METHODS FOR EVALUATING EARTHQUAKE PREDICTIONS

by

Jeremy Douglas Zechar

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(GEOLOGICAL SCIENCES)

August 2008

TO MY FAMILY

This goes out to you. This goes out to you,

and you, and you, and you.

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many individuals. Chief among them is my advisor, Thomas Hillman Jordan. Tom was my staunchest supporter and my toughest critic; his infectious enthusiasm for tackling difficult problems lured me into the geosciences, and he kept me in high spirits throughout my student career.

Each member of my qualifying examination committee and my dissertation committee provided useful feedback and thought-provoking questions; for their service, I thank Charlie Sammis, James Dolan, Thomas Lin, Thorsten Becker, Yehuda Ben-Zion, and Yolanda Gil. For stimulating conversations—scientific and otherwise—and thoughtful papers, I thank Danijel Schorlemmer, Dave Jackson, David Rhoades, Ilya Zaliapin, Jeff McGuire, Max Werner, Peter Shebalin, Volodya Keilis-Borok, and Yan Kagan.

During my studies, a number of people shared their offices and homes with me. For being gracious hosts in Boston, I thank Alan Kafka, Uncle Chris Johnston, Eleanor Sonley, Jennifer Anne Wade, John Ebel, Aunt Pat Johnston, Rachel Scudder, and Sharon Gershman. For my time in New York, I wish to thank Chris Pietroburgo, Melissa Cooper, Rozzie Cooper, and Stanley J. Cooper. I am thankful to Beebe Longstreet, who hosted me in Santa Barbara during the critical time leading up to quals. For many nights at the pink house, I thank Aissa Riley, John Atkinson, and Worthless Rob. I also enjoyed great support and friendship from USC/SCEC staff members Barbara Grubb, Cindy

Waite, Dana Coyle, John McRaney, John Yu, Masha Liukis, Nitin Gupta, Phil Maechling, Sally Henyey, Shelly Werner, and Vipin Gupta.

I owe thanks to the following dear friends for keeping me laughing: Alicia Granse, Amir Khan, Andrea Wolcott, Angelo Antignano, Ben Costanza, Caitlin Longstreet, Catherine Powers, Charles Slomovitz, Cheryl Bautista, Cheryl Eng, Genevieve Rodriguez, Holden Caufield, Iain Bailey, Jennipher Galvan-Speers, Margaret Lee, Peter Powers, Ruben Speers, Ryan Dibble, Sara Pelosi, Suzi Elzie, Tim Bentley, Waldo Johnson, Wenzheng Yang, and Zheqiang Shi.

My visits to Synergy Charter Academy were often my week's highlight and I am indebted to Meg Palisoc and Randy Palisoc for letting me in on the fun. Thanks go to Ms. Aquino, Mrs. Chua, Ms. Maria, and Mr. Santana for creating such a professional and welcome environment, even when you were talking about me and thought I couldn't understand Spanish. Mrs. Bárcenas, Mrs. Deomampo, Mrs. Epps, Mr. Lawton, Mr. Louie, Mrs. Partida, and Mrs. Scanell: your dedication to teaching is inspiring, and I am so pleased to see your hard work resulting in accolades for Synergy. The students at Synergy warmed my heart and made me laugh each time I saw them; I wish them continued success, and I hope to see them at Synergy Quantum Academy.

My memories of grad school will always be inextricably linked to Kurt Frankel and Stephanie Briggs. Kurt turned out to be less of a jerk than I originally estimated, and Stephanie showed me how to throw down on some candy. They shared their unbridled enthusiasm for all things Jack White, and I also learned what little geology I know from traipsing around Death Valley with them. I thank Kurt and Stephanie for their friendship, and I look forward to doubling up on some milkshakes with them soon.

Without the love, support, and arguments I received from Lauren Beth Cooper, it's more likely than not that this work would have gone unfinished. I anticipate more of the same and hope that I can return the favors in the coming year. Lauren, you are the icing on my cake.

My mother Sharon and my biological father Ron have given me everything: their love, encouragement, and a distinguished-sounding middle name. My red-headed sister Ashleigh has been a good friend and source of amusement and my big brother Matthew Scott has always offered his guidance on my mathematical difficulties. To the Zechar family, I thank you for your unending patience and trust.

The following musical artists have helped me enjoy these last few years and an acknowledgements list would be incomplete without them: [[[VVRSSNN]]], 3Ds, Acid Ranch, Adam Green, Afghan Whigs, Airport 5, Andre 3000, Archers of Loaf, The Band, Bardo Pond, Be Gulls, Be Your Own Pet, Beanie Sigel, Beck, Big Black, Björk, Blitzen Trapper, Blood on The Wall, The Blow, Bob Dylan, Bob Marley, Bobby Birdman, Bobby Digital, Bottomless Pit, Brainiac, Breeders, Bright Eyes, Bruce Springsteen, Bubba Sparxxx, Built To Spill, Calvin Johnson, Cam'ron, Cat Power, Catfish Haven, Chavez, Circus Devils, Clipse, Cody ChesnuTT, Common, Counting Crows, Crooked Fingers, Crust Brothers, Cursive, D'Angelo, DangerDoom, David Bowie, David Vandervelde, Dead C, Dead Meadow, Deerhoof, Deerhunter, Desaparecidos, Dinosaur Jr, Dizzee Rascal, DJ Clue, Dr. Dre, Elvis Costello, Eminem, Enon, Everclear, Feist, Flaming Lips, Franklin Bruno, Freeway, Frogs, Fugazi, Funkadelic, Gang of Four, Gary Wilson, Gaunt, Ghostface Killah, Girl Talk, Gnarls Barkley, Go Back Snowball, Golden Animals, Golden Boots, Good Life, Gravediggaz, Green Day, Greenhorn, Grifters,

Guided by Voices, GZA/Genius, Halo Benders, Hazzard Hotrods, Hockey Night, Hold Steady, Hole, Howling Wolf Orchestra, Iceland Symphony Orchestra, Iron Horse, Islands, Jadakiss, James Brown, Jane's Addiction, Jason Anderson, Jason Molina, Jason Traeger, Jay-Z, Jeff Mangum, Jenny Lewis, Jens Lekman, Jimi Hendrix Experience, Joanna Newsom, Johnny Cash, Jonathan Richman, Juicy, Justin Timberlake, Kanye West, Keene Brothers, Killer Mike, Kimya Dawson, Lady Sovereign, Lauryn Hill, Led Zeppelin, Lemonheads, Leonard Cohen, Lexo And The Leapers, Liars, Lifeguards, Lifter Puller, Lil Wayne, Little Wings, Liz Phair, Love Letter Band, LOX, Lupe Fiasco, Magnolia Electric Co, Marvin Gaye, Mates of State, McLusky, Memphis Bleek, Microphones, Mike Watt, Mirah, Mission of Burma, Missy Elliot, Modern Lovers, Modest Mouse, Moldy Peaches, Moping Swans, Mount Eerie, Mountain Goats, My Morning Jacket, Neil Young, Neutral Milk Hotel, Nirvana, Nothing Painted Blue, Notorious B.I.G., Okkervil River, Ol' Dirty Bastard, Otis Redding, OutKast, Paul Westerberg, Pavement, Pixies, Polvo, Portastatic, Preston School of Industry, Prince, R. Kelly, R.E.M., Rapeman, Replacements, Rhymefest, Rilo Kiley, Robert Pollard, Roots, Rosebuds, RZA, Scarface, Scout Niblett, Scrawl, Sebadoh, Seu Jorge, Shellac, Shins, Shyne, Silkworm, Silver Jews, Sleater-Kinney, Slint, Smog, Smokin Hot Bitch, Smudge, Snoop Dogg, Songs; Ohia, Sonic Youth, Spinanes, Spoon, Stephen Malkmus & The Jicks, Stevie Wonder, Strand of Oaks, Strokes, Sugar, Suicide, Superchunk, Swearing at Motorists, Talking Heads, Ted Leo and the Pharmacists, That Dog, The Takeovers, The Thermals, Thomas Jefferson Slave Apartments, Thurston Moore, Timbaland, Times New Viking, Tom Petty & The Heartbreakers, TV On The Radio, Twilight Singers, UGK,

Ugly Casanova, V-3, Walkmen, White Rainbow, White Stripes, Wilco, Wolf Colonel, Wu Tang Clan, and Yo La Tengo. Thank you all.

Parts of this dissertation have been published previously as:

Zechar, J.D., & Jordan, T.H., 2008. Testing alarm-based earthquake predictions, Geophys. J. Inter., doi:10.1111/j.1365-246X.2007.03676.x.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF TABLES x

LIST OF FIGURES xi

ABSTRACT xiii

CHAPTER ONE: INTRODUCTION & OVERVIEW 1

CHAPTER TWO: Formal evaluation of the Reverse Tracing of Precursors earthquake prediction algorithm 7
Abstract 7
2.1 Reverse Tracing of Precursors 8
2.2 Alarms and earthquakes excluded from this study 16
2.3 Contingency table analysis of RTP 17
2.4 Molchan diagram analysis of RTP 21
2.5 Prior probability analysis of RTP 24
2.6 Discussion and conclusions 27

CHAPTER THREE: Testing alarm-based earthquake predictions 30
Abstract 30
3.1 Introduction 31
3.2 Alarm-based prediction 32
3.3 Optimal Molchan trajectories 39
3.4 Alarm functions and the area skill score 42
3.5 Models and data 45
3.6 Models and data 51
3.7 Conclusions 54

CHAPTER FOUR: The area skill score statistic for evaluating earthquake predictability experiments 57
Abstract 57
4.1 Earthquake forecasting with an alarm function 58
4.2 Molchan diagram for testing alarm functions 60
4.3 Area skill score 63
4.4 Area skill score distribution 65
4.4 Higher moments of the area skill score distribution 68
4.5 Experimental discretization 72
4.6 Discussion 77
4.7 Conclusion 78

CHAPTER FIVE: Optimizing earthquake forecasts based on smoothed seismicity 79
Abstract 79
5.1 Smoothed seismicity reference models 80
5.2 Functional forms of smoothing kernels 82
5.3 Smoothing experiments 84
5.4 Fixed test period 86
5.5 Moving test period, growing learning period 95
5.5 Moving test period, moving learning period 101
5.6 Discussion and conclusion 104

CHAPTER SIX: Conclusions, ongoing work, and potential extensions 109
6.1 Introduction 109
6.2 Proposed experiments in southern California 111
6.3 Proposed experiments in Japan and Taiwan 112
6.4 Proposed repeating micro-earthquake experiments 113

BIBLIOGRAPHY 114

APPENDIX A: Reverse Tracing of Precursors alarm specifications 122

APPENDIX B: Finding expected value of Molchan trajectory jumps 126

APPENDIX C: Method for finding area skill score moments 130

APPENDIX D: Number of nonzero sums of a set's elements 134

LIST OF TABLES

Table 2.1: RTP Intermediate Term precursors 9

Table 2.2: RTP study regions..... 14

Table 2.3: RTP alarms and outcomes to date 14

Table 2.4: Target earthquakes..... 15

Table 2.5: RTP contingency table..... 18

Table 2.6: RTP contingency tables measures of success..... 20

Table 3.1: Target earthquakes..... 51

Table 4.1: Moments about origin..... 69

Table 5.1: California study region 85

Table 5.2: Target earthquakes..... 86

Table C.1: Number of tuples..... 133

LIST OF FIGURES

Figure 2.1: Japan RTP natural laboratory. 11

Figure 2.2: California RTP natural laboratory 12

Figure 2.3: Eastern Mediterranean RTP natural laboratory 12

Figure 2.4: Italy RTP natural laboratory 13

Figure 2.5: North Pacific RTP natural laboratory 13

Figure 2.6: Example RTP hit 18

Figure 2.7: Example RTP false alarm. 19

Figure 2.8: Molchan diagram for RTP experiment 23

Figure 2.9: Molchan diagram for RTP experiment (2 extra hits) 24

Figure 2.10: RTP simulation results 26

Figure 3.1: Molchan diagram 35

Figure 3.2 Example alarm function 37

Figure 3.3: Molchan diagram confidence bounds 38

Figure 3.4: Illustration of alarm optimization 40

Figure 3.5: Map view of an example alarm function 43

Figure 3.6: Testing Relative Intensity (RI) alarm function 46

Figure 3.7: Testing Pattern Informatics (PI) alarm function 47

Figure 3.8: Testing National Seismic Hazard Map (NSHM) alarm function 48

Figure 3.9: Results of testing relative to the RI reference model 53

Figure 4.1: Schematic Molchan diagram for N=10. 63

Figure 4.2: Geometric illustration of area skill score 64

Figure 4.3: Comparing exact and approximate ASS probability density 71

Figure 4.4: Illustrative alarm function, shown here in continuous form. 73

Figure 4.5: Hypothetical predictability experiments 74

Figure 5.1: Fixed test period experiment schematic 88

Figure 5.2: Example of fixed test period predictability experiment 89

Figure 5.3: Temporal evolution of fixed test period area skill scores 90

Figure 5.4: Fixed test period area skill scores (σ = 30, 1000 km) 92

Figure 5.5: Fixed test period optimal lengthscale and misfit. 93

Figure 5.6: Growing learning period experiment schematic 96

Figure 5.7: Temporal evolution of growing test period area skill scores. 98

Figure 5.8: Low predictability in growing test period experiment. 99

Figure 5.9: Growing learning period optimal lengthscale and misfit 100

Figure 5.10: Moving learning period experiment schematic. 102

Figure 5.11: Moving learning results relative to uniform reference model 103

Figure 5.12: Low predictability in moving learning experiment 103

Figure 5.13: Moving learning period optimal lengthscale and misfit 105

Figure 5.14: Optimized prospective forecast 107

ABSTRACT

Earthquake prediction is one of the most important unsolved problems in the geosciences. Over the past decade, earthquake prediction research has been revitalized, and predictability experiments are currently active worldwide. In considering these experiments, a number of issues related to prediction evaluation are vital: a detailed experiment specification, the measure of success to be used, and a choice of appropriate reference model(s). Here, we address each of these, with an emphasis on testing prospective earthquake predictions.

We consider a general class of earthquake forecasts for which the forecast format allows a binary interpretation; that is, for any given interval of space and time, we can infer whether or not an earthquake of a given size is expected. This generalization allows us to test deterministic and probabilistic forecasts and compare the results; furthermore, the tests are easily understood because they are essentially the sum of many yes/no questions. As an introduction to binary performance measures and their wide applicability, we considered Reverse Tracing of Precursors (RTP), a recent earthquake prediction algorithm intended to forecast damaging earthquakes. We introduce and analyze several methods for measuring predictive performance but concede that the RTP experiment results are likely unstable due to the small number of earthquakes occurring to date.

In the context of an experiment with three 10 year seismicity forecasts—Relative Intensity, Pattern Informatics, and National Seismic Hazard Map—we introduce the area skill score, a measure of success derived from the Molchan diagram. Using this experiment and applying approaches from statistical hypothesis testing, we illustrate the

importance of choosing an appropriate reference model, and show that added model complexity does not necessarily yield a significant improvement in predictive skill.

Having demonstrated the use of the area skill score as a performance metric, we explore its statistical properties and the related computational procedures in some detail. Based on this work and the previous experiment results, we used the area skill score to explore the evolution of regional seismicity and optimize simple forecast models.

CHAPTER ONE: INTRODUCTION & OVERVIEW

The problem of earthquake prediction has generated a vast body of literature including books (e.g., Press 1965, Wyss & Dmowska 1997, Keilis-Borok & Soloviev 2003), special volumes (e.g., Wyss 1991, Knopoff *et al*. 1996, Hough & Olsen 2007), a multitude of peer-reviewed papers, and even, at one time, a dedicated journal (Rikitake 1982). The majority of this work is aimed at what Jordan (2006) denoted **operational earthquake prediction**: the ability to foretell large, damaging earthquakes with sufficient lead time to reduce potential losses. In particular, a search for precursors to large earthquakes has dominated prediction research programs. These precursors can be broadly divided in two categories: **physical precursors** based on direct observations—for example, radon emission (Haukkson 1981), tidal loading (Yin *et al.* 2000), seismic velocity ratios (Feng, 1975), various hydrologic signals (Roeloffs 1987), etc.; and **statistical seismicity patterns** based on earthquake catalog analysis—for example, quiescence (Wyss & Habermann 1988), temporal changes in Gutenberg-Richter *b*-value (Enescu & Ito 2001), accelerating moment release (Bowman *et al*. 1998), and a number of other patterns (see, e.g., Eneva & Ben-Zion 1997, Keilis-Borok 2002). Unfortunately, in both instances, most of the publications have relied on selective case studies or retrospective experiments that declare that a particular large earthquake could have been successfully predicted because a signal was identified following the earthquake. Certainly such statements are less convincing than results from prospective experiments, where precise forecasts are made in advance of any observations.

Recently, there has been a renewed vigor in earthquake prediction research, with

an emphasis on earthquake predictability through rigorous, systematic experimentation. For example, a large-scale experiment has been developed by the Southern California Earthquake Center (SCEC) Regional Earthquake Likelihood Models (RELM) working group (Field 2007, and references therein), and this experiment is now underway within the international Collaboratory for the Study of Earthquake Predictability (CSEP) (Jordan *et al*., in prep.). Several time-invariant forecasts that estimate seismicity rates in California over a five-year period were submitted and observations are currently being collected (Schorlemmer & Gerstenberger 2007).

To be properly specified, a predictability experiment should include descriptions of the following: the set of earthquakes being predicted and the earthquake catalog to use for verification; the geographical region of interest; the time period of interest; and any experiment- or forecast-specific rules. We call the set of earthquakes being predicted the **target earthquakes**, and this set is typically defined as all earthquakes above some minimum magnitude, although it may be specified in terms of scalar moment and also include focal mechanism information. If using magnitude information, a statement of the magnitude scale to be used is of particular importance, as some catalogs mix magnitude scales and this could lead to ambiguous results. The geographical region of interest is the **study region** (sometimes **natural laboratory**) and its specification should be unambiguous—for example, "the region between 32.5°N and 37.5°N, -121°W and -114°W" rather than "southern California"—and may include depth information. The time period of interest is the **test period** and is characterized by a beginning and ending date. Examples of forecast and experiment-specific rules include discretization parameter values, i.e., size of spatial grid, and declustering methods.

Vital to these experiments—and for that matter, to any rigorous predictability experiment—is the need for appropriate evaluation techniques, methods which quantify the predictive skill of a forecast. To be effective, performance measures should be flexible, easily explained and well-understood, and should not produce counter-intuitive results.

Flexibility refers to the breadth of experiments and forecasts that can be evaluated. If a measure is mathematically sound but does not apply to any existing models, it cannot be practically used. Furthermore, if a measure has extremely narrow requirements for its use and may only be applied to a small set of forecasts, it will not be useful for comparisons.

A performance measure that is too complex to apply or too difficult to interpret is not useful. Particularly in the case of earthquake predictability experiments, one ought to be able to interpret results in physical terms. For a measure to be considered well-understood, one should be able to determine the measure's distribution under certain conditions; for example, how does the measure behave in experiments with very few target earthquakes, and how does this compare to experiments with many target earthquakes? How well can a forecast do by chance and what is the distribution of the measure for random guessing? Along these same lines, one must consider carefully the assumptions and approximations inherent in a given performance measure.

The final requirement—to be consistent with intuition—is difficult to define explicitly but, for example, if one can obtain a statistically significant value of a performance measure with a very naïve approach, it is unlikely to characterize predictive

skill effectively. If, on the other hand, one can never obtain the optimal value of a performance measure, the measure is unlikely to distinguish between forecasts of differing skill.

Coupled with the importance of effective performance measures is the choice of appropriate reference models. A poor choice of reference model can yield results that make a forecast seem more powerful than it is, or as Rhoades & Evison (1984) put it: "The enhanced significance that a prediction might seem to have could be illusory, for it might state in different terms no more than was known before." In earthquake prediction experiments, it is easy to illustrate the choice of a poor reference model. Consider a reference forecast that suggests earthquakes are equally likely everywhere in space and time; we call this the **uniform reference model**. One of the first order observations of seismicity is clustering in space and time. While the form of clustering may not be known exactly, a forecast that includes some reasonable clustering component is likely to significantly outperform the uniform reference model.

The issue of reference models has received some little attention (e.g., Michael 1997, Marzocchi *et al*. 2003, Stark 1997), but it seems that emphasizing the role of the reference model is a prudent approach to advancing our understanding of earthquake predictability. By starting with a simple reference model and an aim to iteratively improve it, we can determine which model enhancements have some effect on performance and which can be disregarded. This will be most effective in prospective testing, where there is no chance for bias or data-fitting, unintentional or otherwise. The importance of prospective testing cannot be over-emphasized. Rhoades & Evison (1989) proposed a procedure by which earthquake precursors should be rigorously analyzed,

including application of the prediction algorithm to an independent data set. They followed this procedure and later published two papers that are singular in the literature (Evison & Rhoades 1993, 1997): they report the failure of their algorithm to significantly outperform the reference model! However discouraging this might be, prospective testing should yield a clear picture of what it is that we understand about earthquake processes, and what it is that we still have to learn.

With these thoughts as motivation, this dissertation addresses issues related to measuring skill in earthquake predictability experiments.

In Chapter 2, we consider Reverse Tracing of Precursors, a particularly complex earthquake prediction algorithm that combines several seismicity patterns. This method differs from the RELM forecasts because it is alarm-based—it does not provide estimates of future seismicity, but rather produces binary statements about the occurrence of target earthquakes. We describe several potential measures of success for alarm-based predictions and illustrate some of the difficulties involved in performing a rigorous, prospective predictability experiment. We also provide an independent evaluation of the results of this ongoing experiment.

In Chapter 3, we consider a very general class of earthquake forecasts based on what we call an **alarm function**, from which alarm-based forecasts can be derived. In this context, we introduce a general performance measure called the area skill score. As an illustration of the area skill score performance measure, we perform a comparative evaluation of three distinct earthquake forecasts: the Pattern Informatics forecast of Rundle *et al*. (2002), the Relative Intensity method of Rundle *et al*. (2002), and the rate

model of the 2002 national seismic hazard map (Frankel 2002). We emphasize the importance of choosing an appropriate reference model and find that, for the pseudo-prospective experiment under consideration, a simple forecast based only on the locations of past earthquakes is difficult to beat.

In Chapter 4, we explore the area skill score's statistical properties in detail and consider the relevant approximations and simulations. While the measure itself is applicable to experiments that are continuous in space and time, we pay particular attention to the case of spatial discretization commonly employed in predictability experiments.

In Chapter 5, we use the area skill score measure for forecast optimization, with an aim to produce a reference model for the RELM forecast experiment. We provide exact analytic solutions for three distinct smoothing kernels; these solutions should aid in comparing different smoothing approaches. We consider a class of simple smoothed seismicity models and three sets of retrospective experiments that explore forecast optimization. We present in detail the results of these experiments and the accompanying interpretation leads us to offer a prospective earthquake forecast for the next five years of seismicity in California.

In Chapter 6, we offer some brief concluding remarks and details of ongoing and potential future work.

CHAPTER TWO:
Formal evaluation of the Reverse Tracing of Precursors earthquake prediction algorithm

Abstract

In 2003, prospective application of the Reverse Tracing of Precursors (RTP) earthquake prediction algorithm began in Japan, California, the Eastern Mediterranean, and Italy; a testing region in the North Pacific was added in 2006. RTP is a pattern recognition algorithm that uses earthquake catalog data to declare alarms which indicate that a moderate to large earthquake is expected in the subsequent months. The spatial extent of the alarms is highly variable and each alarm typically lasts nine months, although alarms may be extended in time and space. RTP garnered the attention of the public and the seismological community when the first two alarms were declared successful. In this chapter, we examine the record of alarms and outcomes since testing began, and we explore a number of measures to characterize the performance of RTP. For each alarm, we estimate the "prior probability" of the corresponding target earthquake using historical seismicity rates. Formally, we do not include the first two successful alarms in our evaluation because they were not fully specified in advance of the earthquakes. The contingency table measures we consider here indicate that RTP is not significantly different from a naïve method of guessing based on the historical rates of seismicity. Likewise, the Molchan diagram analysis indicates that RTP performance to date is not statistically significant. Given the small sample of earthquakes considered, however, the Molchan diagram analysis is unstable; if the first two alarms are considered, RTP does look to be significantly successful. The RTP investigators have dutifully

specified and documented the RTP alarms, and we discuss the benefits of integrating RTP into Collaboratory for the Study of Earthquake Predictability (CSEP) testing center.

2.1 Reverse Tracing of Precursors

Keilis-Borok *et al*. (2004) presented a pattern recognition algorithm called Reverse Tracing of Precursors (RTP), intended to predict large earthquakes in a time window of nine months. RTP comprises two distinct steps: short-term chain recognition and intermediate-term chain confirmation. The first step consists of grouping all events in a declustered regional catalog into **chains**, which are comprised of **neighbors**. Two events are neighbors if one follows the other by less than t_0 days and has an epicenter within r km of the first event, where

$$r = 10^{c(m_{\min}-2.5)} \tag{2.1}$$

Here c is a region-specific model parameter and $m_{\min}$ is the smaller magnitude of the two events. Upon decomposing the entire declustered catalog into chains, RTP determines which chains have 1) more than k_0 events, 2) a spatial extent greater than l_0 km, and 3) proportion $\gamma > \gamma_0$, where

$$\gamma = \frac{N(3.5)}{N(m_c) - N(3.5)} \tag{2.2}$$

Here $N(m)$ denotes the number of events in the chain with magnitude greater than or equal to m, m_c is a region-specific minimum magnitude, and k_0, l_0, and γ_0 are model parameters. Chains that meet these three criteria are considered to exhibit a short-term precursor to a target earthquake. The spatial domain of the chain is defined as the union

of circles of radius *R* centered on each epicenter in the chain, where *R* is a model parameter taken to be either 50 km or 100 km. The second step of RTP seeks to confirm the short-term precursor by searching for intermediate-term precursors within each chain's spatial domain. In this step, values of 8 precursory functions (listed in Table 2.1) are computed. These functions capture four types of precursory behavior: increase in seismic activity, increase of clustering, increase of correlation length, and transformation of the Gutenberg-Richter relation. They have been used in previous prediction studies with limited success (Keilis-Borok 2002).

Table 2.1: RTP Intermediate Term precursors

Precursor	Characterizes	Formulation
Activity	Increase of activity	$N(t) = \sum_{(t-s)\le t_k<t} 1$
Sigma	Increase of activity	$\Sigma(t) = \sum_{(t-s)\le t_k<t} 10^{m_k}$
Acceleration of average magnitude	Increase of activity	$\frac{d^2M(t)}{dt^2} = \frac{\sum_{(t-s)\le t_k<t} m_k}{N(t)} - \frac{\sum_{(t-2s)\le t_k<(t-s)} m_k}{N(t-s)}$
Acceleration of average inter-event time	Increase of activity	$\frac{d^2N(t)}{dt^2} = \frac{\sum_{(t-s)\le t_k<t} \frac{1}{t_k - t_{k-1}}}{N(t)} - \frac{\sum_{(t-2s)\le t_k<(t-s)} \frac{1}{t_k - t_{k-1}}}{N(t-s)}$
Swarm	Increase of clustering	$W(t) = \frac{A_r^{\cap}}{\pi r^2}$
B-micro	Increase of clustering	$b_\mu(t) = \sum_{(t-s)\le t_k<t} \sum_l \frac{10^{m_{kl}}}{10^{m^*}}$
Accord	Increase of correlation length	$Acc(t) = \frac{A_r^{\cup}}{\pi r^2}$
Gamma	Transformation of Gutenberg-Richter relation	$\gamma(t) = \frac{2}{N} \sum_{(t-s)\le t_k \le t; m_k \ge m_{1/2}} (m_k - m^*)$

Table 2.1 RTP precursor definitions, from Shebalin (written communication). The following notation is used to formulate the precursors: $m_{1/2}$ is the median magnitude of a set of earthquakes; $A_r^{\cap}$ is the "total area of intersections of two or more circles of radius *r* in the same sequence"; $A_r^{\cup}$ is the "total area of the union of the circles of radius r centered at the epicenters in the sequence that occurred within interval ($t - s$, t)"; and m_{kl} is the magnitude of the l^{th} aftershock during the first two days after mainshock *k*.

The values of the intermediate-term precursory functions are computed in sliding time windows of length *s*, beginning *T* years before the first event in the chain. For each chain, eight different combinations of *s* and *T* values are used. In total, for each short-term chain of interest, the values of 64 intermediate term precursory functions are computed. Threshold values are defined for each of these 64 precursors. If the threshold is exceeded, this precursor contributes a single positive **vote**. If the number of positive votes exceeds a region-dependent meta-threshold, an alarm that lasts 9 months from the end of the chain's formation is declared within the chain's spatial domain. Alarms can be extended if a chain continues to grow in space and/or time. Declaration of an alarm indicates that one or more target earthquakes—those earthquakes that are targeted for prediction in a given region—are expected to occur within the spatial domain of the alarm within 9 months.

The intermediate-term patterns used in RTP have been employed previously and seek to capture fluctuations in seismicity without being supported by a specific physical mechanism. In contrast, the short-term component of RTP is based on a seismicity pattern that identifies a rapid increase of earthquake correlation range. This pattern was found in seismicity models (Gabrielov *et al*. 1999, Gabrielov *et al*. 2000) and in regional earthquake observations (Shebalin *et al*. 2000, Zaliapin *et al*. 2000). One physical interpretation of this pattern is that fault network elements that are geographically distant begin to interact as they approach a critical state, at which point several fault elements rupture simultaneously and produce a large earthquake. The increase in correlation range is thought to highlight a coalescing instability in the fault network and may represent the range of lengthscales involved in earthquake processes. The ideas that inform RTP have

been explored in a more theoretical context by researchers in the statistical physics community, particularly with respect to critical behavior in complex systems (Blanter & Shnirman 1997, Sornette 2000, Rundle 2003).

Since the summer of 2003, RTP has been applied in an ongoing prospective prediction experiment. Beginning in 2004, RTP alarms were disseminated by email and archived at http://www.igpp.ucla.edu/prediction/rtp. In this chapter, we provide a current evaluation of the experiment results. The geographical regions being studied using RTP are Japan, California, Italy, the Eastern Mediterranean, and the North Pacific. The study regions are shown in Figures 2.1 through 2.5 and the details of each study region are listed in Table 2.2.

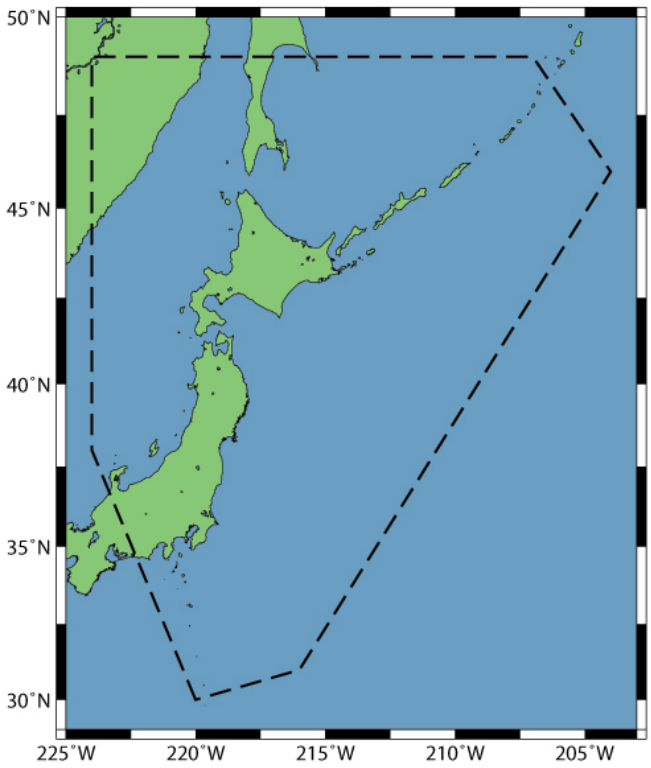

Figure 2.1 Japan natural laboratory considered in this analysis. The dashed polygon delineates the study region.

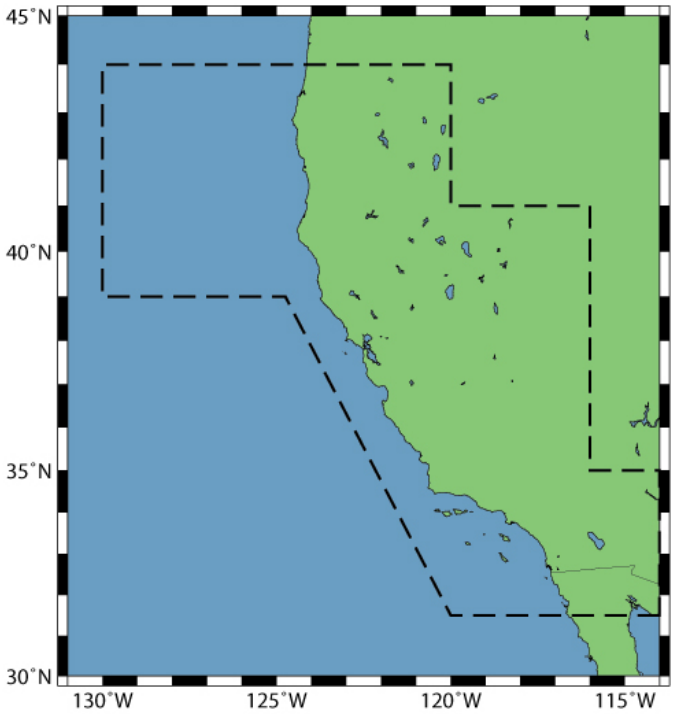

Figure 2.2 California natural laboratory considered in this analysis. The dashed polygon delineates the study region.

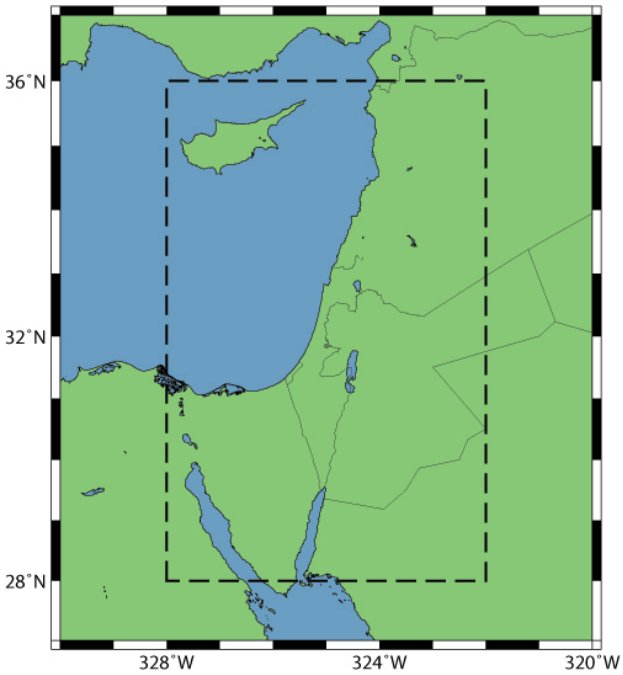

Figure 2.3 Eastern Mediterranean natural laboratory considered in this analysis. The dashed polygon delineates the study region.

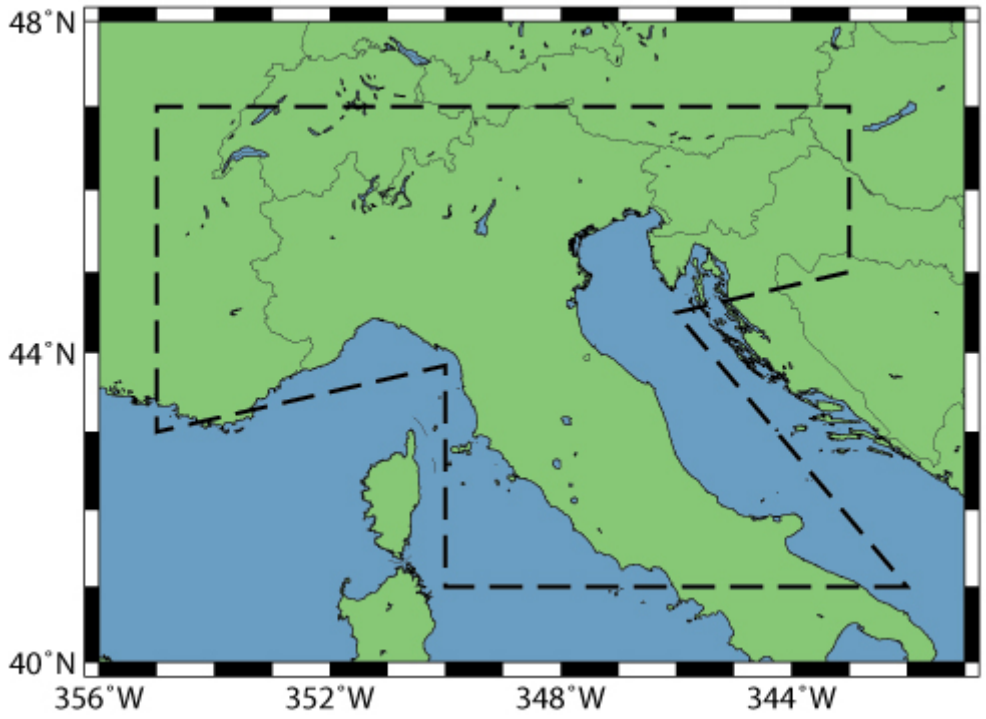

Figure 2.4 Italy natural laboratory considered in this analysis. The dashed polygon delineates the study region.

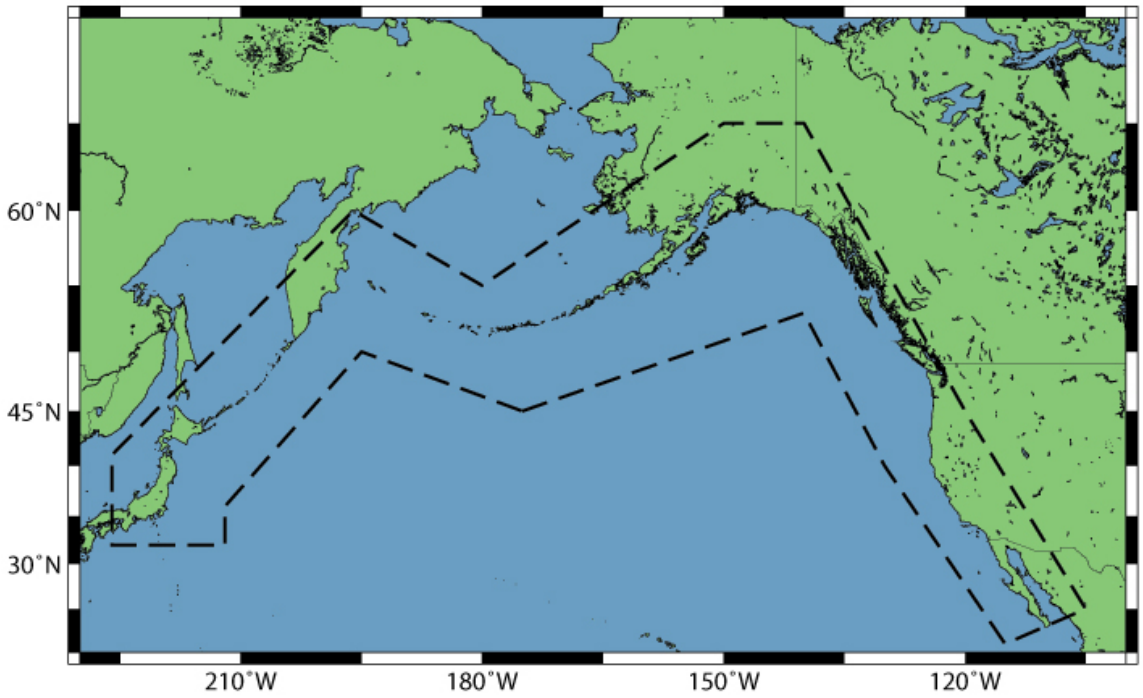

Figure 2.5 North Pacific natural laboratory considered in this analysis. The dashed polygon delineates the study region.

Table 2.2: RTP study regions

Name	Catalog of interest	Polygon enclosing study region
Japan	JMA 1923 – 31 Dec 2002 CMT 1977 – 31 Dec 2002	(30,140) , (38,136), (49,136), (49,153), (46,156), (31,144)
California	ANSS 1932 – 31 Dec 2002	(31.5,-114), (31.5,-120), (39,-124.75), (39,-130), (44,-130), (44,-120), (41,-120), (41,-116), (35,-116), (35,-114)
Eastern Mediterranean	CMT 1977 – 31 Dec 2002	(28,32), (36,32), (36,38), (28,38)
Italy	CMT 1977 – 31 Dec 2002	(41,18), (41,10), (43.84,10), (43,5), (47,5), (47,17), (45,17), (44.5,14)
North Pacific	CMT 1977 – 31 Dec 2002	(-175,45), (-140,53), (-130,40), (-115,21), (-105,25), (-120,45), (-140,65), (-150,65), (180,55), (164,60), (134,41), (134,32), (148,32), (148,36), (165,50)

Table 2.2 List of study regions and corresponding earthquake catalogs. In the case of Japan, the JMA catalog is used for Alarms 1, 4, and 10; this was the catalog used to determine those alarms. For subsequent alarms in the Japanese region, the CMT catalog is used.

RTP has generated twenty-six alarms; the details of each are reported in Table 2.3. To date, eight target earthquakes have been observed during this experiment and these are listed in Table 2.4. We have excluded some of the alarms and target earthquakes from this study; a detailed explanation follows.

Table 2.3: RTP alarms and outcomes to date

#	Region	Alarm Start	Alarm End	Magnitude	Prior (%)	Result
1[♠]	Japan	27 Mar 2003	27 Nov 2003	$M_{JMA} \geq 7.2$[1]	75.1	Hit
2[♠]	California	5 May 2003	27 Feb 2004	$M_{ANSS} \geq 6.4$[2]	2.3	Hit
3	California	13 Nov 2003	5 Sep 2004	$M_{ANSS} \geq 6.4$	9.1	False alarm
4	Japan	8 Feb 2004	8 Nov 2004	$M_{JMA} \geq 7.2$	32.0	False alarm
5	Italy	29 Feb 2004	29 Nov 2004	$M_w \geq 5.5$	2.1	False alarm[3]
6	California	14 Nov 2004	14 Aug 2005	$M_{ANSS} \geq 6.4$	4.1	False alarm
7	California	16 Nov 2004	16 Aug 2005	$M_{ANSS} \geq 6.4$	2.8	False alarm
8	Italy	31 Dec 2004	1 Oct 2005	$M_w \geq 5.5$	8.2	False alarm
9	Italy	6 May 2005	6 Feb 2006	$M_w \geq 5.5$	10.3	False alarm

[♠] Alarm reported here but not used in analysis; see text for details.
[1] The magnitude range was not specified in advance of the target earthquake; see text for details.
[2] The magnitude range was not specified in advance of the target earthquake; see text for details.
[3] Magnitude ambiguously specified in original alarm statement; see text for details.

Table 2.3, Continued

10♦	Japan	2 Jun 2005	2 Mar 2006	$M_{JMA} \geq 7.2$	61.2	Hit[4]
11	California	17 Jun 2005	17 Mar 2006	$M_{ANSS} \geq 6.4$	7.1	False alarm
12	California	18 Mar 2006	18 Sep 2006	$M_{ANSS} \geq 6.4$	2.1	False alarm
13	California	24 Mar 2006	24 Dec 2006	$M_{ANSS} \geq 6.4$	10.1	False alarm
14	California	2 May 2006	2 Feb 2007	$M_{ANSS} \geq 6.4$	9.2	False alarm
15	Italy	2 May 2006	3 Feb 2007	$M_w \geq 5.5$	8.3	False alarm
16	Japan	11 May 2006	11 Feb 2007	$M_w \geq 7.2$	16.1	False alarm
17	California	23 Sep 2006	23 Jun 2007	$7.6 \geq M_{ANSS} \geq 6.6$	0.3	False alarm
18	Japan	30 Sep 2006	30 Jun 2007	$M_w \geq 7.2$	15.9	Hit
19	N. Pacific	28 Oct 2006	28 Jul 2007	$8.1 \geq M_w \geq 7.1$	0	False alarm
20	California	17 Jan 2007	17 Oct 2007	$7.1 \geq M_{ANSS} \geq 6.1$	2.1	False alarm
21	California	3 May 2007	28 Jan 2008	$6.9 \geq M_{ANSS} \geq 5.9$	8.0	False alarm
22	California	18 Oct 2007	14 Jan 2008	$6.8 \geq M_{ANSS} \geq 5.8$	0.3	False alarm
23♦	N. Pacific	29 Jul 2007	28 Jan 2008	$8.1 \geq M_w \geq 7.?$[5]	0.0	False alarm?[6]
24	N. Pacific	24 Aug 2007	24 May 2008	$8.2 \geq M_w \geq 7.2$	1.2	False alarm
25♦	California	29 Jan 2008	26 Sep 2008	$6.7 \geq M_{ANSS} \geq 5.7$	6.3	—
26♦	California	14 Apr 2008	14 Jan 2009	$7.7 \geq M_{ANSS} \geq 6.7$	0.3	—

Table 2.3 List of all RTP alarms to date, including the prior probability of a target earthquake within the alarm region and the outcome of the alarm. Alarms are listed in chronological order according to their start date. The epicenters of each alarm region are listed in Appendix A.

Table 2.4: Target earthquakes

#	Origin Time	Magnitude	Latitude (degrees)	Longitude (degrees)	Outcome
1♦	25 Sep 2003	$M_w = 8.3$	42.21	143.84	Hit
2♦	25 Sep 2003	$M_w = 7.3$	41.75	143.62	Hit
3♦	22 Dec 2003	$M_{ANSS} = 6.5$	35.7002	-121.0973	Hit
4	15 Jun 2005	$M_{ANSS} = 7.2$	41.292	-125.953	Miss

[4] Due to delay of catalog data, the alarm was declared after a satisfactory target earthquake, see text for details.

[5] In the declaration of this alarm, the magnitude range was listed as $M_w \geq 7.2$; in a supplementary technical file, the magnitude range was given as $8.1 \geq M_w \geq 7.1$.

[6] An earthquake occurred within the alarm region on 19 Dec 2007 and was announced as having magnitude M_w=7.2. It has since been downgraded to magnitude M_w=7.1. Given the target magnitude ambiguity mentioned in Footnote 5, the assignment of this alarm to one of the contingencies is difficult. See text for details.

♦ Earthquake reported here but not used in analysis; see text for details.

Table 2.4, Continued

5	17 Jun 2005	$M_{ANSS} = 6.6$	40.773	-126.574	Miss
6	16 Aug 2005	$M_w = 7.2$	38.24	142.05	Miss[7]
7	15 Nov 2006	$M_w = 8.3$	46.71	154.33	Hit
8	13 Jan 2007	$M_w = 8.1$	46.17	154.80	Hit

Table 2.4 Target earthquakes considered in this study. Events with magnitude type M_w are taken from the global CMT catalog.

2.2 Alarms and earthquakes excluded from this study

In this study, we do not consider a number of the alarms listed in Table 2.3. We exclude Alarm 1 and Alarm 2 because both were improperly specified: neither alarm announcement included an explicit statement of target earthquake magnitude or alarm time window, but rather included vague phrases such as "preparing for a major earthquake" (Aki *et al*. personal communication, Shebalin *et al*. 2003, Shebalin *et al*. 2004). To be fair, we do not consider the three target earthquakes corresponding to Alarm 1 and Alarm 2 (Earthquakes 1-3 in Table 2.4). Alarm 10 is also excluded from this study; data from the earthquake catalog used to determine chains was delayed, and therefore the alarm was not declared until after a qualifying target earthquake occurred (Shebalin, personal communication). Accordingly, we also exclude this earthquake from our study (Earthquake 6 from Table 2.4). Alarm 23 is a particularly difficult case. In the official alarm announcement, the target magnitude was listed as $M_w \geq 7.2$. On 19 December 2007, an earthquake occurred within this alarm region and was initially estimated as having moment magnitude $M_w = 7.2$ (Shebalin, personal communication). In the

[7] Due to delay of catalog data, this event was technically missed because no alarm was declared prior to the origin time; see text for details.

subsequent weeks, however, its magnitude was downgraded to moment magnitude M_w = 7.1. Complicating the case, a technical support document that accompanied the alarm declaration listed the target earthquake magnitude as $8.1 \geq M_w \geq 7.1$. Due to the problematic nature of this alarm and the earthquake that followed, we do not consider this alarm or the earthquake. The final two alarms listed in Table 2.3 are also excluded from this study as they have not yet expired.

2.3 Contingency table analysis of RTP

RTP alarms belong to the realm of binary prediction and binary outcome. In this regime, there are four possible results: if a positive prediction is made and the outcome is positive, this is a **hit**; if a positive prediction is made and the outcome is negative, this is a **false alarm**; if a negative prediction is made and the outcome is positive, this is a **miss**; and if a negative prediction is made and the outcome is negative, this is a **correct negative**. In terms of RTP, a positive prediction corresponds to the declaration of an alarm and a positive outcome is the occurrence of one or more target earthquakes. Each target earthquake that occurs within an alarm is considered a hit, and all target earthquakes occurring outside alarms are counted as misses. Typically, these results are grouped into a **contingency table**, a matrix containing the counts of each result for a given experiment. For each RTP alarm, we assign the appropriate result in the final column of Table 2.3. We also list the result for each observed target earthquake in Table 2.4. Using these results, and accounting for the exclusions outlined in the previous section, we present a corresponding contingency table in Table 2.5

Table 2.5: RTP contingency table

		Did a target earthquake occur?	
		Yes	**No**
Was an alarm declared?	**Yes**	2 hits	19 false alarms
	No	2 misses	0 correct negatives

Table 2.5 Contingency table for the RTP alarms listed in Table 2.3.

Figure 2.6 shows an example RTP hit and Figure 2.7 shows an example RTP false alarm.

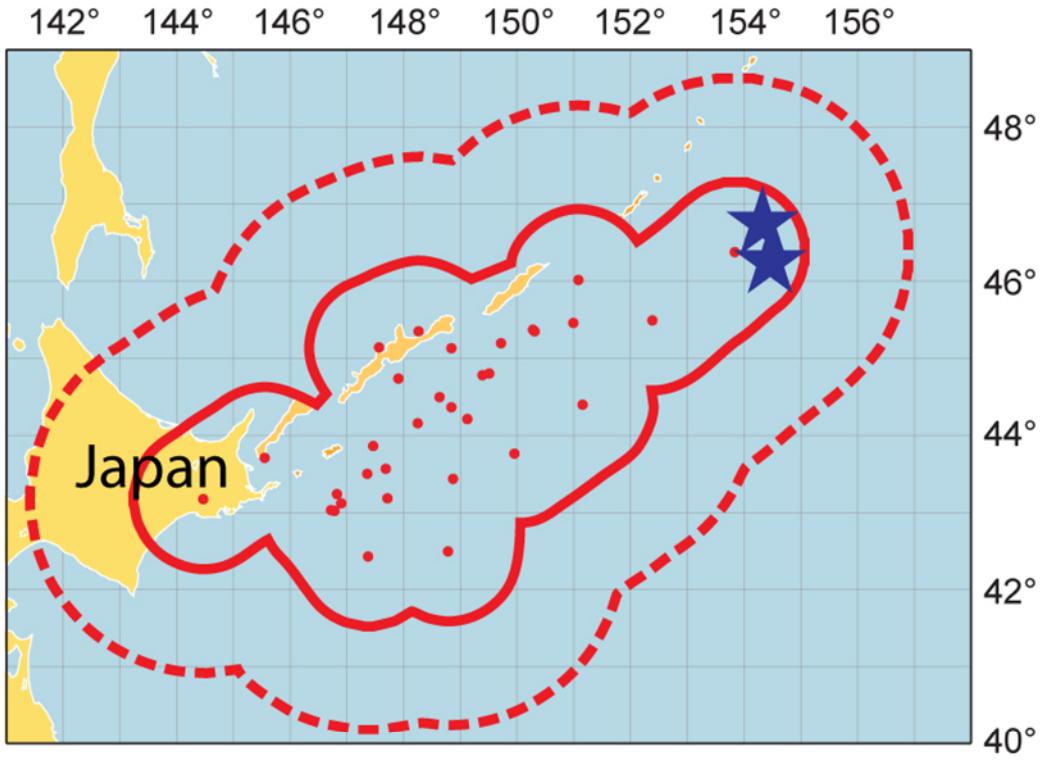

Figure 2.6 Example RTP hit (corresponds to Alarm #18 in Table 2.3 and earthquakes 7 and 8 in Table 2.4); in fact, this is a double hit. Red points are the earthquakes forming the chain, solid red line indicates alarm region with R = 50 km, dashed line corresponds to alarm region with R = 100 km.

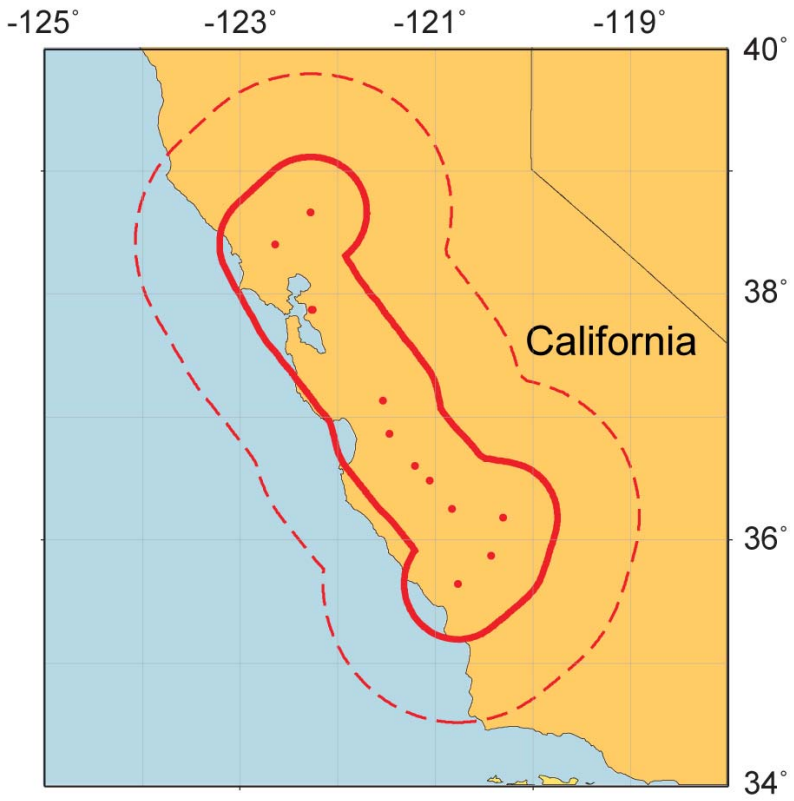

Figure 2.7 Example RTP false alarm (corresponds to Alarm #20 in Table 2.3). Red points are the earthquakes forming the chain, solid red line indicates alarm region with R = 50 km, dashed line corresponds to alarm region with R = 100 km.

Research related to weather forecasting and medical testing has yielded several measures of skill based on the contingency table (Joliffe & Stephenson 2003). No single measure is suited for all types of experiments. Some measures permit an optimal score to be obtained by a simple strategy. For example, if one never declares an alarm, the false alarm rate will be optimized. Some measures consider false alarms to be as important as misses, e.g., the Critical Success Index combines false alarms and misses. Considered alone, none of these measures seem ideal, and thus measures are often reported jointly. To evaluate the performance of RTP, we have chosen some common contingency table measures and computed RTP's scores; these are reported in Table 2.6. Each of these scores, when compared to the optimal score for each measure, indicates that RTP has not

been very effective at predicting target earthquakes in this experiment. To further illustrate this point, we consider scores from an operational weather forecasting algorithm.

Table 2.6: RTP contingency tables measures of success

Name of measure	Definition	Range	RTP Value
Hit rate, *H*	$\frac{a}{a+c}$	[0,1]	0.5
False alarm rate, *F*	$\frac{b}{b+d}$	[1,0]	1
False alarm ratio, *FAR*	$\frac{b}{a+b}$	[1,0]	0.90
Proportion correct, *PC*	$\frac{a+d}{n}$	[0,1]	0.09
Peirce's skill score, *PSS*	$\frac{ad-bc}{(b+d)(a+c)}$	[-1,1]	-0.5
Critical success index, *CSI*	$\frac{a}{a+b+c}$	[0,1]	0.09
Gilbert skill score, *GSS*	$\frac{a-e}{a+b+c-e}, e=\frac{(a+c)(a+b)}{n}$	[-1/3,1]	-0.09
Yule's *Q*	$\frac{ad-bc}{ad+bc}$	[-1,1]	-1

Table 2.6 Various contingency table measures for RTP alarms, using contingencies listed in Table 2.5 and RTP alarms in Table 2.3. Here, the measure ranges are listed in order from least optimal to optimal.

The Aviation Branch Forecast Verification Section of the U.S. National Oceanic and Atmospheric Administration maintains a system in which contingency table scores are automatically computed and archived for various meteorological forecasts (Mahoney *et al*. 1997, Loughe *et al*. 2001, Mahoney *et al*. 2002). For example, Airman's Meteorological Advisories (AIRMETs) are issued daily by the Aviation Weather Center to warn of potentially hazardous weather and are compared with Meteorological Aviation

Reports (METARs) collected by the U.S. National Weather Service to assess their accuracy. AIRMETs have been tested since June 1999 and have obtained an average hit rate of 0.68, average false alarm ratio of 0.64, and average critical success index of 0.31, all of which are far better than the RTP measures reported in Table 2.6.

It is important to note, however, that only a small number of observed target earthquakes are included in this analysis and therefore these contingency table measures are dominated by the large number of false alarms. A further caveat to using contingency table measures to evaluate alarm-based earthquake predictions is the lack of correct negatives. If an algorithm was required to make explicit predictions on a regular schedule, it is likely that "negative" predictions—indicating that no target earthquake was expected—would be made. In this case, it would be trivial to determine the number of correct negatives. In practice, however, alarm-based algorithms such as RTP generally only declare positive alarms, and there is no unambiguous method for inferring negative alarms. Therefore, RTP will never obtain correct negatives. In the following section, we consider other methods for evaluating RTP that do not require the explicit statement of negative alarms.

2.4 Molchan diagram analysis of RTP

The Molchan diagram (Molchan 1991, Molchan & Kagan 1992) is a useful diagnostic because it captures two intuitive measures: miss rate, ν—the proportion of target earthquakes falling outside all alarms—and the fraction of space-time occupied by alarm, τ. Miss rate is the complement of hit rate:

$$\nu = 1 - H = \frac{c}{a + c} \tag{2.3}$$

Here, a is the number of hits and c is the number of misses. For this study, we define the fraction of space-time occupied by alarm as:

$$\tau = \frac{\sum_{i=1}^{x} N_i(m_j) t_i}{\sum_{j=1}^{y} N_j(m_j) t_j} \tag{2.4}$$

In the numerator, x is the number of alarms, $N_i(m_j)$ is the number of past epicenters with magnitude greater than m_j occurring within the spatial domain of the i^{th} alarm, and t_i is the duration of i^{th} alarm. In the denominator, y is the number of distinct study regions under consideration, $N_j(m_j)$ is the number of past epicenters with magnitude greater than m_j occurring in the j^{th} region, and t_j is the duration of the RTP experiment in the j^{th} region.

Figure 2.8 shows the Molchan diagram for RTP. Here, the dashed diagonal represents the average behavior of an unskilled prediction algorithm: one that essentially guesses based on the historical distribution of seismicity. The shaded region corresponds to statistical significance above 95% (see Chapter 3 for details). That the RTP (τ, ν) point does not fall within the shaded critical region indicates that RTP has not done significantly well in this experiment. We concede that this result is not a very stable one because it takes into account only four target earthquakes. For example, if we include the first two alarms in the analysis, the (τ, ν) point does fall within the corresponding critical region (Figure 2.9). Moreover, having only a single point on the Molchan diagram greatly limits an analysis of RTP's predictive skill. For example, it is not possible to know what value of τ

would be required for RTP to successfully predict all of the target earthquakes (i.e., $\nu = 0$). In the following chapter, we revisit the Molchan diagram in the context of many (τ, ν) points for a single prediction algorithm. In the next section, we continue the RTP analysis using simulations based on prior probabilities.

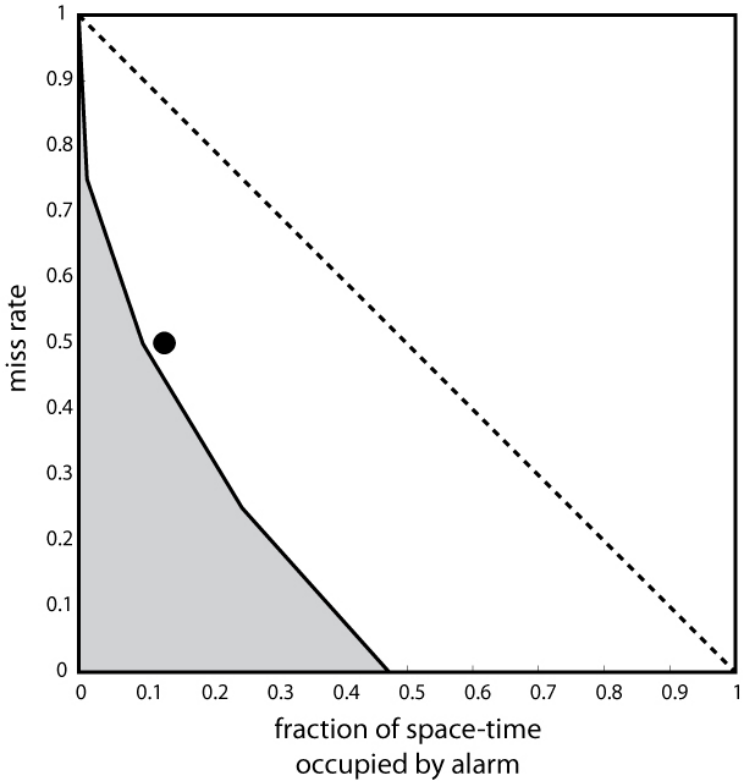

Figure 2.8 Molchan diagram for RTP experiment including 95% confidence bounds. This indicates that RTP has not shown statistically significant performance in this experiment.

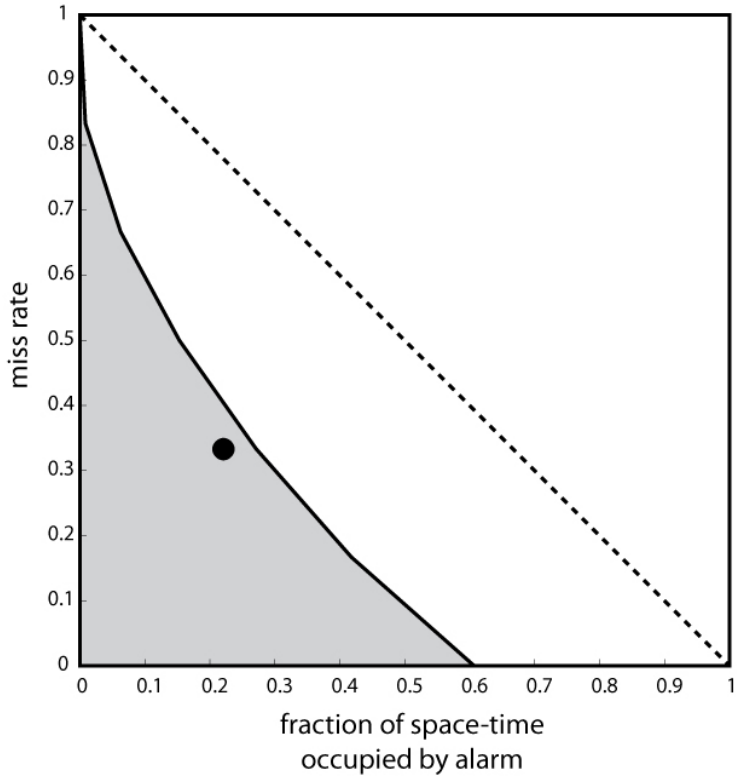

Figure 2.9 Molchan diagram for RTP experiment including 95% confidence bounds. In this case, we use leniency and consider the first two alarms as successes. Here, the diagram indicates that RTP **has** shown statistically significant performance in this experiment.

2.5 Prior probability analysis of RTP

Jackson (1996) outlined a simple procedure to estimate the skill of an alarm-based prediction algorithm. For each alarm and for each missed target earthquake, a prior probability is computed. For an alarm, the prior is the estimated probability of a target earthquake occurring within the alarm. For a missed earthquake, the prior is the estimated probability of a target earthquake occurring anywhere in the study region in a time window equal to the typical alarm. Once these priors are estimated, one can simulate a hit distribution in the following manner: for each prior, a random number is drawn from the uniform distribution on [0, 1], and if the random number is less than the prior, this corresponds to a hit. Iterating over this simulation procedure yields an

empirical hit distribution, where each simulation corresponds to an alarm set from random guessing based on the prior. To characterize the performance of a set of alarms, one can compare the simulated hit distribution with the observed number of hits.

Our preferred method of estimating the prior probability for alarm region A_i is to compute the historical rate of target earthquakes in the spatial domain of A_i and assume that earthquake rates follow a Poisson distribution. In this case, the probability of witnessing one or more target earthquakes in the magnitude range and geographic region specified by A_i, in a time window of duration t_i is given by

$$\tilde{p} = 1 - \exp\left(-r\left(m_1^i, m_2^i\right)t_i\right) \tag{2.5}$$

Here, $r\left(m_1^i, m_2^i\right)$ is the daily rate of epicenters with magnitude in $\left[m_1^i, m_2^i\right]$ occurring within the spatial domain of A_i and t_i is the duration (in days) of the alarm. In the case where $r\left(m_1^i, m_2^i\right) = 0$—that is, no target earthquakes have been observed in the past within the alarm space/magnitude domain—we compute $r\left(m_1^i - \Delta m, m_2^i - \Delta m\right)$, the rate of somewhat smaller earthquakes. We do this because earthquake catalogs are finite and a prior probability of zero seems unrealistic. To obtain the estimated rate of target earthquakes, we assume that the magnitude-frequency distribution follows a Gutenberg-Richter relation with a b-value of unity, and rescale:

$$\tilde{r}\left(m_1^i, m_2^i\right) = \frac{r\left(m_1^i - \Delta m, m_2^i - \Delta m\right)}{10^{\Delta m}} \tag{2.6}$$

For alarms where $r\left(m_1^i, m_2^i\right) = 0$, we let Δm=1 and use Equations 2.5 and 2.6 to estimate the prior probability. These probabilities are reported in Table 2.3.

To estimate the prior probability of a missed target earthquake, we consider the entire study region and modify slightly Equation 2.5:

$$\tilde{p} = 1 - \exp\left(-r\left(m_j, \infty\right)t_j\right) \tag{2.7}$$

Here, $r(m_j, \infty)$ is the daily rate of earthquakes in the j^{th} study region with magnitude greater than m_j —target earthquakes—and t_j is the typical duration (in days) of an alarm in the j^{th} study region; for this study, we take $t_j = 270$ for all study regions. For each of the two misses observed in the California study region, the prior probability is 30.5%.

We present the results of 10,000 iterations of the simulation procedure in Figure 2.10. To date, RTP has obtained 2 hits. From the figure, we note that approximately 30% of simulations that declare alarms randomly based on prior probabilities obtained 2 hits, and 65.4% of the simulations obtained at least 2 hits; this finding supports the contingency table result that RTP has not been exceptionally successful in this experiment.

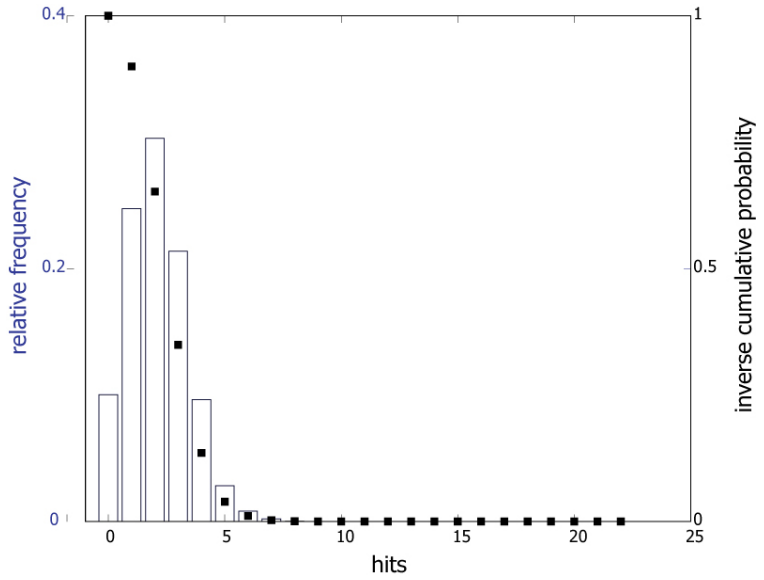

Figure 2.10 Results of 10,000 simulations based on prior probabilities. Bars indicate the relative frequency of the number of hits obtained by random guessing. Squares show the inverse cumulative probability—the fraction of simulations that obtained at least the number of hits.

2.6 Discussion and conclusions

Our analysis of RTP sometimes includes an implicit assumption regarding independence. One might claim that the miss rate we use in the Molchan diagram is incorrect because the second miss in California is obviously an aftershock of the first miss. We argue that physical dependence, beyond being ambiguous, is not relevant in our analysis. Rather, we are concerned with the independence of the ability to predict individual earthquakes. For example, because RTP missed the first earthquake in the California region does not require that it miss the second earthquake. Note that we treat the problem symmetrically: in the case of the two hits in Japan, it is likely that the second hit is physically related to the first earthquake, but it could happen that an algorithm successfully predicted the first earthquake and missed the second. Therefore, we grant RTP credit for both hits. In the next chapter, we consider time-invariant earthquake predictions, in which case the probabilities to predict two earthquakes that occur in the same spatial region are not independent; RTP is time-varying, however, and so we would consider these probabilities independent.

The RTP archive notes that a number of the alarms have been "near misses"; that is, target earthquakes have occurred just outside an active alarm or just outside a study region. As others have discussed, this is a difficulty in making binary predictions. To avoid such unfortunate cases, many forecasts are made in probabilistic, rather than binary terms. We consider a number of these forecasts and methods for evaluating them in the following chapters. Other near misses include improper or incomplete magnitude specification; this highlights the care that is required in making detailed, unambiguous

forecast statements.

Despite the negative results of the performance measures considered here, we emphasize the instability of these results due to the small sample of target earthquakes. Moreover, it seems that the contingency table measures in particular are affected by the form of predictions considered. For example, RTP could be reformulated to declare daily negative alarms in all space/time regions not covered by an RTP alarm. This would yield a large number of correct negatives and drastically change the values of the contingency table measures. Under this reformulation, however, the simulation-procedure and the Molchan diagram results would not change. To accelerate the evaluation of the RTP algorithm, it would be interesting to apply to other study regions, perhaps even on a global scale, and to smaller target earthquakes. The rate of false alarms may be decreased by including known fault structures; if no fault large enough to host a target earthquake exists in the RTP alarm, that alarm may be cancelled. Inclusion of faults might also allow one to shrink the spatial region of the alarms to only include areas surrounding major faults.

The investigators of RTP have made great efforts to officially share, archive, and evaluate their prospective earthquake alarms. It remains difficult, however, to reproduce the RTP results based only on RTP publications and alarm declarations. Additionally, it is difficult to say if the RTP algorithm or any of the myriad model parameters are changing in time, and this effectively makes RTP a moving target. The Collaboratory for the Study of Earthquake Predictability (CSEP) is designed to address these specific problems (Jordan 2006), and RTP can benefit from integration. CSEP testing centers maintain earthquake prediction model codes in an automated, reproducible environment; during

prospective prediction experiments, codes and parameters remain unchanged (Jordan *et al*., in prep.). Moreover, a number of evaluation techniques are in use or under development by CSEP, so the scientific community can monitor how well a given algorithm is doing. Integrating RTP into a CSEP testing center would reduce controversy and ambiguity related to alarm declaration and evaluation. The chains and alarms would be formally reproducible, as would be the results of any testing. RTP is not unique in this respect; many such complex prediction algorithms can benefit from integration in a CSEP testing center. One hope of the CSEP effort is that comparative testing of earthquake prediction algorithms will increase our understanding of the earthquake system, and RTP may yet provide some guidance in this direction.

CHAPTER THREE:
Testing alarm-based earthquake predictions

Abstract

Motivated by a recent resurgence in earthquake predictability research, we present a method for testing alarm-based earthquake predictions. The testing method is based on the Molchan diagram—a plot of miss rate and fraction of space-time occupied by alarm—and is applicable to a wide class of predictions, including probabilistic earthquake forecasts varying in space, time, and magnitude. A single alarm can be simply tested using the cumulative binomial distribution. Here we consider the more interesting case of a function from which a continuum of well-ordered alarms can be derived. For such an "alarm function" we construct a cumulative performance measure, the area skill score, based on the normalized area above trajectories on the Molchan diagram. A score of unity indicates perfect skill, a score of zero indicates perfect non-skill, and the expected score for a random alarm function is ½. The area skill score quantifies the performance of an arbitrary alarm function relative to a reference model. To illustrate the testing method, we consider the ten-year experiment by J. Rundle and others to predict $M \geq 5$ earthquakes in California. We test forecasts from three models: Relative Intensity, a simple spatial clustering model constructed using only smoothed historical seismicity; Pattern Informatics, a model that aims to capture seismicity dynamics by pattern recognition; and the USGS National Seismic Hazard Map, a model that comprises smoothed historical seismicity, zones of "background" seismicity, and explicit fault information. Results show that neither Pattern Informatics nor National Seismic Hazard Map provide significant performance gain relative to the Relative

Intensity reference model. We suggest that our testing method can be used to evaluate future experiments in the Collaboratory for the Study of Earthquake Predictability and to iteratively improve reference models for earthquake prediction hypothesis testing.

3.1 Introduction

Despite the notable lack of success in reliably predicting destructive earthquakes, there has been a resurgence of research on earthquake predictability motivated by better monitoring networks and data on past events, new knowledge of the physics of earthquake ruptures, and a more comprehensive understanding of stress evolution and transfer. However, the study of earthquake predictability has been hampered by the lack of an adequate infrastructure for conducting prospective prediction experiments under rigorous, controlled conditions and evaluating them using accepted criteria specified in advance. To address this problem, the Working Group on Regional Earthquake Likelihood Models (RELM), supported by the Southern California Earthquake Center (SCEC) and U. S. Geological Survey (USGS), has recently established a facility for prospective testing of scientific earthquake predictions in California, and a number of experiments are now underway (Field 2007 and references therein).

The RELM project conforms to the requirements for well-posed prediction experiments (e.g., Rhoades & Evison 1989; Jackson 1996) through a strict set of registration and testing standards. For a five year experiment, models are constructed to predict earthquakes in California above magnitude 4.95 during 2006-2010 by specifying time-invariant earthquake rates in prescribed latitude-longitude-magnitude bins. Three tests based on likelihood measures will be used to evaluate the forecasts (Schorlemmer *et*

al. 2007).

The interest in the RELM project shown by earthquake scientists has motivated an international partnership to develop a Collaboratory for the Study of Earthquake Predictability (CSEP). CSEP is being designed to support a global program of research on earthquake predictability (Jordan 2006), and one of its goals is to extend the testing methodology to include alarm-based predictions. In this chapter, we outline such a methodology and apply it to the retrospective testing of three prediction models for California.

3.2 Alarm-based prediction

Earthquake alarms are a natural construct when we consider the problem of predicting the locations and origin times of earthquakes above some minimum magnitude—target earthquakes. A common approach to this problem is to search for precursory signals that indicate an impending target earthquake in a given space-time window (Keilis-Borok 2003 and references therein, Kossobokov & Shebalin 2003, Keilis-Borok 2002). These signals can be represented by precursory functions, the values of which are computed and analyzed in moving time windows. For example, at the present time t we consider a region R where we wish to predict target earthquakes using precursory function f, which is based on information available up to time t. If $f(t)$ exceeds some threshold value (typically optimized by retrospective testing), an alarm is declared, indicating that one or more target earthquakes are expected in R during the period $(t, t + \Delta t)$, a **time of increased probability** (TIP) (Keilis-Borok and Kossobokov 1990). In practice, pattern recognition algorithms often combine several precursory functions. For

example, the Reverse Tracing of Precursors (RTP) algorithm employs eight intermediate-term precursory patterns and yields alarms with fixed duration but highly-variable spatial extent (Shebalin *et al.* 2006, Keilis-Borok *et al.* 2004), which makes testing difficult.

To place alarm-based testing in the RELM context, we consider spatially-varying but time-invariant prediction models. At the beginning of the experiment, we assume there exists some (unknown) probability P_k that the next target earthquake in the RELM testing region *R* will occur in r_k, the k^{th} subregion of *R*. We further assume that P_k is identical for every target earthquake in the testing interval; i.e., the conditional probability that the n^{th} earthquake locates in r_k after $n-1$ earthquakes have already occurred also equals P_k. We suppose that, prior to the experiment, some reference model of this time-invariant distribution, $\tilde{P}_k$, is available. For example, the prior distribution might be the next-event probability calculated from the smoothed, average historical rate of earthquake occurrence in r_k (Kagan & Jackson 2000, Kafka 2002, Rhoades & Evison 2004, Kossobokov 2004, Helmstetter *et al.* 2007). In this chapter, we will follow Tiampo *et al*. (2002) by calling this the Relative Intensity (RI) forecasting strategy. By definition, the summation of P_k or $\widetilde{P}_k$ over all subregions in *R* is unity.

An alarm-based prediction uses fresh information or insights to identify a "domain of increased probability," $A \subseteq R$, the alarm region, where the true probability is hypothesized to exceed the reference value:

$$P_A \equiv \sum_{r_k \in A} P_k > \sum_{r_k \in A} \tilde{P}_k \equiv \tilde{P}_A . \qquad (3.1)$$

At the end of the testing interval, we observe that *N* target earthquakes have occurred and

that h of these are located in the alarm region A. Under the null hypothesis $H_0 : P_A = \widetilde{P}_A$, the probability of h "hits" in A follows a binomial distribution,

$$B(h \mid N, \widetilde{P}_A) = \binom{N}{h} \widetilde{P}_A{}^h (1 - \widetilde{P}_A)^{N-h} \tag{3.2}$$

H_0 can be rejected in favor of $H_1 : P_A > \tilde{P}_A$ if

$$\sum_{n=h}^{N} B(n \mid N, \tilde{P}_A) \leq \alpha, \tag{3.3}$$

for some critical significance level α; that is, if the probability of obtaining h or more hits by chance is less than or equal to α. Rejection of H_0 in favor of H_1 at a high confidence level ($\alpha << 1$) is evidence that the alarm-based prediction has significant skill relative to a prediction based on the reference model $\tilde{P}_k$.

Following Molchan (1990, 1991) and Molchan & Kagan (1992), we consider how the miss rate, $\nu = (N - h)/N$, varies with the probability-weighted area of alarm region A, $\tau = \tilde{P}_A$. Plots of (τ, ν) where $\tau, \nu \in [0,1]$ are called Molchan diagrams ("error diagrams" in Molchan's terminology). At the end of the testing period, the total number of target events, N, is known, so the value of ν is restricted to the discrete set $\{n/N : n = 0, 1, \ldots, N\}$. The boundary conditions are fixed: if no alarm is declared ($\tau = 0$; the optimist's strategy), all events are missed ($\nu = 1$), whereas if an alarm is declared over the entire testing region R ($\tau = 1$; the pessimist's strategy), no events are missed ($\nu = 0$).

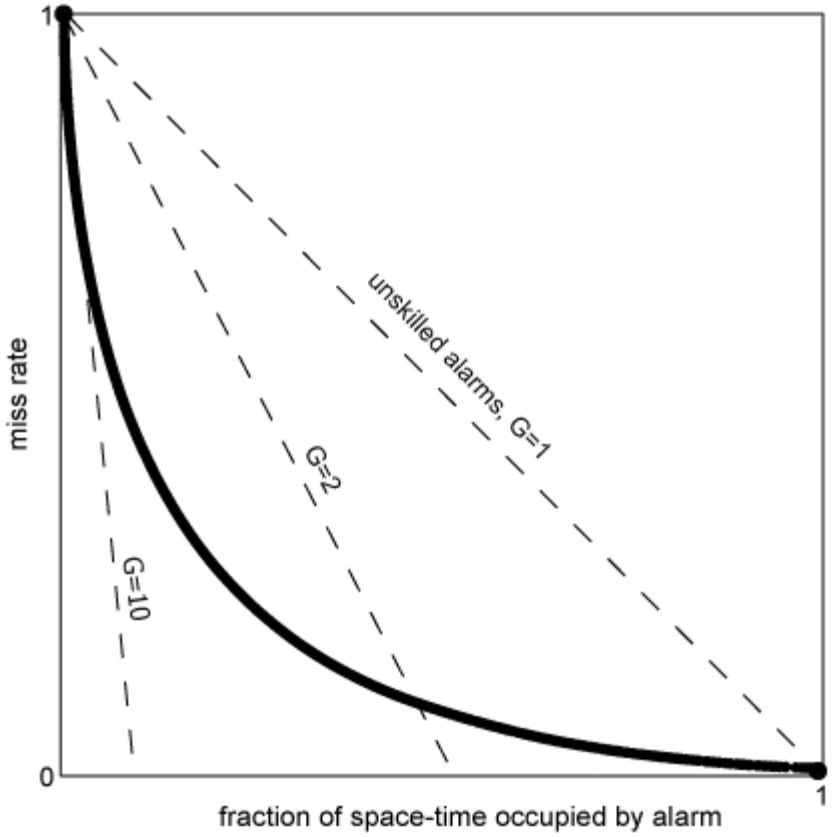

Figure 3.1 Molchan diagram—a plot of miss rate versus fraction of space-time occupied by alarm—with dashed isolines of probability gain. The descending diagonal, corresponding to unit probability gain, represents the expected performance when $P_A = \widetilde{P}_A$. The dark line represents a hypothetical optimal trajectory which depends on the unknown "true" distribution.

Under H_0, the distribution of ν is given by Equation 3.2, and its expected value, $\langle \nu \rangle$, lies on the descending diagonal of the Molchan diagram, $\langle \nu \rangle = 1 - \tau$ (Figure 3.1). More generally, $1 - \langle \nu \rangle$ measures the long-run probability of a subregion being in the alarm region conditional on it containing an event, $P(A \mid E)$, while τ is the prior probability of the alarm region, $P(A) = \tilde{P}_A$. The Bayes identity requires

$$P(E \mid A) = \left[\frac{P(A \mid E)}{P(A)} \right] P(E), \tag{3.4}$$

where the quantity in brackets is called the probability gain (Aki 1981, Molchan 1991, McGuire *et al.* 2005):

$$G \equiv \frac{P(E \mid A)}{P(E)} = \frac{1 - \langle \nu \rangle}{\tau}. \tag{3.5}$$

On the Molchan diagram, the sample value of G is the slope of the line connecting (0,1) to (τ, ν), and (3) provides a test of the null hypothesis $H_0 : G = 1$ against the alternative $H_1 : G > 1$.

For the grid-based RELM models, the values of τ are also discrete, given by summations over the cell values $\tilde{P}_k$. In the continuum limit where the cell size shrinks to zero, $\tilde{P}_k$ becomes a probability density function (p.d.f) $\tilde{p}(\mathbf{x})$ of an event at a geographic location $\mathbf{x} \in R$. The analysis is also simplified by representing target earthquakes as discrete points at their geographic epicenters $\{\mathbf{x}_n : n = 1, 2, \ldots, N\}$. In this point-process limit, τ is a continuous variable on [0,1], and all realizable values of ν are stepwise constant functions of τ. Moreover, as cell size shrinks to zero, τ becomes the measure of false positives (false alarms) and $(1 - \tau)$ becomes the measure of correct negatives. The Molchan diagram then describes the complete contingency table (Figure 3.2) and is equivalent to the Receiver Operating Characteristics (ROC) diagram, a plot of hit rate versus false alarm rate that has been employed in weather forecast verification, medical testing, and signal analysis (Mason 2003 and references therein). In the continuum limit, we can contour constant values of α on the Molchan diagram by finding the minimum value of τ that solves the equality in (3) for each discrete value of ν. These stepwise confidence intervals are illustrated in Figure 3.3.

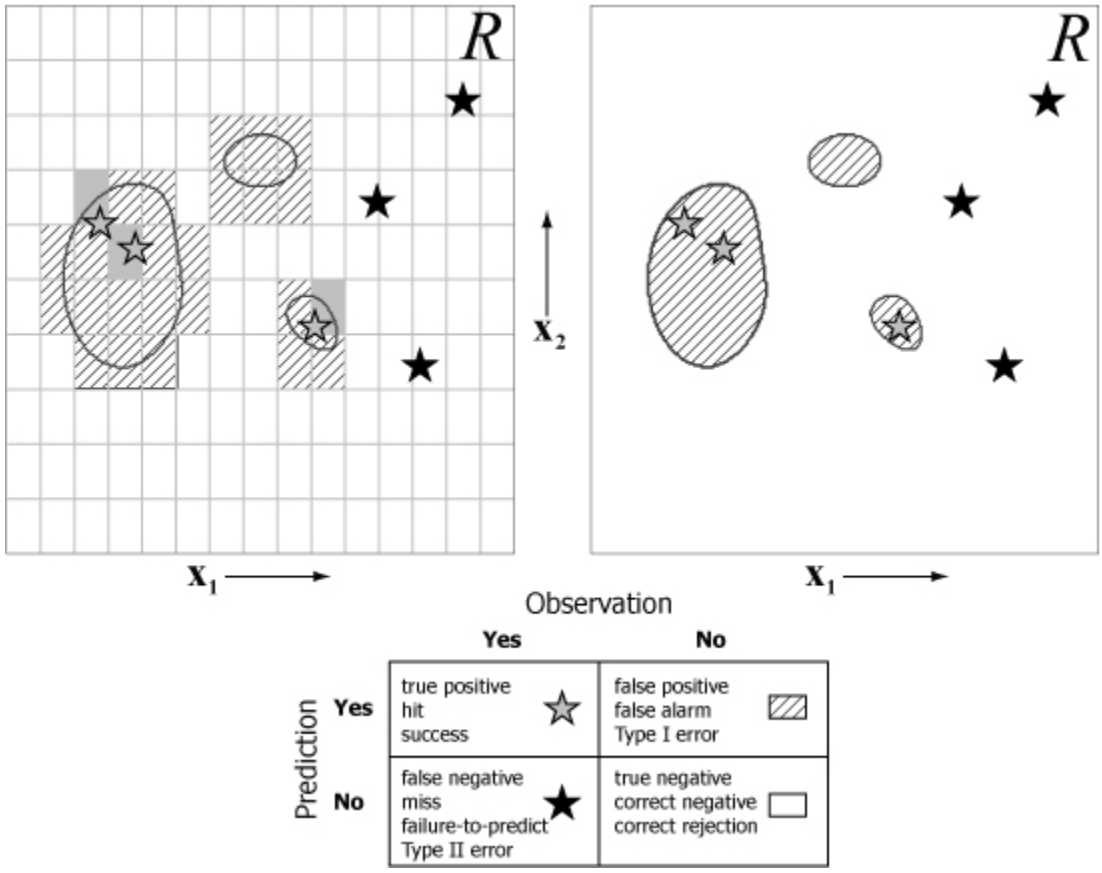

Figure 3.2 Example alarms in the case of coarse spatial discretization (left panel) and in the continuum limit, where cell size approaches zero (right panel). A full contingency table (with alternate names for each contingency) provides a legend. In the left panel, the shaded boxes are hit regions, which are not explicitly represented in the contingency table because hits and misses describe events rather than cells. In the right panel, the hit regions are infinitesimally small, so all alarm regions become false alarms.

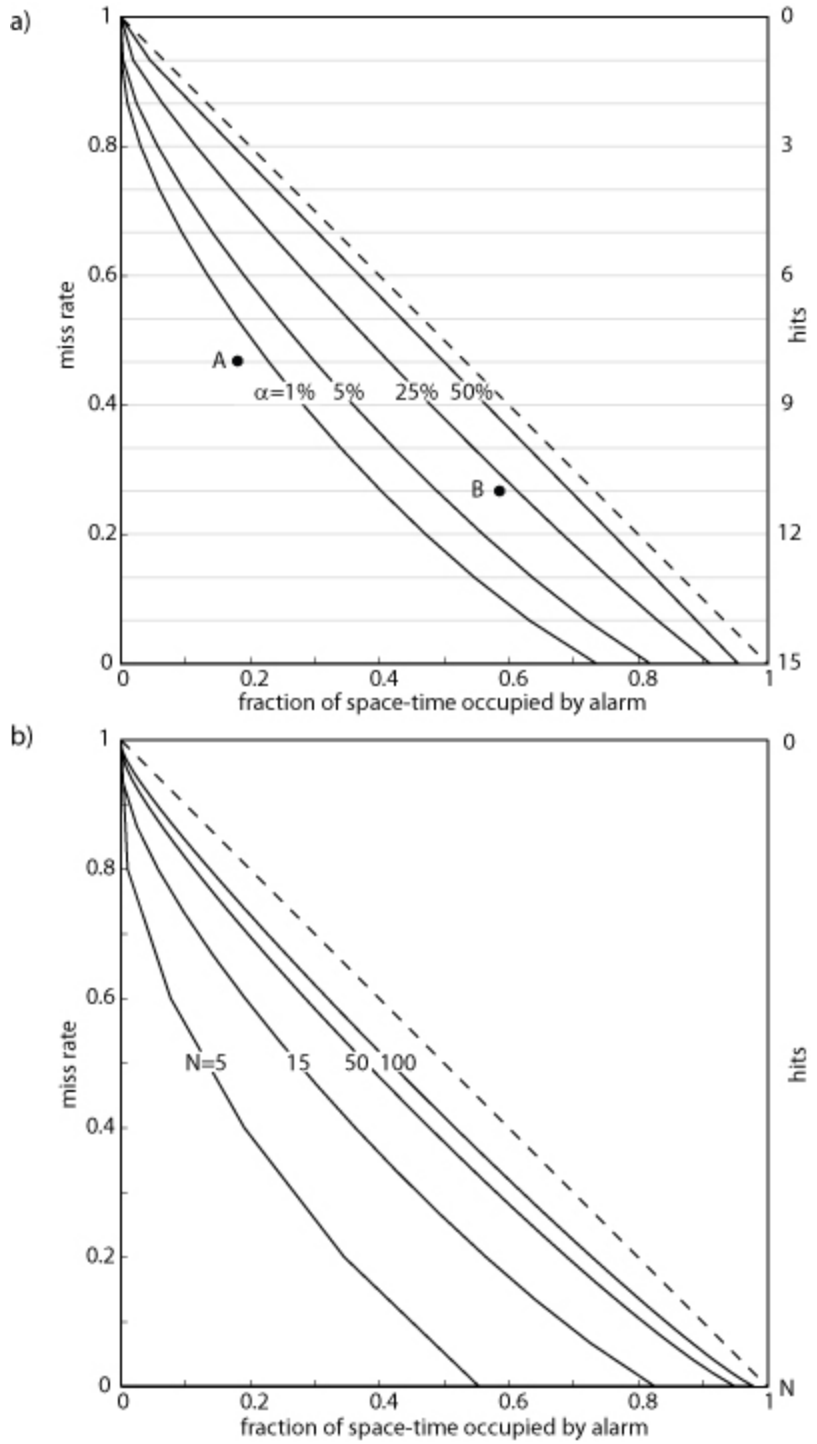

Figure 3.3 Molchan diagram confidence bounds computed by solving Equation 3 for a) fixed N and varying α and b) fixed α and varying N. In a), N = 15 and the curves are contours for α = {1%,5%, 25%, 50%}. Here, point A represents an alarm region that has obtained 8 hits and indicates that the null hypothesis $H_0 : P_A = \widetilde{P}_A$ can be rejected at a confidence level greater than 99% while the point B (11 hits) supports rejection at just above 75% confidence. In b), α = 5% and the curves are contours for N = {5, 15, 50, 100}. As N increases, the contours approach the descending diagonal.

3.3 Optimal Molchan trajectories

Minimizing $\langle \nu \rangle$ for a fixed value of τ yields an optimal alarm region A^*. We consider the special case where two conditions apply:

(a) the prior distribution is uniform over R; i.e., the values of the p.d.f. $\tilde{p}(\mathbf{x})$ are everywhere equal, and τ measures the normalized geographic area covered by an alarm; and

(b) the true distribution has no flat spots; i.e., in general, the contours $\{\mathbf{x}(\lambda) : p(\mathbf{x}) = \lambda\}$ are sets of measure zero (lines, points, or empty) for all contour levels λ.

The optimization problem is then solved by the "water-level principle", which states that a region on a map above a contour level λ has the highest average elevation of any region with the same area (Figure 3.4). In the case we consider here, the optimal alarm is the domain of R where the topography represented by $p(\mathbf{x})$ rises above a water level λ; this alarm can be expressed as $A^*(\lambda) = \{\mathbf{x} \in R : H(p(\mathbf{x}) - \lambda) = 1\}$, where H is the Heaviside step function. Dropping the water level from the maximum value of $p(\mathbf{x})$ to zero traces out an optimal trajectory,

$$\tau^*(\lambda) = \int_R H(p(\mathbf{x}) - \lambda) d\mathbf{x}, \tag{3.6a}$$

$$\nu^*(\lambda) = 1 - \int_R p(\mathbf{x}) H(p(\mathbf{x}) - \lambda) d\mathbf{x}. \tag{3.6b}$$

The optimal trajectory lies on or below the descending diagonal of the Molchan diagram (Figure 3.1).

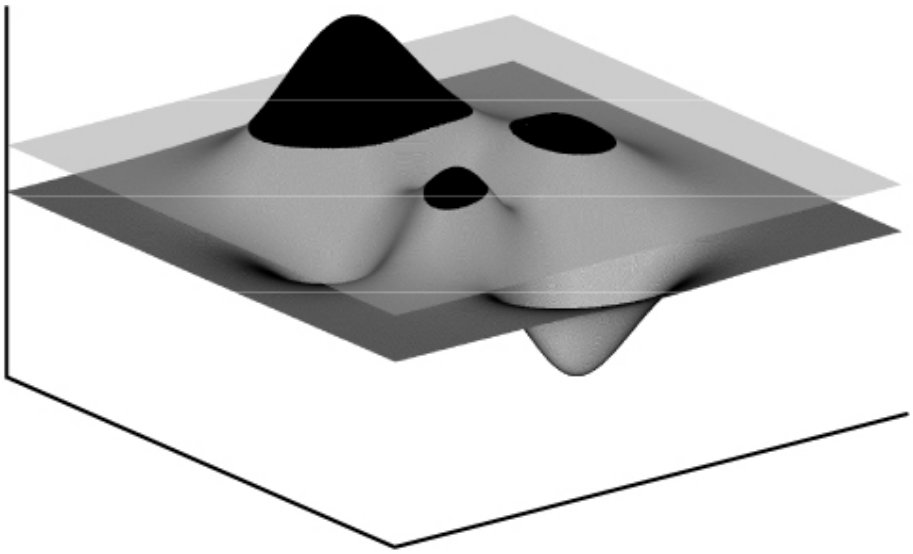

Figure 3.4 Illustration of alarm optimization technique via water-level threshold procedure. Here, the curved surface on R represents a gain function $g(\mathbf{x})$; the plane intersects the surface at a height of λ. The resulting region above this threshold is the optimal alarm region A^*, or the region of this fixed area with the highest average elevation. A map view of this alarm region is shown in Figure 3.2.

We can relax condition **(a)** by considering an arbitrary prior p.d.f. satisfying $\tilde{p}(\mathbf{x}) > 0$ for all $\mathbf{x}$ in R. In this case, the optimal alarm is given by

$$A^*(\lambda) = \{\mathbf{x} \in R : H(g^*(\mathbf{x}) - \lambda) = 1\}, \tag{3.7}$$

where $g^*(\mathbf{x}) = \dfrac{p(\mathbf{x})}{\tilde{p}(\mathbf{x})}$ is the optimal local probability gain at $\mathbf{x}$. The optimal trajectory becomes

$$\tau^*(\lambda) = \int_R \tilde{p}(\mathbf{x}) H(g^*(\mathbf{x}) - \lambda) d\mathbf{x} = \tilde{P}_{A^*}, \tag{3.8a}$$

$$\nu^*(\lambda) = 1 - \int_R p(\mathbf{x}) H(g^*(\mathbf{x}) - \lambda) d\mathbf{x} = 1 - P_{A^*}. \tag{3.8b}$$

We note that the probability gain of the optimal trajectory can be written as the weighted average of the local gain over the optimal alarm region,

$$G^*(\lambda) = \frac{\int_{A^*(\lambda)} g^*(\mathbf{x})\tilde{p}(\mathbf{x})d\mathbf{x}}{\int_{A^*(\lambda)} \tilde{p}(\mathbf{x})d\mathbf{x}} \quad . \tag{3.9}$$

We can relax condition **(b)** by considering an arbitrary optimal local gain function $g^*(\mathbf{x})$ that may contain flat spots. We consider a flat spot domain $D \subset R$ where $p(\mathbf{x}) = \lambda_D \tilde{p}(\mathbf{x})$ for all **x** in *D* and let

$$A^{*\pm}(\lambda_D) = \lim_{\varepsilon \to 0} \{\mathbf{x} \in R : H(g^*(\mathbf{x}) - \lambda_D \pm \varepsilon) = 1\}, \tag{3.10}$$

such that $A^{*+} = A^{*-} \cup D$. Then, the optimal trajectory "jumps" from (τ^{*-}, ν^{*-}) to (τ^{*+}, ν^{*+}) at λ_D. Sampling any two subsets of the same size from *D* yields the same Molchan trajectory point, and therefore relaxing condition (b) can lead to non-unique optimal alarms. We note, however, that such a sampling can only yield points on the line connecting (τ^{*-}, ν^{*-}) to (τ^{*+}, ν^{*+}), and therefore the Molchan trajectory remains unique and no alarm region can achieve a lower value of $\langle \nu \rangle$.

In statistical hypothesis testing, the power of a test is the probability that a false null hypothesis is rejected—in other words, that a Type II error is not committed—and is equal to $1 - \beta$ where β is the Type II error rate (Lehman & Romano 2005). In our problem, where the Type I error rate is measured by τ and the Type II error rate by $\nu(\tau)$, an appropriate measure of the power of an alarm is $1 - \nu(\tau)$. In these terms, $A^*(\lambda)$ is the most powerful alarm of size $\tau(\lambda)$. As the reference model approaches the true distribution, $\tilde{P}_{A^*} \to P_{A^*}$, the power of the optimal alarm approaches the average power of a random alarm, $1 - \nu^*(\tau) \to \tau(\lambda)$, and a larger number of events *N* is needed to

discriminate H_1 from H_0.

When $\tilde{p}(\mathbf{x}) = p(\mathbf{x})$ for all **x** in *R*, $g^*(\mathbf{x})$ is flat—and in particular, equal to unity—throughout *R*, and the optimal trajectory coincides with the descending diagonal of the Molchan diagram. In this case, no alarm-based strategy can reject H_0, and the time-invariant prediction problem for the simple RELM set-up is solved.

3.4 Alarm functions and the area skill score

We consider alarm sets $\{A(\lambda): \lambda \geq 0\}$ that are ordered by $\tau(\lambda)$ such that

$$\tau(A(\lambda)) < \tau(A(\lambda')) \Rightarrow A(\lambda) \subset A(\lambda'); \tag{3.11}$$

and complete on $0 \leq \tau \leq 1$; i.e., sufficient to generate complete Molchan trajectories $\{\nu(\tau): \tau \in [0,1]\}$. A complete, ordered alarm set can be represented as an unscaled contour map on *R* (Figure 3.5). Such a set can be generated from a continuous, positive-semidefinite **alarm function** *g*(**x**) by water-level contouring,

$$A_g(\lambda) = \{\mathbf{x} \in R : H(g(\mathbf{x}) - \lambda) = 1\}. \tag{3.12}$$

An example of an alarm function is the optimal gain function, $g^*(\mathbf{x})$. In fact, any p.d.f. constitutes an alarm function; not all alarm functions, however, specify a p.d.f.

Alarm functions that have Molchan trajectories with the same values as *g*(**x**) form an equivalence class indexed by *g*:

$$C_g = \{f(\mathbf{x}) : \nu_f(\tau) = \nu_g(\tau), \forall \tau \in [0,1]\}. \tag{3.13}$$

An infinite number of alarm functions yield the same alarms as *g*(**x**). For example,

consider any order-preserving functional; i.e., one that satisfies $h(z) < h(z') \Leftrightarrow z < z'$. Then $f(\mathbf{x}) = h(g(\mathbf{x}))$ is also in the equivalence class C_g. Such a functional only rescales the contour map, so that $A_f = A_g$, and the trajectory $\nu_f(\tau)$ is identical to $\nu_g(\tau)$.

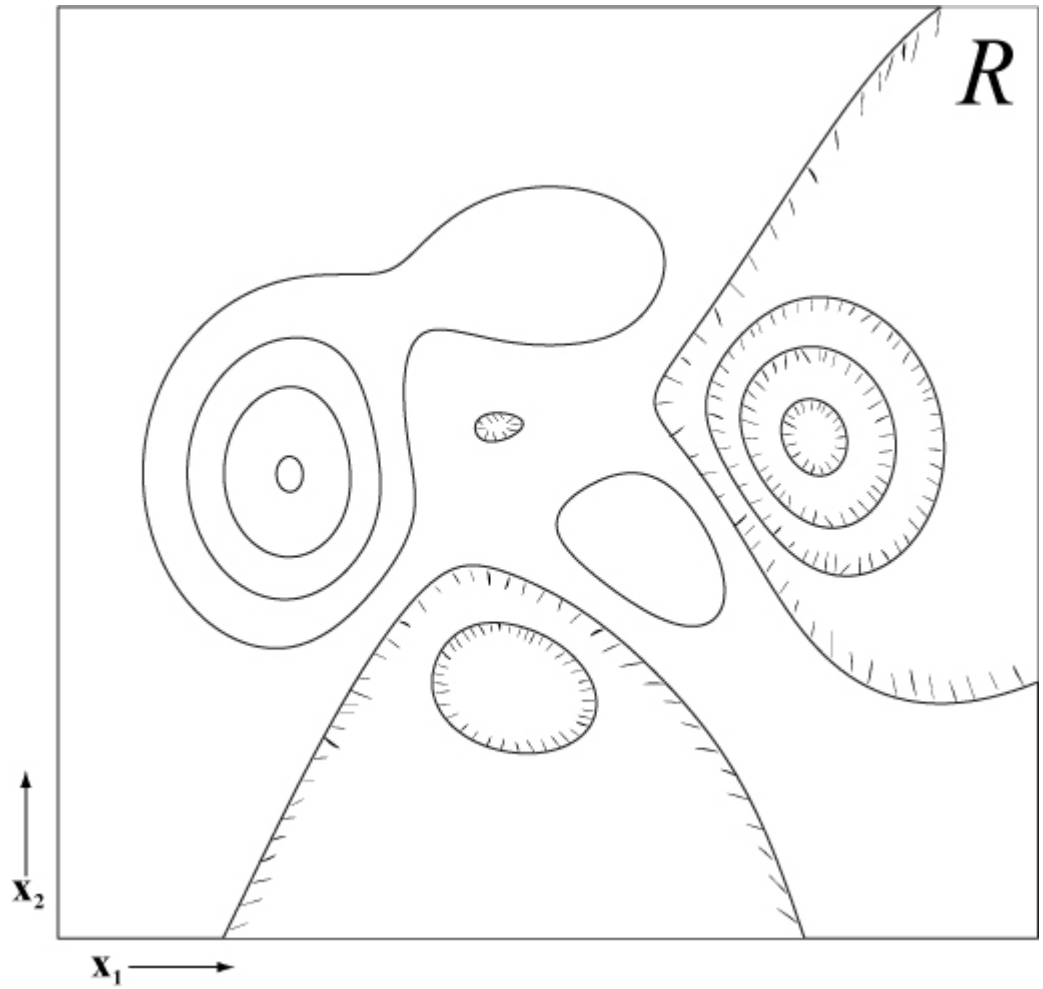

Figure 3.5 Map view of an example alarm function; here, the contour map corresponds to the spatial alarm function shown in Figure 3.4. Hashed contours indicate descending values.

All members of a given equivalence class C_g provide equal probability gain relative to the prior $\tilde{p}(\mathbf{x})$. The interesting extremes are C_1, comprising functions equivalent to $g(\mathbf{x}) = 1$, and C_*, comprising functions equivalent to g^*. Alarm functions belonging to C_1 yield trajectories with expected values lying on the descending diagonal

and thus provide no gain in the long run, whereas alarm functions belonging to C_* yield the optimal trajectory. For any alarm function $f(\mathbf{x})$, we seek to test the null hypothesis $H_0 : f \in C_1$, against the alternative that f performs "better" than alarm functions belonging to C_1. (An alarm function with an expected trajectory above the descending diagonal is "worse" than C_1; however, as noted by Molchan & Kagan (1992), we can use it to generate a set of "anti-alarms" whose complements in R have probability gains greater than unity. We therefore consider only one-sided tests.)

The performance of an alarm function $f(\mathbf{x})$ can be measured by the area above its Molchan trajectory evaluated at a given τ, a statistic we call the **area skill score**:

$$a_f(\tau) = \frac{1}{\tau}\int_0^{\tau}[1-\nu_f(t)]dt \tag{3.14}$$

This statistic is normalized such that its value is between 0 and 1 and under H_0 its expectation is

$$\left\langle a_f(\tau)\right\rangle = \frac{\tau}{2}. \tag{3.15}$$

We can use this statistic to assess the skill of $f(\mathbf{x})$ relative to $\tilde{p}(\mathbf{x})$ by testing the null hypothesis $H_0 : a_f(\tau) = \tau/2$ against the alternative $H_1 : a_f(\tau) > \tau/2$. In the limit of infinitesimal discretization in τ, the area skill score is equivalent to the area under curve (AUC) measure used in ROC analyses (Mason 2003).

In order to use the area skill score for hypothesis testing, we have explored the score distribution of unskilled alarm functions with an arbitrary prior (see Chapter Four). We have an analytic approach for generating moments of the distribution and find that

this distribution is related to the distribution of cross-sectional wedge "area" of an *N*-dimensional hypercube along its principal diagonal. An application of the Central Limit Theorem shows that, in the case of continuous alarm functions, the area skill score distribution at $\tau = 1$ is asymptotically Gaussian with a mean of ½ and a variance of $1/(12N)$. Furthermore, the distribution's kurtosis excess—a factor dependent on the second and fourth central moments and an indicator of deviation from the Gaussian distribution—is equal to $-5/(4N)$. For *N* on the order of a dozen or more, the Gaussian approximation provides an excellent estimate of confidence bounds.

We can estimate the area skill score distribution by simulation for any number of target events *N* at any value of τ. It can be shown that the power of the area skill score, while dependent on the prior, tends to increase with increasing τ, and therefore it is best to use $a_f(\tau = 1)$ for the hypothesis test. In the illustrative experiment described below, we consider discrete alarm functions and the observed seismicity yields multiple earthquakes in a single forecast cell. In this case, the Molchan trajectory and area skill score confidence bounds are most easily estimated by simulation of unskilled alarm functions (i.e., those belonging to C_1).

3.5 Models and data

To demonstrate the area skill score testing procedure, we consider three models of spatial predictability—Relative Intensity (RI), Pattern Informatics (PI), and the United States Geological Survey National Seismic Hazard Map (NSHM)—in a quasi-prospective prediction experiment. For visual comparison, the alarm function values for each model are shown in Figures 3.6-3.8.

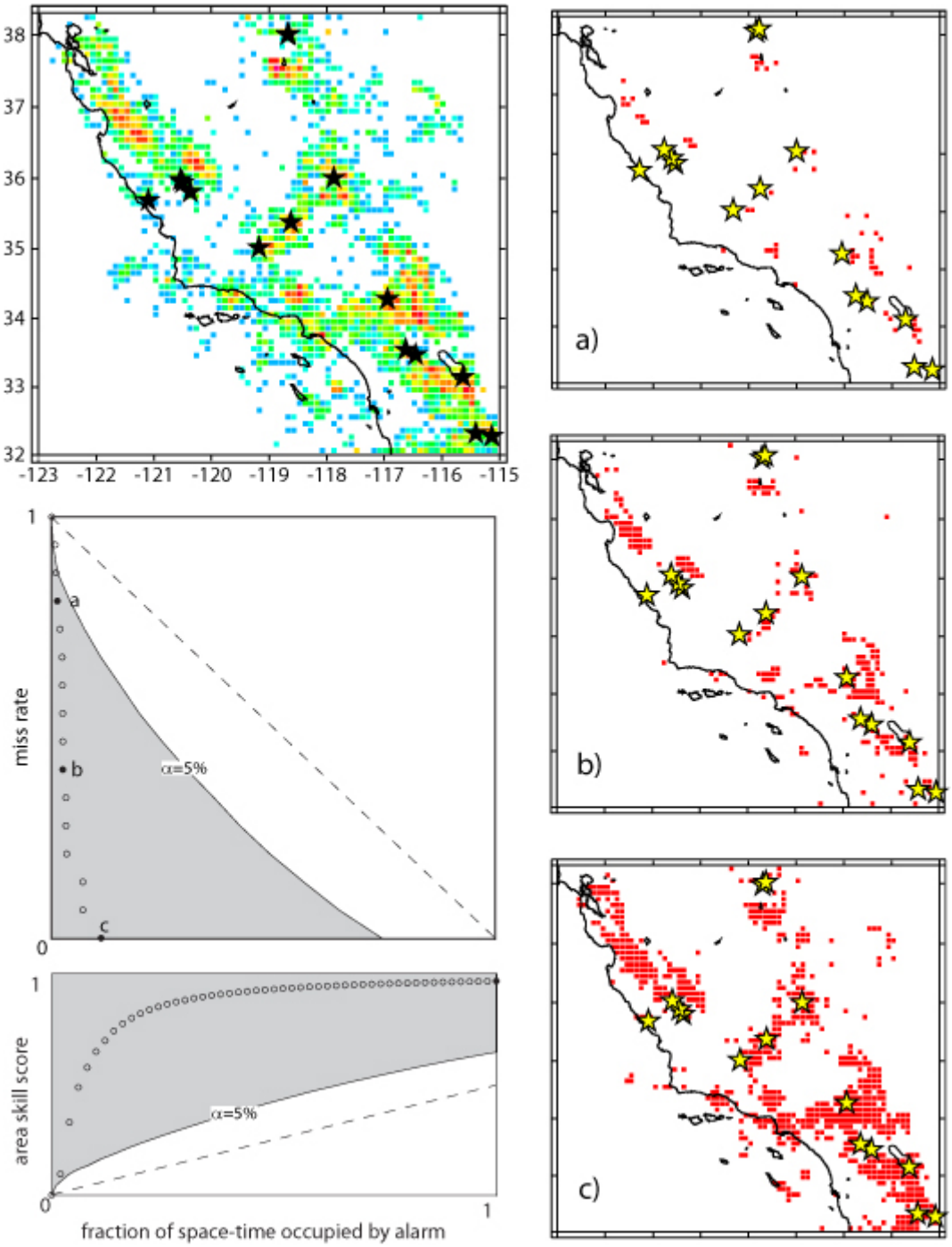

Figure 3.6 Illustration of testing Relative Intensity (RI) alarm function relative to a uniform spatial prior. Top left frame shows map of RI alarm function values and observed target earthquakes (stars), created with Generic Mapping Tools software (Wessel & Smith 1998). Panels a), b), and c) show alarm regions for three decreasing threshold values. The corresponding Molchan diagram points are labelled in the plot on the left of the second row. The left panel in the third row shows the corresponding area skill score trajectory. The shaded areas on the plots are the $\alpha = 5\%$ critical region.

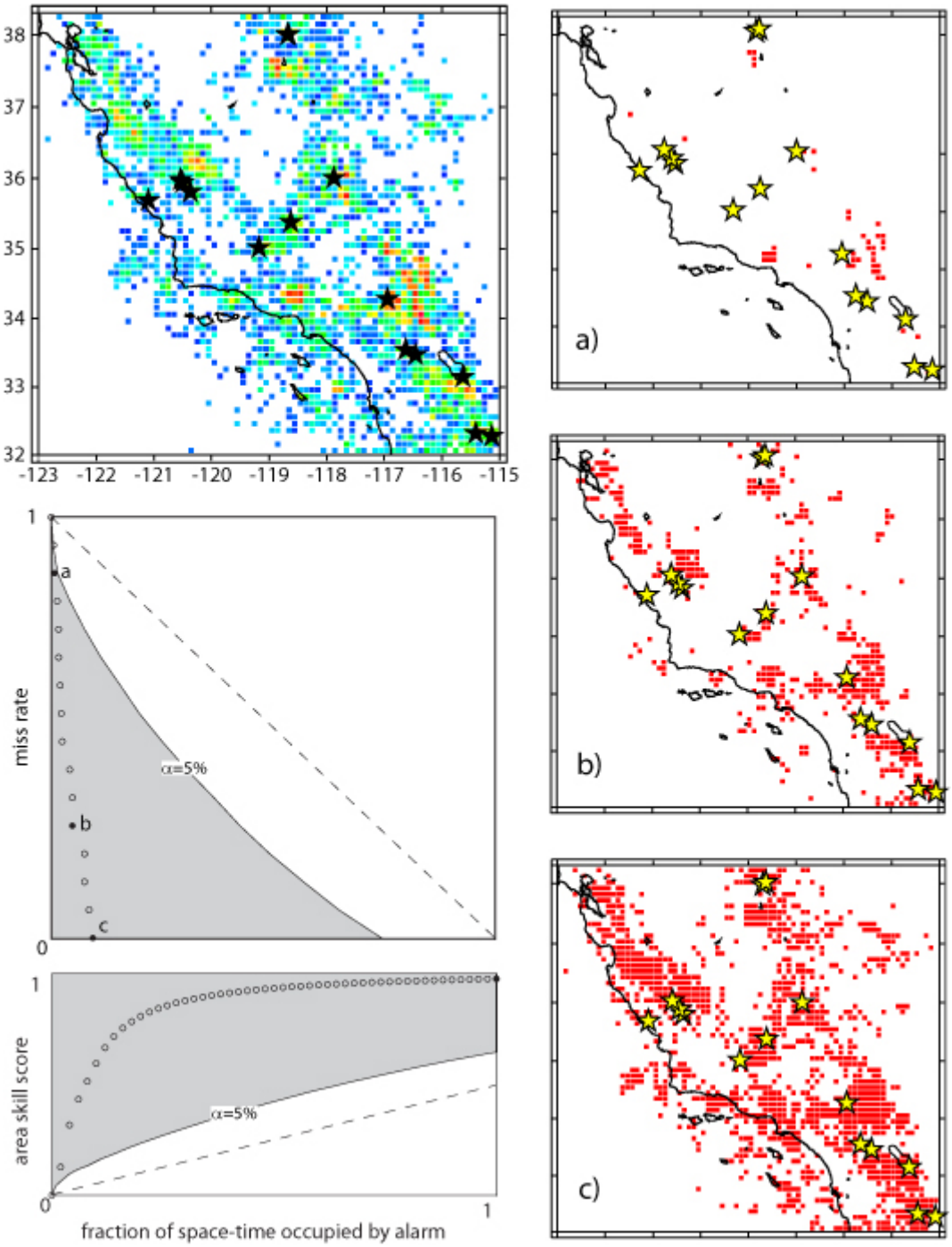

Figure 3.7 Same as Figure 3.6 for the Pattern Informatics (PI) alarm function.

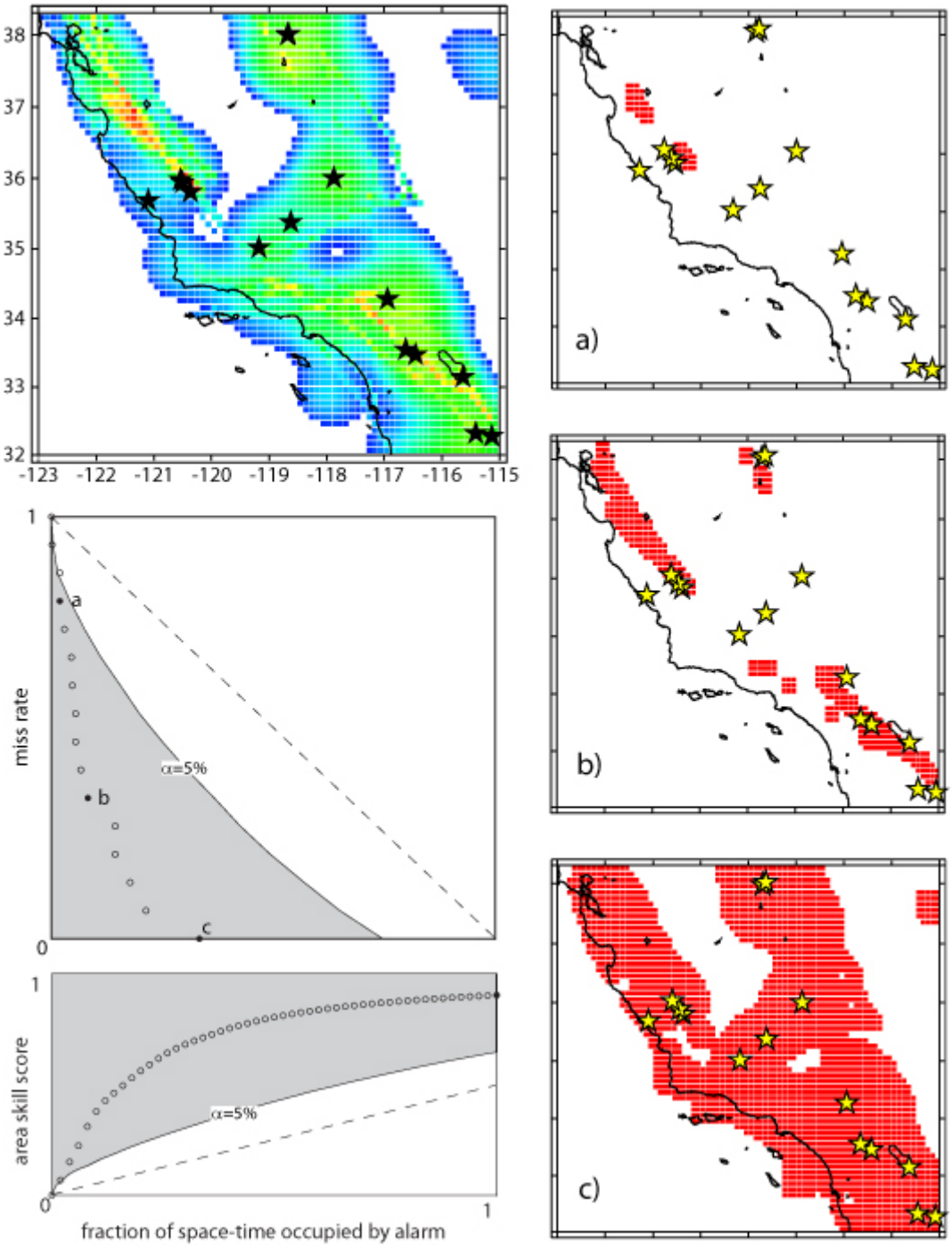

Figure 3.8 Same as Figure 3.6 for the National Seismic Hazard Map (NSHM) alarm function.

These models make for an interesting set of examples because they represent distinct hypotheses about the spatial distribution of earthquakes. RI suggests that future earthquakes are most likely to occur where historical seismicity rates are highest (Tiampo *et al*. 2002). RI uses a particularly simple measure of seismicity—the rate of past earthquakes occurring in each spatial cell—and belongs to a general class of smoothed seismicity models. Proximity to Past Earthquakes (Rhoades and Evison 2004), Cellular Seismology (Kafka 2002), and others (e.g., Kagan and Jackson 2000, Kossobokov 2004, Helmstetter *et al.* 2007) are members of this class that offer slightly different representations of the same basic hypothesis; each has been recommended as a reference model.

PI suggests that the locations of future earthquakes are indicated by anomalous changes in seismic activity. Regions undergoing seismic activation or seismic quiescence are found by computing the short-term seismicity rate in a given spatial cell—say, for the previous ten years—and comparing with the long-term seismicity rate in this cell—say, for the previous fifty years. If the short-term rate is anomalously low/high, the cell is considered to be undergoing seismic quiescence/activation in preparation for a target earthquake in the near future (Tiampo *et al.* 2002). As shown in Figures 3.6 and 3.7, the PI and RI alarm functions are highly correlated—in particular, many regions with a high PI index also have a high RI index.

NSHM suggests that future earthquakes will occur where past earthquakes have occurred, with the qualification that moderate to large earthquakes are likely to occur near mapped faults and some earthquakes will be surprises. Therefore, the NSHM earthquake rate model combines smoothed historical seismicity, fault information, and

"background" zones where a spatially uniform seismicity rate is assumed. According to Frankel *et al*. (1996, 2002), this combination represents the best knowledge of faults and spatial distribution of earthquakes. The NSHM alarm function values in Figure 3.8 reveal a forecast that is much smoother than that of RI and PI.

Rundle *et al.* (2003) issued ten-year "hotspot" maps based on RI and PI . The hotspots are alarms that last the duration of the experiment and are derived from underlying alarm functions. The alarm function of the PI model is not a p.d.f. and does not provide explicit forecasts of earthquake rate; PI simply provides a ranking of cells and therefore the RELM likelihood testing procedures cannot be applied in a straightforward way. We were provided the PI hotspot map values by J. Holliday (personal communication). The RI alarm function constitutes a next-event spatial p.d.f. and, upon assumption of a magnitude distribution and regional seismicity rate, can be tested using the RELM methods (e.g., Zechar *et al*. 2007). For this experiment, we computed the RI values using the parameters suggested by Rundle *et al.* (2002). The alarm function of the NSHM model is a p.d.f. of space and magnitude and yields a forecast of expected seismicity rates; a previous version of the model is currently being tested by RELM (Petersen *et al.* 2007). We computed the 2002 NSHM values using the OpenSHA platform (Field *et al.* 2005). These three models also make for an interesting set of examples as they demonstrate the potential to compare heterogeneous forecasts.

We consider the experiment specified by Rundle *et al.* (2003): to forecast the epicentral locations of $M \geq 5$ earthquakes during the ten year period starting 1 January 2000 in the gridded region with latitude ranging from 32° to 38.3°, longitude ranging from -123° to -115°, and a spatial discretization of 0.1°. We consider this a quasi-

prospective experiment because the PI and RI forecasts were issued in 2002; none of the forecasts, however, use data collected after the beginning of the experiment. Although the magnitude scale and earthquake catalog to be used for verification were not stated in the original experiment specification, we followed the RELM project in taking the ANSS composite catalog to be the authoritative data source for this natural laboratory. We selected all tectonic earthquakes in this region since 2000 that had ANSS reported magnitudes greater than or equal to 5.0, regardless of the reported magnitude scale. This selection process yielded the 15 target earthquakes listed in Table 3.1.

Table 3.1: Target earthquakes

#	Origin Time	Magnitude	Latitude (degrees)	Longitude (degrees)
1	2001/02/10 21:05	5.13 ML	34.2895	-116.9458
2	2001/07/17 12:07	5.17 Mw	36.0163	-117.8743
3	2001/10/31 07:56	5.09 ML	33.5083	-116.5143
4	2002/02/22 19:32	5.70 Mw	32.3188	-115.3215
5	2003/12/22 19:15	6.50 Mw	35.7002	-121.0973
6	2004/09/18 23:02	5.55 Mw	38.0095	-118.6785
7	2004/09/18 23:43	5.40 Mw	38.0187	-118.6625
8	2004/09/28 17:15	5.96 Mw	35.8182	-120.3660
9	2004/09/29 17:10	5.00 Mw	35.9537	-120.5022
10	2004/09/29 22:54	5.03 Mw	35.3898	-118.6235
11	2004/09/30 18:54	5.00 Mw	35.9890	-120.5378
12	2005/04/16 19:18	5.15 ML	35.0272	-119.1783
13	2005/06/12 15:41	5.20 Mw	33.5288	-116.5727
14	2005/09/02 01:27	5.10 Mw	33.1598	-115.6370
15	2006/05/24 04:20	5.37 Mw	32.3067	-115.2278

Table 3.1 Fifiteen target earthquakes occurring in the testing region with latitude ranging from 32° to 38.3°, longitude ranging from -123° to -115°, during the interval 1 January 2000 – 30 June 2007.

3.6 Models and data

Earthquakes cluster in space and time and therefore any forecast that captures this clustering behavior should outperform a uniform reference model (Kagan 1996, Stark 1996, 1997, and Michael 1997). Figures 3.6-3.8 confirm the expectation that RI, PI, and NSHM provide significant gain relative to a spatially uniform prior distribution. From

the area skill score trajectories, and in particular, the points at $\tau = 1$, it can be seen that, at greater than 95% confidence, each forecast obtains an area skill score that is greater than ½.

To include time-invariant spatial clustering in the reference model, we use the RI alarm function values as the prior distribution—in other words, the RI index defines the measure of space for τ—and compute the Molchan trajectory and corresponding area skill score curve for the PI and NSHM forecasts. Computationally, this means that the "cost" of declaring an alarm in a given cell is proportional to the RI alarm function value of this cell; in the case of a uniform prior distribution, the cost is everywhere equal. Figure 3.9 shows the result of testing for the 15 target earthquakes since 1 January 2000. In the calculation of τ and ν, we include the margin of error suggested by Rundle *et al.* (2003); namely, if a target earthquake occurs in an alarm cell or in one of the alarm cell's immediate neighbors (Moore neighborhood), it is considered a hit. Accordingly, all cells in the Moore neighborhood of alarm regions are counted as alarms when computing τ. We note that our method for generating alarms from an alarm function is exact and efficient; we use as the alarm thresholds all of the unique values of the alarm function, rather than iterating the thresholds by some constant. The codes for generating the Molchan and area skill score trajectories are available upon request.

With RI as the reference model, the Molchan trajectories for PI and NSHM are closer to the descending diagonal, indicating much smaller probability gains than in the case of a uniform reference model. The NSHM forecast, however, yields three exceptional trajectory points at low values of τ. These points arise from the fact that

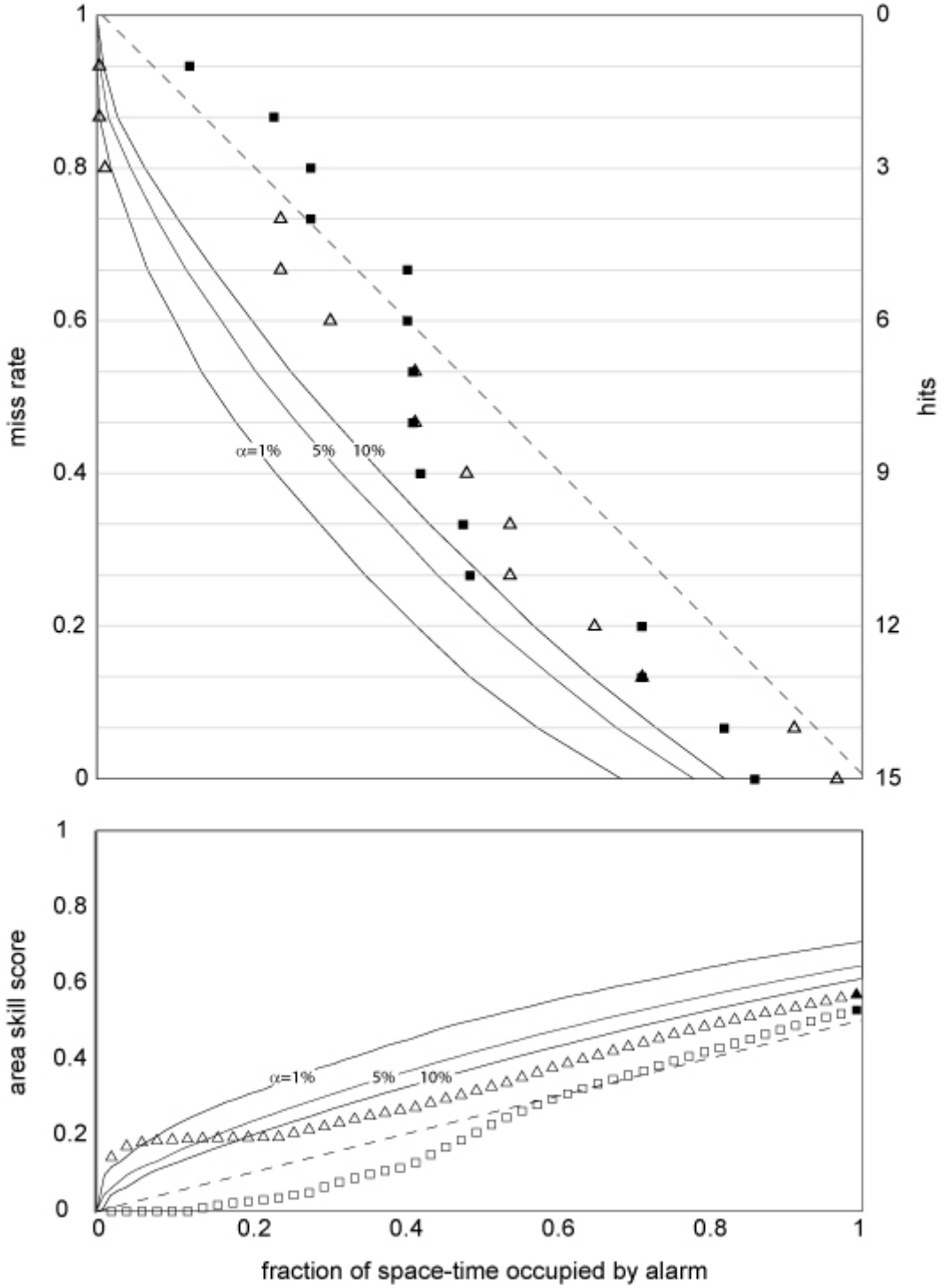

Figure 3.9 Results of Molchan trajectory/area skill score analysis for PI (squares) and NSHM (triangles) relative to the RI reference model. Top panel shows complete Molchan trajectories for both predictions and bottom panel shows corresponding area skill score curve. Each plot also shows the $\alpha = 1\%$, 5%, and 10% critical boundaries. In the Molchan trajectory plot, points below these boundaries reject the alarm region null hypothesis; in the area skill score trajectory plot, points above the boundaries reject the alarm function null hypothesis. We test the area skill score points at unit τ—the filled points on the bottom panel—and find that neither PI nor NSHM provides significant gain relative to RI.

three of the target earthquakes—numbers 3, 8, and 13 in Table 3.1—occurred in cells where NSHM had very high alarm function values and RI had low values. These three hits also manifest themselves in the area skill score trajectory, where NSHM obtains a few exceptional points at small τ. Because the statistical power of the area skill score increases with increasing τ, however, we test the area skill score value at $\tau = 1$; here, neither PI nor NSHM obtain a score that is significantly greater than ½. Given these results and because we test at standard significance values $\alpha = 1\%$, 5% and 10%, we cannot reject at the 90% confidence level the null hypothesis that PI and NSHM belong to C_1. In other words, the observed set of 15 target earthquakes during this experiment is consistent with the spatial distribution forecast by RI and neither PI nor NSHM provide significant gain relative to this simple model of smoothed seismicity.

3.7 Conclusions

In an illustration of an alarm-based earthquake prediction evaluation technique, we have shown that, contradictory to the retrospective testing of Rundle *et al.* (2002, 2003), the Pattern Informatics forecast model does not yield statistically significant performance in a *quasi*-prospective earthquake forecast . In particular, at the 90% confidence level, we are unable to reject the null hypothesis that PI and NSHM provide no gain relative to RI.

With respect to NSHM, we note that this model was constructed to forecast large earthquakes in the long term, and we have tested it for a period of only 7.5 years during which only one damaging earthquake occurred. By increasing either the duration of the experiment or the minimum magnitude of target earthquakes, the fault information

included in the NSHM forecast might provide better spatial resolution and accuracy than purely statistical methods. Both of these increases require more time to collect a meaningful number of events but may offer insight into the spatial predictability of a region's largest earthquakes. Fault-based experiments and testing thereof will be investigated further by CSEP researchers.

While the alarm functions considered here focus on forecasting the geographic location of future earthquakes above a minimum magnitude, our method can be applied to more complex forecasts, including time-varying, magnitude-varying, and fault-based alarm functions. In the experiment considered here, we have disregarded catalog errors. Because the forecasts are time-invariant, timing errors are irrelevant to the test. The spatial discretization of the forecasts is of such a scale that location errors are probably negligible. Because this experiment concerns earthquakes above a minimum magnitude without magnitude discretization, magnitude errors are only relevant for earthquakes close to the minimum target magnitude and are also unlikely to change the result of the hypothesis test. In general, however, it is important to consider catalog errors when testing earthquake forecasts (e.g., Werner & Sornette 2007), and our testing method can account for such errors by using simulations comparable to those planned for the RELM experiments (Schorlemmer *et al.* 2007).

The framework for evaluating multi-level alarms has been described by Molchan and Kagan (1992); applying these principles would allow further disaggregation of testing results. For example, magnitude discretization can reveal that one model accurately predicts small earthquakes and another is better at predicting intermediate size earthquakes. A bootstrap approach where these models are combined may be an effective

way to proceed with earthquake prediction research. Hypothesis testing using the area skill score can be used as a guide in this process.

The consistent failure to find reliable earthquake precursors leads us to believe that a more effective way to advance earthquake prediction is a "brick-by-brick" approach that synthesizes hypotheses, models, and data across space- and time-scales (Jordan 2006). Rigorous testing methods like the one described here are vital in identifying the most robust characteristics of seismicity and improving reference models. Such testing may provide a better means of communicating earthquake forecast performance and progress to the public.

CHAPTER FOUR:
The area skill score statistic for evaluating earthquake predictability experiments

Abstract

Rigorous predictability experimentation requires a profound understanding of the performance metric in use. Here we explore in detail the area skill score metric and issues related to experimental discretization. For the case of continuous alarm functions and continuous observations, we present exact analytic solutions describing the distribution of the area skill score for unskilled predictors, and the approximation of this distribution by a Gaussian with known mean and variance. We also quantify the deviation of the exact area skill score distribution from the Gaussian estimate by specifying the kurtosis excess as a function of the number of observed earthquakes. In the case of discretized alarm functions, and particularly in the case of discretized observations, it is most efficient to simulate the area skill score distribution, and we present detailed analysis of simulations and shortcuts for numerically estimating the distribution. Particular attention is paid to the case in which experiment discretization and/or the target earthquake distribution is such that more than one of the observed target earthquakes occurs within the same space/time/magnitude cell, in which case the probabilities of predicting these events are not independent, thus requiring special attention.

4.1 Earthquake forecasting with an alarm function

Earthquake forecasts can be stated in various forms: one may estimate the time of the next major earthquake on a given fault or fault segment; one might predict that a large earthquake will occur within a specified space/time/magnitude range; or one might forecast the future rate of seismicity throughout a geographical region. In practice, predictions of the first type are difficult to evaluate because they may require decades of waiting for large earthquakes, and fault structures are not uniquely defined, making the assignment of an earthquake to a specific fault or fault segment a subjective procedure. If properly specified, the latter two types of experiments can be evaluated formally, and such experiments are currently underway. For example, the Reverse Tracing of Precursors (RTP) algorithm (Keilis-Borok *et al.* 2004, Shebalin *et al.* 2006) has been used to make predictions of target earthquakes in several regions, and a formal evaluation is presented in Chapter 2. Additionally, many researchers have submitted 5 year seismicity rate forecasts in prescribed latitude/longitude/magnitude bins in California as part of the Regional Earthquake Likelihood Model (RELM) working group project (Field 2007, and references therein). The RELM forecasts are being evaluated within the Collaboratory for the Study of Earthquake Predictability (CSEP) testing center (Jordan *et al*. in prep).

A difficulty arises, however, when we compare forecasts stated in different forms, even when forecasts apply to the same space/time/magnitude domain. For example, RELM likelihood tests used for evaluation require a gridded rate forecast and cannot be used to compare forecasts that are not of this type. One way to address this problem is to consider earthquake forecasts in the basest terms. Most forecast statements can be

reduced to an ordering of space/time/magnitude bins by the expected probability of each bin to host a specified future earthquake (or earthquakes). In other words, most forecasts can be translated to a statement not unlike the following: space/time/magnitude bin r_1 is more likely to host a future earthquake than bin r_2, which in turn is more likely than r_3, and so on. This yields a very general approach by which we can compare forecasts originally stated in different formats: if we consider the region R, each forecast provides an ordering of its bins $r_1, r_2, \ldots, r_j$. In this context, a forecast does well when many earthquakes occur in the most highly-ranked bins and few earthquakes occur in bins with low ranking. We could compare two such forecasts by considering the ten most highly-ranked bins for each forecast and counting the number of earthquakes that occur within these bins, i.e., those that have been successfully predicted. Implicit in this evaluation is the choice of a threshold, below which the rankings are disregarded—this yields a binary prediction. We call any bin above the threshold an **alarm**, where one or more target earthquakes are expected. Furthermore, we call this form of prediction **alarm-based**, and we consider the ranking to be an **alarm function**. We note that an alarm function need not be stated in terms of rank, but the implicit ordering should be unambiguous. For example, each of the RELM forecasts is an alarm function with values specified by expected rates—the bin with the highest forecasted rate is ranked the highest. Likewise, any algorithm that computes a seismicity index provides an alarm function with values specified by the index.

Alarm functions are multidimensional; they can be defined over space, time, magnitude, focal mechanism, etc. To compare two alarm functions, each must be specified on the same parameter space; that is, they should cover the same

space/time/magnitude range, although they need not specify the same discretization. The simple threshold testing method described above can be iterated to consider the entire alarm function by varying the threshold from the highest rank to the lowest. In the rest of this chapter, we explore a testing framework based on alarm functions and a threshold approach to testing.

4.2 Molchan diagram for testing alarm functions

Given an alarm function, a threshold, and an observed target earthquake catalog—that is, a catalog containing the events we wish to predict—we can compute a number of contingency table measures (see Chapter 2 for details). The Molchan diagram (Molchan 1991, Molchan & Kagan 1992) is a useful diagnostic because it captures two such measures and the tradeoff between them: miss rate, ν—the proportion of target earthquakes falling outside all alarms—and the fraction of space-time occupied by alarm, τ. The latter metric requires a reference model $\tilde{p}(\mathbf{x})$ to define the measure of space. The reference model should be a probability density function that estimates the future distribution of target earthquakes; typically, reference model values are computed using the historical distribution of earthquakes. By applying a threshold λ to the alarm function $f(\mathbf{x})$, we obtain an alarm set:

$$A = \{x_i \mid f(x_i) > \lambda\}$$

At the end of a prediction experiment, N—the number of target earthquakes observed during the experiment—is known. For example, if we have gridded a region into j distinct bins, we write

$$N = \sum_{i=1}^{j} N(\mathbf{x}_i) \tag{4.1}$$

where $N(\mathbf{x}_i)$ is the number of target earthquakes bin $\mathbf{x}_i$. The number of hits, h, is the number of target earthquakes located inside A, and the miss rate is:

$$\nu = \frac{N - h}{N} \tag{4.2}$$

The fraction of space-time occupied by alarm is:

$$\tau = \sum_{x_i \in A} \tilde{p}(x_i) \tag{4.3}$$

For any threshold $\lambda \geq \sup\{f(\mathbf{x})\}$, no alarm is declared and all events are missed: $(\tau, \nu) = (0, 1)$. Likewise, for any threshold $\lambda < \inf\{f(\mathbf{x})\}$, all of R is an alarm region and no events are missed: $(\tau, \nu) = (1, 0)$. We can repeat the threshold process for many different thresholds and obtain what we call a **Molchan trajectory**, the set of (τ, ν) points on [0, 1] x [0, 1] that completely characterize the performance of the alarm function during the experiment. Without any loss of information, we can reduce this set to only the set of points where one or more hits occur, points which we call **Molchan trajectory jumps**. We write this reduced Molchan trajectory as the set of minimum τ values such that a given number of hits is obtained. In other words, τ_i is the minimum fraction of space-time that the alarm function must occupy to obtain i hits:

$$\{\tau_i = \inf(T) \mid \nu = \nu_i, i \in [1, N]\}$$

Here, T is the set of τ values from the complete Molchan trajectory, and we use the following indexed notation to specify the miss rate:

$$\nu_i = \frac{N - i}{N} \tag{4.4}$$

We can also express the Molchan trajectory in terms of miss rate as a stepwise continuous function of τ:

$$\nu_f(\tau) = \sup\{\nu_i \mid H(\tau < \tau_i) = 1\} \tag{4.5}$$

Here, H is the Heaviside function. We prove in Appendix B that the expected value for a Molchan trajectory jump for an unskilled alarm function is:

$$\langle \tau_i \rangle = \frac{i}{N+1} \tag{4.6}$$

We note that Equation 4.6 tells us that for any given observational experiment the canonical Molchan trajectory diagonal ($<\tau_i> = i/N$) does not represent the average behavior of an unskilled alarm function for a given experiment; rather, the diagonal should be replaced by a stairstep function starting at (τ, ν) = (0, 1) with stairs of width $1/(N+1)$ and height $1/N$ (see Figure 4.1).

Using the Molchan diagram and its confidence bounds to evaluate an entire alarm function can yield ambiguous results. In particular, an alarm function may yield some alarm sets that demonstrate significant skill (i.e., trajectory points outside the confidence bounds) and some alarm sets that demonstrate otherwise (trajectory points well within the confidence bounds). To naïvely address this problem, one might choose a specific value of τ, or a specific value of ν, at which to examine the trajectory. This process is subjective, however, and does not fully characterize the alarm function. Therefore, in the following section, we suggest a scalar cumulative measure that depends on multiple Molchan trajectory points.

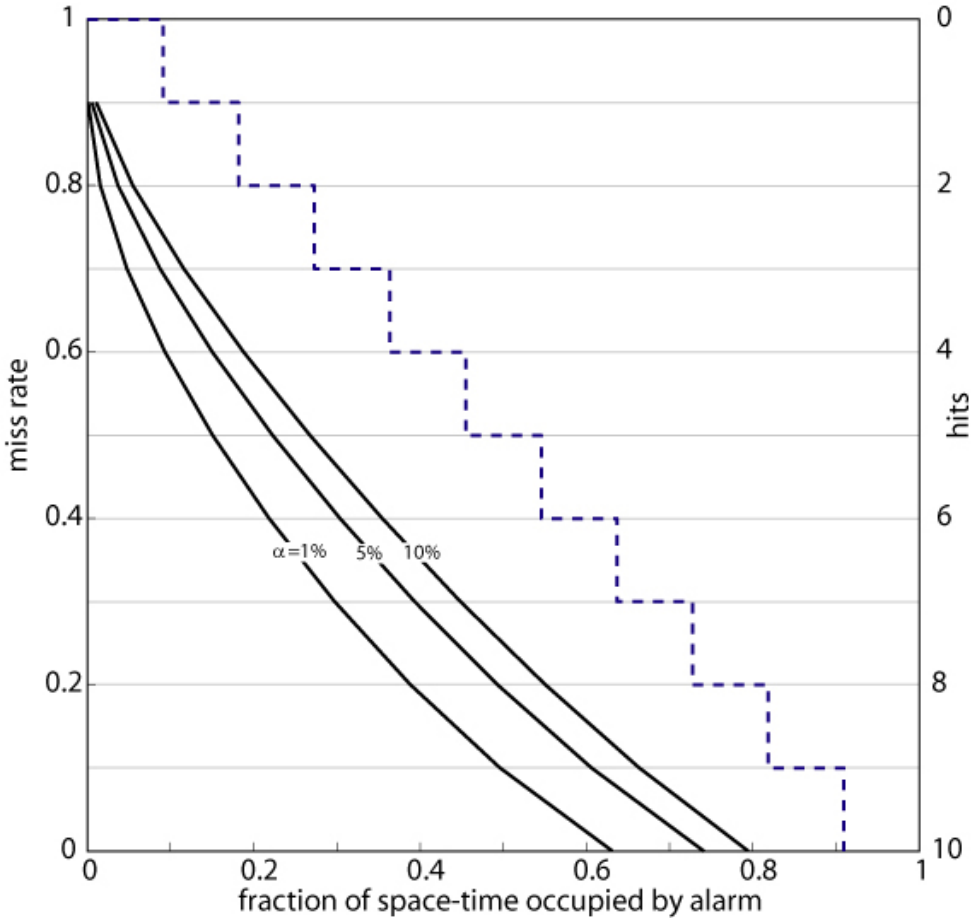

Figure 4.1 Schematic Molchan diagram for N=10. The stairstep proto-diagonal represents the long run average behavior of an unskilled alarm function. Also shown are the 90%, 95%, and 99% confidence bounds.

4.3 Area skill score

Recall from Chapter 3 that we define the area skill score for alarm function *f*:

$$a_f(\tau) = \frac{1}{\tau}\int_0^{\tau}[1 - \nu_f(t)]dt \tag{4.7}$$

This is the normalized area above the continuous Molchan trajectory ν_f up to the given value of τ. For an experiment with *N* target earthquakes, the area skill score evaluated at τ=1 measures the predictive skill of *f* throughout the entire space of the experiment—that is, all *N* target earthquakes and the entire forecast region *R* are considered. Evaluating the

area skill score of the entire trajectory addresses how well an alarm function estimates the distribution of target earthquakes, rather than how well it predicts individual earthquakes. In this case, we can write (see also Figure 4.2):

$$a_f(1) = \sum_{i=0}^{N-1} (\nu_i [\tau_{i+1} - \tau_i]) \tag{4.8}$$

By substituting Equation 4.4 into Equation 4.8 and combining terms, we find

$$a_f(1) = 1 - \frac{1}{N} \sum_{i=1}^{N} \tau_i \tag{4.9}$$

Equation 4.9 shows that the area skill score for an alarm function is proportional to the average of its Molchan trajectory jumps $\{\tau_i\}$.

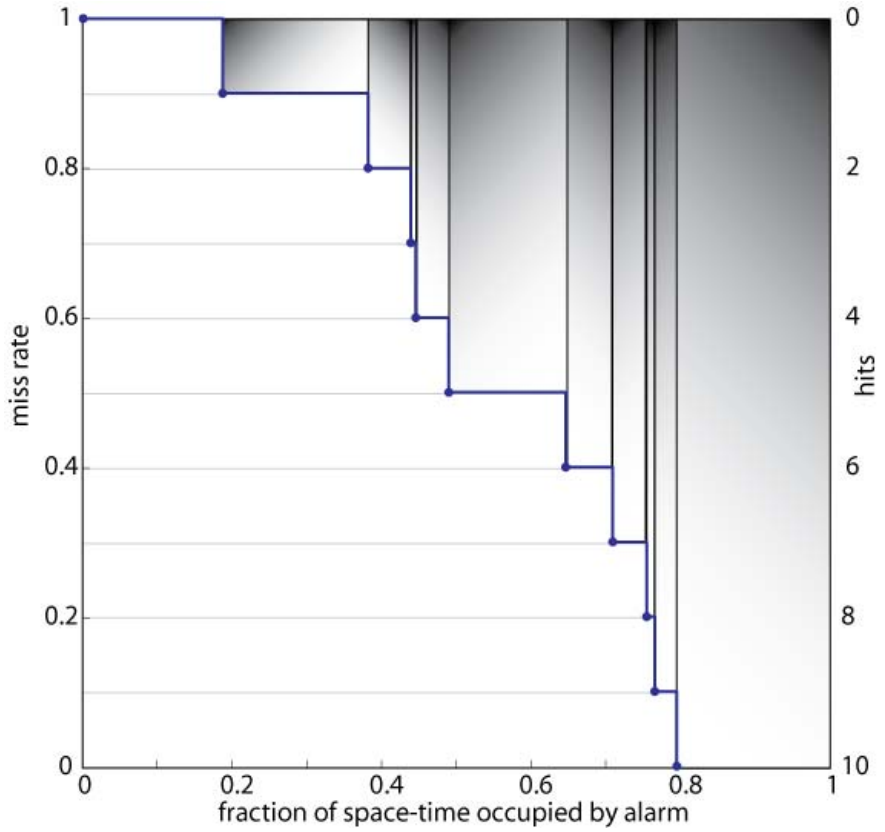

Figure 4.2 Molchan diagram for N=10, shown here with a sample trajectory based on an unskilled alarm function. The shaded region is the area skill score, shown here as a sum over the vertical strips. The dots are the trajectory jumps.

4.4 Area skill score distribution

Hypothesis testing with the area skill score requires knowledge of its distribution for unskilled alarm functions. By unskilled, we mean an alarm function that essentially guesses the future distribution of seismicity according to the reference model, ranking the subregions of R randomly. In practice, we represent an unskilled alarm function by a function whose values are random uniform variables on (0, 1]. A straightforward method for estimating the distribution of the area skill score for a given experiment is to use brute-force simulation: generate a large number of random alarm functions and compute Molchan trajectories and corresponding area skill scores for each random alarm function. This process can become quite computationally cumbersome, particularly as experiment discretization decreases and the number of target earthquakes increases. Fortunately, we can often optimize this simulation method.

Owing to experiment discretization and/or the distribution of target earthquakes, it may occur in some experiments that more than one target earthquake occurs in a single forecast bin. The case of discretized experiments and, in particular, the case in which more than one target earthquake may occur in a single bin, are addressed separately in Section 4.6; for the remainder of this section, we consider experiments wherein the reference model is a continuous function and therefore any value of τ can be realized. In this case, the Molchan trajectory for an unskilled alarm function can be considered as an ordered sequence of independent and identically distributed (i.i.d.) uniform random variables on (0, 1]. That is, rather than simulating many random alarm functions and computing a Molchan trajectory for each alarm function, we can repeatedly select N

uniform random variables on (0, 1], where N is the number of observed target earthquakes. For each simulation, we sort these values in ascending order and analyze their distribution. To understand the equivalence of these methods, consider an experiment with N target earthquakes. In the former method, because the alarm functions randomly rank the sub-regions of the study region R, the resultant Molchan trajectory points will be random samples from (0, 1]. Therefore, the latter method is equivalent and offers a simple computational shortcut.

We can use Equation 4.6 to determine the average area skill score for unskilled alarm functions. We recall the following properties of expectation:

$$\begin{aligned} \langle \mathrm{c}x \rangle &= \mathrm{c}\langle x \rangle \\ \langle x + y \rangle &= \langle x \rangle + \langle y \rangle \end{aligned} \tag{4.10}$$

where c is a constant. By applying these properties to Equation 4.9, we determine the expected value of the area skill score for unskilled alarm functions:

$$\begin{aligned} \langle a_f(1) \rangle &= 1 - \frac{1}{N}\left[\frac{1}{N+1} + \frac{2}{N+1} + \ldots + \frac{N}{N+1}\right] \\ &= 1 - \frac{1}{N}\left[\frac{N(N+1)}{2(N+1)}\right] \\ &= \frac{1}{2} \end{aligned} \tag{4.11}$$

Given that the Molchan trajectory for an unskilled alarm function can be treated as an ordered sequence of i.i.d. uniform random variables on (0, 1], and having shown that the area skill score is proportional to the normalized sum of these variables (Equation 4.9), we write the additive complement of the area skill score, or the area under the Molchan trajectory:

$$1 - a_f(1) = \hat{a} = \frac{1}{N}\left[\tau_1 + \tau_2 + \ldots + \tau_N\right]$$
$$N\hat{a} = u = \left[\tau_1 + \tau_2 + \ldots + \tau_N\right] \tag{4.12}$$

We claim (and prove analytically in the following section) that the area skill score distribution is symmetric. Therefore, to know the distribution of the area skill score, it suffices to know the distribution of the complement $\hat{a}$; in turn, we can obtain the distribution of $\hat{a}$ if we know the distribution of *u*. The distribution of *u*—namely, the distribution of the sum of *N* uniform random variables on (0, 1]—is known (e.g., Sadooghi-Alvandi *et al*. 2007) and, in terms of probability density, is described by the following:

$$f(u) = \frac{1}{(N-1)!}\sum_{k=0}^{\lfloor u \rfloor}(-1)^k \binom{N}{k}(u-k)^{N-1} \tag{4.13}$$

Here, $\lfloor u \rfloor$ denotes the floor function. The variable *u* is defined over (0, *N*] but we seek the distribution of $\hat{a}$, which is defined over [0, 1), so we need to rescale *f*(*u*). In general, if we know $f_1(x)$—the probability density of *x*—and we want to know $f_2(y)$—the probability density of *y*—where $y = g(x)$, then we can use the following:

$$f_2(y) = \frac{1}{g'\left(g^{-1}(y)\right)} f_1\left(g^{-1}(y)\right) \tag{4.14}$$

Here g' is the first derivative of *g*. In terms of Equation 4.14, we know $f_1(x)$ (where $x = u$) and want to know $f_2(y)$ (where $y = \hat{a}$):

$$g(x) = \frac{x}{N} \Rightarrow g^{-1}(x) = Nx \tag{4.15}$$

$$g'(x) = \frac{1}{N} \tag{4.16}$$

By combining Equations 4.12—4.16, we find:

$$f(\hat{a}) = Nf(u)$$
$$\Rightarrow f(\hat{a}) = \frac{N}{(N-1)!}\sum_{k=0}^{\lfloor N\hat{a} \rfloor}(-1)^k \binom{N}{k}(N\hat{a}-k)^{N-1} \tag{4.17}$$

We can use Equation 4.17 to compute the cumulative density for any area skill score for arbitrary N, thereby establishing the statistical significance of any area skill score:

$$F(\hat{a}) = \int_0^{\hat{a}} f(\hat{a})d\hat{a} \tag{4.18}$$

We note that, by applying the Central Limit Theorem to the i.i.d. trajectory values, the area skill score distribution asymptotically approaches a normal distribution with mean μ=½ and variance σ^2 that depends on N; in the next section, we provide an analytic solution for σ^2. For larger values of N, the normal approximation is computationally advantageous compared to the exact solution provided by Equation 4.17 and the simulation methods described above. In the following section, we quantify the accuracy of the Gaussian approximation through a discussion of the moments of the area skill score distribution.

4.4 Higher moments of the area skill score distribution

The exact area skill score distribution described in Equation 4.17 can be better understood through an examination of its moments. Because the Central Limit Theorem allows us to approximate the exact distribution with a normal distribution, we need to determine the second central moment—the variance—in order to fully specify the normal approximation. The fourth moment is also of particular interest because it allows us to

quantify the difference between the exact distribution and the Gaussian approximation.

For any distribution, the n^{th} central moment $\hat{\mu}_n$ can be expressed in terms of moments about the origin, μ_k (Abramowitz & Stegun 1965, p 928):

$$\hat{\mu}_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \mu_k (\mu_1)^{n-k} \tag{4.19}$$

In the case of the area skill score distribution, we have found a general method for computing any moment about the origin; the details are provided in Appendix C, and the first four moments about the origin are listed in Table 4.1.

Table 4.1: Moments about origin

n	μ_n
1	$\frac{1}{2}$
2	$\frac{1}{4} + \frac{1}{12N}$
3	$\frac{1}{8} + \frac{1}{8N}$
4	$\frac{1}{16} + \frac{1}{8N} + \frac{1}{48N^2} - \frac{1}{120N^3}$

Table 4.1 The first four moments about the origin of the area skill score distribution.

By substituting values from Table 4.1, we find

$$\hat{\mu}_1 = \frac{1}{2}$$

This is in agreement with the expected value we found in Equation 4.11; additionally, we find the second central moment to be

$$\hat{\mu}_2 = \sigma^2 = \frac{1}{12N} \tag{4.20}$$

We can combine Equation 4.11 and Equation 4.20 to express the Gaussian approximation to the area skill score distribution, suitable for large N:

$$\tilde{f}(a) = \frac{6N\sqrt{\frac{\pi}{6N}}}{\pi} \exp\left(-6N\left(a - \tfrac{1}{2}\right)^2\right) \tag{4.21}$$

Again using the entries in Table 4.1 and Equation 4.19, we find the third central moment to be:

$$\hat{\mu}_3 = 0 \tag{4.22}$$

Equation 4.22, which describes the skewness of the distribution and measures its asymmetry, serves as proof that the area skill score distribution is symmetric.

To approximate how large *N* must be to use Equation 4.21 (the Gaussian approximation) in the place of Equation 4.17 (the exact distribution), we are interested in quantifying the differences in the distributions these equations describe. These distributions have identical central moments up to and including the third moment. To measure the deviation from normality of the distribution described in Equation 4.17, we therefore derive the kurtosis excess γ_2, which is dependent on the second and fourth central moments and is defined (Abramowitz & Stegun 1965, p 928):

$$\gamma_2 = \frac{\hat{\mu}_4}{\left(\hat{\mu}_2\right)^2} - 3 \tag{4.23}$$

Kurtosis describes the peakedness of a probability distribution and the normal distribution has a kurtosis of 3; the kurtosis excess of a given distribution indicates how much more peaked the distribution is. For example, the kurtosis excess of the uniform distribution is -1.2 and the Laplace distribution has a kurtosis excess of 3. Using Equation 4.19 and the terms in Table 4.1, we find the fourth central moment of the exact area skill score distribution is

$$\hat{\mu}_4 = \frac{1}{48N^2} - \frac{1}{120N^3} \tag{4.24}$$

By substituting Equation 4.20 and Equation 4.24 into Equation 4.23, we determine that the kurtosis excess of the area skill score distribution is

$$\gamma_2 = \frac{-5}{4N} \tag{4.25}$$

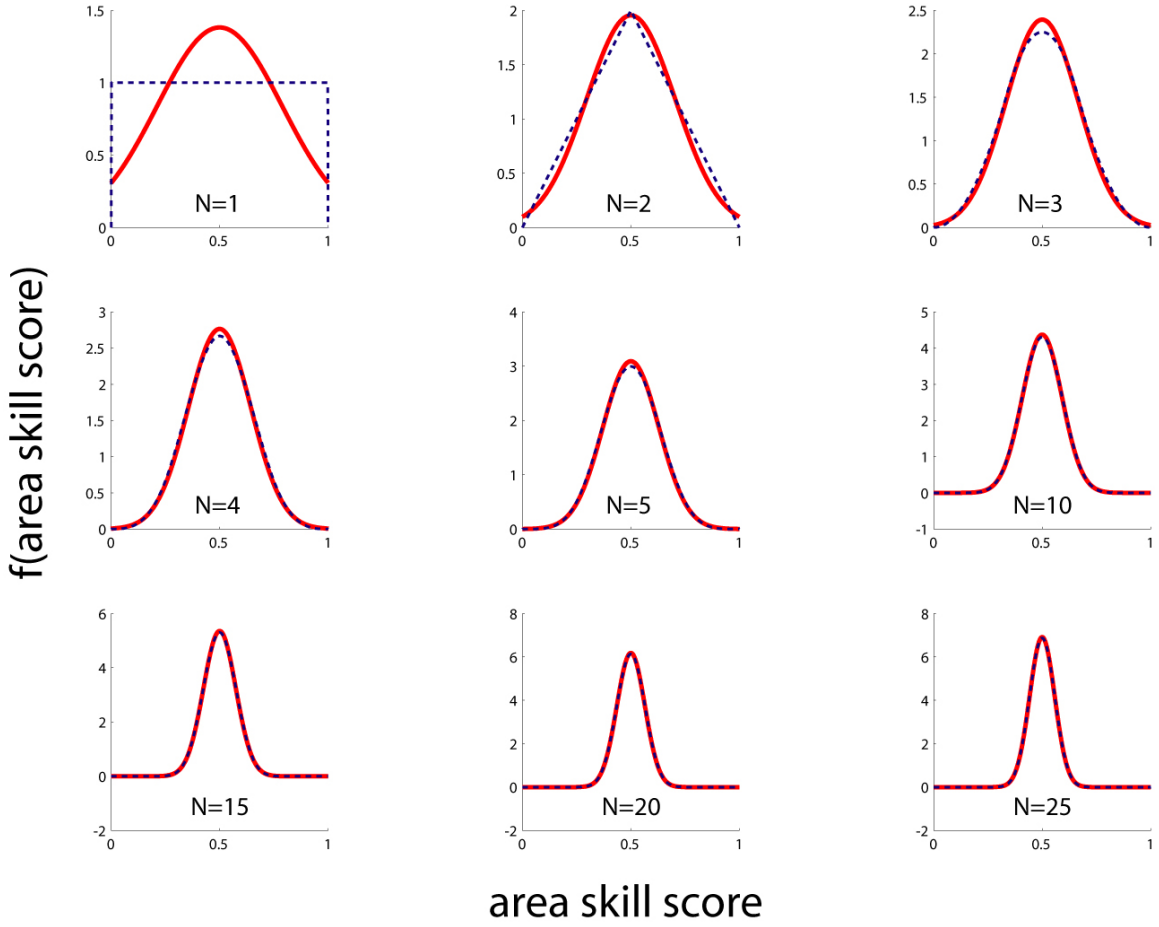

Figure 4.3 Comparison of exact area skill score probability density given by Equation 4.17—shown here as dashed blue curves—and the Gaussian approximation given by Equation 4.21—shown here in red. As N, the number of target events being considered, increases, the Gaussian approximation quickly approaches the exact density.

Equation 4.25 shows that the exact area skill score distribution is platykurtic—that is, it has a negative kurtosis excess—which indicates "thin tails" relative to the normal distribution. Indeed, this must be the case because the range of the area skill

score distribution is [0, 1), whereas the normal distribution has infinite range. This analysis also shows that as the number of observed target earthquakes increases, the kurtosis excess approaches zero, in agreement with a Central Limit Theorem application that suggests the distribution asymptotically approaches normality. Figure 4.3 shows how the approximation differs from the exact solution for several values of N, indicating that for N as small as 5, the normal approximation provides a satisfying estimate.

4.5 Experimental discretization

In the two previous sections, for the purpose of deriving analytic results, we have considered the distribution of the area skill score only in the case where the reference alarm function was assumed to be continuous. In practice, however, this is an unlikely case as predictability experiments almost always deal with discretized regions to reduce computations and to informally address uncertainties (e.g., epicenter uncertainties). Under certain circumstances, despite discretization, the analytic solutions and approximations presented in the previous sections can provide accurate estimates of predictive skill. As experiment cell size decreases toward the continuum limit, and in the case where no one space/time/magnitude cell contains more than one target earthquake, the analysis above becomes increasingly accurate. In the case of more coarsely-grained experiments, however, we rely on simulation methods as previously suggested; here, we discuss some caveats that ought to be considered when computing the significance of a given area skill score. To illustrate these caveats, we will refer to the alarm function shown in Figure 4.4 and the experiment results shown in Figure 4.5.

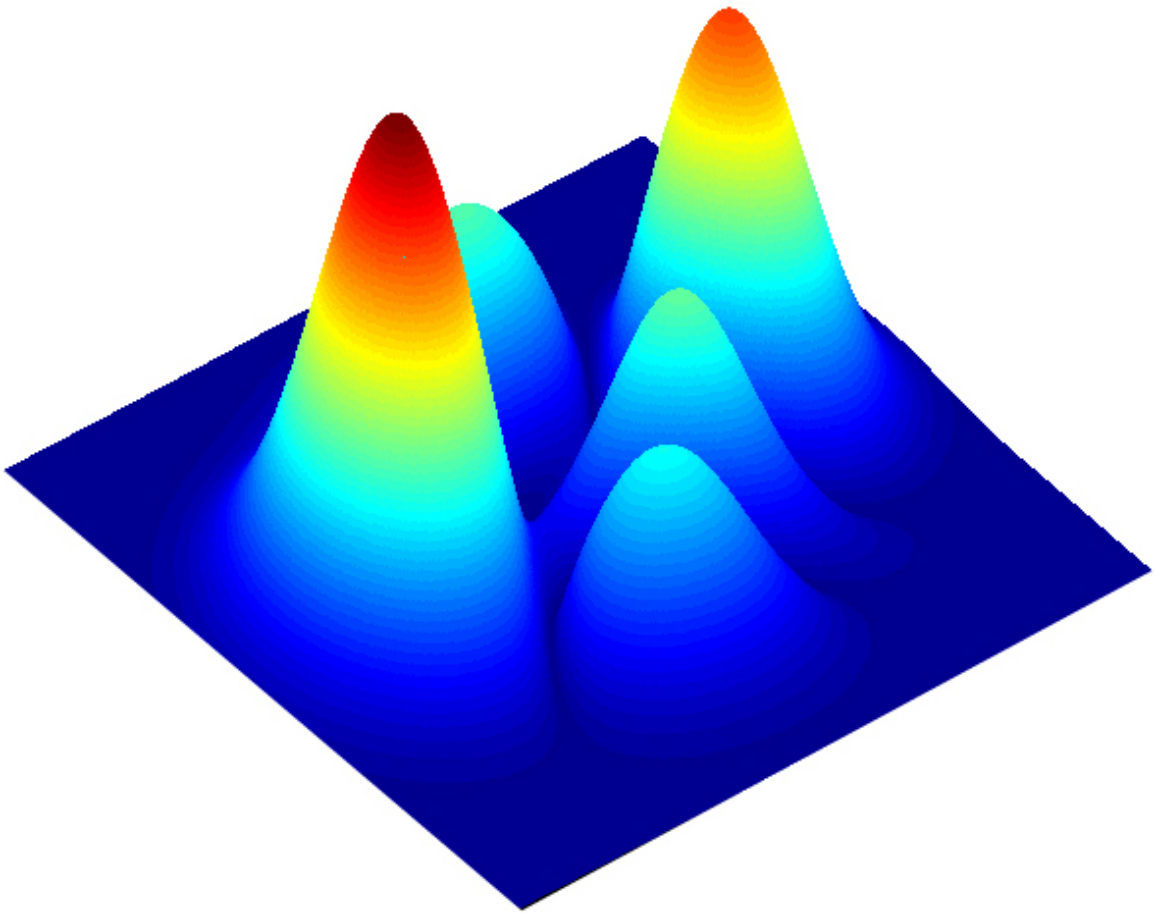

Figure 4.4 Illustrative alarm function, shown here in continuous form.

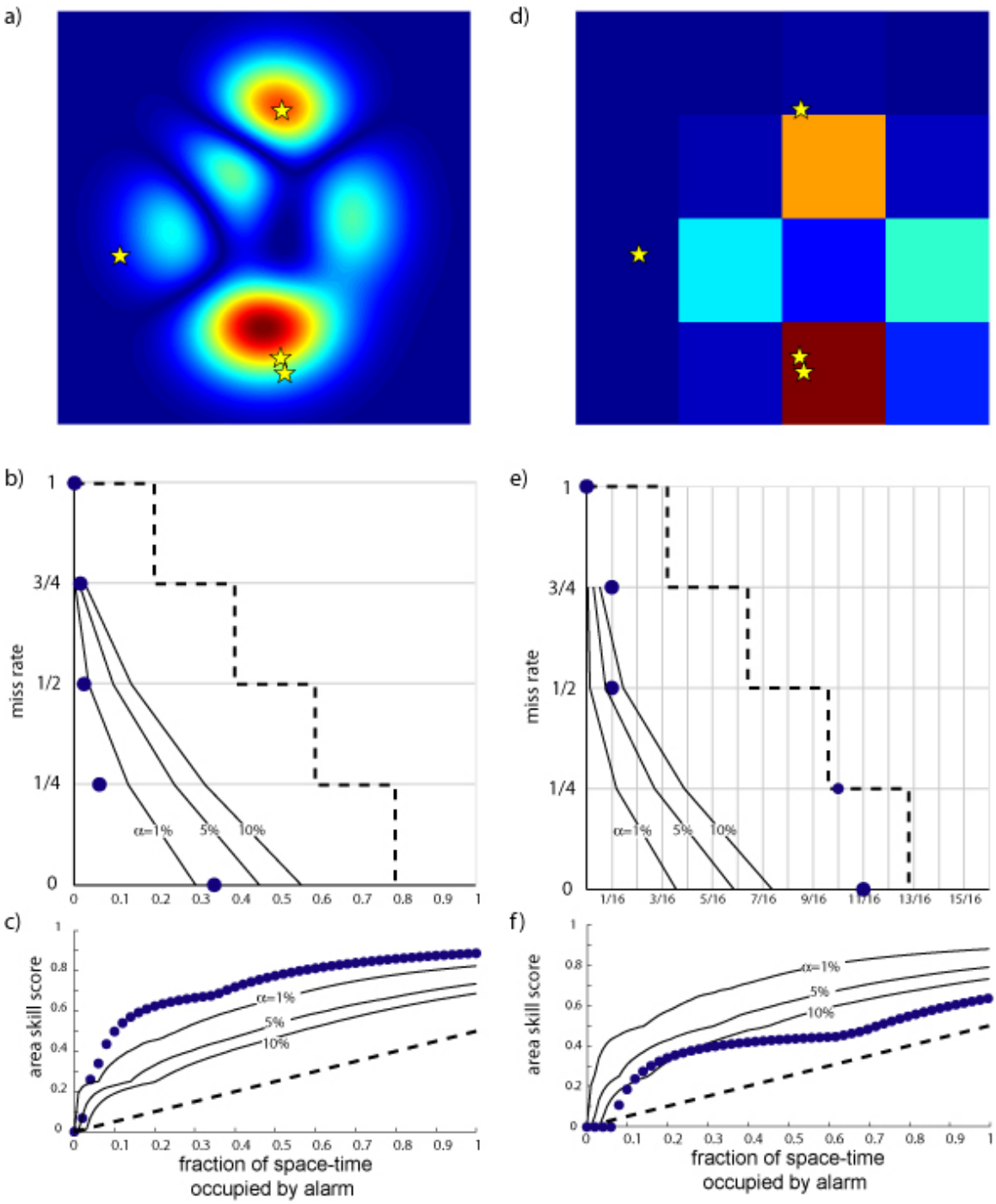

Figure 4.5 Hypothetical predictability experiments comparing the results between the continuous version (left column) and a discretized version (right column) of the alarm function shown in Figure 4.4. **a)** Map view of the continuous alarm function, with four hypothetical target earthquakes denoted by stars. **b)** Molchan trajectory (blue points) for the experiment shown in a), using a uniform reference model. **c)** Area skill score trajectory corresponding to experiment shown in a). **d)** Map view of a very coarsely discretized version of the alarm function. **e)** Molchan trajectory for the experiment shown in d) Note that all values of both τ and ν are now discrete, and that the trajectory has infinite slope at τ=1/16, owing to two target earthquakes being in the same cell. **f)** Area skill score trajectory corresponding to experiment shown in d).

In the case of a uniform reference model and a discrete alarm function f defined on study region R with j distinct sub-regions (e.g., Figure 4.5.d),

$$f(\mathbf{x}) : \mathbf{x} \in R$$
$$R = \{r_1, r_2, \ldots r_j\}$$

the only attainable values of τ are members of the following set:

$$\left\{\tfrac{1}{j}, \tfrac{2}{j}, \ldots, 1\right\}$$

This is illustrated in Figure 4.5.e. If we use the computational shortcut described in the Section 4.4, we violate this constraint and therefore may obtain incorrect results, results which become less accurate as j decreases. We can return to the original simulation procedure, or we can modify slightly the computational shortcut; rather than drawing N random numbers uniformly distributed on (0, 1], we draw N random non-negative integers uniformly distributed on $(0, j]$ and divide each by j.

In principle, we could modify this shortcut for any arbitrary reference model, where the only attainable values of τ are given by the nonzero sums of the reference model values. If we construct the set of all reference model value sums, we could draw N entries from this set to simulate the trajectory from an unskilled alarm function. Constructing this set is prohibitively expensive, however, particularly when dealing with a reference model defined by thousands of cells. As we prove in Appendix D, if our reference model has j values, the set of sums has $(2^j\text{-}1)$ elements. As j becomes large, it is more efficient to use the original alarm function simulation method. There is some trade-off here, though. For a fixed reference model alarm function, as j becomes large, the set of attainable τ values approaches the continuum between 0 and 1. In practice, for j on the

order of a thousand or more, the computational shortcuts we suggest offer a good trade-off between approximation accuracy and speed.

There is one important special case remaining: the case when discretization and the observed target earthquake distribution are such that more than one target earthquake occurs in a single bin (e.g., Figure 4.5.d). When this happens, the probabilities of correctly predicting these events are not independent of each other—this independence is an implicit assumption in the computational shortcuts suggested so far. To correct for this, we can examine the target earthquake distribution and construct the simulated unskilled trajectory appropriately: for a bin containing more than one earthquake, we draw a random number from (0, 1] and append it to the simulated unskilled trajectory. Unlike in the computational shortcut described in Section 4.4, however, rather than appending this random number once and moving on, we append the random number n times, where n is the number of target earthquakes in this bin. This method captures the idea that, when this bin is covered by an alarm, all the target earthquakes within the bin are successfully predicted.

We note that, in practice, almost all predictability experiments take place in a discretized, finely gridded framework. In all cases, the alarm function simulation procedure will be accurate and appropriate but, as we have pointed out, it can be computationally cumbersome. A careful examination of the experimental discretization, the target earthquake distribution, and the reference model should be conducted prior to evaluation using the area skill score; based on the outcome of this examination, it is very likely that one of the shortcut simulation methods discussed here is applicable. In the rare case of the predictability experiment in a continuum—for example, the RTP

experiment discussed in Chapter 2, the analytic solutions are applicable and, as N becomes large, the Gaussian approximation for determining statistical significance is advantageous.

4.6 Discussion

Imagine we have a set of n candidate alarm functions, $\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$, and we want to determine whether any has predictive skill in a given experiment. We can begin with a **uniform test** by selecting a uniform reference model ($\tilde{p}(\mathbf{x}) = c, \ \forall \mathbf{x} \in R$, where c is any arbitrary constant), computing the area skill scores $\{a_{f_1}(1), a_{f_2}(1), \ldots a_{f_n}(1)\}$, and their significance using Equation 4.18. Because earthquakes cluster in space and time and do not occur everywhere, it is likely that most candidate alarm functions will incorporate some form of clustering and thus obtain a statistically significant area skill score relative to the uniform reference model. Therefore, the goal of further testing is to improve the reference model and thereby distinguish the alarm functions. For all those candidate alarm functions that pass the uniform test, we continue with the **self test**. In the self test of alarm function f_1, we take f_1 as the reference model $\tilde{p}(\mathbf{x}) = f_1(\mathbf{x}), \ \forall \mathbf{x} \in R$ and recalculate its Molchan trajectory and area skill score $a_{f_1}(1)$. If f_1 is a reasonable reference model—that is, it approximates the distribution of seismicity well—we expect that $a_{f_1}(1)$ will not deviate significantly from the corresponding area skill score distribution. If it does, this indicates that f_1 is not an appropriate reference model. For all those alarm functions that pass the self test, we proceed with a **round-robin test**. In the round-robin test, each surviving alarm function is fixed as the reference model and the

area skill scores for all other alarm functions are computed. An alarm function is supported as a good reference model if none of the area skill scores deviate from the expected distribution. If any one area skill score deviates significantly, the reference alarm function is considered to be an inappropriate reference model.

4.7 Conclusion

In this chapter, we have explored the concept of an alarm function, a general format for specifying earthquake forecasts that is defined by an ordering of space/time/magnitude regions in terms of their estimated probability to contain future target earthquakes. We described the Molchan diagram and Molchan trajectories, and presented relevant analysis that includes a corrected proto-diagonal describing the behavior of unskilled alarm functions and the explicit use of a reference model. We presented the exact distribution of the area skill score. We also presented a Gaussian approximation of the area skill score distribution and, to determine the applicability of this approximation, performed an analysis of the moments of the exact distribution. We have dealt carefully with potential pitfalls regarding experimental discretization and the special case of more than one target earthquake occurring in a given space/time/magnitude bin.

CHAPTER FIVE:
Optimizing earthquake forecasts based on smoothed seismicity

Abstract

Recently, the momentum of earthquake prediction research has shifted from predicting individual large earthquakes to a slightly different problem: forecasting the distribution of seismicity—including small to moderate events—in space, time, and magnitude. This change has been accompanied by an increased emphasis on rigorous testing and evaluation of earthquake forecasts. The development of reasonable reference models is vital to earthquake forecast evaluation, and improving standard reference models is paramount to improving our understanding of the earthquake system. Perhaps the simplest reasonable reference model is one based on the distribution of past earthquakes; such a model incorporates the intuition that future earthquakes will occur in regions near past earthquakes, or the corresponding notion that each earthquake has some influence on the tectonic system in which it occurs. Taking this approach typically involves smoothing the distribution of observed past earthquakes with a given kernel. In this study, we seek a smoothing kernel that is optimal, in the sense that this kernel yields the best forecast of future seismicity. We present a framework in which time-varying, scale-dependent smoothing kernels can be optimized. We illustrate the procedure with a series of retrospective predictability experiments in California, taking $M \geq 5$ events as the target earthquakes and attempting to predict their locations using previous $M \geq 4$ events. We present analytic solutions for three functional forms of smoothing and consider experiments with a number of smoothing lengthscales. Our results indicate that, for these experiments, the optimal smoothing lengthscale slowly decreases through time. These

experiments have yielded a new five-year forecast that can be considered as a simple reference model for evaluating RELM models currently being tested within the SCEC CSEP testing center.

5.1 Smoothed seismicity reference models

Evaluation of earthquake forecast experiments requires careful consideration of appropriate reference models. For example, because earthquakes cluster in space, a forecast that incorporates some form of spatial clustering is likely to perform significantly better than a reference model based on a uniform seismicity distribution. Therefore, it seems reasonable to expect that a reference model captures the clustering observed in seismicity (e.g., Michael 1997). One method for incorporating observed clustering is to smooth past seismicity; that is, to allow each past earthquake in the catalog to make some smoothed contribution to an estimate of seismic density. Smoothing can take a number of forms and it is not clear which form is best.

In Chapters 3 and 4, we introduced a performance metric called the area skill score for evaluating earthquake forecasts. Testing based on the area skill score explicitly specifies the reference model in terms of an alarm function, which is a general method for ranking regions of space/time/magnitude in terms of the probability that a future target earthquake will occur within a given region. A wide class of earthquake forecasts can be interpreted in terms of alarm functions, and this allows for rigorous comparative testing. One important goal of such testing is an iterative improvement of the reference model and indeed, our understanding of earthquake predictability is likely to proceed in this manner. In experiments such as that described in Chapter 3, we have found that quite

complex prediction algorithms typically do not significantly outperform very crude smoothed seismicity models. This has inspired us to essentially invert the testing problem: rather than building increasingly complex models to forecast seismicity, can we optimize simple smoothed seismicity models? In this study, we consider simple forms of smoothing and take an empirical approach to optimizing them with respect to the area skill score.

Smoothed seismicity is also one of the building blocks of modern seismic hazard analysis, which in a sense is a particular type of earthquake forecast experiment. When estimating seismic hazard in regions with low seismicity rates or where little is known about regional fault structure, the locations of past earthquakes in the region are smoothed. For example, this is the procedure followed by the United States Geological Survey National Seismic Hazard Map (NSHM) working group when estimating hazard in the Central and Eastern United States (Frankel 1995, Frankel *et al*. 1996); in other regions, smoothed seismicity is used to model "background seismicity." While using smooth seismicity reduces the subjectivity inherent in polling expert opinion, the choice to use a Gaussian smoothing kernel seems itself subjective. Moreover, the standard smoothing lengthscale used in the NSHM is justified quite qualitatively, by examining resulting forecasts by eye. Here, we take a more rigorous approach, considering alternative kernels and quantifying the differences between various lengthscale values. The experiments that we describe here could be used to optimize future revisions of such large-scale seismic hazard maps.

5.2 Functional forms of smoothing kernels

To simply smooth seismicity, one might discretize the region of interest and count the number of earthquakes that occur in each cell, allowing each point source epicenter to be smoothed over the cell in which it falls. Rundle *et al*. (2002) called this the Relative Intensity method. The results of such smoothing, however, will be unstable with respect to discretization parameters such as cell size and grid alignment, and the smoothing itself is anisotropic and un-physical. For example, epicenters occurring in opposite corners of a cell are treated as though they both occurred in the center of the cell. To minimize these effects, and to relax the constraint that each epicenter contribute only to the cell in which it occurs, we smooth earthquakes using continuous kernel functions that allow for a wider region of influence. In this study, we explored a simple isotropic two-dimensional smoothing kernel governed by a single lengthscale parameter. It was our goal to optimize the smoothing lengthscales. For a given functional form of smoothing, our primary task is to calculate the contribution of an earthquake epicenter at (x_{eqk}, y_{eqk}) to a grid cell with bounds x_1, x_2, y_1, y_2. In general, for the bivariate smoothing kernel $K(x, y)$, this contribution takes the form:

$$K\left(x_{eqk}, y_{eqk}, x_1, x_2, y_1, y_2\right) = \int_{y_1}^{y_2}\int_{x_1}^{x_2} K\left(x_{eqk}, y_{eqk}\right)dxdy \qquad (5.1)$$

A Gaussian kernel has been used in several previous studies of smoothed seismicity (Stock & Smith 2002a, Stock & Smith 2002b, Helmstetter *et al*. 2007) and it forms the basis for modeling "background" seismicity in the national seismic hazard maps of the United States (Frankel *et al*. 2002) and New Zealand (Stirling *et al*. 2002). Frankel (1995) first formulated Gaussian smoothing for seismic hazard, but that study

formulated the kernel in terms of a correlation distance *c* rather than using a pure Gaussian form, so we will explicitly detail our formulation. Because we are interested in an isotropic kernel with a single smoothing lengthscale, we can write the bivariate Gaussian as:

$$K_\sigma(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{5.2}$$

Here, σ is the classic standard deviation parameter. We substitute Equation 5.2 into Equation 5.1 and solve for the Gaussian contribution of an epicenter to a cell:

$$K_\sigma(x_{eqk}, y_{eqk}, x_1, x_2, y_1, y_2) = \frac{1}{4}\left[\operatorname{erf}\left(\frac{x_{eqk}-x_2}{\sigma\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{eqk}-x_1}{\sigma\sqrt{2}}\right)\right]\left[\operatorname{erf}\left(\frac{y_{eqk}-y_2}{\sigma\sqrt{2}}\right)-\operatorname{erf}\left(\frac{y_{eqk}-y_1}{\sigma\sqrt{2}}\right)\right] \tag{5.3}$$

Analytically, the kernel described by Equation 5.2 extends infinitely in both dimensions. Within the precision provided by modern computers, however, erf(*a*)=sgn(*a*) when |*a*| >~3.86. Therefore, any cell satisfying the following conditions will cause Equation 5.3 to equal zero; that is, the cell will obtain no contribution from the epicenter:

$$\begin{aligned} &|x_{eqk}-x_2| > 3.86\sigma\sqrt{2} \quad \text{and} \quad |x_{eqk}-x_1| > 3.86\sigma\sqrt{2} \\ &\qquad\qquad \text{or} \\ &|y_{eqk}-y_2| > 3.86\sigma\sqrt{2} \quad \text{and} \quad |y_{eqk}-y_1| > 3.86\sigma\sqrt{2} \end{aligned} \tag{5.4}$$

To save computation, we determine which cells will obtain some contribution from the epicenter before evaluating Equation 5.3 everywhere in the gridded regions. Cells will obtain some contribution from the epicenter if:

$$\begin{cases} x_{eqk} - x_1 < 3.86\sigma\sqrt{2} \\ x_{eqk} - x_2 > -3.86\sigma\sqrt{2} \\ y_{eqk} - y_1 < 3.86\sigma\sqrt{2} \\ y_{eqk} - y_2 > -3.86\sigma\sqrt{2} \end{cases} \tag{5.5}$$

Particularly for small σ and large catalogs, using the condition in 5.5 will provide a significant reduction in computing time. For the smoothing experiments described below, we only consider the kernel contributions described by Equation 5.3. In Appendix E, we present comparable analytic solutions for a power-law kernel and an Epanechnikov kernel.

5.3 Smoothing experiments

The Regional Earthquake Likelihood Model (RELM) working group recently began a prospective earthquake predictability experiment in California (Field 2007, and references therein). Nineteen forecasts were submitted, each of which provides the expected number of earthquakes with magnitude greater than or equal to 4.95 for the subsequent five years; the forecasts are specified in latitude/longitude cells. Some of the submitted forecasts include a component of smoothed seismicity (Ebel *et al*. 2007, Holliday *et al*. 2007, Ward 2007) and others are based solely on smoothed seismicity (Helmstetter *et al*. 2007, Kagan *et al*. 2007). These forecasts employ complex and/or adaptive smoothing kernels, and the smoothing parameters have either been arbitrarily fixed or chosen by maximum likelihood. One goal of this work was to develop a simple smoothed seismicity forecast optimized using the area skill score; it will be interesting to compare such a forecast in a prospective test with those based on more complex smoothing. Evaluation of these forecasts, using either the likelihood testing framework

(Schorlemmer *et al*. 2007) or that proposed in Chapters 3 and 4, provides a level of granularity that might allow us to understand why one forecast outperforms another, thereby indicating what model complexities provide significant benefit.

In order to be formally defined, a smoothed seismicity predictability experiment requires the following to be specified: the geographic region of interest and choice of discretization (i.e., cell size); the magnitude range of target earthquakes and the magnitude range of earthquakes to be used for smoothing—including the magnitude scale and earthquake catalog; the period of earthquakes to be used for smoothing (the learning period); the period of earthquakes to be predicted (the testing period); and the form and parameter values of the smoothing kernel.

In each of the experiments described below, we have chosen parameters that closely match the current 5-year forecast experiment designed by RELM. We consider the California natural laboratory, following the RELM parameterization and spatial discretization of 0.1° by 0.1° latitude/longitude cells (see Table 5.1 for details). We use the Advanced National Seismic System (ANSS) catalog and ignore the magnitude type.

Table 5.1: California study region

Latitude (degrees)	Longitude (degrees)
43.0	-125.2
43.0	-119.0
39.4	-119.0
35.7	-114.0
34.3	-113.1
32.9	-113.5
32.2	-113.6
31.7	-114.5
31.5	-117.1
31.9	-117.9
32.8	-118.4

Table 5.1, Continued

33.7	-121.0
34.2	-121.6
37.7	-123.8
40.2	-125.4
40.5	-125.4

Table 5.1 Coordinates of the California study region considered in this study, after the Regional Earthquake Likelihood Models natural laboratory (Schorlemmer & Gerstenberger 2007).

5.4 Fixed test period

Table 5.2: Target earthquakes

#	Origin Time	Magnitude	Latitude (degrees)	Longitude (degrees)
1	22 Feb 2002	$M_w = 5.70$	32.3188	-115.3215
2	17 Jun 2002	$M_L = 5.09$	40.8098	-124.5520
3	07 Feb 2003	$M_L = 5.00$	31.6280	-115.5110
4	22 Dec 2003	$M_w = 6.50$	35.7002	-121.0973
5	18 Sep 2004	$M_w = 5.55$	38.0095	-118.6785
6	18 Sep 2004	$M_w = 5.40$	38.0187	-118.6625
7	28 Sep 2004	$M_w = 5.96$	35.8182	-120.3660
8	29 Sep 2004	$M_w = 5.00$	35.9537	-120.5022
9	29 Sep 2004	$M_w = 5.03$	35.3898	-118.6235
10	30 Sep 2004	$M_w = 5.00$	35.9880	-120.5378
11	16 Apr 2005	$M_L = 5.15$	35.0272	-119.1783
12	12 Jun 2005	$M_w = 5.20$	33.5288	-116.5727
13	02 Sep 2005	$M_w = 5.11$	33.1598	-115.6370
14	24 May 2006	$M_w = 5.37$	32.3067	-115.2278
15	19 Jul 2006	$M_w = 5.00$	40.2807	-124.4332

Table 5.2 Parameters of earthquakes with magnitude greater than or equal to 5.0 in the California study region during the test period T*=1 Jan 2002 to 31 Dec 2006.

We consider a series of experiments in which we attempt to best forecast California earthquakes of magnitude greater than $m_{\text{target}} = 5.0$ during the test period T^* starting 1 Jan 2002 and ending 31 Dec 2006 (duration $\Delta t^* = 5$ years); the target earthquakes for this experiment are listed in Table 5.2. We denote these as the "fixed test period" experiments. In the first of these experiments, we use the Gaussian kernel given by Equation 5.3 to smooth all earthquakes in this region with magnitude greater than $m_{\text{smooth}} = 4.0$ during a preceding learning period T_1 1 Jan 1997 to 31 Dec 2001, inclusive (duration $\Delta t = 5$ years). We generate ten distinct alarm functions by iterating over each smoothing lengthscale element in the set:

$$\Sigma = \{5, 10, 15, 20, 25, 30, 50, 100, 200, 1000\} \text{ km} \tag{5.6}$$

We repeat this procedure for many distinct learning periods, setting the start date of the learning period progressively earlier ($\delta t = 1$ year), until we reach the beginning of the catalog, 1 Jan 1932. The temporal discretization yields 66 distinct, though not independent, experiments. This procedure is illustrated in Figure 5.1.

We provide an example alarm function, the target earthquake distribution, and the resultant Molchan trajectory and area skill score trajectory in Figure 5.2. In this experiment, we have computed the trajectories relative to a uniform reference model, following the procedures described in Chapter 3. We note that the two target earthquakes near 121W, 36N—earthquakes 4 and 7 as listed in Table 5.2—are not forecast well, as they occur in regions that have obtained no contribution from earthquakes in the

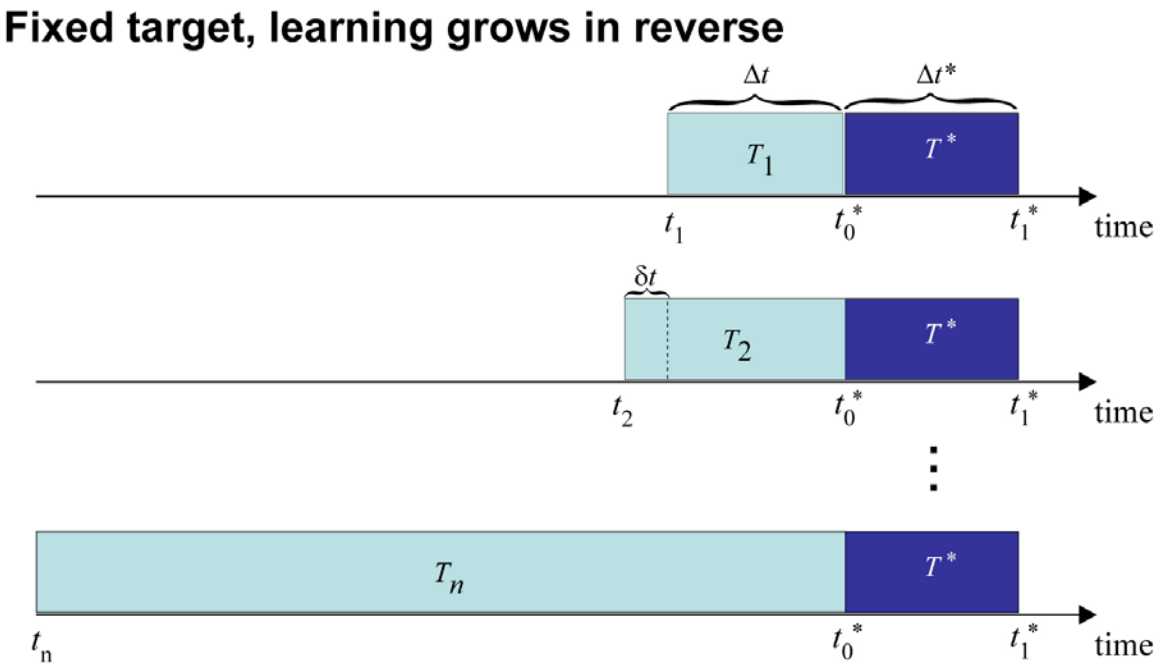

Figure 5.1 Schematic of smoothing optimization experiments with a fixed test period T^*. Light blue indicates a learning period and dark blue indicates the test period—that is, epicenters in the light blue period are smoothed to forecast target earthquakes in the dark blue period. In these experiments, the learning period grows backward in time by δt = 1 year.

preceding learning period. In contrast, the earthquake near 119W, 38N—earthquake 6—occurs near a peak in the alarm function and is thus well forecast (predicted at a small value of τ). This corresponds to a situation where a target earthquake occurred near the location of many previous earthquakes. The prominent alarm function peak near 116W, 35N, however, would yield false alarms because no target earthquake occurs nearby. As indicated by the area skill score at τ=1, this alarm function shows significant skill relative to a uniform reference model.

For each experiment in the series, we want to determine the optimal smoothing lengthscale for that experiment, so we evaluate the performance of each alarm function. We begin by examining the performance of each smoothed alarm function relative to a uniform reference model. For each alarm function, we compute a Molchan trajectory and

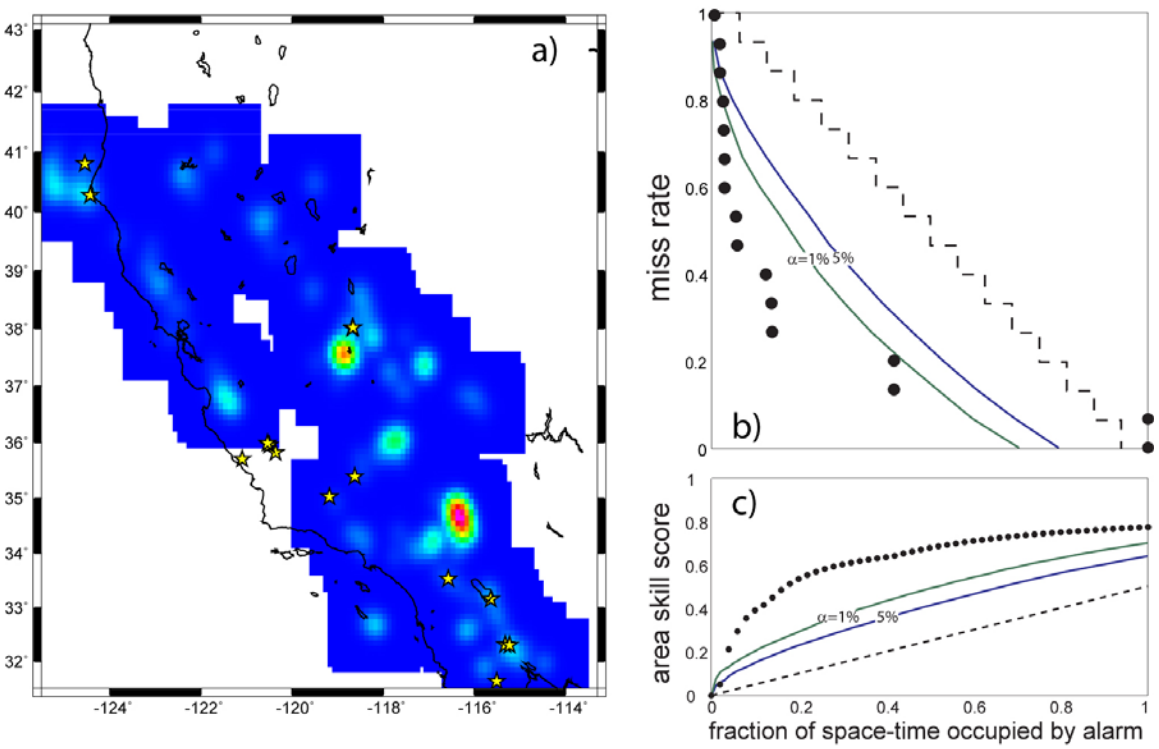

Figure 5.2 Example of smoothed seismicity predictability experiment. **a)** Seismic density alarm function and the target earthquakes for the fixed testing period experiments. In this case, earthquakes during learning period T_1=1 Jan 1997 to 31 Dec 2001 were smoothed with the Gaussian kernel with σ=15km; warm colors indicate high density, white regions indicate zero density. Stars denote target earthquakes that occurred during the test period T^*=1 Jan 2002 to 31 Dec 2006. **b)** Molchan trajectory corresponding to the alarm function and target earthquake distribution shown in a), using a uniform reference mode. Also shown are the 95% and 99% confidence bounds on the random diagonal. **c)** Area skill score trajectory corresponding to the example forecast experiment, with 95% and 99% confidence bounds; the final area skill score indicates statistically significant performance relative to the uniform reference model. We claim that the performance of the alarm function in this experiment can be characterized by the area skill score at τ=1, the right-most point in the area skill score plot.

corresponding area skill score trajectory. As discussed in Chapter 3 and Chapter 4, the performance of an alarm function in a given experiment can be characterized by its area skill score at τ=1; therefore, in Figure 5.3, we report only this measure for each of the 66 experiments. We note that, for every experiment, area skill scores from every smoothed alarm function are better than random with respect to the uniform reference model at 95% confidence. With few exceptions, the smoothed alarm functions obtain area skill scores that are significant at 99% confidence for each experiment. We also note that the area skill scores are monotonically decreasing with increasing smoothing start date. This

indicates that the performance of each smoothing kernel is enhanced by extending the learning period, thus integrating more earthquake data. In other words, these kernels benefit from inclusion of older earthquakes, despite the fact that location uncertainty increases as we approach the beginning of the catalog.

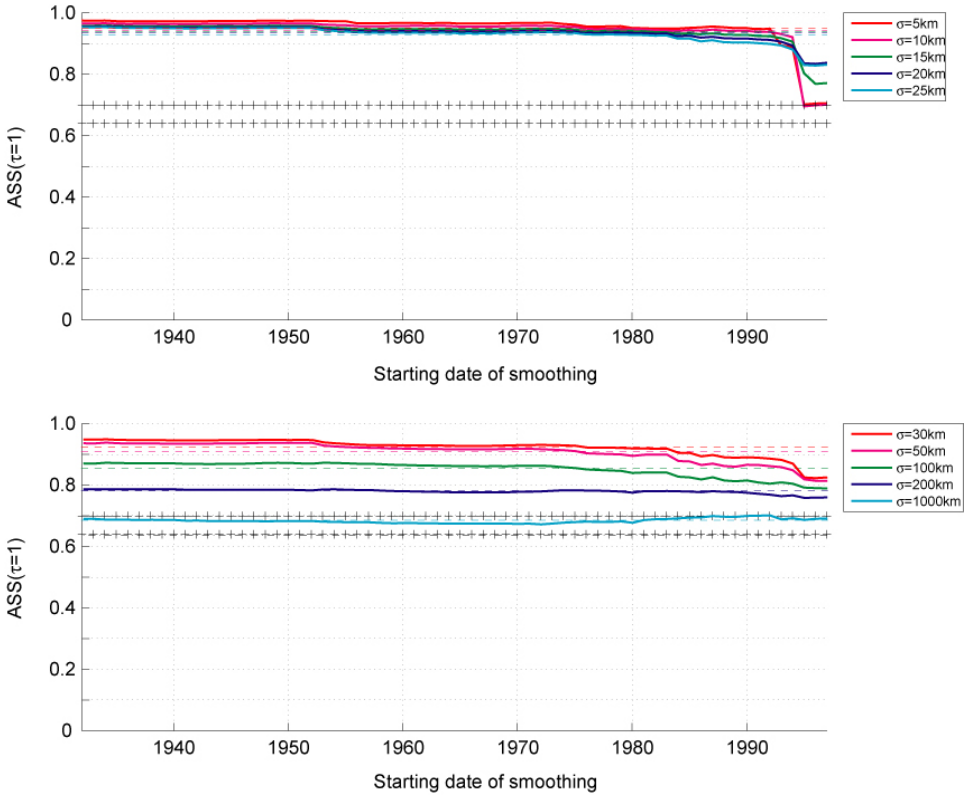

Figure 5.3 Temporal evolution of area skill scores for each lengthscale for each experiment illustrated in Figure 5.1, relative to a uniform reference model. The result of the first experiment in Figure 5.1 is reported at the far right of this plot and subsequent experiments proceed to the left. The two '+' lines demarcate the area skill score confidence bounds corresponding to α=0.05 (bottom) and α=0.01 (top). The dashed lines are the average area skill score at τ=1 over the 66 experiments.

The results in Figure 5.3 indicate that all smoothing lengthscales outperform the uniform reference model, but it is difficult to say from this plot which smoothing lengthscale yields the optimal reference model for each experiment. To address this problem, as before, we compute Molchan trajectories and area skill score trajectories for every experiment; in this step, however, we consider the alarm functions themselves as

reference models, iterating over each one. That is, for the first experiment (learning period from 1997 to 2001 inclusive), we first fix the σ = 5km smoothed alarm function as the reference model and compute trajectories for all smoothed alarm functions, including the σ = 5km alarm functions itself. We next use the σ = 10km smoothed alarm function as the reference model and compute trajectories for all lengthscales, repeating this process until each lengthscale's alarm function has been used as the reference model. We repeat this procedure for the next experiment and for all those remaining, until we have reached the beginning of the catalog. As before, we analyze the results in terms of the area skill score at τ = 1. We provide in Figure 5.4 two examples of results. In Figure 5.4.a we have employed a reference model of σ = 1000km. We note that because this is such a large smoothing value, the results are nearly identical to those of Figure 5.3; in the limit where $\sigma \to \infty$, the smoothed seismicity alarm function becomes the uniform alarm function. In Figure 5.4.b, we set σ = 30km andthe results are quite different due to the differences between this reference model and the uniform distribution. In this case, many of the area skill scores fall below the corresponding critical region and, at least in the case of alarm functions from smaller lengthscales, these results provide a clearer distinction between alarm functions. Again, with little exception, we note a general decrease in performance with an increase of smoothing start date.

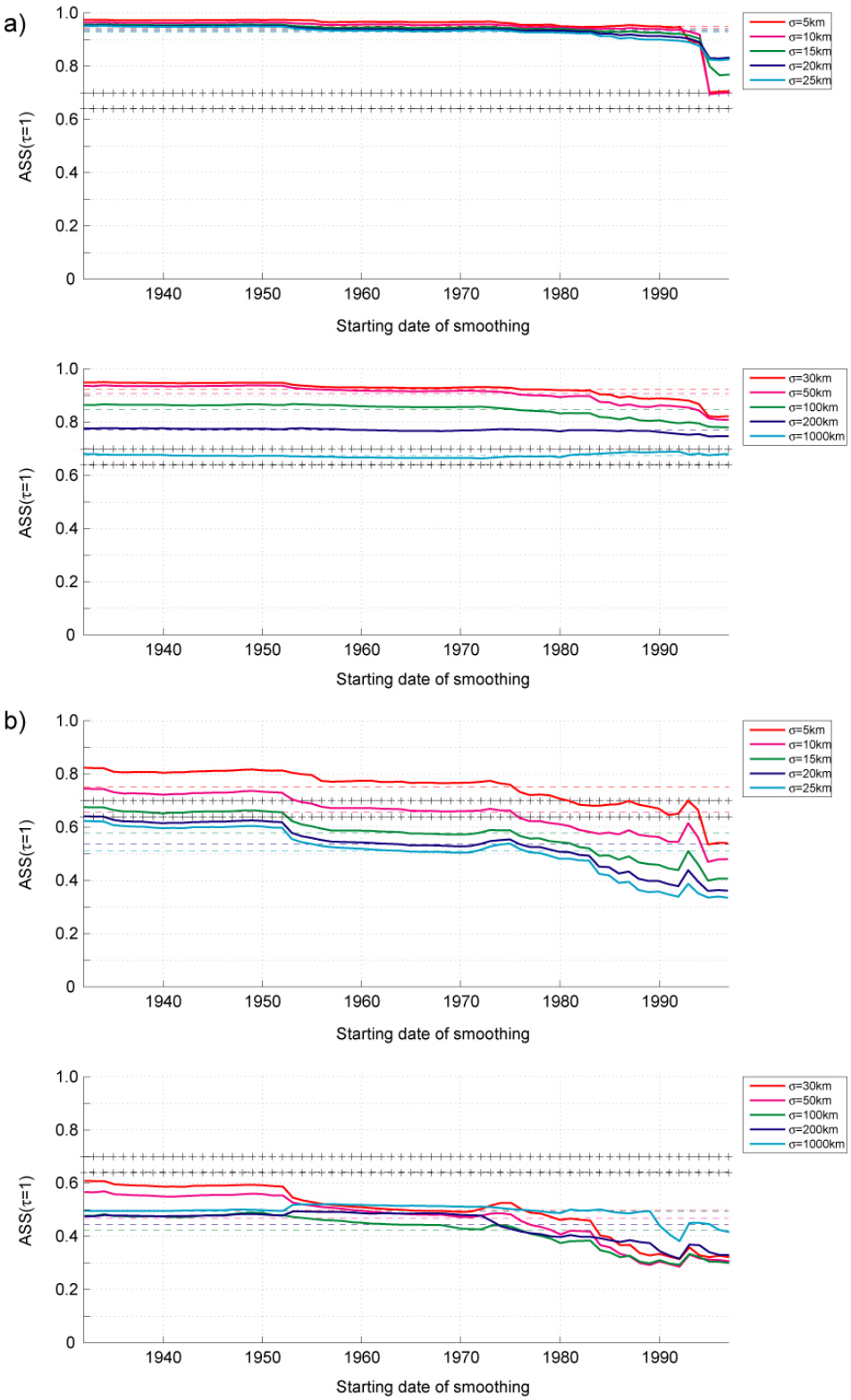

Figure 5.4 Temporal evolution of area skill scores for each lengthscale for each experiment illustrated in Figure 5.1, relative to smoothed seismicity alarm functions with smoothing lengthscale **a)** σ = 1000 km and **b)** σ = 30 km. Plot details are as described in Figure 5.3.

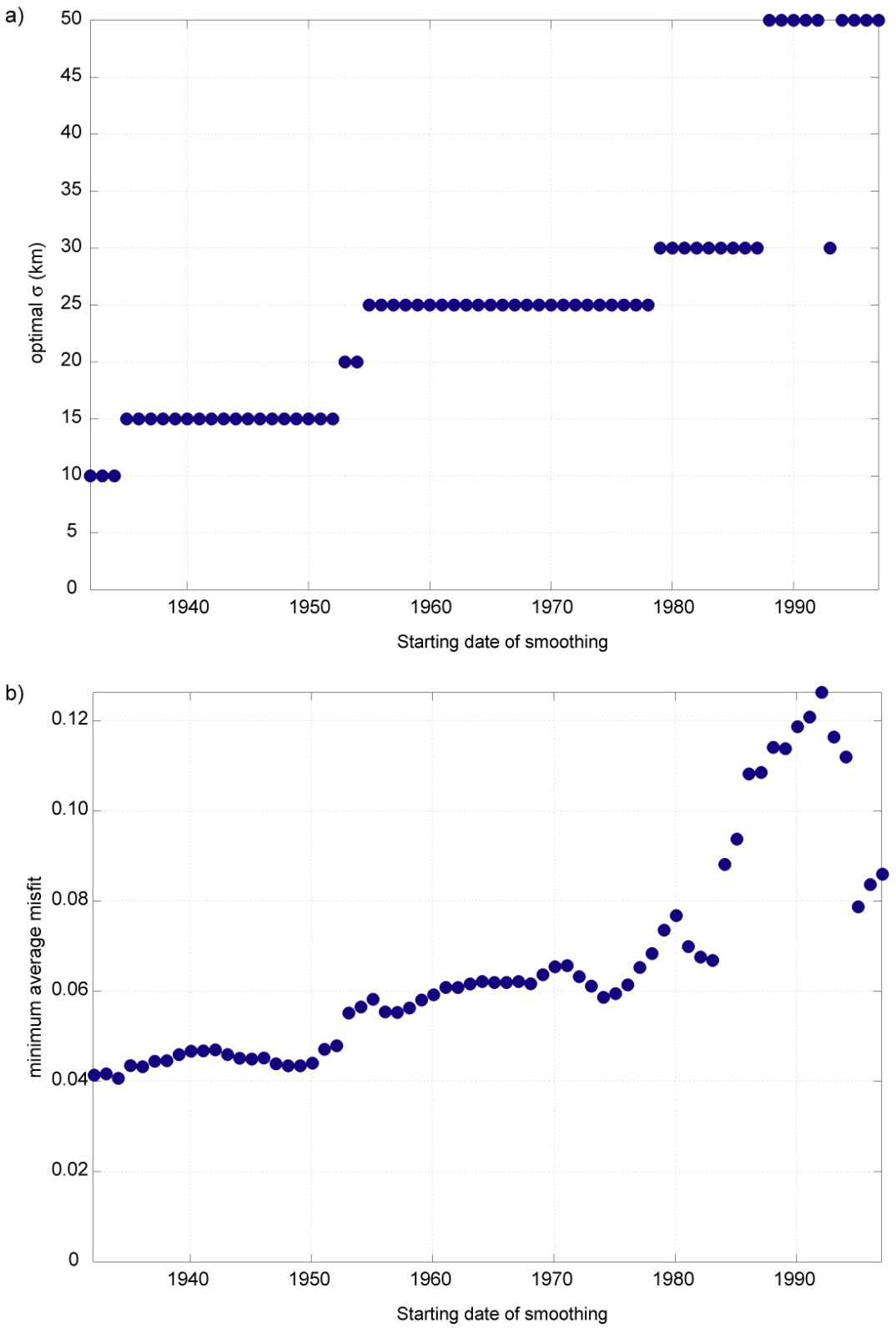

Figure 5.5 Temporal evolution of **a)** the optimal reference model smoothing lengthscale for the experiments illustrated in Figure 5.1 and **b)** the corresponding misfit.

We recall from Chapter 4 that if we have found an appropriate reference model, the area skill scores of competing alarm functions will go to ½. In the present case, we have a set of candidate reference models; to distinguish between them, we define an average misfit parameter that measures the deviation of each alarm function's area skill score from ½:

$$\varphi_{\sigma_j}(t_i) = \frac{1}{n}\sum_{k=1}^{n}\left| ASS(t_i, \sigma_k \mid \sigma_j) - \tfrac{1}{2} \right| \qquad (5.7)$$

Here, brackets denote absolute value, n is the number of competing alarm functions, and $ASS(t_i, \sigma_k \mid \sigma_j)$ is the area skill score at $\tau = 1$, in the experiment where the smoothing begins at t_i, of the alarm function with smoothing lengthscale σ_k relative to a reference model with smoothing lengthscale σ_j. For the experiment where the smoothing begins at t_i, we define the optimal smoothing lengthscale as the value of σ_j that minimizes Equation 5.7. In Figure 5.5, we report the optimal smoothing lengthscale and corresponding minimum average misfit for each of the experiments. We note from Figure 5.5.a a systematic increase in the optimal smoothing lengthscale with increasing start time of smoothing. Viewed in reverse time, the systematic decrease indicates that, as we include more and older earthquakes, it is best to smooth less. This can be understood intuitively: if we had an infinitely long catalog, we would know the exact spatial distribution of seismicity (assuming stationarity), and we could the optimal reference model would be obtained by smoothing the observed epicenters with an infinitesimally small lengthscale. With few exceptions, we observe a systematic increase of misfit in Figure 5.5.b with starting smoothing date. Again, we can consider this in reverse time, in which case we find a systematic decrease in misfit as we include older

earthquakes, which means that these alarm functions benefit from including as many data as possible. This principle guides the experiments described in the following section.

5.5 Moving test period, growing learning period

Previous studies that optimized smoothed seismicity model parameters considered a single retrospective experiment in which epicenters from learning period T_1 were smoothed to forecast epicenters occurring in a subsequent testing period T_2. The results of such an experiment are likely to be unstable to changes in the temporal division; in the following set of experiments, which we denote "moving test period, growing learning period," we iterate over several such retrospective experiments. In addition to addressing instability issues, this also permits an analysis of the temporal evolution of the optimal lengthscale and corresponding misfit.

In the first experiment, we smooth the epicenters of all earthquakes with magnitude at least $m_{\text{smooth}} = 4.0$ occurring in learning period $T_1 = t_0$ to t_1, iterating over the elements in the set of smoothing lengthscales listed in Equation 5.6. Each of these yields a seismic density map, which we take as an alarm function for this experiment; we use each seismic density alarm function to forecast earthquakes with magnitude at least $m_{\text{target}} = 5.0$ occurring in the testing period $T_1^* = t_1^+$ to t_2^*, where t_1^+ indicates the moment immediately following t_1. For each kernel and each lengthscale value, we compute a Molchan trajectory and an area skill score. In the subsequent experiment, we repeat this process for learning period $T_2 = t_0$ to t_2 and testing period $T_2^* = t_2^+$ to t_3^*; that is, we allow the learning period to grow, rather than simply discarding the older events. We repeat this procedure as many times as the earthquake catalog and our temporal discretization

parameters allow, yielding n learning periods denoted $T_i = t_0$ to t_i and n corresponding testing periods denoted $T_i^* = t_i^+$ to t_{i+1}^*. The temporal discretization parameters are δt = t_{i+1} - t_i (for $i > 0$) = 1 year, the amount by which the learning period grows from one experiment to the next, and $\Delta t^* = t_{i+1}^* - t_i^+$ = 5 years, the length of each testing period. This experimental procedure is illustrated in Figure 5.6. These experiments differ from those in the previous section because the test period changes for each experiment and the learning period always captures the entire catalog preceding the test period.

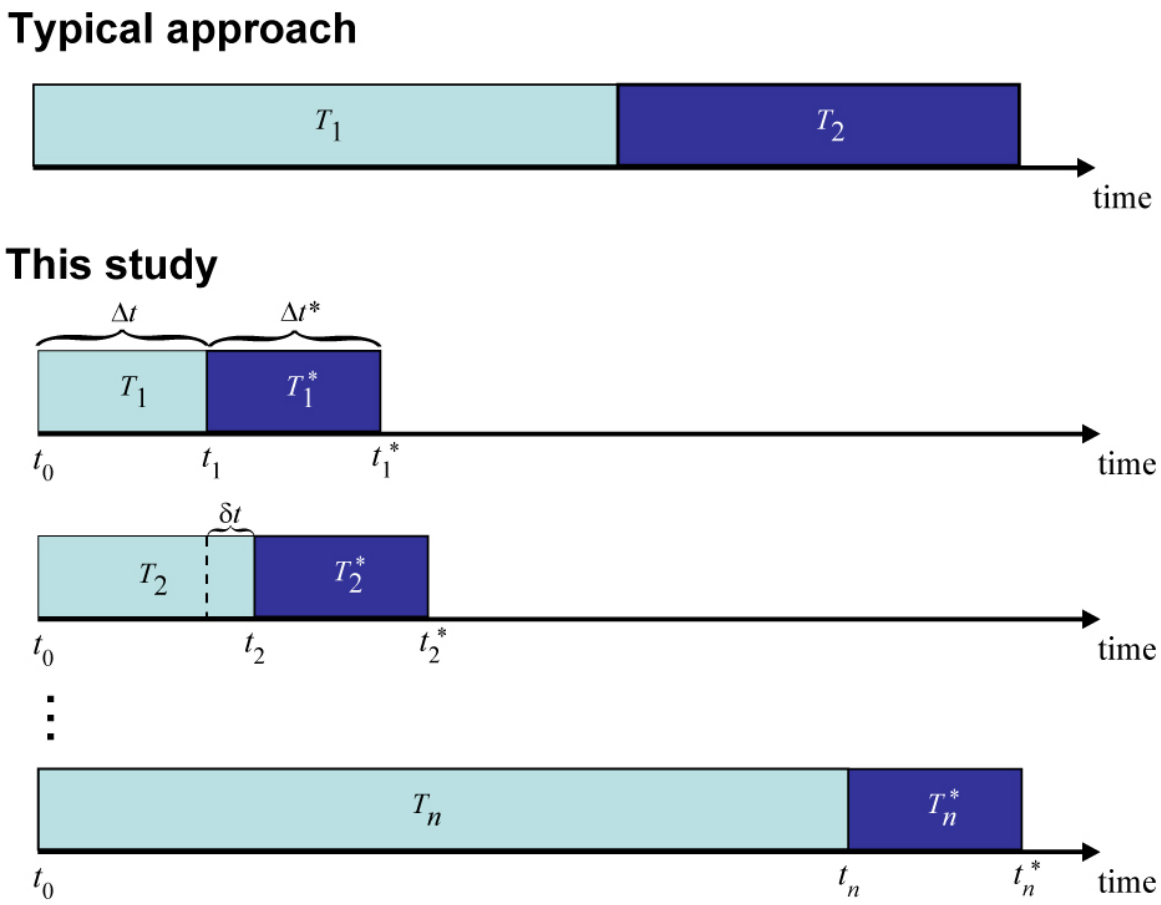

Figure 5.6 Schematic comparison of past seismicity smoothing optimization studies and the present study. Colors are as in Figure 5.1. The typical approach is to optimize smoothing based on a single experiment, whereas this study includes several experiment iterations, where the learning period grows in time while the length of the testing period remains fixed.

Figure 5.7 shows the results of the experiments relative to a uniform reference

model, with t_0 = 1 Jan 1932 and t_1 = 31 Dec 1936 (Δt = 5 years). We note that, as in Figure 5.3, it is difficult to distinguish between the performances of different lengthscales on this plot. It is clear, however, that nearly all of the alarm functions are significantly better than the uniform reference model in every experiment. This is expected due to spatial clustering of seismicity. An exception occurs at 1973, indicating that target earthquakes occurring between 1 Jan 1973 and 31 Dec 1977 were particularly difficult to forecast. We can gain intuition for these experiments by considering this case in detail. In Figure 5.8, we show the alarm function and corresponding trajectories for smoothing lengthscale σ = 1000 km. From the map of the alarm function and target earthquakes, it is clear why the area skill score is relatively low: the cluster of earthquakes west of 124W and between 40N and 41N are located in a region where the alarm function values are relatively low, and few earthquakes occur contiguous to the alarm function peak near 119W, 35.5N. As shown by this example, plots of the area skill score evolution for simple smoothed seismicity alarm functions characterize regional seismicity; a sudden dip in these curves indicates that target earthquakes during the corresponding test period occurred where very few earthquakes occurred in the learning period. On the other hand, a sudden spike in such curves indicates that target earthquakes occurred where a great many earthquakes occurred in the learning period: for example, a high degree of spatial clustering would yield a spike.

Following the same procedure as for the fixed test period experiments, we have computed the optimal smoothing lengthscale and corresponding minimum average misfit for each of the moving test period, growing learning period experiments; these are shown in Figure 5.9. We do not notice any strong systematic patterns in either of the plots in

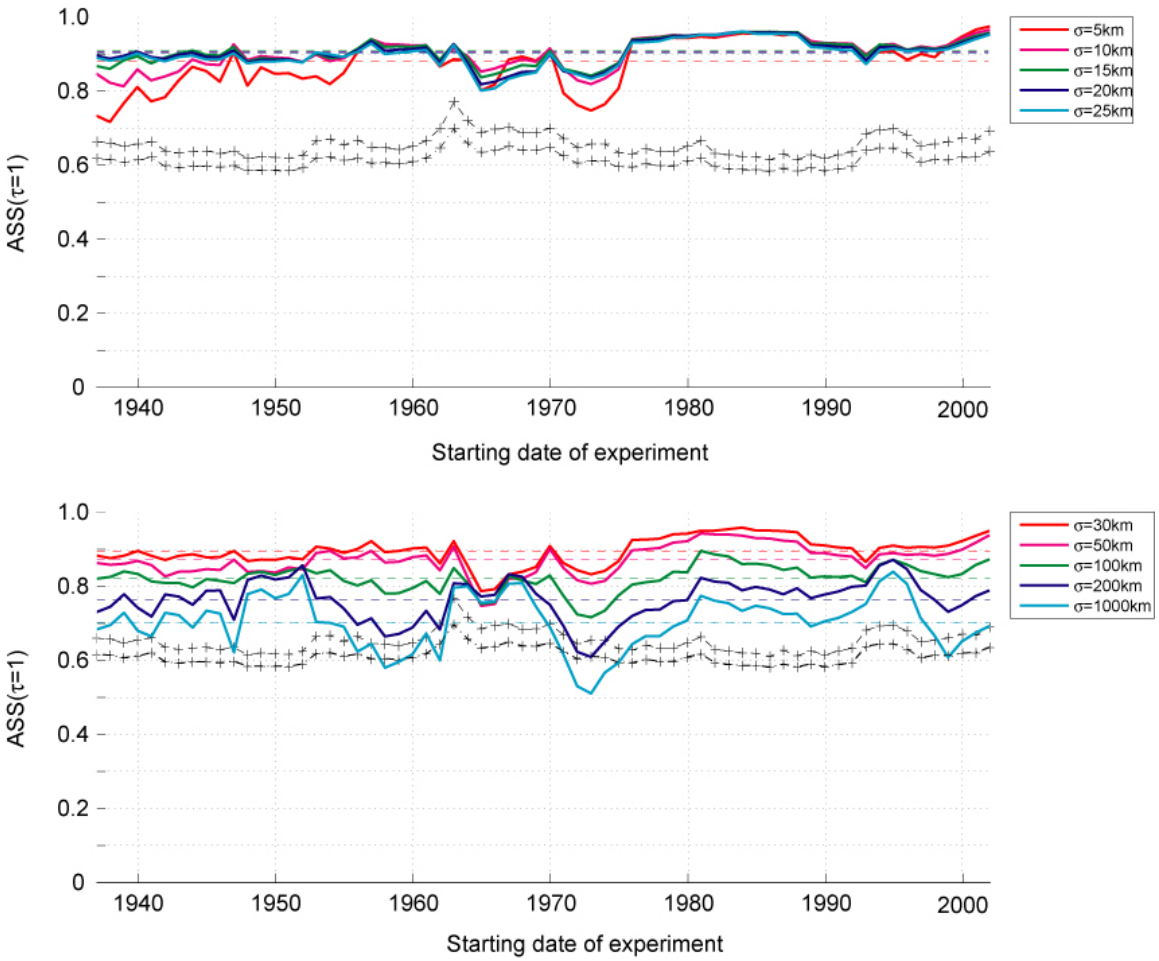

Figure 5.7 Temporal evolution of area skill scores for each lengthscale for each experiment illustrated in Figure 5.6, relative to a uniform reference model. The result of the first experiment in Figure 5.6 is reported at the far left of this plot and subsequent experiments proceed to the right. The two '+' lines demarcate the area skill score confidence bounds corresponding to α=0.05 (bottom) and α=0.01 (top). The dashed lines are the average area skill score at τ=1 over the 66 experiments.

this figure. There is a weak signal indicating that the optimal smoothing lengthscale is lower in the later experiments, but this does not correspond to a systematically decreasing minimum average misfit. In terms of the misfits, however, we do note that the peak value—indicating the lowest predictability—occurs at 1974, immediately following the example experiment analyzed above and corresponding to a relatively low point for all curves in Figure 5.7.

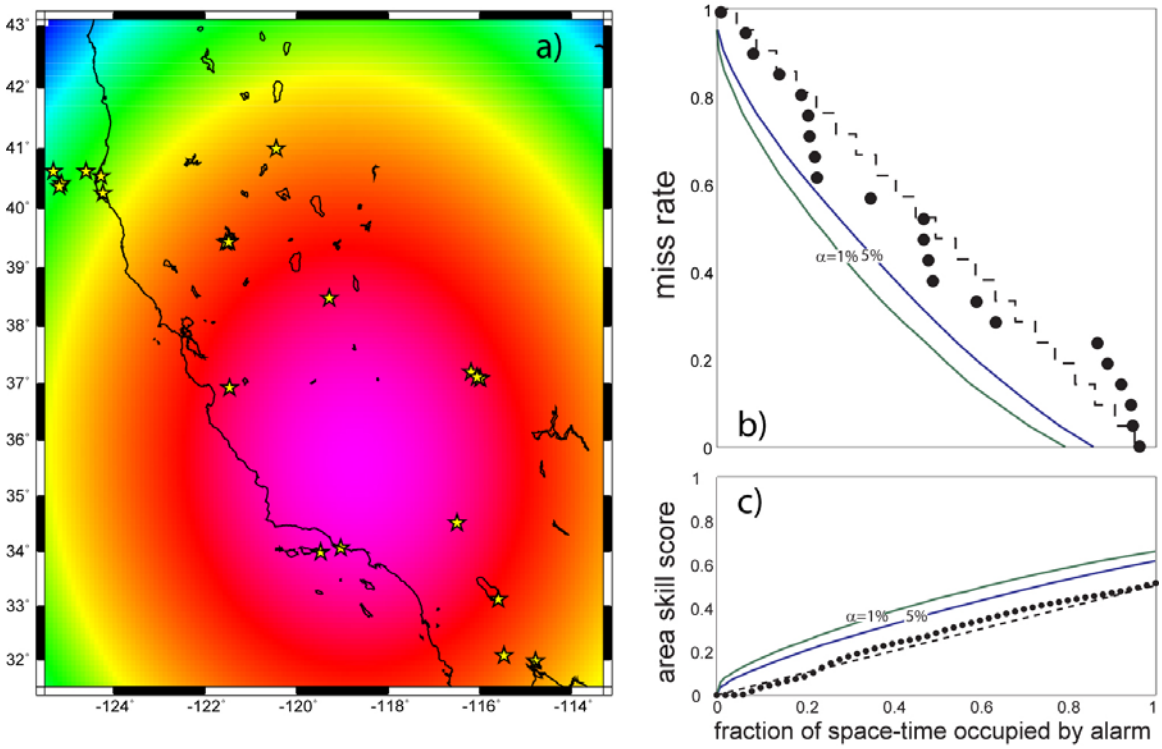

Figure 5.8 Smoothed seismicity predictability experiment for the learning period 1 Jan 1932 to 31 Dec 1972, test period 1 Jan 1973 to 31 Dec 1977. **a)** Seismic density alarm function and the target earthquakes. In this case, earthquakes during learning period were smoothed with the Gaussian kernel with $\sigma = 1000$ km; warm colors indicate high density, white regions indicate zero density. Stars denote target earthquakes that occurred during the test period. **b)** Molchan trajectory corresponding to the alarm function and target earthquake distribution shown in a), using a uniform reference mode. Also shown are the 95% and 99% confidence bounds on the random diagonal. **c)** Area skill score trajectory corresponding to the example forecast experiment, with 95% and 99% confidence bounds.

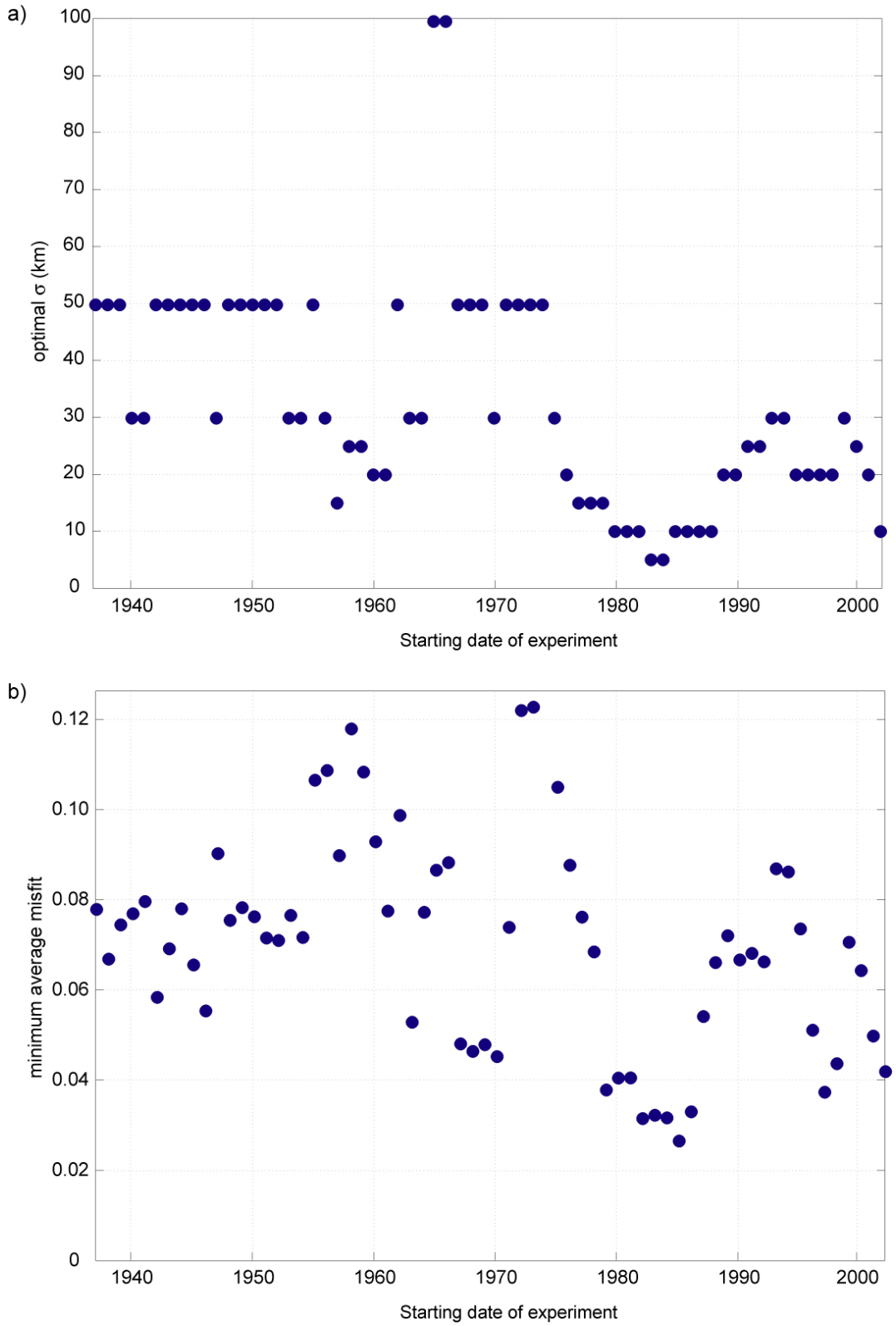

Figure 5.9 Temporal evolution of **a)** the optimal reference model smoothing lengthscale for the experiments illustrated in Figure 5.6 and **b)** the corresponding misfit.

5.5 Moving test period, moving learning period

Re-examining Figure 5.5b, we observe that the most notable exceptions to the trend of decreasing optimal lengthscale with decreasing smoothing start date fall at the right of the plot; we interpret these data points as likely caused by short-term clustering of earthquakes. That is, it seems that several of the earthquakes in the test period occur close to earthquakes from the recent past (1995-1997) and, when the learning period begins somewhat before these events (i.e., between 1992 and 1994), the optimal performance decreases because many events far from the target earthquakes make smoothed contributions. Based on this observation, and the general observation of short-term clustering of earthquakes, we consider a third class of experiments that we call "moving test period, moving learning period." These experiments are very similar to those in the previous section, with the only difference being that the learning period, rather than growing in time, is of a fixed duration and advances with the test period. This procedure is depicted in Figure 5.10. All experiment parameter values are the same as in the previous section. Likewise, the test periods in these experiments coincide with those of the previous section.

In Figure 5.11, we have plotted the experiment results relative to a uniform reference model. As in the other experiments, most smoothed alarm functions are consistently significantly better than the uniform model. As before, however, there are a few sharp dips in the curves. In particular, we consider the absolute minimum in the $\sigma =$ 5 km curve at 1965. In Figure 5.12, we provide the corresponding alarm function, target earthquakes, and trajectories. This is a clear case of undersmoothing; seven out of the

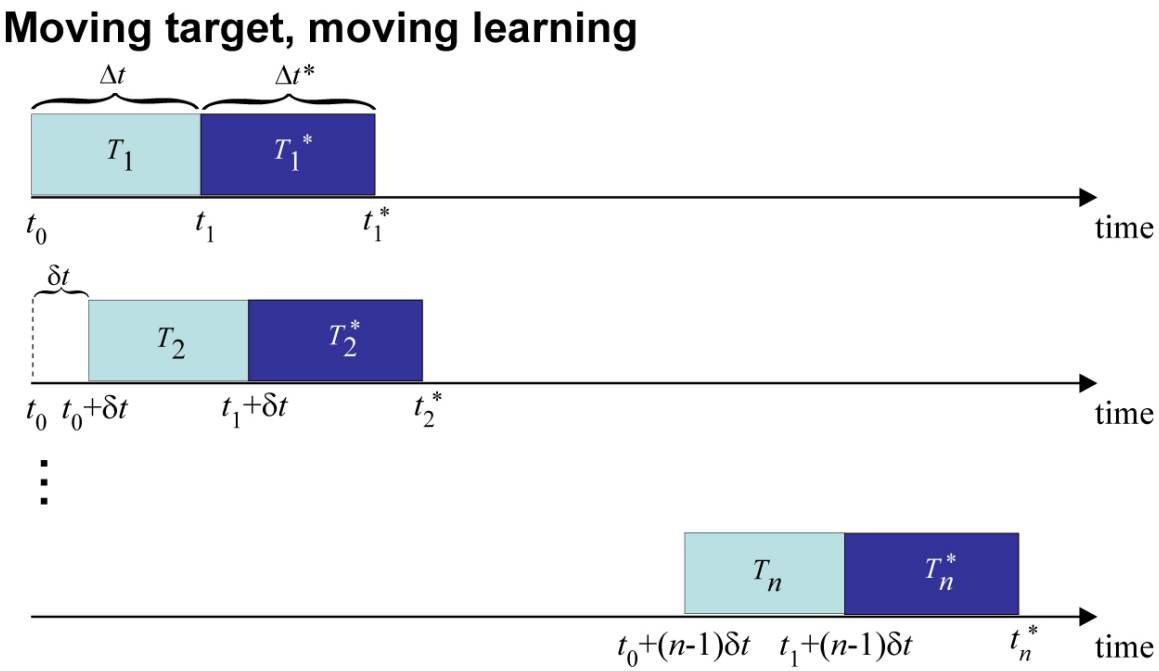

Figure 5.10 Schematic of smoothing optimization experiments with moving learning periods and moving test periods. Colors are as in Figure 5.1. In these experiments, the learning period advances with the test period.

twelve target earthquakes fall in regions obtaining no contribution from smoothed seismicity. Indeed, we see in Figure 5.11 that the $\sigma = 5$ km curve often obtains the minimum area skill score of all alarm functions considered. We attribute this to consistent undersmoothing; it is more pronounced in these experiments than those described in the previous section because the number of events being smoothed is smaller.

In Figure 5.13, we have plotted the optimal smoothing lengthscale and corresponding minimum average misfit for each of the moving test period, moving learning period experiments. As in the previous section, we do not note any strong systematic patterns in either of these plots. The optimal lengthscale plot indicates that small to intermediate lengthscales are preferred, with these optimal values being, on the whole, slightly higher than in the previous section's experiment. Again, this is consistent

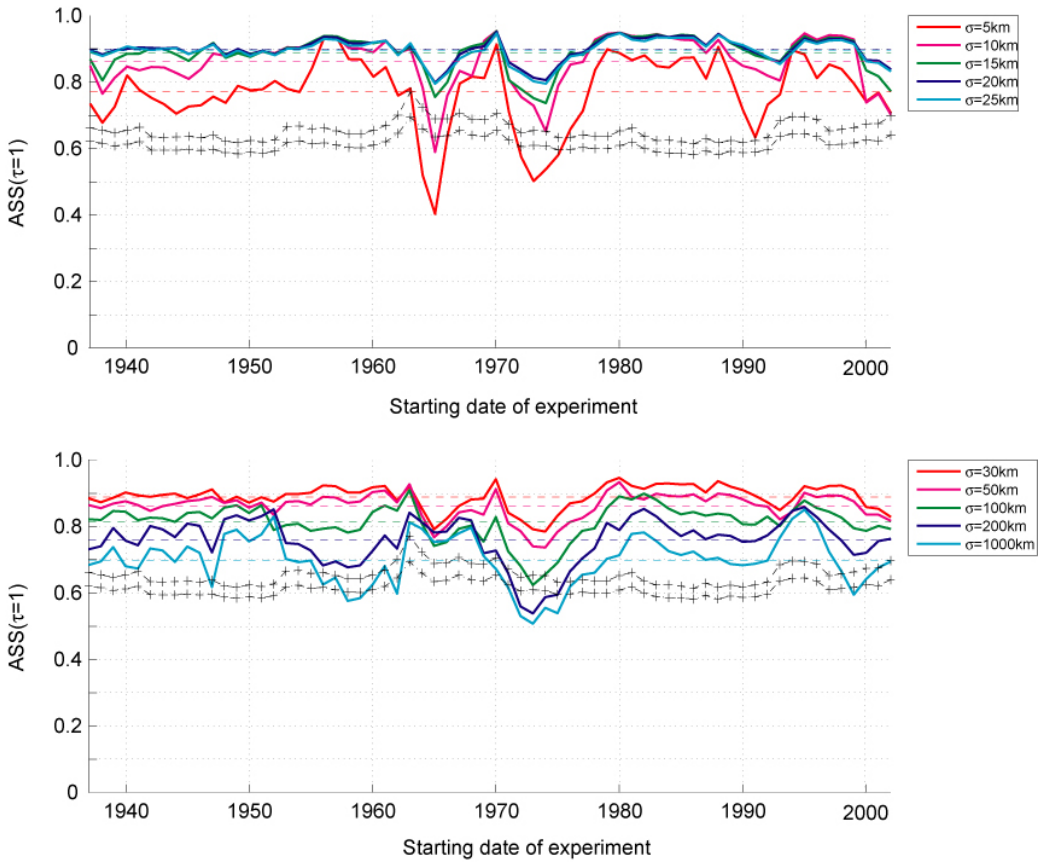

Figure 5.11 Moving learning results relative to uniform reference model

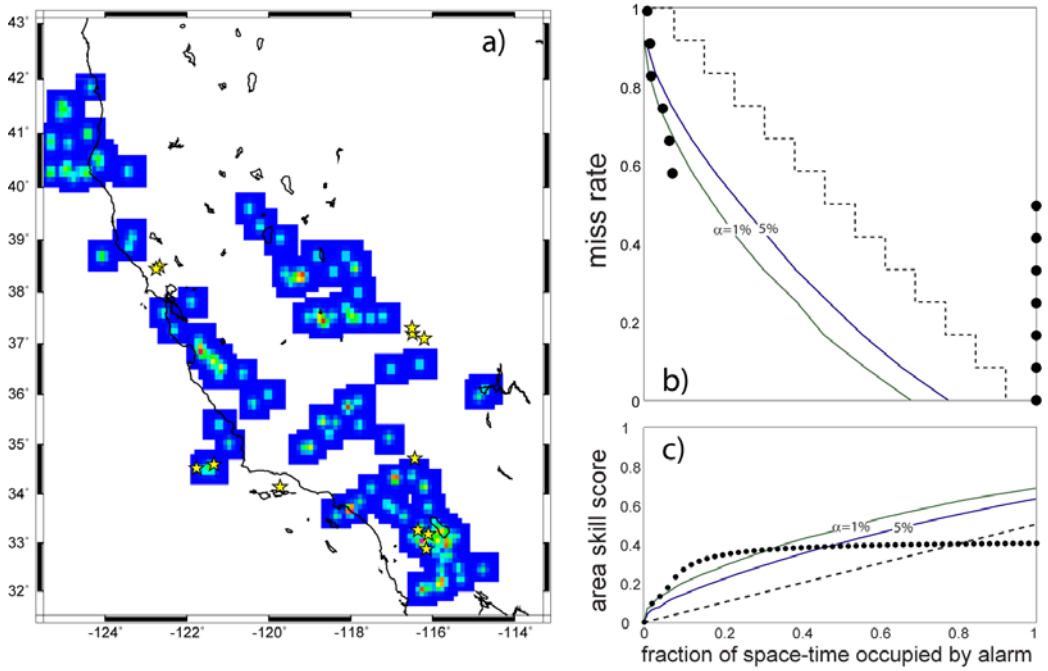

Figure 5.12 As in Figure 5.8, low predictability in 1965 moving test period, moving learning period experiment with σ = 5km.

with the fact that the learning periods in this experiment are smaller and therefore contain fewer earthquakes. To further compare these two different approaches, we have superposed the minimum average misfit data for the moving test period, growing learning period experiments. Considered as a whole, this plot indicates some small advantage to smoothing the entire catalog up to the test date, as opposed to smoothing only the previous five years of seismicity, although this advantage may not be large. There also seems to be some time-varying behavior indicated by this plot: it seems that during the first half of the experiments, the short-term smoothing is superior to the long-term smoothing, whereas the pattern is reversed in the second half of the experiments.

5.6 Discussion and conclusion

In this chapter, we presented exact analytic expressions for three types of spatial smoothing kernels in a gridded natural laboratory, where each kernel is controlled by a single lengthscale parameter. We developed and exercised an empirical framework for exploring optimality within a class of smoothed seismicity forecast models. With the goal of developing a simple, optimized reference forecast, we conducted three series of experiments and analyzed the results: in the first, we fixed the test period and allowed the learning period to grow backward in time; in the second, we moved the test period in a sliding time window and allowed the learning period to cover all previous seismicity; in the third, we moved the test period in a sliding time window and moved the learning period simultaneously.

Based on the results of the experiments described in this chapter, we have computed the prospective smoothed seismicity forecast shown in Figure 5.14, and we

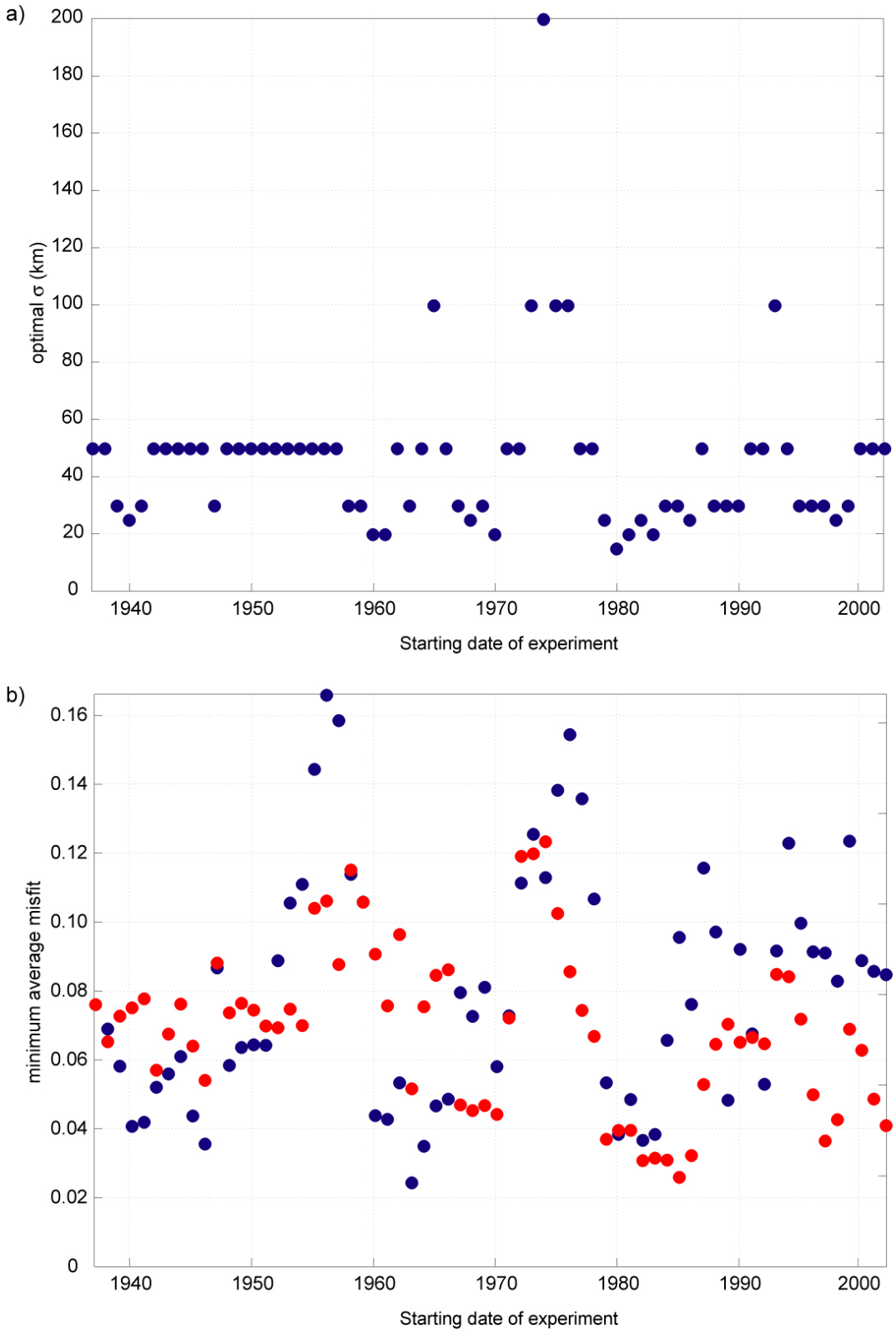

Figure 5.13 Temporal evolution of **a)** the optimal reference model smoothing lengthscale for the experiments illustrated in Figure 5.10 and **b)** the corresponding misfit. Red points are results from the moving test period, growing learning period experiments.

plan to submit such a forecast for evaluation within a CSEP testing center. Here, we smoothed all earthquakes in the California testing region with magnitude greater than 4.0 occurring from 1 Jan 1932 to 31 Mar 2008. We chose to smooth over the entire catalog rather than only the previous five years based on the results of the fixed test period experiments, which indicate that predictability increased with increasing learning period duration (e.g., Figure 5.5.b). We used the Gaussian smoothing kernel described by Equation 5.3 with σ = 10 km. As justification for this lengthscale value, we consider the results shown in Figure 5.9.a. Ideally, we would like to see a stronger signal in this plot, but as it stands, we can most likely rule out σ values larger than 100 km. We note that the optimal smoothing lengthscale, with only two exceptions, does not change by more than 20 km from one experiment to the next and, over the final 25 experiments, it hasn't changed by more than 10 km from one experiment to the next. Therefore, we chose the optimal lengthscale from the ultimate experiment in this series.

In this chapter, we have also quantified the limits of predictability for a given target earthquake distribution and class of smoothed seismicity models. For example, Figure 5.5.b, the plot of the minimum average misfits corresponding to the optimal reference models of Figure 5.5.a, indicate the limits of predictability. That is, these misfits quantify how close we can get to the "true" reference model for the chosen smoothing kernel and set of possible lengthscales. The experimental procedures described in this chapter can be applied to other types of forecast models and indeed, could be used to further explore parameter space (e.g., at a finer discretization) for the very simple model we have considered here. Along these lines, it would be interesting to use these methods to produce simple short-term forecasts to compare with those from

point process models (e.g., Ogata, 1988, Kagan & Jackson 2000, Gerstenberger *et al*. 2005). The pursuit of anisotropic, fault-based kernels should also be considered.

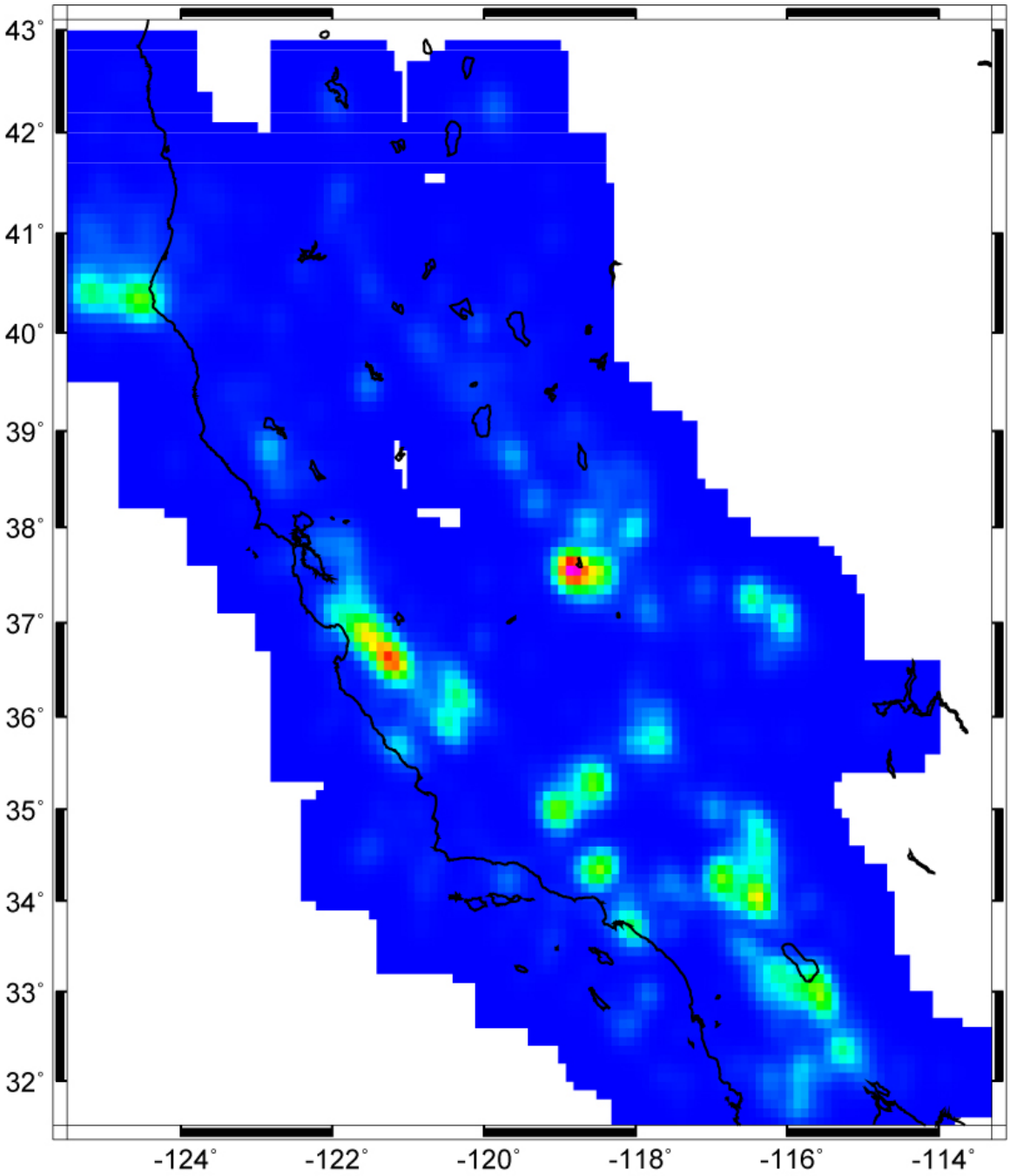

Figure 5.14 Prospective forecast to be submitted to CSEP as reference model. This forecast is the product of smoothing all events in the ANSS catalog with magnitude greater than or equal to 4.0, using the Gaussian smoothing kernel described by Equation 5.3 with $\sigma = 10$ km.

We mentioned briefly that the temporal evolution of the area skill score

characterizes the predictability of regional seismicity. Further study of this evolution could drive new approaches to modeling seismicity dynamics. For example, if the periods wherein simple smoothing techniques perform poorly (i.e., when we observe a sudden spike in minimum average misfit) could be forecast, hybrid models that switch modes to incorporate some other form of forecasting might be developed. It is also likely that the evolution of the area skill score is strongly affected by the occurrence of very large events; in this case, we would expect the number of target earthquakes near the epicenter/fault rupture of large earthquakes to increase. Time-varying, adaptive smoothing that responded to large earthquakes would be beneficial.

CHAPTER SIX:
Conclusions, ongoing work, and potential extensions

6.1 Introduction

The work described in this dissertation has addressed several questions related to earthquake predictability. In this space, rather than making broad generalizations or reiterating the conclusions of each chapter, I will describe ongoing related work and propose some potential extensions of this research.

In the near-term, many of the ideas presented here will be implemented for wide use within the Southern California Earthquake Center (SCEC) Collaboratory for the Study of Earthquake Predictability (CSEP) testing center. In particular, there are plans to integrate the Reverse Tracing of Precursors (RTP) codes and the accompanying testing procedures described in Chapter 2. The testing method outlined in Chapter 3 and illustrated in Chapter 4 will also be incorporated in CSEP testing centers, allowing for comparative testing of a wide class of forecasts. The prospective forecast presented in Chapter 5 will be submitted as a potential reference model for testing in the CSEP California natural laboratory, and an analogous global forecast is under development.

Concurrent with these technical implementations, we plan to compare alarm-based testing methods, such as the area skill score, with probabilistic methods, in particular likelihood-based tests. A comprehensive comparison of these two approaches should include analytic work, numerical experimentation, and exploration with actual forecasts and observations. An analytic comparison can be based on forecasts and observations with exact, known distributions and preliminary analysis along these lines

has been conducted by Kagan (2007). Numerical experiments can be performed using simple synthetic models that include a stochastic component. The final step in this comparison can leverage existing RELM forecasts in both prospective and retrospective observations; this will be an extension of the work begun by Zechar *et al.* (2007).

The smoothed seismicity experiments described in Chapter 5 were kept simple intentionally, but the framework we have established is very flexible. Along these lines, there are a number of potentially productive experiments one could pursue. For example, one may explore a class of hybrid models (e.g., Gerstenberger & Rhoades 2007) in which recent earthquakes are smoothed using a kernel distinct from that used for older earthquakes, thereby explicitly incorporating temporal clustering. To this end, the Omori-Utsu relation could be incorporated directly via a weighting scheme. Alternatively, one may explore the use of a single spatial kernel and supplement it with weights that are time-dependent or magnitude-dependent. One might also consider spatial smoothing kernels with a variable lengthscale; again, this may depend on the time elapsed since the earthquake of interest, local earthquake density, or the magnitude of the earthquake being smoothed. Certainly the parameter space of smoothed seismicity modeling is something that begs for further exploration. Even beyond this class of models, however, the area skill score minimum average misfit statistic can be useful; for example, it may be applied to the 5 year RELM experiments to determine which forecast most closely approximates the observed distribution of seismicity.

The issue of reference models is key to advancing earthquake predictability research. Many assumptions implicit in smoothing seismicity remain as of yet untested. For example, in the case where earthquake data are not numerous, researchers often

estimate the spatial distribution of large earthquakes based on the observed distribution of smaller earthquakes. There are regions for which this procedure will clearly yield incorrect results—for example, the Geysers geothermal region in California—but a quantitative analysis remains to be done. Smoothed seismicity forecasts also operate under a weak assumption of stationarity, which is a hypothesis that should be tested independent of the forecasting problem.

The earthquake predictability experiments discussed here are conceptually very similar to efforts within the earthquake early warning (EEW) community. While some EEW algorithms are presently in operational status, many underlying hypotheses have not been tested rigorously, particularly in a prospective sense. The work described in this dissertation seems ideally suited for EEW evaluation as the relevant algorithms are inherently alarm-based.

6.2 Proposed experiments in southern California

Many current forecast models are based on seismicity patterns and use a regional earthquake catalog as the only input data source. High-precision relocated catalogs are now readily available to describe tectonic activity in southern California (Lin *et al.* 2007), and these ought to be integrated into prediction experiments. By using such catalogs, we can gain spatial resolution and may construct hypocentral—rather than epicentral—forecasts. Moreover, by leveraging the reduced hypocenter uncertainties, we may design experiments with a much finer spatial discretization than those currently under consideration.

Fault networks are critical to understanding the accommodation of plate boundary

motion and, in particular, the spatial distribution of seismicity. By conducting prediction experiments that explicitly integrate the SCEC Community Fault Model (Plesch *et al*. 2007), we may test the hypothesis that seismicity rates correlate with fault density, and also quantify the amount of seismicity that is not readily associated with a known, mapped fault. Moreover, smoothed seismicity forecasts and stochastic triggering methods such as the Epidemic Type Aftershock Sequence model (Ogata 1988) can be extended to use anisotropic kernels that depend on local fault orientations.

Through the Plate Boundary Observatory component of EarthScope, a great number of GPS data covering southern California are now freely available. These data can be used for predictability experiments that systematically test the hypothesis that intermediate-term strain transients, observable over a period of several years, precede large earthquakes (e.g., Ogata 2007). By combining GPS data with seismicity catalogs, there is some hope that forecasts can be developed based on the idea that catalogs tell us where future earthquakes will occur and GPS data tell us when.

6.3 Proposed experiments in Japan and Taiwan

While southern California is well-monitored and studied by many earthquake scientists, prospective prediction experiments of large earthquakes in this region will require years to decades to amass a statistically significant sample size. To accelerate experiments of potentially damaging earthquakes, and to capitalize on a wealth of existing data, Japan and Taiwan should be developed as natural laboratories in which to conduct prediction experiments.

Taiwan experiences a plate deformation rate that is nearly double that of the

boundary between the North America and Pacific plates and a seismicity rate that is estimated to be five to ten times that of California. Likewise, Japanese seismicity presents the opportunity to study predictability of deep earthquakes in a subduction zone setting. One may take advantage of the much higher seismicity rates to test models for intermediate to large earthquakes on a one-year timescale, as compared with the five-year time scale for California. Additionally, this development will allow for testing the rate models and assumptions used in the Japanese national seismic hazard maps.

6.4 Proposed repeating micro-earthquake experiments

The Parkfield Earthquake Prediction Experiment (Bakun & Lindh 1985) has yielded a unique set of repeating micro-earthquake (M_w<2) data (e.g., Nadeau *et al*. 1994, Nadeau & Johnson 1998, Nadeau & McEvilly 1997) and these provide an excellent opportunity for new predictability experiments. Clearly, microrepeaters are in some sense predictable: not only do they occur in roughly the same hypocentral region, they appear to have highly similar source mechanisms and repeatedly rupture the same source material, yielding highly similar seismograms. To date, however, there have been no formal predictions of these events. Parkfield microrepeaters can be used to further push predictability experiment resolution and to develop forecasts that include estimates of focal mechanism. Moreover, they can be used to explore the possibility of deterministic prediction, including occurrence-time forecasts.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. A. (eds.), 1972. Handbook of mathematical functions with formulas, graphs, and mathematical tables. Dover, New York, 1046 pp.

Aki, K., 1981. A probabilistic synthesis of precursory phenomena, in Earthquake Prediction: An International Review, pp 566—574, eds. Simpson, D. & Richards, P., Am. Geophys. Union, Washington, D.C.

ANSS Earthquake Catalog. Produced by Advanced National Seismic System (ANSS) and hosted by the Northern California Data Center (NCEDC), 1932-2007. http://quake.geo.berkeley.edu/anss.

Bakun, W.H., & Lindh, A.G., 1985. The Parkfield, California, earthquake prediction experiment, Science **229**(4714), 619—624.

Blanter, E.M., & Shnirman, M.G., 1997. Simple hierarchical systems: stability, self-organized criticality and catastrophic behavior. Phys. Rev. E **55** (6), 6397–6403.

Bowman, D.D., Ouilon, G., Sammis, C.G., Sornette, A. and Sornette, D., 1998. An observational test of the critical earthquake concept. J. Geophys. Res. **103**, 24359–24372.

Cao, T., Petersen, M.D., & Reichle, M.S., 1996. Seismic hazard estimate from background seismicity in southern California. Bull. Seismol. Soc. Am., **86**(5), 1372-1381.

CMT Earthquake Catalog. Produced by the Global Centroid Moment Tensor (GCMT) group, 1976-2007. http://www.globalcmt.org.

Ebel, J.E., Chambers, D.W., Kafka, A.L., & Baglivo, J.A., 2007. Non-Poissonian earthquake clustering and the hidden markov model as bases for earthquake forecasting in California. Seismol. Res. Lett., **78**(1), 57—65.

Enescu, B., & Ito, K. , 2001. Some premonitory phenomena of the 1995 Hyogo-Ken Nanbu (Kobe) earthquake: seismicity, b-value and fractal dimension, Tectonophys., **338** (3-4), 297-314.

Eneva, M., & Ben-Zion, Y., 1997. Techniques and parameters to analyze seismicity patterns associated with large earthquakes. J. Geophys. Res., **102**, 17785-17795.

Evison, F.F., & Rhoades, D.A., 1993. The precursory earthquake swarm in New Zealand: hypothesis tests. New Zealand J. Geol. and Geophys., **36**, 51-60.

Evison, F.F., & Rhoades, D.A., 1997. The precursory earthquake swarm in New Zealand: hypothesis tests II. New Zealand J. Geol. and Geophys., **40**, 537-547.

Feng, D. Y., 1975. Anomalies of seismic velocity ratio before the Tangshan-Daguan earthquake (M = 7.1) on May 11, 1974. Chinese Geophys., **1**, 47-53.

Field, E.H., 2007. Overview of the working group for the development of regional earthquake likelihood models (RELM), Seismol. Res. Lett., **78**(1), 7—16.

Field, E.H., Gupta, N., Gupta, V., Blanpied, M.L., Maechling, P.J. & Jordan, T.H., 2005. Hazard calculations for the WGCEP-2002 forecast using OpenSHA and distributed object technologies, Seismol. Res. Lett., **76**, 161—167.

Frankel, A., 1995. Mapping seismic hazard in the central and eastern United States. Seismol. Res. Lett., **66**(4), 8-21.

Frankel, A., Mueller, C., Barnhard, T., Perkins, D., Leyendecker, E., Dickman, N., Hanson, S. & Hopper, M., 1996. National seismic-hazard maps: Documentation June 1996, U.S. Geol. Surv. Open-file report 96-532, 41 pp.

Frankel, A., Petersen, M., Mueller, C., Haller, K., Wheeler, R., Leyendecker, E., Wesson, R., Harmsen, S., Cramer, C., Perkins, D. & K. Rukstales, 2002. Documentation for the 2002 update of the national seismic hazard maps, U.S. Geol. Surv. Open-file report 02-420, 33 pp.

Gabrielov, A., Keilis-Borok, V., Zaliapin, I., & Newman, W.I., 2000. Critical transitions in colliding cascades. Phys. Rev. E **62**, 237–249.

Gabrielov, A., Newman, W.I., Turcotte, D.L., 1999. An exactly soluble hierarchical clustering model: inverse cascades, self-similarity and scaling. Phys. Rev. E **60**, 5293–5300.

Gerstenberger, M.C., & Rhoades, D.A., 2007. Natural laboratories: the New Zealand earthquake forecast testing center, Seism. Res. Lett., **78**(2), 237.

Hauksson, E., 1981. Radon content of groundwater as an earthquake precursor: Evaluation of worldwide data and physical basis. J. Geophys. Res., **86**(B10), 9397—9410.

Helmstetter, A., Kagan, Y.Y. & Jackson, D.D., 2007. High-resolution time-independent forecast for M>5 earthquakes in California. Seismol. Res. Lett., **78**(1), 78—86.

Holliday, J.R., Chen, C., Tiampo, K., Rundle, J.B., Turcotte, D.L., & Donnellan, A., 2007. A RELM earthquake forecast based on pattern informatics. Seismol. Res. Lett., **78**(1), 87—93.

Hough, S., & Olsen, K.B. (eds.), 2007. Special issue on: Regional earthquake likelihood models. Seismol. Res. Lett., **78**(1).

Jackson, D.D., 1996. Hypothesis testing and earthquake prediction. Proc. Natl. Aca. Sci. USA, **93**(9), 3772-3775.

JMA earthquake catalog. Produced by the Japanese Meteorological Agency (JMA), 1921-2007.

Jolliffe, I.T. & Stephenson, D.B. (eds.), 2003. Forecast Verification. Wiley, Hoboken, 254 pp.

Jordan, T.H., 2006. Earthquake predictability, brick by brick, Seismol. Res. Lett., **77**(1), 3—6.

Jordan, T.H., Schorlemmer, D., Wiemer, S., Gerstenberger, M.C., Jackson, D.D., Maechling, P.J., Liukis, M., Marzocchi, W., & Zechar, J.D., in preparation. Collaboratory for the Study of Earthquake Predictability: progress and plans.

Kafka, A.L., 2002. Statistical analysis of the hypothesis that seismicity delineates areas where future large earthquakes are likely to occur in the Central and Eastern United States, Seismol. Res. Lett., **73**, 990—1001.

Kagan, Y.Y., 1996. VAN earthquake predictions—an attempt at statistical evaluation, Geophys. Res. Lett. **23**(11), 1315—1318.

Kagan, Y. Y., 2007. On earthquake predictability measurement: information score and error diagram, Pure Appl. Geoph., **164**(10), 1947-1962.

Kagan, Y.Y, & Jackson, D.D., 1994. Long-term probabilistic forecasting of earthquakes. J. Geophys. Res., **99**(B7), 13685-13700.

Kagan, Y.Y, & Jackson, D.D., 1999. Testable earthquake forecasts for 1999. Seismol. Res. Lett., **70**(4), 393-403.

Kagan, Y.Y, & Jackson, D.D., 2000. Probabilistic forecasting of earthquakes. Geophys. J. Int., **143**, 438-453.

Kagan, Y.Y., Jackson, D.D., & Rong, Y., 2007. A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity, Seismol. Res. Lett., **78**(1), 94-98.

Keilis-Borok, V.I., & Kossobokov, V., 1990. Premonitory activation of earthquake flow: algorithm M8, Phys. Earth Planet. Inter., **61**(1/2), 73—83.

Keilis-Borok, V.I., & Soloviev (eds.), 2003. Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer-Verlag, Berlin, 337 pp.

Keilis-Borok, V.I., 2002. Earthquake prediction, state-of-the-art and emerging possibilities, Annu. Rev. Earth Planet. Sci., **30**, 1—33.

Keilis-Borok, V.I., 2003. Fundamentals of earthquake prediction: four paradigms, in Nonlinear Dynamics of the Lithosphere and Earthquake Prediction, pp 1—36, eds. Keilis-Borok, V.I. & Soloviev, A., Springer-Verlag, Berlin.

Keilis-Borok, V.I., Shebalin, P.N., Gabrielov, A. & Turcotte, D.L., 2004. Reverse tracing of short-term earthquake precursors, Phys. Earth Planet. Inter., **145**, 75—85.

Knopoff, L., Aki, K., Allen, C., Rice, J., & Sykes, L. (eds.), 1996. NAS Colloquium on Earthquake Prediction: The Scientific Challenge. Proc. Natl. Acad. Sci. USA, **93**(9).

Kossobokov, V. and Shebalin, P.N., 2003. Earthquake prediction, in Nonlinear Dynamics of the Lithosphere and Earthquake Prediction, pp 141—205, eds. Keilis-Borok, V.I. & Soloviev, A., Springer-Verlag, Berlin.

Kossobokov, V., 2004. Earthquake prediction: basics, achievements, perspectives. Acta Geod. Geoph. Hung., **39**(2/3), 205—221.

Lehman, E.L. & Romano, J.P., 2005. Testing Statistical Hypotheses, 3rd edn, pp. 784, Springer, New York.

Lin, G., Shearer, P.M., & Hauksson, E., 2007. Applying a three-dimensional velocity model, waveform cross correlation, and cluster analysis to locate southern California seismicity from 1981 to 2005, J. Geophys. Res. **112**(B12309), doi:10.1029/2007JB004986.

Loughe, A.F., Henderson, J.K., Mahoney, J.L., & Tollerud, E.I., 2001. A verification approach suitable for assessing the quality of model-based precipitation forecasts during extreme precipitation events. Preprints, Symposium on Precipitation Extremes: Prediction, Impacts, and Responses, January 14-19 2001, Albuquerque, NM.

Mahoney, J. L., Henderson, J.K., & Miller, P.A., 1997. A description of the Forecast Systems Laboratory's real-time verification system (RTVS). Preprints, 7th Conference on Aviation, Range, and Aerospace Meteorology, Long Beach, CA, Amer. Meteor. Soc., J26-J31.

Mahoney, J.L., Henderson, J.K., Brown, B.G., Hart, J.E., Loughe, A.F., Fischer, C., & Sigren, B., 2002. Real-Time Verification System (RTVS) and its application to aviation weather forecast. 10th Conference on Aviation, Range, and Aerospace Meteorology, 13-16 May, Portland, OR.

Marzocchi, W., Sandri, L., & Boschi, E., 2003. On the validation of earthquake-forecasting models: the case of pattern recognition algorithms. Bull. Seismol. Soc. Am., **93**(5), 1994-2004.

Mason, I.B., 2003. Binary events, in Forecast Verification, pp. 37—76, eds. Jolliffe, I.T. & Stephenson, D.B., Wiley, Hoboken.

McGuire, J.J., Boettcher M.S. & Jordan, T.H., 2005. Foreshock sequences and short-term earthquake predictability on East Pacific Rise transform faults, Nature, **434**(7032), 457—461.

Michael, A.J., 1997. Testing prediction methods: earthquake clustering versus the Poisson model, Geophys. Res. Lett., **24**(15), 1891—1894.

Molchan, G.M., 1991. Structure of optimal strategies in earthquake prediction, Tectonophysics, **193**, 267—276.

Molchan, G.M. & Kagan,Y.Y., 1992. Earthquake prediction and its optimization, J. Geophys. Res., **97**, 4823—4838.

Molchan, G.M., 1990. Strategies in strong earthquake prediction, Phys. Earth Planet. Inter., **61**, 84—98.

Nadeau, R.M., Antolik, M., Johnson, P.A., Foxall, W., & McEvilly, T.V., 1994. Seismological studies at Parkfield III: Microearthquake clusters in the study of fault-zone dynamics, Bull. Seismol. Soc. Am., **84**(2), 247—263.

Nadeau, R.M., & Johnson, L.R., 1998. Seismological studies at Parkfield VI: Moment release rates and estimates of source parameters for small repeating earthquakes, Bull. Seismol. Soc. Am., **88**(3), 790—814.

Nadeau, R.M., & McEvilly, T.V., 1997. Seismological studies at Parkfield V: Characteristic microearthquake sequences as fault-zone drilling targets, Bull. Seismol. Soc. Am., **87**(6), 1463—1472.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. J. Am. Stat. Assoc., **83**, 9 –27.

Ogata, Y., 2007. Seismicity and geodetic anomalies in a wide area preceding the Niigata-Ken-Chuetsu earthquake of 23 October 2004, central Japan, J. Geophys. Res., **112**, B10301, doi:10.1029/2006JB004697.

Petersen, M., Cao, T., Campbell, K. & Frankel, A., 2007. Time-independent and time-dependent seismic hazard assessment for the state of California: uniform California earthquake rupture forecast model 1.0, Seismol. Res. Lett., **78**(1), 99—109.

Plesch, A., Shaw, J.H., Benson, C., Bryant, W.A., Carena, S., Cooke, M., Dolan, J.F., Fuis, G., Gath, E., Grant, L., Hauksson, E., Jordan, T.H., Kamerling, M., Legg, M., Lindvall, S., Magistrale, H., Nicholson, C., Niemi, N., Oskin, M., Perry, S., Planansky, G., Rockwell, T., Shearer, P.M., Sorlien, C., Suss, P., Suppe, J., Treiman, J., & Yeats, R., 2007. Community fault model (CFM) for southern California, Bull. Seismol. Soc. Am., **97**(6), 1793—1802.

Press, F. (ed.), 1965. Earthquake prediction: A proposal for a ten year program of research. Ad Hoc Panel on Earthquake Prediction, White House Office of Science and Technology, 134 pp.

Rhoades, D.A., & Evison, F.F., 1984. Method assessment in long-range earthquake forecasting. In Earthquake Prediction: Proceedings of the International Symposium on Earthquake Prediction. Terra, Tokyo, pp. 497-504.

Rhoades, D.A., & Evison, F.F., 1989. Time-variable factors in earthquake hazard, Tectonophysics, **167**, 201—210.

Rhoades, D.A., & Evison, F.F., 2004. Long-range earthquake forecasting with every earthquake a precursor according to scale, Pure and App. Geophys., **161**, 47—72.

Rikitake, T. (ed.), 1982. Earthquake prediction research **1**(1).

Roeloffs, E.A., 1988. Hydrologic precursors to earthquakes: A review. Pure & App. Geoph., **126**(2-4), 177-209.

Rundle, J.B., Tiampo, K., Klein, W. & Sa Martins, J., 2002. Self-organization in leaky threshold systems: the influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting, Proc. Natl. Aca. Sci. USA, **99**, 2514—2521.

Rundle, J.B., Turcotte, D.L., Shcherbakov, R., Klein, W., Sammis, C., 2003. Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev. Geophys. **41**, 1019.

Sadooghi-Alvandi, S.M., Nematollahi, A.R., & Habibi, R., 2007. On the distribution of the sum of independent uniform random variables. Stat. Papers, DOI 10.1007/s00362-007-0049-4.

Schorlemmer, D., & Gerstenberger, M.C., 2007. RELM Testing Center. Seismol. Res. Lett., **78**(1), 30—36.

Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D. & Rhoades, D.A., 2007. Earthquake likelihood model testing, Seismol. Res. Lett., **78**(1), 17—29.

Shebalin, P., Zaliapin, I., Keilis-Borok, V.I., 2000. Premonitory rise of the earthquakes' correlation range: lesser Antilles. Phys. Earth Planet. Inter. **122**, 241– 249.

Shebalin, P.N., Keilis-Borok, V.I., Gabrielov, A., Zaliapin, I., & Turcotte, D., 2006. Short-term earthquake prediction by reverse analysis of lithosphere dynamics. Tectonophys., **413**(2006), 63-75.

Shebalin, P.N., Keilis-Borok, V.I., Zaliapin, I., Uyeda, S., Nagao, T., & Tsybin, N., 2003. Short-term premonitory rise of the earthquake correlation range. IUGG Abstracts, Sapporo, Japan.

Sornette D., 2000. Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization and Disorder. Springer Ser. Synerg., Springer–Verlag, Heidelberg, 432 pp.

Stark, P.B., 1996. A few considerations for ascribing statistical significance to earthquake predictions, Geophys. Res. Lett., **23**(11), 1399—1402.

Stark, P.B., 1997. Earthquake prediction: the null hypothesis, Geophys. J. Int., **131**, 495—499.

Stirling, M.W., McVery, G.H., & Berryman, K.R., 2002. A new seismic hazard model for New Zealand. Bull. Seismol. Soc. Am., **92**(5), 1878-1903.

Stock, C., & Smith, E.G.C., 2002a. Adaptive kernel estimation and continuous probability representation of historical earthquake catalogs. Bull. Seismol. Soc. Am., **92**(3), 904-912.

Stock, C., & Smith, E.G.C., 2002b. Comparison of seismicity models generated by different kernel estimations. Bull. Seismol. Soc. Am., **92**(3), 913-922.

Tiampo, K. F., Rundle, J.B., McGinnis, S., Gross, S.J. & Klein, W., 2002. Mean-field threshold systems and phase dynamics: an application to earthquake fault systems, Europhys. Lett., **60**(3), 481—488.

Ward, S.N., 2007. Methods for evaluating earthquake potential and likelihood in and around California, Seismol. Res. Lett., **78**(1), 121—133.

Werner, M.J. & D. Sornette, 2007. Magnitude uncertainties: impact on seismic rate estimates, forecasts and prediction experiments, in preparation.

Wessel, P. & W. Smith, 1998. New, improved version of Generic Mapping Tools released, Eos Trans. AGU, **79**(47), 579.

Wyss, M. (ed.), 1991. Special issue on earthquake prediction. Tectonophys., **193**.

Wyss, M., & Dmowska, R. (eds.), 1997. Earthquake prediction - state of the art. Birkhauser, Basel, 264 pp.

Wyss, M., & Habermann, R.E., 1988. Precursory seismic quiescence. Pure & App. Geoph., **126**(2-4), 319-332.

Yin, X.C., Wang, Y.C., Peng, K.Y., Bai, Y.L, Wang, H.T., & Yin, X.F., 2000. Development of a new Approach to earthquake prediction: load/unload response ratio (LURR) theory. Pure & App. Geoph., **157**(11-12), 2365-2383.

Zaliapin, I., Keilis-Borok, V.I., Axen, G., 2002. Premonitory spreading of seismicity over the faults' network in southern California: precursor accord. J. Geophys. Res., **B 107**, 2221.

Zechar, J.D., Jordan, T.H., Schorlemmer, D. & Liukis, M., 2007. Comparison of two earthquake predictability evaluation approaches, Molchan error trajectory and likelihood. Seism. Res. Lett., **78**(2), 250.

APPENDIX A:
Reverse Tracing of Precursors alarm specifications

Below we list the epicentral latitudes and longitudes for events forming each alarm

Alarm 1:

lats=38.60,38.60,38.90,39.75,37.40,36.73,34.95,37.89,35.42,35.29,35.28,38.17,35.31,36.40,37.72,37.74,38.05,37.76,36.43,35.83,36.86,40.35,36.36,39.81,37.06,42.87,40.35,38.89,35.92,41.74,41.53,38.70,38.47,38.02,42.50,40.14,39.62,37.32,42.52,42.66,41.73,37.67,37.25,37.62,40.15,40.11,37.16,38.62,39.13,38.00,38.76,40.38,40.39,35.42,33.54,37.01,39.56,35.40,41.03,41.05,35.40,34.20,35.40,36.04,42.08,39.29,38.84,36.73,37.83,33.75,34.05,35.21,41.62,37.39,44.11,36.98,38.47,42.20,41.93,35.63,36.47,41.88,38.04,37.96,37.68,41.88,38.25,41.94,37.18,36.08,36.11,35.36,39.63,40.15

lons=141.14,141.14,142.55,141.83,143.12,141.29,140.10,141.90,139.02,140.60,140.58,141.67,140.55,141.04,142.80,142.11,142.42,142.77,140.69,140.90,141.81,142.97,141.12,139.96,141.20,142.73,142.08,142.14,139.68,143.72,142.90,141.13,140.52,141.84,145.01,142.45,142.10,141.94,145.01,143.96,143.51,141.79,142.16,141.66,142.47,142.45,144.05,143.16,142.89,143.41,143.49,142.07,144.09,139.74,141.62,142.38,143.91,140.42,143.28,143.25,140.81,139.25,140.44,140.10,142.56,144.27,144.26,141.43,138.40,140.81,139.39,140.24,142.17,141.15,141.85,140.51,144.35,141.08,142.44,140.05,140.53,139.21,143.40,139.72,141.78,139.22,138.77,140.83,141.88,139.85,139.31,141.28,142.11,142.42

Alarm 2:

lats=36.60,35.85,36.04,36.64,36.56,35.88,38.80,38.80,36.55,38.45,37.31,36.97,37.76

lons=-120.74,-120.39,-120.61,-121.23,-120.71,-121.42,-122.73,-122.80,-121.10,-122.69,-121.67,-120.18,-122.57

Alarm 3:

lats=33.18,33.21,33.40,34.03,33.05,34.65,33.36,33.69,33.80,33.37

lons=-115.60,-116.15,-116.40,-116.39,-115.90,-116.29,-116.40,-116.03,-116.18,-116.31

Alarm 4:

lats=38.11,37.83,38.29,38.57,37.35,37.72,33.35,32.98,37.83,38.58,38.15,36.45,38.17,38.06,36.43,33.73,32.62,35.06,35.57,34.58,32.97,33.28,33.52,34.48,32.72,32.04,34.27,34.27,34.26,32.25,34.18,31.90,31.42,31.95,31.70,32.72,31.55,32.20

lons=139.27,138.13,140.25,144.34,140.01,144.54,140.58,139.57,142.70,139.94,143.59,140.61,143.58,143.66,141.17,140.91,141.84,141.08,141.13,140.67,141.88,142.03,140.90,140.54,140.42,141.88,139.19,139.20,139.18,141.59,140.42,140.85,140.05,142.62,141.40,140.64,141.40,140.83

Alarm 5:

lats=45.66,45.05,45.67,46.21,46.54,45.22,46.53,45.03,46.43,46.47

lons=14.32,14.55,14.19,13.98,13.85,14.92,13.23,14.83,12.69,12.83

Alarm 6:

lats=33.86,34.34,34.72,35.03,34.72,34.35,34.17,33.85

lons=-117.72,-116.91,-116.04,-116.91,-116.04,-116.84,-117.44,-117.77

Alarm 7:

lats=42.96,43.06,43.76,43.78,44.34,44.31,42.90,42.99,43.66,42.29,43.34,43.14,44.66,42.59,43.47,43.18,43.27,43.53,43.53,43.61,43.90,44.02,43.22

lons=-126.65,-127.04,-127.91,-128.08,-124.49,-124.56,-126.48,-126.35,-127.90,-126.52,-126.81,-127.75,-124.30,-126.81,-128.35,-126.31,-126.69,-127.03,-127.60,-127.28,-128.05,-128.66,-127.88

Alarm 8:

lats=41.68,42.08,41.69,41.87,41.65,42.53,42.51,43.08,41.87,43.46,43.19,43.09,42.81,42.67

lons=14.77,12.74,14.17,12.99,14.84,13.18,13.32,13.51,13.53,14.83,15.15,15.36,13.82,15.80

Alarm 9:

lats=42.52,42.41,41.83,43.08,43.09,43.27,41.92

lons=13.28,12.33,13.57,13.34,13.38,12.73,13.64

Alarm 10:

lats=37.16,36.63,34.79,36.94,36.37,35.41,37.69,35.68,35.79,35.14,33.99,37.41,37.06,37.20,37.37,32.90,31.86,33.61,35.73,36.90,35.49,36.78,33.33,31.80,36.66,32.27,33.35,34.30,36.03,35.84,33.91,35.63,36.63,35.27,36.63,34.29,33.84,38.54,34.51,32.53,33.68,32.20,35.55,35.55,39.29

lons=140.76,139.83,139.69,141.62,137.23,136.27,137.37,140.74,137.17,135.66,140.01,136.91,141.14,139.95,141.75,141.83,142.69,139.24,140.62,137.56,138.96,137.86,141.00,141.51,138.30,142.53,140.87,139.11,139.45,138.19,137.21,139.43,139.49,138.52,139.49,140.25,137.24,140.58,137.70,140.91,140.47,141.02,139.82,139.82,140.36

Alarm 11:

lats=34.11,34.13,34.80,34.84,33.84,33.45,33.22,32.56,34.33,33.49,33.52,34.05

lons=-117.30,-116.84,-116.27,-116.32,-117.05,-116.62,-116.20,-117.52,-116.46,-116.52,-116.57,-117.01

Alarm 12:

lats=34.60,34.33,33.61,35.13,34.24,33.91

lons=-116.36,-116.83,-117.27,-117.56,-117.45,-116.88

Alarm 13:

lats=32.11,32.42,32.44,32.15,32.25,32.34,32.19,32.17,32.19,32.06,32.18,32.62,34.02,32.57,34.29,32.87,34.58,32.29,33.79

lons=-116.43,-115.42,-115.40,-115.89,-115.19,-115.15,-115.07,-115.07,-115.87,-115.88,-115.87,-115.77,-117.56,-115.57,-116.83,-116.22,-116.26,-115.21,-116.18

Alarm 14:

lats=32.04,32.21,32.40,32.31,32.10,32.68,31.80,31.80,34.16,33.24,33.86,33.73,33.69,34.11

lons=-114.97,-115.08,-115.37,-115.23,-116.30,-115.86,-116.27,-116.27,-117.77,-116.04,-117.11,-117.48,-116.80,-117.32

Alarm 15:

lats=44.30,44.26,42.74,44.44,43.02,44.23,43.40,43.98,43.66,44.91,44.06,44.79,44.60

lons=10.67,11.01,12.77,9.89,12.91,10.47,13.49,11.82,10.19,9.23,9.01,11.83,10.41

Alarm 16:

lats=45.16,44.00,44.07,43.26,43.90,45.24,43.62,45.34,43.02,44.28,43.15,44.80,44.83,43.54,43.71,43.12,44.95,44.07,44.58,43.10,44.78,43.05,43.71,43.16,44.64,44.37,43.65,43.40,44.82,43.50,44.70,44.88,45.00,44.23,44.57,44.27,45.88,44.50,45.76,43.91,44.38,42.85,45.56,45.22,43.81,45.03,44.04,45.57,43.90,45.21,44.49,43.92,44.44,45.52,46.01,44.94,43.10,46.66,45.12,43.17,43.48,42.34,44.84,41.78,43.62,41.76,41.71,41.57,42.22,43.16,44.43,44.29,43.84,41.94,45.47,42.82,43.14,45.20,42.37

lons=147.59,148.30,148.17,146.25,146.76,149.13,149.63,150.33,148.46,148.38,146.87,148.30,151.47,146.73,146.40,146.80,147.59,148.18,146.79,146.78,146.73,147.77,147.26,147.33,146.67,147.33,147.26,148.14,149.17,147.25,148.76,148.76,149.63,146.65,147.96,146.10,151.02,150.74,150.90,147.56,149.18,147.26,152.50,149.54,147.72,150.88,147.91,151.26,148.30,151.70,149.72,150.46,149.69,150.94,151.14,150.12,148.46,151.92,149.69,147.65,147.46,144.62,150.36,143.68,144.91,142.89,144.26,142.05,144.41,146.92,148.35,147.36,147.46,142.32,149.60,143.37,146.10,149.66,143.92

Alarm 17:

lats=44.20,44.80,44.24,44.14,44.53,44.10,44.73,43.56,42.99

lons=-128.60,-128.29,-129.87,-129.27,-129.77,-128.37,-129.34,-128.58,-126.59

Alarm 18:

lats=44.78,44.16,43.03,42.43,45.35,45.20,44.80,43.77,45.14,45.35,43.24,44.50,44.37,43.12,44.40,43.87,44.74,45.49,46.38,42.50,44.21,46.02,45.46,43.71,45.37,43.44,43.19,45.13,43.02,43.51,43.18,43.57

lons=149.39,148.25,146.73,147.38,148.26,149.72,149.51,149.96,147.57,150.31,146.83,148.64,148.85,146.91,151.15,147.47,147.91,152.38,153.84,148.78,149.12,151.08,150.99,145.56,150.28,148.88,147.72,148.85,146.78,147.37,144.47,147.69

Alarm 19:

lats=51.52,51.43,52.94,52.16,52.47,52.10,51.64,51.48,51.43,50.17,50.74,51.76,50.02,51.01

lons=184.69,185.72,185.26,184.96,183.99,182.85,186.10,185.19,182.47,181.09,186.89,183.16,181.19,178.37

Alarm 20:

lats=35.86,36.24,35.63,38.65,36.85,36.17,36.17,37.86,36.47,37.86,36.59,38.39,38.39,37.12

lons=-120.41,-120.81,-120.75,-122.26,-121.45,-120.29,-120.29,-122.24,-121.04,-122.25,-121.19,-122.62,-122.62,-121.52

Alarm 21:

lats=32.39,32.40,32.18,32.69,33.45,32.77,32.18,33.07

lons=-115.12,-115.15,-115.88,-116.06,-116.59,-115.44,-115.89,-116.50

Alarm 22:

lats=35.78,37.18,37.86,37.31,37.34,37.90,36.76,36.56,35.52,35.78,37.12,36.25,35.60,36.74

lons=-121.38,-121.96,-122.24,-121.08,-121.71,-122.11,-121.27,-121.15,-120.81,-120.33,-121.52,-120.81,-120.84,-121.34

Alarm 23:

lats=51.32,51.44,51.92,51.40,52.28,51.44,53.19,52.23,52.65,52.89

lons=-178.17,178.28,-179.17,-176.53,-176.25,-177.19,-172.35,-173.94,-176.10,-175.57

Alarm 24:

lats=43.44,43.96,44.43,44.46,44.52,44.40,44.67,43.80

lons=-126.52,-127.61,-129.59,-129.67,-130.19,-129.76,-129.06,-128.29

Alarm 25:

lats=33.36,33.51,33.17,32.11,31.80,32.67,32.41,32.37,32.65,32.72,31.77,32.13

lons=-116.32,-116.47,-116.55,-115.83,-115.55,-116.12,-116.40,-115.22,-116.16,-115.41,-116.18,-115.89

Alarm 26:

lats=43.83,42.41,43.15,44.11,43.95,44.33,44.64,44.38,44.36,44.74

lons=-128.40,-126.84,-126.08,-129.32,-130.08,-129.07,-128.39,-130.43,-129.34,-128.94

APPENDIX B:
Finding expected value of Molchan trajectory jumps

We seek $\langle \tau_i \rangle$, the expectation of the i^{th} Molchan trajectory jump of an unskilled alarm function, where expectation is defined as

$$\langle X \rangle = \int_{-\infty}^{\infty} x f(x) \tag{B.1}$$

Here $f(x)$ is the probability density function; therefore, we need to find the probability density for τ_i.

We can find the probability density by taking the derivative of the cumulative density function. For τ_1, this is the probability that the trajectory has experienced at least one jump prior to reaching τ. In other words, it is the probability of covering τ and obtaining 1, 2, 3,…, or N hits. This probability is given by summing binomial terms:

$$D_{\tau_1}(\tau) = \sum_{j=1}^{N} \binom{N}{j} \tau^j (1-\tau)^{N-j} \tag{B.2}$$

We can express Equation B.2 in the following closed form:

$$D_{\tau_1}(\tau) = \begin{cases} 0, \tau < 0 \\ 1-(1-\tau)^N, \tau \in [0,1] \\ 1, \tau > 1 \end{cases} \tag{B.3}$$

By differentiating Equation B.3 with respect to τ, we obtain the probability density:

$$p_{\tau_1}(\tau) = \frac{dD_{\tau_1}(\tau)}{d\tau} = \begin{cases} N(1-\tau)^{N-1}, \tau \in [0,1] \\ 0, \quad \text{otherwise} \end{cases} \tag{B.4}$$

Now we can substitute Equation B.4 into Equation B.1 to obtain the expectation,

changing the limits of integration to isolate the region where the probability density is nonzero:

$$\begin{aligned}\langle \tau_1 \rangle &= \int_0^1 \tau p_{\tau_1}(\tau) d\tau = \int_0^1 \tau N(1-\tau)^{N-1} d\tau \\ &= -\frac{(1-\tau)^N (N\tau+1)}{N+1}\Bigg|_0^1 \\ &= \frac{1}{N+1}\end{aligned} \tag{B.5}$$

This shows that the first hit is expected to be obtained by unskilled alarm functions when they cover $\frac{1}{N+1}$ of the study region. Similarly, we can express the c.d.f., p.d.f., and expectation for the next jump:

$$D_{\tau_2}(\tau) = 1-(1-\tau)^N - N\tau(1-\tau)^{N-1} \tag{B.6}$$

$$\begin{aligned}p_{\tau_2}(\tau) &= N(1-\tau)^{N-1} - \left[-N\tau(N-1)(1-\tau)^{N-2} + N(1-\tau)^{N-1}\right] \\ &= N(N-1)\tau(1-\tau)^{N-2}\end{aligned} \tag{B.7}$$

$$\begin{aligned}\langle \tau_2 \rangle &= \int_0^1 N(N-1)\tau^2(1-\tau)^{N-2} d\tau \\ &= -\frac{(1-\tau)^{N-1}\left((N-1)\tau(N\tau+2)+2\right)}{N+1}\Bigg|_0^1 \\ &= \frac{2}{N+1}\end{aligned} \tag{B.8}$$

In words, on average, unskilled alarm functions obtain 2 hits once they have covered $\frac{2}{N+1}$ of the study region. Likewise for the following jump, we have

$$D_{\tau_3}(\tau) = 1-(1-\tau)^N - N\tau(1-\tau)^{N-1} - \binom{N}{2}(\tau)^2(1-\tau)^{N-2} \tag{B.9}$$

$$
\begin{aligned}
p_{\tau_3}(\tau) &= N(1-\tau)^{N-1} - \left[-N(N-1)\tau(1-\tau)^{N-2} + N(1-\tau)^{N-1}\right] \\
&\quad - \left[-\binom{N}{2}(N-2)(\tau)^2(1-\tau)^{N-3} + 2\binom{N}{2}\tau(1-\tau)^{N-2}\right] \\
&= N(N-1)\tau(1-\tau)^{N-2} + \binom{N}{2}(N-2)(\tau)^2(1-\tau)^{N-3} - N(N-1)\tau(1-\tau)^{N-2} \\
&= \frac{N(N-1)(N-2)}{2}(\tau)^2(1-\tau)^{N-3}
\end{aligned}
\tag{B.10}
$$

$$
\begin{aligned}
\langle \tau_3 \rangle &= \frac{N(N-1)(N-2)}{2}\int_0^1 (\tau)^3(1-\tau)^{N-3}\,d\tau \\
&= \frac{N(N-1)(N-2)}{2}(1-\tau)^N\left(-\frac{2N+3}{N(N+1)} - \frac{\tau}{N+1} - \frac{3}{(N-1)(\tau-1)} + \frac{1}{(2-N)(\tau-1)^2}\right)\Bigg|_0^1 \\
&= \frac{N(N-1)(N-2)}{2}\left(\frac{2N+3}{N(N+1)} - \frac{3}{(N-1)} - \frac{1}{(2-N)}\right) \Rightarrow
\end{aligned}
$$

$$
\langle \tau_3 \rangle = \frac{3}{N+1} \tag{B.11}
$$

For an inductive proof, we assume that $\langle \tau_{N-1} \rangle = \frac{N-1}{N+1}$ and compute $\langle \tau_N \rangle$. At this point, we can get a compact expression for the c.d.f. by returning to the original formulation in Equation B.2, such that

$$
\begin{aligned}
D_{\tau_N}(\tau) &= \sum_{j=N}^{N}\binom{N}{j}\tau^j(1-\tau)^{N-j} \\
&= (\tau)^N
\end{aligned}
\tag{B.12}
$$

Then,

$$
p_{\tau_N}(\tau) = N(\tau)^{N-1} \tag{B.13}
$$

$$\langle \tau_N \rangle = \int_0^1 N\tau(\tau)^{N-1} d\tau$$

$$= N\int_0^1 (\tau)^N d\tau$$

$$= \frac{N(\tau)^{N+1}}{N+1}\Bigg|_0^1$$

$$\langle \tau_N \rangle = \frac{N}{N+1} \tag{B.14}$$

This completes the proof and thus, for all jumps, the expected value of the jump is described:

$$\langle \tau_i \rangle = \frac{i}{N+1} \tag{B.15}$$

APPENDIX C:
Method for finding area skill score moments

We find the following relation for moments about the origin:

$$\mu_k = \frac{(-1)^k}{N^k} \sum_{t=1}^{T_p} A_t^{(k)} B_t^{(k)} C_t^{(k)}(N) \tag{C.1}$$

Here, N is the number of observed target earthquakes and

$$A_m^{(k)}(k_1, k_2, \ldots, k_m) = \frac{k!}{k_1! k_2! \ldots k_m!} \tag{C.2}$$

$$\begin{aligned} B^{(k)}(k_1, k_2, \ldots, k_m) &= \int_0^1 d\tau_m \ldots \int_0^1 d\tau_1 {\tau_1}^{k_1} \ldots {\tau_m}^{k_m} \\ &= \prod_{j=1}^{m} \frac{1}{k_j + 1} \end{aligned} \tag{C.3}$$

$C_t^{(k)}(N)$ is the number of distinct permutations of each N-tuple of type $t \in (1,2,\ldots T_p)$ with sum k. We're interested in integral tuples that are of a given length, are ordered, and produce a given sum. If we express a tuple as

$$K = (k_1, k_2, k_3, \ldots, k_x)$$

we say its **length** is x (or, alternatively, call it a x**-tuple**), its **sum** is $s = \sum_{i=1}^{x} k_i$, and it is **ordered** if and only if

$$k_i \geq k_{i+1} \;\forall\; i \in [1, x-1]$$

We also restrict our study to integral tuples such that (k_1, k_1, …, k_N) are non-negative integers.

We'll denote the number of ordered x-tuples with sum s as $N_x(s)$, and the number of such

tuples beginning with a given first digit d as $N_x(s \mid d)$. For example, $N_1(1)$ is the number of 1-length ordered tuples with unit sum. It should be apparent that there is only one such 1-tuple: $K = (1)$. Likewise, $N_2(2 \mid 2)$ is the number of 2-tuples with sum 2 and leading digit 2. It should be apparent that there is only one such 2-tuple: $K = (2, 0)$. In general, we make the following observations:

i. We can find the number of ordered x-tuples with sum s by fixing the first digit and reducing by one the length of tuples we seek; we repeat this process for each value between 1 and s and sum the number of tuples found for each value.

ii. For any length x, there is only one ordered tuple with a sum of zero. In particular, this tuple is $K = \left(\overbrace{0,0,0,\ldots,0}^{x \text{ terms}} \right)$.

iii. For any length x, there is only one ordered tuple with unit sum. In particular, this tuple is $K = \left(1, \overbrace{0,0,\ldots 0}^{(x-1)\text{terms}} \right)$.

iv. For any length x, there is only one ordered tuple with sum s having leading digit s. In particular, this tuple is $K = \left(s, \overbrace{0,0,\ldots,0}^{(x-1)\text{terms}} \right)$.

v. For any length $x > 1$, there is only one ordered tuple with sum s with leading digit (s-1). In particular, this tuple is $K = \left(s-1,1, \overbrace{0,0,\ldots 0}^{(x-2)\text{terms}} \right)$.

vi. For any length x, there is only one ordered tuple with sum x with leading digit 1.

In particular, this tuple is $K = \left(\overbrace{1,1,\ldots 1}^{x \text{ terms}} \right)$.

vii. For any length x, there are no ordered tuples with sum s if the leading digit is greater than s.

viii. The number of x-tuples with sum s and leading digit d can be obtained by finding the number of (x-1)-tuples with sum (s-d) minus the number of (x-1)-tuples that have a leading digit greater than d.

Using the notation above, we can express these observations more compactly:

1. $N_x(s) = \sum_{i=1}^{s} N_x(s \mid i)$

2. $N_x(0) = 1$

3. $N_x(1) = 1$

4. $N_x(s \mid s) = 1$

5. $N_x(s \mid s-1) = 1, \forall x > 1$

6. $N_x(s \mid 1) = 1$

7. $N_x(s \mid y) = 0, \forall\ y > s$

8. $N_x(s \mid d) = N_{x-1}(s-d) - \sum_{j>d} N_{x-1}(s-d \mid j)$

To find T_p, the number of ordered p-tuples with sum p, we need only find $N_p(p)$ using

equations 1-8. Table C.1 provides results for small p:

Table C.1: Number of tuples

p	T_p
1	1
2	2
3	3
4	5
5	7
6	11
7	15
8	22
9	30
10	42
11	56
12	77
13	101
14	135
15	176

Table C.1 For p = 1 to 15, the number of ordered p-tuples with sum p

We can use these findings to determine moments about the origin. In the case where n = 1, we have

$$A_1^{(1)} = 1$$
$$B_1^{(1)} = \frac{1}{2}$$
$$C_1^{(1)}(N) = N$$

By substitution, we find that μ_1 = ½.

APPENDIX D:
Number of nonzero sums of a set's elements

In the case of a discretized reference model, the set of attainable values of τ is finite and its elements are the nonzero linear combinations of the reference model values with coefficients equal to zero or one. In other words, for a reference model specified in j bins, the set T_j of attainable τ values is comprised of the nonzero sums of the j reference model values. We denote the cardinality of this set $|T_j|$.

Theorem. $|T_j| = 2^j-1$.

Proof. In the case where j=1, it is clear that there is only one nonzero sum: $T_1=\{1\}$, $|T_1|=1$. In the case where j=2, we represent a reference model as the set $\{n_1, n_2\}$. In this case, the set of attainable τ values is $\{n_1, n_2, n_1+n_2\}$; $|T_2|=3$. When j=3, we represent a reference model as the set $\{n_1, n_2, n_3\}$ and the set of sums is $\{n_1, n_2, n_3, n_1+n_2, n_1+n_3, n_2+n_3, n_1+n_2+n_3\}$; $|T_3|=7$. We assume that the relation holds for all values of j up to and including (x-1) and we consider $j=x$. The set T_x will contain all elements of $T_{(x-1)}$ as well as each of these elements added to $n_{(x+1)}$; the only additional sum in T_x is $n_{(x+1)}$. Thus the cardinality $|T_x|$ is twice the cardinality $|T_{(x-1)}|$ plus one additional element:

$$\left|T_x\right| = 2\left|T_{(x-1)}\right| + 1 = 2\left(2^{(x-1)} - 1\right) + 1 = 2^x - 1$$