Standardized Tests of RELM ERF's Against Observed Seismicity

Danijel Schorlemmer[1], Dave D. Jackson[2], Matt Gerstenberger[3], Stefan Wiemer[1]

[1] ETH Zürich, Swiss Seismological Service, Switzerland
[2] Department of Earth and Space Sciences, UCLA, USA
[3] Institute of Geological & Nuclear Sciences, New Zealand

Introduction

Why do we test?

- Agreement of a group
- Unbiased forward-looking test
- Fully specified in advance
- Rigorous test

What do we test?

- Comparative performance of ERF's
- Consistency of ERF's with observations

How do we test?

Examples from Parkfield

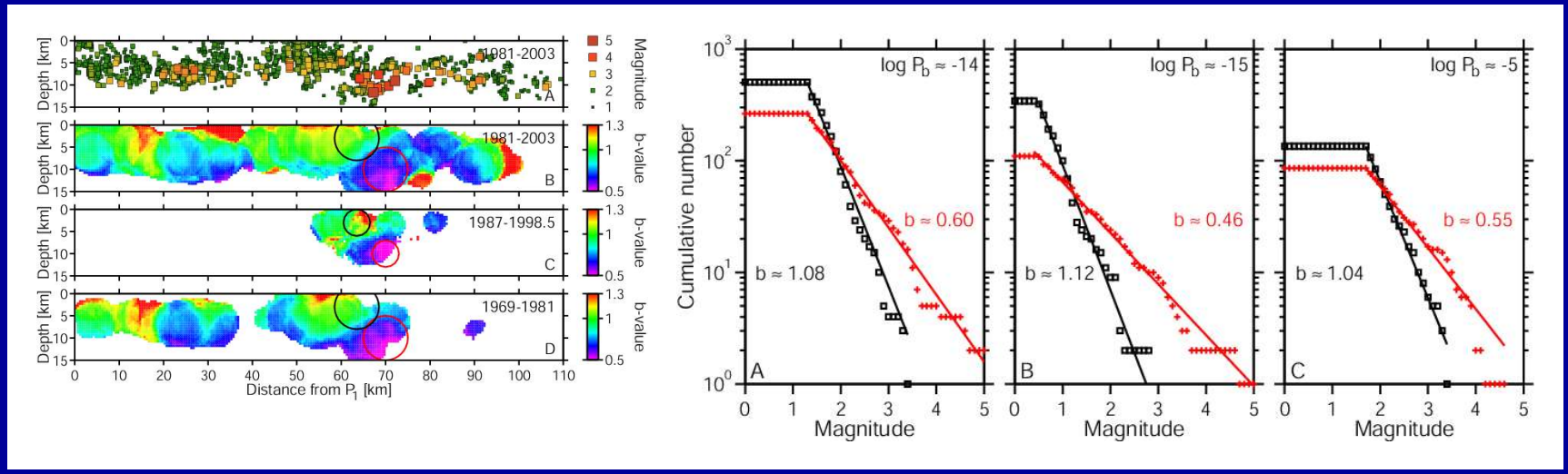
Example – Test

Model 1 (Variable b-value)

Forecast generated by extrapolating the frequency-magnitude distribution using spatially varying a- and b-values

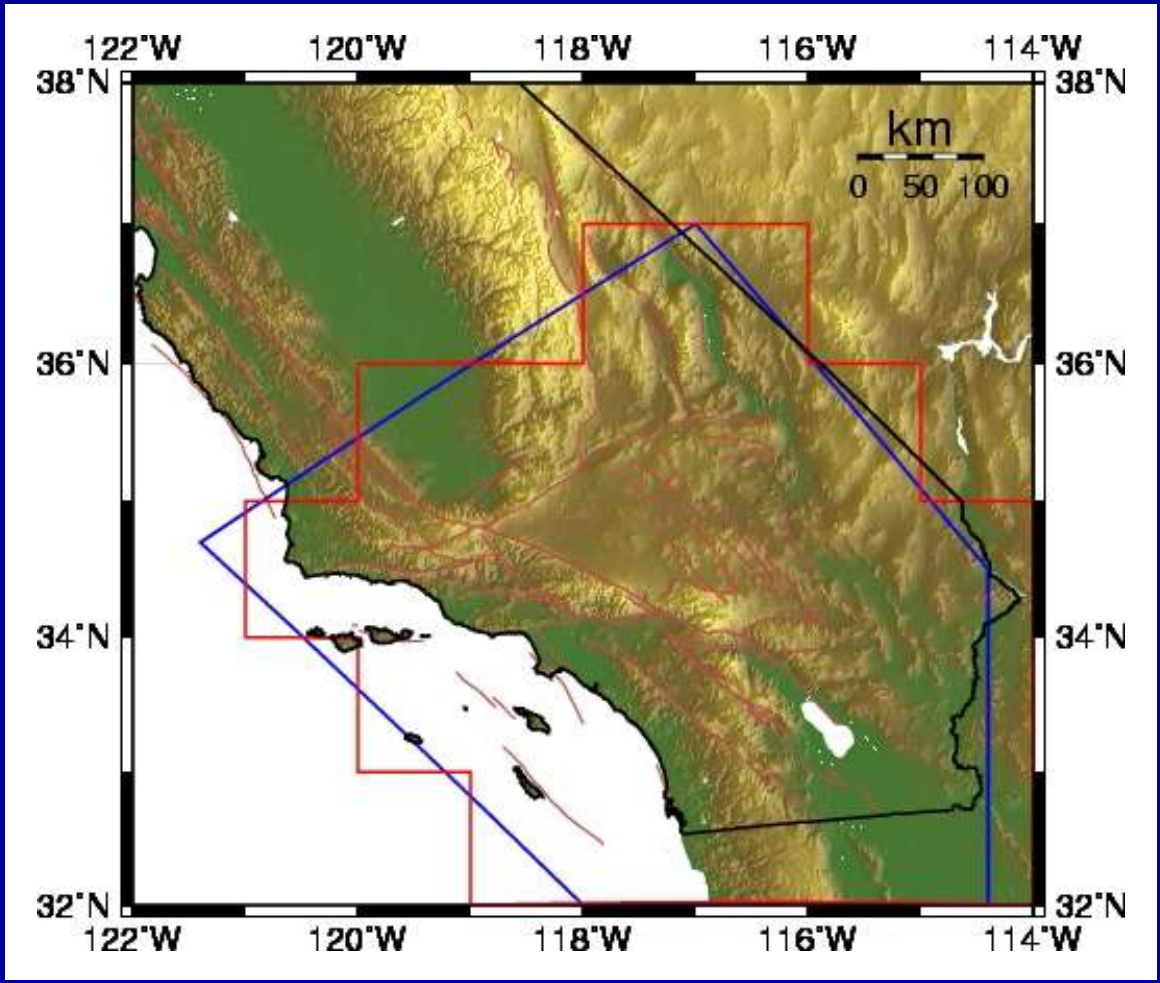
Model 2 (Constant b-value)

Same as model 1, but with regional constant b-value



Testing – Area

Testing area in California



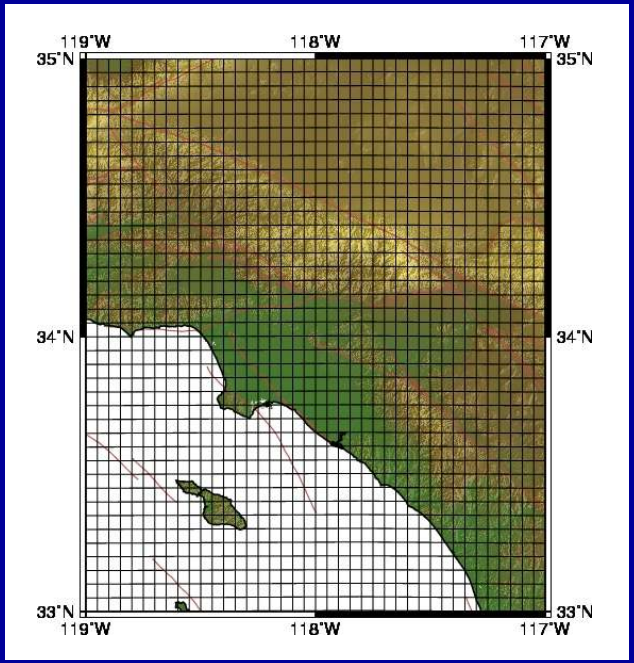
Bins

What is a bin?

A bin b_i defines a volume, magnitude range, period, and range of focal mechanism angles for which a forecast is issued

Our proposed default binning:

Lon/Lat	0.05°x0.05°
Depth	no binning: 0-30km
Magnitude	0.1
Focal M.	30°



Forecast

A forecast of a model j is defined as the expectations λ_i^j per bin b_i

$$\forall b_i | \lambda_i^j := \lambda_i^j(b_i)$$

The total expectation of a model j is the vector Λ^j

$$\Lambda^j = \begin{pmatrix} \lambda_1^j \\ \lambda_2^j \\ \vdots \\ \lambda_n^j \end{pmatrix}$$

Forecast

Forecasts are issued for

1 year for quasi-stationary models

1 day for time-dependent models

Forecasters provide expectations, i.e. numbers of events per bin, not programs

Observations

Observations ω_i (number of earthquakes) per bin b_i are a vector Ω

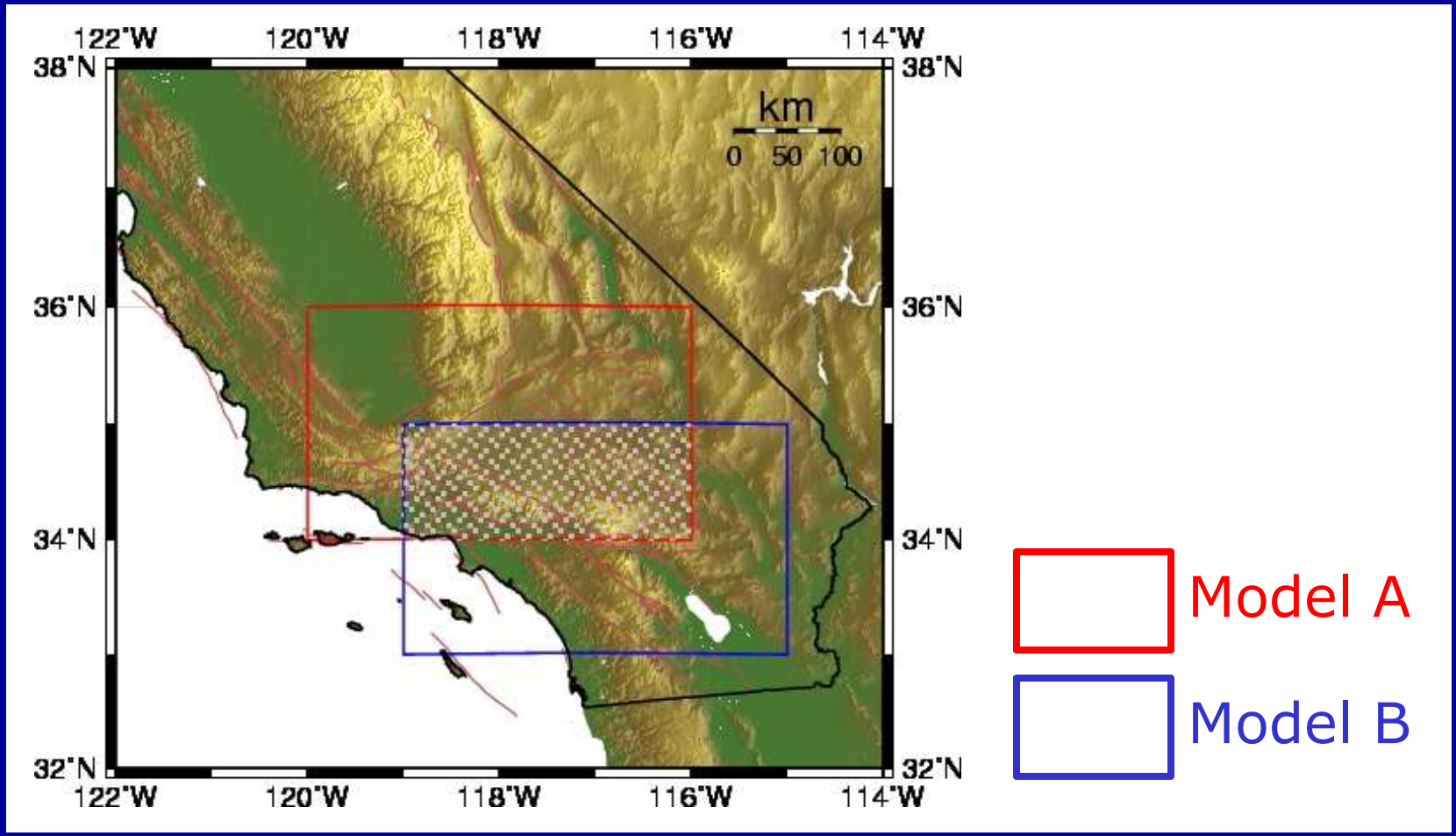
$$\Omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix}$$

We use 2 catalogs:

1. Y. Kagan's RELM catalog for hypocenter, magnitude, focal time
2. Harvard catalog for focal mechanisms

Limitations

We can only test in corresponding bins of two models



Agreements

Forecast is a vector of expectations (rates): quakes per year (or day) per bin

Forecasters provide expectations, not programs

All quakes count: no distinction between foreshocks, main shocks, and aftershocks

Bins of 0.05° x 0.05° x 0.1 magnitude units

Two main "menu items":
- Five year forecast of $M \geq 5$, no updates
- Five year forecast of $M \geq 4$, updated daily

Special orders ok if
- Multiple models use same bins, and
- Sufficient earthquakes for test

Testing – Likelihood

Likelihood is based on the poissonian probability distribution

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

Log-Likelihood of model j in bin b_i of the observation ω_i given the expectation λ_i^j

$$L(\omega_i|\lambda_i^j) = -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i!$$

Testing – Likelihood

The joint log-likelihood L^j of model j

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

Calculation example

Forecast bin			Expectation		Probabilities					
Bin	Region	Mag.	Ann. rate	λ	0	1	2	3	4	etc.
1	1	4.5-5.5	0.40	2.0	0.14	0.27	0.27	**0.18**	0.09	
2	1	5.5-6.5	0.04	0.2	**0.82**	0.16	0.02	0.00	0.00	
3	2	4.5-5.5	0.20	1	0.37	**0.37**	0.18	0.06	0.02	
4	2	5.5-6.5	0.02	0.1	**0.90**	0.09	0.00	0.00	0.00	
Likelihood L = log(0.18 * 0.82 * 0.37 * 0.90)										

Testing – Likelihood-ratio

The log-likelihood-ratio R of two models

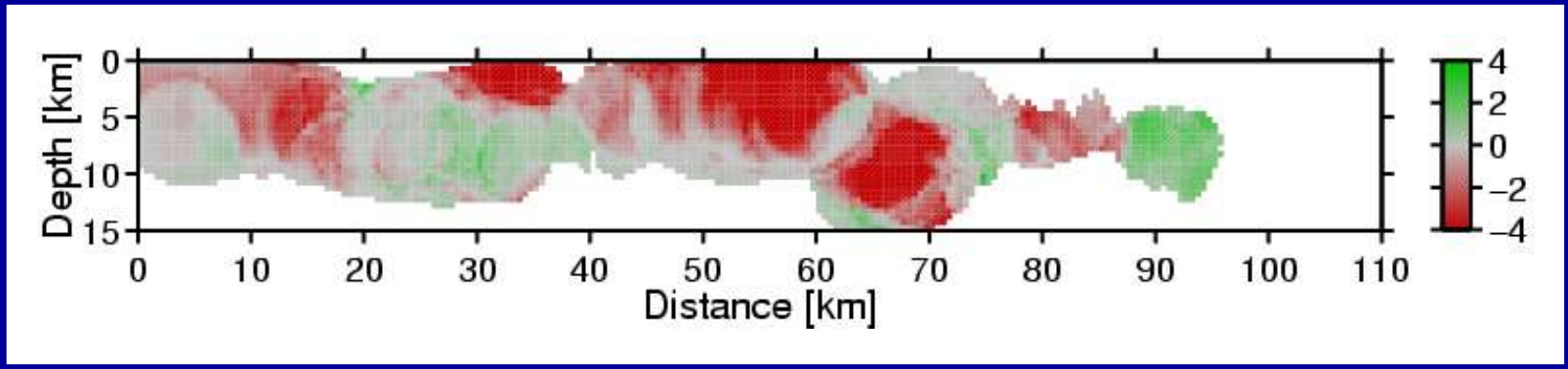
$$R = L^2 - L^1 = L(\Omega|\Lambda^2) - L(\Omega|\Lambda^1)$$

First evaluation of results

R < 0 Higher likelihood of model 1 (Variable b)

R > 0 Higher likelihood of model 2 (Constant b)

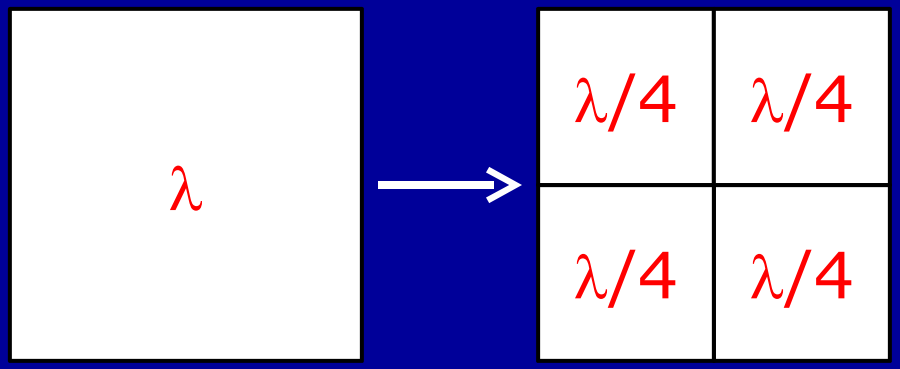
Likelihood-ratio per bin at Parkfield



Testing – Different Resolutions

Testing of models with different resolutions of bins

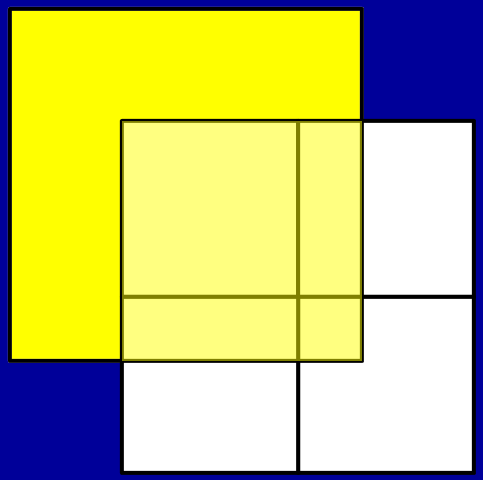
Distribute the expectations of the low resolution forecast over the bins of the high resolution forecast



This approach does not change the likelihood-ratio

Testing – Different Resolutions

Constraints of this approach



Limits of bins need to be aligned

Testing – Different Tests

N-Test

Test if the number of observed events is in the range of the expectation of a model

L-Test

Examines the consistency of a model with the observation

R-Test

Compares 2 models by its log-likelihood-ratio

Testing – Simulations

Create ‘simulated catalogs’

for every bin draw a random number from uniform distribution [0,1]

create simulated number of events based on random number and expectation of the bin (Poissonian distribution)

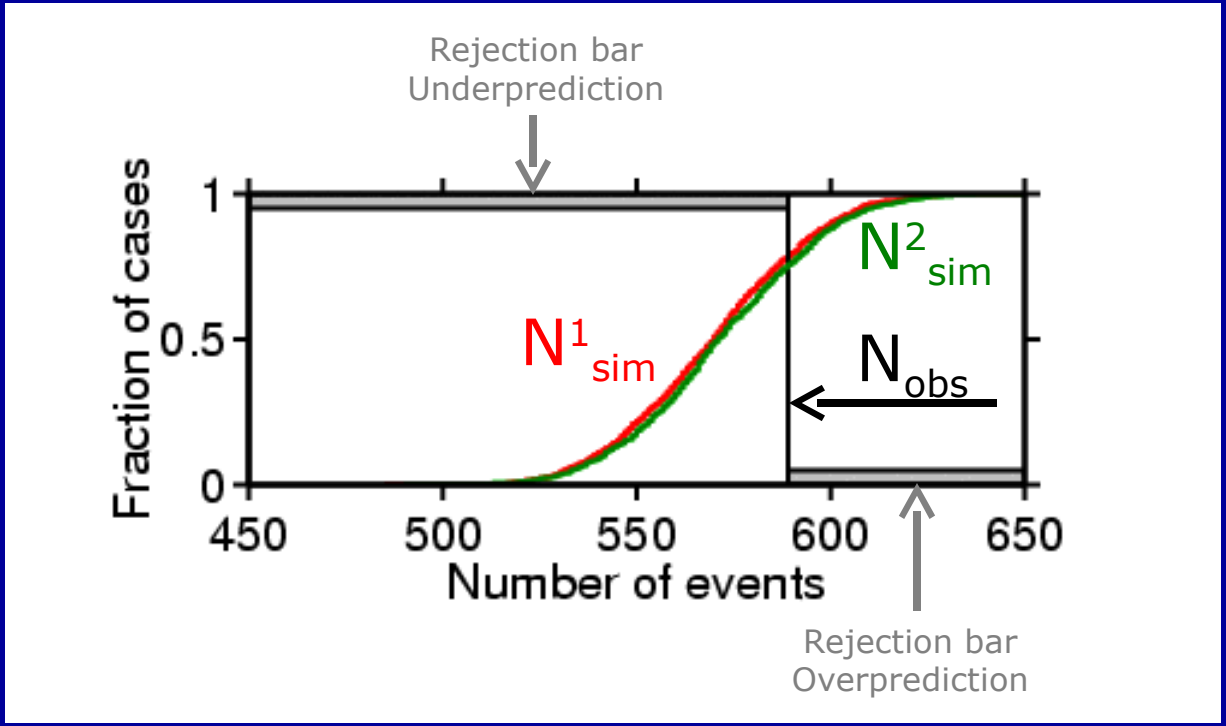
repeat over all bins

repeat 1000 times for the set of bins

→ Simulation based on expectations of a model

N-Test

Compare total number of observed events N_{obs} with number of simulated events N^{j}_{sim}



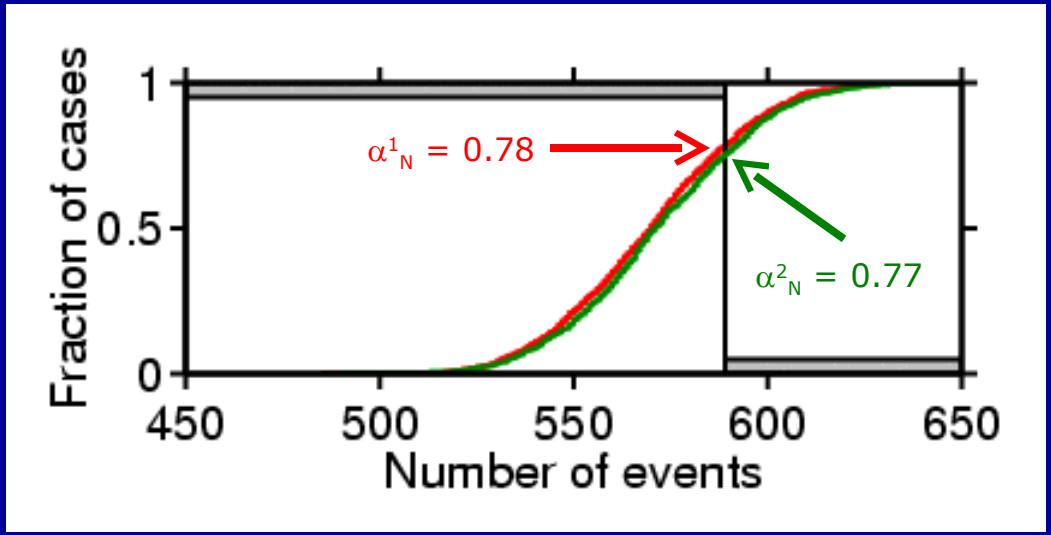
If the distribution N^{j}_{sim} touches one of the rejection bars, the model is over- or underpredicting events

N-Test

Quantify performance of a model with α^j_N, fraction of $N^j_{sim} < N_{obs}$

Reject model if

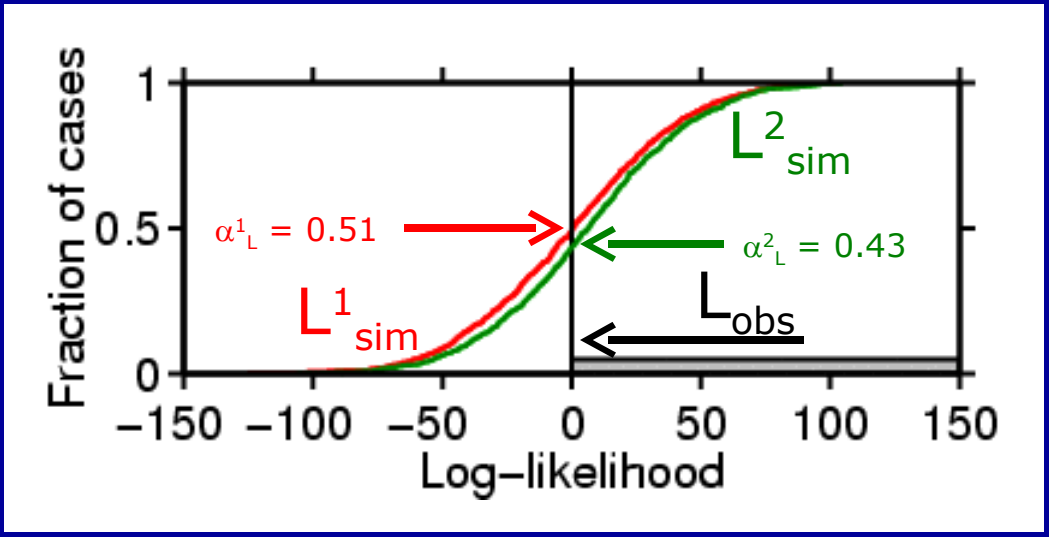
$\alpha^j_N > 0.95$ or
$\alpha^j_N < 0.05$



This test shows whether the observed number of events is in the range of the model's forecast

L-Test

Compare the likelihood L^j_{obs} of a model with the likelihoods L^j_{sim} of simulated catalogs



We normalize the log-likelihoods → $L^j_{obs} = 0$

This test shows whether the model's forecast is consistent with the observation

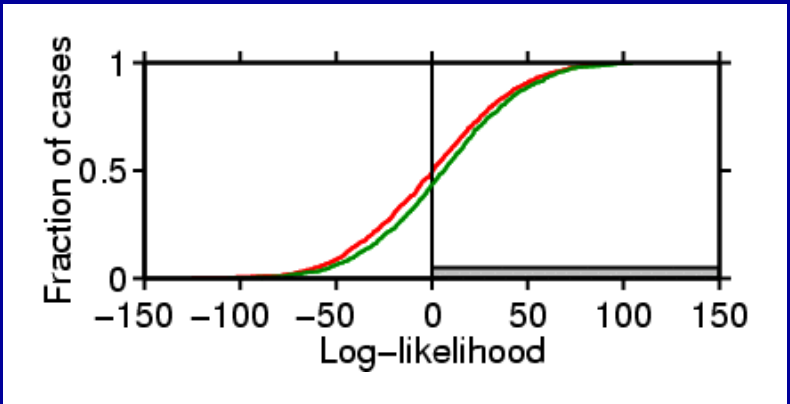
L-Test

Why can't we reject $\alpha^j_L > 0.95$?

Underprediction vs. 'perfect forecast'

A 'perfect forecast' would have a higher likelihood then any of the simulated likelihoods -> We cannot reject a 'perfect forecast'

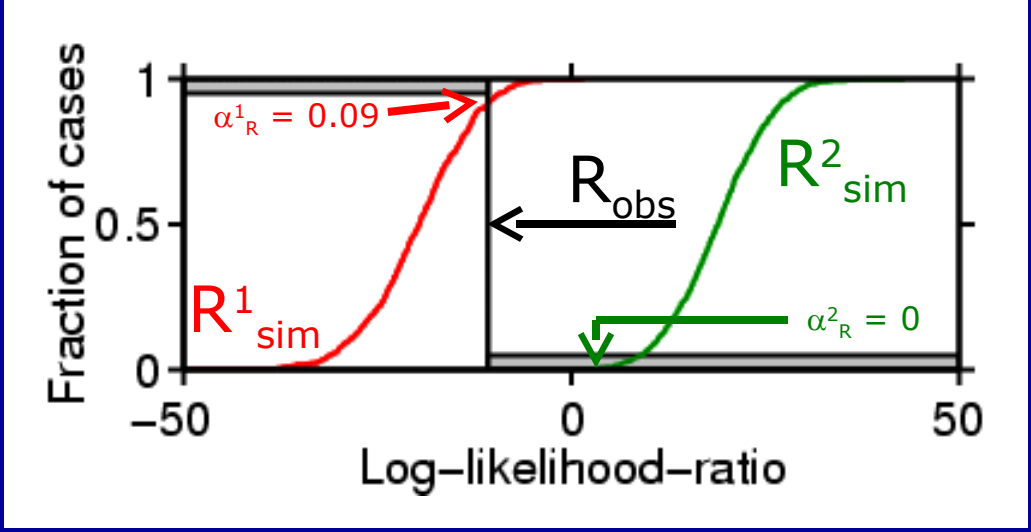
Models underpredicting events may succeed in the R-Test $\rightarrow$ N-Test



R-Test

Compare the likelihood-ratio R_{obs} of two models with the likelihood-ratios R^1_{sim} and R^2_{sim} of simulated catalogs based on the expectations of both models

R^1_{sim} simulation based on model 1
R^2_{sim} simulation based on model 2



Evaluation

Evaluate the likelihood-ratio of pair-wise tests of all models
→ R-Test

Test for consistency of the models with the observation
→ L-Test, N-Test

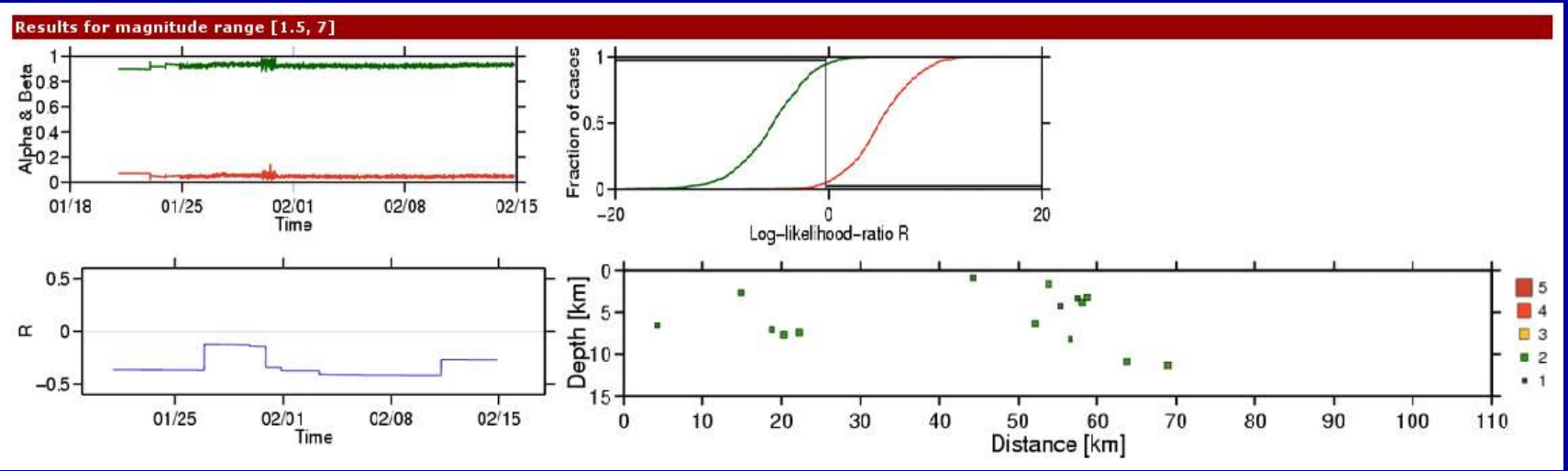
Analyze the spatial performance of models

Conducting the Tests

Compute the tests daily to track the temporal performance

Evaluate yearly to determine the overall and spatial performance
Yearly model investigations

Parkfield realtime test



Open Questions

Is the Poissonian distribution appropriate for likelihood computation?

$\rightarrow$ Negative binomial distribution

Should we use declustered catalogs for quasi-stationary models?

Thank you

Track the Parkfield realtime test from next month on at:

http://cool.ethz.ch