A retrospective comparative forecast test on the 1992 Landers sequence

J. Woessner,[1] S. Hainzl,[2] W. Marzocchi,[3] M. J. Werner,[1,4] A. M. Lombardi,[3] F. Catalli,[3] B. Enescu,[2,5] M. Cocco,[3] M. C. Gerstenberger,[6] and S. Wiemer[1]

Received 14 July 2010; revised 16 February 2011; accepted 7 March 2011; published 26 May 2011.

[1] We perform a retrospective forecast experiment on the 1992 Landers sequence comparing the predictive power of commonly used model frameworks for short-term earthquake forecasting. We compare a modified short-term earthquake probability (STEP) model, six realizations of the epidemic-type aftershock sequence (ETAS) model, and four models that combine Coulomb stress changes calculations and rate-and-state theory to generate seismicity rates (CRS models). We perform the experiment under the premise of a controlled environment with predefined conditions for the testing region and data for all modelers. We evaluate the forecasts with likelihood tests to analyze spatial consistency and the total amount of forecasted events versus observed data. We find that (1) 9 of the 11 models perform superior compared to a simple reference model, (2) ETAS models forecast the spatial evolution of seismicity best and perform best in the entire test suite, (3) the modified STEP model matches best the total number of events, (4) CRS models can only compete with empirical statistical models by introducing stochasticity in these models considering uncertainties in the finite-fault source model, and (5) resolving Coulomb stress changes on 3-D optimally oriented planes is more adequate for forecasting purposes than using the specified receiver fault concept. We conclude that statistical models perform generally better than the tested physics-based models and parameter value updates using the occurrence of aftershocks generally improve the predictive power in particular for the purely statistical models in space and time.

Citation: Woessner, J., S. Hainzl, W. Marzocchi, M. J. Werner, A. M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. C. Gerstenberger, and S. Wiemer (2011), A retrospective comparative forecast test on the 1992 Landers sequence, *J. Geophys. Res.*, *116*, B05305, doi:10.1029/2010JB007846.

1. Introduction

[2] Motivated by the observed spatiotemporal clustering of triggered earthquakes following moderate and large earthquakes a wide range of statistical and physics-based seismicity models has been developed and compared with observations over the past decades. Some models were derived from empirical statistics and the concept of triggering (e.g., the epidemic-type aftershock sequences (ETAS) model [*Ogata*, 1988]; short-term earthquake probabilities (STEP) [*Gerstenberger et al.*, 2005]). Others are based on physical concepts of static or dynamic stress transfer [e.g., *Aoi et al.*, 2010; *Catalli et al.*, 2008; *Gomberg et al.*, 2001; *Harris*, 1998; *King et al.*, 1994; *Steacy et al.*, 2005a; *Toda et al.*, 1998], or viscoelasticity and fluid migration [e.g., *Bosl and Nur*, 2002; *Nur and Booker*, 1972; *Miller et al.*, 2004]. While each model is validated to some extent by observations even in a truly prospective way [e.g., *Marzocchi and Lombardi*, 2009], comparing the predictive power of these models directly is difficult because different studies usually use different data sets, regions and timescales. Such comparisons, however, are important for two reasons: (1) to use the best available models for short-term earthquake forecasts and time-dependent seismic hazard assessment and (2) to enhance our understanding of the physical mechanisms of earthquake interaction and earthquake predictability.

[3] The Collaboratory for the Study of Earthquake Predictability (CSEP, www.cseptesting.org) [*Jordan*, 2006] and the Regional Earthquake Likelihood Models project (RELM, www.relm.org) [*Field*, 2007] highlight the need within the seismological community for comparative testing of models and defined frameworks for prospective testing of models on all scales. Here, we alter the CSEP concept: We evaluate and compare the predictive skill of a range of models for a specific aftershock sequence through statistical testing. We perform a retrospective earthquake predictability experiment

[1]Swiss Seismological Service, ETH Zürich, Zurich, Switzerland.
[2]GFZ German Research Centre for Geosciences, Potsdam, Germany.
[3]Instituto Nazionale di Geofisica e Vulcanologia, Rome, Italy.
[4]Now at Department of Geosciences, Princeton University, Princeton, New Jersey, USA.
[5]Now at the National Research Institute for Earth Science and Disaster Prevention, Tsukuba, Japan.
[6]GNS Science, Lower Hutt, New Zealand.

0148-0227/11/2010JB007846

Table 1. Overview of the Forecast Models That Contributed Forecasts for the Retrospective Testing Experiment[a]

	Model Type/ Model Name	Features	Total/Free Parameters	Modeler/Reference
0	STEP-0 generic STEP	M_{th} = 6 reference model	6/0	Woessner/*Gerstenberger et al.* [2005]
1	STEP-1 modified STEP	M_{th} = 2.5	6/6	Woessner/*Gerstenberger et al.* [2005]
2	ETAS-1	space-independent parameters stationary homogeneous bg.	7/7	Hainzl/*Hainzl et al.* [2008]
3	ETAS-2	K is space dependent stationary homogeneous bg.	7/7	Hainzl/*Hainzl et al.* [2008]
4	ETAS-3	stationary heterogeneous bg.	8/7 q = 1.5	Lombardi/*Lombardi et al.* [2010]
5	ETAS-4 NETAS	nonstationary heterogeneous bg.	9/8 q = 1.5	Lombardi/*Lombardi et al.* [2006]
6	ETAS-5	stationary heterogeneous bg. "effective parameters"	6/0	Werner/*Helmstetter et al.* [2006, 2007]
7	ETAS-6	stationary heterogeneous bg. updating "effective parameters"	6/5	Werner/*Helmstetter et al.* [2006, 2007]
8	CRS-1	space-dependent stressing rate nonuniform reference seismicity	1/1	Catalli/*Catalli et al.* [2008]
9	CRS-2	stationary heterogeneous background	4/1 r not fix	Enescu/*Toda et al.* [1998]
10	CRS-3	stress heterogeneity CV stationary uniform bg.	4/3 t_a fix	Hainzl/*Hainzl et al.* [2009]
11	CRS-4	stress heterogeneity CV stationary uniform bg. poroelastic & coseismic	4/3 t_a fix	Hainzl/*Hainzl et al.* [2009]

[a]The model number, the model class, first-order features, the number of total and free parameters, as well as the modeler and the reference(s) of the models are given. M_{th} is a threshold magnitude that determines which earthquakes are used as triggering events in the STEP model.

[see also *Jordan*, 2006; *Field*, 2007] to forecast the aftershock sequence of the 1992, *M*7.3, Landers earthquake sequence. We investigate the forecasting abilities of state-of-the-art models of clustered seismicity with one of the best available data sets of an aftershock sequence. As a group of modelers and testers, we jointly define the rules of this retrospective forecasting experiment, and we use statistical tests proposed by CSEP [*Schorlemmer et al.*, 2007; *Schorlemmer and Gerstenberger*, 2007; *Zechar et al.*, 2010] to evaluate the forecasts. Recent analyses have shown that such tests may be biased in analyzing clustering models [*Lombardi and Marzocchi*, 2010] as represented in this experiment; however, we emphasize that the main purpose here is to show how model forecasting performances may be evaluated and compared in a rigorous retrospective scientific experiment. Such kind of experiments are very important to provide a first evaluation of models that are not yet tested in a truly prospective experiment. In addition we outline ways to improve testing procedures that are yet not feasible to perform for all models. We report the results for the Landers sequence and we hope to expand this collaborative approach to retrospective testing of time-varying earthquake forecast models to other prominent aftershock sequences to assess the robustness of the results.

[4] Evaluating retrospective forecasts on the scale of aftershock sequences is in our opinion an important addition to the ongoing prospective tests within the framework of CSEP. While prospective testing remains the necessary standard for an unbiased performance evaluation, the evaluation progress and feedback to modelers can be slow because meaningful statistics accrue slowly with the number of events. For example, only first intermediate results for the Regional Earthquake Likelihood Model Experiment in California are now available from prospective testing [*Schorlemmer et al.*, 2010]. Retrospective testing, particularly when done in the spirit of a collaboratory, allows for much faster feedback: poorly performing models can be identified and either rejected or improved [*Mulargia*, 1997, 2001]. Retrospective testing can give pragmatic guidance about the quality and limitations of particular models and lead to recipes on how to best apply a particular model. By focusing on individual aftershock sequences and small magnitudes, we ensure that numerous events are available to evaluate the performance. Moreover, some prominent models of triggered seismicity require data (such as fault orientations or focal mechanisms) that may not be available in near-real time or for experiments at larger scales.

[5] However, evaluating retrospective forecasts may result in biased, overly optimistic results compared to prospective experiments because of already existing knowledge. First, we assume to have a finite-fault source model for the main shock for the first forecast; although finite-fault source models are now rapidly available in about a few hours after an event (e.g., from the USGS at http://earthquake.usgs.gov/earthquakes/), this will never be the case for a forecast immediately starting after a strong event occurred, the time when statistical models are able to generate a first forecast. Second, we select a catalog with locations computed by waveform cross-correlation techniques [*Hauksson*, 2000] that are generally not available in near-real time and have information about parameters that are needed to set up models. Such data may become more rapidly available in the future and might thus benefit near-real time prospective experiments. Third, modelers benefit from their knowledge of the well-known Landers earthquake sequence from earlier studies, and this might bias the forecasts. Finally, some models were explicitly developed on earthquake catalogs that include the Landers sequence. Therefore, it may be that the retrospective forecasts analyzed here provide upper limits of the predictive skill of the evaluated models.

[6] In this first retrospective collaborative experiment, we evaluate the capabilities of 11 different models of triggered seismicity to forecast the spatial and temporal distribution of magnitude $M_L \geq 3$ earthquakes after the Landers earthquake (Table 1). In particular, we investigate whether (1) the observations are consistent with the forecasted rates in terms of total number and spatial distribution, (2) physics-based models based on Coulomb stress changes and rate-and-state

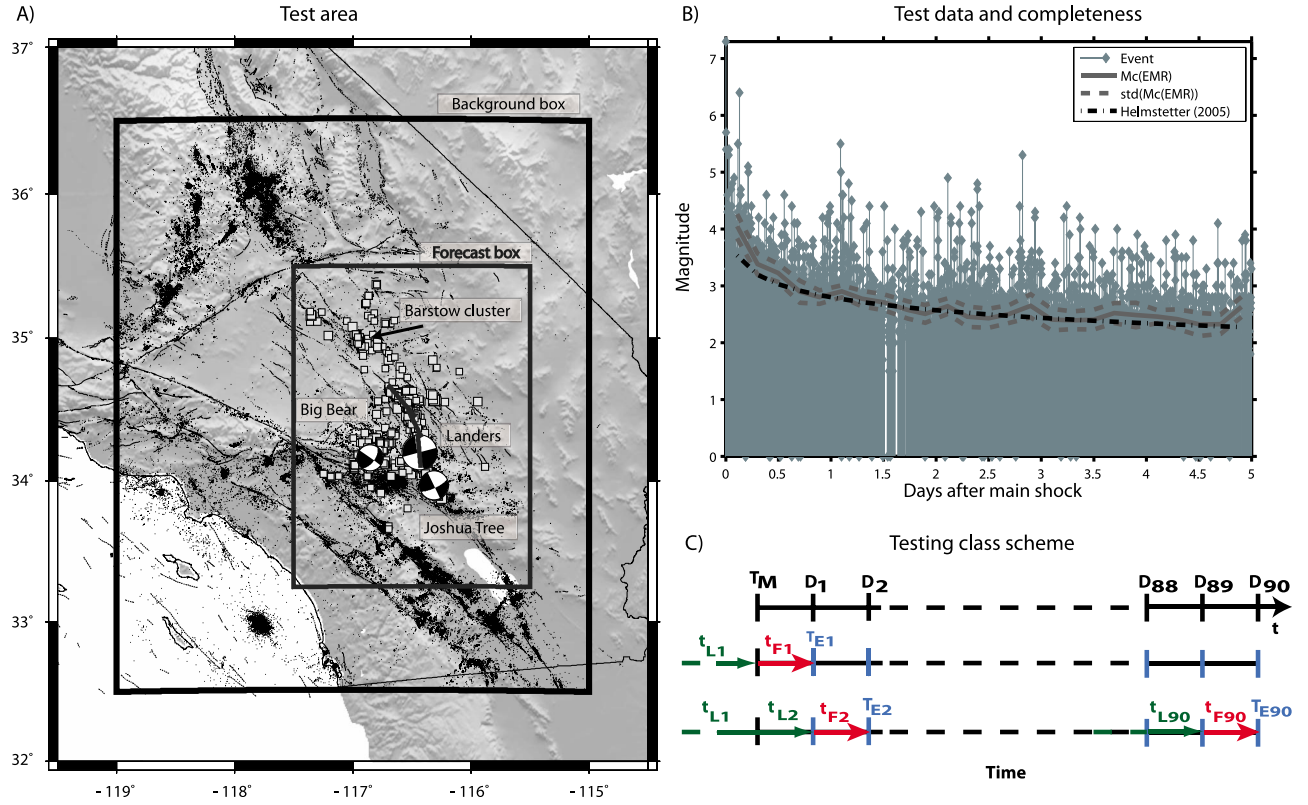

Figure 1. (a) Region of the 1992 Landers aftershock sequence with the experiment's data collection region (−119°W, 32.5°N; −115°W, 36.5°N) and forecast testing region (−117.5°W, 33.25°N to −115.5°W, 35.5°N). Also shown are the surface projection of the *Wald and Heaton* [1994] Landers earthquake fault model; focal mechanisms of the Joshua Tree, Big Bear, and Landers earthquakes; along with earthquakes with magnitudes $M_L \geq 2$ (black dots) and target events with magnitude $M_L \geq 3$ in the testing region (light gray squares). (b) Magnitude of completeness during the first 5 days after the Landers main shock in the testing region estimated using the $M_c(EMR)$ method by *Woessner and Wiemer* [2005] (gray curve with uncertainties based on 500 bootstrap samples) and the method by *Helmstetter et al.* [2006] (black curve). (c) Testing class scheme: T_M, the main shock time; D_i, days after T_M; t_{Li}, learning periods; t_{Fi}, forecast periods; T_{Ei}, times of the test performance.

friction theory (CRS models) perform better than purely statistical models, and (3) different flavors of available statistical models provide substantially different results.

[7] We structure the manuscript by introducing the data provided to all modelers for usage in the experiment. Second, we outline the testing class that defines the testing region, period, and magnitude range. Third, we summarize the models that provided forecasts; each model has previously been published and was altered to match the testing class requirements: the models thus were only changed to use the provided authorized data. Fourth, we introduce the statistical tests and measures that we use to rank the performance of the models in comparison to the observed seismicity. Based on these measures, we discuss the results of our experiment.

2. Data

[8] We use the relocated earthquake and the focal mechanism catalogs [*Hauksson and Shearer*, 2005; *Hauksson*, 2000] provided by the Southern California Earthquake Data Center (SCEDC) (http://data.scec.org/research/altcatalogs.html). We only include shallow (<30 km) earthquakes. We define separate data collection and forecast testing regions (Figure 1a). Models may select earthquakes from the larger collection region as input data for their forecasts within the testing region to minimize boundary effects. The forecast box defines the area earthquakes are forecasted in and tested for.

[9] There are 98879 earthquakes in the relocated catalog with local magnitude $M_L \geq 0.1$ (1563 with $M_L \geq 3.0$) located in the collection region between 1984 and the time T_M of the Landers main shock. In the testing region, 38941 $M_L \geq 0.1$ occurred before Landers, of which 670 events have $M_L \geq 3.0$

Table 2. Number of Learning and Target Earthquakes and Focal Mechanisms Available in the Testing Region[a]

	Relocated Events		Events With Fault Plane Solution	
Period	$M_L \geq 0.1$	$M_L \geq 3$	$M_L \geq 0.1$	$M_L \geq 4.5$
$1984 < T_M$	38941	670	10102	15
$T_M - T_M + 90d$	21647	1245	4354	31

[a] T_M is the main shock time of the 1992 M_L 7.3 Landers earthquake.

Table 3. Parameter Values of the ETAS Models[a]

	ETAS Number					
	1/2	3	4	5[b]	6	Parameter
$\mu(M \geq 3)$, [day^{-1}]	0.046	0.19	0.19	0.221	0.221	background rate
K	0.021	0.021	0.021	0.65	0.65	productivity
α	1.6	1.2	1.2	1.7	1.7	α value
c, [days]	0.0031	0.002	0.002	0.0035	0.0035	c value
p	1.06	1.06	1.06	1.06	1.06	decay parameter
f_d				0.73	0.73	
b	0.91	0.91	0.91	0.91	0.91	b value
γ		0.6	0.6			
d_0, [km]	0.048	0.3	0.3			
q	1.44	1.71	1.5			
ν, [day^{-1}]			0.19			
$M_{\min}$	3	3	3	3	3	min. magnitude for triggering
$f_i(r, Mi)$	f_{pl}	f_{pl}	f_{pl}	f_{gs}	f_{gs}	spatial PDF

[a]Initial values for the first forecast at the time of the Landers main shock. The parameter values are reestimated with available data up to the start date of the subsequent 24 h forecast.

[b]Model ETAS-5 does not reestimate the parameters during the sequence. f_{pl} and f_{gs} refer to power law and gaussian smoothing of seismicity rates, respectively.

(Table 2). In the 90 days after the Landers main shock, the catalog lists 1245 events with $M_L \geq 3.0$; to forecast these target events is the objective of the experiment. Due to computation time considerations, CRS modelers agreed to only use focal mechanisms of events with $M_L \geq 4.5$. In the testing period, there are 31 events with focal mechanisms that could be used.

[10] Some of the models require a regional stress field as input data. We assume a direction of the maximum compressive stress oriented at *N*7°E in agreement with *King et al.* [1994]. The modelers used a differential stress ($\sigma_1 - \sigma_3$) of 10 MPa and set the vertical stress σ_2 to an intermediate value to account for the strike-slip environment. In particular, σ_1, σ_2, and σ_3 were assumed to be 5, 0 and −5 MPa. This is a simplifying assumption as the stress field has undergone multiple changes in this time period [*Hauksson*, 1994] and may be heterogeneous on the scale of the Landers rupture [*Hardebeck and Hauksson*, 2001]. Furthermore, *Steacy et al.* [2005b] showed that predictions based on Coulomb failure stresses are sensitive to the magnitude and direction of the regional stress field when resolving the stress changes onto optimally oriented planes, both in 2-D and 3-D. However, to facilitate the comparison between the models, we selected this simple regional stress field.

[11] There are multiple finite-fault source models available for the Landers earthquake [*Cotton and Campillo*, 1995; *Cohee and Beroza*, 1994; *Hernandez et al.*, 1999; *Wald and Heaton*, 1994; *Zeng and Anderson*, 2000]. Details of the slip distribution play an important role in the stress computations close to the causative fault, but not in the far field. The stress variability resulting from different source models and its impact on parameter estimates and model forecasts was investigated by *Hainzl et al.* [2009]. To make comparisons between models easier, we use the same finite-fault slip model for all CRS models, namely the model by *Wald and Heaton* [1994], who incorporated multiple data sets to constrain their solution. For the 1992 $M_L = 6.5$ Big Bear earthquake, the line source model by *Jones and Hough* [1995] is used. All source models are available from the finite-fault source model database at ETH Zurich (http://www.seismo.ethz.ch/srcmod by M. P. Mai). Due to the resolution and uncertainty of the slip model it cannot generally be expected that the seismicity rate computed at grid cells close to the source model correctly replicate the occurrence of seismicity.

[12] The CRS models may calculate multiple stress steps due to earthquakes with a magnitude $M_L \geq 4.5$ (Table 2). To obtain a simple estimate of the amount of slip and the dimension of the area, we assume that the moment magnitude is equal to the local magnitude [*Clinton et al.*, 2006]. We calculate the source dimensions based on the scaling relations by *Wells and Coppersmith* [1994] that differentiate between faulting style (dip slip or strike slip) based on the events' rake. This information is only used in CRS-1 (Table 1) [*Catalli et al.*, 2008]. We provide the entire data sets and predefined parameter values needed to set up model calculation as auxiliary material.[1] Parameter values and choices for single models are explained in section 4 (see Tables 3 and 4).

3. Definition of the Retrospective Forecasting Experiment

[13] Testing needs to be performed against a complete data set. The completeness threshold in an aftershock sequence varies strongly in time. We estimated the completeness threshold for the period of interest, with a particular focus on the first days. We analyzed the first five days following the Landers main shock with two methods: (1) the Entire Magnitude Range method $M_c(EMR)$ by *Woessner and Wiemer* [2005] and (2) the completeness function by *Helmstetter et al.* [2005, 2006]. For $M_c(EMR)$, we sample the events in successive 0.2 *day* time windows. We find that both methods lead to consistent results (Figure 1b) indicating that the completeness level is $M_c = 4$ for the first day and improves to $M_c = 3$ in the following days. This completeness threshold is achieved throughout the 90 days testing period. Thus we test against magnitudes $4 \leq M_L \leq 8$ on the first day and against $3 \leq M_L \leq 8$ thereafter.

[14] To facilitate a comparative evaluation of different forecast models, the forecasts need to comply with a set of rules. The objective of this retrospective experiment is to forecast the expected number of earthquakes over successive 24 h periods in spatial cells of 0.05 by 0.05 degrees in the magnitude range $3 \leq M_L \leq 8$ with a binning of $\Delta M_L = 0.1$. The last bin includes events above magnitude 8. The first 24 h forecast period starts at the time T_M of the 1992 Landers main shock (including information about the main shock). From then on, the models successively forecast seismicity rates in 24 h windows with the possibility to update parameter values and forecasts based on all hitherto observed earthquakes. The forecasts are computed for a total of 90 days and evaluated after each 24 h period (Figure 1c).

4. Overview of Models

[15] All 11 different models in this study have been published elsewhere (see Table 1) and applied to similar

[1]Auxiliary material data sets are available at ftp://ftp.agu.org/apend/jb/2010JB007846. Other auxiliary material files are in the HTML.

Table 4. Parameter Values in CRS Models[a]

	CRS Number				
	1	2	3	4	Parameter
μ_f	0.75	0.3	0.3	0.75	friction coefficient
$A\sigma$, [MPa]	0.04	0.04	0.04	0.04	
t_a[b], [yr]	$\frac{A\sigma}{\dot{\tau}_r}$	50	27.4	27.4	sequence duration
$r(M \geq 3)$, [day^{-1}]	0.176[c]	0.75	0.75	0.75	background
W, [km]	30				thickness of seismogenic area
B	0.47			0.6	Skempton coeff.
D, [$\frac{m^2}{s}$]				0.1	hydraulic diffusivity
$\tilde{a}$				0.8	dimensionless effective stress coeff.
CV			1	1	coeff. of variation of stress values

[a]Initial values for the first forecast at the time of the Landers main shock. The parameter values are reestimated with available data up to the start date of the subsequent forecast.

[b]Here $\dot{\tau}_r$ and t_a are spatially variable in CRS-1 with $\dot{\tau}_{r(ref)} \cong M_o W^{-1} \frac{b}{1.5-b}[10^{(1.5-b)(M_{max}-M_c)} - 1]$ [*Catalli et al.*, 2008; *Cocco et al.*, 2010]. All models use a b value of $b = 0.91$.

[c]Reference rate.

tectonic regimes; therefore we only provide a brief overview. The number of free parameters is representative for the number of variables in the model equations that are estimated with a maximum likelihood approach and that can be updated during the time sequence; it is however difficult to define the real number of free parameters for models that combine a time-invariant and a time-variable part in a two-step procedure (e.g., for STEP models or ETAS-6) or that estimate the parameter values spatially; the number could be considered much higher. In this sense the numbers in the Tables 3 and 4 represent only a minimum number. For model ETAS-5 based on *Helmstetter et al.* [2006], we consider the number of free parameters as zero: the parameter values were estimated from the entire Southern California catalog including the Landers sequence; the parameters are fixed and partly expert choice, partly from maximum likelihood estimation. We emphasize that we consider the forecast results of each model as having the same degree of freedom. For reference, Tables 3 and 4 list important initial model parameter values estimated from seismicity prior to the Landers sequence or taken from the literature.

4.1. Statistical Forecast Models

[16] The statistical forecast models are based on empirical relationships that emerged from analyzing multiple earthquake catalogs. The most important empirical relations herein are the frequency-magnitude distribution [*Gutenberg and Richter*, 1944], the Omori-Utsu law [*Utsu*, 1961] and the ETAS model [*Ogata*, 1988]; yet empirical studies on the distribution of earthquakes in space contribute equally important to the design of statistical models. The models should be understood as frameworks for which single components can be modified, added and supplemented in future realizations upon success in experiments such as this one.

[17] Some statistical models (the STEP models and ETAS-5 and ETAS-6) merge a time-invariant spatially heterogeneous background model with time-varying seismicity rates. To estimate the background model, we select earthquakes from 1 January 1984 until 31 December 1991 in the collection region (Figure 1). We decluster the catalog with the algorithm by *Reasenberg* [1985] modified by *Helmstetter et al.* [2007]. This background seismicity model is used for models STEP-0, STEP-1, ETAS-5, and ETAS-6.

4.1.1. A Reference Model: STEP-0

[18] The STEP model is a spatially extended version of the simple aftershock model by *Reasenberg and Jones* [1989, 1990, 1994]:

$$\lambda(t, M) = \frac{10^{a'+b(M_m-M_{th})}}{(t+c)^p} \tag{1}$$

where $\lambda(t, M)$ is the rate of aftershocks with magnitudes greater than a threshold $M \geq M_{th}$ occurring at time t, and M_m denotes the main shock magnitude. The constants a' and b are derived from the frequency-magnitude distribution with a' as described by *Reasenberg and Jones* [1989]. The constants p and c result from the Omori-Utsu law [*Utsu*, 1961].

[19] To assess whether the research in the past decades has lead to significantly improved forecast models, we define a simple reference model (STEP-0) based on *Reasenberg and Jones* [1989, 1990, 1994] (Table 1). STEP-0 corresponds to the generic element of the STEP model [*Gerstenberger et al.*, 2004; *Woessner et al.*, 2010] in that it does treat earthquakes always as a point source. STEP-0 only considers earthquakes with magnitude $M_L \geq 6.0$, i.e., the Joshua Tree $M_L = 6.1$, the Landers $M_L = 7.3$ and the Big Bear $M_L = 6.5$ earthquakes, to forecast seismicity rates. The earthquake rates are smoothed radially with a decay of r^{-2} around the epicenters of the causative events, with r being the epicentral distance. The initial or generic parameters for estimating the rate of seismicity in the model are: $a' = -1.67$, $b = 0.91$, $p = 1.08$, $c = 0.05$ days [*Reasenberg and Jones*, 1994; *Gerstenberger et al.*, 2004].

4.1.2. STEP Model: STEP-1

[20] In addition to the simplified reference model STEP-0, we include a full STEP model (STEP-1, Table 1) [*Gerstenberger et al.*, 2005] adjusted to the Landers region and with the same background model as STEP-0. In this model, earthquakes with magnitudes $M_L \geq 2.5$ are assumed to contribute to the triggering of earthquakes in the target magnitude range because small events can significantly increase the probabilities of larger earthquakes [*Helmstetter et al.*, 2005]. Additionally, the parameters of equation (1) are estimated first for the total sequence and then allowed to vary spatially as soon as sufficient seismicity from the aftershock sequence is available to resolve spatial variability. All other parameter settings are the same as those of *Gerstenberger et al.* [2004, 2005].

4.1.3. ETAS Model: ETAS-1 to ETAS-6

[21] The models ETAS-1 to ETAS-6 (see Table 1 and, for parameters, Table 3) are based on the formulations by [*Ogata*, 1988, 1998; *Ogata and Zhuang*, 2006; *Zhuang et al.*, 2002; *Lombardi et al.*, 2006; *Helmstetter et al.*, 2006]. The seismicity rate $\lambda(t, x, y)$ at each point in space (x, y) and time t is given by

$$\lambda(t, x, y) = \mu(t, x, y) + \sum_{i:t_i<t} \frac{Ke^{\alpha(M_i-M_{th})}}{(t-t_i+c)^p} f_i(r, M_i) \tag{2}$$

with the model parameters K, c, α, p, and the background rate μ. f_i (r, M_i) is a normalized function describing the

spatial probability distribution of triggered seismicity that differs between the various ETAS models.

[22] Models ETAS-1 to ETAS-4 use a power law function of the epicentral distance r to the parent event

$$f_{pl}(r) = \frac{(q-1)d_{pl}(M)^{2(q-1)}}{\pi\left[r^2 + d_{pl}(M)^2\right]^q} \quad (3)$$

with the additional parameter q and the magnitude-dependent distance parameter d_{pl}, which varies between ETAS-1 to ETAS-4 as discussed below. In contrast, ETAS-5 and ETAS-6 use a Gaussian kernel [*Helmstetter et al.*, 2006, 2007]:

$$f_{gs}(r) = C_{gs}\exp\left(-\frac{r^2}{2d_{gs}(M)^2}\right) \quad (4)$$

where C_{gs} is a normalization constant and $d_{gs}(M) = 0.5 + f_d \times 0.01 \times 10^{0.5M}$ km, and f_d is a free parameter measuring the size of the rupture zone. Additionally, ETAS-5 and ETAS-6 obtain the spatial aftershock distribution due to earthquakes larger than magnitude 5.5 by smoothing the locations of early aftershocks.

[23] ETAS-2 differs from ETAS-1 only by a spatially variable aftershock productivity K in equation (2), which is estimated by comparing the predicted and observed initial aftershocks: aftershock locations are smoothed according to *Helmstetter et al.* [2007]; then K is replaced in each grid cell by $K_i = K \cdot (N_{obs,i} - N_{back})/(N_{F,i} - N_{back})$, where $N_{back} = \mu \times T \times A_i$ is the number of background events for the forecast of the spatial probability map of future earthquakes. N_{obs} and N_F are the observed and forecasted number of events, T the forecast period and A_i the area of the ith grid cell.

[24] In contrast to the other ETAS models, ETAS-1 and ETAS-2 use only the parameter estimates of the space-independent ETAS model ($f(r, M) \equiv 1$) to forecast the number of future events. *Hainzl et al.* [2008] showed that the space-dependent parameters can be biased (in particular, α in equation (2) may be underestimated) if the anisotropy of the aftershock pattern is not taken into account. The parameter estimates of the space-dependent ETAS model are then only used to calculate the spatial probability distribution of the predicted events (but not their rate).

[25] The models ETAS-1 to ETAS-4 calculate the forecasted number of events for the forecast period T_F by averaging over many Monte Carlo simulations of forecasts. ETAS-1 and ETAS-2 perform 10000 simulations, while 1000 simulations are used by ETAS-3 and ETAS-4, resulting in higher rate fluctuations. The simulations are needed to provide for secondary aftershocks triggered during the target period which can contribute significantly to the activity [*Helmstetter et al.*, 2003].

[26] In ETAS-1 to ETAS-4, parameter values are estimated from preceding seismicity using the maximum likelihood method [*Ogata*, 1998; *Daley and Vere-Jones*, 2003]. The model ETAS-4 assumes that the background rate varies with time. In the present study, we applied the strategy proposed by *Lombardi et al.* [2010], to model the temporal variation of the background rate and of the spatial distribution of background seismicity. These short-term variations can occur in volcanic regions [*Lombardi et al.*, 2006] or in tectonic regimes in which rapid fluid flow can play an important role [*Lombardi et al.*, 2010]. In particular, the background rate is estimated in a moving, nonoverlapping time window of 24 h in the collection region, where all the other parameters are taken equal to the values inferred for the whole sequence. For each forecasting time step, we estimate the background rate, by using all data hitherto occurred. In this way, we have a nonstationary ETAS model with the background varying in space and time, as a piecewise nonhomogeneous Poisson process.

[27] ETAS-5 and ETAS-6 use parameter values inferred from southern California seismicity based on a maximum likelihood estimate; parameter values of ETAS-5 stay fixed for the entire forecasting period. In contrast, ETAS-6 updates its parameters on a daily basis according to the following procedure: retrospective 1 day forecasts in each 0.05 degree cell are evaluated according to their Poisson likelihood; parameters that maximize the likelihood are chosen to forecast the next day in the sequence [*Helmstetter et al.*, 2006]. Thus, the estimated parameters are effective values which account for direct and indirect activity in the 1 day time window and do not require Monte Carlo simulations.

[28] The ETAS models mainly differ in their assumptions about the background seismicity (homogeneous or heterogeneous) and about the spatial probability distribution function of aftershocks. ETAS-1 and ETAS-2 assume a homogeneous and constant background rate μ_0, while the background rate $\mu(x, y)$ is spatially variable for the other ETAS models. ETAS-4 assumes additionally that the background rate $\mu(x, y, t)$ is nonstationary (time varying) [*Lombardi et al.*, 2006]. For the spatial probability distribution $f_i(r, M_i)$ (equation 2), ETAS-1 and ETAS-2 both calculate two forecasts each based on (1) a constant $d_{pl} = d_0$ and (2) a scaling function $d_{pl} = \sqrt{10^{-3.49+0.91M}}$ according to *Wells and Coppersmith* [1994] and then combine the two forecasts with weights determined by their respective Akaike Information Criterion (AIC) scores. In contrast, ETAS-3 and ETAS-4 use the scaling $d(M) = d_0 e^{\gamma(M-M_{th})}$, where γ is an additional free parameter, while the parameter q in equation (3) is fixed to 1.5 because this value is expected in the far field if aftershocks are triggered by static stress changes [*Lombardi et al.*, 2006]. Finally, ETAS-5 and ETAS-6 use $d_{gs}(M)$ defined in equation (4).

4.2. Combined Coulomb Stress Change–Rate and State Models (CRS Models)

[29] CRS models form a class of physics-based models to forecast aftershock seismicity. These models express the hypothesis that earthquakes are trigged by static Coulomb Stress Changes (ΔCFS) induced by a dislocation in an elastic medium [e.g., *King et al.*, 1994]. To obtain seismicity rate forecasts, the Coulomb stress changes are combined with the framework of rate- and state-dependent friction [*Dieterich*, 1994; *Dieterich et al.*, 2000]. The theoretical background of the CRS models and their sensitivity is has been discussed in detail in literature [*Catalli et al.*, 2008; *Cocco et al.*, 2010; *Hainzl et al.*, 2009; *Toda et al.*, 1998, 2005]. Initial parameter values are given in Table 4. Models CRS-2, CRS-3, and CRS-4 update the model parameters to forecast the seismicity rates, CRS-1 uses fixed parameter

values estimated for the seismicity preceding the Landers aftershock sequence.

[30] Rate-and-state friction theory accounts for the dependence of slip on the frictional strength and the time-dependent healing that are observed in laboratory experiments [*Dieterich*, 1994]. For a given stressing history, the seismicity rate hR depends on three model parameters: (1) the background seismicity rate r, (2) the tectonic stressing rate $\dot{\tau}$, and (3) the parameter product $A\sigma$, where A is a dimensionless fault constitutive parameter and σ the effective normal stress. Alternatively to the parameter $\dot{\tau}$, the aftershock relaxation time $t_a \equiv A\sigma/\dot{\tau}$ can be used as free parameter.

[31] While model CRS-1 resolves Coulomb stress changes onto planes with a prespecified strike, dip and rake approximately parallel to the Landers rupture (330°, 90°, 180° according to the convention proposed by *Aki and Richards* [2002]), models CRS-2 to CRS-4 compute stress changes on 3*D* optimally oriented planes.

[32] CRS-1 is identical to the model by *Catalli et al.* [2008] and *Cocco et al.* [2010], calibrated to the Landers sequence. It reduces the rate-and-state model to only estimate one free parameter ($A\sigma$) which is set equal to the optimal value obtained by *Catalli et al.* [2008] for the 1997 M_W5.9 Colfiorito, Italy, earthquake sequence (see Table 4). In particular, the reference seismicity $r(x, y)$ is estimated from the smoothed seismicity prior to the main shock. This reference seismicity differs from the background rate used in the ETAS models because it includes clusters and does not assume that background earthquakes are independent. The tectonic stressing rate $\dot{\tau}(x, y)$ is then fixed by the linear relation $\dot{\tau}(x, y) \propto r(x, y)$, which is derived from the balance of seismic moment release [*Catalli et al.*, 2008].

[33] CRS-2 model builds on the work by *Toda et al.* [1998, 2005]. Coulomb stress changes are determined on 3*D* optimally oriented fault planes, with the algorithm of *Wang et al.* [2006] and assuming a coefficient of friction $\mu_f = 0.3$ [*Hainzl et al.*, 2009]. The maximum Coulomb stress change over the seismogenic depth sampled at 7 and 11 km is calculated assuming that seismicity will occur at the location and depth where stress is most increased toward failure [*Toda et al.*, 2005]. The parameter product $A\sigma = 0.04$ *MPa* is set constant. The aftershock duration $t_a = 50$ *years* is fixed because its value does not influence the forecasts much [*Toda et al.*, 2005]. The daily background rate $r(M \geq 3)$ is assumed to be uniform in space and initially set to 0.75 for the entire study area, based on the observed seismicity rate in the year preceding the Landers main shock in the forecasting region. The background rate is updated during the forecasts using equation (11) of *Dieterich* [1994] considering aftershocks that occurred before the forecast period. Starting with the second-day forecasts, r has only small fluctuations around an average of 0.46. This estimation is robust against variations in the chosen initial background rate.

[34] In model CRS-3, coseismic stress changes are calculated (as for CRS-2) with the algorithm by *Wang et al.* [2006], while the model CRS-4 additionally computes postseismic stress changes due to poroelasticity with the software provided by *Wang and Kümpel* [2003]. The latter model requires additional parameters whose values were set to typical values at all depth levels: (1) the hydraulic diffusivity $D = 0.1\frac{m^2}{s}$, (2) the Skempton ratio $B = 0.6$, and (3) the dimensionless effective stress coefficient $\tilde{a} = 0.8$, which measures the change in pore volume per unit change in bulk volume under drained conditions. By assuming $\mu_f = 0.75$, the effective coefficient of friction $\mu_f' = (1 - B)\mu_f = 0.3$ is equal to the coefficient of friction also used by CRS-3 and CRS-1 [*Hainzl et al.*, 2009].

[35] Both the CRS-3 and CRS-4 models account for uncertainties in the computed stress changes because (1) small-scale slip variability is generally not resolved in the finite-fault source models with a slip patch resolution of 3 km × 2.5 km; (2) lateral heterogeneities of the material properties exist; (3) large uncertainties exist in the inversion of finite-fault source models. Stress fluctuations in each grid cell are computed assuming a Gaussian distribution according to *Marsan* [2006], where the standard deviation is assumed to be proportional to the absolute value of the stress change. In particular, the proportionality constant (or coefficient of local stress variability) CV is for simplicity assumed to be constant in space, which seems to be a good first-order approximation according to *Hainzl et al.* [2009]. In general, both models have four free parameters, but in this study, the aftershock duration is fixed at $t_a = 10000$ *days* $\simeq$ 27.4 *years*, reducing the number of free parameters to three. The aftershock duration does not affect the seismicity decay in the beginning of the seismic activity [*Cocco et al.*, 2010]. The three free parameters are estimated by maximum likelihood estimation based on first aftershock data: the parameter product $A\sigma$, the background rate r and the coefficient of stress variability CV.

5. Definition of Statistical Tests

[36] We quantify the consistency of the observed earthquake sequence with the calculated forecasts using previously published statistical tests [*Jackson*, 1996; *Schorlemmer et al.*, 2007; *Werner et al.*, 2009; *Zechar et al.*, 2010]: (1) the modified N(umber) test compares the total number of predicted and observed earthquakes, (2) the S(pace) test measures the consistency between the spatial distribution of observed and predicted earthquakes, and (3) the likelihood ratio test (R test) to compare the eleven models against the simple STEP-0 reference model. All of these tests are currently used to evaluate prospective short- and long-term earthquake forecasts within CSEP. To establish a ranking of the models, we use the log likelihood scores of the models (Table 6), likelihood gains and rejection ratios as will be defined in sections 5.1–5.3.

[37] We apply statistics that are implemented in ongoing prospective testing experiment of CSEP. All the tests use a Poisson distribution to perform simulations and to establish if a null hypothesis can be rejected at a particular significance level, in general at an effective level of $\alpha_{eff} = 0.025$. *Lombardi and Marzocchi* [2010] have shown that this hypothesis does not hold for ETAS models applied to forecast aftershocks sequences [see also *Werner and Sornette*, 2008]; in brief, the variability of a forecast is certainly larger than the variability estimated through a Poisson distribution. This leads the CSEP tests to reject forecast of ETAS models too frequently. Keeping this in mind, we nevertheless decided to use these tests because we are not yet in the position to quantify exactly this bias as a function

of the expected seismic rate. Other possible biases may arise from the fact that the occurrences and the rejections are not spatially independent.

[38] Therefore, the results we obtain may be overly conservative and may imply that the models perform actually better than we find. In other words, the models might be rejected less frequently as they are in the current tests when using the correct probability density function for simulations.

[39] In an attempt to quantify the gain achieved by the more complex models over the simple STEP-0 model, we calculate likelihood gains for each model compared to STEP-0 and use the likelihood ratio test (R test) to evaluate whether the simple STEP-0 model can be rejected in favor of more complex models, and vice versa [*Jackson*, 1996; *Schorlemmer et al.*, 2007]. By Occam's razor, a simple model is preferable over more complex ones if both explain the data equally well.

5.1. Modified N Test

[40] The modified N test evaluates the consistency between the forecasted and observed total number of events. *Zechar et al.* [2010] proposed the two metrics

$$\delta_1 = 1 - F(N_{Obs} - 1|N_F) \tag{5}$$

$$\delta_2 = F(N_{Obs}|N_F) \tag{6}$$

where $F(x|\mu_e)$ is the right-continuous Poisson cumulative distribution function with expectation μ_e evaluated at x. The two metrics allow us to answer two questions separately: (1) Assuming the forecast is correct, what is the probability of observing at least N_{Obs} earthquakes (δ_1) and (2) assuming the forecast is correct, what is the probability of observing at most N_{Obs} earthquakes (δ_2)? If the observations fall into the far tails of the distributions, then we reject the forecasts as inconsistent with the observations at a chosen significance level.

[41] We apply the N test to each 24 h forecast individually (daily N tests); additionally we perform a cumulative N test for which we cumulatively sum the forecasted number of earthquakes since the main shock and apply the N test consecutively over a growing testing period composed of multiples of 24 h forecasts. For each of the two N tests, we obtain 90 quantile scores and consequently a binary test result, rejection or nonrejection according to the effective rejection level α_{eff}. To summarize the results, we calculate the fractions $R_N(Daily)$ and $R_N(Cumulative)$ of the total number of days (90) during which a model's forecast is rejected because either $\delta_1(t) < \alpha_{eff}$ or $\delta_2(t) < \alpha_{eff}$.

[42] The cumulative number tests are strongly correlated in time as models are penalized for poor performance on single test days which is kept in the memory of the cumulative test. In contrast, the daily N tests can generally be assumed to be independent in time, although this is not entirely correct because of the correlations due to ongoing underlying physical processes (see above discussion).

5.2. Consistency Tests in Space

[43] The S test measures the consistency of the spatial distribution of the daily forecasts with the spatial distribution of the observed earthquakes [*Zechar et al.*, 2010]. To isolate the spatial component of the forecasts, the forecasts are first scaled to the total number of observed earthquakes; then, for each individual spatial cell, we sum over the expected rates in the magnitude bins to end up with a normalized spatial forecast. To measure the consistency between this spatial forecast and the observations, we compare the log likelihood score of the observed locations given the spatial forecast with the log likelihood values that would be expected if the forecast were correct. To obtain the expected likelihood scores, we simulated 10000 synthetic samples consistent with the forecast model based on a Poisson distribution and compute their spatial likelihood values. We then calculate the quantile ζ of the simulated spatial likelihood scores less than the observed score. Low values of ζ indicate that the forecasted spatial distribution is inconsistent with the observations because the measured likelihood value is much lower than expected if the forecast were correct. A high value of ζ indicates that the observed score is much higher than expected, which is not taken as grounds for rejection because, to obtain the high score, the observed events fell right at the peaks of the forecasted distribution. The S test is therefore a one-sided test, and we reject models at the 95% confidence level whenever $\zeta < 0.05$.

5.3. Probability Gain in the S Test

[44] We determine the probability gain per earthquake for the S test, G_{LL_S}, following *Helmstetter et al.* [2006] to rank the models compared to the simple reference model. The probability gain per earthquake for the S test is defined as

$$G_{LL_S} = \exp\left[\frac{(LL_S(H^i) - LL_S(H^r)}{N_{Obs}}\right] \tag{7}$$

H^i refers to the models $i = 1$–11 with their joint log likelihood score $LL_S(H^i)$, H^r denotes the reference model (STEP-0) and its joint log likelihood score $LL_S(H^r)$, N_{Obs} the total number of observed events. Values larger than 1 indicate a probability gain per earthquake, values smaller than 1 a loss.

5.4. Testing Against Reference Model

[45] As final measure, we apply the R test to compare each model with the reference model [*Schorlemmer et al.*, 2007]. In the R tests, the total number and the spatial distribution of the model forecasts are tested against the reference model. The R test is evaluated at the 0.05 significance level and serves the information whether one forecast model may be rejected at the significance level tested against a reference model. This adds an additional piece of information considering the spatial distribution and the amount of forecasted rates: a test hypothesis cannot be rejected in case the forecasted rates in the space-time-magnitude bins obtain a better likelihood score compared to the ones of the null hypothesis.

[46] In the R test, both models, H^r and H^i, are used as null hypothesis to test against. For likelihood ratio tests of model i versus the reference model r, LL^{ir} indicates that the reference model is the null hypothesis and model i is the test hypothesis; vice versa for LL^{ri}. The log likelihood ratio is defined as the difference

$$LL^{ir} = LL(\Omega|\Lambda^i) - LL(\Omega|\Lambda^r) \tag{8}$$

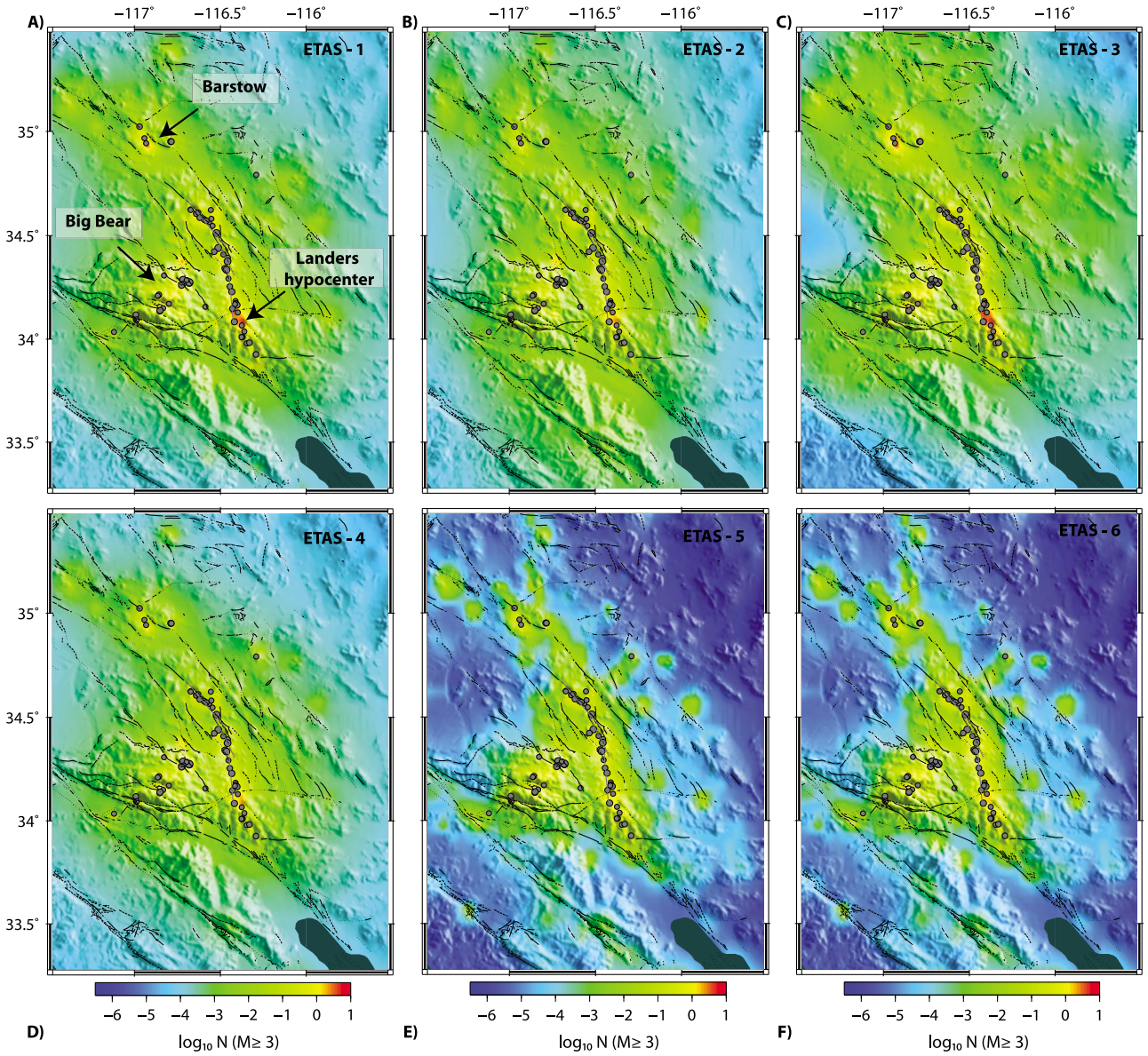

Figure 2. Seismicity rate forecast maps for day 4 after the Landers main shock for six different ETAS models. Shown is the logarithm of the expected number of earthquakes per day per 0.05 by 0.05°. Earthquakes with $M_L \geq 3$ on day 4 are superimposed as gray dots. Surface projections of faults (black lines) and the *Wald and Heaton* [1994] fault model (thick black lines) are shown.

where Ω denotes the vector of observed earthquakes, Λ^i and Λ^r are the forecasts of hypotheses i and r. $LL(\Omega|\Lambda^i)$ and $LL(\Omega|\Lambda^r)$ are the joint log likelihoods of the models. In contrast to the log likelihoods of the S test (LL_S) the log likelihoods of the R test include the information of each space-rate-magnitude bin; in the S test, rates were collapsed to one magnitude bin and normalized. We compute the quantile scores γ based on 10000 simulations following the method of *Schorlemmer et al.* [2007].

[47] As we are testing a time series of forecasts, we define the rejection ratios for both ways of testing, $R(LL^{ir})$ and $R(LL^{ri})$, respectively. The rejection ratios denote the fraction of cases the test hypothesis is rejected at the significance level in favor of the null hypothesis by comparing the quantile score with the predefined significance level. Small values of $R(LL^{ir})$ ($R(LL^{ri})$) imply that the test hypothesis is superior.

6. Results

[48] Prior to diving into the statistical analysis of the forecasts, we describe some model features by comparing snapshots of the model forecasts for day 4 (Figures 2 and 3). By day 4, abundant seismicity is available to which the models can adjust their parameter values. Figures 2 and 3 display the base 10 logarithm of the daily expected number of earthquakes $M_L \geq 3$ for each model along with the actually observed earthquakes of magnitude $M_L \geq 3$.

[49] Models ETAS-1 to ETAS-3 indicate a similar smooth distribution of seismicity rates with some spots of

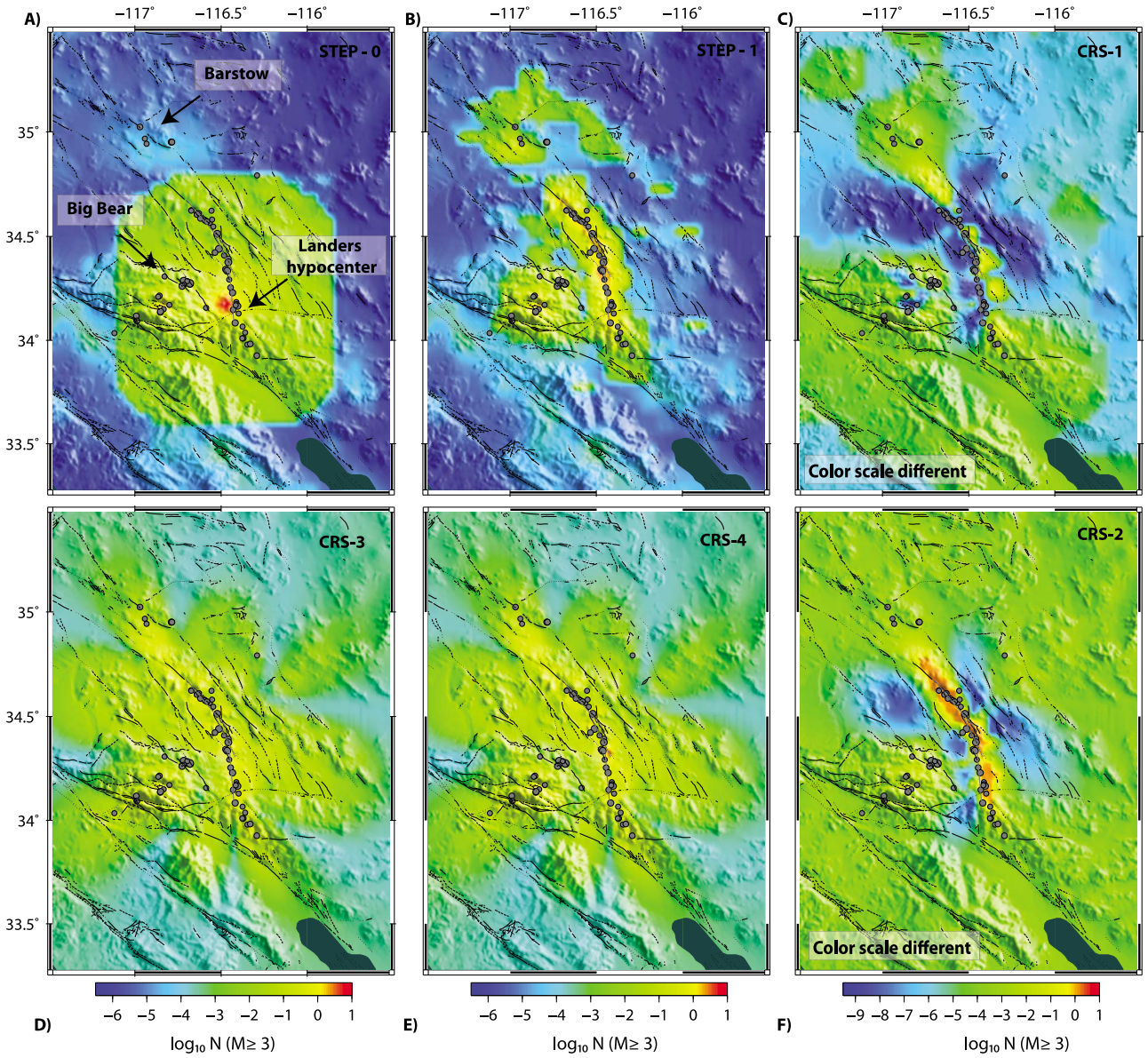

Figure 3. Seismicity rate forecast maps for day 4 after the Landers main shock for models STEP-0, STEP-1, and the four CRS models. Shown is the logarithm of the expected number of earthquakes per day per 0.05 by 0.05°. Earthquakes with $M_L \geq 3$ of day 4 are superimposed as gray dots. Surface projections of faults (black lines) and the *Wald and Heaton* [1994] fault model (thick black line) are shown.

higher rates at the hypocentral areas of the Landers and the Big Bear event and in the Barstow region (Figure 2); model ETAS-4 employs an even smoother distribution while models ETAS-5 and ETAS-6 differentiate strongest between areas of recent activity. The differences are subtle but cause differences as will be discussed in section 6.2 where we discuss the spatial consistency test. The highest forecasted rates appear at the same locations, however, they are smoothed out differently. For example, ETAS-1 and ETAS-3 assign higher rates ($\log_{10} N(M \geq 3) \geq 0.5$) on the spot of the Barstow cluster (red spot indicated with the arrow in Figure 2) compared to the other ETAS models ($\log_{10} N(M \geq 3) \leq -0.5$).

[50] The stationary background model for ETAS-6 and STEP-1 is the same in terms of the spatial distribution while the seismicity rate of the background is smaller compared to other models. The background is defined as $\mu(x, y) = \mu_0 p(x, y)$ where $p(x, y)$ is a normalized distribution and μ_0 is estimated for ETAS-6 and fixed (Table 3).

[51] Including the fault model information of by *Wald and Heaton* [1994], STEP-1 exhibits its ability to adjust to the aftershock sequence (Figure 3b) compared to the reference model STEP-0 (Figure 3a) that does not take advantage of the available fault model information. Seismicity rates are concentrated along the causative fault segments with the rates being influenced by the spatially varying seismicity parameter values. Similarly to the ETAS models (Figure 2), the increased rates in comparison to the background model are not only seen along the Landers fault, but also in the region around the magnitude $M_L = 6.5$ Big Bear aftershock

and less pronounced to the north in the Barstow region. Punctual rate increases spread out through the entire forecast box, however, less intense than in the ETAS models. Model STEP-0 concentrates the rates at the Landers and the Big Bear hypocenters and the rates are radially smoothed away from these two points. The rate variations outside the yellow colored circle results from the background model.

[52] CRS-1 and CRS-2 show the strongest imprints of the Coulomb stress change calculations in the form of the lobes traditionally associated with the Coulomb theory (Figure 3); note that the color scale ranges from −9.5 to 1 for these two models. As CRS-1 calculates the stress changes for aftershock mechanisms identical to the average mechanism of the Landers main shock rupture, the regions of increased rates are much smaller than those of the other CRS models which calculated the stress changes for optimally oriented faults. The observed seismicity on day 4 does not match the forecast of CRS-1 well. CRS-2, on the other hand, appears to forecast the earthquakes relatively well. The model predicts a concentrated region of increased rates along the causative faults. In CRS-3 and CRS-4, the traditional spatial imprint of the Coulomb stress change shadows is no longer existent as uncertainties in the stress change calculations for the forecasts are included. The uncertainties in the stress calculations have the effect of removing stress shadows (regions of negative Coulomb stress changes), resulting overall in an increased rate of seismicity [*Hainzl et al.*, 2009]. Contrary to the statistical models, CRS-3 and CRS-4 concentrate high seismicity rates not at the hypocenters of the Landers and the Big Bear event but close to the ends of the individual fault model segments.

6.1. Evaluating Forecasted Rates: N Test Results

[53] We start the quantitative model evaluation by comparing the daily number of predicted earthquakes N_F with the number of observed earthquakes N_{Obs} as a function of time since the Landers main shock (Figure 4). The observed daily rate varies strongly from day to day during the 90 day testing period but can, on average, be modeled with an Omori-Utsu law decay. We test the forecasts against observed earthquakes $M_L \geq 4$ on the first day because the catalog is not complete down to $M_L \geq 3$, against which we test the remainder of the days (see Figure 1).

[54] In absolute numbers, the forecasts of all models deviate the most from the observed number of shocks in the beginning of the sequence. The first several hours to days after the main shock are the most critical of the sequence in terms of hazard and productivity for which the least information is available due to earthquake detectability issues [*Woessner and Wiemer*, 2005]. Therefore, it is not surprising that the largest differences between N_F and N_{Obs} exist early on. Toward the end of the 90 day testing period, the number of events fluctuates between 0 and 10 per day, and the differences between N_F and N_{Obs} decrease.

[55] Models ETAS-1 and ETAS-2 forecast the same total number of events because they differ only in their spatial distribution function (Figures 4a and 4b). Models ETAS-3 and ETAS-4, the stationary and nonstationary realizations, respectively, by *Lombardi et al.* [2006], display the largest day-to-day variability in the forecasts. These two models seem to be most sensitive to the reestimations of the model parameter values. A numerical reason for the larger variability in ETAS-3 and ETAS-4 might be the smaller amount of 1000 Monte Carlo simulations compared to 10000 used for ETAS-1 and ETAS-2 that can cause larger fluctuations in the mean number. The large fluctuations refer to the heavy tail distribution of ETAS forecasts [*Lombardi and Marzocchi*, 2010]. Comparing STEP-1 with its simplified version STEP-0 (Figures 4a and 4b, black and gray line), the forecasts by STEP-0 decay more smoothly because it considers only events above $M_L \geq 6$ to contribute to earthquake triggering, while STEP-1 includes all magnitude $M_L \geq 2.5$ events.

[56] The N test evaluates the ability of a model to correctly forecast the number of observed earthquakes. The daily tests result in large variations of the quantile scores δ_1 and δ_2 on each test day due to the fluctuations in observed seismicity. Such fluctuations are expected and by defining a significance level of 0.05 to reject a forecast, we would expect a perfect model to be rejected in 5% of the cases; for a 90 day test period one can expect a perfect model to be rejected on four test days. The cumulative tests (Figure 5) provides insight in the overall performance of a model with a memory of the daily performance.

[57] The forecasts for day 1 may partly suffer from poor parameter values that stem from previous seismicity; parameter values could not yet be adapted to the Landers sequence. The only additional information for day 1 is the magnitude of the main shock. The number of forecasted events $M \geq 4$ on day 1 ranges between $N_F = 11.24$ and 73.85 while $N_{Obs} = 62$ events were observed (see Table 5). A model is not rejected at the 0.05 significance level by the N test if the predicted number of events lies between 48 and 79. This range is met only by the two STEP models. The latter two models are the only ones to overestimate the number of events; all other models underestimate N_{Obs}. Except for CRS-1, all models can reduce the difference between N_F and N_{Obs} over time (Figure 4), mainly because the models reestimate parameter values to adjust to the Landers sequence.

[58] For all but two models (ETAS-4 and CRS-1), the rejection ratios for the daily N tests R_N are lower than $R_N = 0.16$ (see Table 5), implying that in less than 16% of the test days (12 days) the forecasted rate of seismicity is rejected. This result is encouraging as one expects from a perfect model with the same test to have about 5% or 4 forecasts in this sequence to be rejected. The rejection ratios of ETAS-4 and CRS-1 are $R_N = 0.31$ and 0.72, respectively; this means these models perform by far worse than the other models in this test.

[59] Furthermore, we observe a strong variability of model scores from day to day. This means that one cannot draw reliable conclusions from a good performance on one day to an also good performance on the next day. One reason for this variability reflected in the quantile scores $\delta_{1,2}$ of the daily N tests, arises from the variability in the number of observed events. All forecast models result in a much smoother time series of N_F compared to N_{Obs} (Figure 4). The ETAS-type models show the strongest fluctuations followed by the modified STEP-1 model.

[60] The cumulative tests provide the most robust insight in the reliability of the models because all available data is used and therefore the N tests gains power with time. In Figure 5, we plot both quantile scores δ_1 and δ_2 to show in

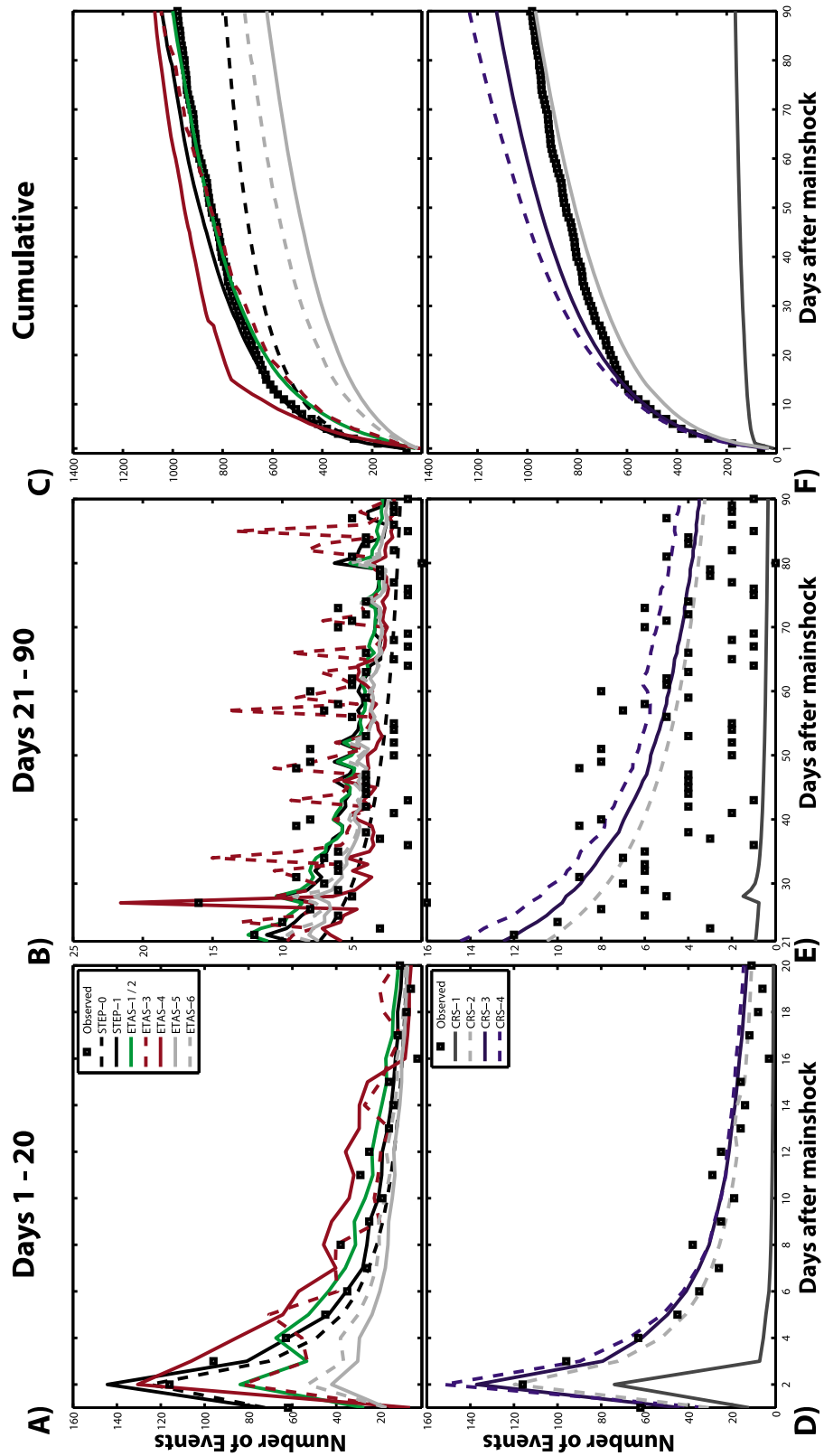

Figure 4. Observed (squares) and expected number of earthquakes $M_L \geq 3$ ($M_L \geq 4$ for day 1) in the testing region versus number of days since the Landers main shock. Daily seismicity rate forecasts by statistical models on (a) days 1–20 and (b) days 21–90; (c) cumulative seismicity (ETAS-1 and ETAS-2 generate identical rates). Daily seismicity forecasts by CRS models on (d) days 1–20 and (e) days 21–90; (f) cumulative seismicity rates.

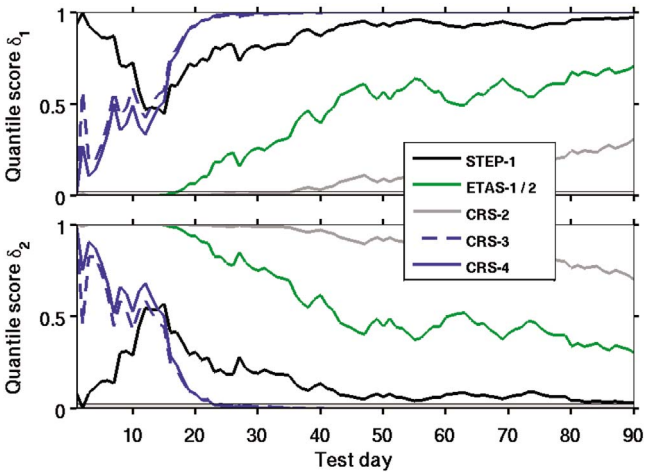

Figure 5. Quantile scores (top) $\delta_1(t)$ and (bottom) $\delta_2(t)$ of cumulative N test as a function of time. Gray patch on Figure 5 (bottom) indicates 0.05 significance level at which a rate forecast are rejected. Models that are rejected during entire period are not shown. Small $\delta_1(t)$ and $\delta_2(t)$ values indicate underestimation and overestimation of the seismicity level by a model forecast, respectively.

which of the one-sided tests the models are rejected. The quantile scores mainly show a complementary behavior ($\delta_2 \approx 1 - \delta_1$) because they become only very different in cases of very low forecasted rates; this case does not occur in this aftershock sequence but often in long-term forecasts or in daily forecasts for large regions [*Woessner et al.*, 2010].

[61] The cumulative number tests reveal the following characteristics of the statistical models (Figure 5): STEP-1 (black) generates forecasts consistent with the total observed number and is only rejected once on day 2. Toward the end of the testing period (starting around day 35), the model seems to overpredict slightly, and therefore $\delta_2(t)$ decreases gradually and comes closer to rejection. ETAS-1 and ETAS-2 are rejected during the first 15 days but model the sequence better with time, although a tendency to overestimate is also visible. The other models (ETAS-3, ETAS-4, ETAS-5, ETAS-6, and CRS-1) poorly forecast the rates during the first days of the sequence and are therefore rejected from the beginning onward and never obtain quantile scores that are larger than the rejection level; thus we do not plot these results. The abundant failing of different flavors of ETAS models may be due to the bias reported by *Lombardi and Marzocchi* [2010] and described above. CRS-2 underpredicts the cumulative seismicity in the entire time period (Figure 4f) but cannot be rejected at the chosen significance level anymore after 35 days. In contrast, models CRS-3 and CRS-4 perform well in the beginning and are not rejected during the first 22 days, but thereafter overestimate the seismicity rate. Common to all the CRS models is the fact that they overestimate the rate of seismicity with advancing time: rate-and-state theory constrains the exponent of the Omori-Utsu law, the p value to be less smaller than 1[*Dieterich*, 1994; *Cocco et al.*, 2010]. This might be the main reason why the CRS models cannot adjust to a faster decay observed during the later stage of the sequence. The effect is even more pronounced in model CRS-4, which includes poroelastic effects in the computations of the stress changes.

[62] In summary, we find that models CRS-3 and CRS-4 and models STEP-1 match the cumulative number of observed events well in the first 15 days while toward the end of the testing period, models STEP-1, ETAS-1, ETAS-2, ETAS-3, and CRS-2 tend to fit best. The comparison also highlights that STEP-0, ETAS-5, ETAS-6, and CRS-1 tend to underestimate the cumulative number of events with time, whereas CRS-3 and CRS-4 and ETAS-3 and ETAS-4 tend to overestimate the seismicity rate.

[63] It is important to remember that the cumulative tests keep a memory of the daily performance of the forecasts which might be an undesirable feature: a strongly inconsistent forecast at any time can lead to a rejection (extreme quantile scores) from which a model might not recover even if subsequent forecasts are consistent with daily observations. This effect is, for example, observed for ETAS models which strongly underestimated the seismicity rate in the first days.

6.2. Testing Data Consistency in the Space

[64] We determine the daily and joint log likelihood score of the S test LL_S for each model to provide insight in its capabilities to match the observed spatial distribution of earthquakes for the entire testing period. The joint log likelihood scores for the S test are used to rank the models for their overall spatial performance (Table 6).

[65] Small negative numbers indicate a better fit to the data than large negative numbers. Model ETAS-5 matches the spatial distribution of events best, showing the highest log likelihood score ($LL_S = -2905.26$), ETAS-6 being not much different with $LL_S = -2907.27$. It is followed by all other ETAS models. Models ETAS-5 and ETAS-6, ETAS-1 and ETAS-2, and ETAS-4 form a group with a difference of about 300 units. Models ETAS-3, CRS-3, STEP-1 and CRS-4 form a group of models with similar LL_S values with a difference of 300 units, but are at least 400 units smaller than the score of model ETAS-4. Model CRS-1 obtains a log likelihood scores of negative infinity because the forecasted rates of the model are very small (Figure 4) and thus the spatially distributed rates become so small that some

Table 5. Rejection Ratios for Daily and Cumulative Modified N Test and Number of Forecasted Events for Day 1 with $M_L \geq 4$[a]

Model	R_N		N_F
	Daily	Cumulative	($M_L \geq 4$, Day 1)
STEP-0	0.17	0.92	72.46
STEP-1	0.09	0.01	73.85
ETAS-1	0.10	0.19	27.22
ETAS-2	0.10	0.19	27.22
ETAS-3	0.13	1.00	11.24
ETAS-4	0.31	1.00	20.13
ETAS-5	0.16	1.00	17.24
ETAS-6	0.13	1.00	17.24
CRS-1	0.72	1.00	12.18
CRS-2	0.10	0.39	30.59
CRS-3	0.08	0.73	43.93
CRS-4	0.10	0.77	32.30

[a]The rejections ratio expresses the percentage of days the daily forecasts are rejected at the effective significance level $\alpha_{eff} = 0.025$ of the modified N test [*Zechar et al.*, 2010]. On day 1, 62 events with $M_L \geq 4$ were observed.

Table 6. Joint Log Likelihood LL_S and Probability Gain Per Earthquake Gain(S) for All Models[a]

Model	LL_S	Gain(S)	Rank
STEP-0	−5187.40	1.00	
STEP-1	−4099.87	3.02	8
ETAS-1	−3160.40	7.86	4
ETAS-2	−3012.83	9.14	3
ETAS-3	−3708.66	4.50	6
ETAS-4	−3308.43	6.76	5
ETAS-5	−2905.26	10.19	1
ETAS-6	−2907.27	10.17	2
CRS-1	-inf	0.00	11
CRS-2	−5351.49	0.85	10
CRS-3	−3932.49	3.58	7
CRS-4	−4298.86	2.47	9

[a]The probability gain is computed against the reference model STEP-0. The rank denotes the comparative ranking based on the spatial predictive power of the models.

obtain a negative infinity value. This is essentially a consequence of predicting stress shadow for the specified receiver faults. The sum of the log likelihood scores then becomes very small. CRS-2 obtains a score that is smaller than the one of STEP-0 that is not ranked.

[66] Based on the ranking of the joint log likelihood score LL_S, we selected models ETAS-6, ETAS-2, ETAS-4, CRS-3, STEP-1, and CRS-2, representing the best performing models with updating parameter values and also different modeling approaches, ranked by the LL_S. We show the joint log likelihood score per spatial bin, i.e., the sum of the likelihoods over the 90 days in each bin (Figure 6). This highlights the locations where the forecasts match the observed data relatively well. The differences are only really of interest in areas (grid cells) where earthquakes actually observed. Light gray to light blue colors (log likelihoods very close to zero) indicate cells in which forecasts and

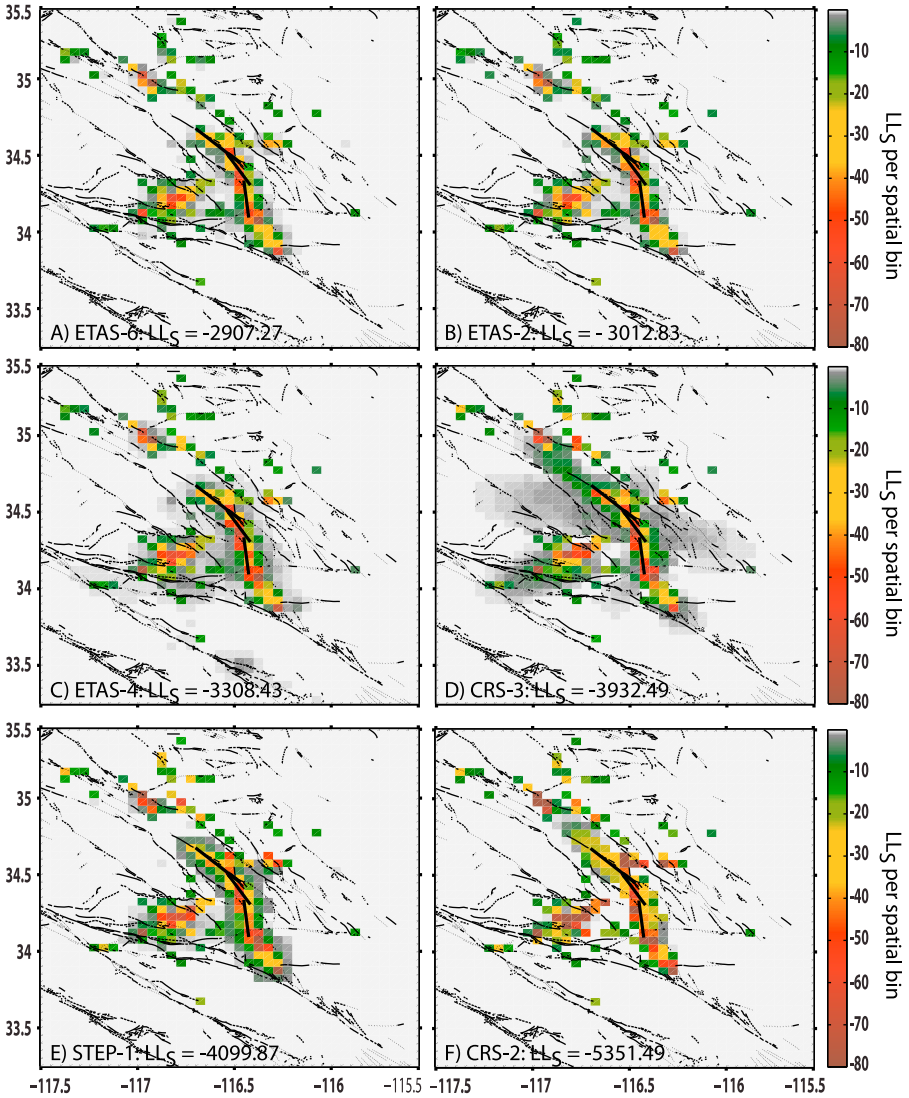

Figure 6. Map of log likelihood sum for each spatial bin (grid cell) at the end of the testing period. Model names and joint log likelihood sum LL_S (see Table 6) is given according to the LL_S ranking: (a) ETAS-6, (b) ETAS-2, (c) ETAS-4, (d) CRS-3, (e) STEP-1, and (f) CRS-2. Color scale is manually saturated at $LL_S = -80$ for comparison reasons; light gray regions indicate log likelihood scores very close to zero.

observed rates match, colors green to red denote large differences; the entire grid is covered, but by using light colors for almost zero numbers we highlight the regions that influence the log likelihood score most. The gray regions show in addition the regions to which most of the rates are distributed by the smoothing kernels. All models show worse fits in the Big Bear and the Barstow region, as well as close to the Landers rupture. In particular, the worst log likelihood scores are determined along the causative fault system and at the ends of the finite-fault model segments of *Wald and Heaton* [1994]. The southern end of the rupture zone in which the hypocenter is located results in a small log likelihood scores. The LL_S become generally better with increasing distance from the causative fault. This feature is expected for CRS models but less so for statistical models.

[67] The joint LL_S in Figure 6 do not consider the temporal evolution and also override the influence of specific days. In a retrospective experiment on an aftershock sequence, the starting date for the first forecast may strongly influence the overall result. Figure 7 shows the log likelihood scores $LL_S(Day\ 1)$ for the first day in the same order as in Figure 6. All the models (including those not shown) result in large negative $LL_S(Day\ 1)$ to the north of the Landers fault, i.e., in the Barstow cluster region, with the best values obtained by models ETAS-1 and ETAS-2. This is likely because they are smoothed out more gradually than the other models. Figure 7 gives a good impression of how the smoothing kernels perform for the first forecast indicated by the gray shading. The statistical models mainly distribute the seismicity from the epicenter radially while CRS models nicely include the fault structure. For ETAS-5 and ETAS-6 the radial components are superimposed by the anisotropic kernel of the large magnitude event. Surprisingly, model CRS-3 obtains the best log likelihood score on day 1 implying that there is a likelihood gain on selected days when using this model type; this points to the fact that only evaluating the overall score is insufficient. Movies of the daily snapshots of the S test for each model are available as auxiliary material.

[68] The relative performance of the ETAS models in the S test may be a direct consequence of the chosen spatial probability distribution function. ETAS-5 and ETAS-6 use a Gaussian kernel (equation (3), Table 3) while all other models apply a power law kernel (equation (4), Table 3), but differences are small. We do not observe a clear dependence on the number of free parameters (Table 1), in particular we cannot state that a more free parameters lead to a better fit; the number of free parameters is difficult to define for the various models because of their strategies to estimate parameters in space or in sequence for different model parts. We observe that the pattern of the CRS-3 leads to the best forecast in space as there is no radial smoothing involved and the rates are concentrated along the causative fault.

[69] The temporal evolution of the joint likelihood scores $LL_S(t)$ during the 90 day testing period presents the range of spatial log likelihood scores that the models achieve on different days (ETAS-6, ETAS-3, ETAS-4, CRS-3, STEP-1, and CRS-2; Figure 8). Days on which the observed log likelihood score falls within the 95% confidence interval of the simulated values (gray error bars) are indicated by green squares; red squares denote days on which the observed score falls outside the confidence limits. All models show an improvement from larger negative to smaller negative values with time starting at quite different levels on day 1, a range between −351 (CRS-2) and −197.79 (CRS-3). The improvement in the log likelihood scores is mainly due to the smaller number of events that is observed. A smaller negative $LL_S(t)$ and a smaller range mean that the spatial predictability of a model is higher compared to other models. On the first day, all models are rejected. Model ETAS-4 adjusts fastest its $LL_S(t)$ values to fall within the confidence interval on day 2. Models ETAS-5 and ETAS-6 need about 10 days to fall within the 95% confidence intervals. Model CRS-3 shows consistent daily forecasts starting at day 4 while model STEP-1 needs about 25 days to match the confidence interval for the first time, but still shows multiple failures afterward. Model CRS-2 remains generally rejected until day 40.

[70] As the final comparison to the observed data, we analyze the cumulative performance of the models to spatially forecast earthquakes in the Landers sequence with the one-sided S test quantile score $\zeta(t)$ (Figure 9). The cumulative quantile score $\zeta(t)$ measures whether a model is rejected at the 0.05 significance level (gray bar). The statistical and the CRS models are separately compared in the panels Figures 9a and 9b, respectively. As indicated by plotting the log likelihood scores $LL_S(t)$, model ETAS-4 is rejected only on the first day at $\alpha_{eff} = 0.025$. In sequence, models ETAS-4, ETAS-5, ETAS-6, ETAS-1, and ETAS-2 as well as CRS-3 are not rejected anymore after 5 days, followed by ETAS-3, STEP-1, CRS-2, and CRS-1. This is in agreement with the findings for the log likelihood scores (Figure 8) that indicate no probability gain in using the CRS models tested when looking at the cumulative scores to improve spatial forecasting ability, however, on single days such as day 1, better scores are observed.

6.3. Testing Against a Simple Reference Model

[71] We evaluate the performance of the eleven forecast models against the reference model STEP-0. In short, the forecasts of STEP-0 are not rejected by the cumulative N test during the first 7 days, but the model never passes the S test. Model STEP-0 achieves a joint log likelihood score of $LL_S = -5187.40$ which is better than the scores of models CRS-1 and CRS-2 (Table 6). The difference between STEP-1 and STEP-0 is $\Delta\ LL_S = 1086$ units, a difference that quantifies how the additional elements in STEP-1 improve the ability to forecast spatial heterogeneous seismicity.

[72] According to the log likelihood scores, we obtain the largest probability gain for model ETAS-5 and ETAS-6 (10.19 and 10.17, Table 6). Models ETAS-2, ETAS-1, and ETAS-4 follow with G_{LLS} ranging between 9.14 and 6.76; models ETAS-3, CRS-3, STEP-1, and CRS-4 result in a small probability gain with values around 3. Models CRS-1 and CRS-2 do perform worse than the reference model.

[73] The daily rejection ratios of the R tests $R(L^{ir})$ in Table 7 show that nine out of eleven models provide better forecasts than STEP-0 as values of $R(L^{ir}) < 0.3$ indicate that the test hypothesis is rejected in less than 30 percent of the tests; exceptions are the models CRS-1 and CRS-2. Including the cumulative rejection ratios $R(L^{ir})$ we find that only models ETAS-3 to ETAS-6 outperform the reference

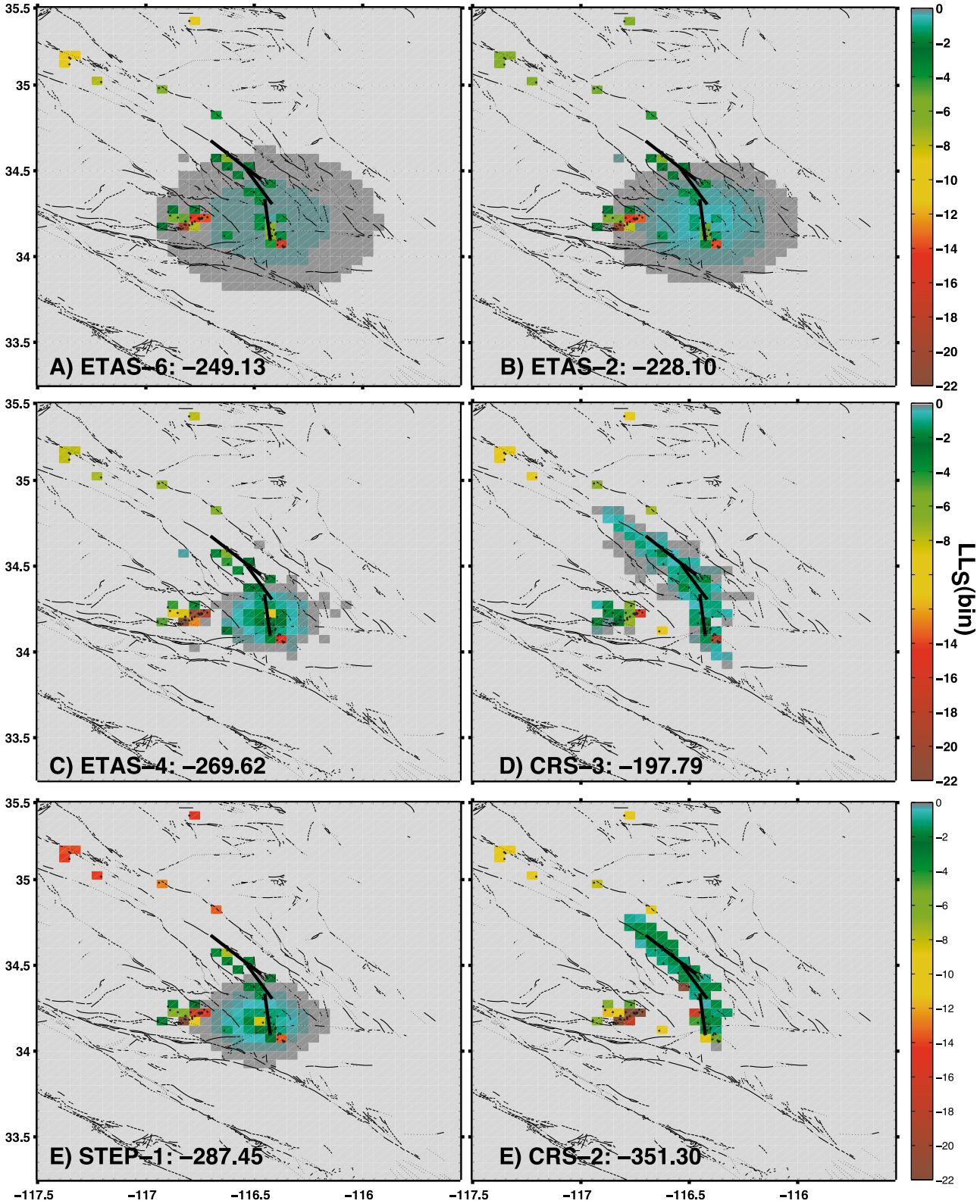

Figure 7. Map of log likelihoods for each spatial bin (grid cell) of test day 1. Model names and log likelihood sum LL_S(Day 1) of day 1 are given. Models are ordered as in Figure 6: (a) ETAS-6, (b) ETAS-2, (c) ETAS-4, (d) CRS-3, (e) STEP-1, and(f) CRS-2. Color scale is manually saturated at $LL_S = -22$ for comparison reasons; light gray regions indicate log likelihood scores very close to zero.

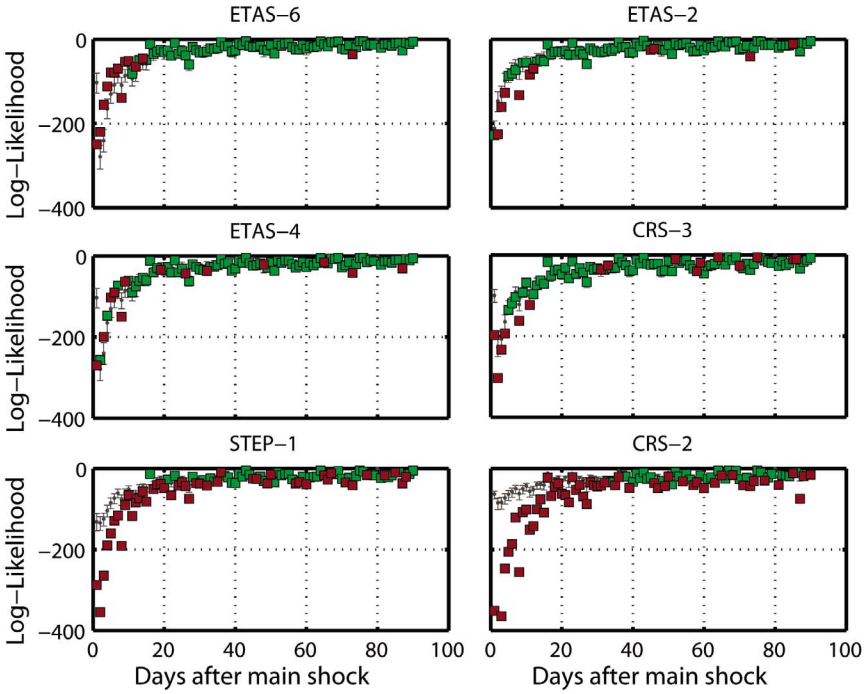

Figure 8. Log likelihood values of the S test as a function of days for model ETAS-6, ETAS-2, ETAS-4, CRS-3, STEP-1, and CRS-2. Displayed are the mean and the 97.5 and 2.5 percentiles (gray dot and bars); days on which a log likelihood value $LL_S(t)$ fall within the percentiles are indicated as green squares and outside with red squares.

model in the R test as the values are small for both. This is true also when the reference model is used as null hypothesis, $R(L^{ri})$. For models STEP-1, ETAS-1, ETAS-2, CRS-3, and CRS-4, we find small rejection ratios for the daily tests when the complex models are used as null hypothesis $R(L^{ir})$, however, this is not found for the cumulative tests. This discrepancy is related to large negative log likelihood scores of these models in the beginning of the testing series. Models CRS-1 and CRS-2 are rejected as null hypothesis at a high percentage of the daily tests (86% and 52% percent,

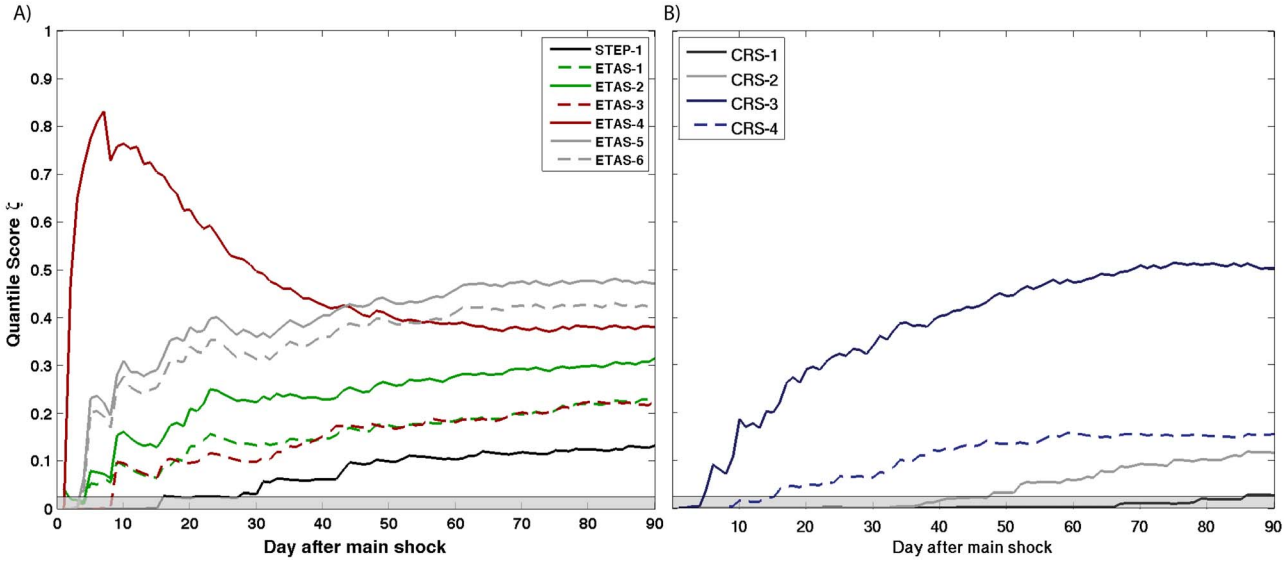

Figure 9. Quantile score $\zeta(t)$ for the cumulative S tests as a function of time for (a) CRS models and (b) statistical models. The significance level $\alpha_{eff} = 0.025$ is indicated as a gray patch at the bottom. In the time sequences, models ETAS-5, ETAS-6, ETAS-1, and ETAS-2, as well as CRS-3 are not rejected anymore after 5 days, followed by ETAS-3, CRS-4, STEP-1, CRS-2, and CRS-1.

Table 7. Rejection Ratios for Daily and Cumulative R Tests[a]

	$R(L^{ir})$		$R(L^{ri})$	
H^i	Daily	Cumulative	Daily	Cumulative
H^1 STEP-1	0.20	1.00	0.84	0.99
H^2 ETAS-1	0.12	0.92	0.90	1.00
H^3 ETAS-2	0.10	0.84	0.93	1.00
H^4 ETAS-3	0.03	0.02	0.87	1.00
H^5 ETAS-4	0.10	1.00	0.99	1.00
H^6 ETAS-5	0.03	0.01	0.92	1.00
H^7 ETAS-6	0.04	0.01	0.92	1.00
H^8 CRS-1	0.86	1.00	0.17	0.06
H^9 CRS-2	0.56	1.00	0.76	1.00
H^{10} CRS-3	0.20	0.99	0.80	1.00
H^{11} CRS-4	0.29	0.99	0.78	1.00

[a]The rejection ratio measure the percentage the null hypothesis can be rejected by the test hypothesis. In column $R(L^{ir})$, the hypothesis H^i forms the null hypothesis: small values of $R(L^{ir})$ denote that hypothesis H^i performs better H^r. Vice versa this is shown in column $R(L^{ri})$ for which the reference model STEP-0 is always the null hypothesis.

respectively) and are always rejected in the cumulative test. Model CRS-1 can also not reject the reference model as null hypothesis at high percentages.

7. Discussion

[74] The retrospective testing experiment shows that an overall ranking of short-term forecast models based on likelihood tests as used in CSEP is challenging because the single tests evaluate specific features. The N tests results indicate that model STEP-1 outperforms all other models when forecasting the total number of events, followed by CRS-3 and CRS-4 that perform well in the first 25 days; contrary, models ETAS-1 and ETAS-2 approximate the number of observed events closely starting on day 15 (Figure 5). Models ETAS-3 to ETAS-6 perform well in the daily N tests, but are rejected throughout the period by the cumulative N test because they cannot recover from a very low forecast during the initial days.

[75] In contrast to the N test, models ETAS-5 and ETAS-6 clearly show the best performance when testing the spatial consistency as found in the sum of the log likelihood scores followed by ETAS-1 and ETAS-2 (Table 6 and Figures 6, 7, and 9) while model STEP-1 has less predictive power in forecasting the spatial distribution correctly. In neither of the applied testing procedures can a CRS model outperform the statistical models, and only model CRS-3 and CRS-4 can compete with the statistical models. By including stress uncertainties computed from uncertain finite-fault source models, the regions of negative Coulomb stress changes are removed and also appear as regions of triggering seismicity [*Hainzl et al.*, 2009]. In other words, there are effectively no stress shadow regions in this model; in fact, higher seismicity rates are forecasted in comparison to the background (see Figure 3).

[76] The major challenge in matching the total number of events is related to finding appropriate initial parameter values for each model. The initial parameter values will become better constrained with time as more and more data of additional sequence becomes available. Thus, the challenge to perform better in the N tests is a question of the amount of data available for calibrating the initial parameter values that might vary with tectonic setting [*Schorlemmer and Wiemer*, 2005].

[77] The present results do not provide conclusive evidence for a best spatial kernel that the empirical models use to forecast the spatial distribution of triggered seismicity. The ETAS models in the experiment use either a power law or a Gaussian kernel, however, there is no systematic improvement of one kernel over the other. In addition, most models include additional components to improve the spatial forecast. For example, ETAS-5 estimates the distribution of aftershocks by smoothing the locations of early aftershocks and thereby obtains an anisotropic forecast. To make progress on this issue, detailed testing of one model using different kernels applied on multiple aftershock sequences and on other scales will be necessary. In addition, studies in which earthquakes are related to faults with quantitative measures such as provided by *Powers and Jordan* [2010] and *Wesson et al.* [2003] are necessary. Simple smoothing with a r^{-x} decay as applied in the generic element of the STEP-1 model prove to be too simplistic without further modifications depending on the background rate distribution as for example implemented in ETAS-4 or in ETAS-2. Spatially mapping seismicity parameters improves forecasting capabilities (STEP-1 compared to STEP-0) and thus further supports an alley that future models may follow (Figure 3) [*Gerstenberger et al.*, 2005] in addition to leverage preexisting fault structures.

[78] Our results show that the traditional fixed-receiver approach of combining Coulomb stress change calculations with rate-and-state theory (CRS-1) cannot compete with other CRS models despite the included effect of $M_L > 4.5$ aftershocks. Instead, it seems essential to resolve stress changes on 3D optimally oriented faults; however, the experiment is not comprehensive enough to conclude onto which fault planes Coulomb stress changes should be resolved for seismicity rate forecasts. We have for example not tested models that resolve stress changes onto the predominant geologic structures and then computed then estimates seismicity rates on those [*McCloskey et al.*, 2003; *Steacy et al.*, 2005a]; we assume that this could have a strong influence on the test results.

[79] The results of model CRS-3 and CRS-4 suggest that, to compete with empirical-statistical models, elements of strong stochasticity need to be included in the physics-based models. In particular, the large uncertainties in stress calculations have to be taken into account. However, both models CRS-3 and CRS-4 do this in a very simplified way which effectively leads to a transformation of the stress shadows into regions of increased seismicity. More sophisticated considerations of the involved uncertainties need to be done in future.

[80] Including poroelastic effects in calculating Coulomb stress changes does not improve the forecasts according to the results. From a physical point of view, this may be surprising as the existence of fluid flow in the Landers region has been suggested [*Bosl and Nur*, 2002]. One reason could be the choice of hydraulic parameters for the crust, yet, we are not able to fully investigate this with the measures we take in this study. In addition, there are multiple

models that describe the effect of earthquake triggering associated with fluid flow that could also be appropriate [e.g., *Miller*, 2002; *Miller et al.*, 2004].

[81] The results of the R tests with which eleven models are compared to a reference model (STEP-0) reveal the superiority of nine models as the rejection ratios are small (Table 7) in the daily tests; exceptions are the models CRS-1 and CRS-2. In the cumulative tests some of the models are likely to suffer from the discrepancy in forecasting the total number of seismicity in the beginning of the sequence. The probability gains for the S test support this result as the spatial forecasting ability of all statistical models clearly outscore the reference model (Table 6).

[82] As pointed out earlier, the simulations to obtain confidence levels for rejecting a hypothesis use the Poisson distribution which may lead to reject a model forecast in more cases than is actually true [*Lombardi and Marzocchi*, 2010]. A solution to this problem would be to simulate, or otherwise provide, the full probability density distribution of seismicity rates in each space-time-magnitude bin by each model. *Werner and Sornette* [2008] proposed two pathways: The first involves propagating model and parameter value uncertainties into forecasts by simulations [*Rhoades et al.*, 1994] while the second solution involves the idea of sequential data assimilation. Forecasts evolve through time the model forecast (prior) taking into account uncertainties in parameters and past data, and correcting the forecast using uncertain data (the likelihood) through Bayes' theorem. Future experiments should consider allowing models to provide the full distributions in each bin. For this experiment, we were not yet in the state to perform this task.

[83] We summarize the model performance by rejection ratios implicitly assuming that the tests are independent. For the daily tests, the assumption that a day-to-day forecast is independent for testing is technically fine; physically independence might not be justified considering that earthquakes are triggered by induced stress changes whatever mechanism is responsible. For the rejection ratios of the cumulative N test, we cannot assume temporal independence. The results may thus only give a crude rule-of-thumb-type information.

[84] In the models provided to the experiment, we did not use some information that might improve forecasts. For example, none of the models available to us include post-seismic relaxation or afterslip in the computation of Coulomb stress changes [e.g., *Perfettini and Avouac*, 2007]. Furthermore, none of the stress-triggering models account for stress changes due to small earthquakes, nor do any of the models include dynamic triggering [*Helmstetter et al.*, 2005; *Felzer et al.*, 2003; *Felzer and Brodsky*, 2006]. Including such models is a desirable extension of the experiments such as we performed and will provide the community with a better insight in the predictive skills of all models.

[85] The 24 h testing class is most suited for models that are using the updated information of seismicity catalogs. For example, STEP-1 takes advantage of updated seismicity to reestimate the spatial distribution of the seismicity parameters in the Gutenberg-Richter relationship and the Omori law. Similarly, ETAS models update parameters and forecasted event rates with incoming event information (except for ETAS-5). In contrast, models that are based on Coulomb stress change computations do not benefit as much from updating information as either only the main shock and the largest aftershock (Big Bear, M_L = 6.5) are used as stress steps (CRS-2 to CRS-4) or only aftershocks above magnitude M_L = 4.5 (CRS-1). None of the models include data down to small magnitudes which might improve forecasts of the physics-based models [*Marsan*, 2005; *Helmstetter et al.*, 2006].

8. Conclusion

[86] The main goal of this paper is to outline a strategy for a rigorous evaluation of short-term forecast models within a retrospective testing experiment. Despite the golden rule that forecast model evaluation and validation can only be done in truly prospective experiments as carried out in the CSEP framework, operational earthquake forecasts around the world may strongly benefit from a retrospective evaluation before standard CSEP experiments are started or in progress. Specifically, in this first community-based retrospective testing experiment we compared the ability of different statistical seismicity models and physics-based models to forecast the seismicity of a complex earthquake sequence. We designed the experiment providing the modelers the entire data set required (available as auxiliary material). We provided more information as would have been available in real time during the Landers earthquake sequence. Thus, we defined an experiment under ideal controlled conditions. We challenge the models by defining a distinct testing class to scrutinize their performance and test in particular the consistency with the observed data in terms of total forecasted number and their spatial distribution.

[87] The total number of observed earthquakes is satisfactorily forecasted only by models STEP-1 and ETAS-1 and ETAS-2 for the entire testing period. Models CRS-3 and CRS-4 perform well in the first 25 days and then start overpredicting. One reason is that the p value is limited to $p \leq 1$ in rate-and-state theory. The results suggest to further investigate the time, space and magnitude dependence of all parameters to improve forecasting abilities considering magnitude dependence of the p value as has been suggested [e.g., *Ouillon and Sornette*, 2005; *Hainzl and Marsan*, 2008].

[88] Epidemic Type Aftershock Sequence Models (ETAS) perform best in terms of forecasting the spatial distribution of data. Model ETAS-5 outperforms all other models in the overall S test log likelihood although parameters are estimated form data before the sequence. Model ETAS-6 performs slightly less well, however, the log likelihood differ only minor. Fluctuations in the parameter estimating procedure contribute to this effect. Models ETAS-1 and ETAS-2 can be considered to perform best overall when looking at the S test, the N test, and the rejection ratios.

[89] All models have particularly problems forecasting seismicity in regions where no events occurred during the learning period or close to the causative fault structure (Figure 6). Research on the distribution of seismicity and its relation to faults, either empirical or based on modeling procedures, are necessary to improve the predictability of the current models [*Wesson et al.*, 2003; *Powers and Jordan*, 2010].

[90] The results of the STEP-1 and ETAS models show that there is need to better define initial parameter values as models are penalized for strong deviations over long testing periods. This need may only be urgent when using these models for short-term forecasting such as in the current experiment and may be of less interest when using the same models for longer period forecasts. Cumulative log likelihood test results can strongly depend on testing results from single testing periods that lead to exceptionally small log likelihood scores. Some models suffer from forecasts of the first few days, for example, ETAS models that are based on parameter values poorly constrained by precursory seismicity.

[91] The model CRS-1 resolves stress changes on specified receiver faults creating large stress shadows. In terms of applying this approach to a sequence that generates seismicity on a variety of fault orientations, we showed that the estimated rates are by far too low and that computing the stress changes on 3D optimally oriented planes performs better. However, the approach did not consider any uncertainties in the receiver fault orientation nor any relation to predominant geologic structures; this should be tested in future experiments as such assumptions might improve the predictive power of CRS models [*McCloskey et al.*, 2003]. Furthermore, estimating parameter values from the first aftershocks instead from precursory seismicity provides better results. More details on the issue of optimally oriented planes versus fixed receiver faults and the sensitivity of CRS models are presented in *Cocco et al.* [2010].

[92] CRS models adding stress heterogeneities as a measure of uncertainty in the ΔCFS computations lead to improved rate forecasts compared to models disregarding these uncertainties. We conclude that it is necessary to include uncertainties when using CRS models for forecasting; however, this added stochasticity may not lead to a better understanding of the physical mechanism. In addition, further research is needed to properly propagate uncertainties through the models [*Hainzl et al.*, 2009].

[93] We investigated the performance of a suite of forecast models for one aftershock sequence which was partly used in the developments of the models themselves; thus, the ranking of the models is specific to the Landers sequence and not universal. To be able to draw conclusions about a models' performance for future sequences it is necessary to run this type of experiments on multiple aftershock sequences in a controlled environment and with various levels of data quality. We are working toward a collaboratory to retrospectively experiment with forecast models targeting prominent aftershock sequences in a similar manner but also other types of highly clustered seismicity such as swarms and human induced seismicity. The aim is to improve forecast models for highly clustered seismicity and to improve our insight in the predictive of the models. This can form the basis for the development of more general models that will ideally perform well in prospective forecast testing.

[94] **Acknowledgments.** M. J. Werner thanks A. Helmstetter for providing her computer code for ETAS-5 and ETAS-6. We thank D. Schorlemmer for pointing out the usefulness of tests against a reference model and for advising on the CSEP tests and M. P. Mai for providing finite-fault source models. We are grateful for the detailed comments of the two anonymous reviews and the Associate Editor. We are indebted to the SAFER-WP5 team for an effective cooperation and fruitful discussions. This work is part of the EU project SAFER, contract 036935. M.J.W. was supported by the EXTREMES project of the ETH Competence Center Environment and Sustainability (www.cces.ethz).

References

Aki, K., and P. G. Richards (2002), *Quantitative Seismology*, 2 ed., Univ. Sci. Books, Sausalito, Calif.

Aoi, S., B. Enescu, W. Suzuki, Y. Asano, K. Obara, T. Kunugi, and K. Shiomi (2010), Stress transfer in the Tokai subduction zone from the 2009 Suruga Bay earthquake in Japan, *Nat. Geosci.*, *3*, 496–500, doi:10.1038/ngeo885.

Bosl, W. J., and A. Nur (2002), Aftershocks and pore fluid diffusion following the 1992 Landers earthquake, *J. Geophys. Res.*, *107*(B12), 2366, doi:10.1029/2001JB000155.

Catalli, F., M. Cocco, R. Console, and L. Chiaraluce (2008), Modeling seismicity changes during the 1997 umbria-marche sequence (central Italy) through a rate- and state-dependent model, *J. Geophys. Res.*, *113*, B11301, doi:10.1029/2007JB005356.

Clinton, J. F., E. Hauksson, and K. Solanki (2006), An evaluation of the SCSN moment tensor solutions: Robustness of the M_W magnitude scale, style of faulting, and automation of the method, *Bull. Seismol. Soc. Am.*, *96*, 1689–1705, doi:10.1785/0120050241.

Cocco, M., F. Catalli, S. Hainzl, B. Enescu, and J. Woessner (2010), Sensitivity study of forecasts based on Coulomb stress calculation and rate-state frictional response, *J. Geophys. Res.*, B05307, doi:10.1029/2009JB006838.

Cohee, B. P., and G. C. Beroza (1994), Slip distribution of the 1992 Landers earthquake and its implications for earthquake source mechanics, *Bull. Seismol. Soc. Am.*, *84*, 692–712.

Cotton, F., and M. Campillo (1995), Frequency domain inversion of strong motions: Application to the 1992 Landers earthquake, *J. Geophys. Res.*, *100*(B3), 3961–3975.

Daley, D. J., and D. Vere-Jones (2003), *An Introduction to the Theory of Point Processes*, vol. 1, *Elementary Theory and Methods*, 2nd ed., Springer, New York.

Dieterich, J. H. (1994), A constitutive law for rate of earthquake production and its application to earthquake clustering, *J. Geophys. Res.*, *99*(B2), 2601–2618.

Dieterich, J. H., V. Cayol, and P. Okubo (2000), The use of earthquake rate changes as a stress meter at kilauea volcano, *Nature*, *408*(6811), 457–460.

Felzer, K. R., and E. E. Brodsky (2006), Decay of aftershock density with distance indicates triggering by dynamic stress, *Nature*, *441*, 735–738, doi:10.1038/nature04799.

Felzer, K. R., R. E. Abercrombie, and G. Ekström (2003), Secondary aftershocks and their importance for aftershock prediction, *Bull. Seismol. Soc. Am.*, *93*, 1433–1448.

Field, E. H. (2007), Overview of the Working Group for the Development of Regional Earthquake Likelihood Models (RELM), *Seismol. Res. Lett.*, *78*, 7–15.

Gerstenberger, M., S. Wiemer, and L. Jones (2004), Real-time forecast of tomorrow's earthquakes in California: A new mapping tool, *U.S. Geol. Surv. Open File Rep., 2004-1390.*

Gerstenberger, M. C., S. Wiemer, L. M. Jones, and P. A. Reasenberg (2005), Real-time forecasts of tomorrow's earthquakes in California, *Nature*, *435*, 328–331, doi:10.1038/nature03622.

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

Gutenberg, B., and C. F. Richter (1944), Frequency of earthquakes in California, *Bull. Seismol. Soc. Am.*, *34*, 185–188.

Hainzl, S., and D. Marsan (2008), Dependence of the Omori-Utsu law parameters on mainshock magnitude: Observations and modeling, *J. Geophys. Res.*, *113*, B10309, doi:10.1029/2007JB005492.

Hainzl, S., A. Christophersen, and B. Enescu (2008), Impact of earthquake rupture extensions on parameter estimations of point process models, *Bull. Seismol. Soc. Am.*, *98*, 2066–2072.

Hainzl, S., B. Enescu, F. Catalli, M. Cocco, R. Wang, F. Roth, and J. Woessner (2009), Aftershock modeling based on uncertain stress calculations, *J. Geophys. Res.*, *114*, B05309, doi:10.1029/2008JB006011.

Hardebeck, J. L., and E. Hauksson (2001), Crustal stress field in southern California and its implications for fault mechanisms, *J. Geophys. Res.*, *106*(B10), 21,859–21,882.

Harris, R. A. (1998), Introduction to special section: Stress triggers, stress shadows, and implications for seismic hazard, *J. Geophys. Res.*, *103*(B10), 24,347–24,358.

Hauksson, E. (1994), State of stress from focal mechanisms before and after the 1992 Landers earthquake sequence, *Bull. Seismol. Soc. Am.*, *84*, 917–934.

Hauksson, E. (2000), Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California, *J. Geophys. Res.*, *105*(B6), 13,875–13,903.

Hauksson, E., and P. Shearer (2005), Southern California hypocenter relocation with waveform cross-correlation, part 1: Results using the double-difference method, *Bull. Seismol. Soc. Am.*, *95*, 896–903, doi:10.1785/0120040167.

Helmstetter, A., G. Ouillon, and D. Sornette (2003), Are aftershocks of large Californian earthquakes diffusing?, *J. Geophys. Res.*, *108*(B10), 2483, doi:10.1029/2003JB002503.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2005), Importance of small earthquakes for stress transfer and earthquake triggering, *J. Geophys. Res.*, *110*, B05S08, doi:10.1029/2004JB003286.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2006), Comparison of short-term and time-independent earthquake forecast models for Southern California, *Bull. Seismol. Soc. Am.*, *96*, 90–106, doi:10.1785/0120050067.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007), High-resolution time-independent grid-based forecast of $m \geq 5$ earthquakes in California, *Seismol. Res. Lett.*, *78*, 78–86.

Hernandez, B., F. Cotton, and M. Campillo (1999), Contribution of radar interferometry to a two-step inversion of the kinematic process of the 1992 Landers earthquake, *J. Geophys. Res.*, *104*(B6), 13,083–13,099.

Jackson, D. D. (1996), Hypothesis testing and earthquake prediction, *Proc. Natl. Acad. Sci. U. S. A.*, *93*(3), 3772–3775.

Jones, L. E., and S. E. Hough (1995), Analysis of broadband records of the 28 June 1992 Big Bear earthquake: Evidence of a multiple-event source, *Bull. Seismol. Soc. Am.*, *85*, 688–704.

Jordan, T. H. (2006), Earthquake predictability: Brick by brick, *Seismol. Res. Lett.*, *77*, 3–6.

King, G. C. P., R. S. Stein, and J. Lin (1994), Static stress changes and the triggering of earthquakes, *Bull. Seismol. Soc. Am.*, *84*(1), 935–953.

Lombardi, A. M., and W. Marzocchi (2010), The assumption of Poisson seismic-rate variability in CSEP/RELM experiments, *Bull. Seismol. Soc. Am.*, *100*, 2293–2300, doi:10.1785/0120100012.

Lombardi, A. M., W. Marzocchi, and J. Selva (2006), Exploring the evolution of a volcanic seismic swarm: The case of the 2000 Izu Islands swarm, *Geophys. Res. Lett.*, *33*, L07310, doi:10.1029/2005GL025157.

Lombardi, A. M., M. Cocco, and W. Marzocchi (2010), On the increase of background seismicity rate during the 1997–1998 umbria-marche, central Italy, sequence: Apparent variation or fluid-driven triggering?, *Bull. Seismol. Soc. Am.*, *100*, 1138–1152, doi:10.1785/0120090077.

Marsan, D. (2005), The role of small earthquakes in redistributing crustal elastic stress, *Geophys. J. Int.*, *163*, 141–151, doi:10.1111/j.1365-246X.2005.02700.x.

Marsan, D. (2006), Can coseismic stress variabiltiy suppress seismicity shadows?, *J. Geophys. Res.*, *111*, B06305, doi:10.1029/2005JB004060.

Marzocchi, W., and A. M. Lombardi (2009), Real-time forecasting following a damaging earthquake, *Geophys. Res. Lett.*, *36*, L21302, doi:10.1029/2009GL040233.

McCloskey, J., S. S. Nalbant, S. Steacy, C. Nostro, O. Scotti, and D. Baumont (2003), Structural constraints on the spatial distribution of aftershocks, *Geophys. Res. Lett.*, *30*(12), 1610, doi:10.1029/2003GL017225.

Miller, S. A. (2002), Properties of large ruptures and the dynamical influence of fluids on earthquakes and faulting, *J. Geophys. Res.*, *107*(B9), 2182, doi:10.1029/2000JB000032.

Miller, S. A., C. Collettini, L. Chiaraluce, M. Cocco, M. Barchi, and J. P. Kaus (2004), Aftershocks driven by a high-pressure CO_2 source at depth, *Nature*, *427*, 724–727.

Mulargia, F. (1997), Retrospective validation of the time association of precusors, *Geophys. J. Int.*, *131*, 500–504.

Mulargia, F. (2001), Retrospective selection bias (or the benefit of hindsight), *Geophys. J. Int.*, *146*, 489–496.

Nur, A., and J. R. Booker (1972), Aftershocks caused by pore fluid flow?, *Science*, *175*, 885–887.

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

Ogata, Y. (1998), Space-time point-process models for earthquake occurrence, *Ann. Inst. Stat. Math.*, *50*, 379–402.

Ogata, Y., and J. Zhuang (2006), Space-time ETAS models and an improved extension, *Tectonophysics*, *413*(1–2), 13–23.

Ouillon, G., and D. Sornette (2005), Magnitude-dependent Omori law: Theory and empirical study, *J. Geophys. Res.*, *110*, B04306, doi:10.1029/2004JB003311.

Perfettini, H., and J. Avouac (2007), Modeling afterslip and aftershocks following the 1992 Landers earthquake, *J. Geophys. Res.*, *112*, B07409, doi:10.1029/2006JB004399.

Powers, P. M., and T. H. Jordan (2010), Distribution of seismicity across strike-slip faults in California, *J. Geophys. Res.*, *115*, B05305, doi:10.1029/2008JB006234.

Reasenberg, P. A. (1985), Second-order moment of central California seismicity, 1969–1982, *J. Geophy. Res.*, *90*(B7), 5479–5495.

Reasenberg, P. A., and L. M. Jones (1989), Earthquake hazard after a mainshock in California, *Science*, *243*, 1173–1176.

Reasenberg, P. A., and L. M. Jones (1990), California aftershock hazard forecast, *Science*, *247*, 345–346.

Reasenberg, P. A., and L. M. Jones (1994), Earthquake aftershocks: Update, *Science*, *265*, 1251–1252.

Rhoades, D. A., R. J. van Dissen, and D. J. Dorwick (1994), On the handling of uncertainties in estimating the hazard of rupture on fault segments, *J. Geophys. Res.*, *99*, 13,701–13,712.

Schorlemmer, D., and M. C. Gerstenberger (2007), RELM testing center, *Seismol. Res. Lett.*, *87*, 30–36.

Schorlemmer, D., and S. Wiemer (2005), Microseismicity data forecast rupture area, *Nature*, *434*, 1086, doi:10.1038/4341086a.

Schorlemmer, D., M. C. Gerstenberger, S. Wiemer, D. Jackson, and D. A. Rhoades (2007), Earthquake likelihood model testing, *Seismol. Res. Lett.*, *87*, 17–29.

Schorlemmer, D., J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, and T. R. W. Group (2010), First results of the regional earthquake likelihood models experiment, *Pure Appl. Geophys.*, *167*, 859–876, doi:10.1007/s00024-010-0081-5.

Steacy, S., J. Gomberg, and M. Cocco (2005a), Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard, *J. Geophys. Res.*, *110*, B05S01, doi:10.1029/2005JB003692.

Steacy, S., S. S. Nalbant, J. McCloskey, C. Nostro, O. Scotti, and D. Baumont (2005b), Onto what planes should coulomb stress pertubations be resolved?, *J. Geophys. Res.*, *110*, B05S15, doi:10.1029/2004JB003356.

Toda, S., R. S. Stein, P. A. Reasenberg, J. H. Dieterich, and A. Yoshida (1998), Stress transferred by the 1995 $M_W = 6.9$ Kobe, Japan, shock: Effect on aftershocks and future earthquake probabilities, *J. Geophys. Res.*, *103*(B10), 24,543–24,565.

Toda, S., R. S. Stein, K. Richards-Dinger, and B. S. Bozkurt (2005), Forecasting the evolution of seismicity in southern California: Animations built on earthquake stress transfer, *J. Geophys. Res.*, *110*, B05S16, doi:10.1029/2004JB003415.

Utsu, T. (1961), A statistical study of the occurence of aftershocks, *Geophys. Mag.*, *3*, 521–605.

Wald, D. J., and T. H. Heaton (1994), Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake, *Bull. Seismol. Soc. Am.*, *84*, 668–691.

Wang, R. J., and H. J. Kümpel (2003), Poroelasticity: Efficient modeling of strrongly coupled, slow deformation processes in a multilayered half-space, *Geophysics*, *68*(2), 705–717.

Wang, R. J., F. Lorenzo-Martin, and F. Roth (2006), PSGRN/PSCMP—A new code for the calculation of co– and post–seismic deformation, geoid and gravity changes based on viscoelastic–gravitational dislocation theory, *Comput. Geosci.*, *32*(4), 527–541.

Wells, D. L., and K. J. Coppersmith (1994), New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, *Bull. Seismol. Soc. Am.*, *84*, 974–1002.

Werner, M. J., and D. Sornette (2008), Magnitude uncertainties impact seismic rate estimates, forecasts and predictability experiments, *J. Geophys. Res.*, *113*, B08302, doi:10.1029/2007JB005427.

Werner, M. J., A. Helmstetter, D. Jackson, and Y. Kagan (2009), High-resolution long-term earthquake forecasts for California and Italy, *Geophys. Res. Abstr.*, *11*, EGU2009-12045.

Wesson, R. L., W. H. Bakun, and D. M. Perkins (2003), Association of earthquakes and faults in the San Francisco bay area using Bayesian inference, *Bull. Seismol. Soc. Am.*, *93*, 1306–1332.

Woessner, J., and S. Wiemer (2005), Assessing the quality of earthquake catalogs: Estimating the magnitude of completeness and its uncertainties, *Bull. Seismol. Soc. Am.*, *95*, 684–698, doi:10.1785/0120040007.

Woessner, J., A. Christophersen, J. D. Zechar, and D. Monelli (2010), Building self-consistent short-term earthquake probability (STEP) models: Improved strategies and calibration procedures, *Ann. Geophys.*, *53*, 141–154, doi:10.4401/ag-4812.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010), Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.*, *100*, 1184–1195, doi:10.1785/0120090192.

Zeng, Y. H., and J. G. Anderson (2000), Evaluation of numerical procedures for simulating near-fault long-period ground motions using

Zeng method, *Tech. Rep., 2000/01*, Pacific Earthquake Eng. Res. Cent., Berkeley, Calif.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002), Stochastic declustering of space-time earthquake occurence, *J. Am. Stat. Assoc.*, *97*, 7369–380.

F. Catalli, M. Cocco, A. M. Lombardi, and W. Marzocchi, Instituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605, I-00143 Rome, Italy.

B. Enescu, National Research Institute for Earth Science and Disaster Prevention, 3-1 Tennodai, Tsukuba, Ibaraki 305-0006, Japan.

M. C. Gerstenberger, GNS Science, PO Box 30-368, Avalon, Lower Hutt 5040, New Zealand.

S. Hainzl, GFZ German Research Centre for Geosciences, Section 2.1, Telegrafenberg, D-14473 Potsdam, Germany.

M. J. Werner, Department of Geosciences, Princeton University, Princeton, NJ 08544, USA.

S. Wiemer and J. Woessner, Swiss Seismological Service, ETH Zürich, Sonneggstr. 5, CH-8092 Zürich, Switzerland. (j.woessner@sed.ethz.ch)