Bulletin of the Seismological Society of America, Vol. 109, No. 1, pp. 1–18, February 2019, doi: 10.1785/0120170188

Ⓔ

Foreshocks and Their Potential Deviation from General Seismicity

by Stefanie Seif, Jeremy Douglas Zechar, Arnaud Mignan, Shyam Nandan, and Stefan Wiemer

Abstract It is still debated whether earthquake occurrence can be described as a single process or whether foreshocks are different phenomena. If foreshocks behaved differently, this would suggest a change of physical processes in the mainshock preparation phase that would boost hopes of forecasting large earthquakes. Most research on foreshocks focuses on case studies or uses global datasets in which recordings of small earthquakes are incomplete and are thus neglected. We do comprehensive foreshock statistics on all mainshocks in a regional earthquake catalog that is complete above **M** 2.5. To detect possible differences between foreshocks and seismicity that follows a uniform triggering model (the epidemic-type aftershock sequence [ETAS] model), we perform a null-hypothesis test. We also estimate the size of the differences between observed and ETAS-simulated foreshocks.

We define different sets of foreshocks using two different methods, because there is no unique definition: a nearest-neighbor declustering technique (Zaliapin *et al.*, 2008) and a variety of space–time windows (e.g., Agnew and Jones, 1991). We use data from southern California, northern California, and Italy. For each region, we first search an appropriate null model: an ETAS model that describes aftershock numbers well. In southern California, we find an appropriate spatiotemporal model that is characterized by a large productivity parameter α. After performing a null-hypothesis test for different mainshock and foreshock magnitudes, we find foreshock signals ($p < 0.05$) for all mainshocks sizes and independent of the foreshock's lower magnitude threshold. Observed mainshocks have more foreshocks than the ETAS model predicts.

Electronic Supplement: Test whether a single epidemic-type aftershock sequence (ETAS) model describes general seismicity, with figures showing p-values of the null hypothesis.

Introduction

Investigating the existence of foreshocks and their spatiotemporal patterns is very important for predicting large earthquakes. A famous example of a successful prediction is the Chinese Haicheng earthquake in 1975, for which correct evaluation of foreshock signals saved thousands of lives (Wang, 2006). Such social impact makes foreshocks an important research topic.

However, the question of which mechanisms are responsible for foreshocks is still disputed. There are two end-member views that can be defined as bottom-up triggering and top-down loading (Mignan, 2014). Bottom-up triggering has already been extensively investigated. Self-organized criticality (Bak *et al.*, 1988) and the epidemic-type aftershock sequence (ETAS) model (Ogata, 1988, 1998), the most frequently used statistical model, are examples of bottom-up triggering. Bottom-up triggering assumes that earthquakes are either triggered by preceding events or are background events that occur at a constant rate. Because all earthquakes have the potential to trigger others, the categories of foreshocks, mainshocks, and aftershocks are not mutually exclusive.

The alternative hypothesis is top-down loading, in which precursors occur due to tectonic loading caused by aseismic slip (Das and Scholz, 1981; Kanamori, 1981; Ohnaka, 1992; Dodge *et al.*, 1995, 1996; Mignan, 2012). Under this view, foreshocks are expressions of minor stress releases on asperities that have less resistive power than the large fault and hence break before it. Foreshocks and mainshocks would thus originate from the same underlying process and not interact, as bottom-up triggering proposes. The idea of tectonic loading as the underlying process is picked up, for example, in the non-critical precursory accelerating seismicity theory model by Mignan (2012).

Since the 1930s, numerous studies investigated the occurrence of foreshocks (e.g., Jones, 1984; Agnew and Jones,

1991; Dodge *et al.*, 1995, 1996; Abercrombie and Mori, 1996; Reasenberg, 1999; Kato *et al.*, 2012; Yagi *et al.*, 2014). Most of the studies that report anomalous foreshock patterns reach this conclusion after investigating single foreshock–mainshock sequences: they are case studies. Their focus is generally to study the nucleation process of large earthquakes and determine a favorable foreshock mechanism.

However, there is skepticism about whether foreshocks can be systematically identified, because it is still unclear whether foreshocks differ from normal seismicity or are actually temporary perturbations thereof. To answer this question, several studies investigated the occurrence and features of foreshocks more systematically by taking account of all mainshocks that happen in a certain area, providing a comprehensive picture of foreshock occurrence. Many of those studies concluded that ETAS is a good model for describing seismicity, including any claimed precursors (e.g., Helmstetter and Sornette, 2003; Felzer, 2004; Hardebeck *et al.*, 2008; Marzocchi and Zhuang, 2011). In short, past research has found: first, some large earthquakes are preceded by foreshocks; and second, there is strong evidence that those foreshocks are just coincidence and cannot be distinguished from normal seismicity. This would mean large earthquakes cannot be predicted based on foreshock activity, because in the ETAS model, the chances of a large earthquake following a small earthquake are very low.

We wish to challenge this conclusion, drawing on features that set our study apart from those mentioned above. Helmstetter and Sornette (2003) and Felzer (2004) could confirm that foreshock patterns are in line with the consequences expected under the ETAS model. However, they did not simulate synthetic ETAS catalogs, which would permit a one-to-one comparison between observations and ETAS. They set relatively high-cutoff magnitudes, above which data are considered in their analysis: $M_c = 3$ and 5.6, respectively.

Marzocchi and Zhuang (2011) compared foreshock occurrence in ETAS-simulated catalogs with observed catalogs for southern California and Italy. However, they used a high $M_c = 4$ that could obscure foreshock patterns, due to insufficient data. They compared the frequency of foreshock–mainshock pairs between ETAS and the catalogs included. By taking pairs as the statistical parameter of interest, in which a pair is defined as a mainshock preceded by at least one foreshock, valuable information about the number of foreshocks is neglected.

There are studies that concluded that the characteristics of foreshocks set them apart from general seismicity. Brodsky (2011) and Shearer (2012) observe an increased aftershock-to-foreshock ratio in comparison to ETAS. They consider relatively small mainshocks in their studies (in comparison to the current study), **M** 3–4 and **M** 2–5, respectively. Their final result could be biased because the conclusion depends on the aftershock rate that could be too low, due to incomplete aftershocks, as the authors state. In addition, their criteria for aftershock removal that removes mainshocks happening 3–4 days after a large earthquake might not be efficient, and a considerable amount of aftershocks might still remain in the foreshock sequences, which could bias the result. Lippiello *et al.* (2012, 2017) found a deviation between the spatial distribution of foreshocks and ETAS; however, they did not perform a statistical test to consolidate that conclusion. The study of Bouchon *et al.* (2013), which shows an exceptionally strong acceleration of seismicity before interplate mainshocks in comparison to ETAS, contains points of criticism that raise doubts about the conclusion. Felzer *et al.* (2015) criticized the study for using incomplete data and inappropriate ETAS parameters that could have caused the apparent differences.

Furthermore, none of the mentioned studies, except for Helmstetter and Sornette (2003) and Felzer (2004), ascertained whether foreshock patterns occur at different mainshock magnitudes. No study, to our knowledge, has investigated the influence that the lowest cutoff magnitudes have on foreshock patterns.

Compared to the aforementioned studies, our study differs in the following ways:

1. We include low-foreshock magnitudes ($M_c = 2.5$).
2. We choose the number of foreshocks as the statistical parameter of interest.
3. We perform a null-hypothesis test in which we compare values of the chosen statistical parameter between synthetic ETAS catalogs and the observed catalog.
4. We choose to evaluate the null hypothesis by comparing distributions, not averages of the number of foreshocks.
5. We also report the effect size: the difference between the number of observed and ETAS foreshocks.
6. We use different foreshock definitions (declustering and space–time windows) in our analyses.
7. We use different mainshock magnitudes ($M = \{4.5-5; 5-5.5; 5.5-6; 6+\}$) and cutoff magnitudes ($M_c = \{2.5, 3, 3.5, 4\}$) in our analyses, to determine whether large mainshocks produce different foreshock patterns and whether the foreshock patterns are more pronounced when potential foreshock data is increased.

By evaluating foreshock distributions instead of the statistics (e.g., mean or median) thereof, we avoid assuming that the statistics are representative of the data. p-values tell us that there is a difference between the null hypothesis and the alternative but do not allow us to conclude whether or not the difference is of practical importance. To provide a complete picture, we therefore also report effect sizes, quantifying both the size and the direction of the effect.

In this study, we are faced with having to define what a foreshock is. Defining a foreshock is not straightforward and, with this article, we do not aim to provide a universal definition, but we want to emphasize that each definition has its meaning and that conclusions should be drawn keeping these in mind. We decided to define foreshocks in two different ways: (1) in terms of varying space–time windows, and (2) with a nearest-neighbor stochastic declustering method

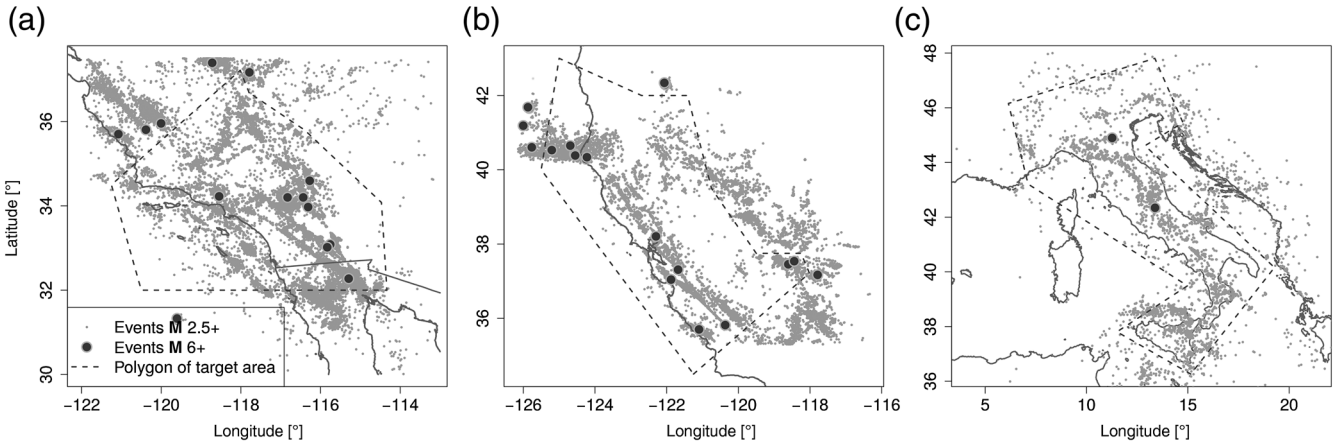

Figure 1. Map of seismicity with $M \geq 2.5$ in the studied regions. (a) Southern California (1981–2014), (b) northern California (1984–2015), and (c) Italy (2005–2016).

(Zaliapin *et al.*, 2008). We provide a more comprehensive description of these definitions in the Foreshock Definition and Aftershock Removal section. The method to identify anomalous foreshocks as presented here may be applied to any other foreshock definition in the future and hence might serve as a tool to further enhance our understanding of foreshocks. By investigating foreshocks of different mainshock magnitudes, we examine potential differences in the nucleation process between small and large mainshocks. By using different cutoff magnitudes, we improve the understanding of the role of small earthquakes.

The goal of this study is to answer the question whether the average foreshock behavior of a region deviates from the rest of the seismicity. With identified deviations, we could, in the future, define peculiar foreshock patterns and raise alarms for impending large earthquakes. For operational purposes, these patterns would ideally look the same in the whole region. Patterns valid for a whole region emerge from comparing observed foreshocks with rather simple ETAS models that describe the whole region with one parameter set. Therefore, we compare two rather simple but also one more complex (spatially variable) ETAS model in the search for a null model. In our study, we consider different regions, namely, southern California, northern California, and Italy. We verify whether our models are adequate to describe seismicity in the Results and Discussion section and in the Ⓔ electronic supplement. Then, with the chosen null model, we verify or disprove the foreshock null hypothesis in a null-hypothesis test. We find that peculiar foreshock patterns exist in southern California for all mainshock sizes, all lower magnitude thresholds of foreshocks, and both foreshock definitions.

Data

Our study draws on three different catalogs: southern California, northern California, and Italy. In Figure 1, we plot the earthquake data of the respective catalogs. In Table 1, we describe the data (catalog duration, spatial polygon, target, and auxiliary data) and indicate the corresponding completeness magnitude M_c and b-value. M_c is calculated using the Clauset method (Clauset *et al.*, 2009) and the b-value with maximum-likelihood estimation (Tinti and Mulargia, 1987) and data above M_c. Auxiliary events in space and time are used to estimate more reliable ETAS parameters (Wang *et al.*, 2010). We choose to investigate the three regions of our study separately, instead of performing a joint analysis of their foreshock data. This should highlight the potential regional differences in foreshock behavior.

Table 1
Description of Earthquake Data and Data Analysis

Catalog (yyyy/mm/dd)	Target Time (Start) (yyyy/mm/dd)	Target Space	Target Number of Earthquakes	Auxiliary Number of Earthquakes	M_c (St. Dev.)	b-Value (St. Dev.)
Southern California 1981/01/01 to 2015/01/01*	1986/01/01	ANSS polygon†	33,789 ($M_0 = 2.5$)	13,063 ($M_0 = 2.5$)	2.4 (0.06)	1.07 (0.01)
Northern California 1984/01/01 to 2015/12/31*	1989/01/01	ANSS polygon†	18,046 ($M_0 = 2.5$)	12,803 ($M_0 = 2.5$)	2.0 (0.51)	0.97 (0.03)
Italy 2005/04/16 to 2015/11/19, Depth < 35 km*	2006/04/16	Italy mainland polygon	5,960 ($M_0 = 2.5$)	1,272 ($M_0 = 2.5$)	2.0 (0.51)	1.01 (0.02)

St. Dev., standard deviation.
*Southern California: catalog by Hauksson *et al.* (2012); Northern California: catalog by Waldhauser (2009); Italy: catalog by Gasperini *et al.* (2013).
†Advanced National Seismic System (ANSS) authoritative regions.

Foreshock Definition and Aftershock Removal

One correct foreshock definition does not exist. Rather there are multiple possible foreshock definitions that reflect different aspects of foreshocks; for example, are they spontaneous or triggered, do they scale with mainshock magnitude, etc? In this article, we apply two possible foreshock definitions. The first one is a window approach that defines foreshocks inside a space–time window around the mainshock (e.g., Agnew and Jones, 1991; Marzocchi and Zhuang, 2011). It does not distinguish between a clustered and spontaneous event that are the two parts that our second foreshock definition consists of. The second definition is based on the declustering approach after Zaliapin *et al.* (2008) that is a nearest-neighbor approach that distinguishes between mainshocks, aftershocks, and foreshocks. The nearest neighbor is defined in means of distance, time difference, and magnitude, according to the following expression:

$$\eta_{ij} = t_{ij} \times r_{ij}^{d} \times 10^{-b(M_i - M_c)}, \tag{1}$$

in which η_{ij} is the nearest-neighbor distance between two events i (earlier event) and j (later event), t_{ij} is the time difference, r_{ij} is the distance, d is the fractal dimension of epicenters, b is the b-value of the Gutenberg–Richter law, and M_i is the magnitude of the earlier event. Earthquakes form clusters when their distances to their nearest neighbor η_{ij} are below a threshold η_0. We optimize for η_0 in each region of this study. The largest earthquake in a cluster is defined as the mainshock. Earthquakes that precede it and are in the same cluster are defined as its foreshocks. Hence, Zaliapin foreshocks are, per definition, clustered. By testing the null hypothesis with foreshocks defined with the window and the Zaliapin definition, we hope to obtain evidence whether anomalous foreshocks are composed of clustered or spontaneous earthquakes.

In both definitions, we consider epicentral distances between fore- and mainshocks. This choice probably has an effect on the result, and while it would be interesting to also study 3D distances, depth is not straightforward to simulate (and estimate) in ETAS, which leaves this investigation beyond the scope of this article. For the former definition, we use different windows, covering combinations of $\{3, 10\}$ days and $\{3, 10, 20, 40\}$ kilometers that include the ranges used in previous studies. We compare the seismicity preceding **M** 6+ earthquakes in southern California and ETAS catalogs in Figure 2 for different space–time windows.

We speculate that foreshocks and aftershocks are the results of different processes. So as not to obscure the foreshock process, we remove foreshock sequences that are biased by potential aftershocks. We define a foreshock sequence to be biased if a large earthquake (**M** 5+) occurs sufficiently close in space and time to the mainshock, such that the mainshock j is considered an aftershock of the large event i, rather than a spontaneous background event. We define this to be the case when the aftershock rate g_{ij} is larger than the background rate μ_j. $g_{i,j}$ is calculated from the triggering capabilities of the ith earthquake at the location and time of the jth event (equation 3), and μ_j is calculated by stochastic declustering and Gaussian kernel smoothing. The definition of large events as earthquakes with **M** 5+ is arbitrary but certainly includes earthquakes powerful enough to produce observable aftershock sequences. Table 2 sums up the number of large mainshocks (**M** 6+), removed mainshocks, and foreshocks for each catalog.

ETAS Model

To evaluate whether foreshocks might originate from a physical process different from the rest of seismicity, we formulate our null hypothesis: "all seismicity follows the same process and is hence described by the same model." We know that the main features of seismicity are well described as a composition of background and triggered seismicity (e.g., Console, 2003; Werner *et al.*, 2011; Zhuang, 2011). This composition also characterizes the ETAS model that we hence choose as our null model. In the following, we discuss shortcomings of the standard ETAS model and propose the choice of three models as potential null models. We introduce the formulation of our ETAS-base model, from which the other two can be derived, and explain the simulation procedure.

Null Models

ETAS is a model that is, by definition, imperfect, and it hence describes seismicity only to a certain degree. Fortunately, we know exactly which features it describes, namely those contained in the model formulation. Deviations between model and data are expected to occur as a consequence of our model's imperfection. We can learn from these deviations as they point to deficits of the model. Understanding the deficits helps us to understand the process better, as, for example, Ogata (1988) showed with his residual analysis.

Many researchers since then searched for reasons that could explain observed deviations. They found that the standard models mostly do not account for the following seismicity features: (1) temporally varying seismicity, (2) a spatially varying triggering rate, (3) missing aftershocks, (4) anisotropic aftershock distribution, (5) finite temporal and spatial triggering, and (6) magnitude-dependent parameters.

Ideally, all of these features should be implemented in the ETAS model. However, this is not straightforward, and to the present day, no model formulation has been developed that implements all features. Furthermore, there exists no study that quantifies different biases that originate from ignoring the features to determine the most important ones. Therefore, regarding our study, it is not feasible to aim for the "best" model as a null model. Our strategy is to define three null models, each considering different, but not all, features from lists 1–6 and compare them in terms of their deviation from the data.

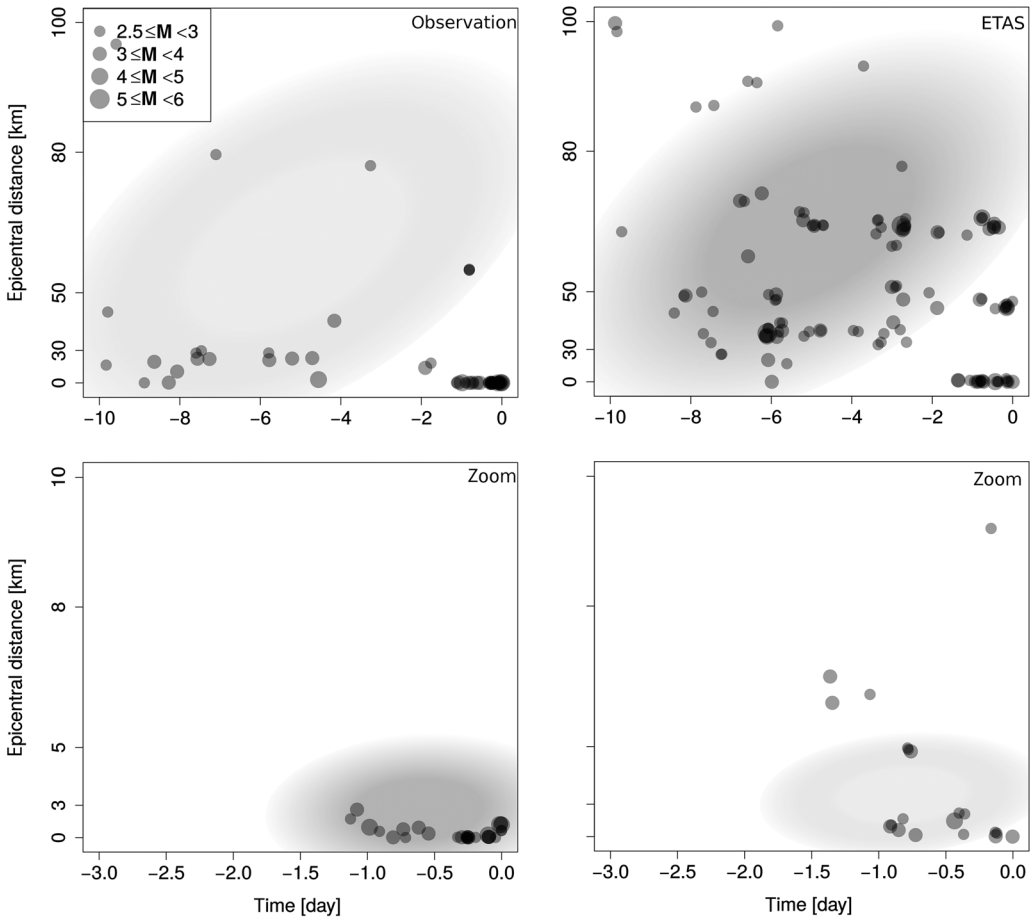

Figure 2. Distance–time plot of observed and epidemic-type aftershock sequence (ETAS) model seismicity that precedes **M** 6+ earthquakes and a qualitative interpretation of high- and low-earthquake density. The data is plotted semitransparently, with high-density areas appearing in darker shades, the shaded ellipses representing a qualitative interpretation of earthquake density. The distance is plotted proportional to its square to account for the 2D epicentral distances. Data concerning mainshocks occurring in a sequence of aftershocks are removed (see the Foreshock Definition and Aftershock Removal section).

These are, first, an ETAS-base with a spatially variable background rate and a uniform triggering function (described in the next paragraph) that does not account for any of the features in lists 1–6. Second, is an ETAS-fixed α, which is an ETAS-base with α fixed to β, with which we intend to account for incomplete aftershocks (e.g., Seif *et al.*, 2017), spatial anisotropy (Helmstetter *et al.*, 2005; Hainzl *et al.*, 2008), and a temporally varying background rate (Hainzl, 2013; Llenos and Michael, 2013). $\alpha = \beta$ also reproduces Båth's law (e.g., Felzer, 2002) and is in agreement with static-stress triggering models (Hainzl *et al.*, 2010). Third, is ETAS-var, which is an ETAS-base but with spatially variable ETAS parameters, as described in Nandan *et al.* (2017).

In the Results and Discussion section, we investigate to which degree each of these models is able to describe our data, through their performance in describing aftershocks. We choose the model with the least deviations from the data as the null model.

Parameter Estimation

Our base model, ETAS-base, is a spatiotemporal ETAS model, with a spatially variable background rate and a

Table 2
Number of Mainshocks, Foreshocks, and Removed Mainshocks

Catalog	Number of Maishocks with **M** 6+	Number of Removed Mainshocks	Number of Foreshocks with **M** 2.5+
Southern California	8 (6)	2 (1)	2,2,12,0,7,10 (2,12,0,7,23)
Northern California	13 (11)	3 (1)	0,0,0,0,0,1,0,0,0,0 (0,0,0,0,0,0,0,0,0,0)
Italy	2 (2)	0 (0)	4,3 (20,3)

The number of mainshocks, removed mainshocks (due to their occurrence in an aftershock sequence) and foreshocks for the different study regions. The values depend on the chosen foreshock definition: foreshocks are defined either within a space-time window of 3 days and 10 km or with the Zaliapin declustering (in parenthesis).

Table 3
Estimated ETAS Parameters at $M_{cut} = 2.5$

Catalog ($M_{cut} = 2.5$)	K_0	α	C	p	d	q	γ	μ	n	n (emp)
SC, ETAS-base	0.462	1.132	0.023	1.270	0.015	1.372	1.355	1.000	0.853	0.868
SC, ETAS-fixed α	0.038	2.450	0.024	1.270	0.006	1.506	1.958	1.001	0.470	0.781
NC, ETAS-base	0.272	1.239	0.013	1.270	0.009	1.539	1.960	0.998	0.600	0.588
NC, ETAS-fixed α	0.051	2.210	0.012	1.270	0.011	1.824	2.087	0.998	0.548	0.586
Italy, ETAS-base	0.395	1.172	0.003	1.167	1.127	2.140	0.734	1.005	0.789	0.803
Italy ETAS-fixed α	0.079	2.290	0.004	1.206	1.106	2.499	0.974	1.004	0.692	0.761

The parameters relate to the different study regions of southern California (SC), northern California (NC), and Italy. They are estimated with different epidemic-type aftershock sequence (ETAS) model formulations: ETAS-base and ETAS-fixed α (described in "ETAS model"). n is the branching ratio calculated with Gutenberg–Richter distributed or empirical magnitudes.

uniform triggering function. The other models, ETAS-fixed α and ETAS-var, can be derived from this formulation. The earthquake rate at time t and location x, y, considering the history $\mathcal{H}_t$ of previous events, is

$$\lambda(t,x,y|\mathcal{H}_t) = \underbrace{\mu \times bg(x,y)}_{\text{background rate}} + \underbrace{\sum_{j:t_j<t} g(t-t_j, x-x_j, y-y_j; M_j)}_{\text{triggering function}}, \quad (2)$$

with the triggering function

$$g(t,x,y;M) = \underbrace{K_0 e^{\alpha(M-M_{cut})}}_{\text{Productivity}} \times \underbrace{(t+c)^{-p}}_{\text{Omori type decay with time}} \times \underbrace{(x^2+y^2+de^{\gamma(M-M_{cut})})^{-q}}_{\text{decay with distance}}. \quad (3)$$

We estimate the ETAS parameters K_0, α, c, p, d, γ, q, μ individually for all regions by maximizing the log-likelihood function of equation (2). We consider data above $M_{cut} = 2.5$, because at this magnitude all catalogs are complete (see Table 1). To estimate the maximum likelihood, we normalize the Omori and distance decay in equation (3). We simultaneously invert for the background rate $bg(x,y)$ by stochastic declustering (Zhuang *et al.*, 2002) with the estimated ETAS parameters. Stochastic declustering results in a background probability for each earthquake that is used to weight Gaussian kernels around each earthquake for which the sum at x, y results in $bg(x,y)$. The Gaussian kernels have a variable bandwidth that is determined by their n_p nearest neighbors. We find the optimal n_p to be close to six, which is consequently used in this study. In southern and northern California, Omori's p is fixed to 1.27, which is the average estimate of p in Werner *et al.* (2011). This is because p is difficult to estimate (Schoenberg *et al.*, 2010; Harte, 2015; Seif *et al.*, 2017) and probably is underestimated (Harte, 2015; Seif *et al.*, 2017), and if this is not fixed, it leads to supercritical parameters for both regions. The estimated ETAS parameters for ETAS-base and ETAS-fixed α are shown in Table 3 for all regions. ETAS-var uses the same triggering function as ETAS-base (compare equation 3) but calculates spatially variable parameters following the procedure of Nandan *et al.* (2017), and their estimates are not summarized in Table 3.

Simulation

We simulate 1000 catalogs for each parameter set in Table 3 and for ETAS-var only in southern California. The simulated magnitudes are drawn from an exponential distribution, the Gutenberg–Richter distribution, with a region-specific b-value given in Table 1. The magnitudes are restricted to a lower and upper bound, $M_{cut} = 2.5$ and $M_{max} = 7.5$, $(7.2, 6.5)$, in southern California, northern California, and Italy, respectively. The upper bound is set to ensure similarity between the observed and simulated catalog and should not be interpreted as an insertion about the physics of earthquakes in the region. There seems to be a misconception that $\alpha = \beta$ induces an explosive process, making simulation impossible. Although this is true when considering magnitudes from a Gutenberg–Richter distribution with no upper magnitude limit, it does not hold for the the truncated case. The branching ratio that defines the average number of offspring when applying an upper magnitude limit becomes subcritical, and hence $\alpha = \beta$ poses no problem for realizing simulations.

Null-Hypothesis Testing

In the null-hypothesis test, we compare the distribution of the number of foreshocks per mainshock of observed and ETAS-simulated catalogs (Fig. 3). Therefore, we first analyze the foreshock distributions and find a model to describe them. Then, we perform a null-hypothesis test and report effect sizes, which are the average differences between ETAS and observed foreshock numbers.

Foreshock Distribution

The distribution of the number of foreshocks is the key element of this study and the null-hypothesis test. We therefore describe the distribution of the number of foreshocks by its best-fitting model. For the evaluation, we consider

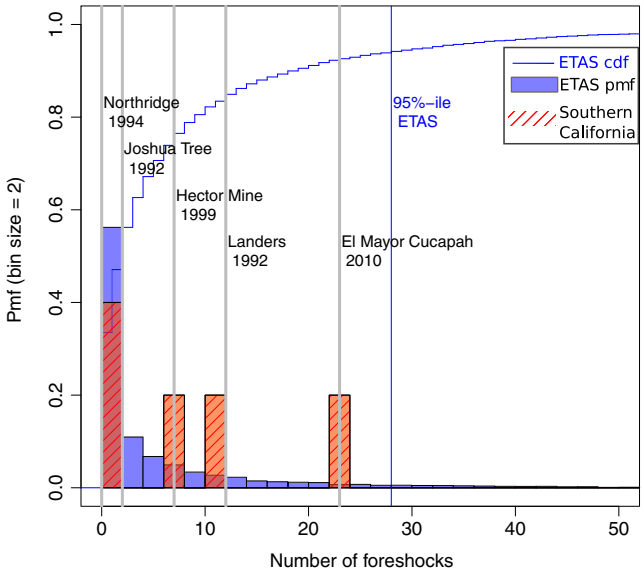

Figure 3. Distribution of observed and ETAS-simulated number of foreshocks ($M_{cut} = 2.5$) of mainshocks with **M** 6+. The foreshock numbers of five Southern Californian mainshocks are plotted (compare Table 2) and compared with a distribution of ETAS foreshocks, obtained from 1000 simulated catalogs. The distribution is normalized and, because foreshock numbers are discrete values, represented in the form of a probability mass function (pmf). Foreshocks are identified using the Zaliapin definition. The color version of this figure is available only in the electronic edition.

different plausible competing models: a lognormal, exponential, power law, and stretched exponential distribution. Because the lognormal, power law, and stretched exponential are either not defined or defined as $y = 0$ at $x = 0$, which both contradict our observations, we fit only foreshock data that are larger than 0 to all models. The best-fitting model is the model with the lowest Bayesian information criterion (BIC). We find that observed foreshocks cannot be fitted well by any model, so we only report the model fits for ETAS foreshocks. Representative for all windows, we show in Figure 4 the observed and ETAS foreshock data of southern California defined in a 3-day and 10-km window and compare different models. We find that the stretched exponential distribution fits the ETAS foreshocks best in 60% of the cases. Considering all window definitions and the Zaliapin definition, the stretched exponential distribution always fits best. The average BIC differences to the second-best model are always higher than 10, which is strong evidence in favor of the stretched exponential. The model parameters seem to be independent of the mainshock magnitude and only dependent on the foreshock magnitude (compare model parameters of the first column in Fig. 4). Observed foreshocks deviate from ETAS foreshocks and their models.

Null-Hypothesis Test

We formulate the null-hypothesis test in a way that it uses the full distribution, rather than just its mean or median value. For each mainshock n, we ask the probability of its number of foreshocks Fs_n or any higher value being observed in ETAS, with $1 - \mathrm{cdf}(X = Fs_n)$. cdf stands for the cumulative density function of the probability mass function's (pmf) of ETAS number of foreshocks. The sum of the logarithms of these probabilities, the joint log-likelihood (JLL) for all mainshocks, is then our test statistic. The JLL is

$$\mathrm{JLL} = \sum_{n=1}^{N} \log(1 - \mathrm{cdf}(X = Fs_n)) \quad \text{with } X \in \mathbb{N} \qquad (4)$$

and N is the number of mainshocks. To evaluate the null hypothesis and obtain a p-value, we repeat 1000 JLL calculations with mainshocks from ETAS catalogs. Each JLL is calculated from N mainshocks that are randomly drawn from the ETAS pmf. Then, we compare the observed JLL with the ETAS distribution of JLL and calculate the p-value with the fraction of JLL(ETAS) $\leq$ JLL(obs).

To detect a possible dependence of foreshock occurrence on mainshock magnitude, we evaluate the null hypothesis at different mainshock magnitudes $M_s = \{4.5 - 5, 5 - 5.5, 5.5 - 6, 6+\}$. To detect the potential importance of small events, we evaluate the null hypothesis at different cutoff magnitudes $M_{cut} = \{2.5+, 3+, 3.5+, 4+\}$. The evaluation of multiple null-hypothesis tests often begs the question of whether p-values should be adjusted (Feise, 2002). With a significance level of 0.05, 1 in 20 independent tests is expected to be (wrongly) rejected if all hypotheses turn out to be true. This is called a type-I error. If avoiding type-I errors is a high priority, one idea would be to adjust p-values. We decided not to adjust the p-values (for our reasons, see the Results and Discussion section). We show an example of the null-hypothesis test in Figure 5 for southern California foreshocks defined within a window of 10 km and 3 days.

We choose to calculate the JLL (equation 4) with $1 - \mathrm{cdf}(X = Fs_n)$ (observing Fs_n or any higher value) and not with $\mathrm{pmf}(X = Fs_n)$ (observing exactly Fs_n), because pmf is not a continuous function and is very sensitive to the number of ETAS simulations from which it is derived. Certain bins of the pmf may contain exceptionally small or large number of events because the distribution is not sufficiently sampled. By taking $1 - \mathrm{cdf}$, the result is less sensitive to sampling and is more continuous. Bins with extremely high or low populations have less influence, because they do not take the full weight. At the same time, evaluating $1 - \mathrm{cdf}$ makes the null-hypothesis test a one-sided test, in which only a surplus of foreshocks is detected.

Effect Size

The drawback of statistical significance tests is that a significant result does not necessarily imply a large effect, and a nonsignificant result does not imply a small effect. Reporting the effect size δ, in our case the difference between ETAS and observed foreshock numbers, helps us to see the whole picture. Although significance tests tell us whether

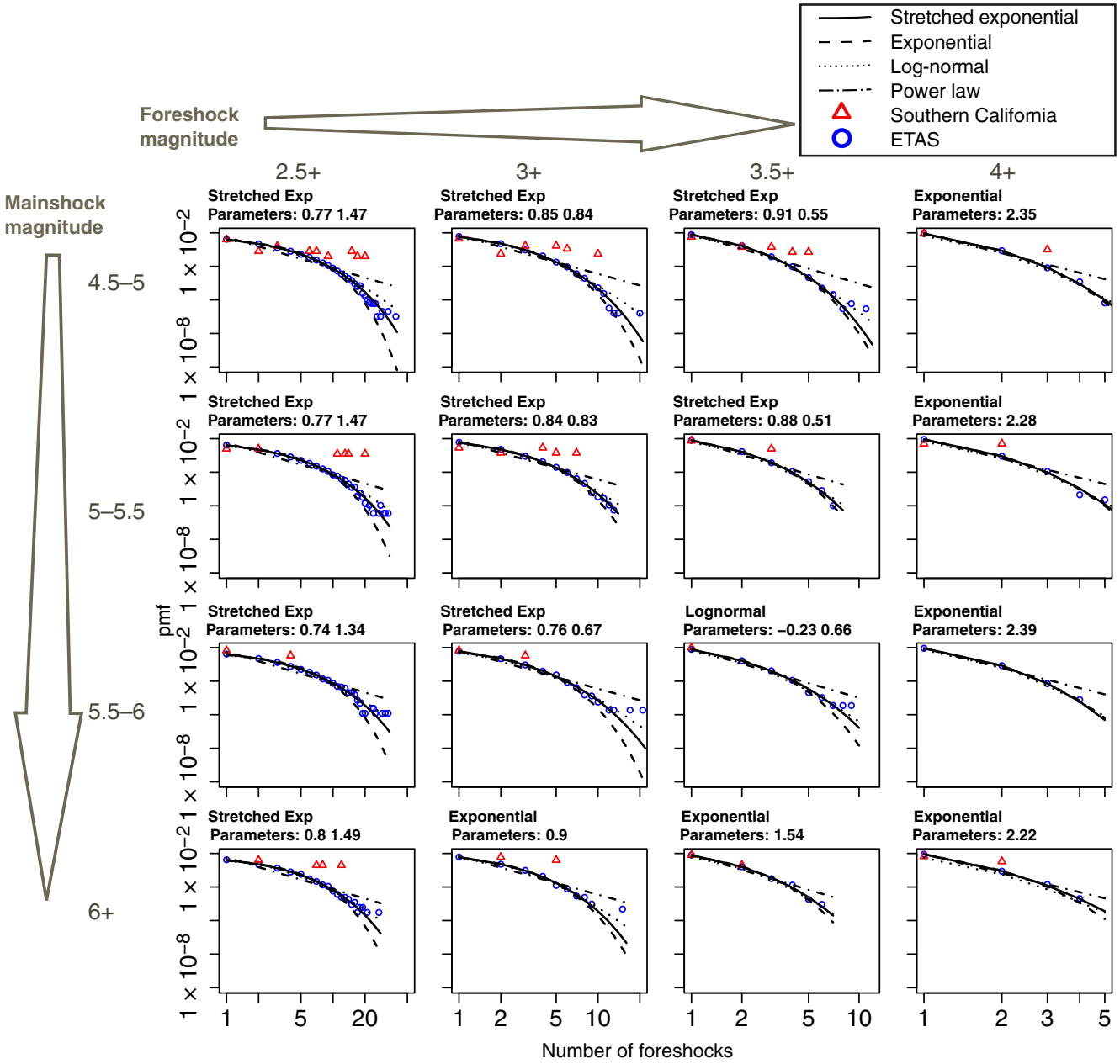

Figure 4. Distribution of observed and ETAS-simulated number of foreshocks and the models fitted to ETAS foreshocks. Rows are different mainshock magnitudes, and columns are different lower thresholds of foreshock magnitudes M_{cut}. Foreshocks are from Southern California, identified within a 3-day and 10-km window. The color version of this figure is available only in the electronic edition.

there is an effect, the effect size tells us how big this effect is. If we know the absolute difference, we can judge whether this effect is of substantive importance, for example, in operational earthquake forecasting.

If data are normally distributed, we measure the effect by comparing the mean values of the null and alternative hypotheses. If data are not normally distributed, as in our case (compare Fig. 3), the Mann–Whitney U test, also known as the Wilcoxon rank sum test, compares tendencies of two independent samples. The requirement for the test is that data must at least be ordinal scaled, which is fulfilled. To estimate the effect size, differences between all possible combinations of the two independent (unpaired) samples, marked with groups 1 and 2, are calculated. Then, the confidence interval is expressed by 95% of the calculated differences.

Results and Discussion

In the present study, we test the null hypothesis that foreshocks behave like general seismicity. We evaluate this hypothesis by comparing the distribution of the number of foreshocks with a uniform triggering model (ETAS) in which all earthquakes are generated by the same process.

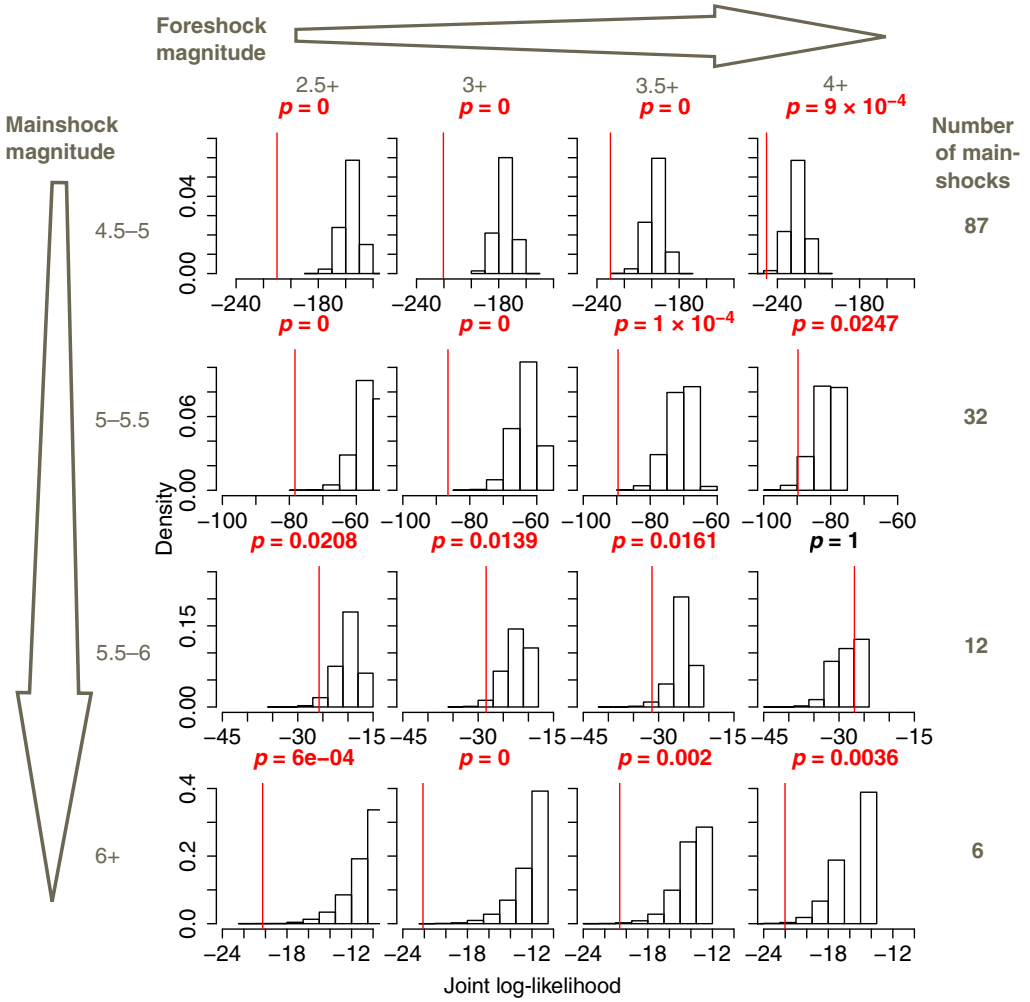

Figure 5. Visualization of the p-value calculation. The JLL distribution obtained from 1000 ETAS-fixed α samples is compared with the observed JLL (vertical line). The p-value is the fraction of JLL(ETAS) $\leq$ JLL(obs). Rows are different mainshock magnitudes, and columns are different lower thresholds of foreshock magnitudes M_{cut}. The null hypothesis is rejected for almost all mainshock magnitudes, and M_{cut}. Foreshocks are defined in a 10-km and 3-day window. The color version of this figure is available only in the electronic edition.

Here, we first report the performance of different ETAS models in describing aftershock numbers. Comparing the performance, we determine the best performing to be the null model. Second, we discuss the interpretation of p-values and provide an estimate of the power of the null-hypothesis test. Finally, we show the resulting p-values of the foreshock null-hypothesis test. Our focus in the discussion is on low p-values, in which the statistical discrepancy between observed foreshock data and ETAS's predictions is substantial. Furthermore, we report effect size, the difference between the observed and ETAS-predicted number of foreshocks. This additional information will be used to judge whether an effect is of considerable size. Here, we discuss our results with regard to the mainshock size, the foreshock definition, the spatiotemporal occurrence of foreshocks, and the importance of small events.

ETAS Null Model

We introduce in "ETAS model" three different models, ETAS-base, ETAS-fixed α, and ETAS-var, as potential null models for the foreshock null-hypothesis test. A null model should describe seismicity under the null hypothesis well, disregarding foreshocks. Here, we test the performance of the three models in describing seismicity by examining their ability to describe aftershock numbers. Deviations of model and data should be most evident in the aftershocks, because the majority of seismicity features that are not accounted for in the models describe aftershock behavior. Unaccounted seismicity features might be (1) temporally varying seismicity, (2) a spatially varying triggering rate, (3) missing aftershocks, (4) anisotropic aftershock distribution, (5) finite temporal and spatial triggering, and (6) magnitude-dependent parameters.

We choose to investigate the aftershock number, because number is also the parameter of interest in the foreshock null-hypothesis test. In doing so, we focus on testing the ability of the productivity parameters to describe the numbers correctly and not whether the spatiotemporal behavior is modeled correctly. In our view, the number

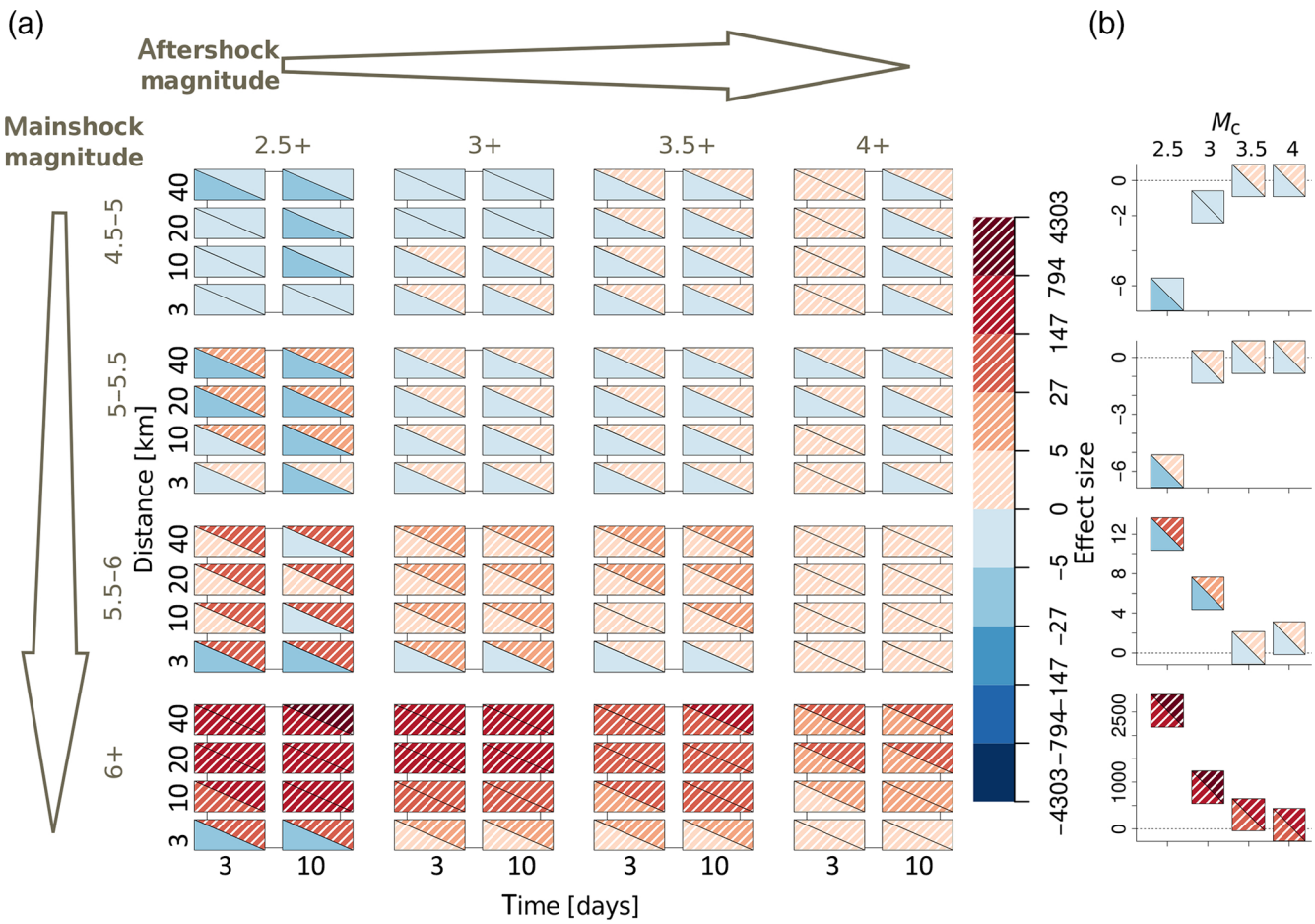

Figure 6. Southern California, ETAS-base: 95% confidence intervals of the effect size between ETAS-base and observed aftershocks. Each rectangle is divided into the lower and upper bound of confidence. Confidence intervals are plotted for different aftershock definitions, (a) Window and (b) Zaliapin, and distinguished for several mainshock (rows) and aftershock (columns) magnitudes. The subplots of the window definitions refer to different space–time windows, in which the *y* axis of each subplot indicates the spatial window (km) and the *x* axis indicates the temporal window (day). The null hypothesis is rejected when the lower- and upper-confidence bounds are either both negative (filled) or both positive (dashed). It is accepted when the confidence bounds include 0. The color version of this figure is available only in the electronic edition.

(of aftershocks or foreshocks) is the parameter of most interest, from which, for example, the probability of a destructive large aftershock can be calculated (independent of its exact location). Regarding foreshocks, the number is of practical importance because potential foreshock patterns that might be derived from our results can be most easily detected by simply counting.

To evaluate the performance, we calculate the difference between observed and ETAS aftershock numbers, the so-called effect size. The number of aftershocks is defined in the same way as are foreshocks, within a space–time window (after the mainshock) and with the Zaliapin definition of aftershocks. Mainshocks that happen in large aftershock sequences are removed from the analysis, according to rules defined in the Foreshock Definition and Aftershock Removal section. We carry out this test, also called the Wilcoxon rank sum test, because it is a two-sided test that detects deviations in both directions. If the confidence interval of the effect size includes 0, we conclude that the ETAS model reproduces the number of aftershocks well. If this is not the case, it likely means, depending on the model, that some or all points (2–6) are not modeled correctly and, depending on the degree of deviation, we should not consider the model as a null model.

Because all of our models are imperfect and only account for a subset of all possible seismicity features, we expect to see deviations between data and model, and we discuss them in the following. We start by evaluating the effect size between ETAS-base and observed aftershock numbers in southern California in Figure 6. Applying the window definition, we see that mainshocks with **M** 5.5–6 and 6+ have more aftershocks than predicted by ETAS-base (indicated by the dashed filling of both the upper- and lower-confidence interval). When evaluating the effect size between observed and ETAS-fixed α aftershocks, we are going to see that this might be the consequence of a too low α. Mainshocks with **M** 4.5–5, depending on the aftershock's lower magnitude level, have either too few (**M** 2.5+) or too many (**M** 4+) aftershocks. This might be related to aftershock

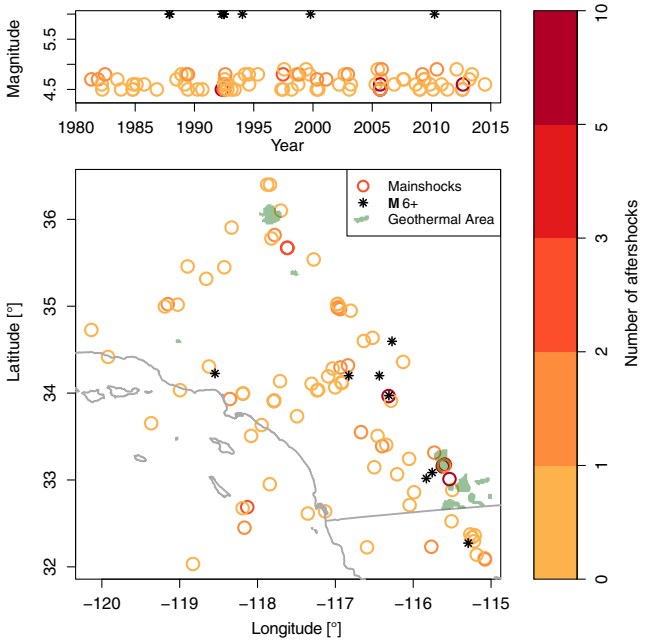

Figure 7. Map of mainshocks (**M** 4.5–5) and their number of aftershocks above $M_c = 4$ indicated in different shades. Aftershocks are defined in a 3-day and 10-km window. Many aftershocks indicate a deviation from, and few aftershocks an agreement with, ETAS-fixed α. The color version of this figure is available only in the electronic edition.

incompleteness that is present at low magnitudes for **M** 2.5+. Or it might be related to swarms, for **M** 4+ that are known to have particular seismicity, modeled with low α or temporally variable background rate (Hainzl, 2013), that we do not account for. Or in both cases, spatially variable ETAS parameters might explain the deviation. Specifically, the variability should be present in the productivity parameters and with them also the branching ratio, because these could model the observed surplus or lack of aftershocks. We will see later that spatial variability only explains the deviations for **M** 2.5+ aftershocks. The deviation of **M** 4+ aftershocks can be explained by swarms and insufficiently removed aftershock sequences. This becomes evident after plotting the location of $4.5 < M < 5$ mainshocks with their corresponding number of **M** 4+ aftershocks. We see in Figure 7 that mainshocks with many aftershocks happen either close to geothermal regions that are known to host swarms or in the vicinity of an **M** 6+ mainshock characterized by numerous aftershocks. Overall, ETAS-base as a null model describes aftershocks in only 44% of the cases, considering all windows and lower magnitude thresholds. Applying the Zaliapin definition leads, in accordance with the window definition, to rejections for **M** 6+ and 4.5–5 mainshocks but no rejections for **M** 5.5–6, that sums up to 63% acceptance with Zaliapin.

In Figure 8, we compare ETAS-fixed α with observed aftershock numbers. Applying the window definition, we see that the null hypothesis is generally accepted (0 is included in the confidence bounds). It is only rejected for small mainshocks (**M** 4.5–5) and **M** 2.5+ and **M** 4+ aftershock magnitudes, identical to the findings for ETAS-base. We suspect the same reasons as mentioned in that context. Overall, the null hypothesis is accepted in 77% of the cases for the window definition. Applying the Zaliapin definition, aftershocks of **M** 4.5–5 mainshocks are similarly rejected. However, in disagreement with the window definition, **M** 5.5–6 mainshocks have too few **M** 3+ and 4+ aftershocks. This sums up to 69% acceptance with Zaliapin.

In Figure 9, we compare ETAS-var with observed aftershock numbers. Generally, we observe too many aftershocks throughout all mainshock magnitudes, both for the window and Zaliapin definition. This tallies with observations for ETAS-base and might hence be explained by an $\alpha < \beta$ that is generally observed in ETAS-var (compare Nandan *et al.*, 2017, Fig. 2d). Aftershocks with **M** 2.5+ of $4.5 < M < 5$ are explained well by ETAS-var. Spatial variable seismicity might hence be the reason for the observed deviation of ETAS-base and ETAS-fixed α in this case. In total, only 43% (38%) of cases accept the null model for the window (Zaliapin) definition.

In summary, for southern California, we find that accounting for aftershock incompleteness and spatial anisotropy of aftershocks, by fixing $\alpha = \beta$, leads to the best performance of the null model in describing aftershock data. This makes the ETAS-fixed α an appropriate null model for seismicity, excluding swarm seismicity at mainshocks with $4.5 < M < 5$ and **M** 4+ aftershocks and spatial variable seismicity for **M** 2.5+ aftershocks.

Regarding northern California and Italy, we find that none of our null models describes aftershock seismicity well. In the Ⓔ electronic supplement, we show that in northern California, when applying the window definition, only 55% (16%) of data subsamples are described by ETAS-base (ETAS-fixed α). In Italy, only 39% (42%) of data subsamples are described by ETAS-base (ETAS-fixed α). The values are similar for the Zaliapin aftershock definition. We will hence not perform a foreshock null-hypothesis test in northern California or Italy.

Interpretation of p-Values, Multiple p-Values, and Power of the Test

We present part of our results in the form of p-values. p-values are the probability of observing data or obtaining a test statistic under the null hypothesis. Our test statistic is the JLL of finding at least the number of observed foreshocks under the ETAS model. Accordingly, a p-value below the significance level ($\alpha = 0.05$), as calculated in a one-tailed test by the fraction of JLL(ETAS) ≤ JLL(obs), provides evidence that more foreshocks are observed in reality than ETAS predicts. When we compare p-values of the same mainshock magnitude for different foreshock definitions (i.e., space–time windows, Zaliapin, foreshock magnitudes), we can conclude the relative effect size between the two. This is possible because the p-value is a confound index

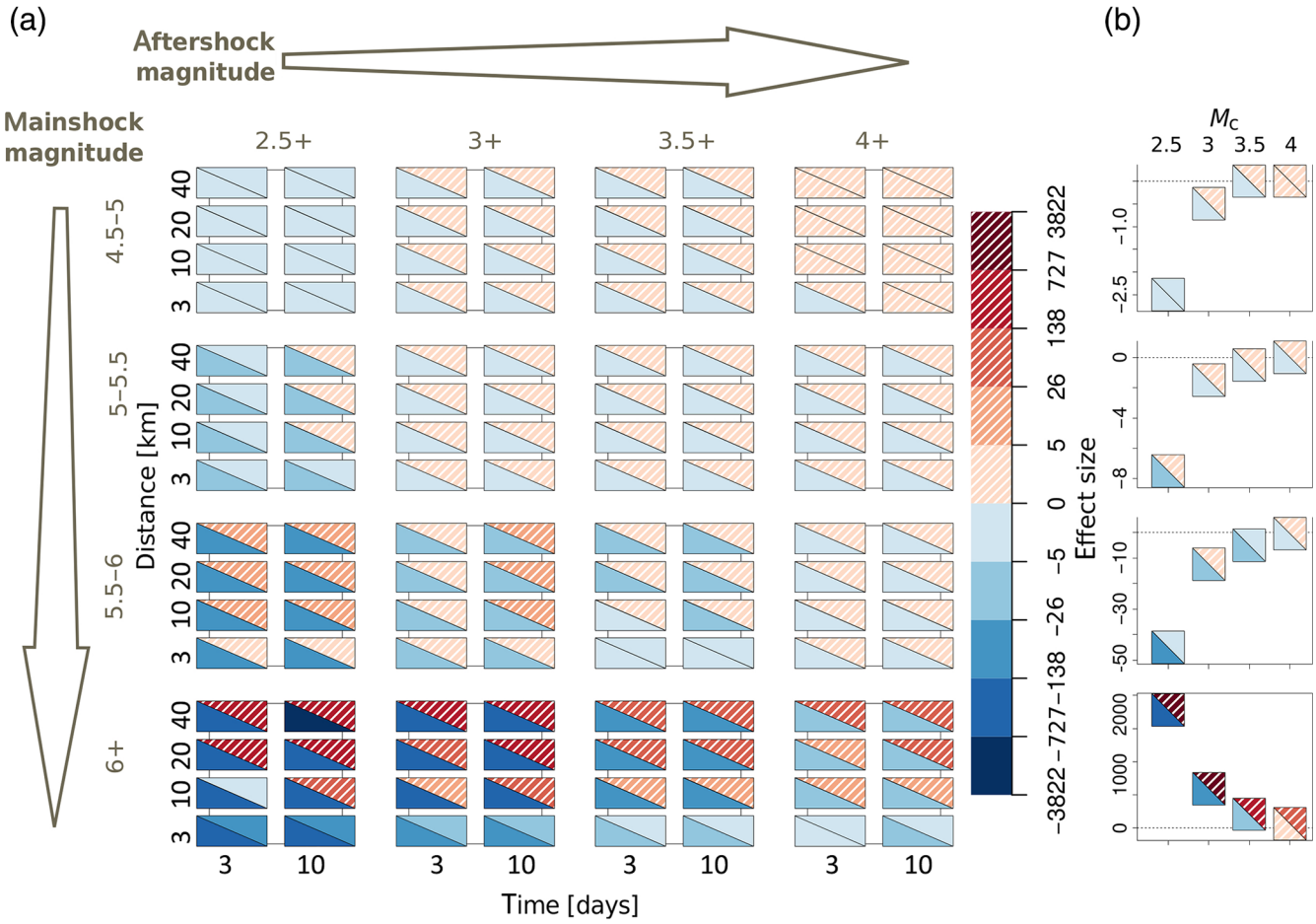

Figure 8. Southern California, ETAS-fixed α: 95% confidence intervals of the effect size between ETAS-fixed α and observed aftershocks. Same as caption of Figure 6. The color version of this figure is available only in the electronic edition.

reflecting both effect size and sample size (Rosenthal and Rosnow, 1991). Because the sample size remains constant for the same mainshock magnitude, changes in the p-value can be attributed to the effect size.

To enable the study of the previously mentioned attributes, we divide our dataset into 16 subdatasets containing different mainshock and foreshock magnitudes. Null-hypothesis tests are performed on each subdataset. Testing multiple null hypotheses in a study often results in the adjustment of the p-value. However, there are several practical objections to adjusting p-values (Feise, 2002). First, the adjustment factor depends on the number of individual tests, which is arbitrary: there is no right number. Second, reducing type-I errors (the chance of finding nonexistent foreshock signals) increases type-II errors (the chance of not detecting existing foreshock signals). Mindful of this, we do not adjust the p-values but discuss our findings aware of the increased chance of type-I errors. We also consider the dependence of subsamples, as defined by different, lower foreshock magnitude thresholds in our discussion.

We calculate the power of the null-hypothesis test (POT) to measure the probability that the test will detect a certain effect δ. δ is a postulated difference between observed and the ETAS-predicted number of foreshocks. In Figure 10, we see that the power is proportional to δ, the sample size, and α (α is not shown, and $\alpha = 0.05$). We calculate the POT, as described in Zechar *et al.* (2010), by comparing two synthetic foreshock samples, one of which is imposed with δ, so we expect the statistical test to reject the null hypothesis. Repeating this comparison 1000 times, the POT is expressed by

$$\text{Power} = \frac{\text{Number of correct rejections}}{\text{Number of tests}}. \qquad (5)$$

The POT is directly related to the type-II error with Power $= 1 - \beta$, in which β is the type-II error. In cases in which the null hypothesis is accepted, the POT tells us how much we can trust this result. The POT tells us how confidently the test would have detected a certain genuinely occurring effect.

Foreshock Null-Hypothesis Test

We perform the foreshock null-hypothesis test (described in the Null-Hypothesis Testing section) for southern Californian foreshocks, using ETAS-fixed α as the null model that we found to describe aftershock seismicity appropriately

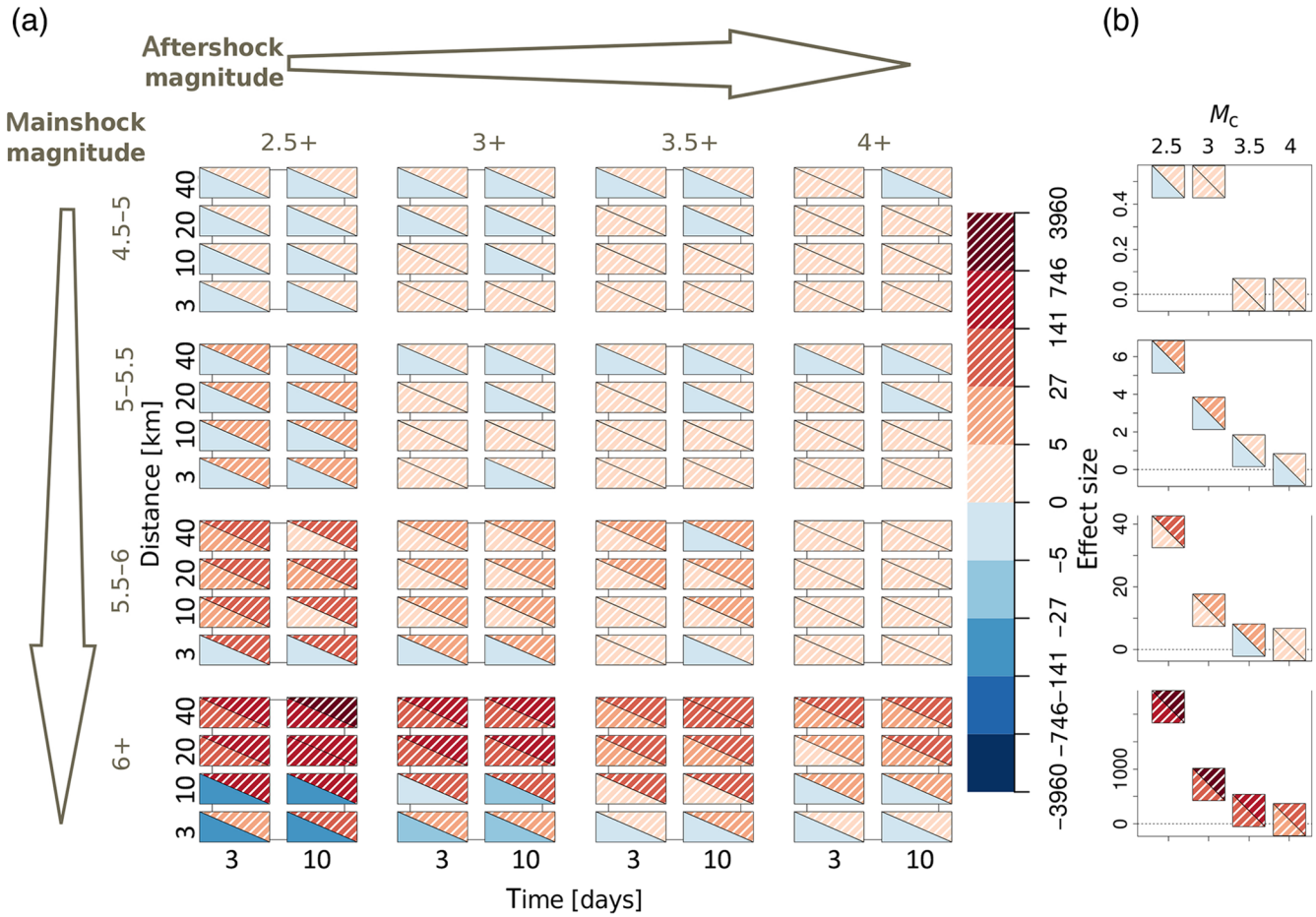

Figure 9. Southern California, ETAS-var: 95% confidence intervals of the effect size between ETAS-var and observed aftershocks. Same as caption of Figure 6. The color version of this figure is available only in the electronic edition.

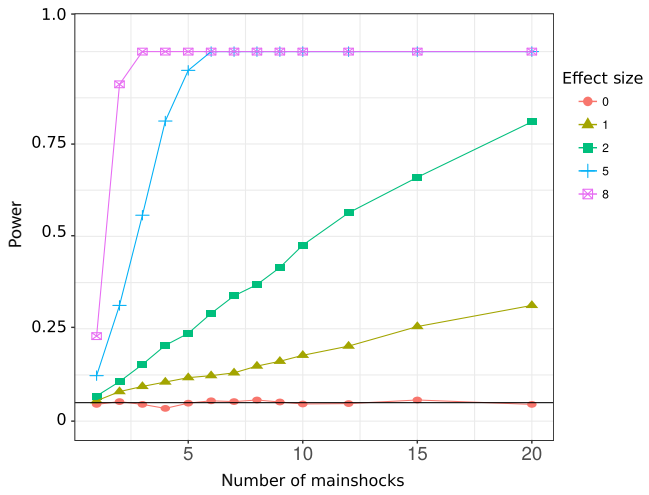

Figure 10. The power of the test is plotted against the number of mainshocks for different effect sizes δ, that is, the differences in the number of foreshocks per mainshock between the null and the alternative hypothesis. The black horizontal line indicates the significance level $\alpha = 0.05$. The color version of this figure is available only in the electronic edition.

in the ETAS Null Model section. The result, that is, the p-values are shown in Figure 11 for different mainshock and foreshock magnitudes and foreshock definitions. The figure is dominated by low $p \leq 0.05$ (i.e., dark shades) for almost all mainshock and foreshock magnitudes and foreshock definitions. For $p \leq 0.05$, the null-hypothesis is rejected, and foreshocks are not explained by ETAS-fixed α. Here, we discuss the results in detail and show how they promote our understanding of the nature of foreshocks.

First, we examine whether peculiar foreshocks (that lead to a rejection of the null hypothesis) depend on the mainshock magnitude. We find $p \leq 0.05$ for all mainshock magnitudes and hence conclude that the foreshock behavior does not depend on the mainshock magnitude. This is in agreement with Brodsky (2011) and Shearer (2012), who also detect no dependence of the foreshock behavior on the mainshock magnitude. Considering the results of window foreshocks, we expect to observe $128 \times 0.05 = 6.4$ random rejections of the null hypothesis (type-I error). We observe more than six rejections: a total of 110 rejections that are hence unlikely to be random. Exceptions, with $p > 0.05$, are only present when analyzing **M** 4+ foreshocks of mainshocks with $5 \leq M < 6$. A reason could be that a potential deviation is not detected;

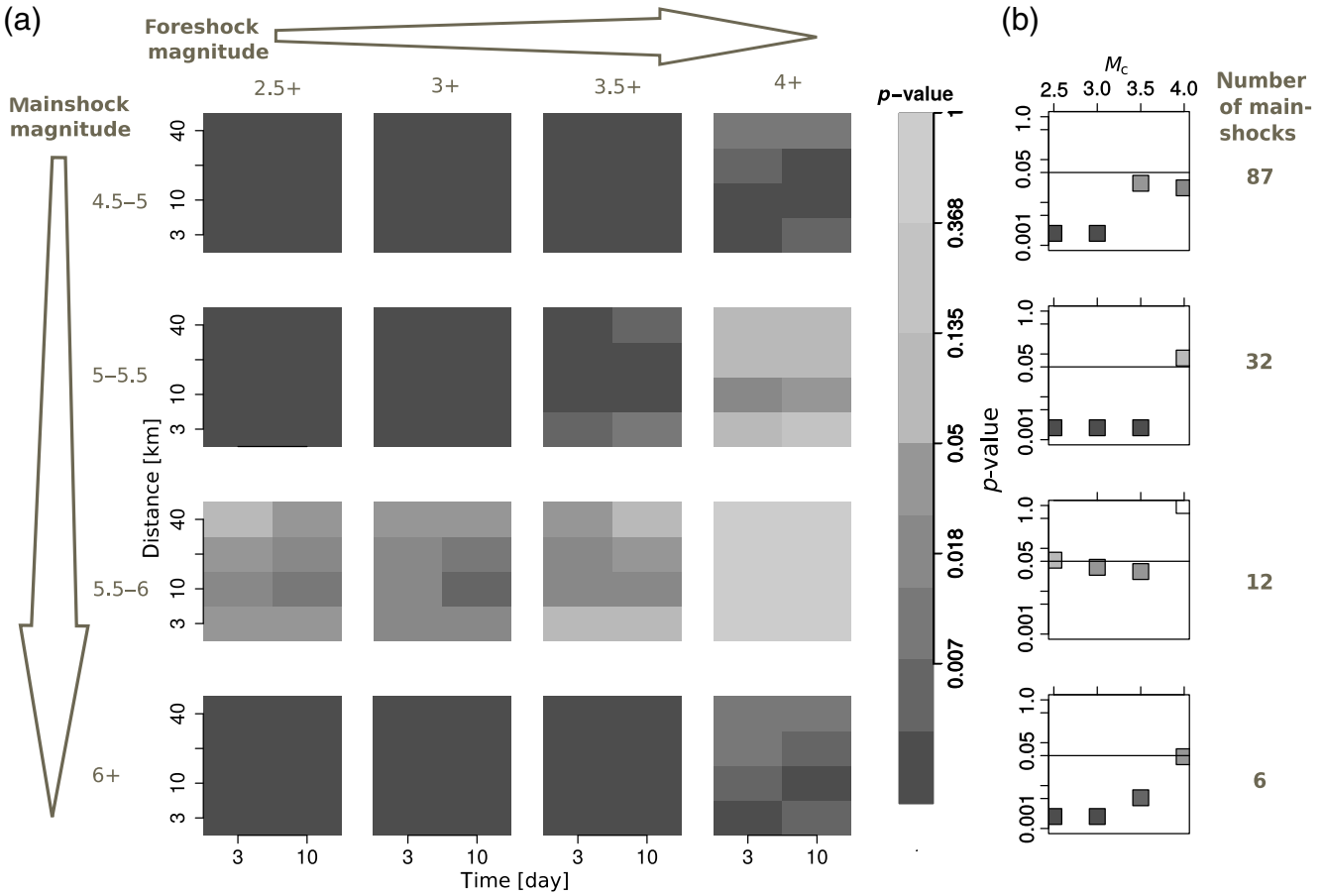

Figure 11. Southern California, ETAS-fixed α: p-values of the null-hypothesis test are plotted for different foreshock definitions and distinguished for several mainshock (rows) and foreshock (columns) magnitudes. Foreshocks are defined (a) within a space–time window and (b) applying the Zaliapin declustering method. The subplots of the window definitions refer to different space–time windows, in which the y axis of each subplot indicates the spatial window (km) and the x axis indicates the temporal window (day). The null hypothesis is rejected for $p \leq 0.05$ (dark shades).

hence, a type-II error occurred. Evaluating the POT at $N = 12$, the number of mainshocks with $5.5 \leq M < 6$, we see that a difference of $\delta = 5$ ($\delta = 2$) would be detected in 100% (58%) of cases. This indicates an increased chance of small effects being present but going undetected. When consulting the effect size, it is revealed that too few **M** 4+ foreshocks are observed. This explains the acceptance of the one-sided null-hypothesis test. The reason why too many **M** 4+ foreshocks for mainshocks with $5.5 \leq M < 6$ are observed remains unclear.

Second, we consider the two different foreshock definitions. The declustering method after Zaliapin *et al.* (2008) distinguishes into spontaneous and clustered earthquakes. By definition, Zaliapin foreshocks are clustered earthquakes that precede the cluster's mainshock. They are part of a chain with small nearest-neighbor distances, and any occurrence of spontaneous earthquakes would interrupt the cluster's chain. In contrast, the window method cannot distinguish between clustered and spontaneous background events. We find that, for most mainshock and foreshock magnitudes, both Zaliapin (81%) and window foreshocks (86%) lead to a rejection of the null hypothesis. This means that Zaliapin and window foreshocks are both not well constrained by the ETAS model. The rejection for foreshocks defined through the window method can be traced back to a surplus of both spontaneous or clustered earthquakes. The rejection for Zaliapin foreshocks can only be traced back to a surplus of clustered earthquakes. Finally, we can be sure of a surplus of clustered earthquakes in the observed catalog, but at the same time cannot exclude a surplus of spontaneous earthquakes. These results are opposed to the ones by Dodge *et al.* (1995, 1996), who found evidence with Coulomb stress modeling that foreshocks do not trigger each other. It might be that the Zaliapin definition of clustered does not coincide with the Coulomb stress definition of triggered, which would explain the result. A clear answer could be obtained by applying the foreshock null-hypothesis test to triggered foreshocks under the Coulomb stress model.

Third, we investigate the spatiotemporal characteristics of foreshocks by examining foreshocks defined in differently sized windows. We find similar p-values, mostly $p \leq 0.05$, for all windows, considering all mainshock magnitudes. That

means evidence against the null hypothesis is present at all spatiotemporal foreshock definitions. This result is opposed to the one obtained by Felzer (2004), who found a break in the distribution of foreshock-to-mainshock distances at about 10 km.

Indifference of the rejections toward the window size for all mainshock sizes also implies that there is no correlation between the spatiotemporal size of the preparatory zone and the mainshock size. This result is also obtained by Felzer (2004) using a correlation test, in contradiction to the findings of Lippiello *et al.* (2012, 2017), who found that the spatial distribution of foreshocks depends on mainshock magnitude. We think that their observation can probably be explained by a sample-size effect at different mainshock magnitudes. Undersampling a power law distribution with $(x + D)^{-q}$ increases D and leaves q constant (Seif *et al.*, 2017), which is precisely what Lippiello *et al.* (2012, 2017) observed.

Fourth, we examine whether peculiar foreshock behavior depends on the lower-magnitude threshold of the foreshocks. Our suspicion is that an increased sample size at a lower-magnitude threshold could make potential foreshock signals more conspicuous. We find however, that the p-value does not depend on the lower-magnitude threshold of the foreshocks; hence, peculiar foreshock behavior is not magnitude dependent. This result does not support the findings of Mignan (2014) that suggest that peculiar foreshock behavior emerges when analyzing foreshocks that are at least three orders of magnitude smaller than the mainshock. Mignan (2014) points out that most studies finding no deviation from general seismicity conclude this after stacking the data. Stacking in this case means averaging both certain observations and time series over multiple mainshocks. We also believe that the methodology is critical for detecting foreshock patterns. In the first case, mean values are not necessarily always representative of a distribution, especially an abnormal one. In the second case, it must be assumed that all time series, that are themselves distributions, follow the same process, so they can be stacked to obtain a single representative distribution. This implies that all mainshocks have to be assumed to experience the same initial processes, in which less-frequent anomalous features may be lost. We deal with the first case, because we have a single observation for each mainshock, the number of foreshocks, and choose a method that takes into account the distribution of the observed data. This could explain why we detect peculiar foreshock behavior at all foreshock magnitudes.

Even though peculiar foreshock behavior is observed for all foreshock magnitude thresholds, small earthquakes, due to their frequent occurrence, might highlight potential differences most. This is confirmed by the effect size in Figure 12 that displays the most marked deviations between the observed and ETAS-predicted number of foreshocks for **M** 6+ mainshocks at the lowest magnitude threshold, **M** 2.5+. Here, the largest effect size of $\delta = 12$ is observed. At the same time, the effect size varies considerably, with confidence intervals between $1 \leq \delta \leq 12$. Given that the largest effect size is observed for the lowest foreshock magnitudes, foreshock signals might be easiest to detect from small foreshocks. The effect size generally supports the findings of the null-hypothesis test, with mainly positive confidence intervals, meaning more observed than predicted foreshocks.

The reader might have noticed, that in comparison to other mainshock-magnitude bins with a size of 0.5, mainshocks summarized by **M** 6+ contain magnitudes as high as $M = 7.3$. To visualize potential differences in the foreshock behavior of the individual mainshocks, we plot their survival function $1 - \mathrm{cdf}(X = Fs_n)$ on a map (Fig. 13). We see that the largest mainshocks with **M** 7+ have the largest number of foreshocks.

Conclusion

The present study provides an answer to the fundamental question of whether foreshocks are caused by a different process than general seismicity. We find that foreshocks deviate significantly from a uniform triggering model in Southern California. We reach this conclusion after performing a null-hypothesis test, in which we calculate the joint probability of drawing observed numbers of foreshocks or higher numbers under ETAS.

The null-hypothesis test requires a null model that describes observed seismicity sufficiently well. After testing three different models in Southern California for their performance in describing aftershocks, we found that ETAS with $\alpha = \beta$ describes 77% of all spatiotemporal windows (69% of Zaliapin). Deviations of the data from the model could be traced back to spatial-variable seismicity and swarm seismicity. In northern California and Italy, we could not determine an adequate null model: not adequate means that the potential null models could only explain aftershocks in about 40% of the cases and that the origin of the deviations could not be traced back. Therefore, northern Californian and Italian foreshocks are not investigated in this study, and all presented results refer to southern Californian foreshocks.

We find that the ETAS foreshock distributions are best represented by a stretched exponential, and its parameters depend on the foreshock definition and the lower-magnitude threshold of the foreshocks but seem to be independent of the mainshock magnitude. Observed foreshocks deviate from ETAS foreshocks and their models, and we could not determine a best-fitting model. This could be due to either too few data or that the data follows none of the considered distributions.

With the use of two different foreshock definitions, we are able to investigate possible mechanisms for foreshock occurrence. We defined foreshocks within a spatiotemporal window and with the Zaliapin *et al.* (2008) declustering method. Although the window foreshocks are defined completely independent of each other, by simply counting earthquakes within a window, the Zaliapin foreshocks reflect the interactions between the foreshocks. Only clustered earthquakes, defined by the spatiotemporal distances between foreshocks, as well as their magnitudes, comprise a chain of

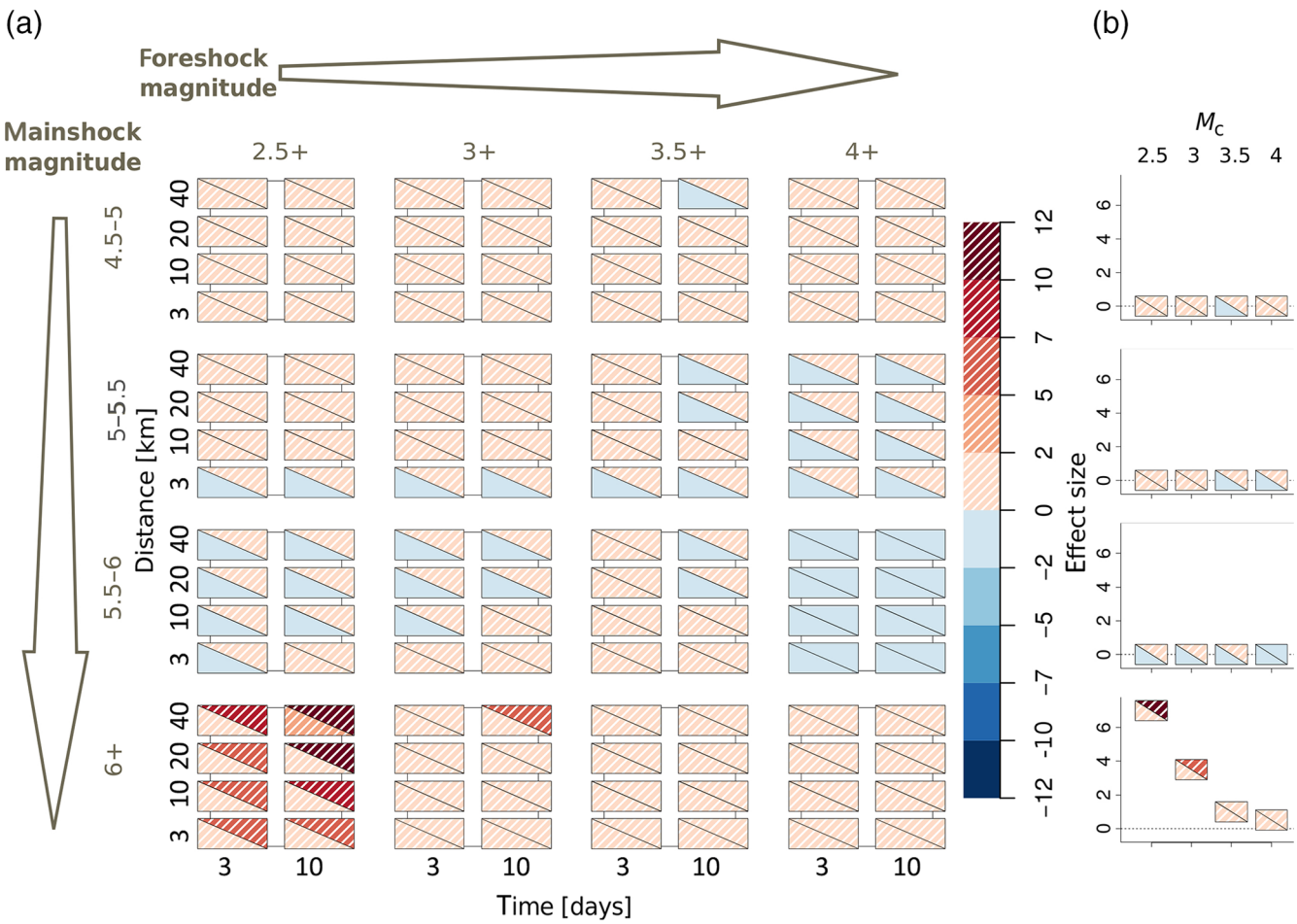

Figure 12. Southern California, ETAS-fixed α: 95% confidence intervals of the effect size between ETAS-fixed α and observed foreshocks. Same as caption of Figure 6. The color version of this figure is available only in the electronic edition.

foreshocks. The observed rejection of the null hypothesis for both window and Zaliapin foreshocks indicates that a surplus of clustered earthquakes is likely responsible for the rejection. However, an additional surplus of spontaneous earthquakes can also not be excluded.

The rejection of the null hypothesis for window and Zaliapin foreshocks does not depend on the foreshock magnitude, so we conclude that small earthquakes do not convey additional information. However, the effect size depends on the lower-magnitude threshold for foreshocks. Smaller thresholds lead to a larger sample size, so the effect size is larger. At the same time, the 95% confidence intervals for small thresholds are large. For **M** 6+ mainshocks, the effect size varies between $1 \leq \delta \leq 12$ at $M_{\text{cut}} = 2.5$ and $0 \leq \delta \leq 1$ at $M_{\text{cut}} = 4$. From an operational perspective, larger effects at smaller M_{cut} are more easily detectable and might be used to identify foreshock patterns in the future.

Regarding the spatiotemporal occurrence of foreshocks, we do not find that the foreshock area scales with the mainshock magnitude. Thus, the earthquake preparation zone provides no indication of how large a future mainshock will be.

The solid performance of our null-hypothesis test in detecting potential differences between the null and alternative hypotheses is confirmed by the power of the test. Differences of $\delta = 5$ ($\delta = 2$) can be detected with a probability of 100% (> 50%) in southern California.

Summing up all the findings, we conclude that foreshocks are generated by a different process than general seismicity and that they are a feature of all mainshock sizes. It still remains unclear, though, whether the foreshock-generating process can be explained by top-down loading or bottom-up triggering. The fact that the general seismicity-triggering model is unable to explain foreshocks does not mean that a model with different parameters would likewise be unable to explain foreshocks. We see this in the example of swarm seismicity: a model with low α or a temporally variable background rate (Hainzl *et al.*, 2008) describes their occurrence well. Similarly, foreshocks might be modeled by temporally variable ETAS parameters. In that case, we see a potential problem of overfitting individual foreshock sequences with a temporary variable ETAS model. ETAS consist of a high number of parameters, in our case seven, and there is little foreshock data.

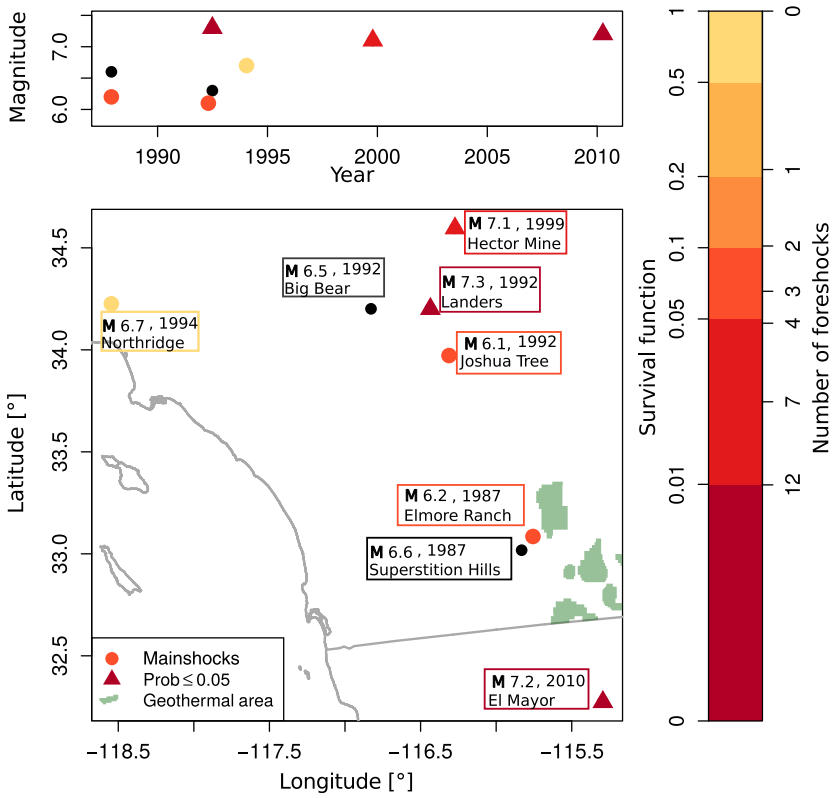

Figure 13. Southern California, ETAS-fixed α: Map of mainshocks (**M** 6–7.3) and their number of foreshocks above $M_c = 2.5$ plotted in different shades. The corresponding value of the survival function $1 - \mathrm{cdf}(X = FS)$ is indicated at the left axis of the color bar (gray scale). Foreshocks are defined in a 3-day and 10-km window. Many foreshocks indicate a deviation from, and few foreshocks an agreement with, ETAS. The color version of this figure is available only in the electronic edition.

The knowledge gained in this study will hopefully provide a basis on which future research on earthquake forecasting can be built.

Data and Resources

Earthquake data for southern California were obtained from the Southern California Data Center (doi: 10.7909/C3WD3xH1) at http://service.scedc.caltech.edu/ftp/catalogs/hauksson/Socal_DD/hs_1981_2014_comb_K4_A.cat_so_SCSN_v01 (last accessed April 2016). Earthquake data for Northern California were obtained from the Northern California Earthquake Data Center (doi: 10.1029/2007JB005479, 2008; doi: 10.1785/0120080294, 2009; last accessed April 2016). The Advanced National Seismic System (ANSS) authoritative region data were downloaded from http://www.quake.geo.berkeley.edu/anss/anss-detail.html (last accessed April 2016).

Acknowledgments

S. Seif would like to thank W. Marzocchi for valuable discussions in the early stage of the work. The authors thank Paolo Gasperini, Barbara Lolli, and Gianfranco Vannucci for providing the Italian catalog (provided April 2016). The authors would like to thank three anonymous reviewers, Associate Editor Matthew C. Gerstenberger, and Editor-in-Chief Thomas Pratt for their helpful reviews that significantly improved the article.

References

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

Agnew, D., and L. Jones (1991). Prediction probabilities from foreshocks, *J. Geophys. Res.* **96,** 11,959–11,971.

Bak, P., C. Tang, and K. Wiesenfeld (1988). Self-organized criticality, *Phys. Rev. A* **38,** no. 1, 364–374.

Bouchon, M., V. Durand, D. Marsan, H. Karabulut, and J. Schmittbuhl (2013). The long precursory phase of most large interplate earthquakes, *Nature Geosci.* **6,** no. 4, 299–302.

Brodsky, E. E. (2011). The spatial density of foreshocks, *Geophys. Res. Lett.* **38,** no. 10, L10305, doi: 10.1029/2011GL047253.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law distributions in empirical data, *SIAM Rev.* **51,** no. 4, 661–703.

Console, R. (2003). Refining earthquake clustering models, *J. Geophys. Res.* **108,** no. B10, 2468.

Das, S., and C. H. Scholz (1981). Theory of time-dependent rupture in the Earth, *J. Geophys. Res.* **86,** no. B7, 6039.

Dodge, D. A., G. C. Beroza, and W. L. Ellsworth (1995). Foreshock sequence of the 1992 Landers, California, earthquake and its implications for earthquake nucleation, *J. Geophys. Res.* **100,** no. B6, 9865.

Dodge, D. A., G. C. Beroza, and W. L. Ellsworth (1996). Detailed observations of California foreshock sequences: Implications for the earthquake initiation process, *J. Geophys. Res.* **101,** no. B10, 22,371.

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Med. Res. Methodol.* **2,** no. 8, doi: 10.1186/1471-2288-2-8.

Felzer, K. R. (2002). Triggering of the 1999 M_w 7.1 Hector Mine earthquake by aftershocks of the 1992 M_w 7.3 Landers earthquake, *J. Geophys. Res.* **107,** no. B9, 2190.

Felzer, K. R. (2004). A common origin for aftershocks, foreshocks, and multiplets, *Bull. Seismol. Soc. Am.* **94,** no. 1, 88–98.

Felzer, K. R., M. T. Page, and A. J. Michael (2015). Artificial seismic acceleration, *Nature Geosci.* **8,** no. 2, 82–83.

Gasperini, P., B. Lolli, and G. Vannucci (2013). Empirical calibration of local magnitude data sets versus moment magnitude in Italy, *Bull. Seismol. Soc. Am.* **103,** no. 4, 2227–2246.

Hainzl, S. (2013). Comment on "Self-similar earthquake triggering, Båth's law, and foreshock/aftershock magnitudes: Simulations, theory, and results for southern California" by P. M. Shearer, *J. Geophys. Res.* **118,** no. 3, 1188–1191.

Hainzl, S., G. B. Brietzke, and G. Zöller (2010). Quantitative earthquake forecasts resulting from static stress triggering, *J. Geophys. Res.* **115,** no. 11, 1–9.

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,** no. 4, 2066–2072.

Hardebeck, J. L., K. R. Felzer, and A. J. Michael (2008). Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant, *J. Geophys. Res.* **113,** no. B08310, doi: 10.1029/2007JB005410.

Harte, D. (2015). Model parameter estimation bias induced by earthquake magnitude cut-off, *Geophys. J. Int.* **204,** no. 2, 1266–1287.

Hauksson, E., W. Yang, and P. M. Shearer (2012). Waveform relocated earthquake catalog for Southern California (1981 to June 2011), *Bull. Seismol. Soc. Am.* **102,** no. 5, 2239–2244.

Helmstetter, A., and D. Sornette (2003). Foreshocks explained by cascades of triggered seismicity, *J. Geophys. Res.* **108,** no. B10, 2457.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2005). Importance of small earthquakes for stress transfers and earthquake triggering, *J. Geophys. Res.* **110,** 1–15.

Jones, L. M. (1984). Foreshocks (1966–1980) in the San Andreas system, California, *Bull. Seismol. Soc. Am.* **74,** no. 4, 1361–1380.

Kanamori, H. (1981). *Earthquake Prediction*, Vol. 4, Maurice Ewing Series, American Geophysical Union, Washington, D. C.

Kato, A., K. Obara, T. Igarashi, H. Tsuruoka, S. Nakagawa, and N. Hirata (2012). Propagation of slow slip leading up to the 2011 M_w 9.0 Tohoku-Oki earthquake, *Science* **335,** no. 6069, 705–708.

Lippiello, E., F. Giacco, W. Marzocchi, G. Godano, and L. de Arcangelis (2017). Statistical features of foreshocks in instrumental and ETAS catalogs, *Pure Appl. Geophys.* **174,** no. 4, 1679–1697.

Lippiello, E., W. Marzocchi, L. de Arcangelis, and C. Godano (2012). Spatial organization of foreshocks as a tool to forecast large earthquakes, *Sci. Rep.* **2,** 846

Llenos, A. L., and A. J. Michael (2013). Modeling earthquake rate changes in Oklahoma and Arkansas: Possible signatures of induced seismicity, *Bull. Seismol. Soc. Am.* **103,** no. 5, 2850–2861.

Marzocchi, W., and J. Zhuang (2011). Statistics between mainshocks and foreshocks in Italy and Southern California, *Geophys. Res. Lett.* **38,** L09310, doi: 10.1029/2011GL047165.

Mignan, A. (2012). Seismicity precursors to large earthquakes unified in a stress accumulation framework, *Geophys. Res. Lett.* **39,** no. 21, L21308.

Mignan, A. (2014). The debate on the prognostic value of earthquake foreshocks: A meta-analysis, *Sci. Rep.* **4,** 4099.

Nandan, S., G. Ouillon, S. Wiemer, and D. Sornette (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to California, *J. Geophys. Res.* **122,** no. 7, 5118–5143.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.* **83,** no. 401, 9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences, *Ann. Inst. Stat. Math.* **50,** no. 2, 379–402.

Ohnaka, M. (1992). Earthquake source nucleation: A physical model for short-term precursors, *Tectonophysics* **211,** nos. 1/4, 149–178.

Reasenberg, P. A. (1999). Foreshock occurrence before large earthquakes, *J. Geophys. Res.* **104,** no. B3, 4755.

Rosenthal, R., and R. L. Rosnow (1991). *Essentials of Behavioral Research: Methods and Data Analysis*, McGraw-Hill, Boston, Massachusetts.

Schoenberg, F. P., A. Chu, and A. Veen (2010). On the relationship between lower magnitude thresholds and bias in epidemic-type aftershock sequence parameter estimates, *J. Geophys. Res.* **115,** no. B04309, doi: 10.1029/2009JB006387.

Seif, S., A. Mignan, J. D. Zechar, M. J. Werner, and S. Wiemer (2017). Estimating ETAS: The effects of truncation, missing data, and model assumptions, *J. Geophys. Res.* **122,** 449–469.

Shearer, P. M. (2012). Space-time clustering of seismicity in California and the distance dependence of earthquake triggering, *J. Geophys. Res.* **117,** no. B10306, doi: 10.1029/2012JB009471.

Tinti, S., and F. Mulargia (1987). Confidence intervals of b values for grouped magnitudes, *Bull. Seismol. Soc. Am.* **77,** no. 6, 2125–2134.

Waldhauser, F. (2009). Near-real-time double-difference event location using long-term seismic archives, with application to northern California, *Bull. Seismol. Soc. Am.* **99,** no. 5, 2736–2748.

Wang, K. (2006). Predicting the 1975 Haicheng earthquake, *Bull. Seismol. Soc. Am.* **96,** no. 3, 757–795.

Wang, Q., D. D. Jackson, and J. Zhuang (2010). Missing links in earthquake clustering models, *Geophys. Res. Lett.* **37,** no. 21, 1–5.

Werner, M. J., A. Helmstetter, D. D. Jackson, and Y. Y. Kagan (2011). High-resolution long-term and short-term earthquake forecasts for California, *Bull. Seismol. Soc. Am.* **101,** no. 4, 1630–1648.

Yagi, Y., R. Okuwaki, B. Enescu, S. Hirano, Y. Yamagami, S. Endo, and T. Komoro (2014). Rupture process of the 2014 Iquique Chile earthquake in relation with the foreshock activity, *Geophys. Res. Lett.* **41,** no. 12, 4201–4206.

Zaliapin, I., A. Gabrielov, V. Keilis-Borok, and H. Wong (2008). Clustering analysis of seismicity and aftershock identification, *Phys. Rev. Lett.* **101,** no. 1, 018501, doi: 10.1103/PhysRevLett.101.018501.

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,** no. 3, 1184–1195.

Zhuang, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model, *Earth Planets Space* **63,** no. 3, 207–216.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002). Stochastic declustering of space-time earthquake occurrences, *J. Am. Stat. Assoc.* **97,** no. 458, 369–380.

Swiss Seismological Service (SED)
ETH Zürich
No. H 51.3, Sonneggstrasse 5
8092 Zürich, Switzerland
stefanie.seif@sed.ethz.ch
(S.S., S.N.)

AXIS Re
Alfred Escher-Strasse 50
8002 Zürich, Switzerland
(J.D.Z.)

Institute of Geophysics
No. H 66, Sonneggstrasse 5
8092 Zürich, Switzerland
(A.M.)

Swiss Seismological Service (SED)
No. H 61, Sonneggstrasse 5
8092 Zürich, Switzerland
(S.W.)

Manuscript received 6 July 2017;
Published Online 4 December 2018