Geophysical Journal International

Geophys. J. Int. (2015) **202,** 1949–1968 doi: 10.1093/gji/ggv259
GJI Seismology

Artefacts of earthquake location errors and short-term incompleteness on seismicity clusters in southern California

Ilya Zaliapin[1] and Yehuda Ben-Zion[2]

[1] *Department of Mathematics and Statistics, University of Nevada, Reno, NV* 89557*, USA. E-mail: zal@unr.edu*
[2] *Department of Earth Sciences, University of Southern California, Los Angeles, CA* 90089*, USA*

Accepted 2015 June 11. Received 2015 June 10; in original form 2015 April 26

SUMMARY

We document and quantify effects of two types of catalogue uncertainties—earthquake location errors and short-term incompleteness—on results of statistical cluster analyses of seismicity in southern California. In the main part of the study we analyse 117 076 events with $m \geq 2$ in southern California during 1981–2013 from the waveform-relocated catalogue of Hauksson *et al.* We present statistical evidence for three artefacts caused by the absolute and relative location errors: (1) Increased distance between offspring and parents. (2) Underestimated clustering, quantified by the number of offspring per event, the total number of clustered events, and some other statistics. (3) Overestimated background rates. We also find that short-term incompleteness leads to (4) Apparent magnitude dependence and temporal fluctuations of *b*-values. The reported artefacts are robustly observed in three additional catalogues of southern California: the relocated catalogue of Richards–Dinger & Shearer during 1975–1998, and the two subcatalogues—1961–1981 and 1981–2013—of the Advances National Seismic System catalogue. This implies that the reported artefacts are not specific to a particular (re)location method. The comparative quality of the four examined catalogues is reflected in the magnitude of the artefacts. The location errors in the examined catalogues mostly affect events with $m < 3.5$, while for larger magnitudes the location error effects are negligible. This is explained by comparing the location error and rupture lengths of events and their parents. Finally, our analysis suggests that selected aggregated cluster statistics (e.g. proportion of singles) are less prone to location artefacts than individual statistics (e.g. the distance to parent or parent–offspring assignment). The results can inform a range of studies focused on small-magnitude seismicity patterns in the presence of catalogue uncertainties.

Key words: Earthquake dynamics; Earthquake interaction, forecasting, and prediction; Statistical seismology.

1 INTRODUCTION

Quantitative characterization of spatio-temporal earthquake clustering in relation to physical properties of the lithosphere and response to different natural and anthropogenic loadings are fundamental problems of seismology. During the past decade, novel statistical approaches, improved background information, and high-quality earthquake catalogues available for southern California and other regions have facilitated a set of complementary studies addressing such problems (e.g. Vidale *et al.* 2006; Vidale & Shearer 2006; Enescu *et al.* 2009; Hauksson 2011; Holtkamp *et al.* 2011; Zaliapin & Ben-Zion 2011; Chen *et al.* 2012; Shearer 2012; Brodsky & Lajoie 2013; Ellsworth 2013; Gu *et al.* 2013; Hainzl 2013; Hainzl *et al.* 2014; Moradpour *et al.* 2014). A notable trend in these works is reduction of magnitude of examined events—sometimes below the completeness threshold (e.g. Vidale & Shearer 2006; Chen & Shearer 2013; Eaton *et al.* 2014). In part this is due to the fact that some problems, like structure of and migration within seismic bursts (Vidale & Shearer 2006; Chen *et al.* 2012) or behaviour of induced seismicity (Brodsky & Lajoie 2013), can be only addressed using small-magnitude events. Another incentive to work with small magnitudes comes from the observation that the structure of small-magnitude clusters differs from that of large events (e.g. $M > 6.5$ in southern California) (Zaliapin & Ben-Zion 2013a). Examining small magnitude events allows one to better sample the spatio-temporal domains of interest, and hence address more comprehensive questions comparing to those answered by analysis of clusters of the largest regional events. On the other hand, this increases the exposure to various catalogue uncertainties and errors that may bias the results.

Uncertainties of earthquake location, occurrence time and magnitude are well documented (e.g. Kuge 1992; Röhm *et al.* 1999; Storchak *et al.* 2000; Kagan 2003). These studies focus on modern routine catalogues such as ISC (1995) or PDE (2015) and primarily

examine the error margins in estimations of different earthquake characteristics. Our study has a different goal. It focuses on effects of catalogue uncertainties on estimated earthquake clustering and/or triggering in recently available relocated catalogues, which provide much more precise information about the event locations than the routine catalogues mentioned above. The earthquake clustering/triggering is the main source of information about dynamics of seismicity, hence understanding effects of catalogue errors is important for further improving both seismicity models and earthquake catalogues. In this study we consider two types of catalogue uncertainties: event location errors and short-term incompleteness (insufficient registration of smaller events immediately after a larger one).

To estimate earthquake cluster properties, we use the approach of Zaliapin & Ben-Zion (2013a,b), which is based on earlier works of Baiesi & Paczuski (2004) and Zaliapin *et al.* (2008). The method uses generalized distances between pairs of earthquakes that combine information on location, time and magnitudes (see Section 3); the partition of events into background and clustered populations is facilitated by bimodal distribution of nearest-neighbour distances that is observed in real data, as well as the Epidemic Type Aftershock Sequence (ETAS) and other cluster models.

Our recent application of this approach to the relocated catalogue of Hauksson *et al.* (2012) in southern California during 1981–2011 led to (i) uniform (using the same parameters and rules) and robust identification of thousands of earthquake clusters, including those of small-to-medium magnitude events, in different seismic environments, (ii) classification of statistically significant earthquake clusters into three main types—singles, burst-like clusters and swarm-like clusters, (iii) relating cluster characteristics to observed spatial distribution of heat flow (that controls the effective viscosity of the crust) and (iv) documenting some robust properties of observed seismicity not simulated by the ETAS model (Zaliapin & Ben-Zion 2013a,b). However, extending these results to other seismically active areas and lower magnitude ranges, as well as using them in studies of induced seismicity and evolutionary earthquake patterns, is impeded by the inferior quality of typical data. Most available catalogues are based on non-uniform recordings/analyses that lead to non-uniform (in space, time and magnitude) location errors, varying magnitude of completeness, and other problems. These non-uniformities may (and do) produce artificial patterns in the identification and classification of earthquake clusters, as well as analysis of *b*-values and other properties of seismicity.

In this paper we document and quantify some of the effects of catalogue errors on inferred cluster properties, including striking seismic patterns that emerge as artefacts of those errors. The artefacts include (i) statistically inflated distance-to-parent, (ii) underestimated offspring productivity, (iii) overestimated background rate, (iv) fluctuations of *b*-values, and some related effects like (v) fluctuations in the proportion of singles, (vi) apparent magnitude dependence and (vii) deflated rate of spatial decay of offspring, etc. The reported effects involve the basic element of clustering—parent–offspring pairs (as opposed to more general clusters that may include offspring of offspring). While our analyses use the nearest-neighbour technique of Zaliapin & Ben-Zion (2013a), the described artefacts are expected to also affect alternative clustering techniques. This is because most artefacts (except *b*-value changes) are caused by relatively high ratio of location errors to rupture lengths observed for small events and their parents, which contaminates parent–offspring identification (Section 5).

The paper is organized as follows. Section 2 describes the used data and Section 3 provides a review of the earthquake clustering method. Section 4 analyses the distribution of the catalogue location errors and discusses their principal spatial controls. The main artefacts of catalogue uncertainties and short-term incompleteness in the relocated catalogue of Hauksson *et al.* (2013) are presented in Sections 5 and 6, respectively. Section 7 extends the analyses to three alternative catalogues of different quality for southern California, which confirm our main findings. The results are discussed in Section 8.

2 DATA

The analyses utilize the following three earthquake catalogues for southern California.

2.1 Catalogue of Hauksson, Yang and Shearer (1981–2013)

The principal data source is the waveform-relocated Southern California earthquake catalogue of Hauksson *et al.* (2013) during 1981–2013 (referred to as HYS), available via the SCEC data centre. The methodology of catalogue construction is described in Hauksson *et al.* (2012). Essentially, the catalogue generation consists of three steps. (1) Initial application of HYPOINVERSE (Klein 2002) with a 1-D model. (2) SIMULPS relocation (Thurber 1983, 1993; Eberhart-Phillips 1990) with a 3-D velocity model of Hauksson (2000). At this step, events are relocated one year at a time, and the model is allowed to vary slightly. Events need 5 or more arrivals within a certain distance to qualify. (3) GrowClust relocation (Matoza *et al.* 2013) to improve locations of events within clusters using the differential travel times from the waveform cross-correlations. This refines only clustered events that meet certain criteria.

The final HYS catalogue has 551 455 events; it lists the SCSN preferred magnitudes in the range $m \in [-0.6, 7.3]$ with 0.01 precision. The magnitude distribution is shown in Fig. 2. The *b*-value of events with $m > 3$ is estimated as 1.01 ± 0.02 (with error margins corresponding to 95 per cent confidence interval) according to the method of Tinti & Mulargia (1987). The *b*-value for events with $m > 2$ is 0.95 ± 0.005. This suggests that the catalogue completeness magnitude is somewhere between 2 and 3, which may differ from region to region depending on the quality of seismic network.

The catalogue events are combined into *similar event clusters* based on waveform cross-correlation analysis (Hauksson *et al.* 2012). There are 3577 similar event clusters with size ranging from 5 to 2745. The similar event clusters combine a total of 408 065 events (74 per cent) and 143 390 earthquakes (26 per cent) do not belong to such clusters. Each event is assigned absolute horizontal and vertical 'standard location errors' corresponding to the standard deviation of a Normal isotropic distribution of the location uncertainty. Events in clusters are also assigned corresponding relative errors with respect to other events in the cluster. The catalogue also reports the number of differential times and the number of *P* and *S* picks used to locate events; these characteristics are used in our analysis. We examine 117 076 events with $m \geq 2$ (Fig. 1). There are 1887 similar event clusters that contain 77 110 (66 per cent) of the events.

2.2 Catalogue of Richards-Dinger and Shearer (1975–1998)

The catalogue of Richards-Dinger & Shearer (2000) uses an alternative approach to event relocation. The catalogue construction starts

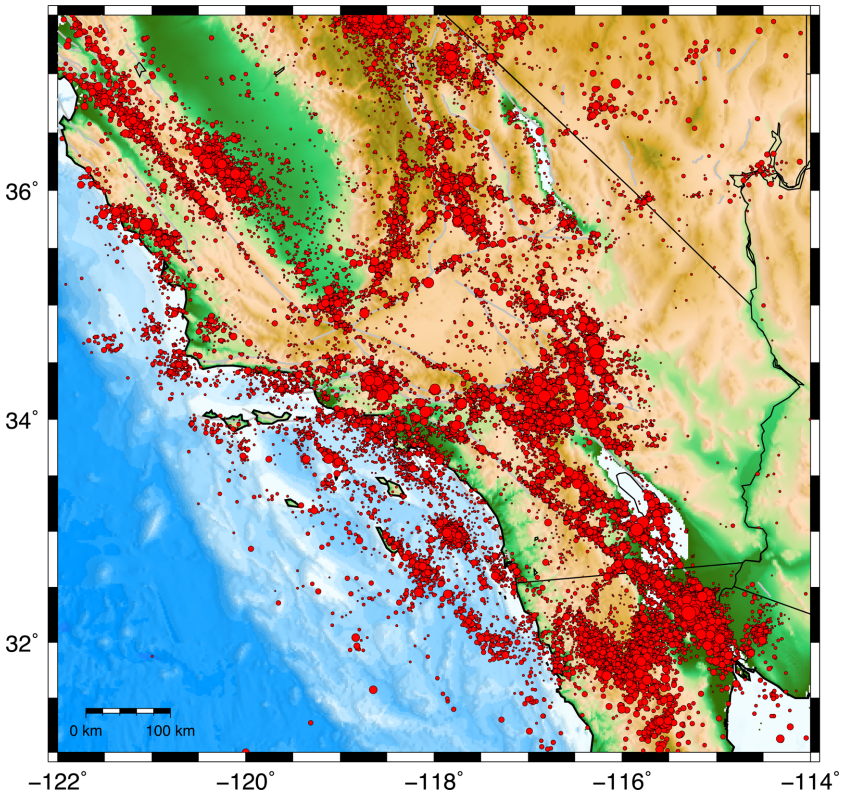

Figure 1. Seismicity of southern California. Epicentres of earthquakes with $m \geq 2$ during 1981–2013 according to the catalogue of Hauksson *et al.* (2013) are shown by red circles. The circle size is proportional to earthquake magnitude. Background colours correspond to topography. Grey lines depict the major faults.

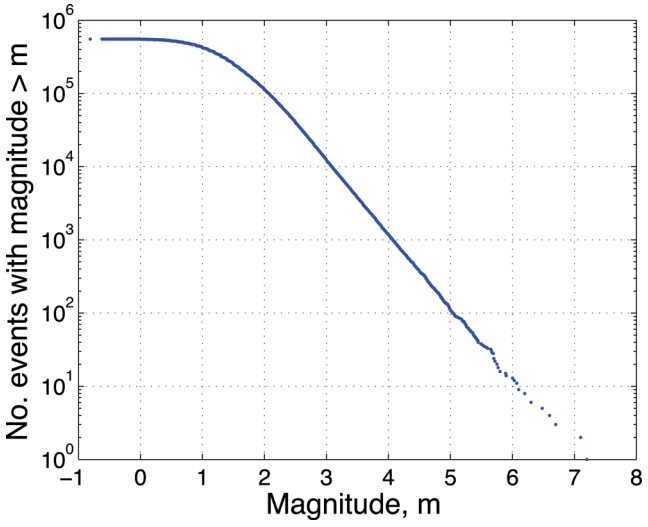

Figure 2. Magnitude distribution in southern California during 1981–2013 according to the earthquake catalogue of Hauksson *et al.* (2013).

with a 1D velocity model and employs station correction terms to account for 3-D velocity structure effects. The catalogue spans the period from 1975 January 01 to 1998 March 18 and covers the southern California region with latitudes between N32.57 and N36.59 and longitudes between W121.3 and W115.23. It includes 58 283 events with magnitudes $m \geq 2$; the magnitude reporting precision is 0.1. The magnitude distribution has fairly linear tail (on semi-logarithmic scale) suggesting that there are no serious problems with completeness (although the completeness magnitude might be above 2.0 in some spatio-temporal areas). The catalogue reports absolute horizontal location errors in both south–north and west–east directions. We use the maximum of these values as a scalar measure of horizontal location uncertainty.

2.3 Catalogue by Advanced National Seismic System (1961–2013)

The ANSS catalogue (ANSS 2015) for southern California uses events provided by the Southern California Seismic Network (SCSN). This catalogue does not apply relocation and is determined using a 1-D velocity model with station corrections; the methodology has varied slightly through the years (Hutton *et al.* 2010). We use two sub-catalogues of the ANSS catalogue that cover the time intervals of 1961–1981 and 1981–2013, referred to as **ANSS-1**, and **ANSS-2,** respectively. The later interval is chosen to coincide with the time coverage of the HYS catalogue. The earlier interval is chosen as a catalogue of comparable duration but with provably lower location quality than that of HYS.

The ANSS-1 catalogue lists 32 585 events with magnitude $m \geq 2$; the magnitudes are reported with 0.01 precision. The catalogue has obviously larger completeness magnitude than the others considered in this study. The overall completeness magnitude is above 4.0, as can be judged from the shape of the magnitude distribution tail (not shown). The ANSS-2 catalogue has 138 426 events with magnitude $m \geq 2$; the magnitudes are reported with 0.01 precision. The completeness magnitude is somewhere between 2 and 3, as can be judged from the shape of the magnitude–frequency distribution (not shown).

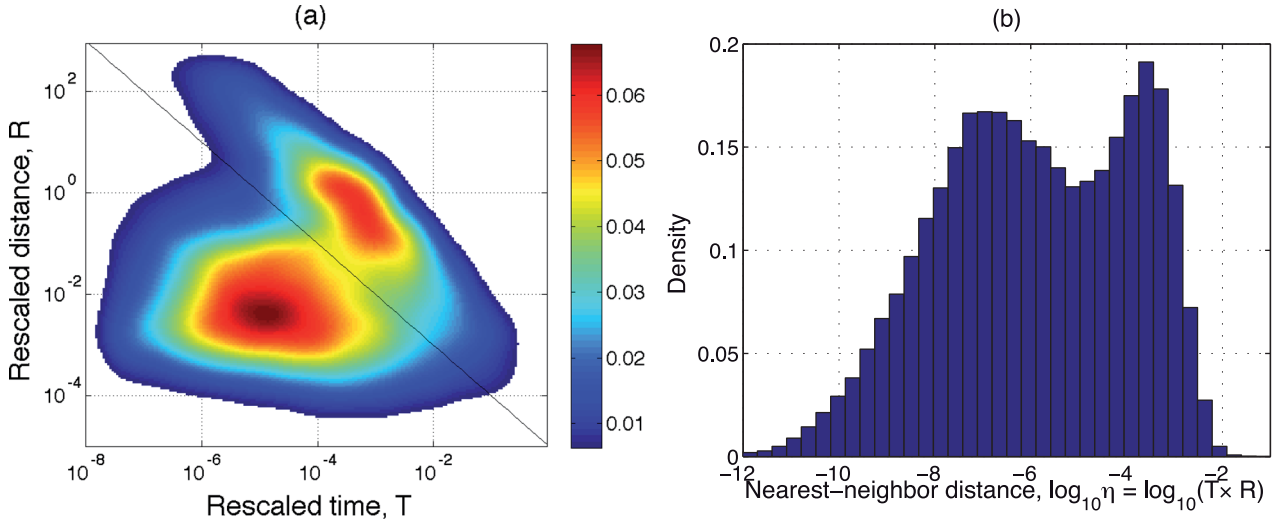

Figure 3. The bimodal distribution of earthquake nearest-neighbour distance. (a) The joint density of the rescaled components ($\log_{10} T$, $\log_{10} R$) of the earthquake nearest-neighbour distance. The lowest 5 per cent of the density values are made transparent to better depict the distribution's support. (b) Density of the nearest-neighbour distance $\log_{10} \eta$. The analysis is done for earthquakes of $m \geq 2$ in Southern California during 1981–2013 from the relocated catalogue of Hauksson *et al.* (2013). The black line $T \times R = 10^{-5}$ in panel (a) visually depicts the separation between two seismic modes. The mode in the upper right part is expected in a Poisson process; it is referred to as the *background* mode. The mode in the left bottom part corresponds to events that occur closer to their parents than expected in a Poisson process; it is referred to as the *clustered* mode.

3 EARTHQUAKE CLUSTERING METHOD

This section reviews the clustering method and key results of Zaliapin and Ben-Zion (2013a,b) used in this study to identify and analyse parent–offspring pairs of earthquakes.

3.1 Offspring and parent identification

The main tool of the offspring–parent identification is a particular earthquake distance in time–space–magnitude coordinates introduced by Baiesi & Paczuski (2004).

3.1.1 Earthquake distance

Each event i in an earthquake catalogue is characterized by its occurrence time t_i, hypocentre (ϕ_i, λ_i, d_i) and magnitude m_i. The distance between earthquakes i and j is asymmetric in time and is defined as (Baiesi & Paczuski 2004):

$$\eta_{ij} = \begin{cases} t_{ij}(r_{ij})^d 10^{-bm_i}, & t_{ij} > 0; \\ \infty, & t_{ij} \leq 0. \end{cases} \quad (1)$$

Here $t_{ij} = t_j - t_i$ is the intercurrence time, $r_{ij} \geq 0$ is the spatial distance between the hypocentres, and d is the (possibly fractal) dimension of the hypocentres.

3.1.2 Nearest-neighbour analysis

For each earthquake j, we find its nearest-neighbour i and the corresponding nearest-neighbour distance η_{ij}. The *nearest neighbour* of an earthquake is called the *parent*. Each event has a single parent and can be the parent of multiple events that are referred to as its *offspring*.

Consider the space and time distances between the pairs of nearest neighbours normalized by the magnitude of the parent:

$$T_{ij} = t_{ij} 10^{-qbm_i}; R_{ij} = (r_{ij})^d 10^{-pbm_i}; q + p = 1. \quad (2)$$

We refer to T and R as rescaled time and rescaled distance to parent, respectively. It is readily seen that $\log \eta_{ij} = \log T_{ij} + \log R_{ij}$. Zaliapin *et al.* (2008) and Zaliapin & Ben-Zion (2013a) demonstrated that a time-stationary space-inhomogeneous Poisson distribution of events with Gutenberg–Richter magnitudes corresponds to a unimodal distribution of ($\log T$, $\log R$) that is concentrated along a line $\log T + \log R = \text{const}$. Notably, observed catalogues show a bimodal joint distribution of ($\log T$, $\log R$). One mode corresponds to *background events* (similar to events in a Poisson process), while the other consists of a large subpopulation of events located considerably closer in time and space to their parents than expected in a Poisson process—we refer to them as *clustered events*. The distribution of the nearest-neighbour earthquake distance η, and the joint distribution of space and time components of the nearest-neighbour earthquake distance in southern California are shown in Fig. 3.

3.1.3 Parameters

We perform all analyses working with event epicentres and using $b = 1, d = 1.6$ and $p = 0.5$. Zaliapin & Ben-Zion (2013a) have shown that the cluster analysis is stable with respect to these parameters. In particular, none of the qualitative findings of this study will change if any of the parameters will fluctuate within reasonable limits.

3.1.4 Close offspring, parents

Each parent–offspring pair is assigned the corresponding nearest-neighbour distance η. We will pay particular attention to *close* offspring and parents separated by earthquake distances $\eta < \eta_0 = 10^{-5}$. The threshold value $\eta_0 = 10^{-5}$ is suggested by the bimodal distribution observed in Fig. 3. Events with no close parent and offspring are called *singles*, while events with close parent and at least one close offspring are called *internal cluster events*.

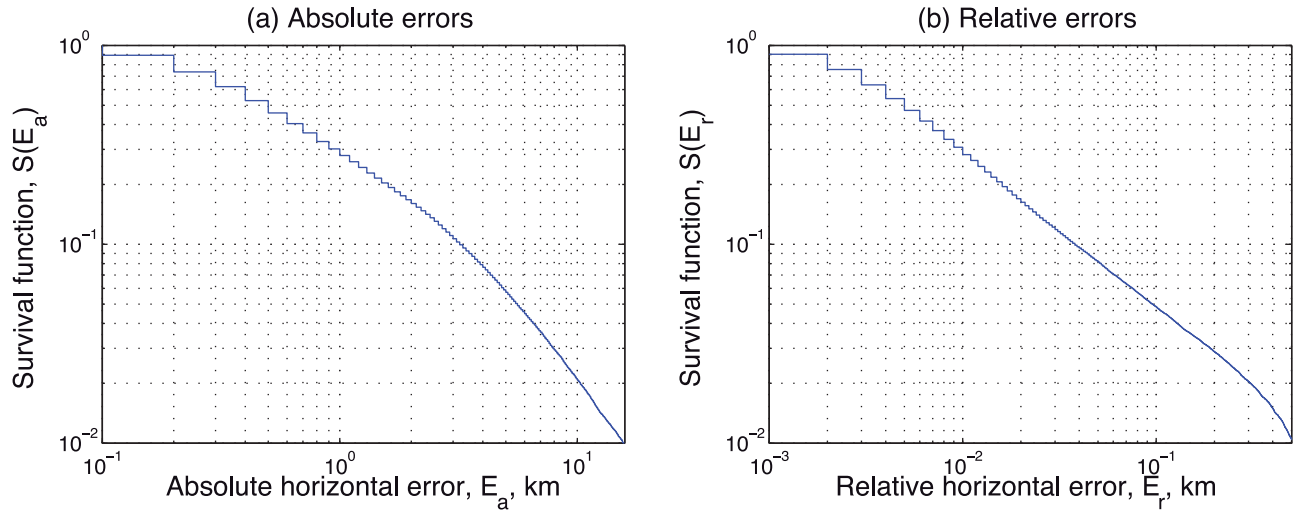

Figure 4. Distribution of location errors for earthquakes with $m \geq 2$ in southern California during 1981–2013 according to the catalogue of Hauksson *et al.* (2013). (a) Absolute horizontal error. (b) Relative horizontal error. The axes are chosen to show only 99 per cent of the smallest errors. The largest 1 per cent corresponds to a very rapid decay of the distribution's tail. The y-axis shows the survival function $S(x) = \text{Prob}(\text{Error} > x)$.

4 CATALOG OF HAUKSSON *ET AL.* (2013): LOCATION ERRORS

4.1 Distribution of error values

The distribution of absolute and relative horizontal errors for earthquakes with $m \geq 2$ in the Southern California HYS catalogue during 1981–2013 is shown in Fig. 4. The absolute errors take values from 0.1 to 99 km; they are reported with 100 m resolution. No absolute error is reported for 2218 events (1.8 per cent); 102 605 events (90 per cent of events with reported error) have absolute errors below 3 km; 24 events have zero error reported. The relative errors take values from 1 m to 1.5 km; they are reported with 1 m resolution. No relative error is reported for 43 186 events (37 per cent), which includes 3220 events within similar event clusters; 66 784 (90 per cent of events with reported relative errors) have relative error below 40 m; 817 events have zero relative error reported (all from similar event clusters). We note that the reported relative errors are significantly smaller than the rupture lengths of the respective events. These values should be interpreted as a convenient measure of relative location quality rather than actual errors (E. Hauksson, personal communication, 2014). For instance, 1 m errors signify the best relative location with respect to similar events, 2 m is the next level of location quality, etc.

Fig. 5(a) shows the joint distribution of absolute and relative horizontal errors, including non-reported values; the largest error values are binned for visual convenience. The distribution does not exhibit any notable dependencies for absolute errors below 10 km and relative errors below 100 m, except the observation that events with high absolute error and low relative error seem to be less probable. Fig. 5(b) shows the joint distribution of absolute horizontal and vertical errors, which exhibit prominent positive nonlinear association. Comparing Figs 5(a) and (b) we conclude that the absolute and relative errors exhibit much weaker dependence than horizontal and vertical errors. Similarly, a weak dependence is seen between

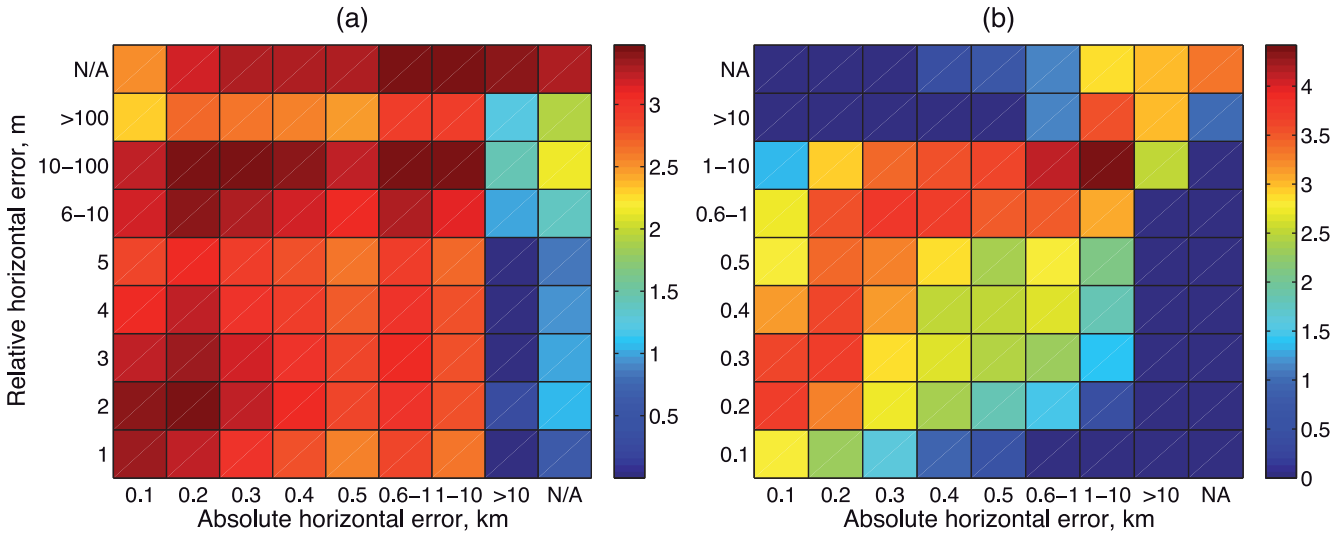

Figure 5. Joint distribution of the location errors in southern California catalogue of Hauksson *et al.* (2013) during 1981–2013. (a) Absolute and relative horizontal errors. (b) Absolute horizontal and vertical errors. The higher error values are binned for visual convenience. The colour bar corresponds to logarithmic (base 10) count of events.

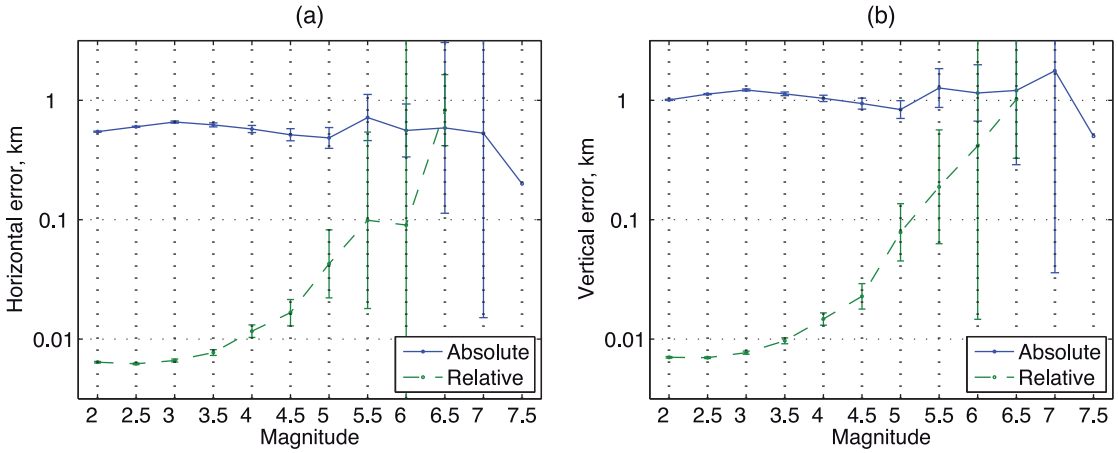

Figure 6. Absolute (solid blue) and relative (dashed green) location error as a function of event magnitude. (a) Horizontal error; (b) Vertical error. The figure shows the mean (lines/dots) and 95 per cent confidence intervals for the mean (error bars).

absolute and relative vertical errors, and much stronger dependence between relative horizontal and vertical errors (not shown). The weak dependence between absolute and relative errors is related to the different methods of absolute and relative error estimation (e.g. Richards-Dinger & Shearer 2000; Hauksson *et al.* 2012; Matoza *et al.* 2013).

Another reflection of different estimation methods for absolute and relative errors is seen in magnitude dependence of location errors. As seen in Fig. 6, both horizontal and vertical absolute errors are only weakly dependent on magnitude. However, the relative errors exhibit significant (over two orders of magnitude) increase as event magnitude changes from 2 to 6.5. This may be related to the fact that the larger is the event magnitude; the worse is the point approximation of the source.

4.2 Spatial controls of location errors

The values of location errors exhibit strong geographical dependence, as illustrated in Fig. 7 for the absolute and relative horizontal errors. The location errors significantly decrease in central southern California with respect to the peripheral areas. The observed variations are closely related to the quality of the seismic network, which also exhibit significant spatial variations. This is illustrated in Fig. 8 that shows the spatial distribution of three catalogue characteristics that can be used as a proxy to the network quality: number of *P* and *S* picks used to locate the event (panel a), number of differential times (panel b) and number of similar events (panel c). Among the three examined statistics, the number of picks exhibits the best spatial correlation with the event location error. At the same time, the spatial distribution for the number of picks is different from that for the number of differential times and number of similar events. The latter two statistics are very similar to each other and, not unexpectedly, have prominently increased values in the most seismically active fault zones: Ventura, San Bernardino, San Gabriel, San Jacinto, Salton Sea and Coso.

5 ARTEFACTS OF LOCATION ERRORS

5.1 Artefact 1: inflated distance-to-parent

The most direct and easily detected effect of location errors is statistical increase of the spatial separation between offspring and

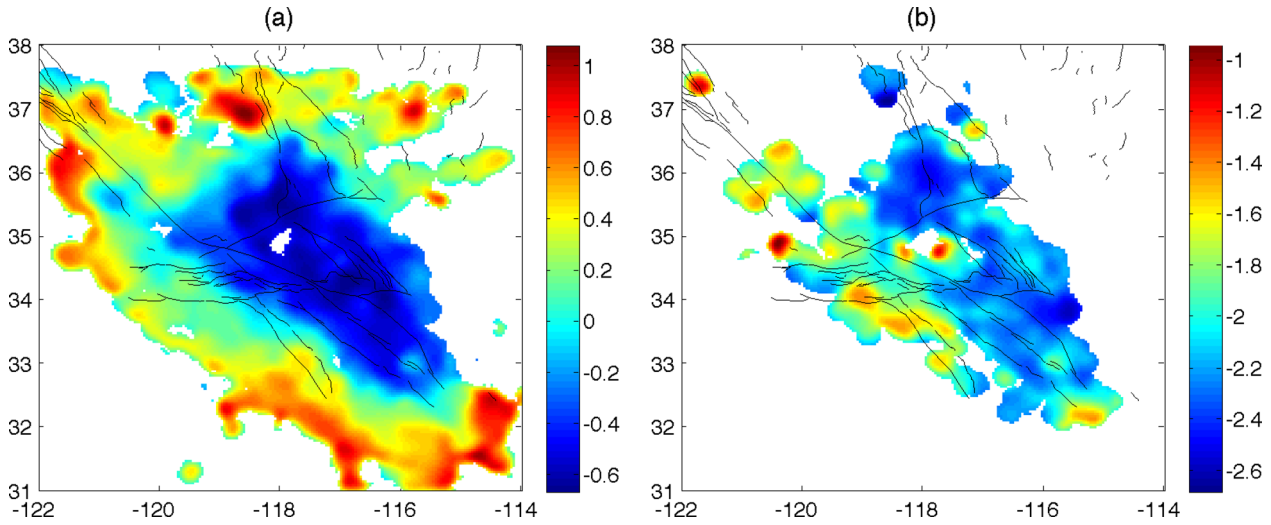

Figure 7. Spatial distribution of absolute (panels a) and relative (panel b) horizontal error (on log10 scale) in southern California according to the catalogue of Hauksson *et al.* (2013). The analysis uses all events with $m \geq 2$ and a valid reported error value. Black lines show the major faults. The figure shows the average values of error, estimated within a circle of radius 20 km; the results are reported for circles that contain 5 or more events with a reported error value.

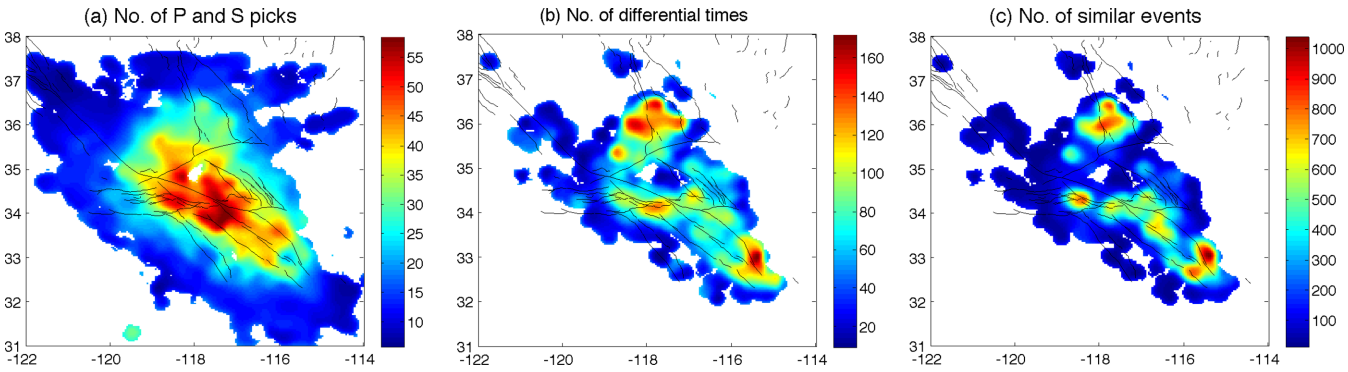

Figure 8. Spatial distribution of various catalogue statistics in southern California, 1981–2013, $m \geq 2$, according to the catalogue of Hauksson *et al.* (2013). (a) Number of *P* and *S* picks, (b) Number of differential times, (c) Number of similar events. The figure shows the average values of selected statistics per event, estimated within a circle of radius 20 km; the results are reported for circles that contain 5 or more events with a reported value of the examined statistic.

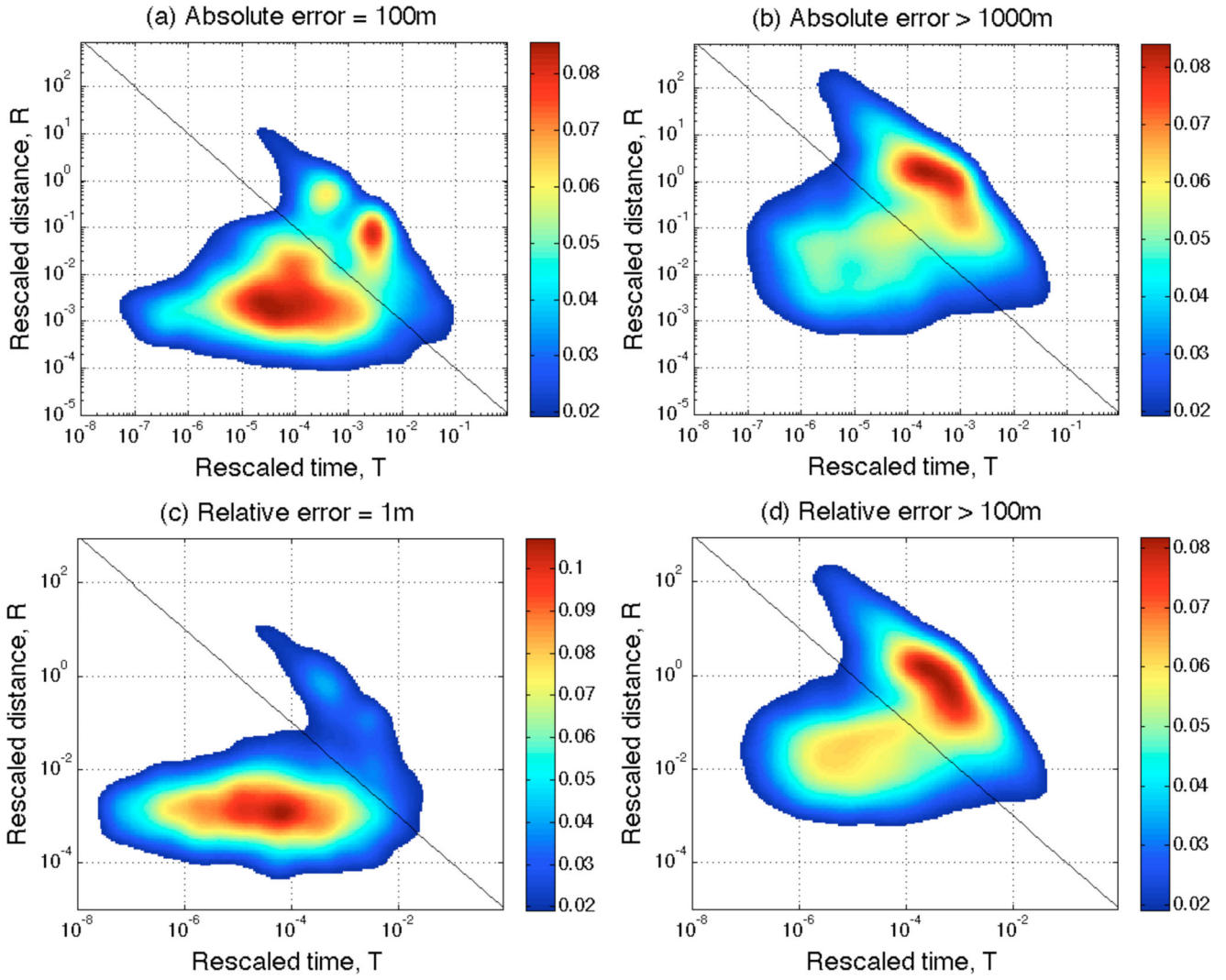

Figure 9. Artefact 1—inflated distance-to-parent. The joint density of the rescaled time (T) and distance (R) to parent as defined in eq. (2). (a) Small absolute horizontal location error (100 m). (b) Large absolute horizontal location error (>1 km). (c) Small relative horizontal location error (1 m). (d) Large relative horizontal location error (>100 m). The estimated distance-to-parent (vertical axis) and proportion of background events (points above the black line) increase with location error. Black line shows the threshold $T \times R = 10^{-5}$ used to separate background and clustered events.

parents. This effect is seen in Fig. 9, which compares the joint distribution of the rescaled time T and distance R of eq. (2) for events with small versus large values of the absolute horizontal location error (panels a, b), and small vs. large relative horizontal error (panels c, d). The estimated distance-to-parent R significantly increases with the values of the horizontal error—this is reflected in higher location of the (T, R) distribution for large-error events (panels b, d) along the vertical axis, compared to that of the

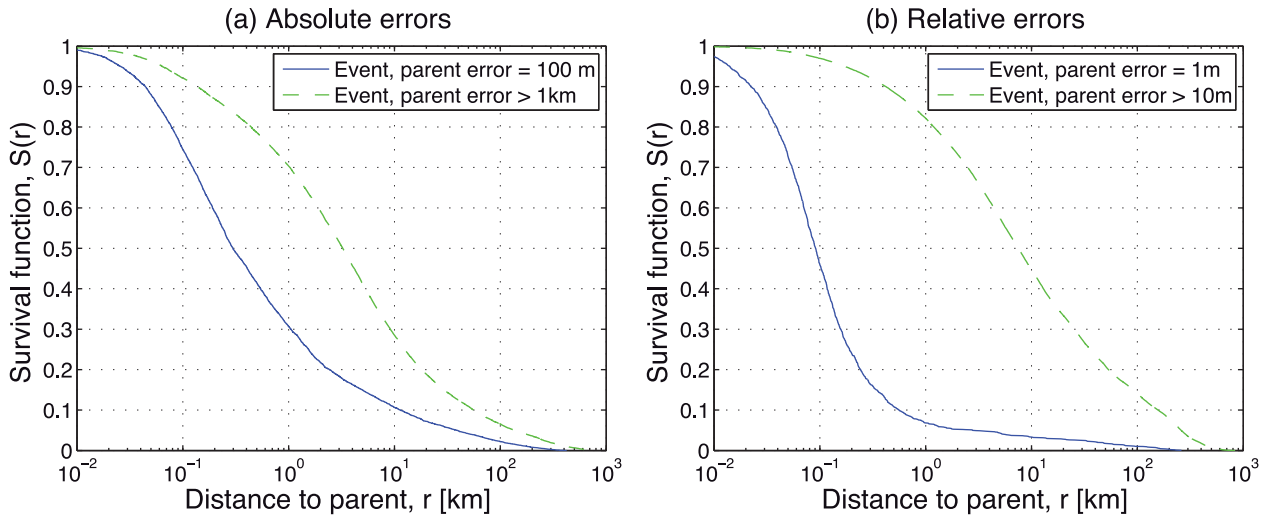

Figure 10. Artefact 1—inflated distance-to-parent. Distribution of the distance to parent for small (solid blue line) and large (green dashed line) location errors for both event and its parent. (a) Absolute horizontal errors. (b) Relative horizontal errors. The y-axis shows the survival function $S(r) = \text{Prob}(\text{Distance} > r)$.

low-error events (panel a, c). The typical rescaled distance-to-parent R increases more than an order of magnitude (from 10^{-3} to $10^{-1.5}$) when the absolute error increases from 100 m to over 1 km, or when the relative error increases from 1 m to over 100 m.

Analysis of the absolute (non-rescaled) distance r to parent with relation to location error is illustrated in Fig. 10. Here we compare the distribution of r for the pairs parent–offspring with small (solid blue) and large (dashed green) absolute (panel a) and relative (panel b) errors. In both analyses there is a shift of the distribution to larger distances as location error increases (from solid blue to dashed green). The size of this effect, measured here by the difference of medians in the small-error and large-error distributions, is an order of magnitude in absolute error analysis (300–3000 m) and two orders of magnitude in relative error analysis (100–10 000 m).

We notice that the results of Fig. 10(b) are affected by relation between the event magnitude and relative location error that was illustrated in Fig. 6. Namely, the larger relative location errors are statistically associated with larger magnitude events. The larger magnitude parents, in turn, are associated with larger distance to offspring due to their larger rupture size, as follows from eq. (1). This leads to a larger proportion of larger distances for large-error parents, and larger proportion of smaller distances for small-error parents. Accordingly, the difference between the two distributions in Fig. 10(b) is due to two related effects: the size of relative location errors *and* increased distance-to-parent for offspring that have a large-magnitude parent. The combination of these two effects makes the difference between the examined distributions much larger than that for the absolute errors in Fig. 10(a) that are not affected by the magnitude dependence.

The described effects might be somewhat affected by varying magnitude of catalogue completeness, as larger number of small events would lead to increase of small offspring–parent distances. However, distance-to-parent analysis repeated for events above $m = 3$ (not shown) leads to same qualitative conclusions. Hence, we conclude that the main driver for inflated distance to parent is imprecise event location. This may result from two alternative scenarios. First, a mislocated event is not able to detect the parent that would be identified in the (hypothetical) catalogue with no errors, and instead finds an alternative parent that is most probably located at a larger distance. Second, a mislocated event finds the same parent as in the catalogue with no errors, but the location errors place it farther away from the parent. Our analysis below (Section 5.2) suggests that both scenarios play an important role in parent identification.

The artificial increase of the distance-to-parent can lead to a cascade of related artefacts in estimated spatial properties of offspring/aftershock/foreshocks. For example, the rates of spatial decay of offspring are expected to decrease with increasing location error, since a larger proportion of offspring are artificially placed farther away from their parents in increased-error areas. A similar effect might be observed in studying spatial foreshock distribution.

The discussed effects do not seem to be intrinsically connected to our particular mechanism of parent identification. Similar artefacts of error locations should exist in results of other cluster approaches.

5.2 Artefact 2: underestimated offspring productivity

This section explores in greater detail the situation when a mislocated event is not able to identify the (mislocated) parent that would have been identified in a catalogue with no errors. This is one of the scenarios considered in Section 5.1 that leads to inflated distance-to-parent.

Recall that we define *close* events as those separated by the combined (rescaled space-time) distance $\eta < 10^{-5}$, with the threshold value suggested by the bimodal distance distribution of Fig. 3. Fig. 11 shows the proportion of events with close offspring (panel a) and those with close parent (panel b) as a function of absolute horizontal location error. Different colours correspond to different magnitude groups indicated in the legend. All the examined proportions significantly decrease as the error increases. In addition, the overall value of the proportion of events with close offspring (panel a) increases with magnitude, which is expected due to the increased offspring productivity. Within each magnitude group in Fig. 11(a), the proportion drops by about 50 per cent as the error increases from 0.1 km to over 10 km. The observed proportion drop is not uniform—the largest decrease is observed at the highest values of the absolute horizontal error (right part of the plot). The proportion of events with close parents might only weakly depend on magnitude—the three curves in Fig. 11(b) are largely overlapping. The proportion decreases from about 0.8 to 0.3–0.4 when the error increases from 0.1 km to over 10 km. Again, the most rapid drop is observed for the largest errors (over 1 km). The results in

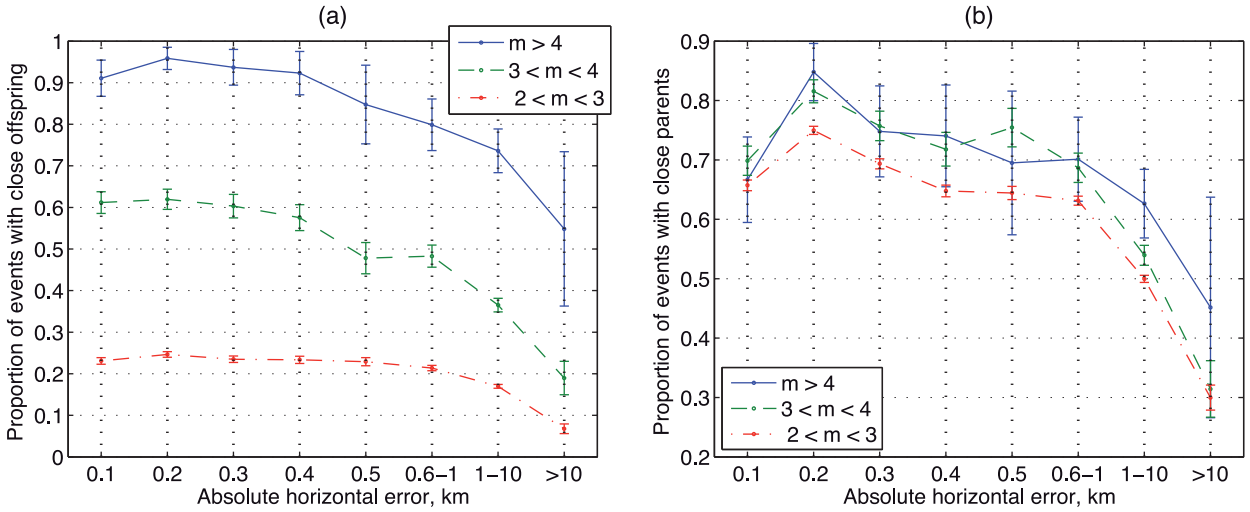

Figure 11. Artefact 2—underestimated offspring productivity. Both the proportion of events with close offspring (panel a) and proportion of events with close parents (panel b) decrease with increasing absolute horizontal errors.

Fig. 11 indicate that location errors corrupt parent–offspring identification leading to underestimation of the offspring productivity and, accordingly, to underestimation of the total number of clustered events.

Similar conclusions can be made by analysing the proportion of *singles* and *internal cluster events* as a function of location error (Fig. 12). The proportion of singles in the entire catalogue significantly increases, while the proportion of internal events decreases, with increasing value of either absolute (panel a) or relative (panel b) horizontal error. This effect also holds within individual regions, as shown in panels (c), (d) for the Salton Sea and San Jacinto regions, respectively. The effect is strong and clearly seen despite significantly different average proportions of singles and internal events in the examined regions—The Salton Sea area has higher proportion of internal cluster events (average of 0.3) and lower proportion of singles (average of 0.2) with respect to San Jacinto region having average proportion of internal events of 0.1 and average proportion of singles of 0.4. This difference of average proportions is likely related to the different levels of heat flow and, more generally, effective viscosity of the two regions, as discussed and further illustrated in Zaliapin & Ben-Zion (2013b).

Fig. 13 presents yet further evidence of offspring productivity deflation by location errors, by showing the average number of offspring per event as a function of relative horizontal error, for events within different magnitude ranges. The average number of offspring steadily decreases as the error increases; the overall drop in the offspring productivity is above 70 per cent as error increases from 1 m to over 100 m. To summarize, the complementary evidence presented in Figs 11–13 suggests that the reported decrease in offspring productivity is primarily due to location errors.

To better understand this effect, we compare the location errors with the rupture lengths of events and their parents. For crack-like events with circular areas, the rupture length can be estimated as (Ben-Zion 2008):

$$L_m = 0.0152 \times 10^{0.42m}.$$

This approximation works well for events with $m < 5.5$ and somewhat underestimates the rupture lengths of larger events (*cf.* Fig. 6 of Ben-Zion 2008). Fig. 14(a) shows the logarithm of the ratios of absolute and relative horizontal location errors to the estimated rupture length L_m for events of different magnitudes m. The average ratio for the absolute error (solid blue) is above unity for $m < 3.5$, which implies that the location error exceeds the rupture length for many such events. Naturally, such large errors might significantly contaminate identification of close neighbours and associated cluster analysis. The relative errors (dashed green) are typically an order of magnitude smaller than the event's rupture length. However, since the scale of the relative errors is somewhat arbitrary (E. Hauksson, personal communication) we refrain from discussing physical meaning of these ratios.

In the cluster analysis used in this study (Section 3), the parent assignment is influenced by the magnitude (and hence rupture area) of the parent, but is independent of the magnitude (and hence rupture area) of a potential offspring as dictated by the definition of the earthquake distance in eq. (1). Fig. 14(b) shows the logarithm of the ratio of horizontal location errors for events of magnitude $2 \leq m \leq 5.5$ to the estimated rupture length of their parents. The events with $m > 5.5$ are not examined as they produce very large error bars due to their small number in each magnitude group. For the absolute errors (solid blue), the ratio stays in the relatively narrow range of

$$10^{[-0.1,-0.6]} \approx [0.25 \quad 0.79],$$

indicating that the event's location error is typically comparable to the half-rupture-length (ratio 0.5) of its estimated parent. This ratio is rather high, confirming that parent identification for many events might be corrupted. Fig. 14(c) shows the logarithm of the ratio of the horizontal location errors to the estimated parent rupture length for different parent magnitudes m. The ratio decreases with parent magnitude in exponential fashion with index of about -0.5. The ratio for absolute errors (solid blue) is above unity for parent magnitudes $m < 3.5$. This suggests that events with $m < 3.5$ might lose large number of offspring due to absolute location errors.

To summarize, the results in Fig. 14 suggest that it is very probable for many events to be mislocated by distance that is sufficiently large to prevent correct identification of the parent. The size of this effect increases with location error and decreases with the magnitude of event (Fig. 14a) and the event's parent (Fig. 14c). In particular, the analysis suggests that offspring misidentification should be negligible for parents with $m > 6$, as for these events the average ratio of offspring absolute location error to parent rupture length

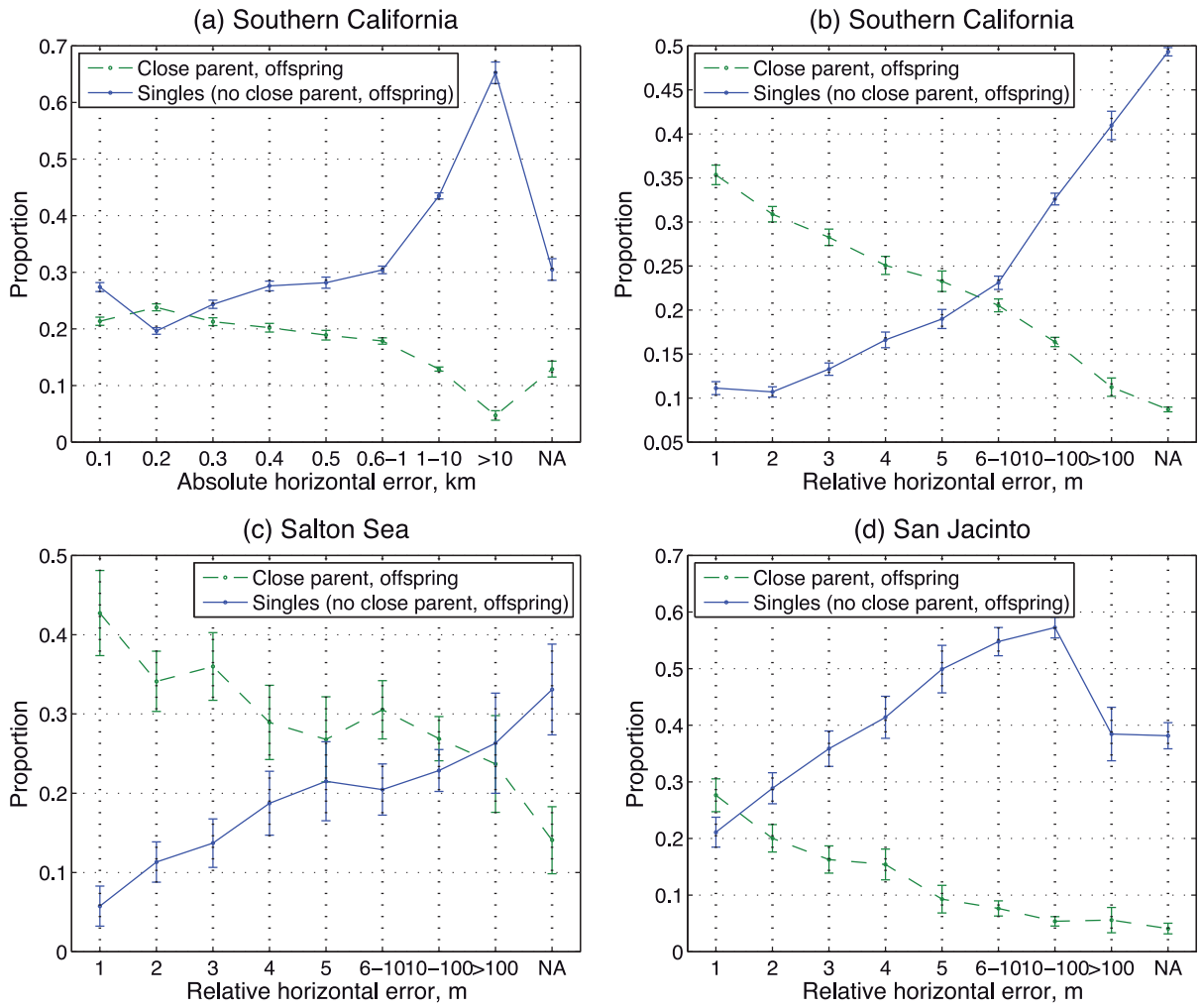

Figure 12. Artefacts 2 and 3—underestimated offspring productivity and overestimated background rate. Proportion of singles—events with no close parent–offspring (solid blue) and internal cluster events—events with close parent and offspring (dashed green) as a function of absolute (panel a) and relative (panel b) horizontal error in entire southern California, as well as the same proportion vs. relative errors in individual regions: Salton Sea (panel c) and San Jacinto (panel d). Points and lines refer to the mean values within each error group; error-bars show 95 per cent confidence intervals for the means.

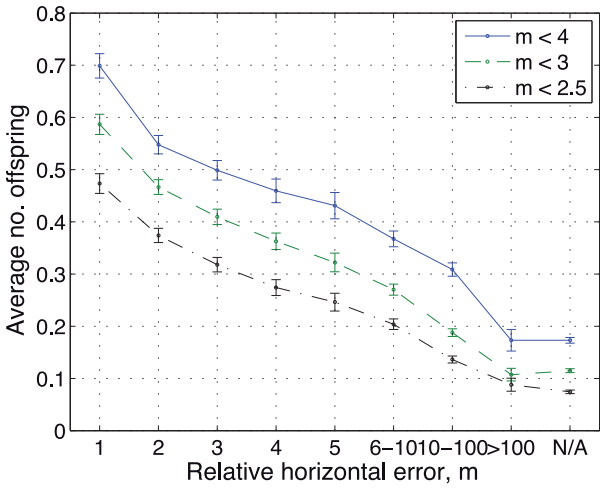

Figure 13. Artefact 2—underestimated offspring productivity. Average number of close offspring per event decreases with the relative horizontal error. Different lines correspond to different range of parent magnitudes, as described in the legend. Points and lines refer to the mean values within each error group; error-bars show 95 per cent confidence intervals for the means.

is below 0.1. Accordingly, typical studies associated with relatively large main shocks have a very small proportion of offspring lost.

It is also interesting to examine how often a parent–offspring pair belongs to the same cluster of similar events in the HYS catalogue. There are 77 110 (65.8 per cent) such events; off those, 37 703 (49 per cent) have parents from the same cluster of similar events and 39 407 (51 per cent) have parents outside of their similar event cluster. Hence, of all 117 076 events with $m \geq 2$, there are 37 703 (32 per cent) with parents from the same similar event cluster and 79 373 (68 per cent) with parents outside of similar event cluster. Since the majority of parent assignments happen outside of offspring's similar event cluster, one should expect the cluster results to be contaminated by the absolute location errors, as we illustrate above in Figs 13–15.

5.3 Artefact 3: overestimated background rate

An overestimation of the background rate originates when a clustered event with close parent is misidentified as a background event (event with a distant parent). This is complementary to the underestimation of offspring productivity examined the previous section. The analysis of Fig. 11(b) suggests that numerous clustered events

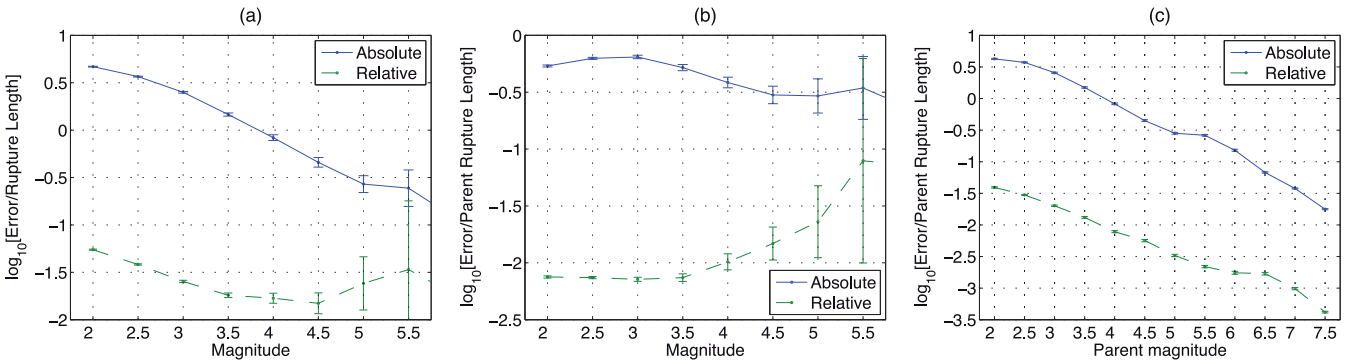

Figure 14. Ratio of the horizontal location error to the estimated rupture length of event (panel a) and event's parent (panels b, c) as a function of the magnitude of event (panels a, b) and magnitude of event's parent (panel c). The figure shows the mean (lines) and the 95 per cent confidence intervals for the mean (error bars) in each group. The axis is cut at $m = 5.5$ in panels (a, b): The larger magnitudes correspond to very large error bars due to small number of events.

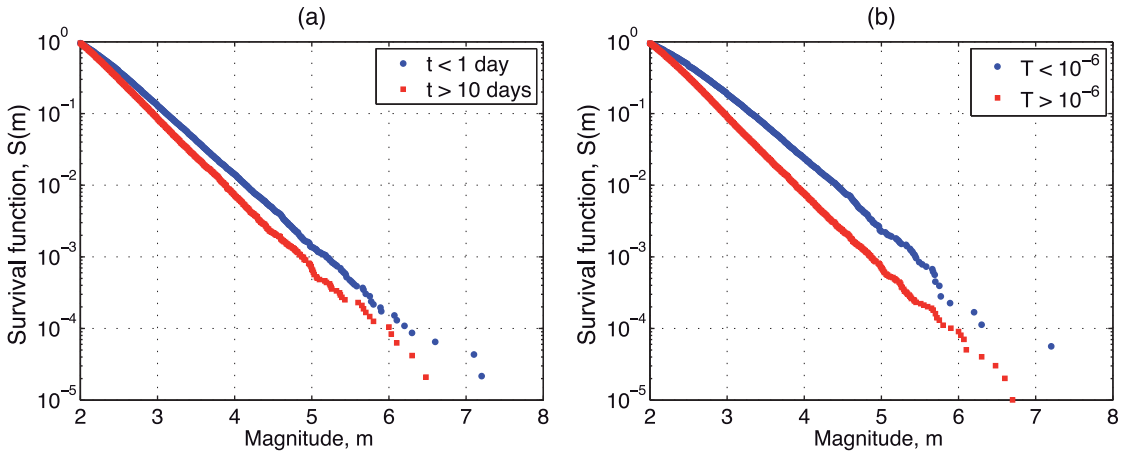

Figure 15. Artefact 4—short-term incompleteness: The *b*-value decreases in the temporal vicinity of the parent. (a) The tail of the magnitude distribution (survival function) for events that happen within 1 d (blue) and after 10 d (red) after the parent. (b) The tail of the magnitude distribution (survival function) for events with the rescaled time T to parent being $T < 10^{-6}$ (blue) or $T > 10^{-6}$ (red).

are misidentified as background events due to location errors. This is further supported by Figs 9 and 12. In Fig. 9 we see an increased proportion of background events (those above the diagonal line) in the population of earthquakes with large location errors (panels b, d). Furthermore, Fig. 12 exhibits at least three-fold increase in the proportion of singles (from about 0.1–0.2 to 0.3–0.6 in different panels) as absolute location errors increase. The effect seems to be even stronger when relative error is considered (Fig. 12b). As mentioned, this increase seems to be mainly caused by location errors rather than other effects like regional differences in completeness magnitude, etc. We notice also that some background events may be misclassified as clustered ones as the results of location errors, although the magnitude of this effect seems negligible.

6 ARTEFACTS OF SHORT-TERM INCOMPLETENESS

6.1 Artefact 4: change of *b*-value

It is well known that many small events are not registered after a large one (e.g. Utsu *et al.* 1995; Wiemer & Katsumata 1999; Narteau *et al.* 2002; Kagan 2004; Lolli & Gasperini 2006; Helmstetter *et al.* 2007; Peng *et al.* 2007). This leads to apparent spatio-temporal changes of *b*-values in time and space after major events. To illustrate this effect in the southern California HYS catalogue, we compare in Fig. 15(a) the magnitude distributions of events with different time delay from their parents: $t < 1$ d (blue) versus $t > 10$ d (red). The temporally close offspring ($t < 1$ d) have visibly lower *b*-value, which reflects the lack of small-magnitude events, or short-term incompleteness. Fig. 15b reports on similar analysis that compares the magnitude distributions of events with different rescaled time to parent defined in eq. (2): $T < 10^{-6}$ (blue) versus $T > 10^{-6}$ (red). Again, the close offspring have visibly lower *b*-value. We note that in both panels the difference in *b*-values is consistently reflected in the maximal magnitude of respective subpopulation: The largest event in both cases belongs to the group with shorter distance to parent. Table 1 confirms this visual impression by more rigorous statistical comparison: It shows the estimated *b*-values and the respective 95 per cent confidence intervals for different selections of times (and rescaled times) to parent for all events (lines 1, 2) and only for the offspring of parents with magnitude $4 \leq m \leq 6$ (lines 3, 4). The estimations and confidence interval computations are done using the method of Tinti & Mulargia (1987) that takes into account magnitude discretization (0.01 in the examined catalogue). All the estimations are done here for events with $m \geq 3$. The results indicate that the *b*-value drops for over 0.1 in the temporal vicinity of a parent, reflecting short-term catalogue incompleteness. This value change is highly significant, as evidenced by non-overlapping 95 per cent confidence intervals of the respective *b*-value estimates.

The examined difference between the close and distant magnitude distributions is more prominent for analysis based on rescaled times

Table 1. Estimating b-value for events with different time to parent in the catalogue of Hauksson *et al.* (2013) (estimator of Tinti & Mulargia (1987); only events with $m \geq 3$ are considered).

		Time to parent		Rescaled time to parent	
		$t < 1$ d	$t > 10$ d	$T < 10^{-6}$	$T > 10^{-6}$
All events	b	0.97	1.07	0.89	1.08
	95 per cent CI	0.95–0.99	1.03–1.10	0.86–0.92	1.05–1.10
Parents with magnitude $4 \leq m \leq 6$	b	0.86	1.02	0.78	1.02
	95 per cent CI	0.82–0.91	0.96–1.07	0.73–0.84	0.98–1.06

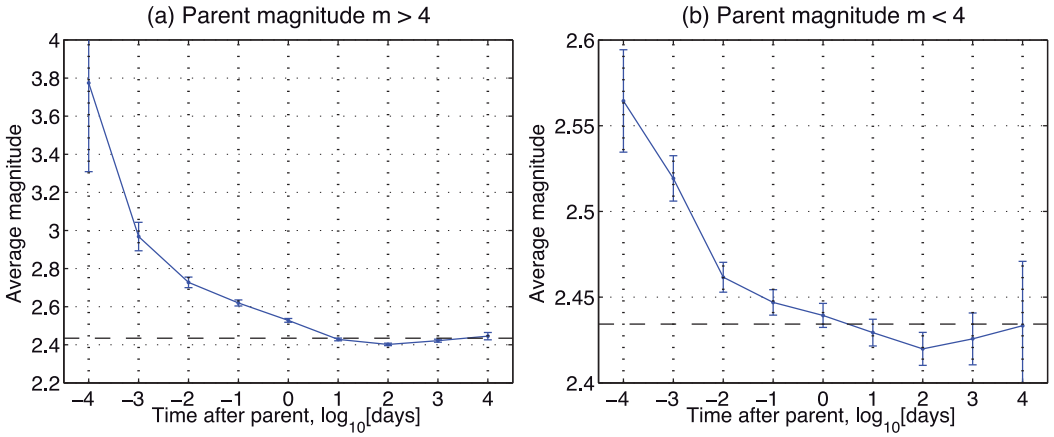

Figure 16. Artefact 4—short-term incompleteness: The average magnitude (a reciprocal proxy for the b-value) decreases with time after the parent. (a) Offspring of parents with $m \geq 4$. (b) Offspring of parents with $m < 4$. The blue line and dots refer to the average within each group, the errorbars correspond to a 95 per cent confidence interval. Dashed black line in each panel shows the theoretical average magnitude of $2 + \log_{10}(e) \approx 2.43$ that assumes $b = 1$ and $m_c = 2$.

(Fig. 15b), as expressed in the following two complementary observations. First, the b-values are significantly lower for the close events defined by $T < 10^{-6}$ than for the close events defined by $t < 1$ d; yet the b-values for the distant events defined under the two approaches are statistically the same (see Table 1). Second, there exists a visible bend in the frequency–magnitude distribution of close events defined via rescaled times (Fig. 15b, blue line) at magnitudes below 3, which indicates reduced number of small-magnitude events. Both observations indicate that the rescaled time approach better identifies time intervals affected by short-term incompleteness. This is because in the rescaled approach the interval of potential short-term incompleteness is scaled by the event magnitude, making it longer for large events and shorter for small events.

Fig. 16 further illustrates the short-term incompleteness effect by tracing how the average offspring magnitude decreases with the time-to-parent. The inverse of the average magnitude is a proxy to the b-value (e.g. Aki 1965; Eneva & Ben-Zion 1997), so decreasing average magnitude corresponds to increasing b-value. Fig. 16(a) refers to offspring of parents with magnitude $m > 4$. The most notable decrease in the average magnitude (increase of b-value) is within 10 d after the parent, after which the examined statistic seems to be at a steady state. Interestingly, a similar incompleteness effect is also seen for small magnitude parents, $2 < m < 4$, as illustrated in Fig. 16(b). The significant decrease of the average magnitude seems to have the same duration (tens of days), although the value drop is much smaller in this case. Similar results are observed for median magnitude (not shown). The results in Figs 15 and 16 are similar to those reported by Hainzl (2013) who used the relocated catalogue of southern California by Lin *et al.* (2007) and examined aftershock sequences of main shocks with magnitudes between 2.5 and 5.5 following the space–time window identification of Shearer (2012).

The short-term incompleteness may explain apparent b-value changes for background vs. clustered events reported recently by Gu *et al.* (2013). Namely, the background events by definition have distant parents ($\eta > 10^{-5}$) which implies larger time-to-parent. On the other hand, clustered events have close parents ($\eta < 10^{-5}$) which implies shorter time-to-parent. Accordingly, the magnitude distribution analysis of Figs 15, 16 and Table 1 explains the b-value changes between background and clustered events.

We notice that in the HYS catalogue there are 62 per cent of clustered events, and 59 per cent of the events occurred within 10 d from their parent. Comparable figures are reported in the other catalogues. Our analyses (Figs 15, 16 and Table 1) suggest that such events are mostly affected by the short-term incompleteness. Accordingly, a range of regional b-value studies might be substantially biased by the short-term incompleteness, even if the analysis is performed for magnitudes exceeding the regional completeness threshold. Indeed, Fig. 16 indicates that the short-term incompleteness affects parents of all magnitudes (not only the largest ones) and hence the artefact of inflated b-value is distributed in space and not confined to the vicinity of the largest main shocks. The overlapping short-term incompleteness regions of events of different magnitudes might create quite complex space-time patterns and result in apparent increase or decrease of the estimated b-value, which in turn may prompt inappropriate physical interpretations.

We note that a deflated b-value is among the best-studied premonitory pattern, reported for decades in observational (e.g. Rikitake 1975; Aki 1981; Habermann 1981; Mogi 1981; Smith 1981, 1986; Hirata 1989; Imoto 1991; Enescu & Ito 2001) and modelling/laboratory (e.g. Hanks 1979; Main *et al.* 1989, 1992; Gabrielov *et al.* 2000; Ben-Zion *et al.* 2003; Zaliapin *et al.* 2003; Schorlemmer *et al.* 2005) studies. The fact that the deviations of the b-value reported here as artefacts coincide with the premonitory

b-value changes that are backed up by theoretical and laboratory studies emphasizes the need to exercise substantial care in observational studies focused on spatio-temporal variations of the magnitude-frequency distribution. Aftershock declustering should make the *b*-value estimation more robust, since it eliminates the spatio-temporal regions most affected by the short-term incompleteness—the spatial vicinities of main shocks during their immediate aftermath. However, our results in Section 5.4 indicate that many clustered events might be misidentified as background events due to location errors. This suggests that short-term incompleteness may well affect *b*-value estimates in declustered catalogues, in particular if the analysis includes small events ($m <$ 3.5 according to our analyses). One should also verify that the declustering itself does not affect the magnitude distribution, as may be the case for window-based methods that assume that the main shock is the largest event in the sequence.

7 ANALYSIS OF ALTERNATE CATALOGUES

In this section, we examine and compare four alternate catalogues of southern California—the relocated catalogue of Hauksson *et al.* (2013) that was studied in detail in the previous sections, the relocated catalogue of Richards-Dinger & Shearer (2000), and two ANSS (SCSN) subcatalogues (see Section 2 for catalogue descriptions). This analysis has two goals. First, we show in Section 7.1 that the artefacts of location errors and short-term incompleteness reported in the previous sections are not specific to the HYS catalogue. The same artefacts are also seen in the catalogue of Richards-Dinger & Shearer (2000) and the two ANSS subcatalogues. Next, we demonstrate how the reported relations between location errors and clustering can be used to order (rank) the four examined catalogues according to the overall event location accuracy.

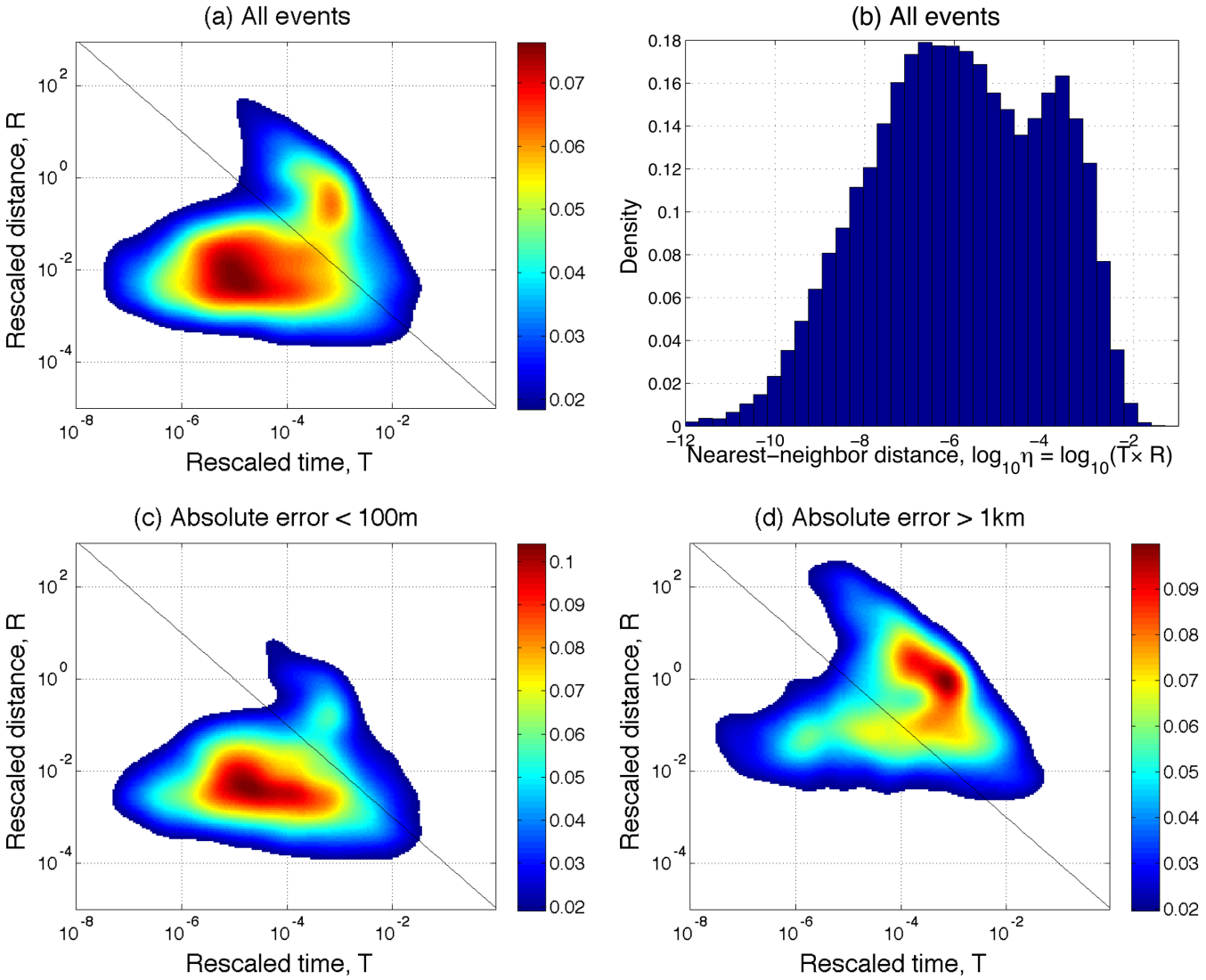

Figure 17. Cluster analysis and location artefacts in the relocated catalogue of Richards-Dinger & Shearer (2000). (a) The joint density of the rescaled time (T) and distance (R) to parent as defined in eq. (2). Notice clearly visible bimodal structure that separates cluster (below the diagonal) and background (above the diagonal) modes. (b) Distribution of the nearest-neighbour distance η. (c) The joint ($\log_{10}T$, $\log_{10}R$) distribution for events with location error below 100 m. (d) The joint ($\log_{10}T$, $\log_{10}R$) distribution for events with location error above 1 km. Notice the difference of the distributions in panels (c, d) that reflects the Artefact 1 (increased distance to parent), Artefact 2 (underestimated offspring productivity) and Artefact 3 (overestimated background rate). In panels (a),(c), and (d) the lowest 20 per cent of the density values are made transparent to better depict the distribution's support.

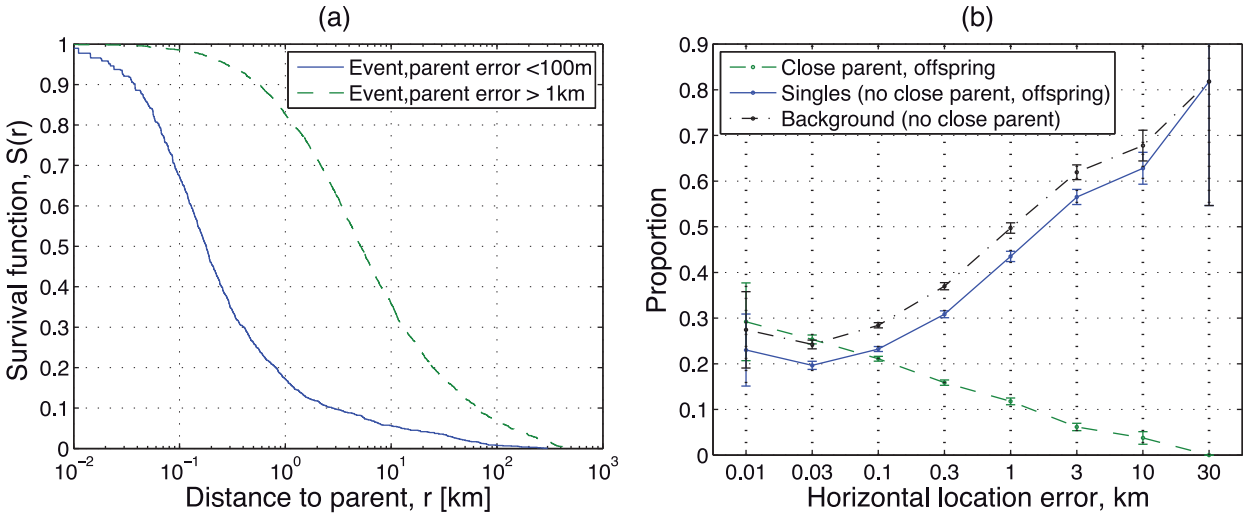

Figure 18. Location error artefacts in the relocated catalogue of Richards-Dinger and Shearer (2000). (a) Artefact 1 (increased distance to parent). Distribution of the distance to parent for parent–offspring pairs with location error below 100 m (solid blue) and above 1 km (dashed green). The distance increases to over an order of magnitude (green line is shifted to the right with respect to blue line) when location error increases. The y-axis shows the survival function $S(r) = \text{Prob}(\text{Distance} > r)$. (b) Artefacts 2 and 3 (underestimated offspring productivity and overestimated background rate). The proportions of internal cluster events (close parent and offspring), singles (no close parent, offspring) and background events (no close parent) as a function of horizontal location error.

7.1 Artefacts in the catalogue of Richards-Dinger and Shearer

Fig. 17 shows the joint distribution of the rescaled time and space components (T, R) of the earthquake nearest-neighbour distance η (panel a) and the distribution of the $\log_{10}\eta$ (panel b) in the catalogue of Richards-Dinger & Shearer (2000). We notice a bimodal structure of the earthquake nearest-neighbour distance, closely resembling that of Fig. 3 for the catalogue of Hauksson *et al.* (2013). One apparent difference is a somewhat less prominent separation of the modes in the catalogue of Richards-Dinger and Shearer, which might reflect overall improvement of event locations in the HYS catalogue.

Next, we reproduce in the catalogue of Richards-Dinger and Shearer (RDS) the three main artefacts studied in Section 5 and collect the results in Figs 17 and 18. The effect of location error on the distance to parent (Artefact 1) is illustrated in Figs 17(c,d) and 18(a). Figs 19(c) and (d) compare the joint distribution of the rescaled time and space components of the nearest-neighbour earthquake distance for events with horizontal errors below 100 m (panel c) and above 1 km (panel d). Several observations are noteworthy: (i) The typical distance to parent increases from $R \approx 10^{-2}$ to $R \approx 10^{-1}$ as the location error increases (from panel c to panel d), (ii) The proportion of background events (those above the diagonal line) increases with error, while the proportion of clustered events (those below the diagonal line) decreases with error, (iii) The separation of the background and clustered modes is better for the low-error events. Fig. 18(a) compares the actual (not rescaled) distance to parent for events with small and large location errors. The median distance increases by more than an order of magnitude, from about 0.2 to 3 km, between the pairs of parents–offspring with small (<100 m) location errors and those with large (>1 km) location errors.

Fig. 18(b) illustrates Artefacts 2 and 3—inflated proportion of background events and singles and deflated proportion of clustered events in high-error areas. The figure shows the proportion of interval cluster events (green dashed line), singles (solid blue line), and background events—those with distant parent (black dash–dotted line) as a function of horizontal location error that increases from 0.01 to 30 km. The proportions of singles and background events significantly increase (from about 0.25 to 0.8) with the error; while the proportion of cluster events decreases from 0.3 to zero.

All these observations closely resemble those reported for the HYS catalogue and support our statement that the discussed artefacts are not caused by a particular (re)location method.

7.2 Comparative analysis of four alternate catalogues

Fig. 19 compares the joint distributions of the rescaled time and distance components (T, R) of the earthquake nearest-neighbour distance in the four examined catalogues. In each catalogue the analysis is done for two complementary groups of events: (A) both event and its parent have magnitude $m \leq 3.5$ (panels a, c, e and g), and (B) event or its parent has magnitude above 3.5 (panels b, d, f and h). The distributions in group B are fairly similar to each other in terms of their overall location in the (T, R) plane. The only noticeable difference is a much higher proportion of background events in the ANSS-1 catalogue. These two observations are consistent with our earlier conclusions that (i) the location uncertainties barely affect the clustering of the parent–offspring pairs with $m > 3.5$ (Section 5.2), and (ii) the decreased location quality leads to increased background rates (Section 5.3). On the other hand, the distributions in group A show significant differences, in agreement with our earlier conclusion that events with $m < 3.5$ are most severely affected by the location errors (Section 5.2).

Specifically, we observe the following systematic changes as we move from HYS (panel a) to RDS (panel c) to ANSS-2 (panel e), and finally to ANSS-1 (panel g): (i) the distribution mode—position of the most numerous (red) population of events—is shifting towards higher rescaled times (T) and distances (R) to parent, (ii) the separation of the clustered (below the diagonal line) and background (above the diagonal line) modes becomes less prominent and (iii) the proportion of background events increases, while the proportion of clustered events decreases. These observations are consistent with the statement that the overall location quality in the examined

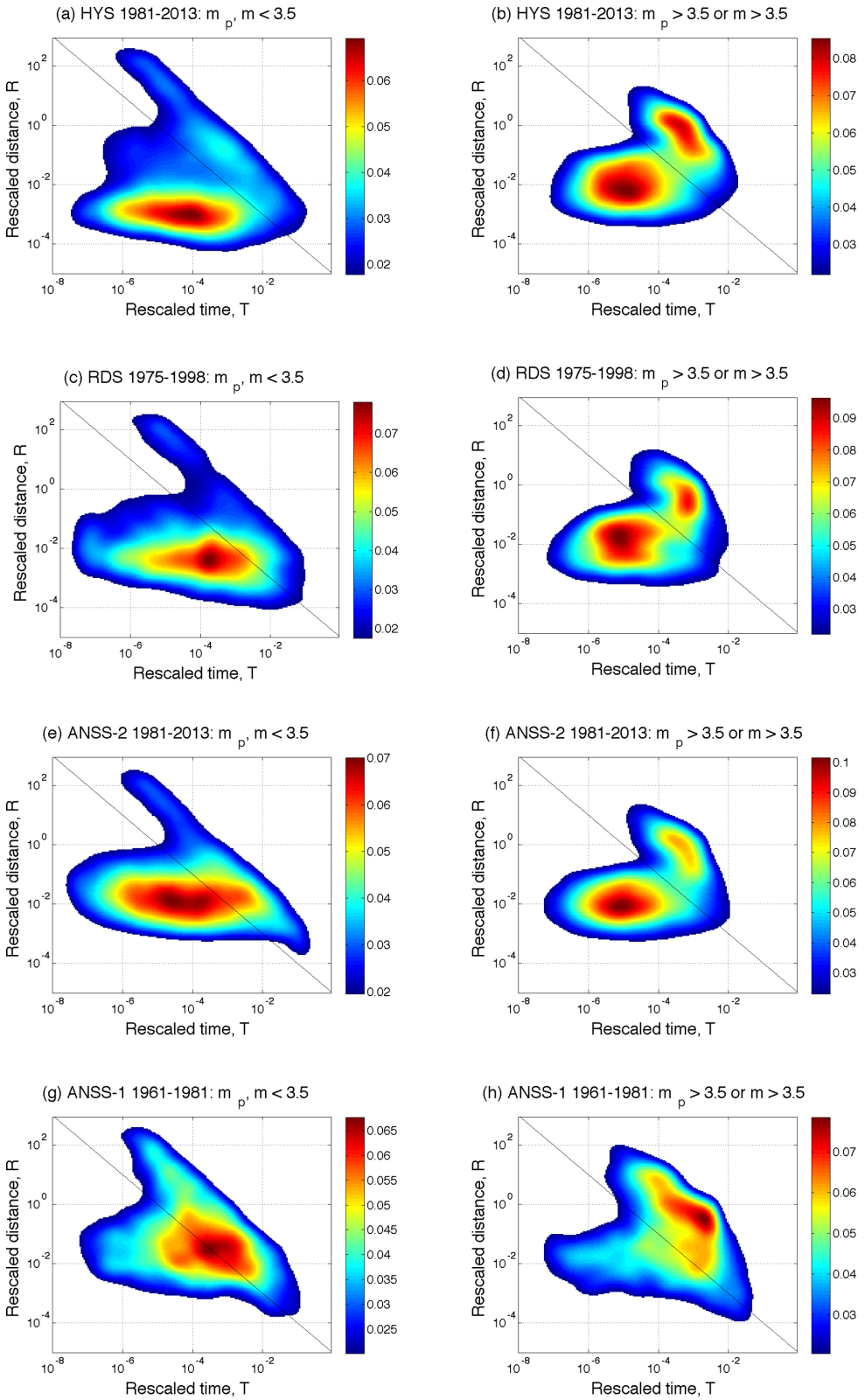

Figure 19. Comparative analysis of the location error effects on earthquake clustering in the four catalogues of southern California (HYS, RDS, ANSS-1 and ANSS-2). Each panel shows the joint density of the rescaled time (T) and distance (R) to parent as defined in eq. (2). (a, c, e and g) Parent–offspring pairs with magnitudes of both events below 3.5. (b, d, f and h) Parent–offspring pairs with magnitude of at least one event above 3.5. (a and b) HYS: Catalogue of Hauksson *et al.* (2013). (c and d) RDS: Catalogue of Richards-Dinger & Shearer (2000). (e and f) ANSS-2: ANSS catalogue during 1981–2013. (g and h) ANSS-1: ANSS catalogue during 1961–1981. The high-magnitude distributions (panels b, d, f and h) are similar, while the low-magnitude distributions (panels a, c, e and g) demonstrate systematic shift to higher times T and distances R as we go from HYS to RDS to ANSS-2, to ANSS-1. In each panel, the lowest 20 per cent of density values are made transparent to better depict the distribution's support.

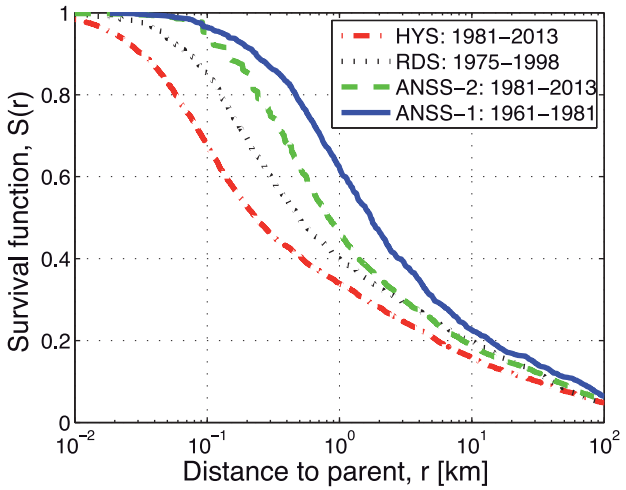

Figure 20. Comparative analysis of the distance to parent within the San Jacinto region in the four catalogues of southern California (HYS, RDS, ANSS-1 and ANSS-2). Here we only consider parent–offspring pairs with magnitudes below 3.5. The distance uniformly increases with the decreasing event location quality in the following catalogue order: HYS (best quality, shortest distance), RDS, ANSS-2, ANSS-1 (worst quality, longest distance). The y-axis shows the survival function $S(r) = \text{Prob}(\text{Distance} > r)$.

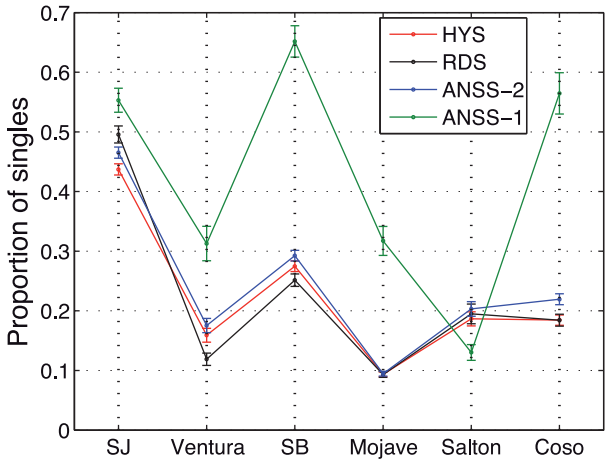

Figure 21. Comparative analysis of the regional proportion of singles in the four catalogues of southern California (HYS, RDS, ANSS-1 and ANSS-2). SJ, San Jacinto; SB, San Bernardino. The proportions are very close in the HYS, RDS and ANSS-2 catalogues, despite significant differences in their quality and cluster characteristics, see Figs 19 and 20. The proportions in ANSS-1 catalogue (green line) differ significantly from the consensus figures, reflecting low quality of this catalogue.

catalogues decreases from HYS to RDS to ANSS-2, and to ANSS-1, which is a reasonable given the used event location techniques. This is supported by Fig. 20 which shows the distribution of the distance r to parent (in km) for the parent–offspring pairs with $m < 3.5$ in the San Jacinto fault zone for the four catalogues. The distribution uniformly shifts from lower to larger distances as we go from HYS to RDS to ANSS-2 to ANSS-1. The increase of the distribution median between each pair of catalogues is about one third of a magnitude unit; accordingly, the difference between the median of the best quality catalogue (HYS) and the worst quality catalogue (ANSS-1) is about an order of magnitude—from 0.2 to 2 km.

Interestingly, the prominent differences in the parent–offspring distance distribution of the alternate catalogues reported in Figs 19 and 20 do not necessarily lead to significant changes in other cluster statistics. For example, Fig. 21 shows the proportion of singles in six individual regions of southern California—San Jacinto, Ventura, San Bernardino, Mojave, Salton Sea, and Coso—for the four catalogues. The proportions change significantly from region to region; yet remain very consistent among the HYS, RDS and ANSS-2 catalogues. However, the ANSS-1 catalogue's proportions significantly deviate from these consensus numbers, emphasizing the low quality of the location information and suggesting that this catalogue should not be used for cluster analysis of seismicity in southern California. The consistency of the proportions of singles and other cluster statistics among the alternate catalogues, despite the differences in distance-to-parent distribution, is consistent with the results of Zaliapin & Ben-Zion (2013a) who demonstrated that large (up to 40 per cent) errors in parent identification have only minor effect on errors in cluster identification (below 10 per cent) in synthetic ETAS catalogues.

Finally, we examine the effects of short-term incompleteness in the four alternate catalogues of southern California. Fig. 22 compares the magnitude distributions of events with different rescaled times-to-parent (panels a, c, e and g). In all examined catalogues the b-value of close offspring (blue) is visibly lower than that for distant offspring (red), despite the differences in the completeness and magnitude resolution among the catalogues. Panels (b), (d), (f) and (h) show how the average magnitude decreases with time after parent, which corresponds to gradual increase of the b-value. In all considered cases, the average magnitude decreases (and the b-value increases) within the first 10 d after the parent and then seems to stabilize. It appears that the decrease takes longer in the ANSS-1 catalogue, although, as discussed above, this low-quality catalogue is prone to additional artefacts with respect to the other examined catalogues; hence the observations not confirmed by the better-quality catalogues probably should be ignored.

The results of this section provide a rough assessment of the size of fluctuations one might expect as a result of varying location quality.

8 DISCUSSION

Improvements in seismic network quality and related development of earthquake relocation techniques during the last decades led to availability of high-quality regional catalogues with unprecedented detail on recorded seismicity. Examples include the catalogues of Hauksson *et al.* (2012) and Richards-Dinger & Shearer (2000) for southern California, Seeber *et al.* (2000) for North Anatolian fault in Turkey, Thurber *et al.* (2006) for Parkfield area in northern California and Waldhauser & Shaff (2008) for northern California. These catalogues allow the community to address problems that are new in scope and more detailed compared to those solved using the routine global catalogues like NEIC PDE (PDE 2015) or ISC (ISC 1995). Such new problems include the existence, structure and dynamics of small-to-medium magnitude clusters (including but not limited to the foreshock/aftershocks sequences and swarms) with relation to heat flow, non-volcanic tremors, slow earthquakes, seismic cycle, and extraterrestrial loadings, as well as dynamics of induced seismicity and discriminating between tectonic and human-induced earthquakes. However, the ability of probing the Earth at fine spatio-temporal scales and addressing important physical and socio-economic problems involves using small events. This requires dealing with a number of catalogue uncertainties and related statistical pitfalls. In the new reality of working with groups of events

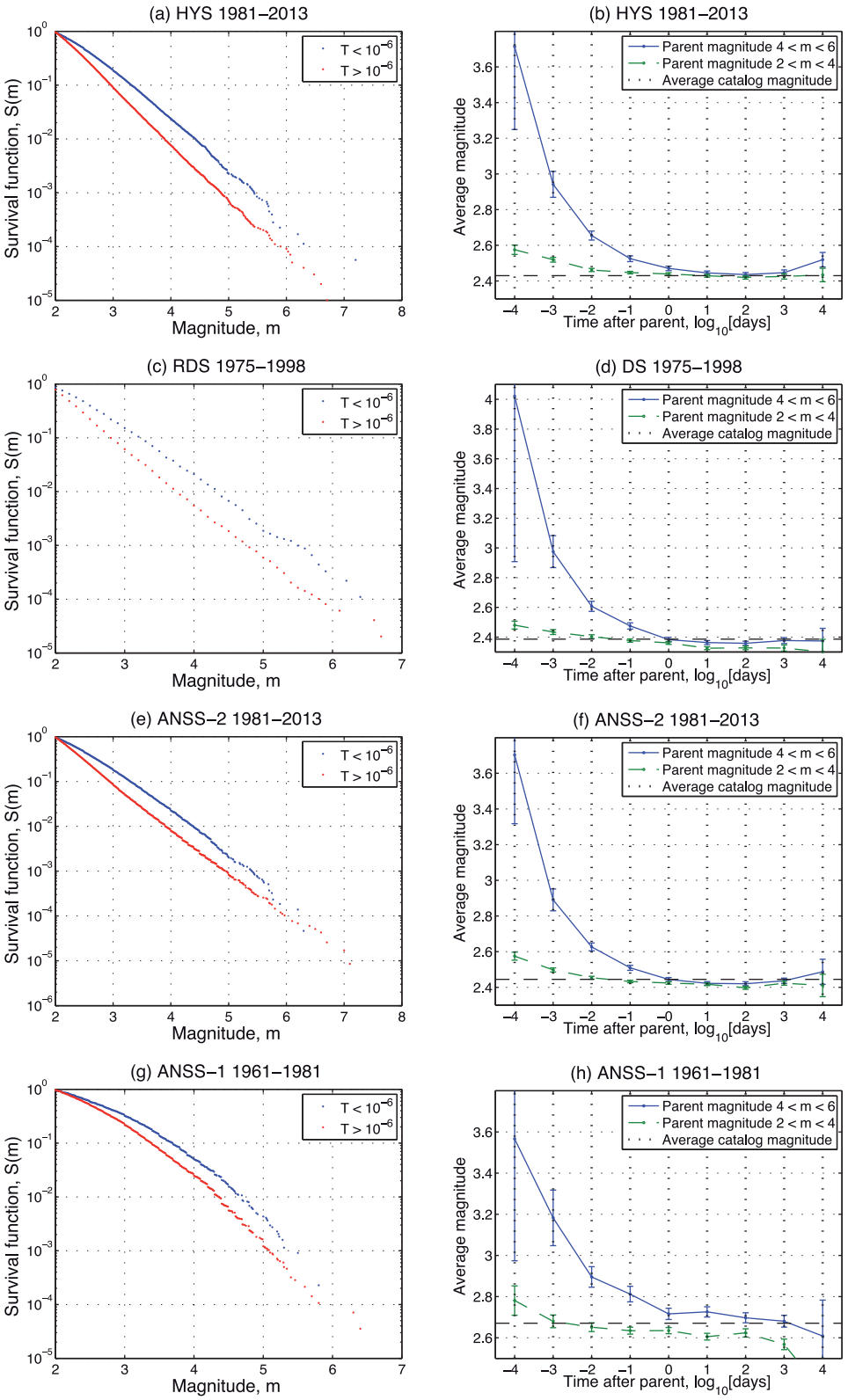

Figure 22. Comparative analysis of the short-term incompleteness effect on earthquake magnitude distribution in the four catalogues of southern California. (a, c, e and g) The tail of the magnitude distribution (survival function) for two group of events depending on the rescaled time to parent: $T < 10^{-6}$ (blue) or $T > 10^{-6}$ (red). The *b*-value is visibly lower for the close offspring (blue), in all examined catalogues. (b, d, f and h) The average magnitude (a reciprocal proxy for the *b*-value) decreases with time after the parent. Solid blue line—offspring of parents with magnitude $4 < m < 6$. Dashed green line—offspring of parents with magnitude $2 \leq m < 4$. Most visible decrease happens within the first 10 d after the parent. Dashed black line shows the average catalogue magnitude. (a and b) HYS: Catalogue of Hauksson *et al.* (2013). (c and d) RDS: Catalogue of Richards-Dinger & Shearer (2000). (e and f) ANSS-2: ANSS catalogue during 1981–2013. (g and h) ANSS-1: ANSS catalogue during 1961–1981.

that are small both in size and number, statistical uncertainties may (and do) significantly bias the results and spoil their interpretations.

In this study, we document and quantify effects of basic catalogue imprecision on essential earthquake cluster statistics. Specifically, we demonstrate that two types of well-known catalogue uncertainties—location errors and short-term incompleteness—affect estimation of earthquake background rates, triggering productivity (Section 5.2), and *b*-value (Section 6.1). Related artefacts may include biased estimations of foreshock and aftershock spatial decay rates and detecting apparent magnitude dependencies. We show that the reported artefacts are robustly observed in several alternative catalogues of southern California, with and without relocation (Section 7), and hence are not caused by a particular event location procedure. We also present evidence that the existence of the reported artefacts should be independent of a particular cluster technique (as long as the technique is based on parent–offspring assignment); although their magnitude may vary across alternative techniques, which is an interesting topic for a further study.

Our main observations can be summarized as follows. The seismicity of southern California is partitioned into a cascade of clusters at different spatial, temporal and energy scales, from the rare yet eye-catching aftershock sequences of the largest regional events (e.g. Kern-County of 1952; Landers of 1992 or El-Mayor Cucapah of 2010, etc.) to tiny yet numerous groups of small-magnitude events whose number keeps increasing with the observation precision. In addition, there is background seismicity that consists of statistically unrelated events in the entire seismogenic volume. The small clusters have rather fine structure, with events occurring at distances comparable or below the event location precision, and magnitudes below the average completeness regional values. Accordingly, the estimated structure of such clusters is easily perturbed by the existing catalogue errors, which results in many events 'losing' their true clusters. Since the spatio-temporal volume occupied by clusters is but a small fraction of the total seismogenic volume, most of the lost events are misidentified as a part of the background seismicity. This leads to overestimated background rates and underestimated clustering—measured by the probability of having a close parent or offspring (Fig. 11), proportion of singles and internal cluster events (Figs 12 and 18b), number of offspring per event (Fig. 13), or any other suitable measure. In addition, the estimated distance to the parent is generally increases with increasing location error (Figs 9, 10, 17 and 18b), which reflects the fact that we are losing the fine cluster structure and replacing it with artificial long-distance connections that correspond to blurred clusters.

In addition to errors of reconstructing cluster structure, there are effects of short-term incompleteness, which are prominently seen for up to 10 d after a parent of any magnitude (Figs 15, 16 and 22). The incompleteness, which affects small-magnitude events more than large-magnitude ones, deflates estimation of the *b*-value and hence can be confused with *magnitude dependence*—change of the magnitude distribution before and after a particular event or in relation to a specific groups of events (e.g. cluster vs. background). We also note that the reported incompleteness has rather long time duration of about 10 d (Fig. 22)—longer than that expected from the known instrumental deficiencies. Understanding the reason for this effect requires further focused study.

In this study, we focused on cluster statistics related to identification of the parent–offspring pairs (distance-to-parent, number of offspring per event, etc.), as opposed to general cluster statistics that refer to whole aftershock/foreshock series and other clusters that consist of events of multiple generations (offspring of offspring). Zaliapin & Ben-Zion (2013b) demonstrated that relatively large errors in individual parent–offspring identification (that can be as high as 40 per cent in the ETAS catalogues examined in that study) in general do not propagate to the global cluster statistics (that stay consistently below 10 per cent). This explains the consistency of general cluster statistics (e.g. proportion of singles, Fig. 21) among alternative catalogues, despite significant differences of the respective parent–offspring statistics (Figs 19 and 20). These observations suggest that considering general regional cluster statistics may somewhat reduce the effects of catalogue uncertainties, although further research is needed in order to make specific practical recommendations. Another avenue of overcoming the reported artefacts is to look for multiple possible parents for each event, which will lead to a more detailed, and hopefully more robust, description of the earthquake-to-earthquake connections. This approach is currently under investigation.

Our results suggest that the reported artefacts mostly affect events with magnitudes $m < 3.5$. This is primarily related to the fact that the location errors for these events are comparable or larger than their rupture length, as well as the rupture length of their parents (Fig. 14). This observation provides strong evidence that the artefacts studied here are not related to our particular method of identifying offspring–parents and should affect any alternative cluster analysis.

ACKNOWLEDGEMENTS

We are grateful to Egill Hauksson for sharing with us the details of the HYS catalog construction. The manuscript benefited from constructive comments of two anonymous referees. The study was supported by the Southern California Earthquake Center (based on NSF Cooperative Agreement EAR-0529922 and USGS Cooperative Agreement 07HQAC0008).

REFERENCES

Aki, K., 1965. Maximum likelihood estimate of *b* in the formula log $N = a - bM$ and its confidence limits, *Bull. Earthq. Res. Inst., Tokyo Univ.,* **43,** 237–239.

Aki, K., 1981. A probabilistic synthesis of precursory phenomena, in *Earthquake Prediction,* eds Simpson, D.W. & Richards, P.G., American Geophysical Union, doi:10.1029/ME004p0566.

ANSS: Advanced National Seismic System Composite Catalog, 2015. Available at: http://www.quake.geo.berkeley.edu/anss/.

Baiesi, M. & Paczuski, M., 2004. Scale-free networks of earthquakes and aftershocks, *Phys. Rev. E,* **69**(6), 066106, doi:10.1103/PhysRevE.69.066106.

Ben-Zion, Y., 2008. Collective behavior of earthquakes and faults: continuum-discrete transitions, evolutionary changes and corresponding dynamic regimes, *Rev. Geophysics,* **46,** RG4006, doi:10.1029/2008RG000260.

Ben-Zion, Y., Eneva, M. & Liu, Y., 2003. Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity, *J. geophys. Res.: Solid Earth,* **108,** 2307, doi:10.1029/2002JB002121(B6).

Brodsky, E.E. & Lajoie, L.J., 2013. Anthropogenic seismicity rates and operational parameters at the Salton Sea Geothermal Field, *Science,* **341**(6145), 543–546.

Chen, X. & Shearer, P.M., 2013. California foreshock sequences suggest aseismic triggering process, *Geophys. Res. Lett.,* **40**(11), 2602–2607.

Chen, X., Shearer, P.M. & Abercrombie, R.E., 2012. Spatial migration of earthquakes within seismic clusters in Southern California: Evidence for fluid diffusion, *J. geophys. Res.: Solid Earth,* **117,** B04301, doi:10.1029/2011JB008973.

Eaton, D.W., Davidsen, J., Pedersen, P.K. & Boroumand, N., 2014. Breakdown of the Gutenberg-Richter relation for microearthquakes induced by hydraulic fracturing: influence of stratabound fractures, *Geophys. Prospect.,* **62**(4), 806–818.

Eberhart-Phillips, D., 1990. Three-dimensional P and S velocity structure in the Coalinga region, California, *J. geophy. Res.: Solid Earth (1978–2012),* **95**(B10), 15 343–15 363.

Ellsworth, W.L., 2013. Injection-induced earthquakes, *Science,* **341**(6142), doi:10.1126/science.1225942.

Enescu, B., Hainzl, S. & Ben-Zion, Y., 2009. Correlations of seismicity patterns in Southern California with surface heat flow data, *Bull. seism, Soc. Am.,* **99,** 3114–3123.

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

Eneva, M. & Ben-Zion, Y., 1997. Techniques and parameters to analyze seismicity patterns associated with large earthquakes, *J. geophys. Res.: Solid Earth (1978–2012),* **102**(B8), 17 785–17 795.

Gabrielov, A., Zaliapin, I., Newman, W.I. & Keilis-Borok, V.I., 2000. Colliding cascades model for earthquake prediction, *Geophys. J. Int.,* **143**(2), 427–437.

Gu, C., Schumann, A.Y., Baiesi, M. & Davidsen, J., 2013. Triggering cascades and statistical properties of aftershocks, *J. geophys. Res.: Solid Earth,* **118**(8), 4278–4295.

Habermann, R.E., 1981. Precursory seismicity patterns: stalking the mature seismic GAP, in *Earthquake Prediction,* eds Simpson, D.W. & Richards, P.G., American Geophysical Union, doi:10.1029/ME004p0029.

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.: Solid Earth,* **118,** 1188–1191.

Hainzl, S., Moradpour, J. & Davidsen, J., 2014. Static stress triggering explains the empirical aftershock distance decay, *Geophys. Res. Lett.,* **41**(24), 8818–8824.

Hanks, T.C., 1979. *b* values and $\omega^{-\gamma}$ seismic source models: implications for tectonic stress variations along active crustal fault zones and the estimation of high-frequency strong ground motion, *J. geophys. Res.,* **84,** 2235–2242.

Hauksson, E., 2000. Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California, *J. geophys. Res.: Solid Earth (1978–2012),* **105**(B6), 13 875–13 903.

Hauksson, E., 2011. Crustal geophysics and seismicity in southern California, *Geophys. J. Int.,* **186**(1), 82–98.

Hauksson, E., Yang, W. & Shearer, P.M., 2012. Waveform relocated earthquake catalog for Southern California (1981 to June 2011), *Bull. seismol. Soc. Am.,* **102**(5), 2239–2244.

Hauksson, E., Yang, W. & Shearer, P.M., 2013, Waveform relocated earthquake catalog for Southern California, 1981–2013, Available at: http://www.data.scec.org/research-tools/downloads.html.

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

Hirata, T., 1989. A correlation between the *b* value and the fractal dimension of earthquakes, *J. geophys. Res.: Solid Earth (1978–2012),* **94**(B6), 7507–7514.

Holtkamp, S.G., Pritchard, M.E. & Lohman, R.B., 2011. Earthquake swarms in South America, *Geophys. J. Int.,* **187**(1), 128–146.

Hutton, L.K., Woessner, J. & Hauksson, E., 2010. Seventy-seven years (1932–2009) of earthquake monitoring in Southern California, *Bull. seismol. Soc. Am.,* **100**(2), 423–446.

Imoto, M., 1991. Changes in the magnitude—frequency *b*-value prior to large (M > 6.0) earthquakes in Japan, *Tectonophys,* **193**(4), 311–325.

ISC: International Seismological Centre, 2012. *On-line Bulletin,* http://www.isc.ac.uk, Internatl. Seis. Cent., Thatcham, United Kingdom.

Kagan, Y.Y., 2003. Accuracy of modern global earthquake catalogs, *Phys. Earth planet. Inter.,* **135**(2), 173–209.

Kagan, Y.Y., 2004. Short-term properties of earthquake catalogs and models of earthquake source, *Bull. seismol. Soc. Am.,* **94**(4), 1207–1228.

Klein, F.W., 2002. User's guide to HYPOINVERSE-2000: a Fortran program to solve for earthquake locations and magnitudes (p. 123), US Geological Survey.

Kuge, K., 1992. Systematic difference in the ISC body-wave magnitude-seismic moment relationship between intermediate and deep earthquakes, *Bull. seismol. Soc. Am.,* **82**(2), 819–835.

Lin, G., Shearer, P.M. & Hauksson, E., 2007. Applying a three-dimensional velocity model, waveform cross correlation, and cluster analysis to locate southern California seismicity from 1981 to 2005, *J. geophys. Res.: Solid Earth (1978–2012),* **112,** B12309, doi:10.1029/2007JB004986.

Lolli, B. & Gasperini, P., 2006. Comparing different models of aftershock rate decay: the role of catalog incompleteness in the first times after main shock, *Tectonophys,* **423**(1), 43–59.

Main, I.G., Meredith, P.G. & Jones, C., 1989. A reinterpretation of the precursory seismic *b*-value anomaly from fracture mechanics, *Geophys. J. Int.,* **96**(1), 131–138.

Main, I.G., Meredith, P.G. & Sammonds, P.R., 1992. Temporal variations in seismic event rate and *b*-values from stress corrosion constitutive laws, *Tectonophys,* **211**(1), 233–246.

Matoza, R.S., Shearer, P.M., Lin, G., Wolfe, C.J. & Okubo, P.G., 2013. Systematic relocation of seismicity on Hawaii Island from 1992 to 2009 using waveform cross correlation and cluster analysis, *J. geophys. Res.: Solid Earth,* **118**(5), 2275–2288.

Mogi, K., 1981. Seismicity in Western Japan and long-term earthquake forecasting, in *Earthquake Prediction,* eds. Simpson, D.W. & Richards, P.G., American Geophysical Union, doi:10.1029/ME004p0043.

Moradpour, J., Hainzl, S. & Davidsen, J., 2014. Nontrivial decay of aftershock density with distance in Southern California, *J. geophys. Res. Solid Earth,* **119,** 5518–5535.

Narteau, C., Shebalin, P. & Holschneider, M., 2002. Temporal limits of the power law aftershock decay rate, *J. geophys. Res.: Solid Earth (1978–2012),* **107**(B12), 2359, doi:10.1029/2002JB001868.

PDE: Preliminary Determination of Epicenters, 2015. ftp://hazards.cr.usgs.gov/NEICPDE.

Peng, Z., Vidale, J.E., Ishii, M. & Helmstetter, A., 2007. Seismicity rate immediately before and after main shock rupture from high-frequency waveforms in Japan, *J. geophys. Res.: Solid Earth,* **112,** B03306, doi:10.1029/2006JB004386.

Richards-Dinger, K.B. & Shearer, P.M., 2000. Earthquake locations in southern California obtained using source-specific station terms, *J. geophys. Res.: Solid Earth (1978–2012),* **105**(B5), 10 939–10 960.

Rikitake, T., 1975. Earthquake precursors, *Bull. seismol. Soc. Am.,* **65**(5), 1133–1162.

Röhm, A.H.E., Trampert, J., Paulssen, H. & Snieder, R.K., 1999. Bias in reported seismic arrival times deduced from the ISC Bulletin, *Geophys. J. Int.,* **137**(1), 163–174.

Schorlemmer, D., Wiemer, S. & Wyss, M., 2005. Variations in earthquake-size distribution across different stress regimes, *Nature,* **437**(7058), 539–542.

Seeber, L., Armbruster, J.G., Ozer, N., Aktar, M., Baris, S., Okaya, D., Ben-Zion, Y. & Field, E., 2000. The 1999 earthquake sequence along the North Anatolia transform at the juncture between the two main ruptures, in *The 1999 Izmit and Duzce Earthquakes: Preliminary Results,* pp. 209–223, eds Barka, A., Kazaci, O., Akyuz, S. & Altunel, E., Istanbul Technical University.

Shearer, P.M., 2012. Self-similar earthquake triggering, Bath's law, and foreshock/aftershock magnitudes: simulations, theory, and results for southern California, *J. geophys. Res.,* **117,** B06310, doi:10.1029/2011JB008957.

Smith, W.D., 1981. The *b*-value as an earthquake precursor, *Nature,* **289,** 136–139.

Smith, W.D., 1986. Evidence for precursory changes in the frequency-magnitude *b*-value, *Geophys. J. Int.,* **86**(3), 815–838.

Storchak, D.A., Bird, A.L. & Adams, R.D., 2000. Discrepancies in earthquake location between ISC and other agencies, *J. Seismol.,* **4**(3), 321–331.

Thurber, C.H., 1983. Earthquake locations and three-dimensional crustal structure in the Coyote Lake area, central California, *J. geophys. Res.: Solid Earth (1978–2012),* **88**(B10), 8226–8236.

Thurber, C.H., 1993. Local earthquake tomography: velocities and Vp/Vs-theory, in *Seismic Tomography: Theory and Practice,* pp. 563–583, eds Iyer, H.M. & Hirahara, K., Chapman and Hall.

Thurber, C., Zhang, H., Waldhauser, F., Hardebeck, J., Michael, A. & Eberhart-Phillips, D., 2006. Three-dimensional compressional wavespeed model, earthquake relocations, and focal mechanisms for the Parkfield, California, region, *Bull. seismol. Soc. Am.,* **96**(4B), S38–S49.

Tinti, S. & Mulargia, F., 1987. Confidence intervals of *b* values for grouped magnitudes, *Bull. seismol. Soc. Am.,* **77**(6), 2125–2134.

Utsu, T., Ogata, Y. & Matsu'ura, R.S., 1995. The centenary of the Omori formula for a decay law of aftershock activity, *J. Phys. Earth,* **43**(1), 1–33.

Vidale, J.E. & Shearer, P.M., 2006. A survey of 71 earthquake bursts across southern California: exploring the role of pore fluid pressure fluctuations and aseismic slip as drivers, *J. geophys. Res.,* **111**(B5), B05312, doi:10.1029/2005JB004034.

Vidale, J.E., Boyle, K.L. & Shearer, P.M., 2006. Crustal earthquake bursts in California and Japan: their patterns and relation to volcanoes, *Geophys. Res. Lett.,* **33**(20), L20313, doi:10.1029/2006GL027723.

Waldhauser, F. & Schaff, D.P., 2008. Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods, *J. geophys. Res.: Solid Earth,* **113,** B08311, doi:10.1029/2007JB005479.

Wiemer, S. & Katsumata, K., 1999. Spatial variability of seismicity parameters in aftershock zones, *J. geophys. Res.: Solid Earth (1978–2012),* **104**(B6), 13 135–13 151.

Zaliapin, I. & Ben-Zion, Y., 2011. Asymmetric distribution of early aftershocks on large faults in California, *Geophys. J. Int.,* **185,** 1288–1304.

Zaliapin, I. & Ben-Zion, Y., 2013a. Earthquake clusters in southern California. I: Identification and stability, *J. geophys. Res.,* **118**(6), 2847–2864.

Zaliapin, I. & Ben-Zion, Y., 2013b. Earthquake clusters in southern California. II: Classification and relation to physical properties of the crust, *J. geophys. Res.,* **118**(6), 2865–2877.

Zaliapin, I., Keilis-Borok, V. & Ghil, M., 2003. A Boolean delay equation model of colliding cascades. Part II: Prediction of critical transitions, *J. Stat. Phys.,* **111**(3–4), 839–861.

Zaliapin, I., Gabrielov, A., Wong, H. & Keilis-Borok, V., 2008. Clustering analysis of seismicity and aftershock identification, *Phys. Rev. Lett.,* **101,** 018501, doi:10.1103/PhysRevLett.101.018501.