AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2014GL060242

Key Points:
- The k-means cluster analysis indicates that the Sumatran fault has 16 segments
- Segment lengths and earthquake potentials range from 22 to 196 km and Mw 6.5–7.8
- Eight great central segments dominate hazard; it is lower in the N and S

Correspondence to:
P. W. Burton,
p.burton@uea.ac.uk

Citation:
Burton, P. W., and T. R. Hall (2014), Segmentation of the Sumatran fault, *Geophys. Res. Lett.*, *41*, 4149–4158, doi:10.1002/2014GL060242.

Received 28 APR 2014
Accepted 29 MAY 2014
Accepted article online 2 JUN 2014
Published online 17 JUN 2014

Segmentation of the Sumatran fault

Paul W. Burton[1] and Thomas R. Hall[2]

[1]School of Environmental Sciences, University of East Anglia, Norwich, UK, [2]National Oceanography Centre, University of Southampton, Southampton, UK

Abstract Segmentation of the Sumatran fault is discerned using an analytical approach in which a *k*-means algorithm partitions earthquakes into clusters of seismicity along the fault. Clusters are tessellated into segment zones from which segment lengths and maximum credible magnitude are estimated. Decreasing the depth of seismicity sampled from 70 to 60 to 50 km reduces interaction with deeper seismicity, and results from the *k*-means algorithm initially suggest that the fault has $K = 14$, 16, and 16 clusters, respectively. After inspection, it becomes clear that the optimum number of clusters is 16. The 16 cluster model developed into zones generates segment lengths ranging from 22 to 196 km and maximum earthquake potentials in the range of M_w 6.5–7.8. The Sumatran fault is dominated by eight great central segments distributed approximately symmetrically about Lake Maninjau. These central fault segments dominate the hazard, which is less in the far north because segments are shorter.

1. Introduction

A fundamental step in seismic hazard analysis since the seminal work of *Cornell* [1968] is the subdivision of seismicity into zones. Geological subdivision of faults into discrete segments is also well established through examination of geomorphological features, fault bends, and fault step overs [e.g., *De Joussineau and Aydin*, 2009]. A method proven effective in grouping earthquakes into clusters and zones, with an element of consistency and objectivity, is *k*-means [*Hartigan*, 1975; *Weatherill and Burton*, 2009]. We think that it is a natural extension of regional seismicity zonation to attempt fault segmentation using predominant analysis of seismicity. It is also to be expected that quasi-linear features, faults, particularly strike-slip faults, might be particularly amenable to seismological segmentation using *k*-means. There has been a substantial work in relation to great earthquakes in subduction zones, fueled by the 2011 Tohuku M_w 9.1 earthquake, and an implication for all subduction zones, when the seismic moment conservation principle is adopted and the maximum rupture length is realized, that M_w 9.0–9.7 is possible in all subduction zones [*Kagan and Jackson*, 2013]. At a lesser level, albeit the disaster of the 2008 Wenchuan M_w 7.9 earthquake in China, fault geometry and slip distribution have been modeled by *Shen et al.* [2009]. The Longmen Shan fault, which contained this earthquake, changes faulting character from thrust in the southwest to dextral strike slip in the northeast. Slip modeling indicates three high-slip concentrations. The rupture lengths supporting each high-slip pattern are termed subsegments, and the fault junctions as barriers "that rarely fail." There is also an ongoing debate concerning the Gutenberg–Richter linear frequency-magnitude recurrence model vis-à-vis the characteristic earthquake model; *Parsons and Geist* [2009] conclude that a preference cannot be made between such models on individual fault zones, there being longevity and completeness issues in earthquake catalogs. Later, *Parsons et al.* [2012] use the Nankai-Tokai subduction zone to aid the demonstration that Gutenberg–Richter constrained by convergence rates is a useful adjunct in forecasts when information is sparse. In a global analysis of seismicity that seeks spatial clustering of earthquakes as one objective, *Kagan and Jackson* [2000] deploy kernel smoothing of seismicity [e.g., *Woo*, 1996] and obtain long-term forecasts and spatial clustering from earthquake history, but not focused to individual faults. Within such a spectrum of studies, our attempt is different and is not focused on earthquake recurrence but on the partitioning of seismicity that denotes underlying fault segmentation. We seek segmentation, as indicated by *Bellier et al.* [1997] and *De Joussineau and Aydin* [2009], but not by their geological means. We wish to improve the modeling of the constituent makeup of the seismogenetic fault by developing the objective *k*-means partitioning of the seismicity itself.

The Sumatran fault (also Great Sumatran fault (GSF); Figure 1) is a 1800 km long geological structure accommodating strain associated with right-lateral shear due to oblique motion northeastward of the

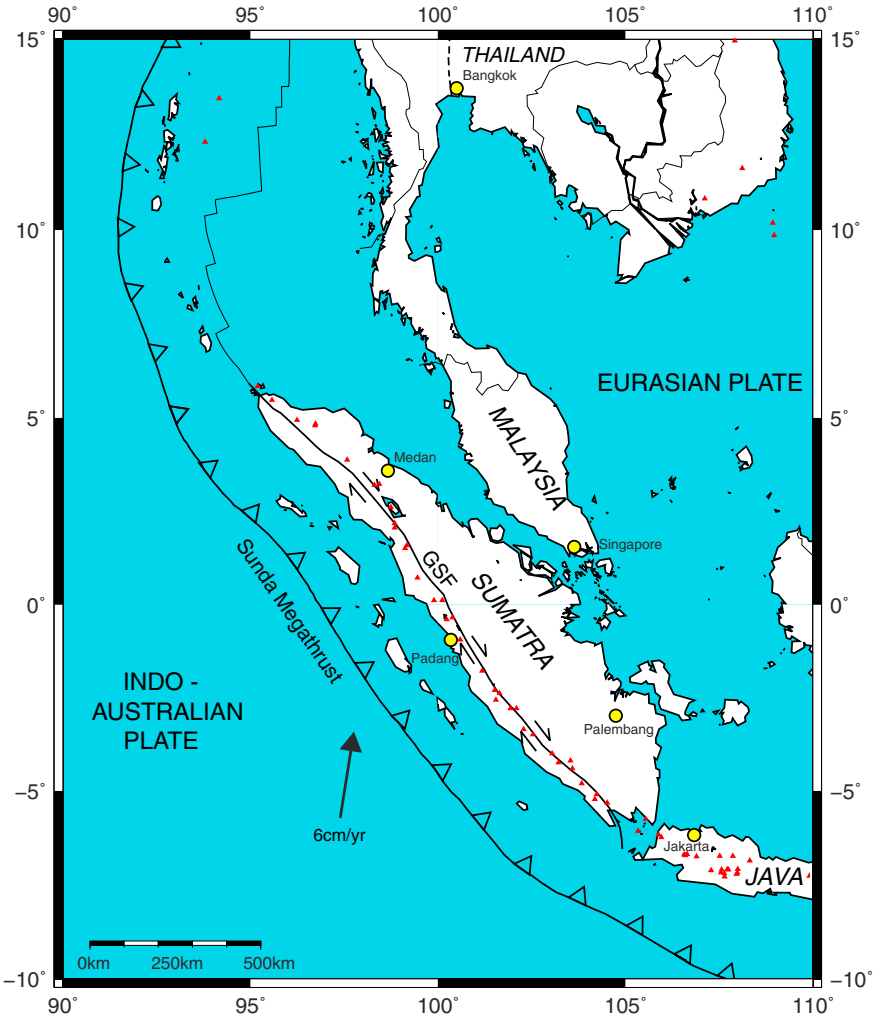

Figure 1. Tectonic and geographic location of the Sumatran fault (GSF). The red triangles mark the volcanoes.

Sumatran subduction zone. Seismic hazard is high [see e.g., *Petersen et al.*, 2004], and the GSF is a key hazard in the area with its seismicity being the largest factor in ground motion hazard for Sumatra. The fault has produced several large historical earthquakes, e.g., 1895 ~M_w 7.6, 1945 ~M_w 7.7, and in recent history Sungaipenuh, Jambi Province, 2009 M_w 6.6. Gaps in seismic activity have been remarked along the fault, and *Sorensen* [2008] suggests that the Sumatra-Andaman 2004 M_w 9.1 earthquake, to the north of the GSF, marked the start of a 200 year seismic cycle. As *Parsons et al.* [2012] succinctly point out, there is a "persistent and difficult issue in seismology" when matching the long-term recurrence rates of largest earthquakes with instrumental recordings of smaller earthquakes. We do not investigate recurrence rates as such but use observed seismicity to map subsurface segmentation. Our model is therefore dependent on earthquakes that have occurred in an incomplete past but not on their rate. Nevertheless, the resulting model accommodates earlier preinstrumental surficial earthquake ruptures and at this time is not contradicted by these in any way; indeed, these complement the model. Slip calculations for oblique subduction west of the GSF indicate trench parallel strain deficits, linkable with raised seismic hazard expectation in North Sumatra or explicable by invoking activity on the West Andaman fault [*Natawidjaja and Triyoso*, 2007] or strike slip within the subducting slab [*Ishii et al.*, 2013], a dilemma that suggests a need for better understanding of the GSF.

The segmented character of strike-slip faulting determines the size of earthquake ruptures [*De Joussineau and Aydin*, 2009]. But identifying fault segments is a difficult business. Primary methods used in identifying

segmented composition of faults include trench excavations and geomorphic expressions in the landscape. A degree of inconsistency between the models proposed for the same fault presents a problem for seismic hazard analysis. Currently, two segmentation models are available for the GSF. *Bellier et al.* [1997] include fault trenching evidence and suggest that the fault has 18 segments. The second model is *Sieh and Natawidjaja*'s [2000], which was built on evidence obtained from stereographic aerial photographs and topographic maps, largely seeking geomorphic expressions in the landscape. They suggest a fault model of 19 segments.

This variety of evidence, observations, and views on the GSF shapes our aims. We shall assess the effectiveness of the *k*-means algorithm at partitioning GSF seismicity into clusters and, given that the *k*-means algorithm does produce realistic results, develop these into a segmentation model founded on the seismicity data. Having obtained a seismogenic GSF segmentation model, maximum credible earthquake magnitudes are estimated for each segment.

2. *K*-Means Methodology

A *k*-means algorithm is a hard-partitioning algorithm which divides N data points in p dimensions into K clusters [*Hartigan*, 1975]. The *k*-means identify groups found in data that correspond to some natural phenomenon; so clusters represent changing seismicity along a fault. Assuming that the different cluster seismicities are directly associated with segments, the number of clusters defined by *k*-means will be the number of segments composing the fault. Euclidean square distance from data point to cluster centroid is used as the metric to define clusters (Figure 2a). When investigating an a priori reasonable range of K, it becomes necessary to identify optimum K among this range. The *Krzanowski and Lai* [1988] (KL) index helps to achieve this, because an optimum K maximizes KL. This is accomplished by KL measuring successive differences between function $K^{2/p}S_K$, with incremental K, where S_K is the sum of the square variance within each cluster model. For all the cluster models ranging from 2 to K clusters, this function essentially remains constant for homogeneous data (which is best modeled trivially by $K=1$). In practice, data are not homogeneous and inherently contain an optimum number of clusters K. For heterogeneous data, with no optimum cluster arrangement, $K^{2/p}S_K$ will follow an exponential decay trend. When the data contain an optimum number of clusters G, *Krzanowski and Lai* [1988] show that there is a strong inflection in the exponential trend at $K=G$. This inflection point manifests as the maximum KL value for the data, which thus discerns an optimum K. *Weatherill and Burton* [2009] give a detailed description of the methodology, and they chose to run a 100-trial ensemble to obtain optimum centroids. For each K, the iterative process converged toward a centroid arrangement minimizing the total within-cluster sum of squares (TWCSS). The clustering error is

$$\mathrm{TWCSS} = \sum_{i=1}^{N} \sum_{k=1}^{K} I(x_i \in C_k) \|x_i - m_k\|^2 \quad (1)$$

where m_k is the mean position (centroid) of cluster C_k partitioned from the N data x_i. Function $I(X)$ is 1 if statement X is true and 0 otherwise. Out of the 100 trials, the one producing the lowest TWCSS provides the optimum arrangement of centroids at that K. Because KL is sensitive to small changes in centroid positioning arising from the initial centroid seeding, the algorithm is modified as follows. In one trial at a specific K, seed centroids change position to locations which minimize TWCSS. This is repeated 100 times. The centroids for the best of these hundred TWCSS, and its KL value, are retained. This entire process is repeated 100 times (10,000 trials), generating 100 independent centroid sets, each with attached KL. The mean KL of these 100 repetitions is calculated, but optimum centroids are returned by one repetition that maximizes KL. More than one competing optimum K value is identified. As the KL results follow a normal distribution, optimum K are defined as values satisfying $KL > 3 \times \sigma$, where σ is the standard deviation of the dispersion of the independent KL values. This was sufficient to identify outliers in the KL results. It is a logical step compatible with our aim to establish a degree of consistency and objectivity in the development of seismic source models.

3. Earthquake Catalog and Selecting Earthquakes Associated With the Sumatran Fault

The foundation of any seismicity and seismic hazard analysis is a high-quality earthquake catalog. The catalog used is *Petersen et al.*'s [2007], which is constructed from four well-known global catalogs

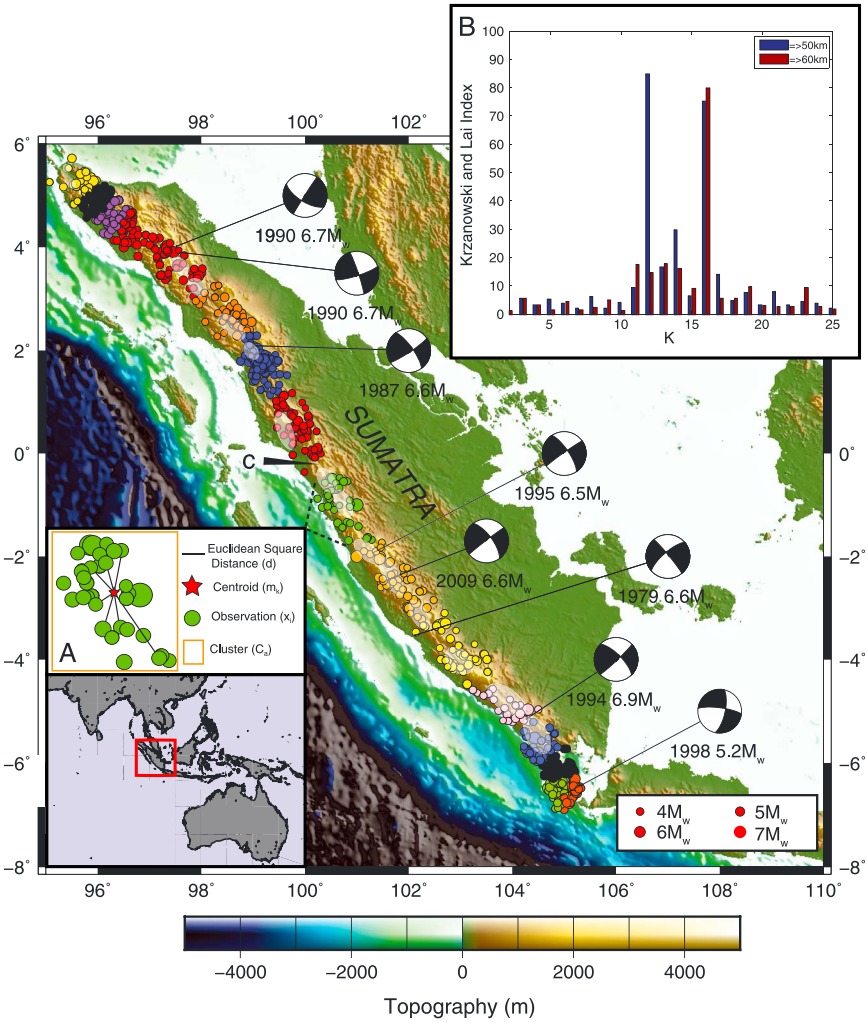

Figure 2. Earthquake clusters along the Sumatran fault. Each individual cluster is colored differently. Cluster distributions are identified using *k*-means analysis, applied to seismicity of depth 50 km or less (the strike-slip fault). This is the global optimum solution, with 16 clusters partitioning Sumatran fault seismicity. The white see-through ellipses depict known surface mapped ruptures [adapted from *Natawidjaja and Triyoso*, 2007]. The position *c* is the demarcator between the north and south clusters. (inset a) A schematic example which visualizes cluster or partition quality estimator as the total within-cluster sum of squares. (inset b) Explores *K*, the number of clusters fitted to Sumatran fault seismicity. Competing optimum *K* are identified by large KL indices [*Krzanowski and Lai*, 1988] in the range $K = 2–25$ for Sumatran fault seismicity consistent with two subsets with focal depth *h* in kilometer: $h \le 50$ km, $h \le 60$ km.

homogenized to moment magnitude M_w. The area covered is 17°S to 26°N by 88° to 122°E, from which the data for Sumatra have been extracted and augmented with events from the U.S. Geological Survey/National Earthquake Information Center preliminary determination of epicenters online catalog [http://neic.usgs.gov], extending the time span to August 2012.

The fault trace was taken as that documented by *Sieh and Natawidjaja* [2000, Plate 1], and it is visible for most of its length in the digital elevation model [*Jarvis et al.*, 2008]. Once the trace had been established, an area of ±1.0° was taken from either side of it and earthquakes collected within this area. Focal depth *h* was limited to 100 km. This was an initial generous swathe of seismicity for preliminary trials. Cross sections of seismicity show that substantial megathrust seismicity is included when focal depth is ≤100 km. Limiting the focal depth to ≤70 km removed the vast majority of this extraneous seismicity, the data sets still being of sufficient size for the *k*-means algorithm to be effective. A recommended criterion for cluster analysis is that the number of data is of order $N = K^2$. Some extraneous seismicity remained present in the southern

Geophysical Research Letters

10.1002/2014GL060242

Table 1. Cluster Parameters, Fault Segments, and Maximum Credible Earthquake Magnitudes Along the Sumatran Fault

Cluster Number (from N to S)	Cluster Centroid (degree latitude, degree longitude)	No. of Earthquakes	Segment *S* latitude from 5.5°N	Segment Length (***L*** km)	M_{cred} (M_w)	Comment[a]
S1	5.21°N 95.65°E	39	5.01°N	82	7.3	Seulimeum: Compatible in $h \le 50$ km and ≤60 km analyses. South end compatible P1.
S2	4.86°N 96.02°E	82	4.71°N	50	7.0	Aceh-north: Compatible in $h \le 50$ km and ≤60 km analyses.
S3	4.59°N 96.26°E	80	4.45°N	45	6.9	Aceh-south: Aceh fragmented. Compatible in $h \le 50$ km and ≤60 km analyses. South end compatible P1. If the throughgoing rupture of combined S2 + S3 hypothesized is possible, M_{cred} rises to M_w 7.4.
S4	4.34°N 96.62°E	58	3.99°N	83	7.3	Tripa-north
S5	3.69°N 97.45°E	51	3.16°N	142	7.6	Tripa-south: Compatible in $h \le 50$ km and ≤60 km analyses.
S6	2.71°N 98.52°E	66	2.23°N	136	7.6	Renun
S7	1.77°N 99.14°E	142	1.18°N	138	7.6	Toru: The $h \le 60$ km analysis divides S6 and S7 into three segments, which would reduce hazard. South end compatible P1.
S8	0.50°N 99.86°E	73	0.27°S	182	7.8	Barumun: Hinge point *c* at south end near Lake Maninjau. Barumun segment subsumes Barumun and all Sumpur and north part of Sianok of P1. Compatible in $h \le 50$ km and ≤60 km analyses.
S9	0.98°S 100.69E	43	1.71°S	194	7.8	Suliti: Suliti segment subsumes south part of Sianok and all Sumani and all Suliti and south end compatible P1. Compatible in $h \le 50$ km and ≤60 km analyses.
S10	2.49°S 101.79°E	47	3.09°S	191	7.8	Siulak: Siulak segment subsumes Siulak and all Dikit and north part Ketaun of P1.
S11	3.85°S 102.88°E	53	4.34°S	196	7.8	Manna: The $h \le 60$ km analysis divides S9, S10, and S11 into four segments, which would slightly reduce hazard. Manna segment subsumes south part Ketaun and all Musi and all Manna and south end compatible P1.
S12	4.91°S 103.86°E	33	5.29°S	141	7.6	Kumering: South end compatible P1
S13	5.69°S 104.48°E	50	5.81°S	73	7.2	Semangko: South end compatible P1
S14	6.09°S 104.89°E	66	6.33°S	62	7.1	Sunda 1
S15	6.56°S 104.85°E	43	6.52°S	22	6.5	Sunda 2
S16	6.55°S 105.17°E	41	6.80°S	33	6.7	Sunda 3: Sunda fragmented, also bifurcated. The $h \le 60$ km analysis combines S14, S15, and S16 into one segment.

[a]Nomenclature in the comments mainly follows *Sieh and Natawidjaja* [2000] = P1 and *Natawidjaja and Triyoso* [2007].

domain of these data when focal depth ≤70 km. A second depth profile with focal depth ≤50 km was judged void of megathrust seismicity. Residual extraneous seismicity perturbed the emerging GSF clusters. In the south, Mentawai fault clusters appeared touching just to the west of the Sumatran fault clusters. To the north, the Toru Fold thrust belt (~1.5°N, west of the GSF) perturbed results. The generous ±1.0° exploratory swathe of seismicity is reduced to ±0.25° across the GSF, and three subset catalogs adopted with focal depths $h \le 50$, 60, and 70 km, respectively, for final modeling.

4. Earthquake Clusters Segmenting the Sumatran Fault

Three subcatalogs explore the possible influence of earthquake depth on clustering and contain ample seismicity to meet the $N = K^2$ criterion: $h \le 50$ km, $N = 967$ earthquakes; $h \le 60$ km, $N = 1034$ earthquakes; and $h \le 70$ km, $N = 1117$ earthquakes. The spread of KL results for $K = 2$–25 is shown in Figure 2b. Analysis was applied through the range of $K = 2$–50. All previous literature suggests that there are less than 25 segments on the Sumatran fault zone. Although we include analysis of large numbers of clusters (26–50), we do not illustrate these as being of no real-Earth relevance. However, the results help to confirm our methodology, and KL values in the range of $K = 26$–50 were indeed found to be low compared to the $K = 2$–25 range. Rarely (once) a high-value KL arises at high K ($K = 46$, $h \le 60$ km subcatalog) when the seismicity becomes overparameterized with too many clusters enabling artificially low TWCSS. The overlapping optimum K range for the three subcatalogs is $K = 12$–16 with individual optimum K values: (a) $h \le 50$ km, $K = 16$; (b) $h \le 60$ km, $K = 16$; and (c) $h \le 70$ km, $K = 14$, with (c) slightly contaminated by deeper megathrust seismicity, whereas (a) and (b) are strongly mutually supportive. The KL optimum for $h \le 50$ km, $K = 12$ does not have corroborative support from $h \le 60$ km analysis. The optimum 16 clusters for focal depths ≤50 km are illustrated in Figure 2. Cluster information is given in Table 1.

5. Tessellation Into Segment Zones and Associated Maximum Credible Earthquakes

Having established optimum cluster centroids, a Voronoi tessellation (VT) algorithm develops the clusters into a geometrical zonation and segmentation model. VT decomposes a metric space determined by distances to a specified discrete set of objects in the space; in our case, the points are cluster centroids. Tessellation constructs a polygon around each cluster, and VT interpolates between centroids so that adjacent clusters share one side. Segment length is taken as synonymous with potential rupture length, and the corresponding maximum credible magnitude is estimated from empirical equations linking this length to magnitude. *Wells and Coppersmith* [1994] used a worldwide distribution of earthquakes to determine the regression of surface rupture length $\boldsymbol{L}$ against earthquake magnitude within M_w 4.3–8.1 and

$$M_{\text{cred}} = 5.02 + 1.19\ \text{Log}(\boldsymbol{L}) \tag{2}$$

allows the estimation of M_w from $\boldsymbol{L}$ for the expected maximum credible earthquake M_{cred}. These tessellated geometric zones model the distribution of seismogenic segments along the Sumatran fault and are illustrated in Figure 3. Traversing each zone is drawn the corresponding seismogenic fault segment, and M_{cred} is labeled. South latitudes terminating each segment, segment lengths, and corresponding M_{cred} are recorded in Table 1.

The results of Figures 2 and 3 show 16 segments with lengths 22–196 km and corresponding magnitudes for M_{cred} range M_w 6.5–7.8. Segment lengths span 45–182 km in the north and 22–196 km in the south. Magnitudes range M_w 6.9–7.8 and M_w 6.5–7.8, respectively. In all trials, the north and south sectors of the Sumatran fault demarcate just south of the equator, with eight segments in the north and eight in the south. Figures 2 and 3 show that long segments dominate the central Sumatran fault, whereas the northernmost and particularly the southernmost segments, are shorter and even appear fragmented in the Sunda Strait segments. Relative to these observations is *Bellier et al.*'s [1997] model with 18 segments ranging in length 45–130 km in the south, but being up to 200 km in the north. The forecast magnitudes spanning all fault segments from *Bellier et al.* [1997] and *Sieh and Natawidjaja* [2000] are M_w 6.7–7.3 and M_w 7.1–7.7, respectively. Both models suggest increased segment size in the northern part of the fault. Taking an average segment length of 111 km from Table 1, the average maximum magnitude for our 16 fault segments is a mean M_{cred} of M_w 7.45, or geometric mean $\boldsymbol{L} = 92$ km gives a mean M_{cred} of M_w 7.35. Although these general

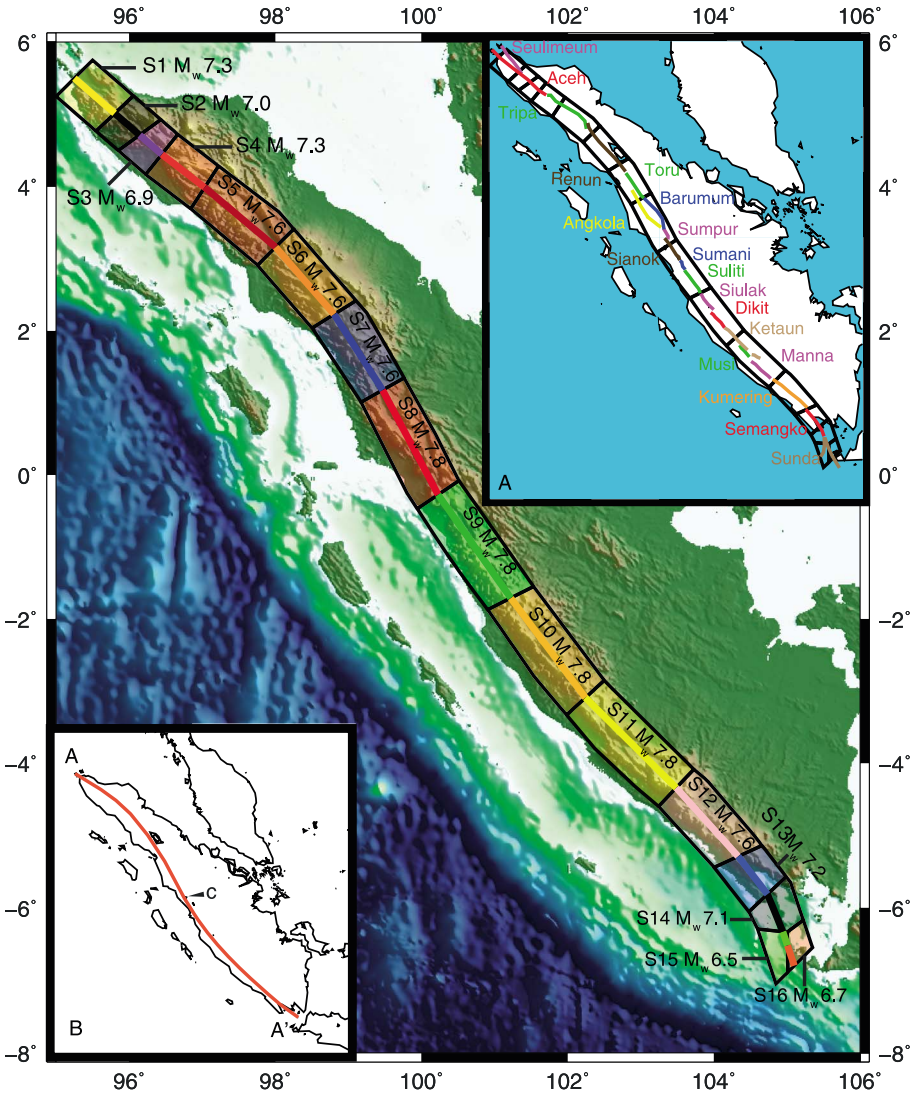

Figure 3. Segments of the Sumatran fault. Segments are obtained by Voronoi tessellation of the centroids of 16 seismicity clusters shown in Figure 2 (and similarly color coded). Maximum credible earthquake magnitudes M_w are obtained from segment length and regressions of *Wells and Coppersmith* [1994]. Segment nomenclature (from the north): S1: Seulimeum, S2: Aceh-north, S3: Aceh-south, S4: Tripa-north, S5: Tripa-south, S6: Renun, S7: Toro, S8: Barumun, S9: Suliti, S10: Siulak, S11: Manna, S12: Kumering, S13: Semangko, S14: Sunda 1, S15: Sunda 2, and S16: Sunda 3 (see Table 1 for more description). (inset a) For comparison, it illustrates the segmentation suggested by *Sieh and Natawidjaja* [2000], overlaid on the *k*-means $K = 16$ segment model. (inset b) Highlights the double-bow shape of the Sumatran fault, from A to A′, hinged at *c*.

issues compare reasonably among these models, they are arrived at by very different techniques, and there are important differences over segment length in the north and in segment distribution.

6. Discussion and Conclusions

Geological description along the fault is of surficial manifestations, symptomatic of process and structure at depth but secondary to those structures. Inspecting seismicity expands this sampling and invokes depth sampling and a volume of seismogenesis sustaining and influencing surficial manifestations of fault rupture. The Sumatran fault has overall shape (Figures 2 and 3) like a reflected letter "S" or double bow. It looks like a compressed linear object AA′ (Figure 3b), pinned or hinged between clusters near the center *c* just south of the equator (Figure 2), and it has bowed out, under horizontal compressive stress (acting ~N-S), eastward in the north and westward in the south. Centroid positions also support this observation. When

looking in the northern domain, 6/8 of the centroid positions locate to the east of the fault trace; conversely in the southern domain, 6/8 of the centroid positions locate to the west of the fault trace. The massive part of the two bows spans eight great segments, four either side of the equator, being slightly longer in the south. These eight segments span 1320 km. Both the north and south ends of the fault consist of several shorter segments and appear relatively fragmented compared to the massive central segments, particularly in the south, where the fault extends into the Sunda Strait and bifurcates and perturbs the end of the reflected S. The northern segmentation also encompasses a degree of fragmentation (and bifurcation [see *Bennett et al.*, 1981] unresolved by *k*-means).

There is a strong cluster demarcator between Barumun (S8) and Suliti (S9) (Table 1 and Figure 3 for nomenclature). This demarcator appears consistently in all the trials. The first reason for this is very obvious; there is a void of seismicity between clusters Barumun and Suliti. Historical ruptures reviewed by *Natawidjaja and Triyoso* [2007] do not break through this void either. Barumun segment finishes at 0°16.5′S, 100°13.6′E. This coordinate in *Kastowo and Leo*'s [1973] map is about 2 km from the NE corner of Danau Maninjau, in formation Qamj, which has the legend "Andesite of Danau Maninjau Caldera—the elongated form of caldera could indicate a prolonged period of eruption during right-lateral displacement on the Great Sumatran fault zone…." But most of Lake Maninjau extends N-S, not NNW-SSE. Only the northernmost extent of Lake Maninjau extends NNW parallel to GSF strike, the mapped surface trace of which is less than 10 km NE from this point. The GSF trace runs between G Singgalang and G Marapi and is mapped up the valley to these volcanoes but not on them. This small apparent disparity between GSF trace and end of our Barumun segment is in fact consistent with that of our analytical results determined using earthquake sampling nucleation and rupture processes at depth, not at the surface. Inspecting the digital elevation model and *Kastowo and Leo*'s [1973] map, the four volcanoes of Lake Maninjau (extinct), Singgalang and Tandikat, and Marapi are inescapable straddling the GSF in a band W-E. The demarcator computed between Barumun and Suliti is at Lake Maninjau. The relationship, if any exists beyond coincidence, between the roots of these volcanoes and the Sumatran fault at depth is unexplained. Body wave tomography maps a strong shear wave velocity decrease (35%) in central Java associated with volcanoes Merapi and Lawu [*Koulakov et al.*, 2009a], and adjacent to the Sumatran fault, *Koulakov et al.* [2009b] map a similar anomaly (reaching 18–7% for 5 km down to 75 km) below the Toba caldera complex. The Sumatran fault surficially avoids the Toba complex, passing to the west of Toba Lake in Renun (S6; Figure 3). It is possible that the reduced shear velocity and modulus influence contemporary seismicity below volcanoes near Lake Maninjau. The structure below the volcanoes near Lake Maninjau merits investigation.

Segment characteristics and junctions can now be described in detail, bearing in mind the nomenclature of *Sieh and Natawidjaja* [2000], who describe segment controlling features along the fault which are mostly dilatational step overs of size 5–10 km, complex fault structures, and contractional bends. Not all of these appear influential in controlling seismicity in our model. Although inspecting earthquakes that have depth, not one surface rupture crosses our GSF cluster-segment boundaries (see-through ellipses in Figure 2). Small dilatational step overs within Barumun and Suliti segments (associated with the ends of Sieh and Natawidjaja's Sianok segments) do not influence our cluster segmentation, and these are absorbed into Barumun and Suliti. Southward, the Suliti segment in Table 1 extends 194 km, terminating at the reported dilatational step over (and Sumani is absorbed into Suliti). Siulak (S10) extends 191 km (Dikit is absorbed); Manna (S11) extends 196 km to a contractional bend; Kumering (S12) extends another 141 km south to the reported dilatational step over. That completes the massive segments. Semangku (S13) extends a smaller distance 73 km to the peninsula in south Sumatra, where Sunda (S14) becomes submarine.

Northward from *c*, Barumun crosses the equator and is 182 km long (Sumpur is absorbed) and joins Toru (S7) at the reported contractional bend then extends 138 km northward; Renun (S6) is 136 km long; Tripa-south (S5) is 142 km long. The demarcator between Renun and Tripa-south at 3.2°N appears consistently for both the $h \leq 50$ and ≤ 60 km analyses. The $h \leq 60$ km analysis confirms much of the $h \leq 50$ km analysis throughout. That completes the northern massive segments. Tripa-north (S4) extends 83 km northward, near a contractional step over. Aceh-south (S3) continues 45 km, Aceh-north (S2) 50 km, and Seulimeum (S1) 82 km; these, discussed later, complete the Sumatran fault on land.

The fault in the north and south has relatively short segments, as it progresses to change in the faulting character as the Andaman Sea spreading center is approached in the north, and the Sunda Strait subduction is approached in the south. In the Sunda Strait, three segments, not one, are determined using $h \leq 50$ km. The

Sunda segments of the fault swing south toward the fore arc and the Sunda Trench. The first segment Sunda 1 (S14) is 62 km long. Segments Sunda 2 (S15) and Sunda 3 (S16) show clear bifurcation emerging from the analysis in Figure 3. However, if deeper seismicity ≤60 km is included in the analysis, then the three segments S14–S16 are resolved as one Sunda segment. Investigation down to 60 or 70 km is counter to initial measures designed to eliminate "foreign" or megathrust seismicity and introduces 10% more seismicity into these southern segments, with the largest increases along the fault. The implication is that the analysis down to 50 km reveals shallower cluster structure as intended for the Sumatran fault, but as the fault swings toward the fore arc, deeper structure influences and coalesces three-cluster resolution into one cluster, as the strike-slip fault character is deteriorated. The slightly different $K = 12$ model (which is not corroborated by the $h \leq 60$ km analysis; Figure 2b) also supports this by absorbing the Sunda 1–Sunda 3 clusters into one. The fault in the north is of different character with four moderate length segments. The division of Tripa into north (S4) and south (S5) is consistent in $h \leq 50$ and ≤60 km analyses and resolves Tripa-north with 83 km length; the resolution of an Aceh-north (S2) and Aceh-south (S3) is also consistently seen.

There is an importance in the lengths of these northern segments. Their M_{cred} are M_w 6.9–7.3 and do not approach the range M_w 7.6–7.8 of the great central segments, being significantly shorter. Large-slip rates are commensurate with large segments and high M_{cred}; this works well for the great segments north of the equator, where slip is ~23 mm/yr, but lower values ~13 mm/yr (maximum) across the GSF in northern Sumatra are reported by *Genrich et al.* [2000] with greater uncertainty. The difficulties accommodating these lower values have been discussed by *Natawidjaja and Triyoso* [2007]; nevertheless, these lower slip rates are compatible with shorter segments through Seulimeum to Tripa-north (S1–S4) and thereby, importantly, lower the hazard expectation in the north (although M_w 6.9–7.3 is far from inconsiderable). If a conservative hazard judgment prevailed and took the combined rupture of Aceh-north with Aceh-south as possible (demarcator S2 and S3 degraded to a rupturable barrier), then M_{cred} would rise to M_w 7.4.

To summarize, *k*-means partitioning of seismicity into earthquake clusters and tessellation of these into adjacent zones straddling the surficial mapped fault strike together suggest that the Sumatran fault consists of 16 segments. These 16 segments range in length 22–196 km, with corresponding maximum credible earthquakes M_w 6.5–7.8. The distribution of segments is approximately symmetrical with eight great central segments encompassing 1320 km (three quarters of the fault) and four smaller segments in the far north and in the far south. Historical surface ruptures are contained within the segments. None rupture across boundaries. The shape traced by the cluster centroids and tessellated segments is a reflected letter S or double bow hinged around 0.27°S at Lake Maninjau, which coincides with a W-E string of four volcanoes across the surficial fault. The north bow bows east, and the south bow bows west, as if squeezed along the trace of the fault. The northern extent of the Sumatran fault shares lower seismic hazard among its four segments, which have shorter segment lengths than the dominating eight great central segments which have potential to rupture in M_w 7.6–7.8 events.

Acknowledgments
We are grateful to Mark Petersen and Charles Mueller (earthquake catalog), Hardy Guchi (scarce geological maps), Zulkifli Nasution (local support), Graeme Weatherill (initial software), and two very helpful reviewers.

The Editor thanks two anonymous reviewers for their assistance in evaluating this paper.

References

Bellier, O., M. Sebrier, S. Pramumijoyo, T. Beaudouin, H. Harjono, I. Bahar, and O. Forni (1997), Paleoseimicity and seismic hazard along the Great Sumatran Fault, *J. Geodyn.*, *24*, 169–183.

Bennett, J. D., et al. (1981), *Peta Geologi Lembar Bandaaceh, Sumatra, Lembar: Banda Aceh 0421, 0521 Sekala 1:250,000*, Geological Research and Development Centre, Ministry of Mines and Energy (with the Institute of Geological Sciences UK), Jl. Diponegoro 57, Bandung, Indonesia.

Cornell, C. A. (1968), Engineering Seismic Risk Analysis, *Bull. Seismol. Soc. Am.*, *58*, 1583–1606.

De Joussineau, G., and A. Aydin (2009), Segmentation along Strike-Slip Faults Revisited, *Pure Appl. Geophys.*, *166*, 1575–1594.

Genrich, J. F., Y. Bock, R. McCaffrey, L. Prawirodirdjo, C. W. Stevens, S. S. Puntodewo, C. Subarya, and S. Wdowinski (2000), Distribution of slip at the northern Sumatran fault system, *J. Geophys. Res.*, *105*, 28,327–28,342.

Hartigan, J. A. (1975), *Clustering Algorithms*, John Wiley, New York.

Ishii, M., E. Kiser, and E. L. Geist (2013), M_w 8.6 Sumatran earthquake of 11 April 2012: Rare seaward expression of oblique subduction, *Geology*, *41*, 319–322, doi:10.1130/G33783.1.

Jarvis, A., H. I. Reuter, A. Nelson, and E. Guevara (2008), Hole-filled SRTM for the globe Version 4. [Available from the CGIAR-CSI SRTM 90m Database, http://srtm.csi.cgiar.org.]

Kagan, Y. Y., and D. D. Jackson (2000), Probabilistic forecasting of earthquakes, *Geophys. J. Int.*, *143*, 438–453.

Kagan, Y. Y., and D. D. Jackson (2013), Tohoku earthquake: A surprise?, *Bull. Seismol. Soc. Am.*, *103*, 1181–1194.

Kastowo, and G. W. Leo (1973), *Peta Geologi Lembar Padang, Sumatra, Lembar: 4/VIII Sekala 1:250,000*, Geological Survey of Indonesia, Ministry of Mines (with the USGS), Jl. Diponegoro 57, Bandung, Indonesia.

Koulakov, I., A. Jakovlev, and B. G. Luehr (2009a), Anisotropic structure beneath Central Java from local earthquake tomography, *Geochem. Geophys. Geosyst.*, *10*, Q02011, doi:10.1029/2008GC002109.

Koulakov, I., T. Yudistira, B.-G. Luehr, and Wandono (2009b), *P, S* velocity and V_P/V_S ratio beneath the Toba caldera complex (Northern Sumatra) from local earthquake tomography, *Geophys. J. Int.*, *177*, 1121–1139, doi:10.1111/j.1365-246X.2009.04114.x.

Krzanowski, W. J., and Y. T. Lai (1988), A criterion for determining the number of groups in a data set using sum-of-squares clustering, *Biometrics*, *44*, 23–34.

Natawidjaja, D. H., and W. Triyoso (2007), The Sumatran fault zone – from source to hazard, *J. Earthquake Tsunami*, *1*, 21–47.

Parsons, T., and E. L. Geist (2009), Is there a basis for preferring characteristic earthquakes over a Gutenberg-Richter distribution in probabilistic earthquake forecasting?, *Bull. Seismol. Soc. Am.*, *99*, 2012–2019, doi:10.1785/0120080069.

Parsons, T., R. Console, G. Falcone, M. Murru, and K. Yamashina (2012), Comparison of characteristic and Gutenberg-Richter models for time-dependent $M \geq 7.9$ earthquake probability in the Nankai-Tokai subduction zone, Japan, *Geophys. J. Int.*, *190*, 1673–1688, doi:10.1111/j.1365-246X.2012.05595.x.

Petersen, M. D., J. Dewey, S. Hartzell, C. Mueller, S. Harmsen, A. D. Frankel, and K. Rukstales (2004), Probabilistic seismic hazard analysis for Sumatra, Indonesia and across the Southern Malaysian Peninsula, *Tectonophysics*, *390*, 141–158.

Petersen, M. D., S. Harmsen, C. Mueller, K. Haller, J. Dewey, N. Luco, A. Crone, D. Lidke, and K. Rukstales (2007), *Documentation for the Southeast Asia Seismic Hazard Maps*, 65 pp., USGS, Reston, Va.

Shen, Z.-K., J. Sun, P. Zhang, Y. Wan, M. Wang, R. Bürgmann, Y. Zeng, W. Gan, H. Liao, and Q. Wang (2009), Slip maxima at fault junctions and rupturing of barriers during the 2008 Wenchuan earthquake, *Nat. Geosci.*, *2*, 718–724, doi:10.1038/NGEO636.

Sieh, K., and D. Natawidjaja (2000), Neotectonics of the Sumatran fault, Indonesia, *J. Geophys. Res.*, *105*, 28,295–28,326.

Sorensen, M. B. (2008), Continued earthquake hazard in northern Sumatra, *Eos*, *89*, 133–140.

Weatherill, G., and P. W. Burton (2009), Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region, *Geophys. J. Int.*, *176*, 565–588, doi:10.1111/j.1365-246X.2008.03997.x.

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

Woo, G. (1996), Kernel estimation methods for seismic hazard area source modelling, *Bull. Seismol. Soc. Am.*, *86*, 353–362.