observations are available to fully constrain the uncertainty of the predicted result. When sufficient observations are collected, predictions can be fully determined by the waveform data. The proposed method does not replace the existing earthquake discrimination algorithms that clearly have important features arising from the use of longer data windows and/or network-based triggers. Moreover, the proposed method provides the potential benefit of faster, more reliable alerts for regions in the vicinity of the epicenter, where the severest shaking is experienced.

The simple model for the likelihood function guarantees rapid processing and alert delivery; the addition of Bayes' prior in seismicity forecast ensures the reliability of the posterior prediction. This method is the first algorithm that bridges the gap between earthquake forecast and earthquake early warning. The probabilistic approach allows users to customize the threshold based on the tolerance of uncertainties towards various applications. The straightforward implementation of the model suggests that it could be incorporated in real-time analysis.

*Chapter 4*

# ETAS Prior application two: Location Estimation

## 4.1 Introduction

Due to the rapid advancement of digital seismic networks, Earthquake Early Warning (EEW) currently uses real-time waveform information to rapidly estimate source parameters such as the magnitude and location of an earthquake (Heaton 1985) (Allen and Kanamori 2003). The source parameters are then used to predict ground motion characteristics at various sites. This waveform analysis approach for source parameter characterization typically requires collecting a few seconds of data after the first detection of P-waves. Since the arrival of S-waves follows immediately after the arrival of P-waves in the epicentral region, processing delays must be minimized if we are to have any hope of providing warnings of potentially damaging S-waves near an earthquake's epicenter.

(Cua 2005) and (Cua and Heaton, The Virtual Seismologist (VS) method: A Bayesian approach to earthquake early warning 2007) described a framework for deriving the probabilities of predicted ground shaking based on information that is available at any given instant for an EEW system. They suggested that the Bayesian probabilistic approach could be used to simulate the type of common-sense analysis that is performed by human experts if they had the time to intervene in EEW. As explained in Chapter 1, Bayesian probabilistic inference provides a natural framework to quantify uncertainties in combining

heterogeneous sources of information. In addition to waveform analysis in the conventional EEW algorithms, they suggested using the fact that many earthquakes occur close to the locations of past earthquakes that occurred in prior minutes to days. If a seismic station detects shaking, then it is natural to inquire whether that station is close to previous activity. In order to use this concept in a working EEW system, we must develop this simple idea into a practical and reliable algorithm that estimates the probabilities of potential earthquake locations, given the seismic data available from the network and the eathquake catalog records, for the time immediately preceding the detected shaking (Gerstenberger, et al. 2005).

We begin with Bayes' theorem, which states that the probability magnitude $M$ and station-epicenter-distances $R$, given seismic data $S(t)$ available at time $t$ is formulated:

$$P(M,R|S(t)) \propto P(S(t)|M,R)P(M,R) \qquad [4.1]$$

$P(M,R|S(t))$ is usually referred as the Bayesian posterior function, $P(S(t)|M,R)$ is called the likelihood function and $P(M,R)$ is called the Bayesian prior. Most existing EEW algorithms focus on the likelihood function; that is, what magnitude and station-epicenter-distances produce seismic data similar to the data just recorded? For the purposes of this study, we assume that the likelihood function is provided by the Gutenberg Algorithm (Meier, Heaton and Clinton 2015). The Gutenberg Algorithm (GbA) uses a filterbank to decompose a real-time waveform into different frequency bands, and then efficiently searches an extensive waveform database to identify past records with similar time-frequency characteristics and uses them to compute probabilistic source parameter estimates for the real-time waveform. The GbA is designed to be optimally fast, first estimates are available using only 0.5 s of data from the closest station. However, earliest estimations could be poorly resolved due to trade-offs between earthquake magnitude and epicenter distance. Introducing prior information to assign relative probability based on

established empirical relationships can reduce this issue. There are several choices for Bayesian prior: perhaps the most general choice is to assume that seismic activity obeys the Gutenberg-Ritcher frequency-magnitude law (GR law), in which case the prior probability can be approximated by $prob(M,R) = 10^{a-bM}R$, where $a$ and $b$ are GR law parameters and $R$ is the scalar distance amplitude. This proposed prior information simply states that the earthquake frequency decreases exponentially with magnitude and the area of a ring at distance $R$ grows linearly with $R$.

In this chapter, we focus on earthquake location parameter estimation. We propose a Bayesian prior that is based on the principle that earthquake sequences tend to cluster in time and space. In particular, we use Epidemic-Type Aftershock Sequence (ETAS) Models to derive a location distribution with far more spatial-temporal information than the simple prior mentioned previously (Y. Ogata 1998). We show how an ETAS prior can provide more accurate source-parameter estimates at short warning times with minimal seismic waveform data. In fact, the use of an ETAS seismicity forecast model as a Bayesian prior often provides accurate estimates of the epicenter location that are available simultaneously with the detection of an event. Even in cases with a highly uncertain ETAS prior, the accurate location estimation is converged immediately with sufficient data processed by GbA.

We collected all 504 M4+ earthquakes in southern California between 1990 an 2015. We evaluated the location estimation performance of Bayesian analysis techniques making use of the GbA as the likelihood function combined with an ETAS model for the Bayesian prior. We also present the earliest estimations for examples of a M5.2 Lone Pine Earthquake and a M5.4 Chino Hills Earthquake in detail. The two examples demonstrate the importance of Bayesian prior and likelihood interaction to ensure the optimum results in different seismic environments.

## 4.2 Method

### *4.2.1 Bayesian Inference in EEW Location Estimation*

To focus on location estimation only, Bayes' Theorem from Eq [4.1] can be simplified to:

$$P(Lat, Lon|S(t)) \propto P(S(t)|Lat, Lon)P(Lat, Lon) \quad [4.2]$$

where $(Lat, Lon)$ are the epicenter location of the earthquake. In this formulation, $P(Lat, Lon|S(t))$ is the Bayesian posterior function, $P(S(t)|Lat, Lon)$ is the likelihood function and $P(Lat, Lon)$ is the Bayesian prior. The following describes how each of the terms is derived.

### *4.2.2 Prior Information – ETAS seismicity model*

The Bayesian prior information is a spatial distribution of earthquake probability produced by ETAS models, where each of the observed earthquakes stochastically generates potential earthquakes in the future. The ETAS model is simply the results calculated in Chapter 2.3. The forecast map for the ETAS model demonstrates the relative probability of expected earthquake occurrences of a region, which provides guidance on where is more likely to have an earthquake nucleation. Thus, normalization is not necessary in the process, as it vanishes in the proportionality property in Eq[4.2].

### *4.2.3 Likelihood Function– The Gutenberg Algorithm*

The GbA is a probabilistic approach to estimate EEW source parameters using real-time waveform information. During an ongoing earthquake, GbA performs real-time time-frequency analysis on the collected waveform using minimum-phase-frequency filter banks, then efficiently search within a catalog of events for similar characteristics. At every time increment after the P-wave arrival, each triggered station computes the relative

probability of magnitude and epicenter distance estimations. The focus is to explore maximum available information during the EEW process. For details of the algorithm, we refer to (Meier, Heaton and Clinton 2015).

In single station location inference, we convert the station-to-source distance probability density function from GbA onto a 2-dimensional spatial distribution that is most likely to produce the recorded waveform $S(t)$, $P^j(S(t)|Lat, Lon)$, according to the distance between the station $j$ to the location $(Lat, Lon)$. Combining location estimation from multiple $K$ stations is straightforward, as the probabilistic formulation of the algorithm simply requires that we multiply the single-station spatial probability functions:

$$P(S(t)|Lat, Lon) = \prod_{j=1}^{K} P^j(S(t)|Lat, Lon) \quad [4.3]$$

Eq [4.3] describes the likelihood function as suggested in Eq [4.2].

## 4.3 Data

We collected all 506 M4.0+ earthquakes in Southern California from 1990 to 2015; from which the catalog locations of the events are then compared to the estimated location parameters at every half-second interval after the event detection. The locations of all the events are shown in Figure 4.1. The data set includes 03 October, 2009 M5.2 Lone Pine Earthquake, located at (-117.86, 36.39); and 29 August, 2008 M5.4 Chino Hills Earthquake, located at (-117.76, 33.95). The details of the two events are provided in Table 4.1. The two events represent two characteristic setting: the Lone Pine earthquake demonstrates the accuracy and speed estimations due to Bayesian prior information during a seismic sequence, while the Chino Hill earthquake demonstrates the importance of Bayesian likelihood function to reduce prior uncertainty during a seismic dominant period.

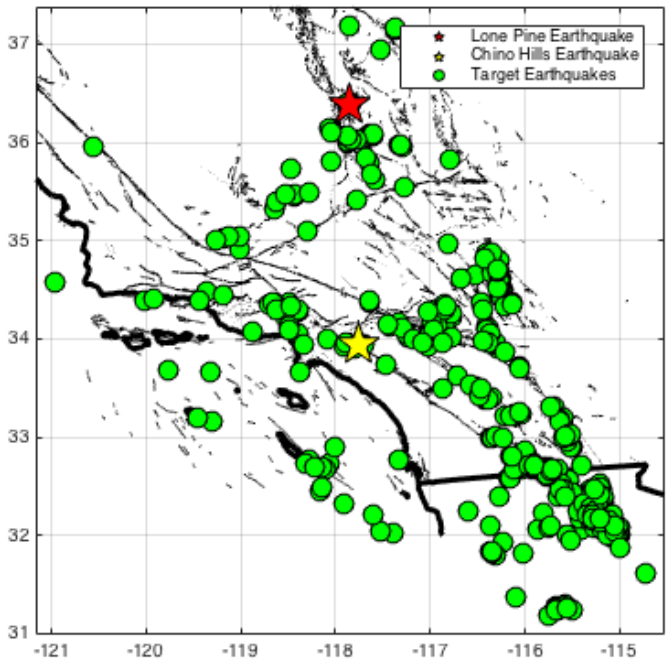



Figure 4.1 Catalog location of the 506 target M4.0+ earthquakes in Southern California from 1990 to 2015. Including 2009 Lone Pine M5.2 Earthquake in red star and 2008 Chino Hills M5.4 Earthquake in yellow star.

| Name | Time | Magnitude | Latitude | Longitude |
|---|---|---|---|---|
| Lone Pine Earthquake | 2009/10/03 01:16:00 | 5.2 | -117.86 | 36.39 |
| Chino Hills Earthquake | 2008/08/29 18:42:15 | 5.4 | -117.76 | 33.95 |

Table 4.1 The detailed Information on Lone Pine Earthquake and Chino Hills Earthquake

The catalog data used in Bayesian prior information was downloaded from the Southern California Earthquake Data Center (http://data.scec.org/). It includes source parameter information, such as origin time, hypocenter location, and magnitude of 567258 historic seismic events in Southern California from 1981 to 2015.

The waveform data set is collected from the global database of waveforms compiled in Meier (2015). All the waveforms have been preprocessed to eliminate poor quality records, including missing channel records, low signal-to-noise ratios, low sampling rates, and clipped records. The subset of events used in this study includes 50750 three-component waveform records from a total of 3523 events.

## 4.4 Results

### *4.4.1 M5.2 Lone Pine Earthquake*

At 01:16:00 on October 03, 2009, the M5.2 Lone Pine earthquake first triggered the vertical channel of station CI.CGO 3.5 seconds after the origin time. The station is about 20km northeast of the catalog event location. The location estimation in GbA resulted in a high initial uncertainty which decreased with time; the location error reaches 18km at 14 seconds after the origin time, as shown in Figure 4.3 a) b) and c).

At the moment of the first P-wave arrival at CI.CGO, the ETAS seismicity forecast map developed an earthquake probability with peak distribution at 15km southwest of the triggered station, shown in Figure 3.2. This was a consequence of the foreshock series recently accumulated in the area. By including the ETAS seismicity map as a Bayesian prior, the maximum posterior location estimation immediately converged to the location error of 2km, as shown in Figure 4.3 b) d) and f).

At every time increment during the ongoing event, the location error estimated from GbA fluctuates around 10 to 20km; with the ETAS seismicity map included as the Bayesian prior, the location error reduced to less than 3km, as shown in Figure 4.4.

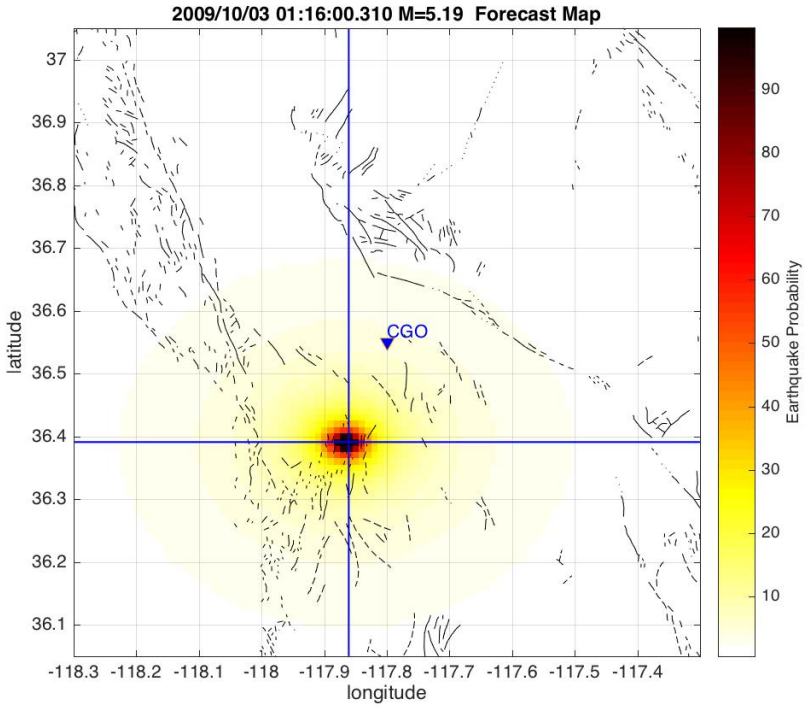
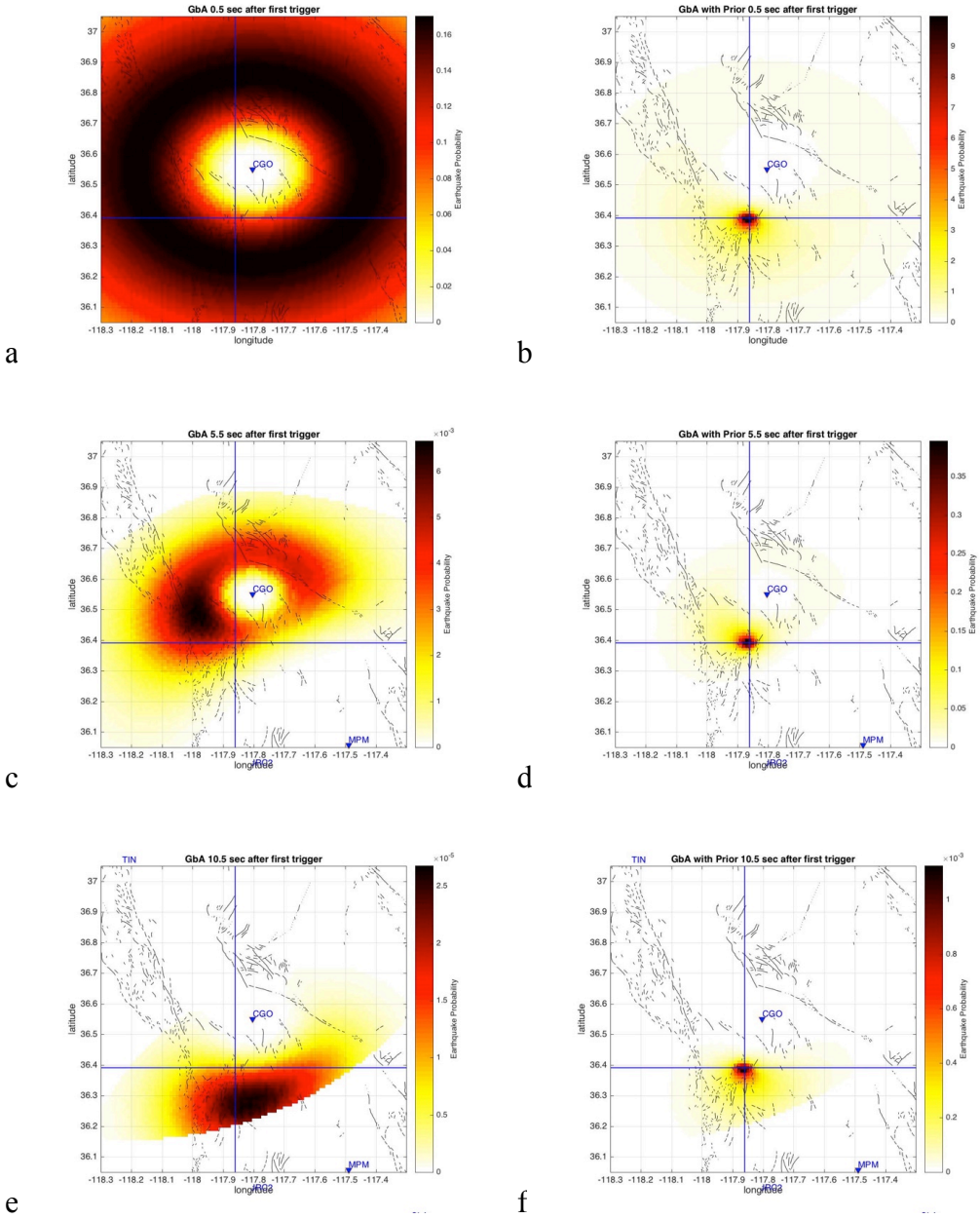



Figure 4.2 Seismicity Forecast Map for Lone Pine M 5.2 Earthquake. It was produced immediately after the first station trigger at CI.CGO. The intersection of the two blue lines is the catalog location




a b c d e f

Figure 4.3 Probabilistic location estimation map of the M5.2 Lone Pine Earthquake at various times after the first station trigger. a) c) and e) are results of Gutenberg Algorithm at 0.5 sec, 5.5 sec, and 10.5 sec after the first trigger, respectively. b) d) and f) are

posterior results of Gutenberg Algorithm with Prior at 0.5 sec, 5.5 sec, and 10.5 sec after the first trigger, respectively. The intersection of the two blue lines is the catalog location.

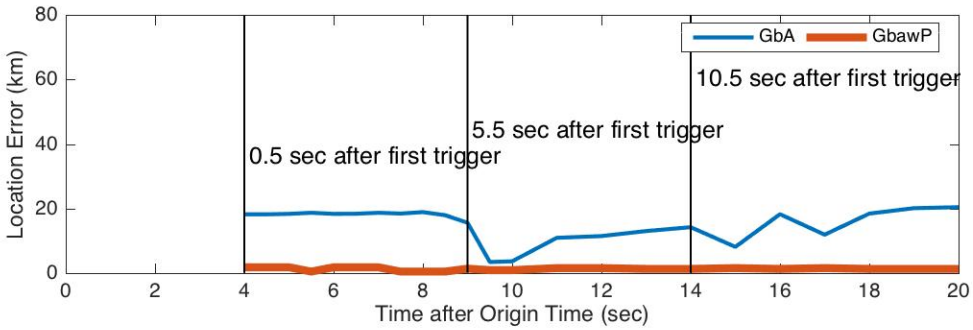



Figure 4.4 M5.2 Lone Pine Earthquake location error as a function of time after the origin time. The blue and red lines are the location error results of the Gutenberg Algorithm, and the Gutenberg Algorithm with ETAS Prior, respectively.

### *4.4.2 M5.4 Chino Hill Earthquake*

At August 29, 2008 18:42:15, the M5.4 Chino Hills earthquake first triggered the vertical channel of station CI.CHN 3 seconds after the origin time. The station is 5 km northeast of the catalog event location. Due to high station density in the area, sufficient waveform data was quickly collected, and GbA initial location estimation shows high accuracy with only 12km error, as shown in Figure 4.6 a) c) and e).

Because there was no observed seismic cluster in the region, the background seismicity greatly influenced the ETAS forecast, shown in Figure 4.5. Although the seismicity prior indicates relatively high earthquake probabilities around 40km east of the station, the hypothesis was immediately updated by the GbA results with the incoming waveforms, shown in Figure 4.6b) d) and f). At 1.0 sec after the first trigger, with 5 triggered stations data, the spatial distribution shows a clear shift from the ETAS to the GbA results. And a half second after that, with 9 stations triggered, the posterior probability is dominated by the GbA results.

Figure 4.7 shows that although the initial location error with the Bayesian prior is 42 km, it quickly reduces to 8 km within the following 1.5 sec, and the error remains low thereafter. At every time step, the posterior results update with the available waveform data.

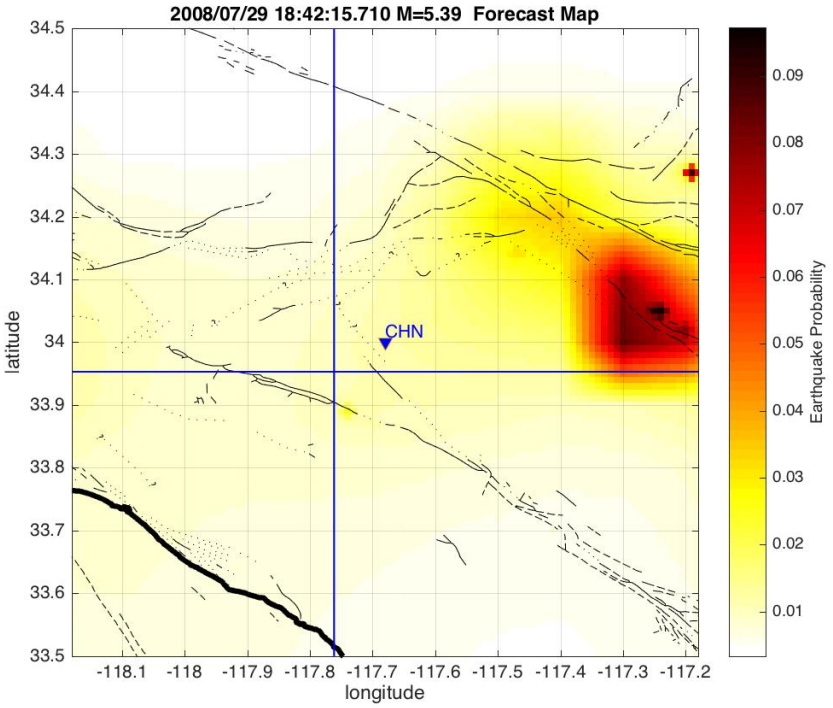



Figure 4.5 Seismicity Forecast Map for Chino Hills M 5.4 Earthquake. It was produced immediately after the first station trigger at CI.CHN. The intersection of the two blue lines is the catalog location

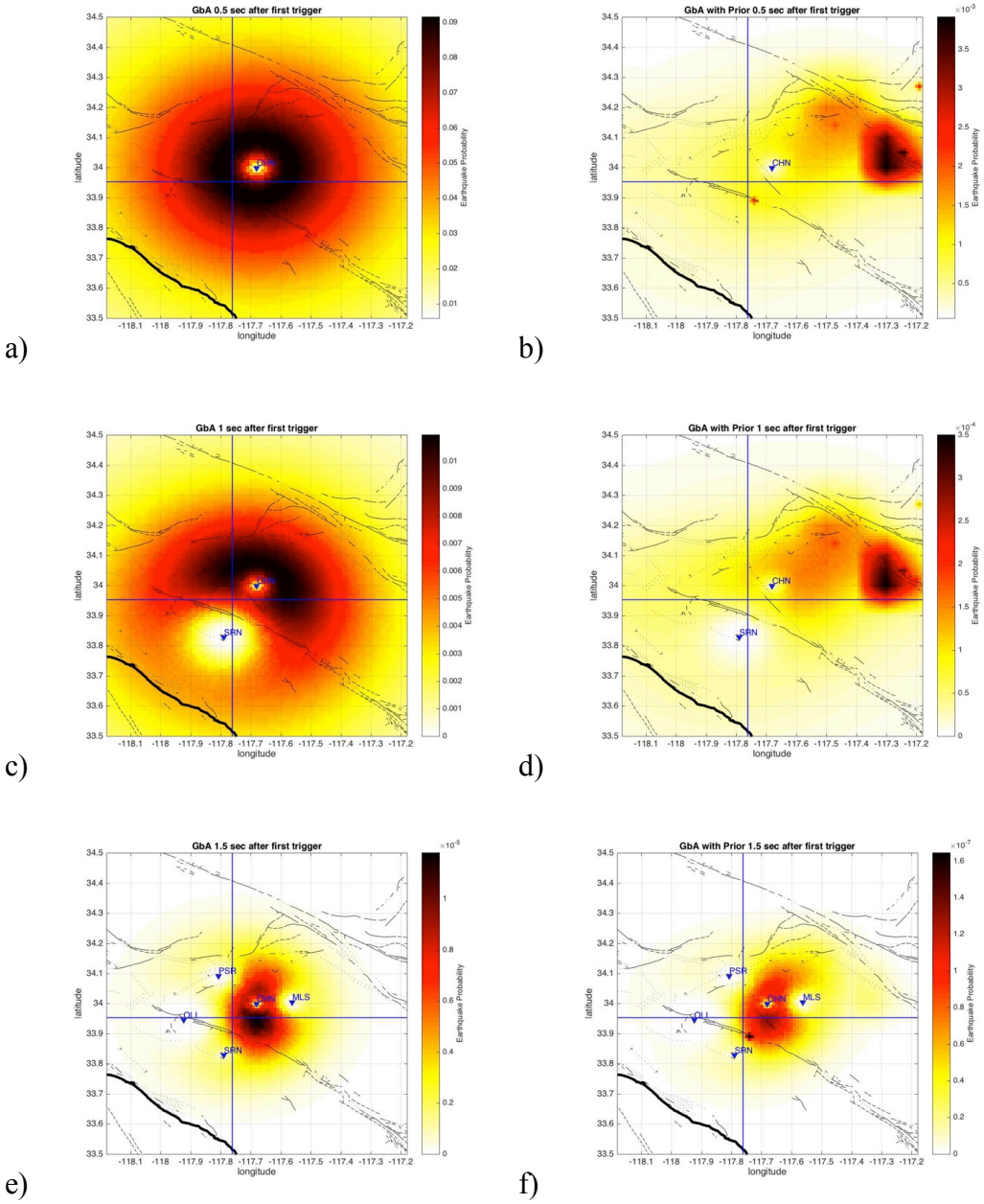



Figure 4.6 Probabilistic location estimation map of the M5.4 Chino Hills Earthquake at various times after the first station trigger. a) c) and e) are likelihood probabilities, results of GA at 0.5 sec, 1.0 sec, and 1.5 sec after the first trigger, respectively. b)

d) and f) are posterior probabilities, results of the GbA with Prior at 0.5 sec, 1.0 sec, and 1.5 sec after the first trigger, respectively. The intersection of the two blue lines is the catalog location.

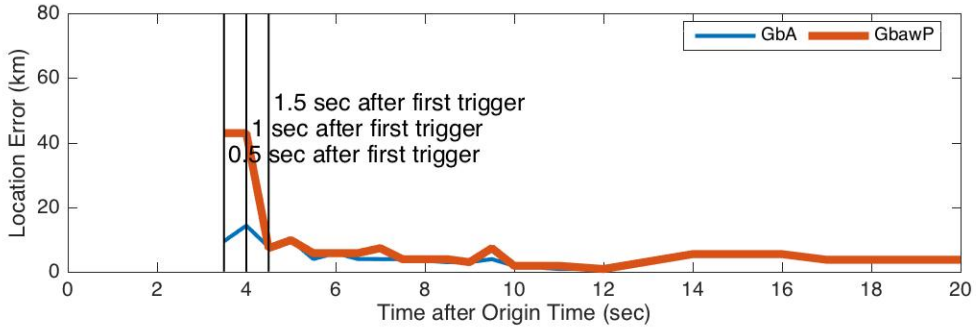
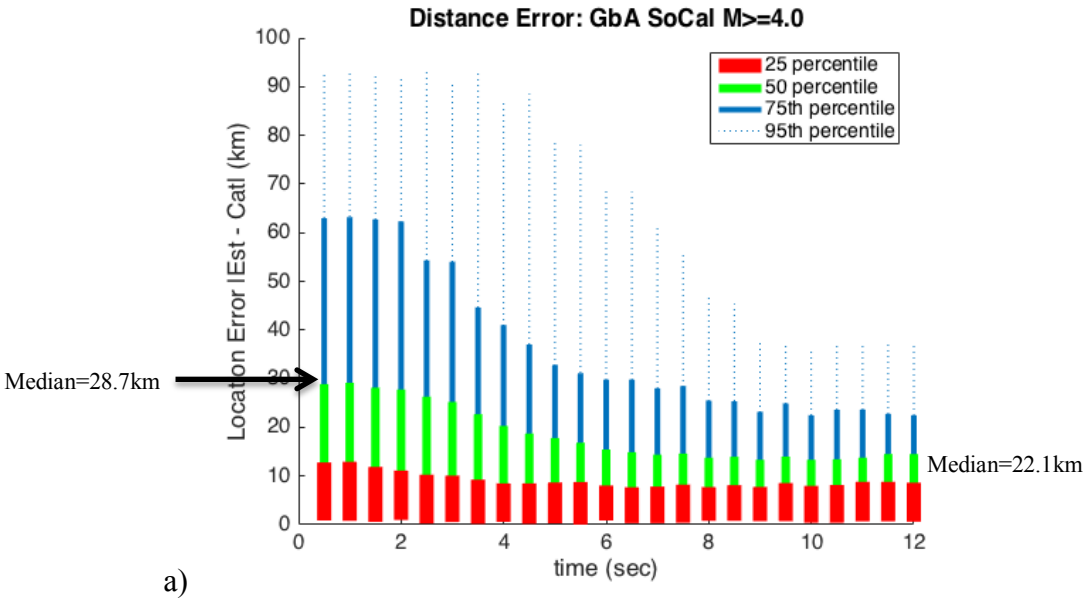



Figure 4.7 M5.4 Chino Hills Earthquake location error as a function of time after the origin time. The blue and red lines are the location error results of the GbA and GbA with ETAS Prior, respectively.

### *3.4.3 Overall Performance*

We evaluated the location error, the distance between the catalog location and the location of maximum probability, as a function of time after the first trigger for all 506 M4+ earthquakes in Southern California from 1990 to 2015, as shown in Figure 3.8. In Figure 3.8 a), only the waveform information is considered, the GbA median location error is 28.7 km, 20 km and 17 km at 0.5 sec, 5 sec, and 10 sec after the first P-wave detection. As expected, the error is initially large and reduced with time when more waveform data is collected. In Figure 4.8b), the location error is calculated using Bayesian Inference by combining waveform and catalog information; the median location error is 12 km, 8 km, and 5 km at 0.5 sec, 5 sec, and 10 sec after the first trigger. The median error reduction has

improved substantially, especially in the first few seconds, reaching a 58% improvement. The distance error at all percentile levels consistently decreased at every time increment.




a)

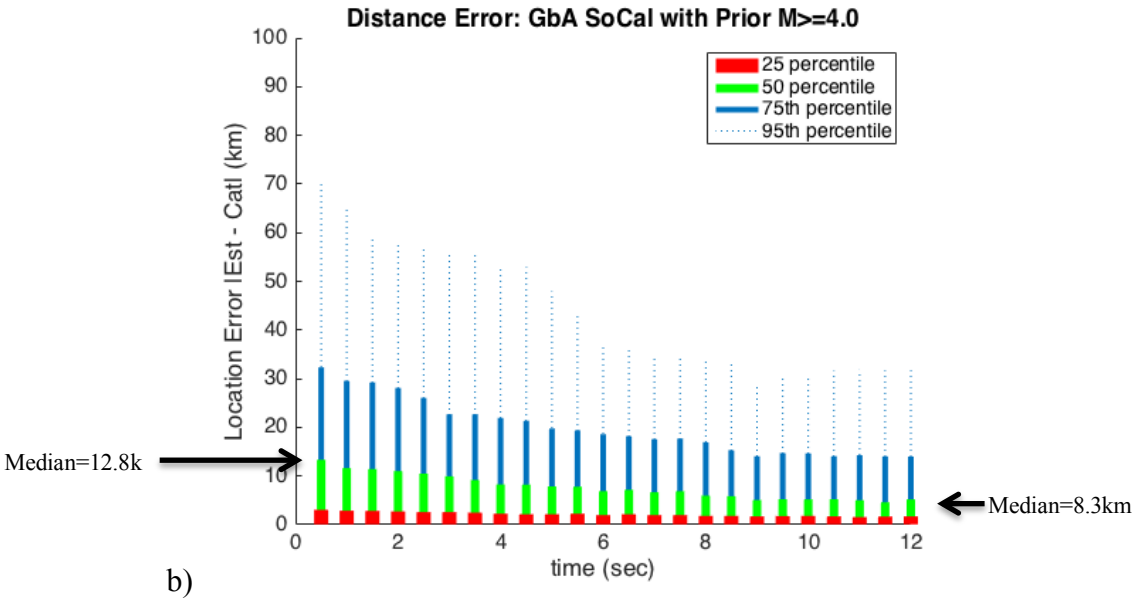



Figure 4.8 Location Error as a function of time after first trigger for 506 M4+ earthquakes in Southern California 1990-2015 a) likelihood performance: GA results b) posterior performance: GA with Prior results. The errors are specified at the 25$^{th}$, 50$^{th}$, 75$^{th}$, and 95$^{th}$ percentile

## 4.5 Discussion

Many of the previous studies have shown that earthquakes tend to occur in places where there has been observed seismic activities; especially so with incidence of foreshocks often observed preceding large earthquakes (P. Reasenberg 1999). While most of the current EEW algorithms focus on time-series analysis of waveforms, they fail to acknowledge that previous events are clearly related to subsequent larger earthquakes. ETAS Bayesian priors exploit the seismicity catalog information to provide optimally fast location approximation, logically using the spatial-temporal clustering property of earthquakes to ensure a higher the level of accuracy.

More significant positive impacts of ETAS Bayesian prior are reflected during aftershock and swarm earthquake sequences. During aftershock sequences, the repetitive ground shaking continuously deteriorates the already weakened infrastructure components. Seismic damage can be even more significant if the aftershocks occur close to a populated urban area. In these cases, the location estimations that use the ETAS Bayesian prior guarantee the delivery of fast and accurate alerts immediately after the first P-wave detection, allowing EEW to offer more alerting time to rescue teams and residences for evacuation during aftershocks.

For the events with no obvious prior seismic activity in the proximity, ETAS produces a smooth spatial distribution of earthquake probability. In such case, the estimations are quickly dominated by the likelihood function of the waveform analysis, as demonstrated in the Chino Hills earthquake. The Bayesian probabilistic approach in this study mimics human behavior in the decision making process during an ongoing earthquake. It first uses scientific intuition of the seismic knowledge to make a quick and rational approximation, and then analyzes real-time waveforms with the assistance of powerful computational tools. The Bayesian framework conveniently combines results from any independent probabilistic algorithm, such as the GbA with ETAS prior. For future development, additional algorithms can be incorporated in this ensemble framework to further enhance the posterior results.

Station topology constraints can be imposed as an additional Bayesian prior in an ideal network with no malfunctioning stations. The concept of the voronoi diagram of a network distribution implies that an earthquake must occur within the voronoi cell of the first triggered station, as the P-wave travel time from any of the points in this voronoi cell to the station is minimized (Rosenberger 2009). However, in our offline study, due to missing records and inconsistency station performance, station topology concept led to poor results.

## 4.6 Conclusion

We proposed a probabilistic approach to obtain faster and more accurate EEW location estimations. We investigated EEW performance combining the GbA and ETAS seismicity models under Bayesian inference. Our results show that Bayesian inference with seismicity priors can reduce overall median location error by 58% for the first few seconds after P-wave arrival at the closest station to the epicenter. In most of the cases evaluated, accurate location estimation is available immediately after the first P-wave detection.

In the current technology, scientists are investigating sophisticated methods to exploit waveform information to estimate source parameters, while neglecting the most fundamental clustering pattern of seismic sequences. In this study, we demonstrated that both the Bayesian seismic prior and waveform likelihood are essential in EEW; only by analyzing various heterogeneous information, could the location estimation in EEW potentially achieve fast results with high confidence.

*Chapter 5*

# Reducing EEW parameter search delays

## 5.1 Introduction

EEW provides useful alerts during earthquakes causing a significant level of ground shaking, so the alert speed is critical to provide a warning to the most strongly affected areas close to the epicenter. Additionally, for high-cost user actions (such as halting industrial processes), the accuracy of ground motion predictions at user sites is important for the widespread adoption and use of EEW (Hoshiba, 2013). In general, the conventional algorithms use trained models to estimate earthquake source parameters (such as magnitude and hypocenter distance) from station ground motion observations, and then apply ground motion prediction equations to estimate the peak ground motion experienced at different user sites (Wu et al., 2007) (Zuccolo et al., 2016) (Kuyuk et al., 2014). The predictive models tend to compress the observed information into a few source parameters, which can overly simplify the behavior of wave propagation through the Earth. Significant error in final prediction results can be accumulated through the uncertainties in the underlying models (Bose et al., 2009) (Allen et al., 2009). As a result, for the purposes of a real-time EEW system, it is a challenge to create a simple model that fully captures all the attributes that influence the peak ground motion in a recorded waveform, such as magnitude, location, depth, soil type, local site condition, directivity, and source radiation.

Fingerprint searching and template match methods are alternative approaches to EEW and have also recently been employed in other areas of seismology (Yoon et al., 2015). In the fingerprint searching method, important waveform characteristics are extracted from each earthquake record to form an extensive database of "earthquake fingerprints". During the

occurrence of an on-going earthquake, the algorithm searches among the database for the most similar "earthquake fingerprints", and then estimates the source parameters or peak ground motions of the new event based on the searched records. A recently developed method, called the Gutenberg Algorithm (GbA) (Meier et al., 2015), applies the fingerprint-searching concept to EEW by abstracting the time-frequency amplitude information of the real-time seismic signal for various filter bands to create a large-scale database, and then estimates the earthquake source parameters such as magnitude and hypocenter distance for on-going earthquakes. In addition, the template-matching method in FinDer (Böse et al., 2012) compares observations with a database of theoretical spatial ground motion patterns to estimate earthquake source parameters and peak ground shaking at various sites. Both methods share the common approach of searching among a pre-processed database.

One of the most important factors required of search algorithms is that the searched database needs to be sufficiently large in order to cover a wide range of potential earthquakes. In other words, if similar data to the target query are not included in the database, the searched result could be significantly off from the true value. As an example, the records in the databases should represent the natural distribution of earthquake occurrence as described by the Gutenberg-Richter relationship; there should be many more small events than large ones because small size earthquakes occur more often than large earthquakes, so the search returns should reflect real earthquake likelihoods. Of course, the best strategy is to include all worldwide earthquakes recorded over a long period of time. While increasing the database promises to improve estimation accuracy, the trade-off is that the processing time of searching among a large database increases significantly due to the rise in comparison operations. A simple search of the ANSS catalog (http://www.quake.geo.berkeley.edu/anss/) reveals that 2090 shallow crustal earthquakes (depth <30km) over magnitude 2 occurred in California during 2015. Similar results are

also indicated with searches on USGS/ComCat, Southern California Earthquake Data Center and other similar earthquake databases. If one wants to include all records from the network over years for all the earthquake events worldwide, the size of the database scales exponentially (Yu, 2016). As a result, the processing delay of the real-time search will significantly increase because the time required to query databases sequentially is proportional to the size of the database. While advances have been made in the development of such algorithms in EEW, very little attention has been paid to optimizing the processing time of large databases.

Database searching is often an application of the Nearest Neighbor (NN) search problem with the Euclidean metric. The problem is commonly encountered in many computational techniques such as event detection, pattern recognition, and data analysis (Bhatia, 2010). In general, we seek for a point in the database that minimizes the Euclidean distance to the target point (sometimes referred as the least square distance). The problem states that for the target point $x = [x_{(1)}, \dots, x_{(d)}]$ and the *i*th training point in the database $y_i = [y_{i(1)}, \dots, y_{i(d)}]$, we define the distance between $x$ and $y_i$ to be

$$d(x, y_i) = \left( \sum_{k=1}^{d} \left( x_{(k)} - y_{i(k)} \right)^2 \right)^{1/2} \qquad [5.1]$$

NN searches for the $\hat{y}$ with the closest distance to the target point, mathematically represented as $\hat{y} = argmin_{y_i}(d(x, y_i))$. In most cases, the *k*-Nearest-Neighbor (k-NN) search method is applied by finding the $k$ closest training points to the target point; this method provides a more robust estimation that avoids outliers in the database. The corresponding parameters associated with the $\hat{y}$ are used to classify or estimate the parameters of interest for the target point.

In this study, we use a data structure, multidimensional binary search tree (KD tree) NN searching concept, to organize the seismic data, and evaluate the reduction of NN searching time for large datasets. KD-tree is a binary tree data structure that links the relative position of all the data points, so data with similar patterns cluster, thereby allowing the search procedure to become faster (Bentley J. L., 1975). Although it requires initial effort to construct the tree data structure, the searching process is quick. The goal is to introduce the concept of data structures in EEW to minimize the processing time for waveform record searching without loss of accuracy, and thereby earthquake alerts can be delivered to the sites of interest much earlier. The effectiveness of fast alerts is especially valuable in the proximity of the epicenter where the strongest damage occurs very quickly after event onset. In this study, we describe a searching procedure that uses the KD tree NN search method that identify the EEW fingerprints characterized by the Gutenberg Algorithm. We 1) evaluate the influence of database size on the prediction accuracy of the earthquake source parameters (magnitude and hypocenter distance) and peak ground motion parameters (PGA, PGV, PGD), 2) estimate the processing efficiency of the KD tree searching for databases with different sizes and extrapolate the future performance by scaling to larger data sets. The KD tree is a well-established NN searching algorithm that has been implemented in a wide range of engineering and database applications (Bentley J. , 1979). Only by overcoming the computational challenges in the processing time can EEW start to adopt the databases for real-time seismology applications, and the fingerprint searching algorithms with big data reveal their full practical potential.

## 5.2 Data

Theoretical analysis of the KD tree searching shows the performance complexity being O(log N) verses O(N) for the linear sequential search, where N is the number of data points in the database (Friedman et al., 1977). Although the theoretical average search time of KD

tree is much shorter than the linear sequential search, the performance varies depending on the distribution of the data. Our goal is to determine the searching efficiency of the KD tree method for our GbA seismic database. We ran a series of offline tests on the earthquake filterbank database to mimic potential performance of EEW using true seismic records. The dataset used is pre-processed by (Meier et al., 2015) for the GbA. The database consists of 182,805 near-site records with 9 feature dimensions in each record. Each of the feature dimensions represents the peak ground velocity in octave-wide frequency bands for a given ground motion record with a fixed time window. The frequency bands used in GbA features are shown in Table 1. GbA creates such a dataset table for every half-second increment in time after the P-wave arrival. In general, EEW tends to consider at least 3 to 4 sec data after the P-wave arrival for the trade-off of accuracy and time delay. For the purpose of this investigation we selected the database for a 10-sec time window because the predictions are stabilize with more data collection. The collected earthquakes cover a large range of magnitude, spanning from M 2.0 to M 8.0, compiled from shallow crustal earthquakes collected from Japan, Southern California, and Next Generation Attenuation-West 1 (Chiou and Youngs, 2008).

| Feature Dimension No. | Frequency Band (Hz) |
|---|---|
| 1 | 0.09375 – 0.1875 |
| 2 | 0.1875 – 0.375 |
| 3 | 0.375 – 0.75 |
| 4 | 0.75 – 1.5 |
| 5 | 1.5 – 3 |
| 6 | 3 – 6 |
| 7 | 6 - 12 |
| 8 | 12 - 24 |
| 9 | 24 - 48 |

Table 5.1 Frequency bands for feature input in Gutenberg Algorithm. The GbA database consists of 9 feature dimensions. Each feature takes the observed peak ground velocity in the given frequency band.

## 5.3 KD Tree and Method

### *5.3.1 KD Tree*

KD tree is a binary tree structure that stores the finite set of database points with k-dimensional feature space. In our case, we have 9 variables corresponding to 9-dimensions. The method involves two steps. First, we construct the tree to organize the information in the database. Then, the NN algorithm is applied on the KD tree to search to the most similar point to the target record during an on-going earthquake. In KD tree implementation, a point in the database is also called a node in the tree.

- Construction of KD-tree

The construction of the KD-tree is a recursive process. Starting with the root of the tree, the first feature dimension (frequency band: 0.09375 – 0.1875Hz) is chosen as the splitting hyperplane. All nodes are ordered with respect to the value in this feature dimension, and the node with the median value is inserted into the root of the tree. All nodes with coordinates less than the median in the splitting hyperplane create the left subtree, and the nodes with coordinates larger than the median in the splitting hyperplane create the right subtree. All the feature dimensions rotate in becoming the splitting hyperplane to create the next level of subtrees.

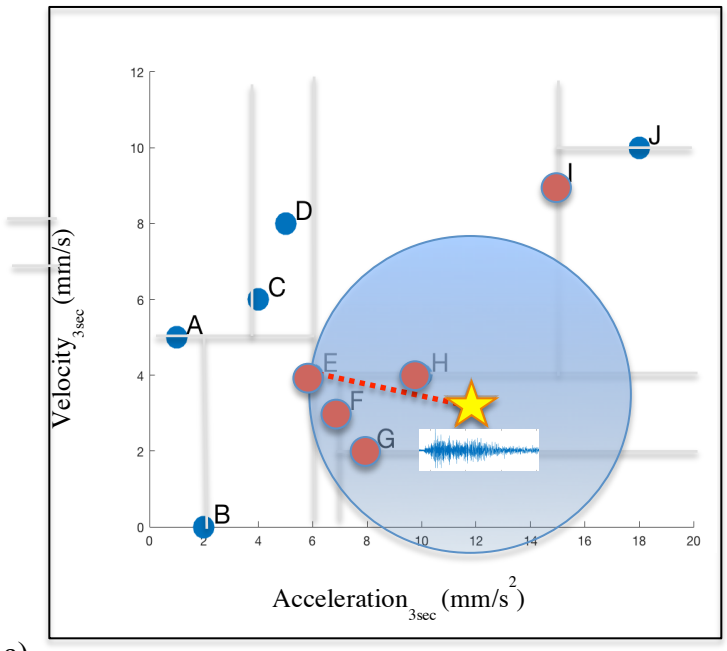
- Nearest Neighbor Search in KD-tree

Starting with the root node of the tree, the nearest distance is initialized to be the distance between the target node to the root. Then recursively move down to the next level in the tree, and checks if the splitting hyperplane intersects with the hypersphere centered at the target record with a radius of the current nearest distance. If the node falls outside of the hypersphere created by the current nearest node (indicating the point is further to the target node than the current nearest node), then this node and any extended child nodes further away can be eliminated from the investigation. The process is repeated, recursively moving down to the next level in the tree until reaching the leaves of the tree. The searching time is reduced since large subsets of the database are not visited. Therefore, the average searching time in a KD tree is significantly lower, especially when the size of the database is large.

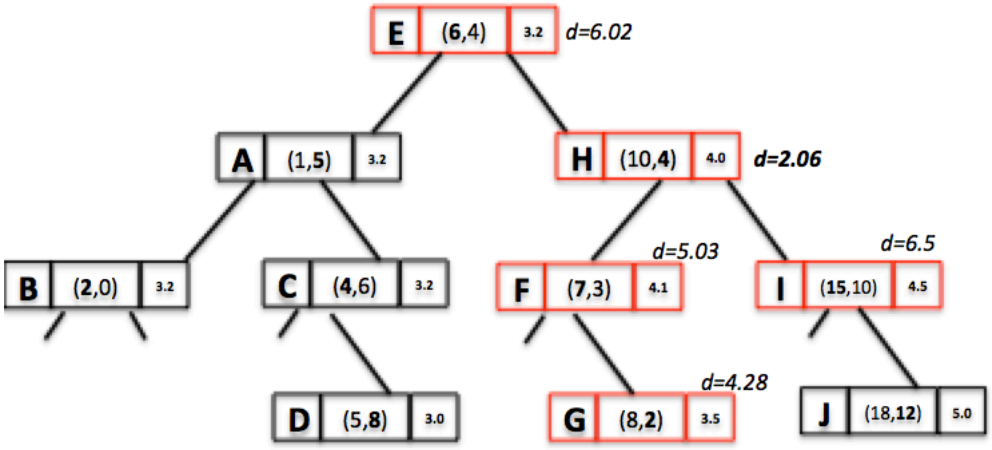
To better visualize the concept, Figure 5.1 demonstrates a KD tree structure for a 2-dimensinal featured database with 10 earthquake records described by the velocity and acceleration at initial 3 sec after triggering a station. The goal is to predict magnitude of the new event based on the velocity and acceleration recorded at the first 3 sec of the p-wave. We start the search process of the nearest neighbor of the target data (the yellow star) with node E, which is the root of the tree. The radius of the initial hypersphere is set between the target data and node E. In a 2D feature space, the hypersphere is simply a circle. Since the left branch (link between node A and E) does not cross the hypersphere, indicating all the nodes in the left subtree (node C, A, B, D) can be eliminated from the search because their Euclidean distance to the target point is clearly further than node E. This eliminates the computational effort of going through almost half of the database at the first step. Since the target node is closest to node H, the magnitude associated with node H (M=4.0) is the prediction result for the target node.

The algorithm can be easily extended to $k$ nearest neighbor (k-NN) search to find $k$ most similar points to the target point in order to give a more probabilistic estimate of target parameters. It requires two modifications. First, we need to keep track of all the current

nearest points in an ordered queue with length k; if the queue contains fewer than $k$ points, the subtrees on both sides need to be visited. Second, instead of comparing the splitting hyperplane with the hypersphere of the nearest point, we should check if the hyperplane intersects with the hypersphere of the last nearest point in the queue. If they intersect, the new node is inserted into the queue of $k$-nearest neighbors to the target point. At the end of the search, the algorithm returns $k$ points from the database that are located with minimum distances to the target point.




a)




b)

Figure 5.1 A 2-dimensional KD tree example: a) visual distribution of the database in feature dimensions, b) tree structure of the database. A database of 10 earthquake records (A - J) is organized using KD tree data structure (grey lines are the branches of the tree). As a new waveform is recorded, the target

record (yellow star) only needs to visit 5 of the data points (red points) to find the record with the most similar record with respect to the select features: initial 3 sec velocity and acceleration of the p-wave. In the KD tree method, only the data points with branches that intersect the hypersphere (shaded circle) are possible candidates; being closer than the current nearest node, other nodes (blue points) can be ignored. As a comparison, the linear sequential search requires going through all 10 records, which doubles the computation effort.

### *5.3.2 Method*

Since one of the ultimate goals of EEW aims to predict ground shaking, we extracted 500 records from the entire database to validate the prediction of earthquake source and ground motion parameters. The validation set was sampled uniformly with even spacing on the Peak Ground Acceleration (PGA) of the records. The reason is to cover the full spectrum of ground shaking intensity, in order to mimic all circumstances that could be encountered in the future. The performance of parameter estimations is evaluated with different dataset sizes. The estimated seismic parameters include station-specific ground motions: Peak Ground Acceleration (PGA), Peak Ground Velocity (PGV), Peak Ground Displacement (PGD), and earthquake source parameters: magnitude, hypocenter distance. The procedure first requires a 30-NN search in the Euclidean distance defined in Eq[5.1] and then a prediction using the Gaussian mean of the corresponding parameters from the 30-NN matched records. The value 30 is chosen to match the original model parameter in the GbA. Later, we compared the searching time of the KD-tree search to the Linear Sequential search, in both the CPU time and the number of operations.

## 5.4 Results

We computed the earthquake parameter estimation error of the validation set for databases with different sizes. Figure 5.2 to Figure 5.4 shows the $100^{th}$, $75^{th}$, $50^{th}$, $25^{th}$, and $0^{th}$ percentile residual errors for the estimated PGA, PGV, and PGD of the 500-validation dataset, respectively. The residual error is defined as the absolute difference between the true observed parameter and the predicted parameter. The $50^{th}$ percentile is the average residual errors; the $100^{th}$ and $0^{th}$ percentile indicate the maximum error and minimum error, respectively. As expected, the residual error decreases as the database size increases on average. The $50^{th}$ percentile is not flattened near the largest given database size showing that the residual errors might not yet reached the global minimum; this suggests that the estimation accuracy could further be improved by increasing the size of the database. The maximum error residual appears to be uncorrelated to the database size, because there is always the possibility of outlier targets regardless how large the database gets. Statistically, there will always be residual on the estimations, unless the features are truly uniquely diagnostic. Of course, if a sufficiently large database were compiled, the probability of encountering outliers would decrease.

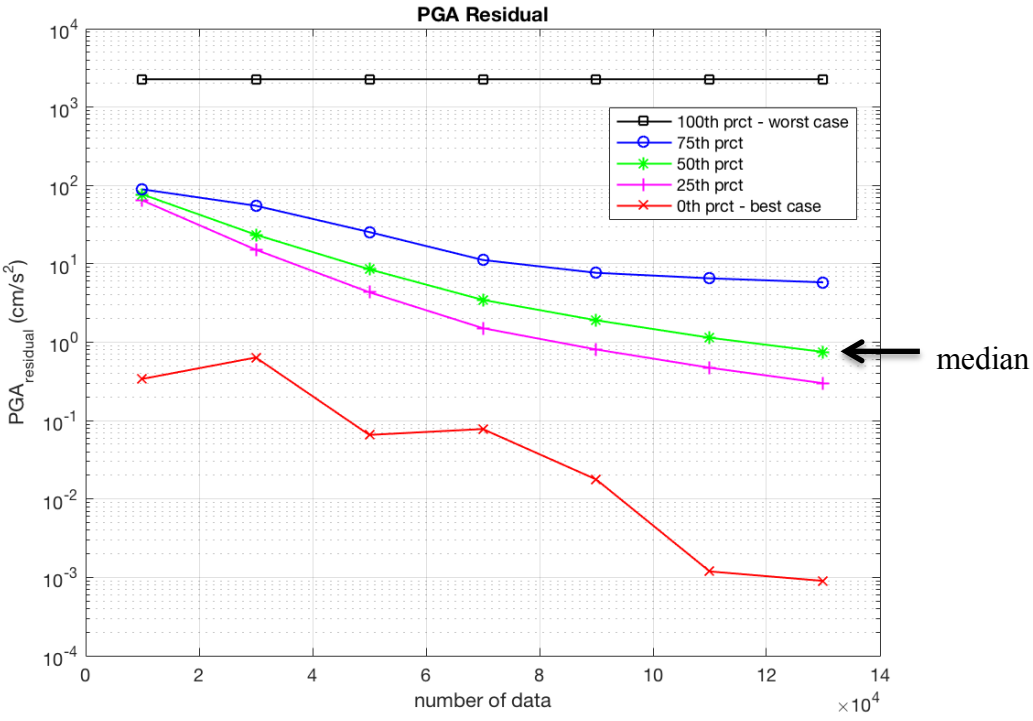



Figure 5.2 Ground motion residuals for the 500-validation dataset with different database sizes. Peak Ground Acceleration residuals are given in absolute ground motion units. The lines show the percentile according to the legend. The $50^{th}$ percentile is the average residual error; the $100^{th}$ and $0^{th}$ percentiles indicate the maximum and minimum errors respectively.

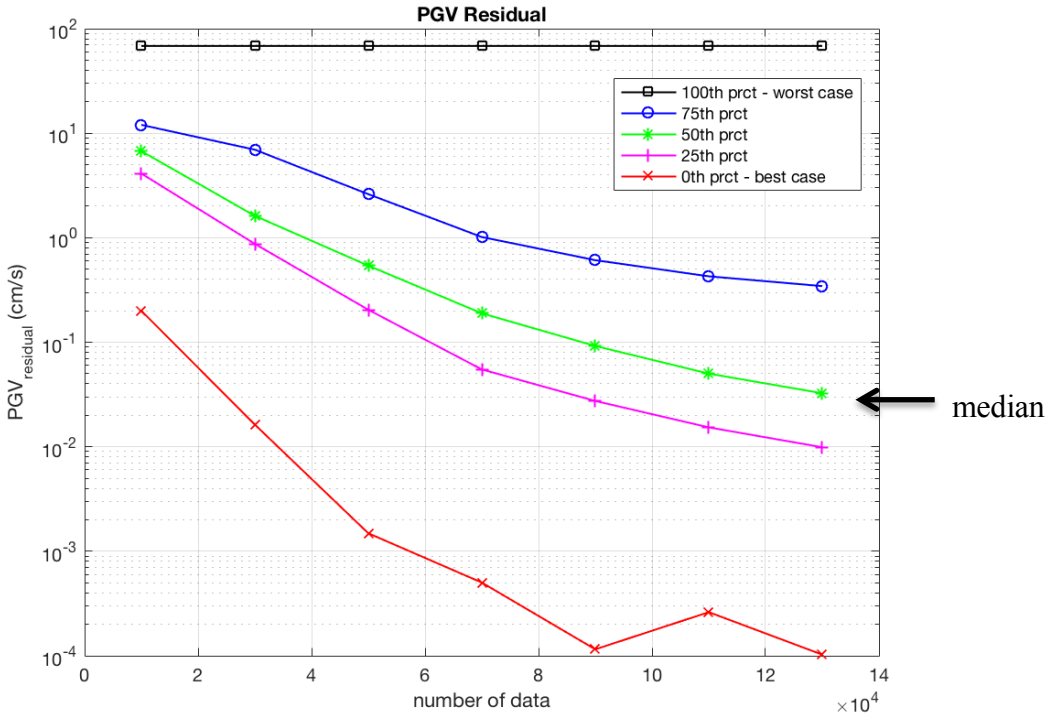



Figure 5.3 Ground motion residuals for the 500-validation dataset with different database sizes. Peak Ground Velocity residuals are given in absolute ground motion units. The lines show the percentile according to the legend. The 50$^{th}$ percentile is the average residual error; the 100$^{th}$ and 0$^{th}$ percentiles indicate the maximum and minimum errors respectively.

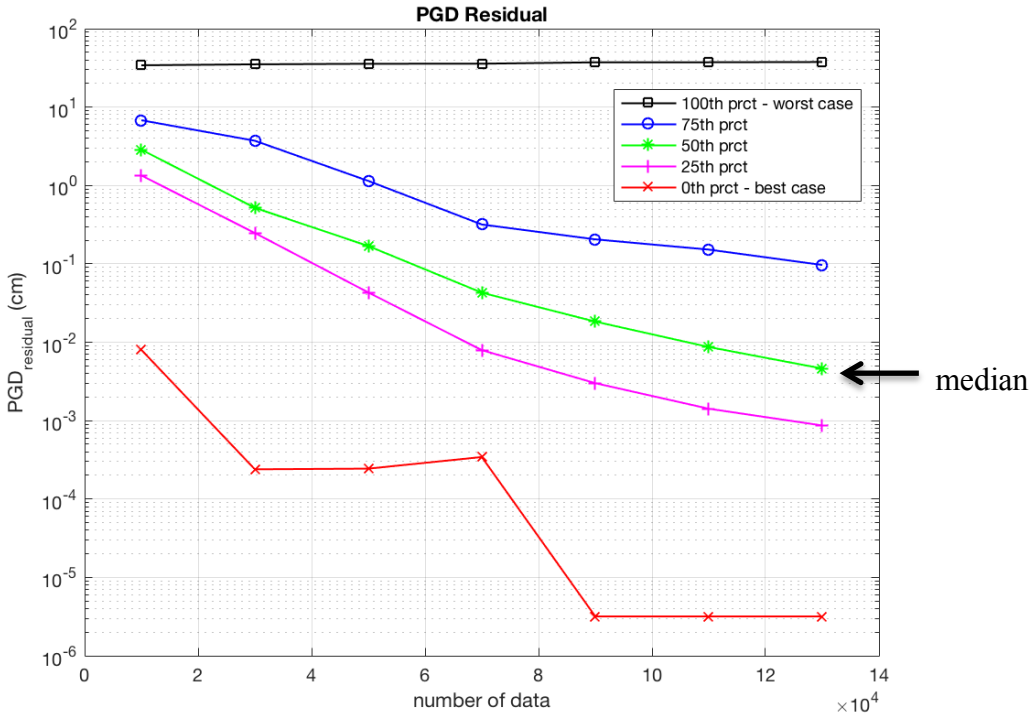



Figure 5.4 Ground motion residuals for the 500-validation dataset with different database sizes. Peak Ground Displacement residuals are given in absolute ground motion units. The lines show the percentile according to the legend. The 50$^{th}$ percentile is the average residual error; the 100$^{th}$ and 0$^{th}$ percentiles indicate the maximum and minimum errors respectively.

We also estimated the earthquake source parameters using the databases: magnitude and hypocenter distance. Although ground motion parameters are more useful outputs for EEW alerts, predicting source parameters is the conventional approach in real-time seismology (Minson et al., 2017). Figure 5.5 and Figure 5.6 show that the size of the database has less impact on hypocenter distance than magnitude estimation. Since hypocenter distance predictions from the observed waveform are a result of source energy and soil properties,

the additional constraints might be necessary. For example, seismicity location forecast could be introduced as prior knowledge to reduce the uncertainties in earthquake location estimation (Yin et al., 2017). This analysis implies that it is essential to select data features intelligently to characterize the parameters we are aiming to predict. Frequency band features might be more suitable to predict the ground motions than source parameters, since local site effects may be implicitly being accounted for.

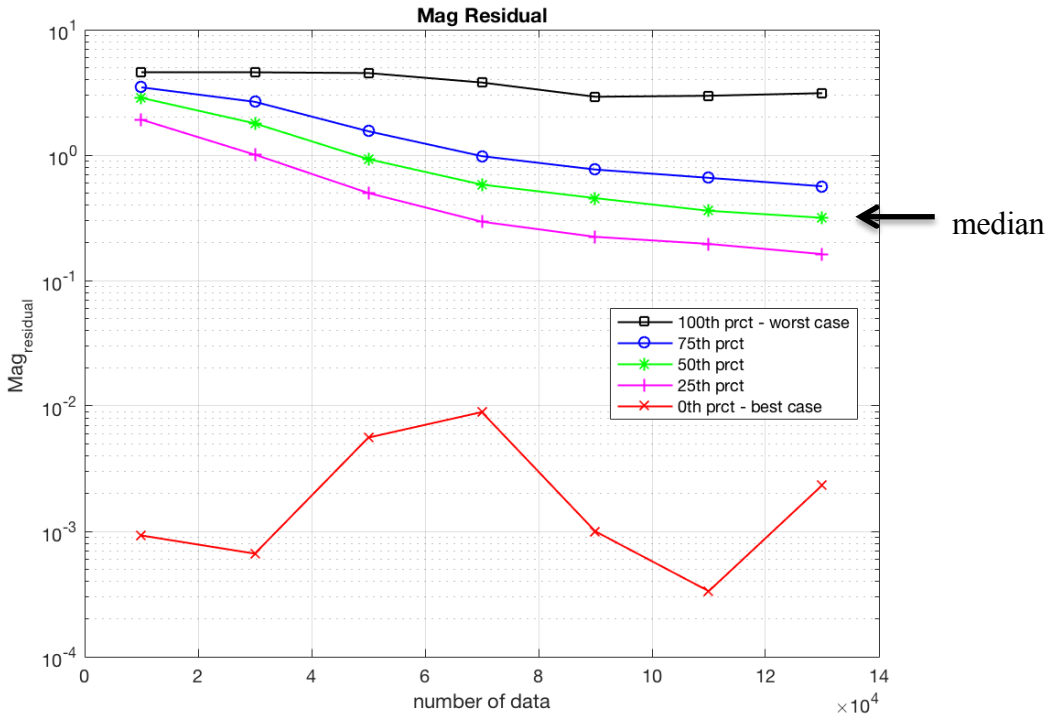



Figure 5.5 Source parameter residual for the 500-validation dataset with different database size. Magnitude residuals are given in absolute units. The lines show the percentile according to the legend. The 50$^{th}$ percentile is the average residual error; the 100$^{th}$

and $0^{th}$ percentiles indicate the maximum error and minimum error respectively.

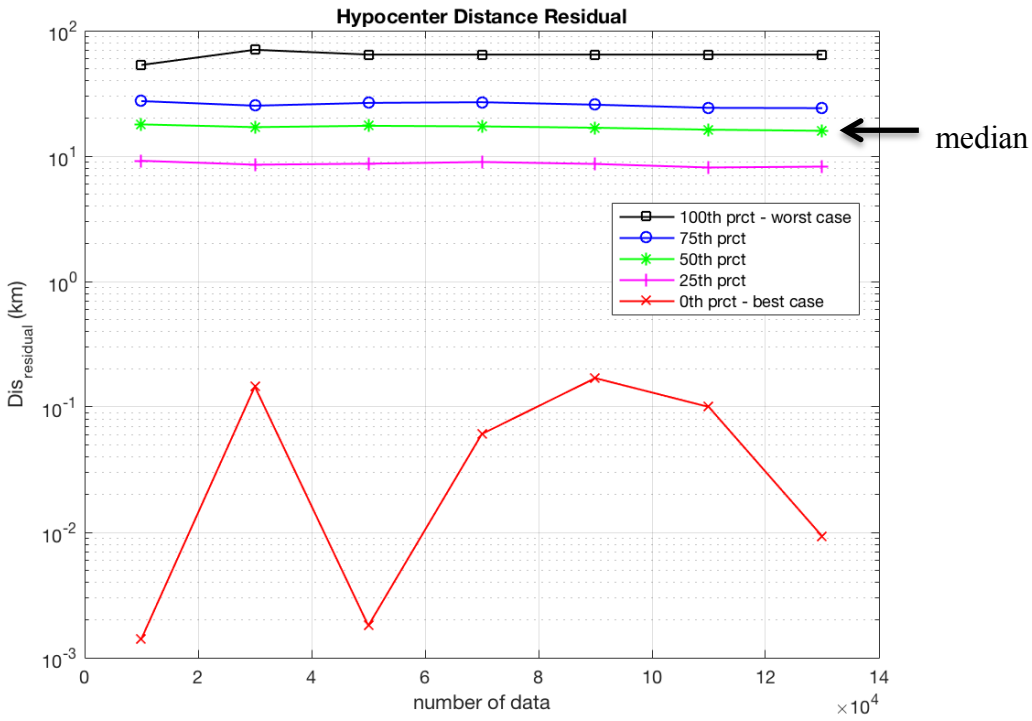



Figure 5.6 Source parameter residual for the 500-validation dataset with different database size. Hypocenter distance residuals are given in absolute units. The lines show the percentile according to the legend. The $50^{th}$ percentile is the average residual error; the $100^{th}$ and $0^{th}$ percentiles indicate the maximum error and minimum error respectively.

Through the performance analysis for databases with different sizes, we conclude that large databases can help to provide more accurate ground motion estimations for EEW. Next, we compare the computational time difference for the 30-NN search using the KD tree

methods for each validation test. The implementation is in Matlab. For comparison, a Linear Sequential search method is also implemented as a base case. The Matlab function follows the pseudo code concept from the Appendix with optimization modules that efficiently process the data. In Figure 5.7, the solid lines show that the average CPU search time of a database with 130, 000 points is about 0.2 sec for the Linear Sequential search method and 0.03 sec for the KD tree search method; the significant reduction in time reduces computational effort by 85%. Although the Linear Sequential search is capable of handling the real-time processing with limited delay using the current size of the database, a significant delay would be introduced as the database size rapidly increases in the future. The dashed lines show extrapolated computational time up to double of the current database size. The results anticipate that the advantages of the KD tree application would be emphasized in the future as global seismic databases are growing significantly (Yu, 2016).

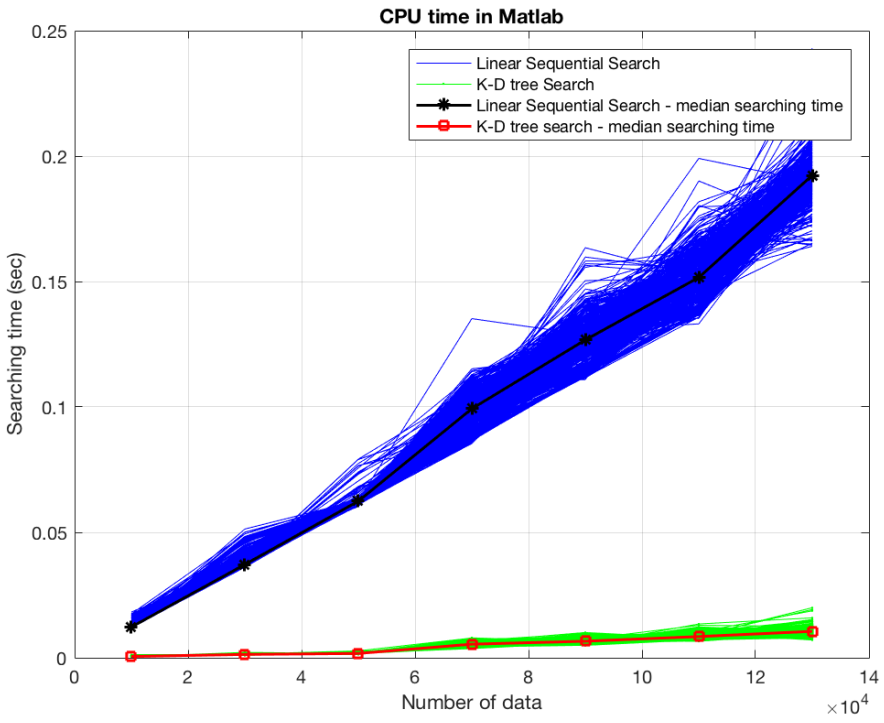



Figure 5.7 CPU searching time for different database sizes using linear sequential search and KD tree search. The implementation is in Matlab.

The measured operational time for the searching process varies significantly between different software languages and implementations; different optimization modules with parallelization might also bias towards one method over another. Implementations in C++ tend to be much faster than Matlab. In order to compare the true efficiency of the method across all platforms, we further compared the number of data points visited for both NN search algorithms. Since the majority of the searching time is made up by the visit to each data point to compute the Euclidean distance to the target point, the fewer data points visited ensures less time effort. In the Linear Sequential search, the operation is required for

all the data in the database in a serial manner. However, in KD tree, subsections of the database can be eliminated depending on the distribution of the tree structure and location of the target point. As shown in Figure 5.8, the number of data points visited in the KD tree for each validation varies; on average, the KD tree approach only visits about 10% of the entire database to find the closest data point to the target, confirming the performance in CPU searching time in Matlab. In the worst-case scenario, all the data points are visited, which leads to the same operational complexity as the exhaustive approach (linear sequential search).

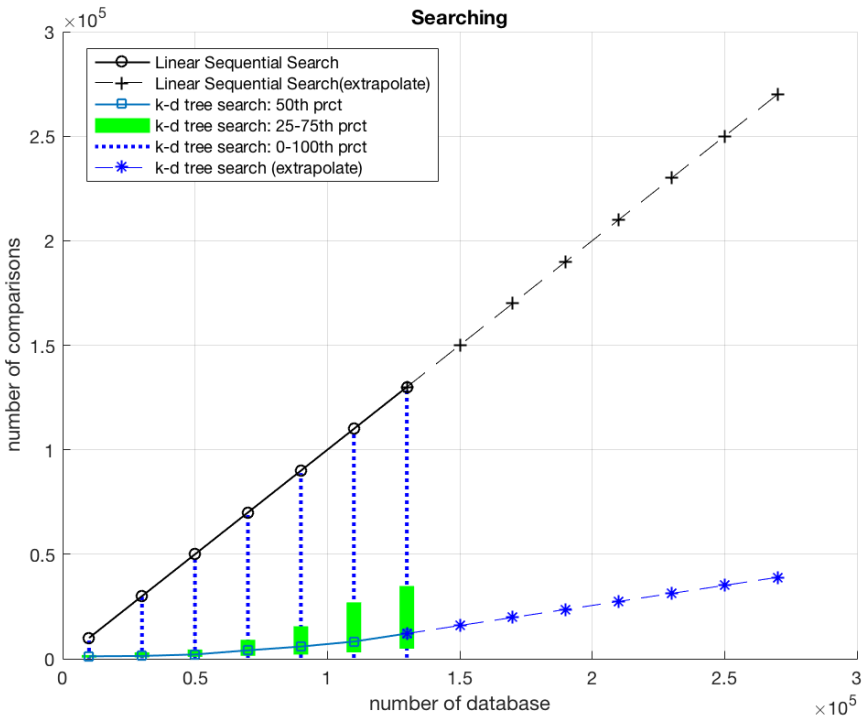



Figure 5.8 Number of data points visited for linear sequential search and KD tree search. The dashed lines are extrapolated to estimate the performance for larger database in the future.

## 5.5 Discussion and Conclusion

In this study, we evaluated the viability of earthquake fingerprint searching methods for EEW, using database structure to reduce searching time for large databases. Specifically, we evaluated the GbA as an example of the EEW fingerprint search algorithm. We found that database size is a critical factor in providing reliable predictions of ground motion (PGA, PGV, PGD) and source parameters (magnitude and hypocenter distance) for EEW. We also present the KD tree approach to reduce the searching time, so that large database searching is feasible for real-time implementations in EEW. By empirical validation, we demonstrated that the searching time using KD tree is about 85% less than the exhaustive approach for the GbA EEW earthquake database.

One of the potential applications of the database searching method is to directly estimate peak ground motions from the observed ground motions for any given site in real-time seismology application such as EEW; it avoids the multi-step modeling errors that could be accumulated through source parameter estimation and the ground motion attenuation relationship, since the final errors can lead to significant uncertainties in the final shaking information. Ideally, the goal of EEW is to serve as an alarm for severe ground shaking in real-time rather than source characterization. The fingerprint searching methodology could also be extended to tackle other challenges in EEW, such as event detection (i.e. earthquake/noise discrimination). In such a problem, characteristics of additional ambient noise and teleseismic records need to be incorporated in the database. This would vastly increase the database size, since incorporating many different types of noise, teleseisms, regional events, and calibration/maintenance signals could potentially be huge. The vision is to be able to accomplish efficient searching for large databases, so that these novel EEW methods are feasible in real-time in the future.

Although we emphasized the importance of having a large number of data, a question is often raised about what should be the minimum size of database in order to get reasonable accurate solutions. Assuming the standard deviation of $\log_{10}$(PGV) estimation of 0.309 by (Kanamori, 2007) is acceptable, the database size needed to achieve this marginal error of ground motion in EEW is about 70 000 to 100 000 data points, as shown in Figure 5.3. The (Kanamori, 2007) study focuses on two EEW parameters, $\tau_c$ and $P_d$, that are extensively used in the existing EEW algorithms, such as Onsite (Bose et al., 2009). The minimum database size calculated varies with geological region, event types, predictive parameters, etc.

Creating a database for real-time seismology is not simple. In addition to the sizes of databases, feature engineering also significantly affects the prediction results. Selecting parameters that correlate to the predictive results requires extensive scientific domain knowledge. In the observation of local earthquake records, the higher frequency band features are more informative than the low frequency features because the high frequency amplitude of ground motion decays rapidly with distance (Hanks & McGuire, 1981) (Kong & Zhao, 2012). A weighted Euclidean distance might be more applicable to emphasize the high frequency information as the significant attributes in the feature space. Continuous monitoring and modifying of the features will help to improve the performance of the system. As the number of features increases, the process time saved by KD tree search decreases (Andoni & Indyk, 2008). For features over 20 or 30 dimensions, alternative approximation to approaches high dimensional searching, such as Locality Sensitive Hashing, would be more appropriate [e.g. (Yoon et al., 2015)].

The accuracy and speed of rapid earthquake source parameter algorithms has significantly improved over the past decade, but are potentially limited by the simplification involved in model parameterization. The earthquake fingerprint searching techniques have the capacity

to guide the development of EEW to a new phase with the assistance of better computational power and data mining techniques.

*Chapter 6*

# Conclusion

## 6.1 Final remarks

Since devastating earthquakes grow rapidly over the time frame of seconds, immediate responses (including automatic alerts and interruption of activities) are essential to mitigate the losses due to the destructions of ground shaking. With this motivation, many scientists and engineers have been continuously focusing on the improvement of faster and more accurate Earthquake Early Warning systems. Particularly, this thesis presents methods to reduce alert latency of EEW system for the earliest alerts while maintaining the accuracy requirements of the estimated earthquake information. This final chapter summarizes the thesis and proposes suggestions for possible directions of the extension research on earthquake early warning.

In Chapter 2, I presented previous studies of the science of earthquake sequences and the modeling earthquake forecasting, especially Epidemic-type Aftershock Sequence (ETAS) modeling, a statistical approach to forecast near future probability of seismic activities. I developed a real-time ETAS algorithm that takes historical seismic catalog and outputs the forecast probabilistic seismic map for near future. The predicted results of the ETAS algorithm are compared to the observed seismic activities for validation. The outputs of earthquake forecasting information are useful to be incorporated into EEW under the Bayesian framework as a source of prior information. The applications for signal discrimination and hypocenter location estimations were presented in Chapter 3 and 4.

In Chapter 3, I presented a classification algorithm that distinguishes near-field earthquake source signals from noise and teleseismic arrivals. This method uses the three-component acceleration and velocity waveform data and Epidemic-Type Aftershock Sequence (ETAS) seismicity forecast information in parallel, producing a posterior prediction by combining the predictions from the heterogeneous sources using a Bayesian probabilistic approach. I collected 2,481 three-component strong-motion records for training and testing. The rapid prediction is available as quickly as 0.5 sec after the trigger at a single station and updates every 0.5 sec up to 3.0 sec, achieving a precision rate of 98% at the first prediction with the classification accuracy increasing with time. The leave-one-out validation method also demonstrates confidence of robust performance for future earthquake signal detections. I compared the method with the $\tau_c - P_d$ EEW classification criterion and find that our prediction is 83% faster with 5% higher precision rate. Since the method evaluates two independent sources of information simultaneously under an ensemble model, the new strategy has shown fast predictions with promising results, and the implementation of this methodology could provide significantly faster and more reliable EEW warnings to regions near the earthquake's epicenter where the strongest shaking is observed.

In Chapter 4, I applied the ETAS results under the Bayesian probabilistic framework to provide optimally fast estimates of earthquake hypocenter location; in many cases, the earthquake location is available as soon as the first P-wave arrival at the station located closest to the epicenter. In order to provide reliable warning as quickly as possible before the arrival of damaging ground shaking, the Bayesian prior of ETAS seismicity forecast model provide a intuitive initial approximation; the analysis of seismic waveform information is incorporated into the solution as data becomes available over time. I have evaluated the algorithm for all 504 M4+ earthquakes in Southern California from 1990 to 2005. For the earliest epicentral location estimation at 0.5 sec after P-wave detection, the median location error using seismicity forecast with waveform analysis improved by 58%

relative to results using waveform analysis only. I also presented location estimations of a M5.2 Lone Pine Earthquake and a M5.4 Chino Hills Earthquake in detail, which highlights the importance of Bayesian seismic prior and waveform likelihood interaction. The new strategy has shown promising results and implementation of this methodology should significantly enhance the performance of EEW systems.

In Chapter 5, I proposed to use a multidimensional binary search tree (KD tree) data structure to organize large seismic databases to reduce the processing time in nearest neighbor search for predictions. Earthquake parameter estimations using nearest neighbor searching among a large database of observations can lead to reliable prediction results. However, in the real-time application of Earthquake Early Warning (EEW) systems, the accurate prediction using a large database is penalized by a significant delay in the processing time. I evaluated the performance of KD tree on the Gutenberg Algorithm, a database-searching algorithm for EEW. I constructed an offline test to predict peak ground motions using a database with feature sets of waveform filter-bank characteristics, and compare the results with the observed seismic parameters. I concluded that large database provides more accurate predictions of the ground motion information, such as peak ground acceleration, velocity, and displacement (PGA, PGV, PGD), than source parameters, such as hypocenter distance. Application of the KD tree search to organize the database reduced the average searching process by 85% time cost of the exhaustive method, allowing the method to be feasible for real-time implementation. The algorithm is straightforward and the results will reduce the overall time of warning delivery for EEW.

## 6.2 Future work

From the first documentation on Earthquake indicator by (Cooper 1868) to the proposal of seismic computerized alert network by (Heaton 1985) to the suggestion of real-time seismology by (Kanamori 2005), scientists have already taken giant steps towards

advancing the technology of seismic early warning and post- earthquake response. While we recognize the contrasting needs for speed and accuracy of information, we still aim to provide reliable estimates (earthquake identification, source parameters and distribution of shaking) with the minimum latency, so users can gain warning time to prepare for the incoming strong shaking. A few of my suggestions for future investigation to extend my studies in this thesis are as follows:

- In order to incorporate the ETAS results into EEW system for practical use, ***real-time streaming of earthquake catalog database*** is necessary. Currently, there are delays in updating the earthquake catalog provided by the authorized agencies due to the validation of information. An ideal approach to this is to directly use the previous earthquake source parameter outputs from EEW as the catalog input to ETAS. Of course, it will be achievable only if the EEW results become reliable and robust in the future.

- ***Additional data source*** can also be incorporated into the Bayesian framework of EEW, such as GPS data or gravity-based sensor (Harms, et al. 2015). Although interpreting the information from heterogeneous data sources under the same metric is challenging, converting all information into probability and then combining under Bayesian framework can be straightforward and convenient.

- The current real-time ETAS model is based on point source event, the ***extension to finite fault implementation*** for large events can better reflect the true observation of seismic activities. This requires the input of geometric information from the known faults. Due to the complication of finite fault calculation, some processing delay might be introduced to improve the accuracy of the predictions.

- In Chapter 3, I collected noise amplitudes from stations across the entire network to develop the ***noise model*** for earthquake-versus-noise discriminator. However, each specific station would experience different ambient noise level depending on their geographic locations. The stations in an urban area often observes a high noise level due to the busy traffics, while stations on the Mt Wilson might record a low noise level. Figure 6.1 shows the ambient noise amplitudes recorded by various stations in the Community Seismic Network. If noise models can be developed based on geometric constraints, instrument conditions and station specifications, the analysis of the signals can be more precise.

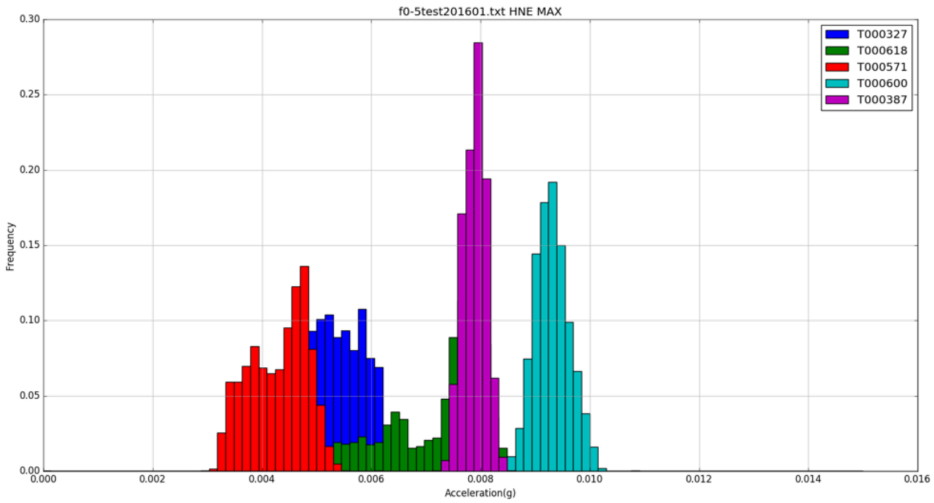



Figure 6.1 Noise amplitude records from a few selected Community Seismic Network stations (provided by the CISN research group)

- In this thesis, I have applied knowledge of prior seismicity to solve the challenges of earthquake detection and hypocenter location estimation. However, it is not possible to directly estimate earthquake magnitude with the assistance of prior

seismic information, since the size of each event is independent and not related to previous seismicity. The vision of this project is to use prior information to get an initial approximation of hypocenter locations, and then characterize the magnitude of the event from the estimated hypocenter and observed waveform amplitudes. As more data becomes available with time, the estimations can be updated accordingly. With the powerful computing power, ***grid search method and parallelization*** can be applied to optimize the efficiency in calculating the maximum likelihood estimations. Computational power can be prioritized to the grid locations with higher prior probability.

EEW is an interdisciplinary project that involves collaboration among different scientific and engineering communities. Only with the contribution of all seismologist, engineers, computer scientists, educators, media and many others' involvements, such a unified system would be successful every time in taking appropriate actions before, during, and after earthquake natural disasters.

# BIBLIOGRAPHY

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

Allen, R V. "Automatic earthquake recognition and timing from single traces." *Bulletin Seismological Society of America* 68 (1978): 1521-1532.

Allen, R., and H. Kanamori. "The potential for Earthquake Early Warning in Southern California." *Science* 300 (2003): 786-789.

Allen, Richard, and Hiroo Kanamori. "The potential for Earthquake Early Warning in Southern California." *Science* 300 (2003): 786-789.

Allen, Richard, Paolo Gasparini, Osamu Kamigaichi, and Maren Bose. "The status of Earthquake Early Warning around the World: An Introduction Overview." *Seismological Research Letters* 80 (2009): 682-688.

Bakun, W H, F G Fischer, E G Jensen, and J VanSchaack. "Early Warning System for Aftershocks." *Bullentin of the Seismological Society of America* 84 (1994): 359-365.

Bose, Heaton, and Hauksson. "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, no. 2 (2012): 738-750.

Bose, M, E Hauksson, K Solanki, H Kanamori, and T H Heaton. "Real-time testing of the on-site warning algorithm in southern California and its performance during the July 29 2008 Mw 5.4 Chino Hills earthquake." *Geophysical Research Letters* 36 (2009): L00B03.

Bose, M, E Hauksson, K Solanki, H Kanamori, Y-M Wu, and T H Heaton. "A Trigger Criterion for Improved Real-Time Performance of Onsite Earthquake Early Warning in Southern California." *Bulletin of Seismological Society of America* 99 (2009): 897-905.

Bose, M, et al. "CISN ShakeAlert - An Earthquake Early Warning Demonstration System for California." In *Early Warning for Geological Disasters*, 49-69. Springer, 2014.

Bose, M, T H Heaton, and E Hauksson. "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, no. 2 (2012): 738-750.

Bouchon, M, V Durand, D Marsan, H Karabulut, and J Schmittbuhl. "The long precursory phase of most large interplate earthquakes." *Nature Geoscience*, 2013: 299-302.

Bouchon, Michel, V Durand, D Marsan, H Karabulut, and J Schmittbuhl. "The long precursory phase of most large interplate earthquakes." *Nature Geoscience* 6 (2013): 299-302.

Cheng, M H, S Wu, T H Heaton, and J L Beck. "Earthquake early warning application to buildings." *Engineering Structures* 60 (2014): 155-164.

Cooper, JD. "Earthquake Indicator." *San Francisco Bulletin*, 11 1868.

Cua. "Creating the virtual seismologist: developments in earthquake early warning and ground motion characterization." Ph.D. Thesis, Department of Civil Engineering, California Institute of Technology, Pasadena, 2005.

Cua, G. "Creating the virtual seismologist: developments in earthquake early warning and ground motion characterization." Ph.D. Thesis, Department of Civil Engineering, California Institute of Technology, 2005.

Cua, G, and T H Heaton. "The Virtual Seismologist (VS) method: A Bayesian approach to earthquake early warning." In *Earthquake Early Warning Systems*, by P Gasparini, G Manfredi and J Zschau, 85-132. New York: Springer, 2007.

Doi. "Earthquake early warning system in Japan." In *Early Warning Systems for Natural Disaster Reductions*, 447-452. Springer, 2003.

Espinosa-Aranda, J, A Jimenez, G Ibarrola, and F Alcantar. "Results of the Mexico City Early warnng system." *11th World Conference on Earthquake Engineering.* 1996. No.2132.

Espinosa-Aranda, J.M., et al. "Mexico City Seismic Alert System." *Seismological Research Letters* 66 (1995): 42-52.

Felzer, K. *Stochastic ETAS aftershock simulator.* 10 31, 2007. http://pasadena.wr.usgs.gov/office/kfelzer/ (accessed 05 13, 2015).

Felzer, K. "Simulated Aftershock Sequences for an M 7.8 Earthquake on the Southern San Andreas Fault." *Seismological research Letters* 80, no. 1 (2009): 21-25.

Felzer, K, and E E Brodsky. "Decay of aftershock density with distance indicates triggering by dynamic stress." *Nature* 441 (2006): 735-738.

Felzer, K, T Becker, R Abercrombie, G Ekstrom, and J Rice. "Triggering of the 1999 Mw 7.1 Hector Mine earthquake by aftershocks of the 1992 Mw 7.3 Landers earthquake." *J. Geophs. Res.* 107, no. 2190 (2002).

Felzer, K., and E. Brodsky. "Decay of aftershock density with distance indicates triggering by dynamic stress." *Nature* 441 (2006): 735-738.

Felzer, Karen. "Simulated Aftershock Sequences for an M 7.8 Earthquake on the Southern San Andreas Fault." *Seismological research Letters* 80, no. 1 (2009): 21-25.

Fujinawa, Y., and Y. Noda. "Japan's earthquak early warning system on 11 March 2011: Performance, shortcomings, and changes." *Earthquake spectra* 29 (2013): S341-S368.

Gerstenberger, M, Stefan Wiemer, L Jones, and P Reasenberg. "Real-time forecasts of tomorrow's earthquake in California." *Nature* 435 (2005): 328-331.

Gomberg, J, P. A. Reasenberg, P Bodin, and R. A. Harris. "Earthquake triggering by seismic waves following the Landers and Hector Mine earthquakes." *Nature* 411 (2001): 462-466.

Gutenberg, B, and C F Richter. "Frequency of earthquakes in California." *Bulletin of the Seismological Society of America* 4 (1944): 185-188.

Hainzl, S, and Y Ogata. "Detecting fluid signals in seismicity data through statistical earthquake modeling." *J. Geophys. Res.*, 2005: 110.

Hanks, T C, and R K McGuire. "The Character of high-frequency strong ground motion." *Bulletin of the Seismological Society of America* 71, no. 6 (1981): 2071-2095.

Harms, J, et al. "Transient gravity perturbations induced by earthquake rupture ." *Geophysical Journal International* 201 (2015): 1416-1425.

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

Horiuchi, Y. "Earthquake early warning hospital applications." *J. Disast. Res.* 4 (2009): 237-241.

Ionescu, C, M Bose, F Wenzel, A Marmureanu, A Grigore, and G Marmureanu. "Early warning system for deep Vrancea (Romania) earthquakes." In *Earthquake Early Warning Systems*, 343-349. 2007.

Kagan, Y Y, and D D Jackson. "Long-term earthquake clustering." *Geophysical Journal International* 104 (1991): 117-133.

Kagan, Y., and L. Knopoff. "Stochastic synthesis of earthquake catalogs." *J. Geophys. Res.* 86 (1981): 2853-2862.

Kanamori, H. "Real-time seismology and earthquake damage mitigation." *Annual Review of Earthquake Planerary Sciences* 33 (2005): 195-214.

Kuyuk, H S, and R M Allen. "Designing a Network-Based Earthquake Early Warning Algorithm for California: ElarmS-2." *Bull. Seismo. Soc. Am.*, no. 104 (2014): 162-173.

Kuyuk, H S, S Colombelli, A Zollo, R M Allen, and M O Erdik. "Automatic earthquake confirmation for early warning system." *Geophysical Research Letter* 42 (2015): 5266-5273.

Kuyuk, S, R M Allen, H Brown, M Hellweg, I Henson, and D Neuhauser. "Designing a Network-Based Earthquake Early Warning Algorithm for California: ElarmS-2." *Bulletin of the seismological Society of America* 104, no. 1 (2014): 162-173.

Lee, W, and J Espinosa-Aranda. "Earthquake early-warning systems: Current status and perspectives." In *Early Warning Systems for Natural Disaster Reduction*, 409-423. Springer, 2003.

Lohman, R B, and J J McGuire. "Earthquake swarms driven by aseismic creep in Salton Trough, California." *J. Geophys. Res.*, 2007: 112.

Lomnitz, C. "Magnitude stability in earthquake sequences." *Bulletin of the Seismological Society of America* 56 (1966): 247-249.

Meier, M, T Heaton, and J Clinton. "The Gutenberg Algorithm: Evolutionary Bayesian Magnitude Estimates for Earthquake Early Warning with a Filter Bank." *Bulletin of the Seismological Society of America* 105, no. 5 (2015).

Mogi, K. "Some dicussions on aftershocks, foreshocks and earthquake swarms: The fracture of a semi-infinite body caused by an inner stress origin and its relation to earthquake phenomena." *Bull. Earthquake Res Inst.* 41 (1963): 615-658.

Nakamura. "Development of earthquake early-warning system for the Shinkanse, some recent earthquake engineering research and practice in Japan." *Proceeding of The Japanese National Committee of the International Association for Earthquake Engineering*, 1984: 224-238.

Nakamura, Y, J Saita, T Araya, and T Sato. "The fastest o-wave warning system freql, UrEDAS and compact UrEDAS with actual situations." *100th Anniversary earthquake Conference Commemorating the 1906 San Francisco Earthquake.* 2006.

Nakamura, Y., and Tucker. "Japan's Earthquake Warning System: Should it be imported to California." *California Geology* 41, no. 2 (1988): 33-40.

Normile, D. "Earthquake preparedness: Some countries are betting that a few seconds can save lives." *Science* 306 (2004): 2178-2179.

Ogata, Y. "Space-time point-process models for earthquake occurrences." *Annals of Statistics* 50 (1998): 805-810.

Ogata, Y. "Statistical models for earthquake occurrence and residual analysis for point processes." *J. Am. Stat. Assoc.* 83 (1988): 9-27.

Reasenberg, P. "Foreshock occurrence before large earthquakes." *Journal of Geophysical research* 104 (1999): 4755-4768.

Reasenberg, P, and L Jones. "Earthquake Hazard After a Mainshock in California." *Science* 243 (1989): 1173-1176.

Reasenberg, P, and L. Jones. "Earthquake Hazard After a Mainshock in California." *Science* 243 (1989): 1173-1176.

Reasenberg, P. "Foreshock occurrence before large earthquakes." *Journal of Geophysical research* 104 (1999): 4755-4768.

Rosenberger, Andreas. "Arrival-Time Order Location Revisited." *Bulletin of the Seismological Society of America* 99, no. 3 (2009): 2027-2034.

Satriano, C, L Elia, C Martino, M Lancieri, A Zollo, and G Iannaccone. "PRESTo, the earthquake early warning system for Southern Italy: Concepts, capabilities and future perspectives." *Soil Dynamics and Earthquake Engieering* 31 (2011): 137-153.

Shearer, P. M. "Space-time clustering of seismicity in California and the distance dependence of earthquake triggering ." *Journal of Geophysical Research* 117, no. B10 (2012).

Strauss, Jennifer, and Richard Allen. "Benefits and Costs of Earthquake Early Warning." *Seismological Research Letters* 87, no. 3 (2016): 765-772.

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

Vere-Hones, D. "A Markov model for aftershock occurrence." *Pure App. Geophys.* 64 (1966): 31-42.

Wald, D J, V Quitoriano, T H Heaton, and H Kanamori. "Relationships between Peak Ground Acceleration, Peak Gound Velocity and Modified Mercalli Intensity in California." *Earthquake Spectra* 15(3) (1999): 557-564.

Wu, and Kanamori. "Rapid assessment of damage potential of earthquakes in Taiwan from the beginning of P waves." *Bellutin of the Seismological Society of America* 95, no. 3 (2005): 1181-1185.

Wu, S, J Beck, and T H Heaton. "Decision criteria for earthquake early warning applications." *Proceedings of the 15th World Conference on Earthquake Engineering.* 2012. 0973.

Wu, S., J Beck, and T Heaton. "ePAD: Earthquake probability-based automated decision-making framework for earthquake early warning." *Computer-aided Civil and Infrastructure Engineering*, 2013: 737-752.

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

Wu, Y, H Yen, L Zhao, B Huang , and W Liang. "Magnitude determination using initial p waves: A single-station approach." *Geophysical Research Letters* 33, no. 5 (2006).

Wu, Y., T Shin, and Y. Tsai. "Quick and reliable determination of magnitude for seismic early warning." *Bulletin of the Seismological Society of America* 88, no. 5 (1998): 1254-1259.

Yamada, Masumi, T H Heaton, and J L Beck. "Real-Time of Fault Reputure Extent Using Near-Source versus Far-Source Classification." *Bulletin of the Seismological Society of America* 97, no. 6 (2007): 1890-1910.