2018 SCEC Report

Implementing rapid, probabilistic association of earthquakes with source faults in the CFM for southern California

John H. Shaw (PI) & Andreas Plesch
Department of Earth and Planetary Sciences
Harvard University
20 Oxford St.
Cambridge, Massachusetts 02138
shaw@eps.harvard.edu

Egill Hauksson (PI) & Men-Andrin Meier
Seismological Laboratory
California Institute of Technology
Pasadena, CA 91125
hauksson@gps.caltech.edu

Proposal Categories:	Collaborative Proposal: Data Gathering and Products
Science Objectives:	P3a, P2a, P3e
Duration:	1 February 2018 to 31 January 2019

Summary
This project has developed a new, statistically robust way to identify the fault (or sets of candidate faults) in the Community Fault Model (CFM) that generated an earthquake using information typically provided soon after these events occur (Evans, 2016). This effort effectively bridges the information provided by increasingly sophisticated near real-time seismograph networks with comprehensive 3D Community Fault Models (CFM's) developed by SCEC (Plesch et al., 2007). Our method of earthquake-to-fault association was developed using comprehensive earthquake hypocenter and focal mechanism datasets in California through 2016 (after Hauksson et al., 2012; Yang et al., 2012) and the southern California Community Fault Model (CFM) (Plesch et al., 2007), to assess what properties of earthquakes serve as the best predictors of the fault on which they occurred. We used a series of training datasets for earthquakes that were known to have occurred on faults within the model, and established that proximity (distance), focal mechanism (nodal plane orientation), and earthquake history (spatial and temporal clustering) can be combined in a robust way to assign probability that a given earthquake was associated with one or more source faults in the model (or on a fault not included in the model). Notably, these training datasets were comprised of earthquakes that occurred in the decade since the release of CFM 2.0, to ensure that they did not influenced the modeled fault geometries.

The method is implemented as an R script which calculates distances between earthquakes and CFM faults, and compares nodal plane orientations (see Figure 1). We have tested this approach on a series of previous CFM versions, and subsequently applied it to a current model version (5.1). For each earthquake in the catalog, the code outputs the five highest probabilities of association with a CFM fault, as well as the probability that the event is not associated with any source within the model. The majority of earthquakes (> 75%, above M 3) have a high probability of association with one or two faults in the model.

Objective earthquake-to-fault associations are of value as they provide an important measure of the activity of faults within the southern California plate boundary. The associations will facilitate detailed studies of whether and to what extent on-fault and off-fault earthquakes differ in their behavior. Moreover, as the method is implemented in near real-time, it will prove helpful in identifying clusters of small earthquakes that may be foreshocks of a larger, imminent event on a major fault. They will also assist in communicating objective information about the faults that source earthquakes to the scientific community and general public. Ultimately, more accurate identification of source faults for large earthquakes may help responders know which planned scenario is most like the situation they are facing. Such guidance should lead to more effective responses that can help save lives.

Method
Identifying the fault that generated an earthquake is often not straightforward. Most moderate to small earthquakes, as well as larger blind events, do not involve direct surface rupture. Epicentral locations may be at significant distances from surface traces of dipping source faults. Moreover, many earthquake sequences occur near fault junctures, or along fault zones that involve several closely spaced segments, providing several options for the earthquake source. In order to develop an objective approach to defining earthquake source faults, we evaluated a training dataset consisting of more than 600 earthquakes in southern California that had known associations with CFM faults (based on surface rupture, source inversions, or other seismologic, geodetic, or geologic studies) (Evans, 2016). In addition, we considered more than 20 events that were not associated with a fault in the model. We initially compared the properties of a subset of these events with known associations with a version of the CFM (2.0) completed in 2005, prior to the occurrence of these earthquakes. This addressed concerns related to the fact that CFM representations are developed, in part, using earthquake information.

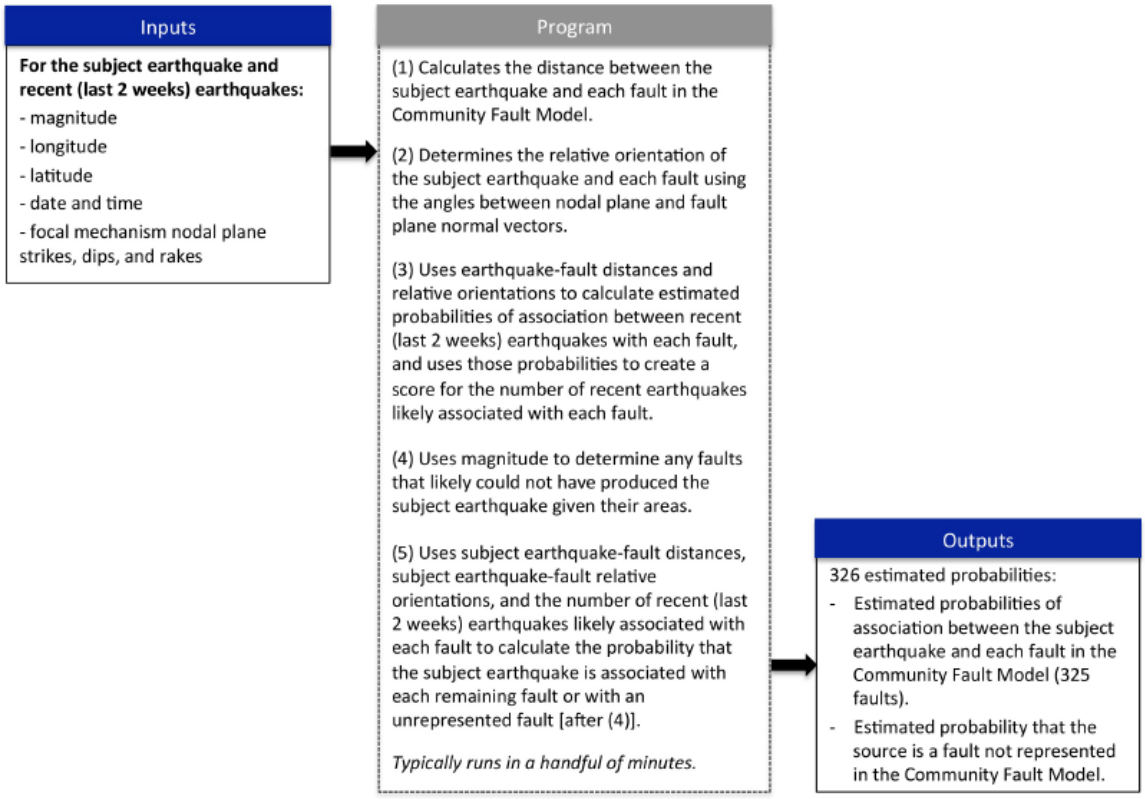

Figure 1: Earthquake-to-fault association workflow.

We design and evaluate four different models that use catalog data to predict the causative source fault(s) and optimize the model parameters in a maximum likelihood sense. First, we employed a method to calculate the distance between earthquake hypocenters and the closest part of faults within the CFM, which are represented as triangulated surfaces (tsurfs). For subsequent assessments, we only considered the faults that are within 20 km of the hypocenters as candidate source faults. Based on typical uncertainties in earthquake locations, we feel that this is a conservative assumption which does not exclude any viable source faults from consideration. We found that the distance to the true source fault is, on average, significantly lower than that with unassociated candidate faults, confirmed our expectation that distance is a useful predictor. The first, simplest model posits that the fault association only depends on the shortest distance between the fault and the hypocenter, with the probability decaying exponentially with distance. This Model 1 assigns the highest probability of association to 83.6% (as opposed to a 40% random probability) of events in our training dataset to the proper source fault. Figure 2a shows an example of applying this model to the Laguna Salada fault. Despite this success, many earthquakes were not assigned to the proper source fault in regions where two or more faults intersected forming a complex junction. Thus, we also explored comparisons of earthquake focal mechanisms with respect to the geometries of nearby faults represented in the model. For this assessment, we compared the minimum angle θ between the normal to the faults and the focal mechanism nodal planes, and used the plane of the two nodal planes that was closest to the orientation of the fault. We see that θ is, on average, smaller for true associations than between events and unassociated candidate faults. Thus we extended our statistical model to include both distance and orientation (Model 2). The revised model assigns the highest probability of association to 85.6% of events in our training dataset to the proper source fault, a 2% improvement over the distance-only assessment.

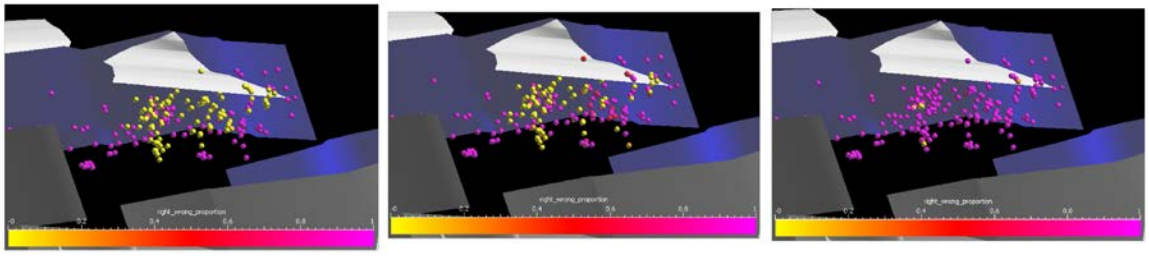

Figure 2: Effect of adding predictors to association model trained on CFM 2 faults, shown for the Laguna Salada fault; yellow events are miss-associated; (left) using only distance (Model 1); (middle) using distance and nodal planes (Model 2); (right) using distance, nodal planes and clustering (Model 3).

As a next step, we considered using earthquake clustering in our model. This approach suggests that if a set of events preceding a candidate earthquake were associated with a given fault that the candidate event should have a higher probability of being associated with that same fault. For this, we examined for each event the sum of the predicted probabilities from Model 2 between events within the last 14 days (an arbitrarily-selected time period) and each candidate fault. In order to model this effect, we included an increase in the modeled probability of an association with fault j exponentially in proportion with the sum of the earlier predicted probabilities from Model 2. This new Model 3 correctly associates the highest probability to true source faults for 98% of the events, a significant improvement over previous models (Figure 2). We then extended our analysis to assign probabilities to the case that a source fault was not included in the CFM. In Model 4, we directly used the model structure from Model 3 by considering the sum by which the probabilities assigned are normalized in order to make them equal to one; this sum is the total sum of the un-normalized probabilities, based on the predictors in the model, that the given event is associated with each of the candidate faults in the CFM. With higher probabilities of being associated with individual faults in the CFM (and therefore a higher normalizing sum), an event should be less likely to be associated with a fault not represented in the CFM. Model 4 is thus represented as;

$$\mathbf{P(y_{i,j}=1)} \propto \frac{\eta * P_{i,j}}{\eta * \sum_j P_{i,j} + \iota * D_i}$$

$$\mathbf{P(y_{i,Not\ in\ the\ CFM}=1)} \propto \frac{\iota * D_i}{\eta * \sum_j P_{i,j} + \iota * D_i}$$

where

$$S_{i,j} = \sum_k e^{(\delta * d_{k,j} + \epsilon * \theta_{k,j})}$$

$$P_{i,j} = e^{(-\alpha * d_{i,j} - \beta * \theta_{i,j} - \gamma * S_{i,j})}$$

$$D_i = e^{(-\zeta * \sum_j P_{i,j})}$$

and $\mathbf{P(y_{i,j}=1)}$: likelihood that event i is associated with fault j

$d_{k,j}$: distance between event k preceding event i up to 14 days and fault j

$\theta_{k,j}$: smallest angle between nodal planes of event k preceding event i up to 14 days and fault j

$d_{i,j}$: distance between event i and fault j

$\theta_{i,j}$: smallest angle between nodal planes of i and fault j

This model correctly assigned the highest probability to "not in the CFM" for 38% of the training dataset. However, when we subsequently applied Model 4 to a larger training dataset (27 versus 88 earthquakes known to be not associated with a CFM fault) and a more complete model version (CFM 5.0), 80% of the "not in the CFM" events were properly assigned. We speculate that the larger dataset was required for the model parameters to be tuned more effectively.

To further test the method, we applied Model 4 (after returning with the complete training data set) to a historical earthquake catalog (Hauksson et al. 2015) consisting of 1,660 $M \geq 3$ events to a more recent version of the Community Fault Model (CFM 5.0). Notably, CFM 5 includes more faults (325 vs. 153) with more detailed representations. As a result, less than 4% of earthquakes had only one candidate source fault within the 20 km distance, and many events had more than 10 faults, lowering significantly the probability of correct assignment by random choice. The results included probabilities of association with each candidate fault, as well as the likelihood that the event

occurred on a fault not represented in the model. 76% of these historical events had a primary association with a CFM 5 fault, leaving 24% of events with the highest probability of not be sourced by a fault in the model. Based on magnitude to moment scaling relationships, more than 99% of the moment release is associated with associated source faults represented in the CFM 5.0.

Enhancements to the Code

In the past year, we made a series of enhancements to the products that are delivered with the earthquake-to-fault associations. These include providing the distance of the event to the associated fault, as well as the common fault name (the code currently delivers the CFM fault name). Ultimately, these will be linked to other resources, such as the USGS Quaternary Fault and Fold Database (Qfault) ID, such that users can explore other resources that help to constrain fault parameters (e.g., MRE, slip rate). We also have installed the code in the SCEDC development environment and developed scripts to incorporate into SCEDC processing so that this information will be automatically calculated for new events.

In addition, we applied the code to a comprehensive catalog of more than 7000 earthquakes M > 3.0 (focal mechanisms from Yang et al., 2012, extended to 2013, joined with Hauksson et al., 2012, extended to 2013) and a current CFM model version (5.1). This analysis associated 76% of the earthquakes to a model fault, leaving 24% without a primary fault association. 42% of earthquakes had higher than 90% association probabilities. 17% of earthquakes were without a primary fault association but had a secondary association with a model fault. 7% of earthquakes could not be associated with any fault. The large size of the catalog revealed performance bottlenecks which are now being addressed.

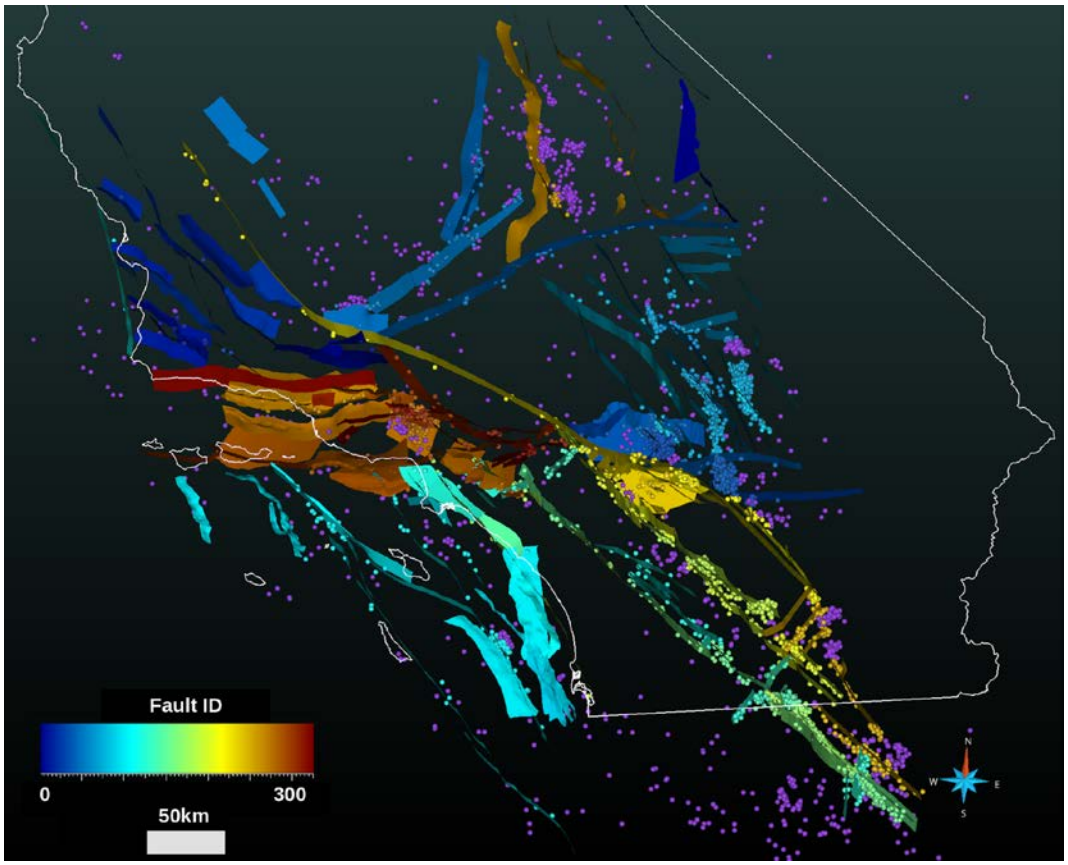

Figure 3: Catalog of M > 3 earthquakes, associated with CFM 5.1 faults. Colors for both earthquakes and faults represent a fault identification number. Thus, colors of events are shared with the fault that has the highest probability of association. Purple color for events indicates the lack of primary fault association.

References

Evans, W., (2016), Determining Earthquake Source Faults in Southern California: Statistical Models for Historical and Future Earthquakes, Senior thesis in Statistics, Harvard University, Cambridge, MA (in prep for submission to BSSA).

Hauksson, E., W. Yang, and P. M. Shearer, (2012), Waveform relocated earthquake catalog for southern california (1981 to June 2011), *Bull. Seismol. Soc. Am.*, *102*(5), 2239–2244.

Plesch, A., J. H. Shaw, C. Benson, W. A. Bryant, S. Carena, M. Cooke, J. Dolan, G. Fuis, E. Gath, L. Grant, E. Hauksson, T. Jordan, M. Kamerling, M. Legg, S. Lindvall, H. Magistrale, C. Nicholson, N. Niemi, M. Oskin, S. Perry, G. Planansky, T. Rockwell, P. Shearer, C. Sorlien, M. P. Süss, J. Suppe, J. Treiman, and R. Yeats, (2007), Community Fault Model (CFM) for Southern California, Bulletin of the Seismological Society of America, Vol. 97, No. 6, doi: 10.1785/012005021

Yang, W., E. Hauksson, and P.M. Shearer, P. M., (2012), Computing a large refined catalog of focal mechanisms for southern California (1981–2010): Temporal stability of the style of faulting. BSSA, 102(3), 1179-1194.