Earthquake statistics and likelihood model testing in California

Author(s):
Schorlemmer, Danijel

Publication Date:
2004

Permanent Link:
https://doi.org/10.3929/ethz-a-004937819

Rights / License:
In Copyright - Non-Commercial Use Permitted
Diss. ETH No. 15741

Earthquake Statistics and Likelihood Model Testing in California

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY

for the degree of

DOCTOR OF NATURAL SCIENCES

presented by

DANIJEL SCHORLEMMER

Diploma of Geophysics (Dipl. Geophys.)
Technical University Berlin
born November 23, 1969
citizen of Germany

accepted on the recommendation of

Prof. Dr. Domenico Giardini, examiner
Dr. Stefan Wiemer, co-examiner
Prof. Dr. David D. Jackson, co-examiner
Prof. Dr. Max Wyss, co-examiner

Zürich, 2004
Contents

1 Zusammenfassung .................................................. 1

2 Abstract .................................................................. 3

3 Introduction .................................................................. 5

4 Stability and Significance Tests for \( b \)-value Anomalies ................................................................. 11
   4.1 Abstract .................................................................. 12
   4.2 Introduction .......................................................... 12
   4.3 Data and Method ..................................................... 13
   4.4 Results .................................................................... 17
   4.5 Discussion and Conclusion ....................................... 21
   4.6 Acknowledgments .................................................... 22

5 Earthquake-Size Distribution: A New Stressmeter? ............................................................... 23
   5.1 Abstract .................................................................. 24
   5.2 A New Stressmeter? .................................................. 24
   5.3 Acknowledgments ..................................................... 32

6 Earthquake Statistics at Parkfield I .......................................................... 33
   6.1 Abstract .................................................................. 34
   6.2 Introduction .......................................................... 35
   6.3 Data ....................................................................... 37
   6.4 Method ................................................................. 42
6.5 Results ................................................. 45
  6.5.1 Mapping $b$-values ............................. 45
  6.5.2 Stationarity of $b$-values ................. 51
  6.5.3 Analysis of stationarity for different periods . 57
6.6 Discussion and Conclusions ..................... 57
  6.6.1 Investigating stationarity of $b$-values at Parkfield ... 60
  6.6.2 Implications for earthquake hazard ........ 65
6.7 Acknowledgments .................................... 67

7 Earthquake Statistics at Parkfield II 69
  7.1 Abstract ............................................ 70
  7.2 Introduction ....................................... 70
  7.3 Method ............................................. 73
    7.3.1 Parameter Scan ............................... 80
  7.4 Data ............................................... 81
  7.5 Results ............................................ 81
    7.5.1 Test with cylindrical sampling volumes .... 82
    7.5.2 Test with cuboidal sampling volumes ...... 85
    7.5.3 Parameter space scan ......................... 88
  7.6 Discussion and Conclusions ..................... 91
  7.7 Acknowledgments .................................. 98

8 Earthquake Likelihood Model Testing 99
  8.1 Abstract ............................................ 100
  8.2 Introduction ....................................... 100
  8.3 Basic Ideas and Definitions .................... 102
    8.3.1 Definitions ................................... 107
  8.4 Computation ....................................... 108
    8.4.1 Uncertainties in Earthquake Parameters .... 109
    8.4.2 Simulation and Evaluation ................... 111
      Data-consistency test or L-Test ................. 112
Contents

Number test or N-Test ......................... 113
Hypotheses Comparison or R-Test ............ 114
8.5 Evaluation .................................. 117
8.6 Definitions ................................ 118
8.7 Examples .................................. 121
   8.7.1 Model H$^0$ ........................... 122
   8.7.2 Consistency of Model H$^0$ with the Observation .... 123
   8.7.3 Model H$^1$ ........................... 124
   8.7.4 Consistency of Model H$^1$ with the Observation .... 125
   8.7.5 Model comparison ..................... 125
   8.7.6 Tests over different time periods ............ 127
8.8 Discussion ................................ 130
8.9 Appendix .................................. 132
   8.9.1 Likelihood ratio independence on bin-sizes ....... 132

9 Conclusions ................................. 135
  9.1 Statistical Methods and $b$-values .......... 135
  9.2 Earthquake forecast testing .............. 136
  9.3 Discussion and Outlook ................... 137

List of Figures .............................. 139

Bibliography ................................. 145

Curriculum Vitae ............................. 157

Acknowledgments ............................. 159
Kapitel 1

Zusammenfassung


Die Abhängigkeit des $b$-Wertes von Spannungen wird in einem statistischen Seismizitätsmodell ($H^1$) ausgeschöpft. Dieses Modell extrapoliert die zukünftige Seismizität aus den räumlich unterschiedlichen Magnituden-Häufigkeitsverteilungen. Hierbei nehme ich an, daß räumlich variierende $b$-Werte wichtige Informationen über zukünftiges Erdbebenauftreten enthalten. Der erste Schritt, um diese Hypothese zu beweisen, ist eine Überprüfung der Stationarität von $b$-Werten. Als Testregion habe ich das Parkfield Segment der San Andreas Verwerfung gewählt, da qualitativ hochwertige Katalogdaten verfügbar sind und es bereits intensiv durch verschiedene geophysikalische Verfahren untersucht worden ist. Für einen Querschnitt entlang der Verwerfung zeige ich, daß die $b$-Werte zum überwiegenden Teil und über drei Jahrzehnte stationär sind. Der zweite Schritt umfaßt einen rigorosen Likelihood-Test dieses Modells gegen ein alternatives Modell ($H^0$), das zukünftige Seismizität mittels Extrapolation räumlich variierender $a$-Werte bestimmt. Bei diesem Modell wird allerdings ein konstanter $b$-Wert angenommen, der dem Mittelwert der gewählten Region ($b \approx 0.9$) entspricht. Ausführliche Tests, in denen die Lern- und Beobachtungsperiode sowie die $b$-Wertbestimmungsparameter variiert werden, zeigen die höhere Prognosequalität des Modells $H^1$. Das Modell $H^0$ kann für die meisten Parameterkombinationen zu Gunsten des Modells $H^1$ für das gegebene Signifikanzniveau von 0.05 verworfen werden.

Chapter 2

Abstract

I develop advanced statistical methods to extract more information from earthquake catalogs. One important parameter of earthquake statistics is the $b$-value in the frequency-magnitude distribution $\log N = a - bM$. Its spatial and temporal variations have been established in numerous publications, however, computing $b$-values requires high quality datasets. Here I introduce a number of methods for estimating the significance of $b$-value anomalies for low seismicity regions, exemplarily shown for the subducting slab in the Tyrrhenian Sea.

Although variations in $b$ have been observed, their physical causes still remain unclear. Acoustic emission experiments and microearthquakes in mines show a clear stress dependence of the $b$-value. $b$-values as a function of the rake angles of earthquake focal mechanisms show a clear dependence: Normal events exhibit high $b$-values, thrust events low $b$-values, and strike-slip events intermediate ones. Because the three classes of events are representing different stress regimes and because the investigated five earthquake catalogs span the entire magnitude range of instrumental data for regional and global catalogs ($M2 - M8$), this study establishes that stress is the major contributor to different size distributions of earthquakes.

The dependence of $b$ on stress is exploited in a statistical seismicity model ($H^1$) describing future seismicity by extrapolating the spatially varying fre-
quency-magnitude distributions in time and magnitude. I assume that spatially varying $b$-values contain important information about future earthquake occurrence. The first step in evaluating this hypothesis is establishing the stationarity of $b$-values. As a test region, I investigate the Parkfield segment of the San Andreas fault in California. Parkfield is chosen because it has been extensively studied with other geophysical techniques and possess a high quality earthquake catalog data. I show that $b$-values, mapped in cross-section, remain mostly stationary for about three decades (1969–2003). In a second step, a rigorous likelihood test of this model against an alternative model ($H^0$), which extrapolates future seismicity using also spatially varying $a$-values but a constant regional $b$-value ($b \approx 0.9$), is performed. Extensive testing, for varying learning and observation periods of the models as well as for varying parameters for $b$-value computation, show the superior performance of model $H^1$. For most parameter selections, $H^0$ can be rejected in favor of $H^1$ at the given significance level of 0.05.

This testing method is further developed into a full testing suite covering different aspects of earthquake likelihood model testing: Data-consistency as well as spatial comparative performance tests. It will be used in the RELM (Regional Earthquake Likelihood Models) framework, evaluating the forecasting capabilities of a variety of quasi-stationary and fully time-dependent models for California. My model $H^1$ will also be tested in a prospective sense during a testing period of 5 years (2005-2010).
Chapter 3

Introduction

The $b$-value from the Gutenberg-Richter relation $\log N = a - bM$ describes the slope of the frequency-magnitude distribution of earthquakes [Ishimoto and Iida, 1939; Gutenberg and Richter, 1944]. This power-law is equivalent to a fractal distribution $N = C/A^{D/2}$ with a fractal dimension $D = 2b$ [Aki, 1981; Huang and Turcotte, 1988]. Although many scientists think that the $b$-value is a constant [Frohlich and Davis, 1993; Kagan, 1999], spatial and temporal variations have been observed. The idea of a constant $b$-value is supported by the assertion that $b$-value computations are mainly influenced by the corner magnitude $m_c$ [Kagan, 2002], network reconfigurations, or magnitude computation changes. This would indicates that varying $b$-values are artifacts rather than reflecting real changes in seismic activity. In contrast to this, Scholz [1968], Wyss [1973], and Urbancic et al. [1992] linked low $b$-values to high-stress environments.

This link has further been supported by various studies. At the Parkfield segment of the San Andreas fault, Wiemer and Wyss [1997] discovered strong spatial heterogeneities of $b$-values along the fault. A patch with a strong $b$-value anomaly was identified as asperity. This technique was extended to the San Jacinto-Elsinore fault zone by Wyss et al. [2000]. They proved that local recurrence time, derived from the $a$- and $b$-value, can be used to identify asperities. Further studies include $b$-value imaging of volcanoes to identify
magma chambers [Wyss et al., 1997; Wiemer and McNutt, 1997; Murru et al., 1999] and of subducting slabs to find locations of slab dehydration and the path- way of the magma through the mantle wedge [Wiemer and Benoit, 1996; Wyss et al., 2001]. Also, variations of \( b \)-values with depths have been observed [Mori and Abercrombie, 1997; Gerstenberger et al., 2001].

The \( b \)-value is a property of a set of earthquakes. Therefore, it is sensitive to changes in this set of events. Different statistical methods have been developed to ensure stable results when determining \( b \)-values. They range from formulating the maximum-likelihood estimation of the \( b \)-value [Utsu, 1965; Aki, 1965; Bender, 1983] and its error estimate [Shi and Bolt, 1982] to different completeness estimation methods [Wiemer and Wyss, 2000, 2002].

All methods require large datasets for reliable determination of \( b \)-values. The next step in \( b \)-value statistics should be the development of more sophisticated statistical approaches to overcome the problems of strongly varying completeness and low seismicity. A promising target for this task was the subducting slab in the Tyrrhenian Sea. In Chapter 4, I introduce new statistical methods, including bootstrap procedures for estimating errors in \( b \)-value computations and methods to establish the significance of mapped anomalies. This work also continues the aforementioned \( b \)-value studies on subducting slabs, mapping loci of slab dehydration as anomalies of high \( b \)-values. This work was published 2003 in *Geophysical Research Letters*.

Besides the statistical problems of \( b \)-value determination, the meaning of \( b \)-values is a debated issue. In various studies, \( b \)-values have been correlated with stresses [Scholz, 1968; Wyss, 1973; Urbancic et al., 1992; Amitrano, 2003], material heterogeneities [Mogi, 1962], and pore pressure [Wyss, 1973; Wiemer et al., 1998]. Evidence for these correlations have been produced in laboratory experiments [Urbancic et al., 1992; Amitrano, 2003], however, the extrapolation from samples to earthquakes in general is questionable. Because nobody can adjust stress conditions in the Earth’s crust to create defined environments for measuring how \( b \)-values and stresses correlate, one must let nature
break the secret. Different styles of faulting can differentiate events into classes representing different stress conditions. In Chapter 5, I investigate the dependence of $b$-values on different styles of faulting in five high-quality datasets. In each catalog, normal events show high value, strike-slip events intermediate, and thrust low $b$-values, thus proving that stress is the major contribution to different size distributions of earthquakes, closing the gap in the magnitude range for which this relationship has been established. This work is submitted to *Nature*.

Given this relation between stress and $b$-values, forecasting events on basis of spatially varying $b$-values is a promising research target. If $b$-values are representing the state of stress in the crust and if we assume the conditions to be relatively stable, today’s $b$-values should be a governing parameter for future earthquake-size distributions. The necessary first step to undertake is the analysis of spatial and temporal behavior of $b$-values. I continued the work of *Wiemer and Wyss* [1997] and analyzed quantitatively the stationarity of the $b$-value pattern along the Parkfield segment of the San Andreas fault in California. In Chapter 6, I spatially computed the Utsu-test [Utsu, 1992] for different division of the earthquake catalog, searching for changes in $b$-values over both periods. For most of the catalog divisions, $b$-values exhibit perfect stationarity. For some divisions, patches do show changes in $b$-values which I partly correlated with a creep episode in the locked section of the fault. Nevertheless, the portion on non-stationary behavior is minor and $b$-values should be considered stationary at Parkfield, thus qualifying for a parameter describing long-term seismic activity. This work was published 2004 in *Journal of Geophysical Research*.

Given the stationarity of $b$-values, one can propose that a forecast generated using spatially varying $b$-values (hypothesis $H^1$) must be superior to a same forecast which only ignores the variations in $b$-values (hypothesis $H^0$). $H^0$ is one of the most common approaches in seismic hazard analysis. Therefore, the next step must be a rigorous test whether this hypothesis holds and
whether the model using spatially varying $b$-values does show a significantly superior performance in forecasting events. This test can also be used to explore the optimal computational parameters for $b$-values. In Chapter 7, I introduce likelihood-ratio testing of earthquake forecasts and found the basis for statistical evaluation of these tests. Similar to the investigations in Chapter 6, I scanned the entire parameter space, hereby varying learning and observation periods in the pseudo-prospective tests. Furthermore, the parameters for $b$-value computations have been scanned. This multi-dimensional parameter study lead to the result, that $H_1$ shows a significant superior performance in forecasting events compared to $H_0$, allowing to reject the latter. This work was published 2004 in *Journal of Geophysical Research*.

A hypothesis showing higher performance than another hypothesis does not mean that it is consistent with the observation. The forecasts can under- or overpredict the seismicity. The proposed forecasting model tests need to be extended to cover this problem. A considerable effort in comparing and evaluating a variety of forecasting models will start 2005 within the scope of the ambitious RELM framework (Regional earthquake likelihood models) in California. RELM is committed to developing a variety of physical and statistical forecast models for California, test them against each other in a truly prospective test, and use the results to further develop seismic hazard analysis. To achieve this goal, spatial performance, total forecasting capabilities, magnitude dependend forecasting performance, and implications on hazard analysis need to be evaluated. In Chapter 8, I developed the full test suite for earthquake likelihood models for RELM. It consists of data-consistency and comparative tests and covers location, magnitude, and focal mechanisms errors. Also, it describes procedures for testing quasi-stationary models (as in Chapter 7) as well as for time-dependent models. This work will be submitted to a special volume covering RELM in *Seismological Research Letters*.

Summing up the parts of my thesis, improvement of statistical methods to gain more information about the physics of earthquakes appears as the main
issue. While the main focus is on $b$-values, the results may improve not only $b$-value statistics but also hazard analysis. Evaluation of earthquake forecast models is the second main target of my research. Earthquake forecasting and earthquake prediction work has in the past been plagued by the lack of rigor in testing such forecast. It is my opinion that only rigorous statistical testing, following methods such as the ones outlined in this thesis, will in the long run allow seismologists to move towards improved seismic hazard assessment. This is particularly important when considering time-dependent hazard assessment.
Chapter 4

Stability and Significance Tests for $b$-value Anomalies: Example from the Tyrrhenian Sea

Danijel Schorlemmer$^1$, Giancarlo Neri$^2$, Stefan Wiemer$^1$, and Antonino Mostaccio$^3$

$^1$ Swiss Seismological Service, ETH Zürich, Zürich, Switzerland.
$^2$ Dipartimento di Scienze della Terra, Università di Messina, Salita Sperone, 31 - 98166 Messina, Italy.
$^3$ Istituto Nazionale di Geofisica e Vulcanologia - Sezione di Catania, Piazza Roma, 2 - 95123 Catania, Italy.

4 Stability and Significance Tests for $b$-value Anomalies

4.1 Abstract

We spatially map $b$-values in the Southern Tyrrhenian Subduction Zone (Italy). A high $b$-value anomaly 90–200 km beneath the Aeolian Island volcanoes is interpreted as being related to the origin of magma for the Aeolian volcanic district. This result is independent of the sample size, width of the cross-section or its orientation. To assess the stability of these findings, we develop a new, more realistic, uncertainty assessment based on a bootstrap method. Our approach explicitly takes into account the uncertainty in the assessment of the completeness in magnitude reporting. To quantify the significance of the results when compared to a chance occurrence, we compute Monte Carlo style simulations based on synthetic catalogs as well as on catalogs with permuted magnitudes. Our entire tests show that the high $b$-value anomaly is stable and highly significant.

4.2 Introduction

Several studies have investigated volcanic areas with $b$-value imaging ($b$-value from the frequency-magnitude distribution of earthquakes, also called the Gutenberg-Richter relation, \cite{GutenbergRichter1944}). \cite{Wyss1997}, \cite{WiemerMcNutt1997}, and \cite{Murru1999} presented $b$-value imaging of volcanoes proposing locations of magma chambers. Investigations of subducting slabs beneath volcano chains were conducted by \cite{WiemerBenoit1996} in New Zealand and Alaska and \cite{Wyss2001} in Japan. These two studies report high $b$-value anomalies on top of the subducting slabs at depths between 100 km and 150 km. They proposed these anomalies to be related to the magma sources of volcanic chains above them. At such locations, dehydration of the subducting oceanic crust increases pore pressure, lowering the effective stress and thus increasing the $b$-value of the frequency-magnitude distribution \cite{Scholz1968}. \cite{Wyss2001} combined $b$-value imaging with
4.3 Data and Method

Seismic tomography to correlate a high $b$-value anomaly with low velocities in the mantle wedge in the subduction zone beneath northeastern Japan. They proposed a contorted pathway of the magma, but for Alaska and New Zealand Wiemer and Benoit [1996] proposed vertical magma paths.

The Tyrrhenian Sea is a roughly triangular-shaped basin elongated towards Calabria, bounded by the Corsica-Sardinia block on the west, Sicily on the south and the Italian peninsula on the northeast. The southeastern and southern margins of the basin show a horst-graben structure with numerous superimposed volcanic edifices forming the ring-shaped Aeolian volcanic district. Deep earthquake activity along a northwest dipping Wadati-Benioff plane is a feature of the Southern Tyrrhenian area, [Peterschmitt, 1956; Ciminini, 1999; Lucente et al., 1999]. Many investigators ([Malinverno and Ryan, 1986], [Faccenna et al., 2001], [Savelli, 2001] among others) have proposed a geodynamic model assuming southeast rollback of the Ionian lithosphere subducting beneath the Tyrrhenian lithosphere (Figure 4.1). Rollback is widely believed to have been a primary tectonic source for opening the Tyrrhenian basin.

In this paper we present $b$-value imaging at a new location. We have chosen the Tyrrhenian Sea (Figure 4.1) with its Aeolian volcanic arc. One of our main goals is to present new qualitative and quantitative methods of validating data, testing result stability, and establishing significance of results in $b$-value imaging. Data validation includes recording quality and completeness assessments as well as rigorous error estimates. We also simulated $b$-value imaging using randomly permuted catalogs to establish the significance of the results.

4.3 Data and Method

The earthquake catalog provided by the Istituto Nazionale di Geofisica e Vulcanologia (INGV) covers all of Italy and its surrounding regions from 1975 to 2001.2. Duration magnitudes $M_D$ are used throughout the catalog and lo-
Figure 4.1: Location map of the study area in the Tyrrhenian Sea. Epicenters of earthquakes (INGV catalog 1988.5–2001.2, $M > 2.0$) are marked with dots, volcanoes are marked with black triangles, and seamounts are marked with white faced triangles. The rectangle marks the volume of the cross-section perpendicular to the strike of the trench (from P1 to P2).
cal magnitudes $M_L$ are available for 1985–2001. However, these magnitude scales do not match, [Giardini et al., 1997; Malagnini et al., 2000]. Due to such differences between the two scales, we decided to use only the later part of the catalog with local magnitudes to avoid errors in computing $b$-values.

An important issue in assessing data quality is the magnitude of completeness $M_c$. We computed an overall $M_c = 2.4$ for the study region (Figure 4.1). Although completeness is a function of space, we can utilize this value for detecting changes in recording quality. Examining the cumulative number of events ($M \geq 2.4$) shows a change in the rate of detected events at 1988.5 with a steeper slope. Therefore, we cut the catalog at 1988.5. We mapped the spatial distribution of $M_c$ for the Tyrrhenian Sea and Sicily as proposed by Wiemer and Wyss [2000] (radius = 40 km with at least 40 events per grid node, 0.1° latitude/longitude spacing) and estimated $M_c = 2.0–2.3$ in northern Sicily and $M_c \geq 2.5$ offshore. Previous studies of $b$-values at subducting slabs used catalogs with similar completeness ($M_c \sim 2.0–2.5$ in [Wiemer and Benoit, 1996; Wyss et al., 2001]) but with more events. Due to heterogeneous completeness and fewer events, we applied here different techniques of data validation to ensure stable results and eliminate any possibility of artifacts. We cut the catalog at the lowest detected completeness value ($M_c = 2.0$) but at every grid node we estimated $M_c$ and cut the catalog accordingly. Simply using the highest $M_c$ detected is a possible approach but looses many events in volumes with lower $M_c$. Additionally, we cross-checked our results with the INGV recompiled catalog (1981–1996) which was obtained by joining all the data of the INGV stations with the data of the stations belonging to the local seismic networks managed by CNR, universities and local institutions [ING-GNDT, 2001].

We computed maximum likelihood $b$-values according to Utsu [1965] and Aki [1965]. Shi and Bolt [1982] gave an error estimate for maximum likelihood $b$-values. We considered this error estimate (standard deviation) too low due to its assumption of a correctly estimated $M_c$. Instead, we propose a bootstrap
approach to estimate errors in \( b \) and \( M_c \), [Chernick, 1999]. At every grid node we draw its detected number of events from its population, allowing any event to be selected more than once. From these events we compute \( M_c \) and \( b \) for \( M \geq M_c \) and repeat this for every grid node 1000 times. We then estimate the errors as the standard deviation of the \( b \)-values, taking into account the imperfection of the local frequency-magnitude distribution.

For mapping the subducting slab in the Tyrrhenian Sea, we defined a cross-section perpendicular to the strike of the trench with a width of 150 km from P1: 39.31N, 14.45E to P2: 37.91N, 17.19E (rectangle in Figure 4.1). We chose a width of this size due to the overall low number of events. Mt. Etna seismicity is not included in the cross-section. All 2402 events in this volume were projected onto the vertical cross-section plane. We spanned a grid over this plane with 2 \( \times \) 2 km grid node spacing. At each grid node we sampled all events within a radius of 40 km and determined a \( b \)-value if at least \( N_{\text{min}} = 40 \) events were available, as described by Wiemer and Wyss [2002]. These parameter settings were selected as the best overall fit for coverage, resolution, and reliability of results. Using smaller radii or a higher number of required events would have led to unsatisfactory coverage of the subducting slab. Therefore, we must examine the frequency-magnitude distributions of interest to ensure correctly computed \( b \)-values. Still, we lack the inherent errors of the computed \( b \)-values. Here, bootstrapping gives us good estimates of errors in \( b \)-value computations for interpretations. All investigations were performed using ZMAP, [Wiemer, 2001].

We performed three additional tests to ensure the stability and significance of our results. Firstly, we rotated the strike of the cross-section around its center by 10 degrees in both directions and computed the \( b \)-values for comparing with the original imaging. Secondly, we changed the width of the cross-section for detecting possible small volumes with a large influence on the result. Thirdly, we computed the significance level \( \alpha \) or Type-I Error of a hypothesis that the observed \( b \)-value contrast (the maximum difference in de-
Results

In the $b$-value cross-section of the slab (Figure 4.2B) we found an overall $b$-value of 1.64: the lowest $b$-value is 0.7 and the highest is 3.3. Therefore, we consider a volume with $b > 2$ as a high $b$-value anomaly. This range of detected $b$-values is higher than typically expected. As has been observed previously by Gasperini [2002], in the INGV catalog the lower magnitudes are overestimated while the higher are underestimated, shrinking the available magnitude range and raising $b$-values overall. However, anomalies can still be detected by investigating relative $b$-value differences.

We observed a high $b$-value anomaly extending from 90 km to about 200 km depth with the highest ($b > 3$) at about 120 km depth on top of the subducting slab. In the center of the slab we observed $b = 2.3 \pm 0.2$ and at the bottom $b = 1.5 \pm 0.4$. Due to the poor event population, we lack sufficient lateral resolution.

The INGV recompiled catalog shows the same anomaly with its peak value at the same location, but with lower $b$-values overall. The $b$-values of the recompiled catalog are on average 0.5 units smaller than in the original. $M_c$ is also lower in the recompiled catalog. We observed completeness down to $M_c = 1.6$ in the shallower parts of the slab and 2–2.6 in deeper parts of the slab. In depths greater than 200 km, the recompiled catalog lacks enough events to compute either $b$-values or completeness because it covers a shorter time period. The INGV catalog shows completeness in the range of 2.2–2.8
Figure 4.2: Cross-sections of the subducting slab. Top row based on the INGV catalog and bottom row on the recompiled catalog: (A) Spatially mapped $M_c$ of the INGV catalog (B) $b$-values of the INGV catalog (C) Error in $b$-values based on Shi and Bolt [1982] for the INGV catalog (D) Error in $b$-values for the INGV catalog using our bootstrap approach (E) Frequency-magnitude distribution for the marked volumes in frame B (F) Spatially mapped $M_c$ of the recompiled catalog (G) $b$-values of the recompiled catalog (H) Error in $b$-values based on Shi and Bolt [1982] for the recompiled catalog (I) Error in $b$-values for the recompiled catalog using our bootstrap approach (J) Frequency-magnitude distribution for the marked volumes in frame G. Hypocenters are marked with black crosses. Volcanoes are marked with red triangles (Seamounts have white face). Distance axis measures distance from point P1 (see Figure 4.1).
increasing with depth (Figure 4.2A).

We compared the errors based on the method of Shi and Bolt [1982] and our bootstrap approach that also considers $M_c$ uncertainties. The error-distribution obtained by the bootstrapping method shows differences when compared with the error of the maximum-likelihood estimation given by Shi and Bolt [1982]. For the INGV catalog we found errors in $b$-value in the range of 0.05–0.15 based on Shi and Bolt [1982] (Figure 4.2C). We observed higher errors of 0.2 only at a depth of 200 km. The recompiled catalog shows a very similar distribution and values of the same order. No significant difference was detected. We have to state that the errors in calculating $b$-values, given as the standard deviation according to Shi and Bolt [1982], are too small when using a small population of events with high $b$-values (especially in the INGV catalog), because uncertainties in $M_c$ and magnitude estimation are not considered. Therefore, the errors are strongly underestimated. In contrast, our bootstrap method shows much larger $b$-value errors. For the INGV catalog we obtain errors of 0.2–1.05. In the shallow and very active parts, we observed errors of about 0.2 units, increasing with depth and at maximum in the high $b$-value anomaly. For the INGV recompiled catalog we observe almost the same pattern but with smaller errors (0.2–0.8) due to smaller $b$-values overall. From these observations we conclude that the quality of the recompiled INGV catalog is higher. Not only are its errors in $b$-values smaller, but the $b$-values are also closer to the usually observed range.

To test for the stability of the results in the INGV catalog, we changed the width of the cross-section in 25 km steps up to 200 km and down to 50 km. We found that the $b$-value distribution does not change (see Figure 4.3D, 4.3E, and 4.3F). The anomaly is visible in all cross-sections and shows similar $b$-values. Only the coverage of the slab changes with width. When choosing widths smaller than 100 km, we cannot map the slab’s lower part due to the low number of events. Also, rotating the strike of the cross-section by 10° in both directions does not change the $b$-values and the anomaly (see Figure
Figure 4.3: $b$-value cross-sections of Tyrrhenian Sea computed for stability test. Extension and color scaling are identical to the $b$-value cross-sections in Figure 4.2. (A) Cross-section rotated $-10^\circ$ (B) Original cross-section as in Figure 4.2B for comparison (C) Cross-section rotated $+10^\circ$ (D) Cross-section with a width of 50 km (E) Cross-section with a width of 100 km (F) Cross-section with a width of 200 km.
4.5 Discussion and Conclusion

We found a high $b$-value anomaly beneath the Aeolian Islands and its eastern chain of seamounts. According to Wiemer and Benoit [1996], a high $b$-value anomaly indicates the location of dehydration of a slab and, therefore, we propose this anomaly to be a magma source of the volcanoes. High $b$-value anomalies have been found at different depths on top of subducting slabs, [Wiemer and Benoit, 1996; Wyss et al., 2001]. Wyss et al. [2001] correlated such an anomaly in the subduction zone beneath northeastern Japan with
low velocities in the mantle wedge and proposed a contorted magma pathway. Correlation with tomography results \cite{Cimini1999} in the Tyrrhenian Sea is currently not possible due to the larger scale of the velocity data there. Therefore, we could not find a velocity-based indication for a magma path.

In the previously mentioned studies quality of the datasets and regional activity level was higher. Therefore, data quality validation was an important issue in our work. Through careful computation of $M_c$, removal of older data and use of a uniform magnitude definition, we ensured the highest possible quality of data for $b$-value imaging. Additionally, we proposed new qualitative and quantitative methods to test for the stability and to assess the significance of our results. Applying these methods can exclude erroneous results caused by artifacts. Furthermore, bootstrap calculation of the $b$-value error takes into account the obvious imperfection of the catalog and prevents misinterpreting $b$-values due to underestimated errors. We consider these processing steps inherently necessary when dealing with catalogs of low to medium quality or low number of events. They enable us to get the most out of the available data without putting the result at risk.

\section*{4.6 Acknowledgments}

The authors would like to thank the INGV for supplying the catalogs for this study, and Dr. Cimini (INGV, Rome) for the tomography results. We also wish to thank an anonymous reviewer and C.-I. Trifu for their valuable comments. This work was financed by Istituto Nazionale di Geofisica e Vulcanologia in the framework of a collaboration agreement between INGV and Messina University, and related sub-contract between Messina University and ETH Zurich. This paper is contribution number 1295 of the Geophysical Institute, ETHZ.
Chapter 5

Earthquake-Size Distribution: A New Stressmeter?

Danijel Schorlemmer\textsuperscript{1}, Stefan Wiemer\textsuperscript{1}, and Max Wyss\textsuperscript{2}

\textsuperscript{1} Swiss Seismological Service, ETH Zürich, Zürich, Switzerland.
\textsuperscript{2} World Agency of Planetary Monitoring and Earthquake Risk Reduction, Genève, Switzerland.

Submitted to Nature.
5.1 Abstract

We demonstrate that the magnitude-frequency distribution (b-value) of earthquakes systematically varies for different styles of faulting. We universally find that normal faulting events have the highest b-values, thrust events the lowest, and strike-slip events show intermediate values. This requires that b acts as a stressmeter, depending inversely on differential stress. This is offering a framework that explains a range of observational data. Errors in seismic hazard assessment can result, unless the critical parameter b is correctly determined.

5.2 A New Stressmeter?

The earthquake size distribution follows, in most instances, a power-law [Ishimoto and Iida, 1939; Gutenberg and Richter, 1944], one of the most basic equations of seismology: \( \log N = a - bM \), where \( N \) is the cumulative number of events above a given magnitude \( M \), \( a \) is a constant, measuring the rate of occurrence, and \( b \) is a constant that describes the size distribution of events. A high b-value indicates a relatively larger proportion of small events, and vice versa.

Statistically significant variations of b-values have been measured in laboratory experiments [Mogi, 1962; Scholz, 1968; Amtrano, 2003], in underground mines [Urbancic et al., 1992], and in various tectonic regimes such as subducting slabs [Wyss et al., 2001], near magma chambers [Wiener et al., 1998], along fault zones [Amelung and King, 1997; Wiener and Wyss, 1997], and in aftershock zones [Wiener and Katsumata, 1999]. The interpretation of some of these differences as due to different stress regimes has been uncertain because other explanations were possible, and because it was questionable that samples in small volumes (0.1 m in laboratory specimens, 100 m in mines and 5 km in the Earth’s crust) are representative for earthquakes in general. Given the lack of physical understanding of these differences, the observation that
Table 5.1: Selection parameters of the catalogs.

<table>
<thead>
<tr>
<th>Network</th>
<th>Period</th>
<th>$M_c$</th>
<th>Depth [km]</th>
<th># of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEID F-Net</td>
<td>4.1.1997–24.3.2004</td>
<td>4.5</td>
<td>0–50</td>
<td>1354</td>
</tr>
<tr>
<td>NEID Kanto-Tokai</td>
<td>1.1.1982–1.7.2003</td>
<td>3.0</td>
<td>all (0–50)</td>
<td>5403 (2337)</td>
</tr>
</tbody>
</table>

$b = 1 = \text{const.}$, if large volumes are sampled \cite{FrohlichDavis1993}, was interpreted by some as indicating that $b = 1$ is a universal constant for earthquakes in general \cite{Kagan1999}.

To resolve the issue of influence of stress on $b$, we search several high quality, large data sets, including a global one, for systematic dependence of $b$ on the style of faulting. Because different faulting styles are uniquely associated with a stress regime\footnote{The stress regimes that lead to the three types of faulting are fundamentally different. Three principal stresses, $\sigma_1 \geq \sigma_2 \geq \sigma_3$, define the stress tensor $\sigma$ which leads to earthquake ruptures (Total stress $\sigma = \text{mean stress } \sigma_m + \text{deviatoric stress } \sigma'$). In all three typical modes, one of the principal stresses is oriented approximately vertically. In normal faulting, the greatest principal stress, $\sigma_1$, is oriented vertically, in strike-slip the vertical is $\sigma_2$, and in thrusting it is $\sigma_3$. The vertical stress, however, is approximately constant and equal to the weight of the rock column per unit area $\sigma_v = \rho gh$ ($\rho$ is density, $g$ is gravitational acceleration, $h$ is depth). This means, that the confining pressure, and also the differential stress $\Delta \sigma$ necessary for faulting \cite{Scholz1990}, is highest in the thrust, and lowest in the normal faulting regime, respectively. Strike-slip is an intermediate case.}, the differences in $b$ between regimes documented here allows us to unambiguously establish $b$ as stressmeter with an universal inverse dependence on differential stress $\Delta \sigma$.

To estimate $b$-values, we use the maximum-likelihood method \cite{Bender1983}, with uncertainties estimated by the method of Shi and Bolt \cite{ShiBolt1982}. For consistency in the analysis, we binned magnitudes in all catalogs to $\Delta M = 0.1$. Also, we selected periods of homogeneous recording \cite{HabermannCraig1988}, determined their overall completeness $M_c$ \cite{WiemerWyss2000} for each catalog, and cut the catalogs accordingly (Table 5.1). The source depths were restricted to 50 km, because deep events may obey different laws of faulting.

Earthquake focal mechanisms (faulting styles) are now determined rou-
tinely for thousands of events. For this study, we use the best sources of focal mechanism available worldwide: The global Harvard CMT moment tensor catalog, a catalog of recently relocated events of southern California (SCSN) [Hauksson, 2000], a northern California dataset (NCSN), a dataset of events from the Kanto-Tokai region of Japan, and the F-Net data for all of Japan. From the Kanto-Tokai catalog we excluded volcanic and geothermal areas.

The two quality descriptors available for focal mechanisms in the Californian catalogs were used to limit the data set to the high quality events: The solution misfit and station distribution ratio (STDR) [Reasenberg and Oppenheimer, 1985]. In the NEID Kanto-Tokai catalog, we computed the misfit value from the ratio of the number of stations whose polarity does not match the solution to those whose polarity is reported and used only events with a misfit \( \leq 0.2 \) and STDR \( \geq 0.5 \).

For classifying events as strike-slip, thrust, or normal events, we use the rake angle \( \lambda \) with a given range of \( \pm \gamma \) (Aki-Richards convention: \( \lambda = 0 \) or \( \lambda = \pm 180^\circ \) as strike-slip, \( \lambda = 90^\circ \) as thrust, and \( \lambda = -90^\circ \) as normal events). The Californian data contain the angles of only the preferred nodal plane. In catalogs where rake angles of both nodal planes are present, both \( \lambda \) have to fall within the given range.

The \( b \)-values for \( N \geq 100 \) events were computed as a function of rake \( \lambda \), using a fixed range of \( \gamma = 20^\circ \) and a moving window with step size of \( 5^\circ \) (Fig. 5.1A). The SCSN catalog is probably the highest quality dataset of focal mechanisms available worldwide, because of the high density of stations, the relocation of events and re-computation of take-off angles using a three dimensional velocity model. These data show that the differences in \( b \)-values for the three classes of events are about a factor of 10 larger than the errors of individual samples (Fig. 5.1A). Normal events (green) show the highest \( b \)-values \( (b_{NR} \approx 1.1) \), strike-slip events (red) show intermediate values \( (b_{SS} \approx \)

2Polygon: 136.76E/35.73N, 137.70E/35.72N, 139.47E/35.72N, 139.47E/36.48N, 140.40E/37.03N, 141.39E/36.67N, 140.91E/35.72N, 140.03E/34.93N, 139.27E/34.79N, 138.47E/34.61N, 137.05E/34.58N, 136.64E/35.23N, 136.76E/35.73N.
Figure 5.1: (A) $b$-value as a function of the rake $\lambda$ of events in southern California. The vertical bars indicate the standard error [Shi and Bolt, 1982]. The green, red, and blue lines mark the $b$-values of mainly normal, strike-slip, and thrust events, respectively. The gray line marks the average $b$-value. The circles at the top of the frame show the rake $\lambda$ used for computing the $b$-values ($\lambda = -90^{\circ} \pm \gamma$, $\lambda = 0^{\circ} \pm \gamma$, and $\lambda = 90^{\circ} \pm \gamma, \gamma = 20^{\circ}$). (B) $b$-value as a function of the range $\gamma$ of the rake $\lambda$ for three classes of events in southern California. The vertical bars indicate the standard error [Shi and Bolt, 1982]. Solid bars are used for samples with $N \geq 200$, dashed bars for samples with $200 > N \geq 100$. The green, red, and blue lines mark the $b$-values of normal, strike-slip, and thrust events, respectively. The gray line marks the average $b$-value. The circles at the top of the frame show the range of rake $\lambda$ used for computing the $b$-values ($\gamma = 15^{\circ}, 45^{\circ}, 75^{\circ}$). (Inset B) Circle explaining the rake values and the corresponding colors of the classes of events. (C) Frequency-magnitude distributions for pure normal (green) and pure thrust (blue) events of the SCSN and Harvard catalogs. (D) Same as (A) for the Harvard catalog. (E) Same as (B) for the Harvard catalog (Only considering rake $\lambda$ of the first nodal plane). (Inset E) Same as (E) but considering rakes of both nodal planes.

0.9), and thrust events (blue) the lowest ($b_{TH} \approx 0.7$). The transition between the average $b$-value (strike-slip) and the extreme $b$-values (normal, thrust) is pronounced.

To further explore the separation into classes, we plotted $b$-values as a function of the range, $\gamma$, used to define a class (Fig. 5.1B). The $\gamma$-value was varied from $175^{\circ}$ in steps of $10^{\circ}$ down to $5^{\circ}$. For $\gamma \geq 45^{\circ}$, events may fall into two or more classes. One can see that, for ranges of $\gamma \geq 90^{\circ}$, almost no differences exist. With smaller $\gamma$, $b$-values of normal and thrust events start to deviate systematically from the regional average and reach their extreme for pure event classes with range $\gamma \leq 5^{\circ}$. Strike-slip events show almost no deviation from the regional average $b$-value, regardless of the range $\gamma$. The differences in the frequency-magnitude distribution between pure thrust and
pure normal events is statistically highly significant and clear to the eye (Fig. 5.1C).

The $b$-value variations of the Harvard catalog, computed with the same parameters as for the SCSN catalog, are less pronounced; however, they show the same dependence on faulting style (Fig. 5.1D and 5.1E). The separation is clearer when using only the first nodal plane for classifying events (Fig. 5.1E) rather than both nodal planes (Inset in Fig. 5.1E). Although each catalog covers a different range of magnitudes, the frequency-magnitude distributions emphasize the similarity of the $b$-values for the two classes of events (Fig. 5.1C).

The data in all three additional catalogs show essentially the same systematic dependence of $b$-value on faulting style (Fig. 5.2). The absolute level of $b$ does not concern us here, because it can vary in different regions. The important observation is that in all cases, normal events exhibit the highest $b$-values, followed by strike-slip and thrust events. In some data sets, the separation is no longer clear for the narrowest ranges of $\lambda$, because the error bars increase due to small sample sizes.

The differences between the data sets are: (1) The magnitudes of earthquakes available are smallest in the three local catalogs ($M_c = 2.5$ in SCSN and $M_c = 3$ in NCSN and NEID), intermediate in the regional catalog ($M_c = 4.5$ in F-Net), and largest in the global catalog ($M_c = 5.5$ in Harvard CMT). (2) The depths of events in California are concentrated in the top 20 km (95%), whereas the other data sets contain substantial percentages of events between 20 and 50 km. (3) The tectonic setting covered by the data sets from California and Japan are essentially restricted to a strike-slip and a convergent plate boundary, respectively. In the CMT data set, all types of tectonic settings are represented. Nevertheless, all data sets show the same dependence of $b$ on focal mechanism.

The uniformity of the separation of $b$-values according to faulting type in these tectonically different data sets requires a universal interpretation. We
Figure 5.2: $b$-value as a function of the range $\gamma$ of the rake $\lambda$ for three classes of events. The vertical bars indicate the standard error [Shi and Bolt, 1982]. Solid bars are used for samples with $N \geq 200$, dashed bars for samples with $200 > N \geq 100$. The green, red, and blue lines mark the $b$-values of normal, strike-slip, and thrust events, respectively. The gray lines mark the average $b$-values. (A) NCSN catalog. (B) NEID Kanto-Tokai catalog. Inset shows $b$-values computed without depth constrains. (C) NEID F-Net catalog.
propose that the differential stress, and indirectly the confining pressure (to which the differential stress is tied), is the parameter most strongly controlling faulting types, thus influencing differences in $b$. This inverse relationship follows because we show that $b_{TH} < b_{SS} < b_{NR}$, and it is known that for a given vertical stress $\sigma_v$ the mean stresses obey the relationship $\sigma_{TH} > \sigma_{SS} > \sigma_{NR}$ (see footnote 1). Thus, we consider the inverse relationship between differential stress and $b$-value as established. It covers all magnitude ranges available for focal mechanism analysis and matches observations of different orders of magnitude in size [Scholz, 1968; Urbancic et al., 1992; Amitrano, 2003]. Therefore, the range extends from submillimeters to hundreds of kilometers of rupture length. This does not exclude that, locally, secondary effects can come from material properties [Mogi, 1962; Scholz, 1968] and that the stress tensor may be modified by pore pressure [Wyss, 1973].

The idea that depths may be the controlling factor, in general, can be ruled out because the average depths of the three classes of events are not systematically ordered according to the $b$-values. Depth is also not a good candidate to be a universal factor controlling $b$, because the trends reported are opposite in California [Mori and Abercrombie, 1997; Gerstenberger et al., 2001] and Japan [Wyss and Matsumura, 2002].

The degree of material heterogeneity [Mogi, 1962] as a fundamental cause of these universal differences can also be ruled out, because the observation is valid for small as well as larger earthquakes, and for all existing tectonic settings.

Other observations that are explained by, or support the inverse relationship between stress and $b$-value, are: In laboratory rock specimens the inverse relationship between differential stress (and confining pressure) is well established [Scholz, 1968; Amitrano, 2003]. For underground mines, there is one data set that shows the same relationship [Urbancic et al., 1992]. The interpretation that increased pore pressure (which decreases the differential stress) leads to increased $b$-values is compelling in at least two data sets [Wyss, 1973;...
Increased temperature is likely to affect the rheology locally, and hence may affect the nature of failure. However, this parameter is poorly known and it is an open question, how much it may locally modify the $b$-value. For the Los Angeles Basin, Hauksson [1990] found that $b_{\text{SS}} = 1.1$ and $b_{\text{TH}} = 0.7$ for strike-slip and thrust events, respectively. Although this result was based on only 144 and 78 earthquakes, respectively, it supports our findings. Finally, the strong differences in $b$-value between locked parts of faults, where $b$-values are low, and creeping sections with much higher $b$ [Amelung and King, 1997; Wiemer and Wyss, 1997; Schorlemmer et al., 2004], are consistent with the model. In a locked section, the shear strength is assumed to be higher than in creeping sections. This results in high $\Delta \sigma$ and low $b$-values for locked sections, while creeping sections exhibit low $\Delta \sigma$ and high $b$-values. This conceptional model relating between $b$-values and fault locking was recently tested and supported by the 2004, M6.0 Parkfield, California event [Schorlemmer and Wiemer, in press], which almost exclusively ruptured areas of the San Andreas fault previously mapped as low $b$-value regions.

The physical model offered by Scholz [1968] as an explanation for the stress dependence of $b$ emphasizes the role of differential stress. He argues that in a rock mass with varying resistance to faulting (and varying stress concentration) a rupture is likely to connect one high-stress sub-volume to the next, because the system contains more energy. Thus, earthquakes grow larger in a high-stress environment. This finding is additionally supported by Huang and Turcotte [1988] in their fractal simulations. Amitrano [2003] proposed a more geometrical model, emphasizing the role of confining pressure. Based on a numerical model, he argues that, at low confining pressure, the internal friction angle is such that failure in a given element can influence other elements only in restricted directions. At high confining pressure, the angle of internal friction decreases and interaction with elements containing defects are possible in all directions. Thus, failures grow larger in a system under high confining pressure.
The proposed dependence of $b$ on differential stress predicts a systematic decrease of the $b$-value toward the end of the seismic cycle. However, the amplitude of the differential stress increase during the loading cycle is probably an order of magnitude smaller than in laboratory experiments and in underground mines. The periods of high quality data sets, necessary to test this hypothesis [Smith, 1981; Lei et al., 2004], are not yet available.

One of the implications of the universal dependence of $b$ on differential stress is that earthquake hazard analysis should be rethought. The earthquake size distribution is a critical parameter in studies of seismic hazard. Because events of different faulting styles produce significantly different ground motions [Oglesby et al., 2000], estimating the recurrence for each faulting style separately should be attempted if data density allows. In addition, hazard studies should pay attention to spatial heterogeneity in $b$-values, because it is directly related to the distribution differential stresses in the Earth’s crust.

### 5.3 Acknowledgments

The authors would like to thank the Northern California Seismic Network, U.S. Geological Survey, Menlo Park and the Berkeley Seismological Laboratory, University of California, Berkeley for the catalog including phase data, the National Research Institute for Earth Science and Disaster Prevention for mechanism solutions of the Kanto-Tokai area. The authors profited from discussions with E. Hauksson, D. Giardini, M. Mai, M. Gerstenberger, D. Jackson, J. Woessner, and G. Hillers. This paper is contribution number XXXX of the Geophysical Institute, ETHZ.
Chapter 6

Earthquake Statistics at Parkfield I: Stationarity of $b$-Values

Danijel Schorlemmer$^1$, Stefan Wiemer$^1$, and Max Wyss$^2$

$^1$ Swiss Seismological Service, ETH Zürich, Zürich, Switzerland.
$^2$ World Agency of Planetary Monitoring and Earthquake Risk Reduction, Geneva, Switzerland.

6.1 Abstract

In this Paper I, we quantitatively show that the heterogeneous pattern of \( b \)-values (of the Gutenberg-Richter relation) in the Parkfield segment of the San Andreas fault is to a high degree stationary for the past 35 years. This prepares the grounds for Paper II, where we test the hypothesis that our model of spatially varying \( b \)-values forecasts future seismicity more accurately than the approach, in which one assumes a constant \( b \)-value equal to the average regional value. The method we develop to measure stationarity in the presence of spatial heterogeneity consists of the following steps. (1) Determine the optimal dimensions of the sampling volume by mapping \( b \)-values with a wide range of radii and selecting the largest radius that gives the most detailed resolution of the \( b \)-value heterogeneity. Along the selected fault segment, the high data density permits the definition of the dominant dimensions of the seismotectonic fabric, which is about 8–10 km. (2) Map the difference in \( b \)-value between two periods, selecting numerous possible catalog divisions. (3) Identify significant changes of \( b \)-values by the Utsu test [Utsu, 1992]. Along the studied fault segment of 110 km length, only one patch of radius 5 km showed a significant increase in \( b \), from below average to above, as a function of time. This change in \( b \) initiates around 1993 and thus correlates in space and time with a well documented episode of creep at depth. Using the derived spatial variable \( b \)-value distributions, we find that the highest probability for earthquakes with magnitude \( M \geq 6 \) is in the Middle Mountain asperity, where the 1966 Parkfield earthquake nucleated and where all \( M \geq 4.5 \) events in the dataset occurred. In contrast, if only the regional average \( b \)-value of 0.92 is used to predict future seismicity, the creeping segment north of Parkfield should produce major earthquakes most frequently, a conclusion that contradicts the observations.
6.2 Introduction

Probabilistic forecasting of earthquakes attempts to deliver the most accurate estimate of future seismicity at a given location and for a given magnitude range and period. In most cases, this task is attempted by sampling the observed seismicity of the past and extrapolating it into the future, using the assumption that the frequency-magnitude distribution of earthquakes can be approximated with a power law \[ \text{log } N = a - bM \] (log denotes the common logarithm), where \( N \) is the cumulative number of events of magnitude \( M \) or greater, while \( a \) and \( b \) are constants. This relationship has been investigated extensively (for an overview, see [Utsu, 1999] and [Wiemer and Wyss, 2002]) and works well, in most cases, as a first order approximation. It is also the backbone of probabilistic seismic hazard assessment [Cornell, 1968]. While this forecasting is simple, and has been used in numerous statistical seismology and hazard related studies, there remain, in our opinion, a number of fundamental unresolved issues related to spatial heterogeneity and stationarity of \( b \)-values.

The basic questions that we address in this paper can be stated as follows: (1) For which volume should one determine \( a \)- and \( b \)-values to obtain a robust and accurate forecast? Is it better to average over large spatial volumes or should one use small volumes to take into account heterogeneity in some detail? (2) How can the assumption that the frequency-magnitude distribution is stationary be verified?

The question of the spatial heterogeneity in seismicity parameters is intrinsically linked to the question of stationarity in these parameters: If temporal changes in seismicity parameters occur, then larger sample volumes may average out the fluctuations and lead to a reasonable forecast. Spatial heterogeneity in seismicity parameters is by now a well established fact. It has long been obvious that seismicity rates, or productivity (\( a \)-value) varies strongly as a function of space and time—any map of seismicity reveals this. Only in the
past 10 years, approximately, has it been firmly established that the $b$-value, or earthquake size distribution, varies also considerably and significantly as a function of space and, possibly to a lesser effect, of time, e.g. [Gerstenberger et al., 2001; Wiemer and Wyss, 2002]. One clear example of spatial heterogeneity of $b$-values stems from the Parkfield section of the San Andreas fault [Wiemer and Wyss, 1997]. Volcanic and geothermal regions also show changes in $b$-values over length scales of 1 km or less. Significant spatial heterogeneity of $b$-values can be shown to exist on length scales of millimeters in laboratory samples [Amitrano, 2003], to thousands of kilometers as in the case of the India-Asia collision zone [Wiemer and Wyss, 2002].

The issue of how to best resolve spatial heterogeneity in seismicity parameters, particularly in $b$-value, has two facets. First of all, it is a question of resolution of the available data, which is primarily determined by the sample size. The uncertainty in estimating model parameters increases with decreasing sample sizes, resulting in a trade-off between accuracy and resolution. Given a global dataset of $M > 5$ events one can hardly expect to resolve variations in $b$-values of less than tens of kilometers, because in order to establish a reliable $b$-value, one needs a sample size of roughly 50–100 events. The second issue has to do with scaling of earthquake properties: Different size sampling volumes may well measure different intrinsic scaling length in the Earth. A large scale continental collision zone and its resulting stress field leaves a large scale imprint on the frequency magnitude distribution, at least of larger events [Wiemer and Wyss, 2002]. Nested within this regional behavior could be any number of smaller scale heterogeneities that reflect local changes in the earthquake size distribution, for example in the vicinity of volcanic or geothermal systems. However, it is also observed that a given area with seismic activity is characterized by a certain size of heterogeneities, due to the seismotectonic fabric [Wyss et al., 2000].

The question of the spatial heterogeneity in $b$-values is closely related to hazard estimates. Even contemporary hazard mapping projects differ in their
approach between assuming a constant $b$-value, or a spatially varying one. It also relates to understanding the underlying physics of the system.

We choose the Parkfield section of the San Andreas fault to address these questions, because it is a simple fault setup with limited interaction between neighboring fault strands, because it has been extensively studied with various geophysical techniques [Roeloffs and Langbein, 1994; Gwyther et al., 1996; Gao et al., 2000; Roeloffs, 2000, 2001; Murray and Segall, 2002] and because it has been monitored with a dense network of seismographic stations for more than 30 years. In addition, we know that here strong spatial differences in $b$-values have been documented in 1997, allowing a fully prospective test of the hypothesis of spatially varying $b$-values that are stationary. There is a final important aspect to Parkfield: To convince ourselves that the spatial heterogeneity in $b$-values are not related to artifacts in catalog properties, we can compare the observed patterns with an independent data set of moment magnitudes derived from a network of borehole stations near Parkfield [Karageorgi et al., 1992].

### 6.3 Data

We use the NCSN catalog from the Northern California Earthquake Data Center (NCEDC) spanning the period 1967–2003. From this catalog we separated the events of the Parkfield region, defined as a 5 km wide cross-section extending from $P_1$: $121^\circ W$, $36.4^\circ N$ to $P_2$: $120.2^\circ W$, $35.64^\circ N$ (Figure 6.1). This cross-section has a length of about 110 km and includes the asperity, the locked segment, and a large portion of the creeping section. We cut the catalog at a depth of 16 km, thereby excluding only a few events that may well be erroneous locations.

Several processing steps have been performed to ensure data quality and selection of a reliable dataset. In the given dataset, 451 out of 10673 events are of magnitude $M = 0$, indicating events without assigned magnitudes. We removed these events from the catalog, thereby deleting the years 1967 and
Figure 6.1: Seismicity map of the Parkfield region in central California. Gray dots mark the epicenters of the events with $M \geq 1.3$ and depth $D \leq 16$ km from 1981–2003, red lines mark mapped faults. The investigation area (Parkfield segment) along the San Andreas fault is marked by a black rectangle. The cross-section extents from $P_1$ to $P_2$ with a width of 5 km. The small red box marks the small Middle Mountain alert box, the medium-sized red box marks the Middle Mountain alert box, and the large red box marks the Parkfield alert box [Michael and Jones, 1998].
1968 which consist only of events without magnitude information. We also re-binned the catalog from magnitude bins with $\Delta M = 0.01$ into new magnitude bins with $\Delta M = 0.1$. This step is necessary because computing magnitude of completeness is based on the non-cumulative frequency-magnitude distribution [Wiemer and Wyss, 2000]. With finer binning, deviations from the assumed Gutenberg-Richter distribution become larger, therefore affecting the computation of magnitude of completeness $M_c$, possibly resulting in its over-estimation.

We investigated whether the dataset is contaminated with quarry blasts by mapping the daytime to nighttime activity ratio as described by Wiemer and Baer [2000]. We have not found any evidence for contamination of the dataset in the selected region.

To identify periods of different recording quality and catalog completeness, we investigated changes in the slope of the plot of cumulative number of events per time, [Habermann and Craig, 1988]. Plotting the cumulative number of all events in the catalog shows only minor changes in the slope at 1981 (see Figure 6.2B, solid line), but when taking only events with magnitudes $M \leq 1.1$, we can clearly identify periods of lower recording quality (see Figure 6.2B, dashed line). The kink in the slope at 1981 indicates the well-known improvement in recording quality at 1981. From 1995 to 2000, recording completeness dropped significantly for small events but was restored in 2000 to the same level as in the period before 1995. When plotting cumulative events with magnitudes $M \leq 1.5$ over time, we can detect only a slight change of the slope in the period from 1995 to 2000. This indicates an overall completeness level of $M_c \leq 1.5$ for the catalog in the period from 1981 to 2003.

We applied the GENAS algorithm for detecting magnitude dependent rate changes, [Habermann, 1983]. We found the aforementioned drop in detected events in 1995.06 and increase in 2000.67 (decimal years). Although additional increases and decreases are visible in the GENAS result, those cannot be related to changes in recording quality and are rather reflecting natural rate
changes in seismic activity.

We also compared rates per magnitude for the two periods 1990–1995 and 1995–2000, as described by Zúñiga and Wiemer [1999]. For the first period we detect an overall completeness of $M_c = 1.1 \pm 0.1$, while for the second period the completeness is higher with $M_c = 1.3 \pm 0.1$ (see Figure 6.3). This change in completeness can hardly be seen in the cumulative frequency-magnitude distribution (Figure 6.3A). In the non-cumulative distribution (Figure 6.3B) however, one can clearly see a higher $M_c$ for the period 1995–2000 as a drop in rates for magnitudes $M < 1.3$. In the earlier period, the rates drop for events with magnitudes $M < 1.1$. This again shows the loss in completeness in the period 1995–2000.

Computing the magnitude of completeness $M_c$ as a function of time [Wiemer and Wyss, 2000] using sample sizes of 500 events with a step size of 10 events reveals gradual improvements in completeness from $M_c \approx 1.7$ in the year 1970 to $M_c \approx 1$ in the year 1990 (see Figure 6.2A, gray line). In the years 1995 to 2000, we can detect $M_c \approx 1.3$, corresponding to the different slope in the cumulative number curve for events with $M \leq 1.1$. From 2000 onwards, com-
6.3 Data

Figure 6.3: (A) Cumulative frequency-magnitude distributions for the periods 1990–1995 (dashed line) and 1995–2000 (solid line). (B) Non-cumulative frequency-magnitude distributions for the periods 1990–1995 (dashed line) and 1995–2000 (solid line).

Completeness improves again to $M_c \approx 1.2$. The windowed average (50 data points) smoothes the changes in $M_c$ (see Figure 6.2A, black line) but also shows the aforementioned development of $M_c$.

In summary, we use as our primary data source the period from 1981–2003. We spatially map completeness along a cross-section for this period in order to confirm that $M_c$ does not show strong spatial variability. Having found no strong spatial variability of $M_c$, we assume for simplicity and stability reasons a spatially homogeneous completeness of $M_c = 1.3$. This leaves a total number of 3780 events in our primary catalog. This catalog has an overall $b$-value of $\bar{b} = 0.92$. As a secondary dataset, we also analyse the period 1967–1981, cut at a higher overall completeness of $M_c = 1.7$.

When bootstrapping to obtain the standard deviation of $b$-values we have to compute $M_c$ at each node for taking into account uncertainties in $M_c$. For this task, we created an additional catalog cut at magnitude $M = 0.8$ instead of $M = 1.3$. We have not found $M_c$-values lower than 0.9 in our spatial mapping. By cutting the catalog at $M = 0.8$ we allow $M_c$ to drop to 0.8 in the bootstrapped samples.
For the Parkfield region, an independent dataset exists that allows verification of the results obtained with the NCSN catalog. This dataset was obtained by the High-Resolution Seismic Network (HRSN), established in 1986 [Karakostas et al., 1992], and is superior in location accuracy and extends to smaller magnitudes because it is derived from borehole seismometers installed as part of the Parkfield earthquake prediction experiment [Bakun and Lindh, 1985; Malin et al., 1989; Roeloffs and Langbein, 1994; Nadeau and McEvilly, 1997, 1999]. However, this catalog spans only the period 1987–1998.5 and covers only a short stretch of the San Andreas fault near Parkfield. The catalog provides moments which have been converted to magnitudes with 3 significant digits by Wyss et al. [2004]. We rebinned this catalog to magnitude bins of $\Delta M = 0.1$. We applied the aforementioned quality analysis to this catalog also. We found no indication of artificially introduced rate changes or magnitude shifts. Because of its limited spatial and temporal extent, we are using this catalog only for comparing $b$-values with the NCSN catalog.

6.4 Method

To investigate the heterogeneity of $b$-values along the Parkfield segment of the San Andreas fault, we map in cross-section $a$- and $b$-values along the fault segment shown in Figure 6.1. We are applying the gridding technique [Wiemer and Wyss, 2002] using the software package ZMAP [Wiemer, 2001]. We compute maximum-likelihood $b$-values using the equation

$$b = \frac{1}{\bar{M} - M_{\text{min}}} \log e$$

[Utsu, 1965; Aki, 1965; Bender, 1983]. $\bar{M}$ denotes the mean magnitude and $M_{\text{min}}$ the minimum magnitude of the given sample. The sample is considered complete down to the minimum magnitude $M_{\text{min}}$. $M_c$ has to be corrected by $\Delta M/2$ to compensate the bias of rounded magnitudes to the nearest $\Delta M$ bin,
thus $M_{\text{min}} = M_c - \Delta M/2$ [Utsu, 1965; Guo and Ogata, 1997]. The confidence limit of this estimation is given by

$$\sigma(b) = 2.30 b^2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} (M_i - \overline{M})^2 / n(n-1)} \quad (6.2)$$

[Shi and Bolt, 1982], where $n$ is the total number of events of the given sample. This $\sigma(b)$ tends to underestimate the true standard deviation of the $b$-value due to its assumption of a complete catalog and a correctly determined magnitude of completeness $M_c$. Therefore, we compute the standard deviation of $b$-values using the bootstrap approach as described by Schorlemmer et al. [2003].

For computing $b$-values, the knowledge of completeness of a sample is important. $M_c$ has to be computed either for every sample or defined assuming a homogeneous recording quality for the entire volume. Here we assume a homogeneous overall $M_c = 1.3$ and do not compute $M_c$ at each node. However, we confirm that our results are not dependent on this choice.

For sampling of earthquakes, we use cylindrical volumes perpendicular to the cross-sectional plane and centered at the nodes spaced at $0.5 \text{ km} \times 0.5 \text{ km}$. The length of these cylindrical volumes is defined by the width of the volume along the cross-section, which is $5 \text{ km}$. In each sample, we also require a minimum number of events with $M \geq M_c$, $N_{\text{min}}$, in order to determine a reliable $b$-value. For samples containing fewer than $N_{\text{min}}$ events, we do not compute $b$-values. Here, we arbitrarily set $N_{\text{min}} = 50$, because below this value the uncertainty in $b$ increases rapidly. To estimate the largest radius for sampling, which is not obscuring the $b$-value contrasts and anomalies, therefore providing the highest possible coverage, we computed $b$-value cross-sections using different radii, varying them between $r = 2 \text{ km}$ and $r = 20 \text{ km}$.

Based on $a$- and $b$-values, one can compute the probabilistic recurrence time $T_r$ for an earthquake with magnitude equal or greater than a chosen $M'$,
given as

\[ T_r = \frac{\Delta T}{10^{a-bM'}} \]  

(6.3)

Here \( \Delta T \) denotes the length of the recording period from which the \( a \)-value is derived. When spatially mapping the probabilistic recurrence time \( T_r \), we denote \( T_r \) at each node as 'local recurrence time' \( T_L \). Giving the rate \( \lambda \) for events of magnitude \( M' \) as \( \lambda = 1/T_r \), we can compute the probability \( P_r \) of occurrence of one or more events of magnitude \( M' \) using the cumulative Poissonian distribution:

\[ P_r(\lambda) = 1 - e^{-\lambda} \sum_{i=0}^{\lambda} \frac{\lambda^i}{i!} = 1 - e^{-\lambda} \]  

(6.4)

For testing stationarity of \( b \)-values, we divide the catalog in two abutting periods at different divisions in time and spatially map \( b \)-values for both periods. If in both periods the sample size is greater or equal \( N_{\text{min}} \) and thus \( b \) can be computed, we also compute the probability \( P_b \) of the hypothesis that the \( b \)-values of the two periods are coming from the same population, i.e. exhibit stationary behavior. This probability \( P_b \) is derived from the Akaike Information Criterion (AIC) [Akaike, 1974]. Comparing the AIC\(_0\) for both periods having the same \( b \)-value \( b_0 \) and the AIC\(_{12}\) for both periods having two different \( b \)-values \( b_1 \) and \( b_2 \) leads to the difference \( \Delta \text{AIC} \) of these two AIC scores as given by Utsu [1992].

\[ \Delta \text{AIC} = -2(N_1 + N_2) \ln(N_1 + N_2) + 2N_1 \ln(N_1 + N_2b_1/b_2) + 2N_2 \ln(N_1b_2/b_1 + N_2) - 2 \]  

(6.5)

The probability \( P_b \) that the \( b \)-values are not different is given by

\[ P_b = e^{-\frac{\Delta \text{AIC}}{2}} \]  

(6.6)
Following Utsu [1999], we consider the difference in $b$-values not significant if $\Delta AIC < 2$. If $\Delta AIC > 2$, the difference is significant. $\Delta AIC = 2$ corresponds to $P_b \approx 0.05$. The difference is considered highly significant if $\Delta AIC > 5$. This value corresponds to $P_b \approx 0.01$. Applying the logarithm leads to log-probabilities of $\log P_b \leq -1.3$ for significantly different $b$-values and $\log P_b \leq -1.9$ for highly significant differences in $b$-values.

\section*{6.5 Results}

\subsection*{6.5.1 Mapping $b$-values}

To find the appropriate radius for resolving $b$-value contrasts at Parkfield, we mapped $b$-values with sampling radii varying between 2 km and 20 km (Figure 6.4). Sampling with radii from 2 km to 5 km (frames A-D) show essentially the same pattern but with different coverage. The smaller the radius, the fewer volumes match the requirement of at least $N_{min} = 50$ events for computing $b$-values. Radii smaller than 5 km resolve additional detail in only few locations. At the southern end of the creeping section (at a distance of 50–60 km from $P_1$) we see that the samples at a depth of approximately 4 km lose their high $b$-values when computed with a radius of 5 km compared to 2 km. When sampling these shallow volumes with small radii, events with small magnitudes are predominant, increasing the $b$-value. Sampling with larger radii tends to mix the shallow volumes with deeper ones and results in average $b$-values.

The observation of a nearly identical pattern of $b$-values when sampled with radii from 2 km to 5 km suggests that using smaller radii than 5 km is not revealing details which are obscured when sampling with larger radii, but it only reduces coverage. The selection of the optimal radius for Parkfield is not applicable to other areas because it depends on the local seismotectonic fabric and data availability.

Sampling with radii greater than 5 km (frames E-I) results in smoothed
<table>
<thead>
<tr>
<th>Distance from P</th>
<th>Depth [km]</th>
<th>( b )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>20 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>30 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>40 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>50 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>60 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>70 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>80 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>90 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>100 km</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>110 km</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Figure 6.4:** Distribution of \( b \)-values in the Parkfield segment of the San Andreas fault computed with the NCSN catalog from 1981–2003 using different radii \( r \) and \( N_{\text{min}} = 50 \). (A) \( r = 2 \) km. (B) \( r = 3 \) km. (C) \( r = 4 \) km. (D) \( r = 5 \) km. (E) \( r = 6 \) km. (F) \( r = 7 \) km. (G) \( r = 8 \) km. (H) \( r = 10 \) km. (I) \( r = 20 \) km.
6.5 Results

$b$-values. Using a radius of $r \geq 10$ km is obscuring any $b$-value contrast at Parkfield. We can also see that the low $b$-value zone (at a distance of 70 km from $P_1$) loses its crispness with sampling radii of $r \geq 6$ km and moves towards southeast (to the right). This is because the small volume containing the information about the very low $b$-value is located at the edge of the active volume. Earthquakes have not been recorded below or in the southeastern vicinity of it. Therefore, with increasing radii the center of the cylindrical volumes has to move southeast to remain dominated by the low $b$-value distribution. At the locus of the $b$-value anomaly, the sample becomes a mixture of different distributions when increasing the radius and the anomaly vanishes.

Thus, the optimal radius for sampling events to map $b$-values at the Parkfield segment is in the range of 4–5 km. Sampling with smaller radii reduces the coverage while sampling with larger radii obscures the anomalies and contrasts. For this study, we decided to sample events at the Parkfield segment using a radius of $r = 5$ km.

Figure 6.5B shows the high resolution mapping of $b$-values along the Parkfield segment of the San Andreas fault based on the NCSN catalog of the period 1981–2003. At every node (grid spacing: 0.5 km × 0.5 km) we computed the $b$-value, using all events within a cylindrical volume of radius $r = 5$ km. If a volume contained less than the minimum number of $N_{min} = 50$ events, no $b$-value was computed.

The $b$-value distribution (Figure 6.5B) shows strong spatial variations. Different anomalous patches along this segment can be distinguished: At a distance from point $P_1$ of about 70 km, the Middle Mountain asperity, the rupture initiation point of the 1966 Parkfield event, correlates with a patch of very low $b$-value ($b \approx 0.5$). The low $b$-value zone extends from the asperity about 25 km southeastwards and matches the locked part of the fault (Figure 6.5).

The southern end of the creeping zone (distance from point $P_1$: 30–60 km) shows high $b$-values. These $b$-values range from $b \approx 1$ up to $b \approx 2$. At a distance from point $P_1$ of 15 km we can identify a volume with high $b$-values
Figure 6.5: Distribution of $b$-values in the Parkfield segment of the San Andreas fault. The circles mark volumes for which frequency-magnitude distributions are shown in Figure 6.6. The radii correspond to the sampling radii. The red circles mark the asperity (distance from $P_1$: 70 km, depth: 10 km), the black circles mark a part of the creeping section (distance from $P_1$: 63.5 km, depth: 3 km). The bars at the top mark the extension of the creeping section, Middle Mountain asperity, and the locked part of the fault. (A) Seismicity distribution of the years 1981–2003. All earthquakes with $M \geq 1$ and depth $D \leq 16$ km are plotted. (B) $b$-values of the NCSN catalog from 1981–2003. (C) $b$-values of the HRSN catalog from 1987–1998.5. (D) $b$-values of the NCSN catalog from 1969–1981.
extending to a depth of 10 km. In accordance with Gerstenberger et al. [2001]
and Wiemer and Wyss [1997], shallower parts show in general higher \( b \)-values
while deeper parts exhibit lower \( b \)-values.

The independently derived HRSN catalog shows approximately the same
pattern (Figure 6.5C), although only for a shorter stretch of the fault. In
the part covered by the HRSN data, the Middle Mountain asperity again
correlates with low \( b \)-values (\( b \approx 0.5 \)). The \( b \)-values in the barely covered
creeping section show similar values as in the NCSN catalog (\( b \geq 1.1 \)). To
take advantage of the superior locations and considering the fact that this
catalog contains more earthquakes per unit volume, we mapped \( b \)-values using
radii of \( r = 3 \) km. Using the same radius \( r = 5 \) km as in the NCSN catalog
smoothes the \( b \)-value contrast. We found strong variations in completeness,
ranging from \( M_c = 0.4 \) to \( M_c = 0.9 \), therefore, we computed \( M_c \) in this case
for every node separately. We also constrained the magnitude of completeness
to the range of \( M_c \in [0.4, 0.9] \) to compensate for occasionally instability of the
\( M_c \)-computation algorithm.

As a third dataset, we analyse the NCSN catalog for the period 1967–1981.
We cut the catalog at the overall completeness level of \( M_c = 1.7 \) (see Figure
6.2A) and again use a spatially homogeneous \( M_c \). The resulting \( b \)-value map
(Figure 6.5D) shows the same general pattern with almost the same absolute \( b \)-
values. Because of the higher \( M_c \) threshold and the shorter period, we actually
resolve fewer nodes with \( N_{\text{min}} \geq 50 \) than in the 1981–2003 period.

The frequency-magnitude distributions based on these three catalogs (Fig-
ure 6.6) for the asperity and southern end of the creeping section (marked
volumes in Figure 6.5) illustrate the large \( b \)-value contrast between these two
volumes. Even though the \( b \)-values of the asperity and the creeping section
are not identical in the three samples, they are remarkably similar and con-
voy the same information. In the asperity, the \( b \)-values are in the range of
\( b \in [0.46, 0.60] \), indicating very low \( b \)-values. Because the observation periods
are different between the three sets, the \( a \)-values also differ. The lowest \( b \)-value
Figure 6.6: Frequency-magnitude distributions of the two selected volumes marked in Figure 6.5. Red crosses mark the frequency-magnitude distributions of the Middle Mountain asperity, black squares of the creeping section. (A) NCSN catalog from 1981–2003 (r = 5 km), frame B in Figure 6.5. (B) HRSN catalog (r = 3 km), frame C in Figure 6.5. (C) NCSN catalog from 1969–1981 (r = 5 km), frame D in Figure 6.5.

is detected in the HRSN catalog with a sampling radius of r = 3 km. Using a radius of r = 5 km would increase the b-value to 0.57, slightly smoothing the large contrast. In the creeping section, we observe relatively high b-values in the range of b ∈ [1.04, 1.12]. Consequently, the Utsu test establishes that it is statistically highly unlikely that the frequency-magnitude distributions of the creeping and asperity section come from the same population (log \( P_b \leq -5 \)).

The probability of a future earthquake of a given magnitude at any location along the fault segment investigated can be estimated according to equations 6.3 and 6.4. In Figure 6.7, we compare the resulting earthquake probabilities for an \( M \geq 6 \) event given the assumption of a spatially varying b-value, and the assumption of a regionally constant \( \overline{b} = 0.92 \), as it is often used in seismic hazard assessment. The forecasts issued by these two approaches are strikingly different: The volume of the asperity at a distance from point \( P_1 \) of 70 km shows a high probability \( P_r \approx 0.025 \) for one or more events of magnitude \( M \geq 6 \), when computed with local b-values. This observation has previously been documented by Wiemer and Wyss [1997] as a \( T_r \)-minimum. When computing the \( P_r \) with the constant overall \( \overline{b} \)-value, the location of the highest probability is in the creeping section, because here the activity of microearthquakes is
6.5 Results

The probability is with \( P_r \approx 0.004 \) overall lower.

### 6.5.2 Stationarity of \( b \)-values

The stationarity of \( b \)-values and the stationarity of the aforementioned strong spatial variations in \( b \) is the next question we address. Our method to detect local temporal changes in \( b \) is based on mapping \( \Delta b \), the difference between the values in two periods [Wiemer et al., 1998]. If \( \Delta b \) is statistically significant at some locations, we plot \( b \) as a function of time for those volumes to further investigate the cause of change. As a first test of stationarity, we compare \( b \)-values by simply subdividing the dataset into two periods of equal length, obtaining two 11-year periods (1981–1992 and 1992–2003). In Figure 6.8, frames A and B show the \( b \)-value cross-sections for these two periods. Overall, both periods show the same pattern. The differences \( \Delta b = b_{1992-2003} - b_{1981-1992} \) are small (frame C). They range from \( \Delta b = -0.49 \) to \( \Delta b = 0.36 \), while 90% of the values are in the interval \( \Delta b \in [-0.1, 0.2] \).

We do not consider changes in \( b \)-value significant if they do not exceed the standard deviation \( \sigma \) (obtained by bootstrapping) of the \( b \)-values, shown in frame D and E. The standard deviation of the first period’s \( b \)-value (frame D) is on average \( \sigma = 0.116 \) and 90% of the values are smaller than 0.187. For the second period (frame E), the average standard deviation is \( \sigma = 0.122 \).
while 90% of the values are smaller than 0.209. Comparing the values of $\Delta b$ with $\sigma$ shows that the differences in $b$-values mostly equal the standard deviation. Therefore, we can say that except for a few samples, the $b$-values remain stationary in the bounds of their standard deviation $\sigma$. This statement is further quantified by applying at each node the test proposed by Utsu [1992] (frame F). Log probabilities of $P_b \leq -1.3$ (auburn colors) indicate significant changes in $b$-value. Only 34 of 2950 computed samples show significant changes in $b$-value, equivalent to less than 1.2%.

The result of this stationarity investigation depends on the two periods
selected. A second sensible division of the data is at 1996, because this allows us a prospective test of the $b$-value distribution at Parkfield published by Wiemer and Wyss [1997]. Figure 6.9, analyzing the periods 1981–1996 and 1996–2003, reveals a different picture from the previous division in 1992. We again plot the $b$-values in the two respective periods, their standard deviation, the difference and finally the result of the Utsu test. While most of the fault segment shows stationary behavior, we can identify several volumes with significant $b$-value changes according to the Utsu test. The largest changes are located at distances from $P_1$ of 15 km, 25 km and 90 km (frame C). At a distance of 15 km, the $b$-value dropped by a maximum of $\Delta b = -0.37$. The largest change in this area happened at depth of $D \approx 5$ km. Even though the standard deviation $\sigma$ is rather large at this depth (frames D and E), the Utsu test shows significant changes in $b$-value (frame F). In the volume at a distance of 25 km, the change in $b$-value shows similar values ($\Delta b \approx 0.2$–0.4) but the standard deviation $\sigma$ is smaller (frame D), about 0.1 or less. The Utsu test shows highly significant changes in $b$-values for these volumes. At a distance of 65 km, we can detect changes in $b$-value of $\Delta b \approx 0.2$, while the standard deviation shows smaller values. The Utsu test shows significant changes in $b$-values for this volume also. While in the previous volume the seismicity rate stayed constant, we can detect a higher rate in the last years of the second period for this volume. An extreme change in $b$-values can be detected in the volume at a distance of 90 km with changes up to $\Delta b \approx 1$. In this volume, we also found a change in seismicity rates, increasing in the second period. In total, 488 of 2580 computed nodes show significant changes in $b$-value, corresponding to 18.9%.

Although both stationarity tests (Figures 6.8 and 6.9) cover in total the same period (1981–2003), the results are different. In the first test, almost no volumes showed significant $b$-value changes, in the second test, the portion of volumes with significant changes in $b$-values amounts to 18.9%. To explore this difference in more detail we inspect the frequency-magnitude distributions of
the different periods and the development of $b$-values as a function of time at nodes that showed a change. The $b$-value for the volume at a distance from $P_1$ of 25 km (volume 1) remains stationary and below the regional average $\overline{b}$-value during 1981–1995, but it increases from 1996 on (Figure 6.10 frames 1A and 1B). It reaches a maximum in the period 2000–2003. When selecting 1992 for dividing the dataset into two periods, the $b$-value of the second period of the catalog (1992–2003) is the average of the observed development during that time, thus the difference of $b$-values between the two periods is insignificant. The green indicators in the Figure show the $b$-values of the two periods and their difference.

The same effect can be seen in the volumes at distances of 65 km (volume 3 and frames 3A and 3B) and 90 km (volume 4 and frames 4A and 4B). In the latter case, no $b$-value for the pre-1992 period can be computed but the development of the $b$-value with time shows a clear increase, starting about 1995. Volume 3 shows a dip in $b$ limited to the years 1990–1995. During other times, the $b$-value remains stationary with values $b \approx 0.8$. Although this volume exhibits fluctuations in $b$, these values are at all times smaller than the average value of $\overline{b} = 0.92$, preserving the information of comparatively low recurrence time. In contrast, the volume at a distance of 50 km (volume 2 and frames 2A and 2B) shows perfect stationary behavior with $b$-values around 1, which is above the regional average of $\overline{b} = 0.92$.

Examining the results for volume 3 in Figure 6.10 frames 3A and 3B, we see only slight variations in $b$-values. Considering the changes in completeness over the period of investigation and errors in magnitude estimates, we interpret this $b$-value pattern as nearly stationary. Also, the $b$-value over time keeps presenting the information that it is lower than 1 and is not scattering around the overall $b$-value of $\overline{b} = 0.92$.

In contrast, in volume 4, we see a larger $b$-value difference accompanied by a probability in the range of $\log P_b \in [-4, -2.5]$. This volume shows clearly a significant change in $b$-values, as demonstrated by the green bar for the year
Figure 6.10: Stationarity test comparing the periods 1981–1996 and 1996–2003. (Center frame) Log-probability of \(b\)-values having non-stationary behavior according to Utsu [1992]. Circles mark volumes 1–4 (Location of volumes in distance from point \(P_1/depth\). Volume 1: 26 km/7 km. Volume 2: 50 km/5 km. Volume 3: 64 km/8 km. Volume 4: 92 km/7 km.). CRR1 indicates the location of the creep meter at the Carr Ranch site [Roeloffs, 2001] (A frames) \(b\)-value with time of volumes 1–4. For volumes 1–3 a sample size of 100 events is used, for volume 4 a sample size of 50 events is used. The green indicator bars indicate the \(b\)-value changes between both the periods 1981–1992 and 1992–2003 and the periods 1981–1996 and 1996–2003. In frame 4B only the latter change is indicated. (B frames) Cumulative frequency-magnitude distributions of both periods of volumes 1–4. Red squares mark the distribution of the period 1981–1996, green squares of the period 1996–2003. The \(b\)-values of both periods are marked by the accordingly colored lines.
1996 in frame 4A. We conclude that, even blurred by scattering, the $b$-value distribution in most volumes contains information about low or high $b$-values.

### 6.5.3 Analysis of stationarity for different periods

To systematically assess the stationarity of $b$, we repeat the above analysis for seven midpoints, starting in 1986 and moving in two year steps (1986, 1988, 1990, 1992, 1994, 1996 and 1998, requiring at least 5 years in each period). We only show the final results of the Utsu test, because in the maps of probabilities $P_b$, the nodes where $b$ is stationary can be readily identified (Figure 6.11). We can clearly see that no significant change in $b$-values occurred until 1992 (frame D). From 1994 (frame E) on, three patches where significant changes in $b$-values can be detected. Two patches (at a distance from point $P_1$ of 65 km and 90 km, volumes 1 and 3) show an increase in activity of events with smaller magnitudes ($M \leq 2$) starting around 1996–1998. This caused a higher $b$-value in these volumes. Catalog subdivision at 1998 again shows few significant deviations from stationarity. This shows that $b$-variations are averaged out when longer periods are used for the calculation of $b$.

We also applied this test on changes of $b$-values as a function of time using a moving window technique. We took five year periods before and after the dividing date to compute $b$-values. Using periods of fewer than five years left too few events in most of the volumes for computing $b$-values. This test showed similar results to those shown in Figure 6.11.

### 6.6 Discussion and Conclusions

The $b$-values along the Parkfield segment of the San Andreas fault vary significantly. The asperity region beneath Middle Mountain exhibits an anomalously low $b$-value of $b \approx 0.5$, (Figures 6.5 and 6.6). The neighboring southernmost part of the creeping section at shallower depth is characterized by values of $b > 1.1$. Using the HRSN dataset based on borehole instruments gives the rare
opportunity to independently confirm a seismicity pattern (Figures 6.5C and 6.6B). This dataset has superior location accuracy, confirming that location errors cannot be the cause for the spatial differences in $b$. It also is based on magnitudes derived from moments [Wyss et al., 2004], confirming that systematic biases in magnitudes cannot explain the spatial variations in $b$. Finally, by analyzing the period 1996–2003 (Figure 6.9B), data collected after the hypothesis that $b$-values at Parkfield vary spatially has been proposed, we confirm the basic pattern in $b$ also in a prospective test mode. Prospective testing is considered the ultimate test of a hypothesis, because only in this way inadvertently biasing the analysis can be excluded [Jackson, 1996; Mulargia, 2001]. In summary, we cannot think of any possible artifact that may create the strong spatial heterogeneity at Parkfield, and thus must accept that it is a natural phenomenon. The assertion that $b$-value computations are mainly influenced by the corner magnitude $m_c$ [Kagan, 2002], but are essentially everywhere the same, cannot hold for the Parkfield segment.

When mapping spatial variability in $b$ one has to strive for a suitable balance between available resolution, uncertainty of the estimate and the size of the seismotectonic feature under investigation. We illustrate this important yet sometimes neglected point in Figure 6.4, where we image $b$ using sampling radii ranging from 2–20 km. Sampling with $r > 5$ km mixes populations of dissimilar frequency-magnitude distributions and thus cannot resolve the apparent intrinsic structure, or seismotectonic fabric at Parkfield, resulting in average values of $b \approx 0.9$. For $r < 5$ km no significant additional heterogeneity of $b$-values was detected. Thus we strengthen the conclusion of Wiemer and Wyss [1997] that a radius in the range of 4–5 km is the best choice for this fault segment and dataset for identifying and resolving contrasts in $b$-values.

Establishing the optimal sampling dimensions is the first step in the method we propose here to test for stationarity of $b$. The optimal radius varies from region to region, depending on the seismotectonic target and the density of catalog information. In a worldwide study where $M_c$ might be 5.5, the min-
imum dimensions of volumes containing enough events for analysis are often 100 km, which means that details like asperities of earthquakes of magnitude $M = 6$ cannot be resolved, even if such asperities existed. Thus, studies in which it cannot be demonstrated that smaller radii than a certain optimal value yield no additional information have only limited value because they do not penetrate to the depth appropriate for the dimensions of the local seismotectonic fabric. Nevertheless, the amount of data is constraining the targeted dimensions of seismotectonic fabric which can be resolved.

### 6.6.1 Investigating stationarity of $b$-values at Parkfield

We have performed in this paper for the first time an in depth investigation of stationarity of $b$ in the Parkfield region. This question is important because it relates to the physical understanding of $b$-values and transients in the Earth crust, and is also highly relevant for the probabilistic forecasting of seismicity. It is complicated by the intrinsic coupling of stationarity and spatial heterogeneity. Stationarity can only be established relative to a given spatial volume. We use as our reference framework the aforementioned 5 km sample radii, as they are the upper limit that resolve the spatial $b$-value contrasts.

As a second step in $b$-value analyses, we recommend that one tests the stationarity of $b$-value patterns that may exist as a function of space. The method we propose here is outlined in Figures 6.8 through 6.11. Because one looses information by mixing samples with different magnitude distributions, mixing must be identified and avoided if possible. This applies to both, mixing dissimilar samples in space and time. Thus, we propose that mapping of $\Delta b$ (comparing data from two periods) and judging its significance by Utsu’s test (Figures 6.8 and 6.9) is a way to find possible changes of $b$ at the time selected for the comparison. The amount and exact time of change can then be determined by plots of $b$-value as a function of time (Figure 6.10) for the locations where maps of Utsu probabilities have shown a change. Because
not all changes are identifiable by a single cut in time, one needs to plot the probabilities $P_b$ according to the Utsu test for all possible divisions of the data set in time (Figure 6.11).

Only a minute percentage of the nodes (1.2%) showed temporal changes when subdividing the data in half (Figure 6.8). When analyzing the data for the periods 1981–1996 and 1996–2003 (Figure 6.9), a larger percentage (18.6%) of nodes displays a change in $b$. The analysis of stationarity for different periods (Figure 6.11) essentially leads to the same conclusions, but it establishes in addition that the change in $b$-values initiates between 1992 and 1994.

The test proposed by Utsu [1992] assumes that the frequency-magnitude relationship perfectly obeys a power law. Thus, it may interpret two earthquake populations as having significantly different $b$-values, when in reality the difference may stem from errors, not from a real difference in mean magnitude. The problem with this test is that it does not allow for errors in magnitudes and completeness and, thus, underestimates the probability $P_b$.

Figure 6.12 shows clearly that the observed changes in $b$-value are not an effect of random scattering of $b$-values around the regional average. Each cross in one of the frames of Figure 6.12 represents the two $b$-values (first and second period) of a single node and their standard deviation. For the stationarity test with periods 1981–1992 and 1992–2003 (frame A) most datapoints are aligned along the stationarity line (gray line in Figure 6.12). Performing the stationarity test with periods 1981–1996 and 1996–2003 still shows a majority of datapoints aligned along the stationarity line, although 18.9% of the nodes showed significant changes in $b$-values. If these changes would indicate random scattering around the regional average, the crosses in Figure 6.12 would have shape a circular cloud with its center at the regional average $\bar{b}$-value.

A number of other geophysical transient changes around 1993 have been detected and described in the Parkfield region. The change in $b$-value in volume 4 (Figure 6.10) correlates with the region of increased creep at the Carr Ranch site (location of creep meter CRR1, Figure 6.10), as detected by EDM
Figure 6.12: Changes in $b$-values in the Parkfield segment of the San Andreas fault. Each cross represents one sample of the cross-section. The position of each cross indicates the $b$-values of the first and second period. The crossbars indicate the error in $b$-values for both periods. The color of each cross corresponds to $\log P_b$ of the test proposed by Utsu [1992]. The gray line marks the stationary behavior. (A) Changes in $b$-values for the periods 1981–1992 and 1992–2003. (B) Changes in $b$-value for the periods 1981–1996 and 1996–2003.

and creep meter data. This geodetic transient has been investigated at length by Gao et al. [2000] and Roeloffs [2001], their conclusion is that it represents most likely a tectonic signal rather than a rainfall induced artifact. Our observation of increasing $b$-values in this location after 1993 corroborates this interpretation. The most viable interpretation to us is that a transient decrease in locking strength started around 1993 in this fault segment, increasing the creep at depth. The seismicity reacts to this strain transient by producing relatively more small events and fewer large events (Figure 6.10 frame 4A), or a $b$-value increase from $b \approx 0.7$ to $b \approx 1.2$.

Transients in $b$ that are established with high significance and correlate with other geophysical signals are rare and important in order to enhance the physical understanding of the frequency-magnitude distribution. Increases in $b$-value have been reported in volcanic regions [Wyss et al., 1997; Wiemer et al., 1998] where magma or fluids are believed to have migrated and caused these changes. In a recent study in the Tokai region of Japan, Wiemer et al. [2004] were able to show a clear correlation of $b$-value and subsidence measure
6.6 Discussion and Conclusions

by leveling data. Additional evidence for understanding $b$-value changes is based on the observation that $b$ decreases significantly with depth in California [Mori and Abercrombie, 1997; Wiemer and Wyss, 1997; Gerstenberger et al., 2001]. This observation is consistent with laboratory experiments by Amitrano [2003], who suggests that the increase in confining pressure at greater depth is the reason for lower $b$-values.

We feel that the increase of $b$ at Parkfield is consistent with the conceptual model of the striking spatial differences in $b$ between the asperity and creeping section of the fault (Figure 6.5): A largely locked fault is characterized by low $b$-values, and this fact can be used to map asperities [Wiemer and Wyss, 1997; Wyss et al., 2000; Öncel and Wyss, 2000; Zúñiga and Wyss, 2001; Wyss and Matsumura, 2002]. Creeping sections of faults on the other hand display the opposite kind of behavior, relatively higher $b$-values [Wiemer and Wyss, 1997; Amelung and King, 1997]. When the southern part of the locked section (volume 4 in Figure 6.10) started to creep, a corresponding change in $b$-value occurred. An important conclusion from this observation is that the governing factor that determines the $b$-value cannot be material heterogeneity engrained in the volume under investigation, because this could not be changed readily. To be able to change the $b$-value so readily after a comparatively minor creep episode suggests that the local stress/strain environment and/or fluid interaction are the main causes. However, the ultimate physical reasons for the difference of $b$-values between the asperity and the creeping section is not known. The SAFOD drilling hole might help to open new perspective on processes at the fault. Unfortunately, this hole will only be drilled to a depth of 4 km, it will only reach the high $b$-value zone in the creeping part of the fault not the more interesting asperity at a depth of approximately 10 km, Figure 6.13.

The frequent occurrence of events in the asperity (1857, 1881, 1901, 1922, 1934, and 1966 [Bakun and Lindh, 1985]) suggest that the properties of the asperity remain stationary over decades to centuries. Taking the $b$-value as
Figure 6.13: Distribution of \( b \)-values in the Parkfield segment of the San Andreas fault computed with the NCSN catalog from 1981–2003 (\( r = 5 \) km, \( N_{\text{min}} = 50 \)). Earthquakes with \( M \geq 1 \) are plotted as circles. Events with \( M > 4.5 \) are marked by red stars. The boxes denoted with 1, 2, and 3 are the small Middle Mountain alert box, the Middle Mountain alert box, and the Parkfield alert box, respectively [Michael and Jones, 1998]. The surface break of the 1966 event is marked by an arrow. The SAFOD drilling hole is marked by a vertical black line with a derrick on top.

A quantity related to the stress level, the stationarity of \( b \)-values along the Parkfield segment, and especially around the asperity and the locked segment, supports the concept of the asperity being the nucleation zone of larger earthquakes which propagate southeastwards. A preferred southeastward propagation of moderate and large earthquakes in the Parkfield area is expected from the known velocity contrast across the San Andreas fault there [Ben-Zion, 2001]. We have shown that the \( b \)-values remained stationary, although in one fault patch a significant change has been observed (volume 4 in Figure 6.10).

We conclude that in most (90%) of the fault surface mapped the pattern of disproportionately higher/lesser production of small earthquakes in the unlocked/locked segments, respectively, remains stable for more than 30 years, and therefore should be considered stationary. However, transient changes in subvolumes do occur if significant changes in the environmental conditions take place.

We want to elaborate on the question how one should sample a dataset of earthquakes that contains variations of \( b \) as a function of space and time. If one uses the entire data set to calculate \( b \), the result is usually near \( b = 1 \), and one gains no information. One can even argue that such a selection makes no sense because dissimilar groups of earthquakes are arbitrarily mixed. The same is true for any subset of the data that is strongly heterogeneous. It follows
that, ideally, one wishes to subdivide the data set into all subsets that are homogeneous and stationary. Although it is difficult to detect local changes of $b$ as a function of time in the presence of strong spatial heterogeneity, it is possible by maps of $\Delta b$ and probabilities $P_b$ (Figures 6.8 and 6.9). Also, variations of $b$ with time are usually not frequent, thus it is not necessary to subdivide into samples of short periods (Figure 6.10 A frames). Therefore, we recommend that the following steps should be used in a detailed analysis of the mean magnitude, or $b$: (1) Map (in cross section or in normal map view) variations of $b$ as a function of sampling radius (as in Figure 6.4) to establish the most suitable sampling radius. (2) Using this sampling radius, map $\Delta b$ and $P_b$, comparing the data from two periods, for a suite of separation dates. (3) Select samples with homogeneous and stationary distribution of earthquake sizes for tectonic analysis. This suggested procedure is idealized and approximations have to be exercised in datasets that usually are complex.

6.6.2 Implications for earthquake hazard and probabilistic forecasting

The pattern of probability of future earthquakes along the fault is strikingly different in our model (varying $b$-values) and the approach of applying the overall $\bar{b}$-value to all locations, asperity and creeping segments, alike (Figure 6.7). Our model predicts that major earthquakes are most likely to initiate in the Middle Mountain asperity, whereas a model with $b = \text{const.}$ predicts that major ruptures are most likely in the seismically highly active, creeping segment of the fault. Based on the facts that the Middle Mountain segment is recognized as an asperity and that low $b$-values are linked to high stress environments [Scholz, 1968; Wyss, 1973; Urbancic et al., 1992] we believe that our model takes account of physical conditions along the fault, and is more appropriate for forecasting seismicity. This expectation has been tested in Paper II.
The low $b$-values in the asperity and the locked segment together with their stationarity strongly support our hypothesis that the asperity is the nucleation zone of earthquakes which ruptured the locked part of the fault. The asperity matches the Middle Mountain alert box (marked with 2 in Figure 6.13) by Michael and Jones [1998]. The Small Middle Mountain alert box (marked with 1) is not specially characterized by $b$-value features compared to the Middle Mountain alert box but is also located in the asperity. The event on November 14th, 1993 of magnitude $M = 5$ occurred in this alert box while all three events with $M > 4.5$ are located in the Middle Mountain alert box (red stars in Figure 6.13). This low $b$-value zone continues about 25 km southeast and corresponds to the 1966 rupture zone according to Segall and Du [1993] and matches perfectly the Parkfield alert box by Michael and Jones [1998]. The observed postseismic surface break (Figure 6.13, arrow) according to Smith and Wyss [1968] extends from the asperity over the locked part into a high $b$-value volume where it stopped.

When computing the probabilistic recurrence time $T_r$ of the entire Parkfield volume for a magnitude $M = 6$ event, we get $T_r = 116.6$ years. This is a clear overestimate based on the proposed Parkfield earthquake cycle of about 22 years [Bakun and McEvilly, 1979, 1984]. In contrast, the local recurrence time of the asperity volume, $T_L \approx 30$ years agrees well with the observed recurrence times of the earthquake cycle.

The local recurrence time is not simply another representation of the $b$-value. It takes into account the $b$-value as well as the $a$-value of a particular volume. The low $b$-value zone extents from the asperity (distance about 70 km in Fig 6.13) southeast along the rupture zone (to a distance point about 90 km). However, only the nucleation point of the Parkfield earthquakes shows the low values of local recurrence time or high values of probabilities $P_r$. Only a suitable combination of $a$- and $b$-values produces low local recurrence times $T_L$. On the other hand, using the regional average $\overline{b}$-value when computing the probability $P_r$ of one or more magnitude $M = 6$ events, selects the creeping
section as the locus with the highest probability (Figure 6.7A). In this case, the probability is proportional to the $a$-value. This result contradicts the observation of the Parkfield earthquake cycle and also the observation that the three events with magnitude $M > 4.5$ which occurred in the period 1981–2003 are located in the Middle Mountain alert box (Figure 6.13). Computing the probability $P_r$ for one or more events of magnitude $M = 6$ using spatially varying $b$-values shows the highest probability in the Middle Mountain alert box. This result again emphasizes the concept of the asperity being the nucleation point of larger earthquakes at Parkfield.

## 6.7 Acknowledgments

We would like to thank the Northern California Earthquake Data Center (NCEDC) for providing us with catalog data, the Northern California Seismic Network, U.S. Geological Survey, Menlo Park for the NCSN catalog, and the Berkeley Seismological Laboratory, University of California, Berkeley for the HRSN catalog. We profited from discussions with M. Gerstenberger, D. Jackson, and R. Zúñiga. We also want to thank J. Woessner, J. Ripperger, G. Hillers, and Y. Ben-Zion for their useful hints. This paper is contribution number 1370 of the Geophysical Institute, ETH Zürich. This research was supported by the Southern California Earthquake Center. SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 786.
Chapter 7

Earthquake Statistics at Parkfield II: Probabilistic Forecasting and Testing

Danijel Schorlemmer\textsuperscript{1}, Stefan Wiemer\textsuperscript{1}, Max Wyss\textsuperscript{2}, D. D. Jackson\textsuperscript{3}

\textsuperscript{1} Swiss Seismological Service, ETH Zürich, Zürich, Switzerland.
\textsuperscript{2} World Agency of Planetary Monitoring and Earthquake Risk Reduction, Geneva, Switzerland.
\textsuperscript{3} Department of Earth and Space Sciences, UCLA, USA.

7.1 Abstract

In Paper I [Schorlemmer et al., 2004] we showed that the spatial $b$-value (of the Gutenberg-Richter relation) distribution at the Parkfield segment of the San Andreas fault remained stationary for the past 35 years. In this paper (II), we extend those results, construct two probabilistic forecasts ($H^1$ with a spatially varying $b$-value, and $H^2$ with a uniform $b$-value) and test these hypotheses against each other. Both hypotheses use a spatially varying seismicity level ($a$-value) determined from past seismicity. We used a range of sampling parameters (magnitude threshold, cell size, etc.) to assure robust results. We found that in most of the tests, hypothesis $H^1$ showed a higher likelihood than $H^2$, although neither provided a consistent fit to the seismicity data. The most positive results for $H^1$ are obtained for testing magnitude ranges down to $M = 1.5$ and with sampling radii as defined in Paper I as appropriate for Parkfield. The superior performance of $H^1$ suggests that spatially varying $b$-values should be considered in earthquake forecasts.

7.2 Introduction

Strong spatial heterogeneity of $b$-values exist at Parkfield, which we investigated in detail in Paper I [Schorlemmer et al., 2004]. This heterogeneity is not an artifact of station distribution, attenuation or corner magnitude. We demonstrated in Paper I the stationary behavior of $b$-values by applying a detailed quantitative analysis based on the test proposed by Utsu [1992].

The key question addressed in this paper is whether or not using spatially varying $b$-values improves probabilistic seismicity forecasts, compared to forecasts based on regional average $b$-values, as commonly used in contemporary seismic hazard estimates [Grünlthal and Bosse, 1996; Giardini et al., 1999]. In most probabilistic hazard studies, the $b$-value is either assumed to be constant for the entire study region [Frankel et al., 1996, 1997], or constant for specific
source zones. The dimensions of these zones are generally 20–100 km, thus exceeding by far the scale of heterogeneity in $b$ that we have documented at Parkfield, and which we believe is associated with the physics of asperities in the Earth's crust. However, even though strong heterogeneities of $b$ have been observed at Parkfield, it is not obvious that they translate into significantly improved probabilistic forecasts. In fact, it is not clear how to measure and establish such significant differences in the first place, and consequently much of this paper is devoted to developing a method of how to compare two probabilistic seismicity models in the presence of small scale heterogeneities.

In this work, we do not attempt to model temporal fluctuations in either $a$-value or $b$-value. If these values are not stationary, then future earthquakes might not match our forecasts well. Furthermore, we have to address the question of which combination of sampling parameters leads to the most successful forecasts. Even though we found that sampling radii of 4–5 km result in largest coverage while still revealing all $b$-heterogeneities and contrasts, different sampling parameters may show better performance in the full probabilistic test applied here. This can be due to minor changes in $b$, which are obscured with larger radii and possibly result in a more stable forecast.

In this paper, we test probabilistic seismicity forecasts of two different hypotheses against each other. One hypothesis, $H^2$, generates forecasts based on spatially varying $a$-values but assuming the constant regional average $\bar{b}$-value. The alternative, challenging hypothesis $H^1$ uses spatially varying $a$- and $b$-values for its forecasts. To accomplish this task, we first need to develop the necessary method for testing probabilistic forecasts. This includes the development of a likelihood-ratio test and rules for creating simulations of real data for establishing significance of results. The testing method is implemented similarly to standard likelihood-ratio tests. However, in our case, both hypotheses have the same degrees of freedom, because neither forecast is adjusted to future seismicity. Thus, the standard approach with $\chi^2$-distributions for establishing significance is not applicable and we need to create a probability distribution
for the likelihood score for judging the significance of the results.

Different approaches to probabilistic earthquake forecast testing have been proposed [Jackson, 1996; Evison and Rhoades, 1999; Console, 2001]. These testing methods have been used for the west Pacific regions [Jackson and Kagan, 1999; Kagan and Jackson, 2000], China [Rong, 2002; Rong and Jackson, 2002], California [Kagan and Jackson, 1994], as well as worldwide [Kagan and Jackson, 1995; Rong et al., 2003]. Most of these papers deal only with the spatial distribution of earthquakes above a certain magnitude threshold. In our case we are interested in testing the entire magnitude range in magnitude bins of $\Delta M = 0.1$. Therefore, modifications to the existing testing method are necessary. Our development of a new extended method represents the first step towards a community agreed suite of tests to be used in the Regional Earthquake Likelihood Models (RELM) project in Southern California (www.relm.org). For RELM, the method will be further extended to test quasi-stationary and time-dependent models.

The tests we present here are ‘pseudo-prospective’, meaning that we divide the presently available catalog into a ‘learning’ and ‘testing’ part. We have varied several sampling parameters, such as the magnitude threshold and the size of regions for which we estimate separate $b$-values, in such a way that the forecast based on the learning catalog best fits the testing catalog. We find the conclusions quite robust, in that they are not very sensitive to the sampling procedures. Nevertheless, the ultimate test will be a fully prospective test in which all parameters are fixed in advance. We describe here such a prospective test to be based on earthquakes after January 1st, 2004.

In this study, we do not address the important question of how to determine the frequency-magnitude distribution for the largest events in a given region, e.g., the characteristic versus the truncated Gutenberg-Richter forecast models [Schwartz and Coppersmith, 1984; Pacheco et al., 1992; Wesnousky, 1994; Kagan, 1993; Kagan and Jackson, 1994]. For forecasting accurately the recurrence rate of the very largest events, which often are the main contributors to
seismic hazard at a given site, this difference between the various forecast models becomes important. However, to test a characteristic versus a truncated Gutenberg-Richter model requires a dataset that contains at least several of these large events, meaning either a large spatial [Kagan and Jackson, 2000] or temporal [Wesnousky, 1994] extent of the data set under investigation. Our tests are based on the entire frequency-magnitude distribution above magnitude 1.5. The largest earthquake in our testing catalog is of magnitude 5.0.

Parkfield is selected as the study area, because it is one of the best monitored and studied fault segments, and it is known to exhibit large $b$-value contrasts on small scales. In addition, this test extends the stationarity investigation in Part I. By choosing the well controlled Parkfield segments for developing the testing method, we strive to build a solid understanding of the testing, which then can be applied to extended regions, such as all of California.

### 7.3 Method

In both models, we forecast earthquakes by extrapolating the Gutenberg-Richter relation to a predefined maximum value ($M_{\text{max}} = 7.0$) using spatially varying $a$-values of the recent seismic activity. Only our hypothesis $H^1$ extrapolates using spatially varying $b$-values. The standard hypothesis $H^2$ also considers a $b$-value, but only a constant regional average $\bar{b}$-value.

We extrapolate the observed seismicity to a forecast or expectation $\lambda$ for every magnitude bin $M_i$. The magnitude bins constitute the set $\mathcal{M}$

$$\mathcal{M} := \{M_1, M_2, \ldots, M_n | M_i < M_{i+1}\}$$

The cumulative (over magnitudes) expectation $\lambda'$ is given by the Gutenberg-Richter relation as

$$\lambda'(M_i) = 10^{a-bM_i} \frac{T_\omega}{T_\lambda}$$

where $T_\omega$ denotes the length of the forecast or observation period and $T_\lambda$ the
length of the learning period, for which the given $a$-value was computed. Extracting the expectations $\lambda$ of the single bins from the cumulative expectations $\lambda'$ leads to

$$\lambda(M_i) = \begin{cases} 
\lambda'(M_i) - \lambda'(M_{i+1}), & i < n \\
\lambda'(M_i), & i = n 
\end{cases}$$

where $n$ denotes the number of magnitude bins in $\mathcal{M}$. For creating the expectations of the hypothesis $H^1$, at every node the local $a$- and $b$-values are used, whereas for $H^2$ the local $a$-values but regional $b$-value are used.

Testing probabilistic earthquake forecast models requires determining the probabilities of observations with respect to the expectations of both models. Expectations are the issued forecasts, thus expected numbers of events for a particular magnitude, location, and period. Therefore, we first have to define the extent and resolution (nodes per unit distance) of the testing volume, as well as the range and resolution of magnitudes for testing (bins per unit magnitude). In the case of the Parkfield segment, we define the testing volume as a cross-section along the San Andreas fault. We call the combination of spatial cells and magnitude bins space/magnitude bins or simply bins. Any forecast has to be given as an expected rate of earthquakes per bin for the defined forecast period. Thus, expectations are a function of a bin defined by its spatial extent, magnitude range and period for the forecast. All defined bins $c_i$ for which forecasts are issued constitute the set of bins $\mathcal{C}$

$$\mathcal{C} := \{c_1, c_2, \ldots, c_n\}$$

while every bin $c_i$ is defined as function of its describing parameters

$$c_i := c_i(V, M, T)$$

where $V$ denotes the volume definition, $M$ the magnitude range, and $T$ the period for which the forecast will be issued.

In Paper I and other previous papers [Wiemer and Wyss, 2002, and ref-
7.3 Method

We have selected earthquakes from overlapping cylinders or other overlapping regions to estimate $a$- and $b$-values. The overlap provides a useful smoothing and assures a sufficient sample size from which to estimate $a$- and $b$-values at each grid point. However, our testing method assumes that the outcomes for each bin are independent, which requires that the bins be disjoint. In most tests described here, we estimated the $b$-values from overlapping regions, because we need a sufficient set of earthquakes for each estimate, and we believe the $b$-value will be relatively smooth at the scale of the cells. We estimated the $a$-values, and tested the results, on disjoint cells. For a fully prospective test on disjoint volumes, the smoothing of the $b$-values causes no problem, because the method simply tests the predictive power of the model, no matter how the parameters were derived. The cells are cuboids and additionally have dimensions of 0.1 magnitude unit. The length of these cuboids is the width of the cross-sectional volume perpendicular to its strike, while their width is the distance along the cross-section and their height the depth range, respectively.

In some tests described below, we also computed the test results for overlapping regions (cylinders). The resulting images are useful for understanding the test performance and for delineating the regions in which $b$-value variations significantly improve forecasting effectiveness. In those cases the requirement for independence is not met, and some test earthquakes may be counted more than once as they may appear in more than one test region. For these cases the probabilistic tests are less rigorous than the ones for disjoint cells. However, we use these less rigorous methods primarily for illustration, and our main results are fully supported by the more rigorous tests on disjoint cells.

The expectation $\lambda^J_i$ of a hypothesis $H^J$ for a single bin $c_i$ is a function of the bin, given by

$$\lambda^J_i := \lambda^J_i(c_i), c_i \in C$$

For every hypothesis, we can set up the vector of the expectations $\Lambda^J$ of all
bins as

\[ \Lambda_j := \begin{pmatrix} \lambda^j_1 \\ \lambda^j_2 \\ \vdots \\ \lambda^j_n \end{pmatrix}, \quad \lambda^j_i := \lambda^j(c_i), c_i \in C, n = |C| \]

For issuing a forecast, we have to compute every single expectation \( \lambda^j_i \) of the vectors \( \Lambda^j \) for both hypotheses. Therefore, we need to compute the \( a \)- and \( b \)-values for each cell (spatial bin or cuboid). At every node defined, we select all earthquakes within a cylindrical volume with fixed radius \( r \). This is the same procedure as described in Paper I and by Wiemer and Wyss [2002].

Why are we using cylindrical volumes for computing \( b \)-values and not the cuboids? The cuboids have to be quite small in order to resolve the variations in seismicity. We choose cuboids of \( 2 \text{ km} \times 2 \text{ km} \times 5 \text{ km} \). Many do not contain enough events for reliably computing \( b \)-values. Therefore, we have to choose larger volumes for estimating \( b \). We sample cylinders, rather than enlarged cuboids, because the former have no corners and better represent the earthquake distribution in the neighborhood of the central point. This approach smoothes \( b \)-values across cells and reflects our level of knowledge about the spatial distribution of \( b \)-values. We consider our approach to be a good compromise between testing resolution and the requirements for reliable high resolution \( b \)-value estimates. If a volume contains the necessary minimum number of events per sample, \( N_{\text{min}} \), we compute the \( b \)-value for this particular node using the maximum likelihood method (see Paper I for details); otherwise, we remove it from the set of bins \( C \) used for testing. The \( a \)-values computed from cylindrical volumes cannot be used for testing, because they do not reflect the activity of the cuboid for which forecasts are issued. Therefore, we compute the maximum-likelihood \( a \)-value for the cuboid sample, constrained by the computed \( b \)-value (either the computed \( b \)-value for the hypothesis \( H^1 \) or the overall \( \overline{b} \)-value for the hypothesis \( H^2 \)) for the cylinder. In the case where we test using cylindrical volumes (overlapping bins), however, the \( a \)-value is
directly determined from these volumes.

With the computed $a$- and $b$-value at each node, we then compute the expectation by extrapolating the Gutenberg-Richter relation into the future, assuming stationary $b$-values and a constant rate of earthquakes per unit time. Iterating through all the spatial bins yields the full vectors of expectations $\Lambda^1$ and $\Lambda^2$ for the two hypotheses $\mathcal{H}^1$ and $\mathcal{H}^2$.

To determine the likelihood of the expectations $\Lambda^1$ and $\Lambda^2$ with respect to the observations, we set up the vector of observations $\Omega$ as

$$\Omega := \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix}, \omega_i := \omega_i(c_i), c_i \in \mathcal{C}, n = |\mathcal{C}|$$

where $\omega_i$ denotes the observation (number of earthquakes) in bin $c_i$. For computing the likelihoods, we need a suitable test statistic. Because we are testing Poissonian models \cite{Gardner1974, Reasenberg1985, Wyss2000}, the test statistic is given by the Poissonian probability density function (also called the probability mass function) $p$

$$p(\omega|\lambda) = \frac{\lambda^\omega}{\omega!}e^{-\lambda}$$

This function gives the likelihood of an observed number $\omega$ assuming an expected number of $\lambda$. The log-likelihood $L$ is the logarithm of this function

$$L(\omega|\lambda) = -\lambda + \omega \log \lambda - \log \omega!$$

For a bin $b_i$ with observation $\omega_i$ and expectation $\lambda_i^j$ of a hypothesis $\mathcal{H}^j$ the log-likelihood is given as

$$L(\omega_i|\lambda_i^j) = -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i!$$
The joint log-likelihood $L^j$ of all bins is given as the sum of the single log-likelihoods $L$ of all bins

$$L^j = L(\Omega|\Lambda^j) = \sum_{i=0}^{n} L(\omega_i|\lambda^j_i)$$

where $n = |\mathcal{C}|$ is the number of bins. The log-likelihood-ratio $R$ of the two models is defined as

$$R = L^2 - L^1 = L(\Omega|\Lambda^2) - L(\Omega|\Lambda^1)$$

where $L^2$ denotes the joint log-likelihood of the hypothesis $H^2$ for the given observation and $L^1$ denotes the joint log-likelihood of the hypothesis $H^1$, respectively. Thus, a log-likelihood-ratio of $R < 0$ indicates a higher likelihood for the hypothesis $H^1$, whereas a log-likelihood-ratio of $R > 0$ indicates a higher likelihood for the hypothesis $H^2$.

The log-likelihood-ratio allows a first judgment of the models, but we need to know whether this result is accidental or significant and whether we can accept or reject any of the hypotheses $H^1$ or $H^2$ at a given significance level. Because we do not know an analytical solution for the probability distribution of our likelihood score, we derive it by Monte Carlo simulations. Therefore, we draw random numbers from an uniform distribution in the interval $[0, 1]$ for every bin and simulation run. With the inverse cumulative Poissonian probability density function and a given expectation $\lambda^j_i$, we can derive simulated observations $\hat{\omega}^j_i$ (simulated values are denoted with a hat) for each bin $c_i$ and each hypothesis $H^j$. This leads to a vector of simulated observations $\hat{\Omega}^j$ per simulation run

$$\hat{\Omega}^j := \left(\begin{array}{c}
\hat{\omega}^j_1 \\
\hat{\omega}^j_2 \\
\vdots \\
\hat{\omega}^j_n
\end{array}\right), \hat{\omega}^j_i := \hat{\omega}^j_i(c_i), c_i \in \mathcal{C}$$
Vectors of \( m \) simulation runs will be denoted \( \hat{\Omega}_1, \hat{\Omega}_2, \ldots, \hat{\Omega}_m \). Subscripts of \( \hat{\Omega} \) are the number of simulations.

How can we interpret the log-likelihood-ratio \( R \)? If \( R = 0 \), both hypotheses match the data equally well (or equally poorly). The question whether any result \( R \neq 0 \) is significant cannot be answered without simulations. Therefore, we assume \( H^1 \) to be correct and we create simulations \( \hat{\Omega}_k \) based on the expectation \( \Lambda^1 \) of the hypothesis \( H^1 \). With these simulated observations, we can compute log-likelihood-ratios

\[
\hat{R}^1_k = L(\hat{\Omega}_k^1|\Lambda^2) - L(\hat{\Omega}_k^1|\Lambda^1)
\]

obtaining the set \( R^1 = \{\hat{R}^1_1, \hat{R}^1_2, \ldots, \hat{R}^1_m\} \). Let \( \alpha^1 \) be the fraction of simulated values of \( \hat{R}_k^1 \) less than the observed \( R \).

\[
\alpha^1 = \frac{|\{\hat{R}_k^1 \leq R, \hat{R}_k^1 \in R^1\}|}{|R^1|}
\]

Then \( \alpha^1 \) is the probability that \( H^1 \) would be falsely rejected, if it were true. Thus \( H^1 \) can be safely rejected if \( \alpha^1 \) is small. We use a two-sided significance level of \( \alpha_0 \). Therefore, values between \( \alpha_0/2 \) and \( 1 - \alpha_0/2 \) support \( H^1 \), because it cannot be rejected based on the given significance level.

To create simulated observations, based on the expectation vector \( \Lambda^2 \) of the hypothesis \( H^2 \), we repeat this computation to get the set \( R^2 \) of log-likelihood-ratios \( \hat{R}_k^2 \). Here again, we define \( \alpha^2 \) as the fraction of simulated values of \( \hat{R}_k^2 \) less than the observed \( R^2 \). A low value of \( \alpha^2 \) then supports the hypothesis \( H^1 \). The advantage of this approach is its symmetry with respect to the models. When swapping \( H^1 \) and \( H^2 \), simply \( \alpha^1 \) and \( \alpha^2 \) are swapped.

Neither \( H^1 \) nor \( H^2 \) receives any favoritism as the ‘null hypothesis’, if \( \alpha^1 \) and \( \alpha^2 \) are evaluated using the same criteria. The false rejection probabilities \( \alpha^1 \) and \( \alpha^2 \) can be compared against each other, as well as against the standard \( \alpha_0 \) (usually taken as 0.05). Thus we ‘prefer’ \( H^1 \) over \( H^2 \) if \( \alpha^1 \) exceeds \( \alpha^2 \). Even so,
we might safely reject neither hypothesis \((\alpha_0 < \alpha^1 < \alpha^2)\), \(H^2 \ (\alpha^2 < \alpha_0 < \alpha^1)\), or both \((\alpha^2 < \alpha^1 < \alpha_0)\). Similarly we would prefer \(H^2\) if \(\alpha^2 > \alpha^1\), even as we reject \(H^1\) only, or both \(H^1\) and \(H^2\).

### 7.3.1 Parameter Scan

Three main free parameters exist in our analysis: radius \(r\), minimum number of events per sample \(N_{\text{min}}\) and duration of the learning period. Their choice influences the outcome of the test. Choosing too large radii for the cylindrical volumes, for example, smoothes the \(b\)-value contrasts. Too small radii on the other hand cause problems in not supplying enough events for computation of \(b\)-values. Therefore, we have to define the parameters based on our knowledge on the spatial distribution of \(b\)-values and the activity in the volume of interest. Usually, an expert judgment can be made. To estimate the influence of the choice of free parameters, we performed the test for a large portion of the parameter space. We can compare the results by the calculated \(\alpha^1\)- and \(\alpha^2\)-values.

If the entire distribution of log-likelihood-ratios based on the simulation shows higher values than the log-likelihood-ratio of the real observation, we cannot judge anymore the different parameter combinations, using the \(\alpha\)-values, because \(\alpha \leq 0.001\) (in the case of 1000 simulation runs). Therefore, we define a measure \(\Delta^\sigma\) for the performance of a forecast as

\[
\Delta^\sigma = \frac{R - \overline{R}^1}{\sigma^1}
\]

where \(\overline{R}^1\) denotes the mean log-likelihood-ratio of the distribution based on simulations and \(\sigma^1\) its standard deviation, given as the second moment of \(R^1\). Thus, \(\Delta^\sigma\) is effectively the mean represented in multiples of \(\sigma^1\). \(\Delta^\sigma\) is only another representation of the significance level \(\alpha^1\). In cases where we cannot perform sufficient simulations to obtain an \(\alpha^1 > 0\), we can either treat the set \(R^1\) as normally distributed and compute the significance level under
this assumptions, or we use $\Delta_1^\sigma$, which is a more precise definition, without assumptions about the distribution of $R^1$. $\Delta_2^\sigma$ is defined accordingly. We use the $\Delta_\sigma$-measure for the relative performance of the different parameters for $b$-value computations used for the test hypothesis model. The relationship between $\alpha$ and $\Delta_\sigma$ cannot be derived analytically. However, comparing these values obtained by all our computations shows that for $\Delta_\sigma > 2.67$ all values of $\alpha$ are below 0.01, indicating that results with $\Delta_\sigma > 2.67$ are always highly significant.

We implemented a pseudo-prospective test, using the earlier part of an earthquake catalog as the learning period from which we compute the expectations of the models, and the latter part of the catalog as the observation period, which we want to compare with the expectations to judge the models. We focus on two time divisions: The middle of the available high quality data in 1992, and on 1996, when the first publication on Parkfield $b$-values proposed a suitable parameter set for analysis.

### 7.4 Data

We use the NCSN catalog from the Northern California Earthquake Data Center (NCEDC) spanning the period from 1981 to 2003 as prepared in Paper I [Schorlemmer et al., 2004]. The catalog contains 3780 events of $M \geq 1.3$. The area of interest is again defined as a 5 km wide cross-section extending from $121^\circ W, 36.4^\circ N$ to $120.2^\circ W, 35.64^\circ N$ (Figure 1 in Part I).

### 7.5 Results

We introduce the test results with a detailed description of tests performed with the same parameters as the stationarity tests shown in Figures 8 and 9 in Paper I. Then we present results of tests where we varied the parameters for $b$-value computation together with variations of the magnitude ranges of the
test, hereby testing 6 different ranges: $M \in [1.5, 7]$, $M \in [2, 7]$, $M \in [2.5, 7]$, $M \in [3, 7]$, $M \in [3.5, 7]$, and $M \in [4, 7]$. Because the extrapolation of $b$-values to higher magnitudes than used for determining the $b$-values is still a debated issue [Aki, 1987; Pacheco et al., 1992; Abercrombie and Brune, 1994; Sornette et al., 1996; Knopoff, 2000; Main, 2000], these tests will show whether varying $b$-values have an impact on determining future seismicity distributions for higher magnitudes ($M > 3$).

7.5.1 Test with cylindrical sampling volumes

For the first probabilistic forecast test, we use the same sampling parameters as in the first stationarity test carried out in Paper I (cylindrical volumes with $r = 5$ km, $N_{\text{min}} = 50$). The catalog was divided into periods 1981–1992 and 1992–2003. Here we use cylindrical instead of cuboidal sampling volumes to show the performance of both models in high spatial resolution. Hereby we can create a denser grid with spacing of $0.5$ km $\times$ $0.5$ km instead of the $2$ km $\times$ $2$ km spacing, which we use with cuboidal grid cells.

The spatial distribution of likelihood-ratios $R$ (Figure 7.1) shows that forecasts of the test hypothesis $H^1$ have a higher likelihood in the volumes with either high or low $b$-values shown in Figure 5 in Paper I. Especially the patch with the lowest $b$-value (asperity) (volume 4 in Figure 7.1) shows the largest likelihood difference between the two hypotheses. In addition, the hypothesis $H^1$ performs also better at the southern end of the creeping section with its high $b$-value. At 2460 out of 2990 computed nodes ($\approx 82\%$) the hypothesis $H^1$ achieves a higher likelihood. In three volumes, the hypothesis $H^2$ performs better (distance from $P_1$ of 20 km, 35 km, and 45 km). This test cannot replace the test with cuboidal sampling volumes, because events sampled with cylindrical volumes are counted multiple times, therefore biasing the overall result when summing over the log-likelihoods. Nevertheless, the results taken node by node are correct for a test designed to be performed on cylindrical
7.5 Results

sampling nodes, only the summation over all nodes would introduce a bias.

We now examine the test results in detail for 4 selected nodes. The log-likelihood-ratios show the largest negative value (that is, most favorable to spatially varying $b$-value) at the asperity marked as volume 4 in Figure 7.1. Even though the activity of this volume increased in the observation period compared to the learning period, the hypothesis $H_1$ performs much better (frame 4B in Figure 7.1). This is due to the very low and stationary $b$-value of 0.56 in the learning period and 0.59 in the observation period, causing the hypothesis $H_2$ ($b = 0.89$) to significantly underestimate the number of larger events. In volume 3, situated between the two $b$-value anomalies, the asperity and the southern end of the creeping section, the $b$-value is averaged ($b = 0.88$) and therefore close to the overall value $\bar{b} = 0.89$ of the entire catalog. This creates virtually identical expectations of both hypotheses, resulting in a log-likelihood-ratio of $R \approx 0$ (frame 3B). In volumes 1 and 2 we can identify changes of $b$-values around the overall $\bar{b}$-value. This favors the hypothesis $H_2$, as can be seen in frames 1B and 2B, respectively.

The results of the same test for a learning period of 1981–1996 and an observation period of 1996–2003 are shown in Figure 7.2. As in the previous test, the hypothesis $H_1$ performs better in the creeping section and the asperity. In the volumes 1 and 4 marked in Figure 10 in part I, the $b$-value shows non-stationary behavior. Consequently, the hypothesis $H_2$ performed better in the probabilistic forecast test. In volume 1 the $b$-value changed from 0.71 to 0.97, thus around the overall $b$-value of $\bar{b} = 0.89$. A similar behavior of the $b$-value can be detected in volume 4, changing from 0.71 to 1.21. In total, 2682 out of 3920 computed nodes show a log-likelihood-ratio of $R < 0$ ($\approx 68\%$), thus a higher likelihood for the hypothesis $H_1$. 
Figure 7.1: Results of the probabilistic forecast test with learning period 1983–1992 and observation period 1992–2003, using cylindrical volumes. (Center frame) Log-likelihood-ratio $R$ of the two models. Negative and positive ratio indicates higher likelihood for hypothesis $H_1$ and $H_2$, respectively. Circles mark volumes 1–4. (A frames) Non-cumulative frequency-magnitude distributions of volumes 1–4. Black squares mark the distribution of the observation period, the red lines mark expectations $\lambda^2$ of hypothesis $H_2$, the blue lines mark expectations $\lambda^1$ of hypothesis $H_1$. (B frames) Cumulative log-likelihoods as function of magnitude. The red lines mark the log-likelihood of hypothesis $H_2$, the blue lines mark the log-likelihood of hypothesis $H_1$. 

<table>
<thead>
<tr>
<th>Distance [km]</th>
<th>Depth [km]</th>
<th>Log-likelihood</th>
<th>Number of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10^5</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>-20</td>
<td>10^3</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>-40</td>
<td>10^2</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>-60</td>
<td>10^1</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>-80</td>
<td>10^0</td>
</tr>
</tbody>
</table>

Note: The figures depict the results of the probabilistic forecast test, showing the log-likelihood-ratio $R$ for the two models, $H_1$ and $H_2$. The graphs illustrate the non-cumulative frequency-magnitude distributions and cumulative log-likelihoods as functions of magnitude for different volumes.
7.5 Results

Figure 7.2: Results of the probabilistic forecast test with learning period 1983–1996 and observation period 1996–2003, using cylindrical volumes. Log-likelihood-ratio $R$ of the two models. Negative and positive ratio indicates higher likelihood for hypothesis $H^1$ and $H^2$, respectively.

7.5.2 Test with cuboidal sampling volumes

We compare the spatial distribution of the test results for a learning period of 1981–1992 and an observation period of 1992–2003 using cuboidal and cylindrical volumes (Figure 7.3). The first impression is that the patches have paler colors, compared to the result using cylindrical volumes. This is due to lower number of events per node, thereby generating lower absolute values of log-likelihoods and, consequently, relatively lower log-likelihood-ratios. However, the distribution of blue and red patches is similar to the one resulting from the test with cylindrical volumes. Results differ only at a few nodes. To pinpoint the impact of differently shaped testing volumes and sampling parameters, we select a pronounced example of differing results: In the left frames of Figure 7.3 the test results of the selected volume (marked by a square in the bottom frame) are displayed. This node shows a higher likelihood of the hypothesis $H^2$ when testing using cuboidal volumes (Figure 7.3 frame RB). This outcome is dominated by the low number of events of magnitudes $M < 2$ and the comparatively high number of events with magnitude $M \geq 2.4$. Even though the hypothesis $H^1$ gives a lower expectation for events with magnitudes $M > 2.5$, the likelihood is dominated by the occurrence of events with magnitudes $M \leq 2.5$. The log-likelihood-ratio is not changing much for events of magnitudes $M > 2.5$. Looking at the result obtained using cylindrical volumes shows a different picture. Here the hypothesis $H^1$ shows a higher likelihood. It can be seen that in the case of cylindrical volumes the event...
Figure 7.3: Results of the probabilistic forecast test with learning period 1983–1992 and observation period 1992–2003, using cuboidal volumes (size and spacing: 2 km × 2 km). (Bottom frame) Log-likelihood-ratio $R$. Negative and positive ratio indicates higher likelihood for hypothesis $H^1$ and $H^2$, respectively. (A frames) Non-cumulative frequency-magnitude distributions of the selected volume (marked with a black rectangle in the bottom frame). Black squares mark the distribution of the observation period, the red line mark expectations $\lambda^2$ of hypothesis $H^2$, the blue line mark expectations $\lambda^1$ of hypothesis $H^1$. (B frames) Cumulative log-likelihoods as function of magnitude. The red lines mark the log-likelihood of $H^2$, the blue lines mark the log-likelihood of $H^1$. Frames RA and RB show the results using a cuboidal volume as marked in the bottom frame, frames CA and CB the results using cylindrical volumes as shown in Figure 7.1

distribution aligns itself with the expectations marked by the colored lines. It has a sufficient number of events to be shaped according to the Gutenberg-Richter distribution with a local $b$-value, thereby supporting $H^1$. In the case of cuboidal volumes no alignment according to the proposed Gutenberg-Richter distributions is present due to the small number of events per cell.

The overall result of the test on disjoint cuboidal volumes is that the sum over all log-likelihood-ratios of all nodes, is $R = -30.13$. This value indicates, that the hypothesis $H^1$ performs better than the hypothesis $H^2$, but is this a significant difference? Can we accept hypothesis $H^1$ and reject hypothesis $H^2$? To address this question, we investigate the $\alpha^1$- and $\alpha^2$-values obtained
Figure 7.4: Results of the likelihood-ratio test (A) for the periods 1983–1992 versus 1992–2003, and (B) for the periods 1983–1996 versus 1996–2003. The vertical line indicates the observed log-likelihood-ratio $R$. The gray patches indicate the rejection bars for a two-sided significance level of 0.05. The dashed and solid curves indicate the cumulative distribution $R^1$ and $R^2$, respectively. If either distribution touches one of the rejection bars, the corresponding hypothesis can be rejected.

from simulations (Figure 7.4). We can see that no log-likelihood-ratio $\hat{R}^2$ of the simulations based on the expectations of the hypothesis $H^2$ is smaller or equal the obtained log-likelihood-ratio $R$ of the test with real data, resulting in $\alpha^2 \leq 0.001$ (we performed 1000 simulations, therefore we know that $\alpha^2$ must be smaller than 0.001) or $\Delta^2 = 8.05$. For this parameter combination of earthquake sampling ($r = 5$ km, $N_{min} = 50$, dividing year 1992) we can reject the hypothesis $H^2$ at the given significance level of 0.05.

The distribution of simulated log-likelihood-ratios $\hat{R}^1$ based on hypothesis $H^1$ lies closer to the line of the observed log-likelihood-ratio $R$ (Dashed line in Figure 7.4A). This underlines the overall better performance of $H^1$, compared to $H^2$. However, given $\alpha^1 = 0.023$ we still can reject $H^1$ at the given two-sided significance level of 0.05. The result of this test is that the hypothesis $H^1$ shows a superior performance, but still does not explain the observed data in the period 1992-2003 sufficiently well to not be rejected.

A similar picture emerges in the probabilistic forecast test for a learning period of 1981–1996 and an observation period of 1996–2003 (Figure 7.4B). Here again, we observe $\alpha^2 < 0.001$, but $\Delta^2 = 5$, reflecting a 'weaker' rejection
of $H^2$, compared to the test with periods 1981–1992 versus 1992–2003. The log-likelihood-ratios $\hat{R}^i$ based on simulations of the expectations of hypothesis $H^1$ is again closer to the line indicating $R$, in fact it is intercepting it without touching the rejection bars. For this parameter combination and catalog division, we cannot reject $H^1$ based on the observation of $\alpha^1 = 0.931$, but we reject $H^2$.

### 7.5.3 Parameter space scan

Extending the two tests using cuboidal sampling volumes to different magnitude ranges allows us to investigate the magnitude range dependence of the test. We use six different magnitude ranges for the tests ($M \in [1.5, 7]$, $M \in [2, 7]$, $M \in [2.5, 7]$, $M \in [3, 7]$, $M \in [3.5, 7]$, and $M \in [4, 7]$). It can be seen that for most of the tested parameter combinations (different dividing years of the catalog, different magnitude range) the hypothesis $H^1$ performs better and that in most of the cases the alternative hypothesis $H^2$ can be rejected (Figure 7.5). When testing only higher magnitudes ($M \geq 3.5$) with a dividing year of 1996, we cannot reject $H^2$. When dividing the catalog in 1996, we cannot reject $H^1$ for any magnitude range tested. In the case of a dividing year of 1992, we can reject $H^1$, except for the higher magnitude ranges ($M \geq 3.5$). The alternative hypothesis $H^2$ can be rejected for almost every magnitude range. Only for a dividing year of 1996 and magnitudes of $M \geq 4$ the hypothesis $H^2$ shows a higher likelihood, however, both hypotheses cannot be rejected.

The development of the test results for different dividing years and magnitude ranges shows that for 91% of the parameter combinations the log-likelihood-ratio $R$ is negative, preferring hypothesis $H^1$ (Figure 7.6). Only a few patches (9%) in Figure 7.6, mainly for magnitude ranges with $M_{\text{min}} \geq 3.5$, show a higher likelihood for $H^2$. For dividing years of the catalog between 1985 and 1992, $H^1$ shows a higher likelihood but can be rejected. Only when
Figure 7.5: Results of the likelihood-ratio test (A frames) for the periods 1983–1992 versus 1992–2003, and (B frames) for the periods 1983–1996 versus 1996–2003. The gray patches indicate the rejection bars, reflecting the two-sided significance level of 0.05. The dashed and solid curves indicate the cumulative distribution $R^1$ and $R^2$, respectively. If either of the distributions touches one of the rejection bars, the corresponding hypothesis can be rejected. The vertical line indicates the observed log-likelihood-ratio $R$. 
Figure 7.6: Results of the likelihood-ratio test for varying periods and varying minimum magnitude $M_{\text{min}}$ of testing ($M \in [M_{\text{min}}, 7]$). The four colors mark four different outcomes of the test. The darkest gray indicates that $H^1$ is consistent with the observation, while $H^2$ can be rejected at the given significance level. White patches do not occur.

testing magnitudes in the range of $M \in [3.5, 7]$, $H^1$ cannot be rejected and, additionally, $H^2$ can be rejected. For smaller minimum magnitudes of testing and division years from 1993 on, $H^1$ wins the test, based on its higher likelihood and the rejection of $H^2$, while $H^1$ cannot be rejected. All of the cases where $H^2$ is preferable are characterized by low expectations and few events for testing because of the high minimum magnitude or the short observation period. With this test we have shown that the results obtained, so far, are not caused by a biased choice of periods or magnitude ranges.

We performed these tests for a variety of sampling parameters. We varied the radius $r$ of the cylinders and the minimum number of events $N_{\text{min}}$ used for sampling earthquakes for $b$-value computations. The minimum number of events per node defines the lower limit of the number of events to calculate a $b$-value. Increasing the radius of the sampling volumes, while keeping the necessary minimum number of events per volume fixed, increases the number of computed nodes. Increasing the minimum number of events per node while keeping the radius fixed decreases the number of computed nodes. Additionally, decreasing the radii tends to separate the volumes of different seismic
behavior, while increasing the radii tends to mix volumes with different properties and therefore smoothes $b$-value contrasts. We also varied the range of tested magnitudes, changing the minimum magnitude from $M = 1.5$ to $M = 4$ in 0.5 magnitude unit steps. The upper limit was always set to $M = 7$. Its choice has no influence on the test. We tested every magnitude bin between the lower and upper magnitude limit in 0.1 magnitude unit steps.

Performing this parameter test with a division year of 1992 shows, that hypothesis $H^1$ performs better in most ($\approx 98\%$) of the parameter combinations (Figure 7.7). Only when using minimum magnitudes of $M_{\text{min}} \geq 3.5$ do a few ($\approx 2\%$) parameter combinations show higher likelihood for hypothesis $H^2$. As in the previous test with a division year of 1992, this test reveals that for most of the parameter combinations, $H^1$ can still be rejected. However, it can be seen that $H^1$ cannot be rejected for radii of $r \leq 6\ km$. This corresponds to the observation we made in Part I of this paper, where we found that sampling radii of $r \geq 6\ km$ are smoothing $b$-values at Parkfield. Also, the overall result corresponds to the trend (increasing minimum magnitude of testing prohibits rejection of $H^1$) visible in Figure 7.6.

Repeating this test with a division year of 1996 shows a different picture (Figure 7.8). Here, hypothesis $H^1$ wins for most of the parameter combinations, such that $H^2$ has to be rejected while $H^1$ cannot be rejected. When testing only the higher magnitude range, $H^2$ shows the higher likelihood, but $H^1$ cannot be rejected.

### 7.6 Discussion and Conclusions

Probabilistic forecasting using spatially varying $b$-values improves the forecasting quality significantly at the Parkfield segment of the San Andreas fault, compared to forecasting with a spatially constant $b$-value. In many tests, $H^1$ (variable $b$-values) is preferred over $H^2$, even though $H^1$ could be confidently rejected.
Figure 7.7: Results of the likelihood-ratio test for the periods 1983–1992 versus 1992–2003 with varying radius $r$, minimum number $N_{\text{min}}$, and varying minimum magnitude $M_{\text{min}}$ ($M \in [M_{\text{min}}, 7]$). The four colors mark four different outcomes of the test. The darkest gray indicates that $H^1$ is consistent with the observation while $H^2$ can be rejected at the given significance level. White patches do not occur. White areas indicate parameter combinations for which no computation has been performed.
7.6 Discussion and Conclusions

\[ R < 0, H_1 \text{ consistent, } H_2 \text{ can be rejected} \]
\[ R > 0, H_1 \text{ cannot be rejected} \]
\[ R > 0, H_1 \text{ consistent, } H_2 \text{ can be rejected} \]

Figure 7.8: Results of the likelihood-ratio test for the periods 1983–1996 versus 1996–2003 with varying radius \( r \), minimum number \( N_{\text{min}} \), and varying minimum magnitude \( M_{\text{min}} \) (\( M \in [M_{\text{min}}, 7] \)). The four colors mark four different outcomes of the test. The darkest gray indicates that \( H_1 \) is consistent with the observation while \( H_2 \) can be rejected at the given significance level. White patches do not occur. White areas indicate parameter combinations for which no computation has been performed.
The strongest results supporting H\(^1\) over H\(^2\) have been found for sampling parameters which preserve the \(b\)-value distribution and do not smooth it. Here, the radius of the cylindrical sampling volumes plays the main role. The \(b\)-value distribution at Parkfield can be mapped best with radii of 4-5 km (Paper I). Smaller radii do not reveal additional information, at least not down to the smallest applicable radius of about 2 km. Decreasing the radius only decreases the coverage. Tests with a division year of 1992 (Figure 7.7) showed that we cannot reject H\(^1\) for radii of \(r \leq 5\) km if we use \(N_{\text{min}} \leq 100\). This result is in accordance with the proposed size of characteristic volumes at Parkfield according to Wiemer and Wyss [1997] and matches our observation in Paper I.

Testing only the higher magnitude ranges (\(M \in [3.5, 7]\) and \(M \in [4, 7]\)) shows that too few events happened at Parkfield in the catalog’s period to reveal significant results (Figures 7.6 and 7.8). The smaller the number of observed events, the closer the cumulative distributions \(\hat{R}\) to the observation \(R\) and the smaller their difference in likelihoods (Figure 7.5). In most of the cases, we are not able to reject either of the hypotheses. Still, H\(^1\) more often showed a higher likelihood. The problem with testing only higher magnitudes is the small number of earthquakes forecasted. If this number is too low, the range of log-likelihood-ratios based on simulated events will likely be in the range of the observed log-likelihood-ratio, thus preventing the rejection of any of the hypotheses. Successfully rejecting either hypothesis requires longer observation periods for testing (Figure 7.6).

Despite the period’s length, the cell size for sampling earthquakes for activity estimates (\(a\)-value) is also a relevant quantity. Here we used cell sizes of \(2 \times 2\) km. The larger the cells, the more events per cell can be sampled and the more stable the result will become. On the other hand, large cell sizes mean a coarser resolution of the seismicity forecast. Also, the larger the cells, the more the \(b\)-values are smoothed and the smaller the difference between the forecasts of the two hypotheses will become. To be able to reject one and ac-
cept the other hypothesis, we need sufficient cell sizes, large magnitude ranges, sufficient length of the observation period for getting seismicity distributions, which are aligned to the Gutenberg-Richter distribution. Only then, can $H^1$ be successful. This applies especially to the nodes where the differences between the overall $b$-value and local $b$-value is small. We have seen, that for the amount of data available at Parkfield, we can achieve significant results best with testing a magnitude range of $M \in [1.5, 7]$ (Figures 7.5 and 7.6). With a minimum magnitude of testing of $M = 3.5$ we reached the limit for significant results.

The results support the idea that non-stationary behavior of $b$-values is only favoring $H^2$, if the values scatter around the average $\bar{b}$-value; otherwise, $H^1$ performs better. In Paper I, we considered volumes to have stationary $b$-values, if the Utsu test [Utsu, 1992] did not show significant changes, but we also accepted smaller scattering, if the information of a relatively low or higher (below or above the average) $b$-value was preserved. This specification is supported by the spatial tests, where $H^2$ showed only higher likelihoods in volumes with either average $b$-value or large (from below to above the average or vice versa) changes in $b$-value between the learning and observation period.

Parkfield may be a special case because of its pronounced $b$-value contrasts. We consider our analysis a first successful case study that has to be extended to larger areas and larger scaling of seismic volumes. The testing routines developed in this paper are valid universally. Therefore, we will extend this study to all of California, consequently changing the size of the investigated volumes and introducing areas of different behavior. Especially, forecast testing on a grid spanning a map and not a cross-section is an important future task because Gerstenberger et al. [2001] showed that the $b$-value varies systematically with depth in California. The question will be whether these variations obscure the information in the $b$-values and thus minimize the effect on forecasts based on spatially varying $b$-values.

The method of testing probabilistic earthquake forecasts introduced in this
paper differs from past approaches in a number of ways. Jackson [1996] proposed a method of testing earthquakes in a binary way. This method judges the occurrence or non-occurrence of events for magnitude larger than a given threshold magnitude. It is suitable for testing larger events only. For testing lower magnitudes, we need to judge the number of observed events by computing the likelihoods of occurrence for every magnitude bin down to a low threshold magnitude (e.g. $M_{\text{min}} = 1.5$). For this purpose, we developed an extension of the existing test methods. Although it seems that we used a standard likelihood-ratio test, we have to compute the significance by random simulations to construct a suitable probability distribution. In the standard approach, the significance level is given by the difference in degrees of freedom of the two models. Here, both models have zero degrees of freedom. Furthermore, none of the models is including the other, which renders the application of a standard likelihood-ratio test inappropriate.

We have scanned the parameter space in order to assure that a significant result is not obtained due to a lucky or even tuned choice of sampling parameters. Because we use pseudo-prospective testing only, parameter space scans are needed for guaranteeing an unbiased analysis. Taking all this into account, we conclude that forecasts ignoring the spatial distribution of $b$-values at Parkfield do not match the spatial future seismicity distribution. Using a constant $b$-value, only an overall estimate of the future seismicity distribution is possible. According to previous studies at Parkfield [Wiemer and Wyss, 1997] and at the San Jacinto-Elsinore fault zones [Wyss et al., 2000], we see the need to spatially determine the future seismicity for roughly estimating recurrence times of the asperity patches. Overall estimates mix the information contained in the different smaller volumes and, therefore, obscure the largely different seismic behavior of these patches.

We found only few ($\approx 10\%$ in total, none when testing magnitude ranges with $M_{\text{min}} \leq 3$) combinations of parameters of computing $b$-values and performing the test for which hypothesis $H^2$ showed a more likely forecast. The
main differences between the parameter settings are the degree of smoothing of the existing $b$-value contrasts, still keeping the extreme $b$-values persistent. This again underlines the importance of $b$-values and their impact on forecasting future seismicity and magnitude distributions. It also emphasizes the hypothesis that local recurrence times give solid estimates of future main shock locations and may contain valuable information about their recurrence.

Nevertheless, we also found limitations of our proposed testing procedure. First of all, for testing only higher and seismic hazard related magnitudes we need longer periods of instrumental catalogs. The 22 years of excellent catalog quality are not long enough for this task in this area because only eight events with magnitude $M \geq 4$ occurred in this period. Testing forecasts with observation periods of a few years may not show any significant preferences for either model because the amount of testable data is too small. Consequently, we cannot counter the argument that the size distribution of small events may not be connected with the occurrence of larger events of $M \geq 5$. Such breaks in scaling have been proposed by a number of studies at a number of magnitudes; however, they have not been tested in a rigorous fashion. Our study cannot address this question in a statistically rigorous fashion, however, we point out that using spatially variable $b$-values result in recurrence times of $M = 6$ event in the Parkfield region that is consistent with the historical record [Wiemer and Wyss, 1997].

Using the parameter scans shown in Figures 7.6, 7.7, and 7.8, we are now able to select the free parameters such that they give a favorable forecast ($r = 5$ km, $N_{\text{min}} = 50$). In doing so, we create a single forecast model for the Parkfield region that contains no free parameters. We used as learning period the years 1981–2003. This model can now be tested in a prospective forward looking test, which is the final validation of any hypothesis. We have started prospective testing on January 1st, 2004. We automatically download the seismicity in the testing regions from the NCEDC, with a time lag of 7 days to allow final processing of the locations and magnitudes. The observed
seismicity of M1.5 and above is then tested against our forecast model, and also against a model that uses the overall $b$-value. Results of the likelihood-ratio test and their significance are displayed on automatically updated web-pages (www.seismo.ethz.ch).

7.7 Acknowledgments

The authors would like to thank the Northern California Earthquake Data Center (NCEDC) for providing us with catalog data, the Northern California Seismic Network, U.S. Geological Survey, Menlo Park for the NCSN catalog. The authors profited from discussions with M. Gerstenberger. We also want to thank J. Woessner, J. Ripperger, and G. Hillers for their useful hints. This paper is contribution number 1371 of the Geophysical Institute, ETH Zürich. This research was supported by the Southern California Earthquake Center. SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 787.
Chapter 8

Earthquake Likelihood Model Testing

Danijel Schorlemmer\textsuperscript{1}, Matt Gerstenberger\textsuperscript{2}, Stefan Wiemer\textsuperscript{1}, D. D. Jackson\textsuperscript{3}

\begin{enumerate}
\item Swiss Seismological Service, ETH Zürich, Zürich, Switzerland.
\item United States Geological Survey, Pasadena, USA.
\item Department of Earth and Space Sciences, UCLA, USA.
\end{enumerate}

In preparation for a special volume about RELM in Seismological Research Letters.
8.1 Abstract

The Regional Earthquake Likelihood Models (RELM) project aims to produce and evaluate alternate models of earthquake potential (probability per unit volume, magnitude, and time) for California. Based on differing assumptions, these models are produced both to test the validity of their assumptions and explore which models should be incorporated in seismic hazard and risk evaluation. Tests based on physical and geological criteria are useful but here we focus on statistical methods using future earthquake data only. We envision two evaluations: a self-consistency test, and comparison of every pair of models for relative consistency. Both tests are based on the likelihood ratio method, and both would be fully prospective (that is, the models are not adjusted to fit the test data). To be tested, each model must assign a probability or probability density to any possible event within a specified region of space, time, and magnitude. For our tests the models must use a common format: earthquake rates in specified “bins” with location, magnitude, time and in some cases focal mechanism limits.

8.2 Introduction

To predict the behavior of a system is the desired proof of a model of this system. Seismology cannot predict earthquake occurrence, however, it should seek for the best possible models to forecast earthquake occurrence as precise as possible. This paper describes the rules of an experiment to examine or test earthquake forecasts in a statistical way. The primary purposes of the tests described below are to evaluate physical models for earthquakes, assure that source models used in seismic hazard and risk studies are consistent with earthquake data, and provide quantitative measures by which the models might be assigned weights in a future consensus model or be judged as suitable for particular areas.
8.2 Introduction

To test models against one another, we require that forecasts based on them can be expressed numerically in a standard format. That format is the average rate of earthquake occurrence within pre-specified limits of hypocentral latitude, longitude, magnitude, and time. For some source models there will also be bins describing depth, the inclination of P-axis (axis of maximum compression), declination of P-axis, and inclination of the T-axis (axis of least compression). There will be only a few choices for all of these limits, so that the models to be tested are not separated into too many different categories. Forecasts specified in this way are clear, easy to archive, and comparable between models.

This kind of forecasts are grid based. Because the test is designed to examine grid-based models, fault-based models cannot compete without modifications. In order to test models against each other, it is required that they provide a comparable description of the system’s behavior. This rule is violated if we try to compare a grid-based with a fault-based model. The fault-based model provides detailed description of possible future main shocks along certain fault segments while the grid-based model fully describes the probability distribution for earthquake occurrence at any location. Therefore, we require a transformation of fault-based models into grid-based models by gridding and smoothing the available forecast data according to the error assessments and probability distributions of the forecasted events. This means, that a fault-based model should be able to give an average rate of earthquakes for a specified location (even off-fault), magnitude range and range of focal mechanism angles.

We envision two time frames for our tests, based on the common applications of forecasts. Quasi-stationary models assume that earthquake rates are relatively stable over about a year. Short-term models assume that rates vary from day to day because of stress changes and other variations possibly resulting from past earthquakes.

Quasi-stationary models are relevant for public policy, construction plan-
ning, and setting insurance rates and priorities for remediation, all of which require seismic hazard estimates valid for years or more. Thus the leaders of the RELM (Regional Earthquake Likelihood Models: www.relm.org) project decided early on to develop a suite of source models for earthquakes over magnitude 5.0 in southern California over a five-year period. Some quasi-stationary models are fundamentally time-dependent. For example, some renewal models assert that large-earthquake probability decreases substantially after a large event and only gradually recovers over a period of decades or centuries. We will evaluate the quasi-stationary forecasts once each year. We will also allow updates of the forecasts to exploit information gained from the yearly tests.

Short-term models are need for emergency response and public information announcements. These models incorporate probabilities that depend on time and distance from previous earthquakes, usually exploiting power-law time dependence evident in aftershock sequences. It is difficult to apply these models in any fixed time interval because of the implicit scale invariance. Because earthquake rates change so fast after an earthquake, only an automatic algorithm for updating rates is adequate to implement the full time-dependence of these models and allow for replays with revised data.

The methods for testing the forecasts will be the same for quasi-stationary and short-term models. In both cases the earthquake rates will be assumed to remain relatively constant over the test period, and the test will be based on the number of events forecasted, and observed, in each interval of location, magnitude, time, and sometimes focal mechanism.

8.3 Basic Ideas and Definitions

We refer to a model as a concept of earthquake occurrence, composed of theories, assumptions, and data. Models can be rather general and need not be testable in a practical sense. A hypothesis is a more formal, testable statement derived from a model. The hypothesis should follow directly from the
model, so that if the model is valid, the hypothesis should be consistent with data used in a test. Otherwise, the hypothesis, and the model on which it was constructed, can be rejected.

For tests described here, we treat earthquakes as point sources with eight parameters: hypocentral latitude and longitude, depth, magnitude, origin time, P-axis inclination, P-axis declination, and T-axis inclination. The last three describe the focal mechanism, and they would not necessarily be included in all forecasts. The P- and T-axes define orthogonal fault and auxiliary planes, but in general we cannot determine from cataloged data which of the two is the fault plane. Thus we will use in our tests the orientations of the P- and T-axes but not of the fault plane orientation. Also, depth information is not required in issued forecasts. If no depth information is given by a forecast, the depth range will be set to 0–30 km to allow for comparisons with models providing depths information.

We will use the CISN (www.cisn.org) part of the ANSS catalog (www.anss.org). It will provide consistent reporting of locations and magnitudes for all of California and it will contain focal mechanisms for events down to magnitude 3.7 on a regular basis. The errors of any parameter are given as $1\sigma$-bounds of a normal distribution.

Of course, earthquakes are too complex to be fully described as point sources with eight parameters. Some earthquake models, especially those based on active faults, describe likely rupture length, area, end points, asperity location, etc. However, at this stage we use only the eight hypocentral parameters and their full error distribution because other qualities are not precisely defined nor consistently reported in earthquake catalogs. Adopting the eight parameter description means that fault-based models must be adapted to express probable hypocentral locations, magnitudes and focal mechanisms, thus become a grid model through smoothing the fault-based information onto a grid.

Two timeframes for testing are representing the two different classes of
models. Quasi-static models have to issue their forecasts for a 5 years period starting January 1st, 2005. Tests will be undertaken every year after an evaluated version of the catalog is available. We do not want to perform these tests with preliminary catalog data. The second timeframe applies to the short-term models. Here we test on a daily basis, also starting January 1st, 2005. The tests will be performed with the preliminary catalog as soon as all events of each particular day are processed and the catalog is made available.

For the quasi-stationary models, we need to decluster the catalog. Otherwise, aftershock sequences are dominating the results, obscuring the desired information about the model’s performance. This procedure is strongly debated and we do not foresee an agreement of every participant on a particular declustering algorithm and its necessary parameters. Therefore, we decided to use the best performing short-term model as a declustering algorithm for the final evaluation of quasi-stationary models. Hereby, we are going to decluster the final catalog with the newly yielded knowledge about event dependences and evaluate the models with this declustered catalog. Only for preliminary result we will decluster the catalog with the method by Reasenberg [1985]. The short-term models will be evaluated on undeclustered catalogs.

Each participant has to deposit the algorithm used for creating the forecasts, so that the computations can be performed as soon as catalog data is available. Especially for time-dependent models, replays with evaluated catalog data are necessary and, therefore, it must be possible to generate the according forecasts at any time automatically. The algorithms will be freely available and part of the OpenSHA-framework (www.opensha.org).

In the RELM project we express a hypothesis as a forecast, which we define as a vector of earthquake rates corresponding to the specified bins. Any bin is defined by intervals of the location, time, magnitude, and focal mechanism, thus a multi-dimensional interval. The resolution of a model corresponds to the bin sizes. The smaller the bins, the higher the resolution.

From the rates specified in each forecast we calculate a vector of "expec-
8.3 Basic Ideas and Definitions

...ations”, or expected number of events within the time interval for all bins, each element of the vector corresponding to a particular bin. The expected number is just the earthquake rate multiplied by the volume in parameter space of the bin. An expectation need not be a whole number nor must it be less than 1. The expectations are dimensionless, but they correspond directly to earthquake rates per unit area, magnitude, time, and possibly depth and orientation of angles because the bin sizes are specified.

In some texts the expectation is referred to as the ”prediction” or predicted number of events for the bin. While the term ”prediction” has a fairly standard meaning in statistics, it has a different meaning in earthquake studies. ”Earthquake prediction” usually refers to a single earthquake and implies both high probability and imminence. We consider ”earthquake prediction” as a special case of a forecast in which the forecast rate is temporarily high enough to justify an exceptional response beyond that appropriate for normal conditions. One can also adopt the definition of prediction by Main [1999]. We will avoid using the term prediction to avoid confusion. None of the forecasts are predictions in either sense nor are they meant to be. It is all about scientific testing.

The vector of expectations has to be compared with the vector of observations, based on the same binning, to score a given forecast. The observed number of events must be integers, and for the tests envisioned here they will usually be 0 or 1.

A useful measure of the agreement between a hypothesis and an earthquake record is the likelihood, defined as the joint probability of realizing the observed number of events, given the expectations in each of the bins. By joint probability we mean the probability of realizing the observed number in bin 1 and bin 2, etc. In all of the models proposed to date, the expectations for the various bins are assumed to be independent, in which case the likelihood is the product of the probabilities. The logarithm of the joint probability, sometimes called the ”log-likelihood” or ”log-likelihood score,” is simply the sum of the...
logs of the probabilities for all bins. If the expectations are not independent, the joint probability can be calculated as a product of conditional probabilities.

By comparing the observed events to a model’s expectations, we derive the hypothetical probability of the observed events occurring in our model. This probability is called the likelihood and is calculated assuming a Poissonian distribution of events in each bin. The Poisson model is strictly valid only if the forecast rate is truly constant during the test interval, but it is a good approximation when the rates do not vary much within the time interval.

The log-likelihood score depends on both the earthquake record and the forecast rates, and higher values imply better agreement between the two. But how large is large enough? We answer this question with two comparisons. First, in what we call the "consistency test", we compare the observed likelihood score with its expected value, assuming that the hypothesis is true. Second, in the "relative consistency test", we compare the observed likelihood with the value obtained using the same data, but forecast probabilities from another hypothesis. In this project, we will compare likelihood scores from all pairs of hypotheses defined on the same bins.

Besides these strict definitions on how to test forecasts against each other to match the defined needs (quasi-static and short-term models), we want this procedure to be the skeleton for any tests, modelers envision. We only want to set requirements on additional or changed rules: The test needs overall expectations of enough earthquakes in total to make it meaningful. If the chance for even only one earthquake to occur is very low, this test cannot be carried out yielding significant results because there is no chance that this test can be performed on a long enough time period. Also, we only want to accept expectation which can be tested based on reliable data. Therefore, tests on magnitudes far below the completeness level do not make any sense. The last requirement is the use of meaningful objective data. This means, we only want to allow data that is published on a regular basis based on unambiguous definitions.
8.3 Basic Ideas and Definitions

8.3.1 Definitions

Expectation The forecasted number \( \lambda \) of earthquakes for any given bin \( b \), equal to the earthquake rate times the binsize.

Model The methodology used to express a scientific idea.

Hypothesis A model with all functions, parameters, etc. completely specified. In the framework of RELM a hypothesis must generate a well defined forecast of future earthquakes including location, magnitude and time.

Forecast A set \( \Lambda \) of numerical estimates of the expected number of earthquakes in each bin, based on a hypothesis.

Bin A bin \( b \) is defined by intervals of the location, time, magnitude, and focal mechanism, thus a multi-dimensional interval.

Likelihood The joint probability of observing \( \omega_1 \) events in bin \( b_1 \) and \( \omega_2 \) events in bin \( b_2 \), etc., given the expectations \( \lambda_1, \lambda_2 \), etc.

Likelihood ratio The ratio of likelihood values for two forecasts evaluated using the same catalog, or two catalogs using the same forecast.

Test Contrary to the standard null hypothesis tests, where a test hypothesis competes against a given null hypothesis, we test each hypothesis against all other hypotheses. Hereby, each hypothesis acts as both a null and a test hypothesis in two tests against every other hypothesis of its category. This is necessary because it is possible that all tests between a hypothesis of a RELM model and the null hypothesis will result in rejection of the null hypothesis. However, significance between two competing RELM hypotheses may be much more difficult to establish. Therefore, without this test, the first model to test against the null hypothesis could become the de facto null hypothesis even if it does not forecast significantly better.
than later models. We simulate earthquake rupture catalogs and follow a similar method to the standard approach used in likelihood-ratio testing to obtain the significances of our results.

8.4 Computation

As outlined above, any hypothesis is expressed as a forecast of earthquake rates per specified bin. Any bin is defined by intervals of location (volume), magnitude, time, and focal mechanism angles, thus defining the resolution of a forecast. We denote bins with $b$ and all bins constitute the set $B$ defined as

$$B := \{b_1, b_2, \ldots, b_n\}, n = |B|$$

where $n$ is the number of bins $b_i$ in the set $B$.

A forecast of a model $j$ is issued as expectations $\lambda^j_i$ per bin $b_i$. We set up a vector $\Lambda^j$ of all expectations as

$$\Lambda^j = \begin{pmatrix}
\lambda^j_1 \\
\lambda^j_2 \\
\vdots \\
\lambda^j_n
\end{pmatrix}, \lambda^j_i := \lambda^j_i(b_i), b_i \in B$$

Expectations have units of earthquakes per year for quasi-static, and earthquakes per day for short-term forecasts. We also set up the vector $\Omega$ of observations $\omega_i$ per bin $b_i$ based on the same binning as the vector $\Lambda$ to be

$$\Omega = \begin{pmatrix}
\omega_1 \\
\omega_2 \\
\vdots \\
\omega_n
\end{pmatrix}, \omega_i = \omega_i(b_i), b_i \in B$$

Assuming that earthquakes are independent, the likelihood of observing $\omega$
events in a bin with an expectation $\lambda$ is the Poissonian probability $p$

$$p(\omega | \lambda) = \frac{\lambda^\omega}{\omega!} e^{-\lambda}$$

The log-likelihood $L$ for observing $\omega$ earthquakes at a given expectation $\lambda$ is defined as the logarithm of the probability $p(\omega | \lambda)$, thus

$$L(\omega | \lambda) = \log p(\omega | \lambda) = -\lambda + \omega \log \lambda - \log \omega!$$

and for a model $j$ at a bin $b_i$

$$L(\omega_i | \lambda^j_i) = -\lambda^j_i + \omega_i \log \lambda^j_i - \log \omega_i!$$

The joint likelihood is the product of the individual bin likelihoods, so its logarithm $L(\Omega | \Lambda^j)$ is the sum of $L(\omega_i | \lambda^j_i)$ over all bins $b_i$

$$L^j = L(\Omega | \Lambda^j) = \sum_{i=1}^{n} L(\omega_i | \lambda^j_i) = \sum_{i=1}^{n} -\lambda^j_i + \omega_i \log \lambda^j_i - \log \omega_i!$$

To compare the joint log-likelihoods of two models we compute the log-likelihood-ratio, defined as

$$R = L(\Omega | \Lambda^0) - L(\Omega | \Lambda^1) = L^0 - L^1$$

where $\Lambda^0$ denotes the vector of expectations of model $\text{H}^0$, $\Lambda^1$ denotes the vector of expectations of model $\text{H}^1$. $L^0$ and $L^1$ are the joint likelihoods of models $\text{H}^0$ and $\text{H}^1$, respectively. If the log-likelihood-ratio $R$ is less than 0, model $\text{H}^1$ provides a more likely forecast; if $R > 0$, model $\text{H}^0$ performs better.

8.4.1 Uncertainties in Earthquake Parameters

None of the earthquake parameters (location, focal time, etc.) can be estimated without uncertainties. Therefore, each parameter is accompanied by
its uncertainty distribution, given as the 1σ-bound of a normal distribution. Additionally, when testing stationary models, every event has assigned a probability $p_I$ of being independent. The lower $p_I$ the more likely the event is an aftershock.

To account for these uncertainties, we generate modified observations of each event. We draw random numbers according to the given normal distribution for each parameter of each earthquake to obtain a modified parameter. Also, we draw a random number from a uniform distribution between 0 and 1 to decide whether each event will be considered independent. If not, it will be deleted from the record. This gives a modified observation $\tilde{\Omega}$ (Modified observations are denoted with a tilde).

$$\tilde{\Omega} = \begin{pmatrix} \tilde{\omega}_1 \\ \tilde{\omega}_2 \\ \vdots \\ \tilde{\omega}_n \end{pmatrix}, \tilde{\omega}_i = \tilde{\omega}_i(b_i), b_i \in B$$

Repeating this procedure $s$ times yields a set of modified observations of the event record $\{\tilde{\Omega}_1, \tilde{\Omega}_2, \ldots, \tilde{\Omega}_s\}$, representing its uncertainty and its possible realizations. Although the uncertainty of each parameter is sufficiently described by the given normal distribution and the 1σ-bound, we need the set of modified observations for evaluation of significances, as described later. Parameter uncertainties cause events to be associated with different bins while event independence probabilities $p_I$ may change the total number of events in a record $\Omega$.

To represent the uncertainties of earthquake parameters in the results, we compute $s$ times the log-likelihoods $L^j$ and the log-likelihood-ratios $R$ using the modified observations $\tilde{\Omega}$, obtaining sets of log-likelihoods $\hat{L}^j = \{\hat{L}_1^j, \hat{L}_2^j, \ldots, \hat{L}_s^j\}$, total numbers of events $\hat{N} = \{\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_s\}$, and log-likelihood-ratios $\hat{R} = \{\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_s\}$. The log-likelihood of a model $j$ is the mean value of $\hat{L}^j$ and its standard deviation is given by the second moment of
Corresponding to that, the log-likelihood-ratio between two models is the mean value of $\bar{R}$ and the standard deviation is given accordingly.

### 8.4.2 Simulation and Evaluation

How can we know the expected value of the likelihood? Furthermore, if the likelihood for the observed earthquake record exceeds the expected value, how can we know whether the result is truly significant rather than accidental? To answer these questions, we need to derive a probability distribution for the likelihood score. The likelihood score is a statistic (i.e., a quantity measurable from any sample of the underlying forecast distribution), so it has its own probability distribution. In some simple cases the distribution of likelihood scores might be derived analytically from the rates in the forecast. However, the analytic solution is not practical here, so we derive the distribution of expected likelihood scores by simulation. That is, we draw random numbers according to the probabilities implied by the forecast to generate random earthquake records $\hat{\Omega}_k$ (simulated values are denoted with a hat) consistent with the forecast. Then we compute the likelihood score $\hat{L}^i_k$ for each of these simulated records, obtaining the set $\hat{L}^i = \{\hat{L}^i_1, \hat{L}^i_2, \ldots, \hat{L}^i_m\}$. From this distribution, we then can compute the significance as quantiles of the observed values compared to the distribution of simulated values.

To create the simulated observations, we draw random numbers from a uniform distribution in the interval $[0; 1]$, for every bin and every simulation run. We use this random number as the probability of the inverse cumulative Poissonian probability density function. This yields a simulated number of observed events $\hat{\omega}^i_i$ for each given bin $b_i$ assuming the expectations $\lambda^i_i$ of model $H^i$. Iterating through all bins creates a vector of simulated events $\hat{\Omega}^i$ based on
model $H^j$. 

$$\hat{\Omega}^j = \begin{pmatrix} \hat{\omega}_j^1 \\ \hat{\omega}_j^2 \\ \vdots \\ \hat{\omega}_j^n \end{pmatrix}, \hat{\omega}_j^i = \hat{\omega}_j^i(b_i), b_i \in B$$

We will denote multiple simulated vectors with $\hat{\Omega}_1^j, \hat{\Omega}_2^j, \ldots, \hat{\Omega}_m^j$. The subscript of $\hat{\Omega}$ is the number of the simulation.

**Data-consistency test or L-Test**

Consider first the data-consistency test, and assume that the hypothesis is true. This is showing whether the observed likelihood of the hypothesis is consistent with likelihoods obtained from simulations. A useful measure for this comparison are the quantile scores $\gamma_q$, or the fraction of simulated likelihood values $\hat{L}^j = \{\hat{L}_1^j, \hat{L}_2^j, \ldots, \hat{L}_m^j\}$, $m = |\hat{L}^j|$ less than the observed likelihoods $\tilde{L}_j^q$

$$\gamma_q^j = \frac{|\{\hat{L}_k^j \leq \tilde{L}_q^j, \hat{L}_k^j \in \hat{L}^j, \tilde{L}_q^j \in \tilde{L}^j\}|}{|\hat{L}^j|}$$

Here $\hat{L}_k^j$ denotes the log-likelihood of the $k$-th simulation and $\tilde{L}_q^j$ the log-likelihood of the $q$-th modification of the event record. Thus, we perform this computation $s$ times iterating through all modifications of the event record. This results in a distribution of quantile scores $\{\gamma_1^j, \gamma_2^j, \ldots, \gamma_s^j\}$. The quantile score $\gamma^j$ is the mean of this distribution and its standard deviation is given as second moment of this distribution.

If $\gamma^j$ is low, then the observed likelihood score is less than most of the simulated values, and the record is not consistent with the forecast. If the observed likelihood is in the middle of the simulated values, then it looks like it ought to, according to this one measure. A problem arises when considering results with a high $\gamma^j$. It means that the likelihood of the real observation is higher than the likelihood scores of the simulations. There are different scenarios under which this can happen. In a catalog with overall very low
expectations for any event to happen, the outcome of 0 events is the most likely one. Nevertheless, the sum of all given rates may exceed 1 or even higher numbers, expecting in total some events to occur. In this case, the outcome of 0 events would show a much higher likelihood than the average simulation because the simulation will reflect the total number of expected earthquakes, distributed over the cells. In contrast, a forecast with expectations exactly matching the observations would also have a too high likelihood compared to the likelihood scores of the simulations, because every simulation will in general add poissonian scatter to the expectations, thus generating observations that do not match the expectations any more. This will result in lower likelihoods for the simulations (high $\gamma^j$). As can been seen, a model should not be rejected based on too high likelihoods in the data-consistency test. We want to use this test only as an one-sided test, rejecting forecasts with significantly too low likelihood compared to the simulations. Matching or too high likelihoods are not giving a measure of goodness of match of expectations with observations. In most cases, however, too high likelihoods will indicate an inconsistent model. Therefore, we will additionally apply the number test (N-Test).

**Number test or N-Test**

The N-Test is an addition to the L-Test. It also tests the consistency of a model with the observation. Instead of comparing the observed likelihoods with likelihoods obtained from simulations, the N-Test compares the observed total number of events with the number of events in the simulated catalogs. Again, we use a quantile score $\delta_q$ for this comparison. The total number $N^j$ of expected events of a model $j$ is simply the sum over all expectations $\lambda_i^j$

$$N^j = \sum_{i=1}^{n} \lambda_i^j$$
while the total number of observed events $N$ is the sum over all $\omega_i$

$$N = \sum_{i=1}^{n} \omega_i$$

Simulating earthquake records according to the probabilities of model $j$, as done in the L-Test, leads to a set of total numbers of earthquakes records $\hat{N}^j = \{\hat{N}^j_1, \hat{N}^j_2, \ldots, \hat{N}^j_m\}$. The quantile score $\delta^j_q$ is defined as the fraction of $\hat{N}^j$ smaller than the observed number of events $\tilde{N}$.

$$\delta^j_q = \frac{|\{\hat{N}^j_k | \hat{N}^j_k \leq \tilde{N}_q, \hat{N}^j_k \in \hat{N}^j, \tilde{N}_q \in \tilde{N}\}|}{|\hat{N}^j|}$$

As in the L-Test, we require that the observed number of events $N$ is in the middle of the distribution $\hat{N}^j$ for a model $j$ to be consistent with the observation. Although this test is weaker than the L-Test, because only the total number of events instead of their location, magnitude and other parameters are tested, however, it is necessary to overcome the aforementioned problem in the L-Test. If a model is underpredicting the total number of events, it may not be rejected in the L-Test, but it will fail in the N-Test. If we can reject models in the L-Test, the N-Test is not necessary to perform. Only if a model cannot be rejected in the L-Test, the N-Test may show that a model is underpredicting events and can be rejected.

**Hypotheses Comparison or R-Test**

In many works (e.g. [Kagan and Jackson, 1994]), a "test hypothesis" is compared to a "null hypothesis." The null hypothesis is presumed simpler and the test hypothesis is only accepted if an observed statistic would be very improbable under the null hypothesis. Evaluating that probability requires knowledge of the distribution, usually estimated by simulation, of the relevant test statistic under the null hypothesis.

In many cases the null hypothesis is similar to the test hypothesis except
that it is missing one or more interesting features. Then rejecting the null hypothesis is equivalent to stating that a model must have the "interesting features" to fit the data well. A special case is when the null hypothesis is a constrained version of the test hypothesis. Both hypotheses are adjusted to fit the data, but the null hypothesis, being constrained, can fit no better than the test hypothesis, and usually it fits worse. Even if the null hypothesis were true, the test hypothesis would generally fit better, and the test measures whether that improvement is too large to be consistent with the null hypothesis. Again, the test is based on the distribution of a relevant test statistic under the null hypothesis.

It has to be mentioned here that additional parameters in a model do not correspond to additional degrees of freedom of models in our experiment, making the use of the Akaike Information Criterion (AIC) [Akaike, 1973, 1974] or any other related method (e.g. AICc, BIC, TIC, etc. [Chow, 1981]) impossible. In these methods, the models are judged based on their likelihoods but also on the number of degrees of freedom. In all models tested in the RELM framework, the number of degrees of freedom is 0 because every forecast is issued in advance of the observation period and is not readjusted during the observation period.

Our study differs from most textbook cases because all models we consider are fully specified in advance. Some hypotheses may be derived by using more degrees of freedom during the "learning" period, but these parameters are then fixed before the test, so all hypotheses have exactly the same number of free parameters: none. Furthermore we have no null hypothesis that we believe should be accepted over others in case of doubt. Nevertheless, we wish to exploit the methods used for testing against null hypotheses, without necessarily choosing a favorite a priori.

This test covers comparisons of models. In this kind of testing we cannot test any hypothesis against a null hypothesis and repeat this test subsequently with all hypotheses. It is most likely that the first hypothesis to test against a
'dumb' null hypothesis will win this test and the null hypothesis get rejected in favor of the tested hypothesis. Unfortunately, it will be also very likely that none of the remaining hypotheses will be able to beat the new null hypothesis at the given significance level. Therefore, we compare all hypotheses against the others with a different definition of the test statistic. In the test hypothesis vs. null hypothesis one uses a simple likelihood ratio

\[ R = L(\Omega|\Lambda^0) - L(\Omega|\Lambda^1) \]

and obtains the significance level \( \alpha \) by computing log likelihood ratios \( \hat{R}_k \) of simulated observation \( \hat{\Omega}_k \).

Now consider the comparative likelihood test, in which we commit to accept one hypothesis and reject the other. Suppose we use the same observed record to compute likelihood scores for two hypothesis, say \( H^1 \) and \( H^2 \). We call these likelihood scores \( L^1 = L(\Omega|\Lambda^1) \) and \( L^2 = L(\Omega|\Lambda^2) \), and let the log-likelihood-ratio \( R^{21} = L^2 - L^1 \). If \( R^{21} \) is large it would seem to support \( H^2 \), but how can we know whether the result is significant? The likelihood ratio is a statistic, as described above, and we can derive its probability distribution by simulation. We assume \( H^2 \) is correct, generate many synthetic records, and score each using both \( \Lambda^1 \) and \( \Lambda^2 \) separately (as we did for the observed record), obtaining the set \( \hat{R}^{21} = \{ \hat{R}_1^{21}, \hat{R}_2^{21}, \ldots, \hat{R}_m^{21} \} \) with

\[ \hat{R}_k^{21} = L(\hat{\Omega}_k^2|\Lambda^2) - L(\hat{\Omega}_k^1|\Lambda^1) \]

Let \( \alpha^{21} \) be the fraction of simulated values of \( \hat{R}_k^{21} \) less than the observed \( \tilde{R}^{21} \). Large values support \( H^2 \).

\[ \alpha^{21}_q = \frac{|\{ \hat{R}_k^{21} | \hat{R}_k^{21} \leq \tilde{R}^{21}_q, \hat{R}_k^{21} \in \hat{R}^{21}, \tilde{R}^{21}_q \in \hat{R}^{21} \}|}{|\hat{R}^{21}|} \]

We perform this computation \( s \) times iterating through all modifications of the event record, thus all \( \hat{R}_q^{21} \). This results in a distribution of quantile scores
The quantile score $\alpha_{21}$ is the mean of this distribution and its standard deviation is given as second moment of this distribution.

So far we have focused on $H^2$, but we should of course focus on $H^1$ as well. We derive the distribution $\hat{R}^{12}$ assuming that $H^1$ is correct by simulating records using $H^1$, then score them using both $\Lambda^1$ and $\Lambda^2$ separately as above. Let $R^{12} = L^1 - L^2$ for both observed and simulated catalogs, and compare the observed and synthetic using $\alpha^{12}$ (fraction of synthetics less than observed).

The advantage of this approach is its symmetry in respect to the models. When swapping $H^1$ and $H^2$, simply $\alpha^{21}$ and $\alpha^{12}$ are swapped. For interpretation of the outcome of this test, we want to use the result matrix or result table containing all computed $\alpha$-values. Consider a test run with $n$ hypotheses $H^1$, $H^2$, $\ldots$, $H^n$. Each of these hypotheses will play the role of a null hypothesis against the others as well as the role of a test hypothesis against the others. Performing the aforementioned test will lead to a set of $\alpha$-values and the result table, where the 'null hypotheses' are displayed on the left side and the 'test hypotheses' on top.

<table>
<thead>
<tr>
<th></th>
<th>$H^1$</th>
<th>$H^2$</th>
<th>$\ldots$</th>
<th>$H^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^1$</td>
<td>$\alpha_{11}$</td>
<td>$\alpha_{21}$</td>
<td>$\ldots$</td>
<td>$\alpha_{n1}$</td>
</tr>
<tr>
<td>$H^2$</td>
<td>$\alpha_{12}$</td>
<td>$\alpha_{22}$</td>
<td>$\ldots$</td>
<td>$\alpha_{n2}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ddots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$H^n$</td>
<td>$\alpha_{1n}$</td>
<td>$\alpha_{2n}$</td>
<td>$\ldots$</td>
<td>$\alpha_{nn}$</td>
</tr>
</tbody>
</table>

### 8.5 Evaluation

The described procedure tests the relative performance of each model against the others as well as the consistency of each model with the observation. The model with the smallest $\alpha$-values shows the weakest comparative performance. Still, any model needs to be consistent with the observation to be considered for future forecasting or hazard assessment.
Any result obtained using this procedure is a statement about any model’s performance over the full magnitude spectrum and the entire testing area. Therefore, we propose a more detailed investigation of any model’s performance by testing the spatial and magnitude performance. Hereby, the tests are carried out for every spatial bin separately and the significances of the model’s performances are computed for each bin. This results in maps from which areas can be identified for which certain models show a strong or weak performance. This kind of secondary tests can help understanding how and why models perform as they do.

Especially for hazard related studies, testing with magnitude constraints may also be very helpful. Hereby, only a portion of the magnitude spectrum is tested. Both, the magnitude and spatial tests, can be combined.

\section{Definitions}

Although we do not wish to enforce any rules that impede how a model generates a forecast, it is necessary to define several rules so that we are able to compare models. We define initial bin sizes and grids for the following variables: location, time, magnitude, and focal mechanism. The bins are not limited to the predefinition, however, they must be a multiple of the default as defined in the algorithms below.

\textbf{Test area} We define the test area in southern California as the following cells:
Figure 8.1: Testing area in southern California. The red line encloses the testing area ($1^\circ \times 1^\circ$ boxes).
Grid  The starting grid uses the above southern California definition with nodes centered at every whole degree. It is important to note that in a likelihood ratio test a coarser grid forecast can be resampled to a finer

<table>
<thead>
<tr>
<th>Latitude range</th>
<th>Longitude range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min  Max</td>
<td>Min  Max</td>
</tr>
<tr>
<td>36  37</td>
<td>-118 -117</td>
</tr>
<tr>
<td>36  37</td>
<td>-117 -116</td>
</tr>
<tr>
<td>35  36</td>
<td>-120 -119</td>
</tr>
<tr>
<td>35  36</td>
<td>-119 -118</td>
</tr>
<tr>
<td>35  36</td>
<td>-118 -117</td>
</tr>
<tr>
<td>35  36</td>
<td>-117 -116</td>
</tr>
<tr>
<td>35  36</td>
<td>-116 -115</td>
</tr>
<tr>
<td>34  35</td>
<td>-121 -120</td>
</tr>
<tr>
<td>34  35</td>
<td>-120 -119</td>
</tr>
<tr>
<td>34  35</td>
<td>-119 -118</td>
</tr>
<tr>
<td>34  35</td>
<td>-118 -117</td>
</tr>
<tr>
<td>34  35</td>
<td>-117 -116</td>
</tr>
<tr>
<td>34  35</td>
<td>-116 -115</td>
</tr>
<tr>
<td>34  35</td>
<td>-115 -114</td>
</tr>
<tr>
<td>33  34</td>
<td>-120 -119</td>
</tr>
<tr>
<td>33  34</td>
<td>-119 -118</td>
</tr>
<tr>
<td>33  34</td>
<td>-118 -117</td>
</tr>
<tr>
<td>33  34</td>
<td>-117 -116</td>
</tr>
<tr>
<td>33  34</td>
<td>-116 -115</td>
</tr>
<tr>
<td>33  34</td>
<td>-115 -114</td>
</tr>
<tr>
<td>32  33</td>
<td>-119 -118</td>
</tr>
<tr>
<td>32  33</td>
<td>-118 -117</td>
</tr>
<tr>
<td>32  33</td>
<td>-117 -116</td>
</tr>
<tr>
<td>32  33</td>
<td>-116 -115</td>
</tr>
</tbody>
</table>
grid without changing the results (see Appendix). However, it is required that finer cells do not overlap cells of a coarser grid. Therefore, any modeler using higher resolution must divide each bin into 100 new equally spaced bins and so on. Equivalently, any modeler using a lower resolution must resample the results to the minimum 1° by 1° resolution.

**Depth** Depth binning is implemented, however, the default is no binning. If necessary, bins are defined to be 10 km, 1 km, 0.1 km, etc. This test is considering only earthquakes with depths between 0 km and 30 km.

**Magnitude range** We require that all models provide a forecast of events between an $M_{\text{min}}$ and a bin containing the expectation for all earthquakes with magnitudes $M \geq M_{\text{max}}$. For quasi-static models, $M_{\text{min}} = 5$, and for time-dependent models, $M_{\text{min}} = 4$. For both types $M_{\text{max}} = 9$. The default binning is 0.01 units. It is allowed to use finer magnitude bins if necessary. In this case, the resolution in bins should be increased by a factor of 10.

**Focal mechanisms** Focal mechanisms are defined by 3 angles, inclination of P-axis (axis of maximum compression), declination of P-axis, and inclination of the T-axis (axis of least compression). Our initial binning of these angles is 30 degrees for each angle. The next binning step is 10 degrees for each angle. For further higher resolution we propose a factor of 1/10.

### 8.7 Examples

To illustrate possible testing scenarios and to give a feeling of test performances we have undertaken a few tests with models which potentially will be part of the RELM testing framework. We performed all tests (L-Test, N-Test, and R-Test) with two models. The first model $H^0$ is derived from the USGS 1996 model [Frankel et al., 1996] and the second model $H^1$ is a stationary
model designed by Helmstetter et al. [submitted]. We tested these models against the RELM catalog provided by Yan Kagan (http://moho.ess.ucla.edu/~kagan/relm_index.html) as the observation. The advantage of this catalog is the presence of an independence probability $p_I$ for each event. Thus, we can modify the observations by considering events as aftershocks or main shocks depending on the drawn random numbers. In one set of tests we use a catalog $C_M$ containing only events with $p_I = 1$, thus only main shocks. In the second set we use a different catalog $C_A$ which includes all events. Thus we create modifications of the observation based on the independence probabilities. We also tested using three different time frames, a 20-years period (1981–2000), a 70-years period (1932–2001), and a 5-years period (1998–2002). For any test, we set a two-sided significance level of 0.1.

8.7.1 Model $H^0$

The main model $H^0$ for the tests is a stationary grid-based model derived from the USGS 1996 model [Frankel et al., 1996]. We have derived the daily background rate from the long-term rate of this model, interpolated the characteristic fault information onto our grid, and extrapolated all rates $\Lambda^0$ into $\Delta M = 0.1$ magnitude bins down to magnitude $M = 5$ using the Gutenberg-Richter relationship. The characteristic fault information was interpolated to the given grid by: 1) projecting the fault to the surface, 2) distributing a large number of points over the surface projection of the fault, 3) counting the number of points that fell within a grid node, and 4) assigning the appropriate percentage of rates to each grid node, based on the percentage of overall points that fell within the node. The nodewise sums of expectations $\lambda^0_i$ over all magnitude bins for a 20-years period are shown in Figure 8.2.
8.7 Examples

8.7.2 Consistency of Model $H^0$ with the Observation

We first performed the L-Test to show whether $H^0$ is consistent with the observation for the 20-years period 1981–2000 using the catalog $C_M$ ($p_I = 1$). We performed 10000 simulations obtaining $\hat{\Omega}$. The result is shown in Figure 8.3A. As can be seen, the model is consistent with the observation. The curve of log-likelihoods based on simulated observations $\hat{\Omega}$ (green curve in Figure 8.3A) is intersecting the log-likelihood of the real observation (black vertical line) at $\gamma^0 = 0.389$.

The N-Test shows almost the same picture of consistency of the model with the observation ($\delta^0 = 0.314$). The total number of events in the simulated records range from 17 to 60 while 33 events have been observed. Therefore, we can state that model $H^0$ is consistent with the observations in the given period.

The result changes if we use catalog $C_A$ and modify the observations (10000

---

Figure 8.2: Nodewise sum of the expectations $\lambda^0_i$ over all magnitude bins for a 20-years period of the model $H^0$ (based on the USGS 1996 model [Frankel et al., 1996]).
Figure 8.3: Result of the data-consistency test of model H\(^0\) for the period 1981–2000 using catalog C\(_M\). The gray patches mark the rejection bars. (A) L-Test. The green curve indicates the cumulative distribution of log-likelihoods based on simulated events \(\hat{\Omega}\). The vertical black line indicates the log-likelihood of the real observation \(\Omega\) (\(\gamma^0 = 0.389\)). (B) N-Test. The green curve indicates the cumulative distribution of numbers of simulated events \(\hat{\Omega}\). The vertical black line indicates the number of observed events \(\Omega\) (\(\delta^0 = 0.314\)).

runs) based on the independence probability \(p_I\), thereby introducing potential main shocks into the catalog. Figure 8.4A shows that almost all log-likelihoods computed using simulated observations \(\hat{\Omega}\) are higher than any of the log-likelihoods computed using the modifications \(\tilde{\Omega}\) of the observation record. The \(\hat{L}\) range from -409.58 to -126.56 while simultaneously the \(\tilde{L}\) span only the range from -419.93 to -335.72. This results in a low \(\gamma^0 = 0.0018 \pm 0.0016\). The N-Test gives an explanation for this result (Figure 8.4B). Model H\(^0\) is underpredicting the total number of events (\(\delta^0 = 0.990 \pm 0.007\)), thus showing higher likelihoods in the simulations than expected considering the real observation. The number of events in the modified observation records ranges from 44 to 57, while the total expectation of the model is 36.52.

Therefore, we can state that model H\(^0\) is consistent with the observation in catalog C\(_M\) of events with \(p_I = 1\) while when including the uncertainties of events being main shocks or aftershocks, the model underpredicts the total number of events. This results in too high log-likelihoods.

### 8.7.3 Model H\(^1\)

The second model H\(^1\) is created by Helmstetter et al. [submitted]. It has the same total expectation as model H\(^0\) but a different spatial distribution of expectations. The nodewise expectations are shown in Figure 8.5. The expectations are more concentrated along the active faults and less smoothed
8.7 Examples

Figure 8.4: Result of the data-consistency test of model $H^0$ for the period 1981–2000 using catalog $C_A$. The gray patches mark the rejection bars. (A) L-Test. The green curve indicates the cumulative distribution of log-likelihoods based on simulated events $\hat{\Omega}$. The vertical solid black line indicates the the median of the log-likelihoods computed with the modifications of the observation record $\tilde{\Omega}$. The vertical dashed lines indicate the 5 and 95 percentile of log-likelihoods computed with the modifications of the observation record $\tilde{\Omega}$. (B) N-Test. The green curve indicates the cumulative distribution of numbers of simulated events $\hat{\Omega}$. The vertical solid black line indicates the the median of the numbers of events in the modified observation records $\tilde{\Omega}$. The vertical dashed lines indicate the 5 and 95 percentile of the numbers of events in the modified observation records $\tilde{\Omega}$.

over the area.

8.7.4 Consistency of Model $H^1$ with the Observation

Repeating the same tests for model $H^1$, shows a very similar picture, see Figure 8.6. Using catalog $C_M$, the model is perfectly consistent with the observation, while with catalog $C_A$, the model is underpredicting the total number of events. This results in too high log-likelihoods as we observed in the consistency test for model $H^0$.

To quantify the result, we computed the $\gamma$- and $\delta$-values of the tests. Performing the test with catalog $C_M$ gives $\gamma^1 = 0.531$. This shows the perfect consistency of the model with the observation. Because the total expectation of this model is the same as of model $H^0$, the N-Test gives the same result of $\delta^1 = 0.314$. Using catalog $C_A$ results in $\gamma^1 = 0.009 \pm 0.006$ and $\delta^1 = 0.990 \pm 0.007$.

8.7.5 Model comparison

We have seen that both models are consistent with the observation when using catalog $C_M$. Using catalog $C_A$, makes both models inconsistent with the data.
Figure 8.5: Nodewise sum of the expectations $\lambda_i$ over all magnitude bins for a 20-years period of the model $H^1$ [Helmstetter et al., submitted].

Figure 8.6: Result of the data-consistency test of model $H^1$ for the period 1981–2000 using both catalogs. See description of Figures 8.3 and 8.4. The red color indicates results of model $H^1$. (A) L-Test using catalog $C_M$ ($\gamma^1 = 0.389$). (B) N-Test using catalog $C_M$ ($\delta^1 = 0.314$). (C) L-Test using catalog $C_A$ ($\gamma^1 = 0.009 \pm 0.005$). (D) N-Test using catalog $C_A$ ($\delta^1 = 0.990 \pm 0.007$).
So far, we can not decide which model has a higher forecast performance. Here we want to investigate their comparative spatial performance using the R-Test.

Figure 8.7 shows the results using both catalogs. In both cases, log-likelihood-ratios based on the expectations of model H$^1$ are in the range of the observed log-likelihood-ratio, giving $\alpha^{10} = 0.179$ and $\alpha^{01} = 0.321 \pm 0.049$. On the same time, model H$^0$ can be rejected favoring the alternative model H$^1$ at the given significance level, because $\alpha^{01} = 0$ and $\alpha^{01} = 0 \pm 0$.

Evaluating our results for the given time period, we reject model H$^0$ in favor of model H$^1$ due to its spatial performance. We also state, that both models do forecast the total number of events equally well or badly. Using catalog C$^A$, both models fail to forecast the average seismicity while with catalog C$^M$ both model’s forecast match the average seismicity.

8.7.6 Tests over different time periods

We repeated this tests for two additional time periods of the catalogs. Figures 8.8 and 8.9 show the distributions of all tests and Tables 8.1 and 8.2 give the quantitative results. The period of 70 years shows a similar result as the previous results for the 20-years period. The L-Test and N-Test show consistency of both models with catalog C$^M$ (Figures 8.8A and 8.8B) and an
Figure 8.8: Results of all tests for the 70-years period 1932–2001. (Left column) Catalog \( C_M \). (Right column) Catalog \( C_A \). (A, B) L-Test. The median, 5, and 95 percentile lines are drawn in the color corresponding to the model. (C, D) N-Test. (E, F) R-Test. Quantitative results are listed in Table 8.1.

underprediction of events with catalog \( C_A \) (Figures 8.8C and 8.8D). Here again, we reject model \( H^0 \) in favor of model \( H^1 \) due to their spatial performance in the R-Test (Figures 8.8E and 8.8F).

In the 5-years period 1998–2002 the results look quite different. While both model overpredict the number of events in catalog \( C_M \) (Figure 8.8C), they are consistent with catalog \( C_A \) (Figures 8.8B and 8.8D). The comparative R-Test again shows that model \( H^1 \) has a better forecasting performance than model

<table>
<thead>
<tr>
<th>Test</th>
<th>Catalog ( C_M )</th>
<th>Catalog ( C_A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Test</td>
<td>( \gamma^0 = 0.295 )</td>
<td>( \gamma^0 = 0.001 \pm 0.001 )</td>
</tr>
<tr>
<td></td>
<td>( \gamma^1 = 0.569 )</td>
<td>( \gamma^1 = 0.018 \pm 0.009 )</td>
</tr>
<tr>
<td>N-Test</td>
<td>( \delta^0 = 0.235 )</td>
<td>( \delta^0 = 0.974 \pm 0.013 )</td>
</tr>
<tr>
<td></td>
<td>( \delta^1 = 0.233 )</td>
<td>( \delta^1 = 0.974 \pm 0.014 )</td>
</tr>
<tr>
<td>R-Test</td>
<td>( \alpha^{01} = 0 )</td>
<td>( \alpha^{01} = 0 \pm 0 )</td>
</tr>
<tr>
<td></td>
<td>( \alpha^{10} = 0.019 )</td>
<td>( \alpha^{10} = 0.731 \pm 0.050 )</td>
</tr>
</tbody>
</table>

Table 8.1: Results of all tests for the 70-years period 1932–2001. Distributions are shown in Figure 8.8.
8.7 Examples

Figure 8.9: Results of all tests for the 5-years period 1998–2002. (Left column) Catalog $C_M$. (Right column) Catalog $C_A$. (A, B) L-Test. The median, 5, and 95 percentile lines are drawn in the color corresponding to the model. (C, D) N-Test. (E, F) R-Test. Quantitative results are listed in Table 8.2.

$H^0$. $H^0$ can be rejected in favor of $H^1$ at the given significance level.

This last tests over the 5-years period are a good test case for the forecast model testing in the RELM framework. They show, that we can distinguish the forecast capabilities of models after a 5–years period. This is further emphasized by the fact that the two models used here have the same total expectation of events. In the RELM framework, the models will not necessarily exhibit the same total expectation, making their differences among each other

<table>
<thead>
<tr>
<th>Test</th>
<th>Catalog $C_M$</th>
<th>Catalog $C_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Test</td>
<td>$\gamma^0 = 0.948$</td>
<td>$\gamma^0 = 0.771 \pm 0.074$</td>
</tr>
<tr>
<td></td>
<td>$\gamma^1 = 0.962$</td>
<td>$\gamma^1 = 0.816 \pm 0.054$</td>
</tr>
<tr>
<td>N-Test</td>
<td>$\delta^0 = 0.047$</td>
<td>$\delta^0 = 0.230 \pm 0.069$</td>
</tr>
<tr>
<td></td>
<td>$\delta^1 = 0.048$</td>
<td>$\delta^1 = 0.229 \pm 0.070$</td>
</tr>
<tr>
<td>R-Test</td>
<td>$\alpha^{01} = 0.002$</td>
<td>$\alpha^{01} = 0 \pm 0$</td>
</tr>
<tr>
<td></td>
<td>$\alpha^{10} = 0.128$</td>
<td>$\alpha^{10} = 0.719 \pm 0.114$</td>
</tr>
</tbody>
</table>

Table 8.2: Results of all tests for the 5-years period 1998–2002. Distributions are shown in Figure 8.9.
even bigger.

8.8 Discussion

This dry matter is the necessary evil to complete the next step of new generation hazard assessment. Multiple forecast models are available, each of them covering different aspects of the physics of earthquakes or their pattern of occurrence. To extend our forecasts abilities we must evaluate these models unprejudiced and unbiased. The only way to achieve this goal is to test the forecasts of all models in a truly prospective test against observed seismicity. During a 5-years period, the RELM project will undertake testing of a suite of forecasting models. The performance of all competing models will be determined and evaluated.

The primary target of the RELM project is defining existing uncertainties in seismic hazard assessment and identifying the research topics needed to reduce these uncertainties. They are not limited to only earthquake occurrence and distribution. However, without the knowledge of earthquake occurrence, further assessments lack a major contribution of hazard. It is needed to assess the implications of each model on seismic hazard and loss estimate. Even more, it is one aspect of seismic hazard which is readily recorded in a standardized manner, thereby already allowing prospective tests. These tests will allow to select a set of viable models describing the seismicity in California and to identify which models are exportable to other regions.

The tests described here are the statistical instruments for the aforementioned effort. Although, the tests cover data-consistency evaluation as well as spatial comparative performance tests, they lack the ability to judge models on only their performance for large ($M \geq 7$) events. This problem is not due to insufficient sophistication of the developed procedure but inherent to tests covering only a short period of time. 5 years are simply not enough time to make significant statements for $M \geq 7$ events. In the case of the two tested
models, the total expectation for any event of magnitude $M \geq 7$ is $0.436$. This means, that even less than one event of this magnitude range is expected in a 10-years period. If we would only test this high magnitude ranges, the result cannot be significant at all, assuming an average occurrence of this kind of events. This is a clear drawback for testing hazard related forecasting abilities. It is most likely that we will not be able to know which model has the highest performance in forecasting this kind of events. We will for sure have a winner, but it will not be significantly better than the other models, meaning that this result can easily be by chance and not due to the better forecasting generation.

Therefore, testing down to magnitude 5 for quasi-stationary models and down to magnitude 4 for time-dependent models is proposed. Although we know, that hazard related studies are mostly interested in magnitudes $M \geq 7$, this approach makes significant results very likely, as shown in the example tests. Considering aftershock hazard, the extension to lower magnitudes is justified.

We have to consider this test and its results as a small step towards physical understanding of earthquakes and their occurrence. This test cannot solve the problem of determining the hazard. Basically, hazard depends on forecasting ground motions or accelerations and not simply earthquake occurrence. Unfortunately, we cannot easily test these parameters. Testing our knowledge of earthquake occurrence is therefore only the first, however an ambitious and necessary step.
8.9 Appendix

8.9.1 Likelihood ratio independence on bin-sizes

Let $P$ be the likelihood for observing $x$ events for a given expectation (rate) $\lambda$:

$$\log P = -\lambda + x \log \lambda - \log x!$$

Let there be a cell $C$ with a given rate $\lambda$ and one observed event. The likelihood for this observation is

$$\log P = -\lambda + \log \lambda - \log 1 = -\lambda + \log \lambda$$

Now let’s divide the cell $C$ into $n$ equally sized subcells $C_1, C_2, \ldots, C_n$. Since the event can only happen in one of the subcells, the likelihood of the observation is:

$$\log P = 1(-\lambda^* + \log \lambda^* - \log 1) + (n-1)(-\lambda - \log 1)$$

Because

$$\lambda^* = \frac{\lambda}{n}$$

and $\log 1 = 0$, we can write the likelihood of the observation as

$$\log P = (-\frac{\lambda}{n} + \log \frac{\lambda}{n}) + (n-1)(-\frac{\lambda}{n})$$

Rearranged:

$$\log P = -\frac{\lambda}{n} + (n-1)(-\frac{\lambda}{n}) + \log \frac{\lambda}{n}$$

$$= n(-\frac{\lambda}{n}) + \log \frac{\lambda}{n}$$

$$= -\lambda + \log \frac{\lambda}{n}$$

$$= -\lambda + \log \lambda - \log n$$
The likelihood changed only by the term $\log n$. Thus, in the likelihood ratio this term will vanish because it does not depend on the $\lambda$ and the likelihood ratio will be the same for the case with one cell as well as for the case with $n$ cells.

Now let us assume $m$ observed events. The likelihood for the case of only one cell is

$$\log P = -\lambda + m \log \lambda - \log(m!)$$

Regardless of the distribution of these $m$ events over the given $n$ subcells, the likelihood will be

$$\log P = -\lambda^* + m \log \lambda^* - X$$

where $X$ is based on the original term $\log x!$ and reflects the distribution of the $m$ events of the $n$ cells. The likelihoods of all possible cases may differ but in the likelihood ratio the term $X$ vanishes, making the likelihood ratio the same as in the one-cell case.
Chapter 9

Conclusions

My thesis has two major parts: I developed and applied statistical methods and analyzed seismicity concentrating on $b$-values. These two parts are usually tied together as $b$-value analyses always imply rigorous statistical approaches to the data. Focusing on the physical meaning of $b$-values lead to generation of forecasting models for California based on spatially varying $b$-values, considering them being a major contribution to the heterogeneity of earthquake occurrence. For testing this forecast model against observed seismicity, I developed a likelihood testing suite which will also be used in the RELM framework in California to evaluate different forecasting models.

9.1 Statistical Methods and $b$-values

In Chapter 4, I investigated the seismicity of the subducting slab in the Tyrrhenian Sea. The limited amount of data in this area made it necessary to introduce new statistical methods to ensure the results obtained by $b$-value mapping. The high $b$-value anomaly at about 150 km depth indicates the location of slab dehydration. To resolve all doubts that the discovered anomaly is not an artifact or stems from errors in computing $b$-values, I tested the anomaly in multiple ways and established the significance level of the anomaly comparing it to randomly achieved anomalies using magnitude-permuted catalogs.
This work is showing how one can extract valuable information about seismicity parameters even in areas of low seismicity and strongly heterogeneous completeness.

Investigating more the physical meaning of $b$-values, I was searching for a dependence of $b$-values on earthquakes of certain types of faulting. In Chapter 5, I computed $b$-values for various sets of earthquakes differing in their focal mechanisms. Normal faulting events show the highest $b$-values while simultaneously thrust events have the lowest $b$-values and strike-slip events intermediate ones. This is a clear proof of the concept of $b$-values depending on differential stress.

As a second target of rigorous application of statistical methods, in Chapter 6 I investigated the stationarity of $b$-values along the Parkfield segment of the San Andreas fault. To answer the question if the previously found strong heterogeneities in $b$-values [Wiemer and Wyss, 1997] are stationary is the first step towards a seismicity model assuming $b$-values containing information about future frequency-magnitude distributions. I spatially applied the Utsu-test [Utsu, 1992] for different divisions of the catalog to quantify the percentage of stationary nodes. I have shown that most of the nodes do not exhibit significant changes in $b$-values over time, thus opening the way to a forecast model based on spatially varying $b$-values. This model preserves the information in anomalous patches.

9.2 Earthquake forecast testing

The model forecasting future seismicity based on spatially varying $a$- and $b$-values is the connecting point between $b$-value analyses and earthquake forecast model testing. In Chapter 7, I introduce this model as well as a likelihood-ratio test to test it along the Parkfield segment against a null hypothesis model which forecasts future seismicity only using spatially varying $a$-values but a constant $b$-value. In the test, I have show that the $b$-value model has a higher
discussion and outlook

In Chapter 8, I extended the aforementioned test. It now additionally includes data-consistency tests to show whether a model’s forecast is consistent with the observation. Also, error distributions of earthquake parameters are included in the tests. This testing method will be used in the RELM framework for testing a variety of forecasting models for California against each other and for investigating their consistency with the observation. This promising effort may lead to new understanding of earthquake physics and earthquake occurrence. It will be one basis for the next generation of earthquake hazard assessment.

9.3 Discussion and Outlook

Some limitations do apply to the work done. The forecast test at the Parkfield segment of the San Andreas fault was only done in a pseudo-prospective manner. Although I scanned the entire parameter space, the true test has to be prospective. Only prospective tests can prove the forecasting capabilities of any model right. Therefore, I implemented the real-time test, however its results are not yet significant due to the short testing period.

The test at Parkfield has shown the superiority of a model using spatially varying $b$-values. Still, the consistency of this model with the data has not
be proven. Considering that this model is a quasi-stationary, therefore not including any time-dependence of earthquake occurrence, it may very likely not be consistent with the undeclustered data. Nevertheless, the likelihood-ratio test as well as the data-consistency tests should be extended to all of California. Only this approach can show that using spatially varying $b$-values improves earthquakes forecasts and therefore hazard assessments in California.

This model can also lead to better understanding of earthquake occurrence in asperities. I have already shown for the asperity at Parkfield, characterized by a low $b$-value, that this model shows higher performance over a model using only a constant $b$-value. Extending this test to all of California will can prove the idea that asperities can be mapped by recurrence times, continuing the work of Wyss et al. [2000], Öncel and Wyss [2000], and Wyss and Matsumura [2002].

This approach should also be accompanied by differentiating the $b$-values according to focal mechanisms. This would lead a new model, which forecasts future seismicity using spatially varying $a$- and $b$-values for different styles of faulting. I am planning to contribute this model to the suite of quasi-stationary models in the RELM framework.
List of Figures

4.1 Location map of the study area in the Tyrrhenian Sea. . . . . . . 14
4.2 Cross-sections of the subducting slab. . . . . . . . . . . . . . . 18
4.3 $b$-value cross-sections of Tyrrhenian Sea computed for stability test. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
5.1 (A) $b$-value as a function of the rake $\lambda$ of events in southern California. The vertical bars indicate the standard error [Shi and Bolt, 1982]. The green, red, and blue lines mark the $b$-values of mainly normal, strike-slip, and thrust events, respectively. The gray line marks the average $b$-value. The circles at the top of the frame show the rake $\lambda$ used for computing the $b$-values ($\lambda = -90^\circ \pm \gamma$, $\lambda = 0^\circ \pm \gamma$, and $\lambda = 90^\circ \pm \gamma, \gamma = 20^\circ$). (B) $b$-value as a function of the range $\gamma$ of the rake $\lambda$ for three classes of events in southern California. The vertical bars indicate the standard error [Shi and Bolt, 1982]. Solid bars are used for samples with $N \geq 200$, dashed bars for samples with $200 > N \geq 100$. The green, red, and blue lines mark the $b$-values of normal, strike-slip, and thrust events, respectively. The gray line marks the average $b$-value. The circles at the top of the frame show the range of rake $\lambda$ used for computing the $b$-values ($\gamma = 15^\circ, 45^\circ, 75^\circ$). (Inset B) Circle explaining the rake values and the corresponding colors of the classes of events. (C) Frequency-magnitude distributions for pure normal (green) and pure thrust (blue) events of the SCSN and Harvard catalogs. (D) Same as (A) for the Harvard catalog. (E) Same as (B) for the Harvard catalog (Only considering rake $\lambda$ of the first nodal plane). (Inset E) Same as (E) but considering rakes of both nodal planes. . . . . . . . . . . . . . 27
5.2  $b$-value as a function of the range $\gamma$ of the rake $\lambda$ for three classes of events. The vertical bars indicate the standard error [Shi and Bolt, 1982]. Solid bars are used for samples with $N \geq 200$, dashed bars for samples with $200 > N \geq 100$. The green, red, and blue lines mark the $b$-values of normal, strike-slip, and thrust events, respectively. The gray lines mark the average $b$-values. (A) NCSN catalog. (B) NEID Kanto-Tokai catalog. Inset shows $b$-values computed without depth constrains. (C) NEID F-Net catalog.

6.1  Seismicity map of the Parkfield region in central California.

6.2  Magnitude of completeness $M_c$ as a function of time; Cumulative number of events.

6.3  Cumulative and non-cumulative frequency-magnitude distributions.

6.4  Distribution of $b$-values in the Parkfield segment of the San Andreas fault computed using different radii $r$.

6.5  Distribution of $b$-values in the Parkfield segment of the San Andreas fault.

6.6  Frequency-magnitude distributions of the two selected volumes.

6.7  Annual probabilities $P_r$ for one or more events of magnitude $M = 6$ at the Parkfield segment.


6.11 Maps of log-probabilities $P_b$ of the stationarity test for different catalog divisions.
6.12 Changes in $b$-values in the Parkfield segment of the San Andreas fault. .................................................. 62
6.13 Distribution of $b$-values in the Parkfield segment of the San Andreas fault. ........................................... 64
7.4 Results of the likelihood-ratio test. .................................................. 87
7.5 Results of the likelihood-ratio test. .................................................. 89
7.6 Results of the likelihood-ratio test for varying periods and varying minimum magnitude $M_{min}$ of testing. .................................................. 90
7.7 Results of the likelihood-ratio test for the periods 1983–1992 versus 1992–2003 with varying radius $r$, minimum number $N_{min}$, and varying minimum magnitude $M_{min}$. .................................................. 92
7.8 Results of the likelihood-ratio test for the periods 1983–1996 versus 1996–2003 with varying radius $r$, minimum number $N_{min}$, and varying minimum magnitude $M_{min}$. .................................................. 93
8.1 Testing area in southern California. .................................................. 119
8.2 Expectations for a 20-years period of the model $H^0$. ................. 123
8.3 Result of the data-consistency test of model $H^0$ for the period 1981–2000 using catalog $C_M$. ....................... 124
8.4 Result of the data-consistency test of model $H^0$ for the period 1981–2000 using catalog $C_A$. ....................... 125
8.5 Nodewise sum of the expectations $\lambda^i$ over all magnitude bins for a 20-years period of the model $H^1$ [Helmstetter et al., submitted]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
8.6 Result of the data-consistency test of model $H^1$ for the period 1981–2000 using both catalogs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .126
8.7 Result of the R-Test comparing the performance of models $H^0$ and $H^1$ for the period 1981–2000 using both catalogs. . . . . . . . .127
8.8 Results of all tests for the 70-years period 1932–2001. . . . . . . .128
8.9 Results of all tests for the 5-years period 1998–2002. . . . . . . .129
Bibliography


Main, I., Nature debate: Is the reliable prediction of individual earthquakes a realistic scientific goal?, [http://www.nature.com/nature/debates/earthquake/equake_contents.html](http://www.nature.com/nature/debates/earthquake/equake_contents.html), 1999.


Savelli, C., Two-stage progression of volcanism (8-0 Ma) in the central Mediterranean (southern Italy), *J. Geodyn.*, **31**(4), 393–410, 2001.


Bibliography


Wyss, M., K. Shimazaki, and S. Wiener, Mapping active magma chambers by


Curriculum Vitae

Danijel Schorlemmer

born November 23, 1969, Croatia

citizen of Germany

1989–1992: University of Göttingen, Germany, Physics

1992–1993: FU Berlin, Germany, Geophysics

1993–2000: Technical University Berlin, Germany, Geophysics


Acknowledgments

First of all, I want to thank Stefan Wiemer and Domenico Giardini for offering me the opportunity to write this thesis. Stefan should get very special thanks as being a superb advisor, offering the necessary freedom I needed for my research, and supporting me in numerous ways.

I would like to thank my referees Max Wyss, who introduced me to seismology during my master’s thesis and Dave Jackson, who offered me countless opportunities to discuss problems in statistical seismology.

How can I thank Matt Gerstenberger? Besides being a great colleague, he discussed almost every issue of my thesis with me in great detail. I profited from these discussions very much. Thanks Matt!

Special thanks to Alexander Körner (he knows why) and Jo-Chen.

Many thanks to everybody at the institute: Adrian, Andre, Cécile, Chrissy, Christoph, Daniel, Donat, Edi, Elisabeth, Fabio, Gabrielle, Gregor, Fabrizio, Ivo, Jan, Jérôme, Jochen B., Klaus, Manfred, Markus, Martin, Nico, Peter, Philipp, Rippi, Sarah J., Sarah H., Sarah S., Silvio, Sjonni, Stefan H., Stephan, Steve, Sybille, Thomas S., Thomas v. S., and Urs.

Thanks also to Ned Field, Lucy Jones, Egill Hauksson, Yan Kagan, Yehuda Ben-Zion and SCEC in California for their great hospitality.
Thanks to Karolina and Lucas for having great fun in the philosophy classes.

Thanks to Zürich for constantly foggy days in winter but great days in summer.

The times in Messina have been great. Many thanks to Giancarlo Neri and Antonino Mostaccio.

A million thanks to my family, Uta, Hannah, and Sarah. A thesis is always a team effort. They have been a great team!