Earthquake Early Warning ShakeAlert System: Testing and Certification Platform

by Elizabeth S. Cochran, Monica D. Kohler, Douglas D. Given, Stephen Guiwits, Jennifer Andrews, Men-Andrin Meier, Mohammad Ahmad, Ivan Henson, Renate Hartog, and Deborah Smith

ABSTRACT

Earthquake early warning systems provide warnings to end users of incoming moderate to strong ground shaking from earthquakes. An earthquake early warning system, ShakeAlert, is providing alerts to beta end users in the western United States, specifically California, Oregon, and Washington. An essential aspect of the earthquake early warning system is the development of a framework to test modifications to code to ensure functionality and assess performance. In 2016, a Testing and Certification Platform (TCP) was included in the development of the Production Prototype version of ShakeAlert. The purpose of the TCP is to evaluate the robustness of candidate code that is proposed for deployment on ShakeAlert Production Prototype servers. TCP consists of two main components: a real-time *in situ* test that replicates the real-time production system and an offline playback system to replay test suites. The real-time tests of system performance assess code optimization and stability. The offline tests comprise a stress test of candidate code to assess if the code is production ready. The test suite includes over 120 events including local, regional, and teleseismic historic earthquakes, recentering and calibration events, and other anomalous and potentially problematic signals. Two assessments of alert performance are conducted. First, point-source assessments are undertaken to compare magnitude, epicentral location, and origin time with the Advanced National Seismic System Comprehensive Catalog, as well as to evaluate alert latency. Second, we describe assessment of the quality of ground-motion predictions at end-user sites by comparing predicted shaking intensities to ShakeMaps for historic events and implement a threshold-based approach that assesses how often end users initiate the appropriate action, based on their ground-shaking threshold. TCP has been developed to be a convenient streamlined procedure for objectively testing algorithms, and it has been designed with flexibility to accommodate significant changes in development of new or modified system code. It is expected that the TCP will continue to evolve along with the ShakeAlert system, and the framework we describe here provides one example of how earthquake early warning systems can be evaluated.

Electronic Supplement: Tables of test suite events used to assess the ShakeAlert system and a thorough description of the nondeterministic behavior of the system during test runs.

INTRODUCTION: PURPOSE AND CONTEXT

ShakeAlert is an earthquake early warning (EEW) system being developed to provide real-time alerts for moderate to large earthquakes in the western United States (Given *et al.*, 2014). Following nearly 10 years of research and development, a Production Prototype ShakeAlert system was released in California in 2016, and across the West Coast of the United States (California, Oregon, and Washington states) in 2017 (see Kohler *et al.*, 2017). Alerts are currently being used by a selected group of community participants, but as the system is developed, the number and types of end users will expand. End users are determining best practices for the use of earthquake alerts by individual companies or across industry sectors, and in some cases pilot applications that take actions based on alerts have been initiated, in preparation for broader alert dissemination.

The ShakeAlert system is not a static software system; it is being continuously developed and tested to improve the system's capability to send alerts for an evolving earthquake rupture. The goal of the system is to provide alerts that can be used to initiate procedures that mitigate the effects of expected ground shaking. New and modified ShakeAlert code modules are periodically introduced that enhance or replace the versions currently in production. All code revisions must perform robustly upon implementation; thus, a testing and certification architecture is necessary to quantitatively determine whether new or enhanced code modules improve overall system performance. System performance includes assessing code optimization and stability as well as the quality of alert products distributed to users.

EEW operates as a subdiscipline of network seismology, and as such requires a unique and customized test environment. Although EEW systems exist in other cities and countries, such as Mexico City, Mexico (Espinosa Aranda *et al.*, 1995), Istanbul, Turkey (Erdik *et al.*, 2003), Japan (Hoshiba

 doi: 10.1785/0220170138

et al., 2008), and Taiwan (Hsiao *et al.*, 2009), there is essentially no literature that describes a procedure for testing operational EEW system robustness to a wide range of earthquake and signal types. Here, we introduce the current version of the ShakeAlert Testing and Certification Platform (TCP) that was developed for the ShakeAlert Production Prototype system, but it is expected to continue to evolve and mature.

The TCP design is driven by several key requirements. First, the testing process must be independent of code development to provide an objective evaluation of the software. Additionally, this ensures that the software can be installed and run in an environment that mimics the production environment. Second, a large dataset consisting of historic earthquakes and anomalous seismic signals is used to evaluate the system. The testing framework allows for the flexibility to use either the entire test dataset or only a subset of the dataset that is most relevant to the candidate code. Third, a centralized testing framework ensures that common inputs (including the type and number of test events), environments, and processing capabilities are used in the comparison testing. Fourth, assessments use a uniform set of metrics that can be compared across systems and code versions. Finally, the ShakeAlert TCP is designed to be transparent to allow independent use of the datasets and verification of the results described here.

DATA AND TESTING PROCEDURE

EEW uses techniques that are different from traditional earthquake detection and location methods to estimate earthquake source information and determine expected intensity of ground shaking across a region (see Given *et al.*, 2014; Kohler *et al.*, 2017). Because the goal of the ShakeAlert system is to provide information about expected ground shaking before it arrives at an end user's location, the tests to verify system and algorithm performance must take into account latency of alert information as well as the accuracy of point-source information and predicted ground shaking. The ultimate objective is to ensure operational functionality and provide assessments that estimate the usefulness of the EEW alerts to end users. Testing comprises two components: (1) *in situ* tests to ensure functionality in a real-time environment and (2) performance assessment by retrospective testing using a test data suite. Results of both testing components are compiled into summary reports. Then, these findings are used to develop recommendations on whether to deploy the candidate code onto the production servers.

Real-Time *In Situ* Test Servers

The first TCP component of testing new and modified algorithms that are candidates for ShakeAlert is a real-time *in situ* test to assess algorithm behavior and performance in the current seismic network operational state. The real-time test implements a system and software environment that replicates the production system, except for the inclusion of the new or revised candidate code. Through the real-time *in situ* tests, operational problems can be identified before implementation on production servers (for details, see Kohler *et al.*, 2017). The real-time test servers are collocated with the production servers in U.S. Geological Survey (USGS)-Pasadena/California Institute of Technology, USGS-Menlo Park, University of California at Berkeley, and the University of Washington. Like the Production Prototype servers, configuration management software is used to ensure a uniform build across all test servers, including uniform installation, updating, and patching of the Linux operating system. The same configuration management software is used to deploy ShakeAlert algorithms to be tested across all servers. The test servers ingest exactly the same waveform data, and they exchange messages in the same manner as the Production Prototype system.

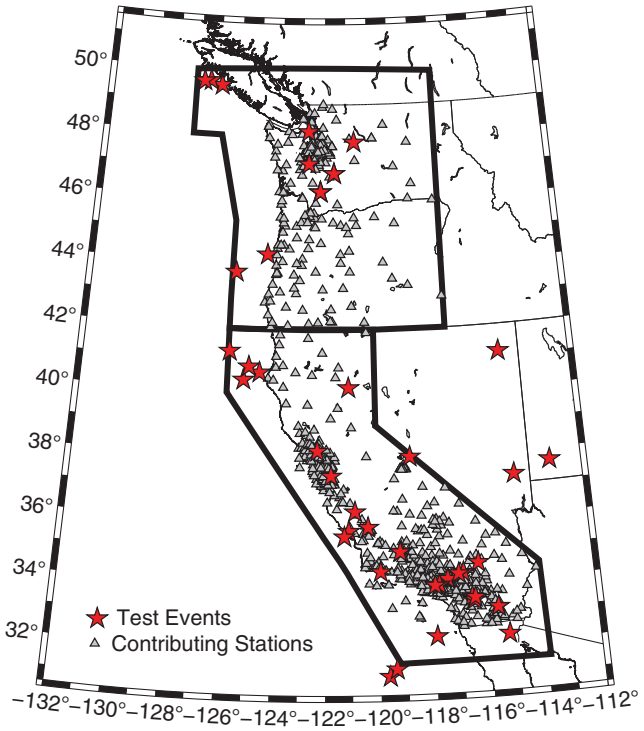
Test candidate codes are run in real time for a minimum of two weeks, although the total length of time for the *in situ* test depends on the features being tested. During the tests, aspects related to the hardware, software, and network as well as the ShakeAlert software execution are monitored. After the software has been installed via the configuration management system, the system resources are monitored to compare the central processing unit (CPU) load, network interface buffers, memory usage, and disc I/O against a baseline metric defined by performance on the production servers. Both improvements and degradations in software performance are investigated, and the findings are included in a summary report.

Historic Event Test Suite

The second component of the TCP consists of playback testing of significant historic earthquakes and signals associated with false alerts. The test suite provides a stress test of the mission-critical ShakeAlert system by inclusion of both expected and potentially anomalous waveform cases. Additionally, these tests are designed to measure how the algorithms, and the system as a whole, handle errors and loads that are beyond the scope of normal operation. The test events described below are by no means considered comprehensive and complete; as new earthquake and anomalous signal datasets become available, they will be added to the historic event suite. Additionally, the TCP does not currently consider network latencies (observed latencies and effects of late data), changing station densities (e.g., removal of stations to replicate a network outage), and other important test cases; such features will be added in the future.

The events that comprise the TCP historic test suite consist of significant southern California, northern California, and Pacific Northwest earthquakes that have occurred since 1999. Their significance is defined by their relatively large magnitudes, epicenter locations relative to seismic network station locations, or close timing of foreshocks and aftershocks. Events prior to 1999 are not included because network density was significantly lower, and these events predate the installation of a significant number of broadband sensors by the regional networks (i.e., Southern California Seismic Network [SCSN], Northern California Seismic System [NCSS], and Pacific Northwest Seismic Network [PNSN]).

In addition, anomalous events are used in testing because these may be mistaken for earthquake ground motion and result in alerts. These include calibration and recentering events



▲ **Figure 1.** Map showing the distribution of local and regional earthquakes (stars) that are used in retrospective testing of system performance. Triangles show current real-time stations that are contributing to ShakeAlert; station distribution is sparser for older historic events. Thick black lines show the boundary of the alerting regions for California and the Pacific Northwest.

from SCSN and NCSS as well as anomalous signals recorded by PNSN due to noisy or dead channels, leap-second corrections, or waveform processing restarts. A fourth subset of events includes regional and teleseismic events, some of which generated false alerts in the real-time system. Finally, very deep or very large teleseismic events are also included because they have the potential to trigger an alert due to an abundance of high-frequency relatively large-amplitude energy with waveform characteristics similar to local earthquake waveforms. Figure 1 shows a map of the local and regional historic events used in the retrospective tests; tables of each category of test events are provided in the Ⓔ electronic supplement to this article.

For testing of local mainshock events, waveforms are windowed such that they start 2 min prior to and end 5 min after the event origin time. The pre-event window is chosen to provide adequate time for waveform processing computations (e.g., baseline removal, filtering, and trigger identification) to initiate before the *P*-wave arrival at the first station. The post-event window is chosen to allow enough time for a *P* wave to be recorded on the majority of stations across the region. Every test uses a compilation of all available waveforms from southern California, northern California, and the Pacific Northwest. The 7-min-long window may include additional local events, including aftershocks of the mainshock that may also be considered in the performance assessments.

For the calibration, recentering, or other anomalous (non-earthquake) signal test events, data durations vary and depend on how long the signal is observed at the specific sensor being calibrated or recentered, with appropriate pre-event and postevent time. For the anomalous and teleseismic event test cases, waveform start times are based on the estimated *P*-wave arrival time at the closest ShakeAlert station that was in operation at the time of the event, with 2 min of pre-event time applied. Total waveform durations of 7 min are used because this adequately allows for the recording of *P*-wave arrivals at the more distant stations.

The historic event suite is run on a Linux server using a tank player module, provided as part of the Earthworm open-source software (Johnson *et al.*, 1995; Friberg *et al.*, 2010; see Data and Resources) to replay the test suite. The tank player allows previously recorded waveforms in tank-file format to be injected into the ShakeAlert software. One-second data packets from the tank file are placed into an Earthworm ring and processed as if they were being received in real time. The method employed during testing is a full replication of the real-time system behavior with the exception that telemetry latencies (generally ∼1–3 s) and alert-distribution latencies (unknown) are neglected. Thus, alerts generated by the replay system have lower overall latencies (from when the earthquake starts to when an alert is issued) than the real-time alerts. During each test, the output of the algorithms is logged and then analyzed to assess alert performance, as described in the Alert Assessment Parameters section.

NONDETERMINISTIC BEHAVIOR

Multithreading (execution of concurrent processes on a single CPU) is used to increase the computation speed of ShakeAlert algorithms. Multithreading ensures that codes rapidly analyze the high-volume real-time data packets and minimize the time to issue an alert. However, multithreading results in some minor nondeterministic behavior such that processing of an event may produce slightly different results each time the algorithm is tested, even under exactly the same processing conditions. Based on our examination of system log files, we repeatedly observe small variations in estimated alert parameters (magnitude, location, origin, and alert time) between tests.

The waveform processing component of the system uses multithreading to assign each station or channel to a different processing thread. The waveform processing calculations are deterministic in every run, that is, the picks made on each waveform dataset are identical in terms of pick time, pick amplitude, and pick ratio. However, multithreading may affect the order in which channels are processed such that the relative (logged) time at which the data are processed varies between runs, and the station order in which packets are processed varies.

The nondeterminism in the waveform processing changes the number and order of triggers available to the event associator at any particular time. As a result, the information used by the associator to calculate event parameters may differ. For example, if stations close to a new event hypocenter happen to be processed sooner than more distant stations, then the alert time will be slightly earlier than if the opposite were true, with

differences much less than the 1-s packet size. Generally, the effect of nondeterminism on alert parameters (magnitude, location, origin time, and alert time) is small as seen over repeated calculations, with values generally varying by less than 5%. However, the number of alerts issued from the event associator can sometimes change between test runs because simple time and distance criteria are used to determine whether a trigger should be associated with an existing event alert or whether a new event alert gets created.

To accommodate the nondeterministic behavior in the testing, our algorithm testing is undertaken in parallel with multiple, typically between four and eight, instances of the system. The results for each event and each run are logged and averaged over all runs per event. For more details on nondeterministic behavior and tests, see the Ⓔ electronic supplement.

ALERT ASSESSMENT PARAMETERS

Quantitative assessment of early warning system alerts can be complex because alerts can take different forms (e.g., point-source solutions, finite fault, etc.) with alerts updated rapidly as the earthquake evolves; additionally, assessments must consider both the timeliness of the alerts and the accuracy of earthquake source estimates and/or predicted ground motions. The ShakeAlert TCP assesses performance of the overall system alerts issued by the decision module and those generated by individual algorithms (e.g., onsite and Earthquake Alarm Systems [ElarmS]). The event logs generated during historic test suite runs are parsed to identify alerts that match with Advanced National Seismic System (ANSS) Comprehensive Catalog events. For each alert, we search for any events that occur up to 4 min prior to the alert and then undertake a series of comparisons to define whether alerts match to catalog events and to assess algorithm performance.

We implement two assessment methods that are based on point-source parameters and ground motions. The parameters critical to the assessment of a point-source alert are magnitude, epicenter, origin time, and time to first alert. Complementary to this, shaking intensities on a grid across the alert region are used to assess point-source solutions as well as accommodate finite fault and other non-point-source parameter solutions. Both assessment methods consider the alert latencies (i.e., how long after an earthquake starts an alert is issued, or conversely, whether an alert reaches a site prior to shaking exceeding a defined threshold). The alerts for a given event often include several updates to key parameters, in which later values presumably better reflect the final event magnitude and fault rupture extent than earlier values. Currently, the point-source assessments described below only make use of the first estimated parameters issued by the system. This approach is justified for moderate earthquakes for which the first alert is critical, but subsequent alert updates are likely to be issued with little or no time to be useful as an alert. The ground-motion-based assessment is undertaken by considering an alert and all its associated updates.

Point-Source Assessment

For the point-source assessment, the alert parameters examined are the magnitude, epicentral location, origin time, and alert latency. In the case of magnitude, epicenter, and origin time assessments, we compare against the ground-truth values from the ANSS Comprehensive Catalog that contains earthquake source parameters (e.g., magnitudes, hypocenters, and origin times) contributed by the regional seismic networks. Below we define thresholds for each parameter to declare a match between catalog events and an alert; these thresholds are chosen such that they associate alerts and events with broadly similar point-source parameters.

To measure goodness of magnitude M_g, we compare the first calculated magnitude of an alert M_a and the ground-truth magnitude obtained from the ANSS earthquake catalog M_c. Specifically, M_g is measured as a linearly scaled percentage that is a function of the absolute value of the difference between M_a and M_c ($\Delta M = \mathrm{abs}(M_a - M_c)$) and a maximum allowed magnitude difference $\Delta M_{\max}$,

$$M_g = \begin{cases} 100\% \times \left(\frac{\Delta M_{\max} - \Delta M}{\Delta M_{\max}}\right) & \text{for } \Delta M < \Delta M_{\max} \\ 0\% & \text{otherwise,} \end{cases} \quad (1)$$

in which $\Delta M_{\max}$ is defined to be 2.0.

To measure goodness of epicenter E_g, we examine the difference between the first estimated epicenter of an alert E_a and the ground-truth epicenter obtained from the ANSS catalog E_c. The distance between E_a and E_c (ΔD) is the great circle distance using an Earth radius value corresponding to the World Geodetic System 1984 ellipsoid reference model coordinate system. E_g is measured as a linearly scaled percentage that is a function of ΔD and a maximum allowed distance $\Delta D_{\max}$,

$$E_g = \begin{cases} 100\% \times \left(\frac{\Delta D_{\max} - \Delta D}{\Delta D_{\max}}\right) & \text{for } \Delta D < \Delta D_{\max} \\ 0\% & \text{otherwise,} \end{cases} \quad (2)$$

in which $\Delta D_{\max}$ is defined to be 100 km. We do not assess the depth estimate of an alert.

To measure goodness of origin time O_g, we examine the difference between the first origin time of an alert O_a and the ground-truth origin time obtained from the ANSS earthquake catalog O_c. O_g is measured as a linearly scaled percentage that is a function of the difference between O_a and O_c ($\Delta O = \mathrm{abs}(O_a - O_c)$) and a maximum-allowed origin time difference $\Delta O_{\max}$,

$$O_g = \begin{cases} 100\% \times \left(\frac{\Delta O_{\max} - \Delta O}{\Delta O_{\max}}\right) & \text{for } \Delta O < \Delta O_{\max} \\ 0\% & \text{otherwise,} \end{cases} \quad (3)$$

in which $\Delta O_{\max}$ is defined to be 15 s.

To assess the alert time T_a, we measure goodness of alert time T_g, that provides an estimate of how timely an alert is, given both network topology and the extent of moderate shaking during an event. First, we define the shortest alert time $T_{\min}$ as the average of the estimated P-wave travel times to

the four closest reporting stations. The travel times to each station are estimated using hypocentral distances and a 1D velocity model (iasp91; Kennett and Engdahl, 1991). The choice of four stations is used because the detection algorithms typically require four stations to issue an alert. Telemetry latencies are not included in our historic test suite datasets.

Second, we consider the longest alert time $T_{\max}$ that is used to assess when an alert is so late that it is no longer useful to end users of alert streams. Here, we consider alerts useful for any regions that experience shaking levels with modified Mercalli intensities (MMI) ≥ IV. MMI ≥ IV is chosen because it is typically considered to be the minimal shaking level likely to be felt and is sometimes associated with minor damage (Wood and Neumann, 1931; Richter, 1958; Dewey *et al.*, 1995). We use the *S*-wave travel time corresponding to the largest distance from the origin with MMI ≥ IV. Predicted ground motions are calculated assuming soil environment ($V_{S30} \leq 434$ m/s) for a range of distances using the ground-motion prediction equation (GMPE) of Cua and Heaton (2007). Then, the predicted peak ground acceleration and peak ground velocity values are converted into MMI values using Worden *et al.* (2012). The *S*-wave travel time to the maximum distance with MMI ≥ IV is then computed assuming $V_S = 3.5$ km/s. A schematic representation of the parameters T_a, $T_{\max}$, and $T_{\min}$ is shown in Figure 2.

T_g is given by

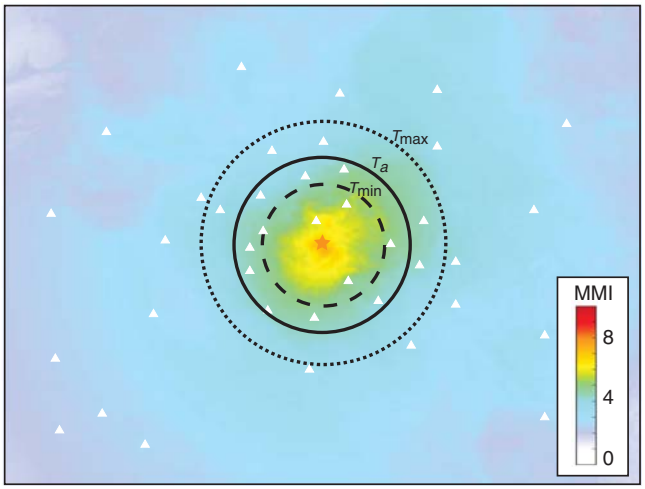
$$T_g = \begin{cases} 100\% \times \left(\frac{T_{\max} - T_a}{T_{\max} - T_{\min}}\right) & \text{for } T_a < T_{\max} \\ 0\% & \text{otherwise.} \end{cases} \quad (4)$$

It is possible for an algorithm to issue an alert faster than $T_{\min}$ if the alert is based on information from only the closest one or two stations or if our estimated *P*-wave travel times are slower than the actual *P*-wave travel times in a particular region. Therefore, it is possible for the T_g to be larger than 100%. It is also possible for $T_{\max}$ to be less than T_a in cases where the region of predicted ground shaking with MMI ≥ IV is small and/or where station density is low. This occurs in a very small number of test event cases where the epicenter was far from the nearest stations, for example, offshore earthquakes with the nearest reporting stations occurring at large distances. In this case T_g is assigned 0%.

Finally, we combine M_g, E_g, O_g, and T_g into an alert/event-specific assessment A_g:

$$A_g = 100\% \times [2/3(W_M \times M_g + W_E \times E_g + W_O \times O_g)/3 + 1/3(W_T \times T_g)], \quad (5)$$

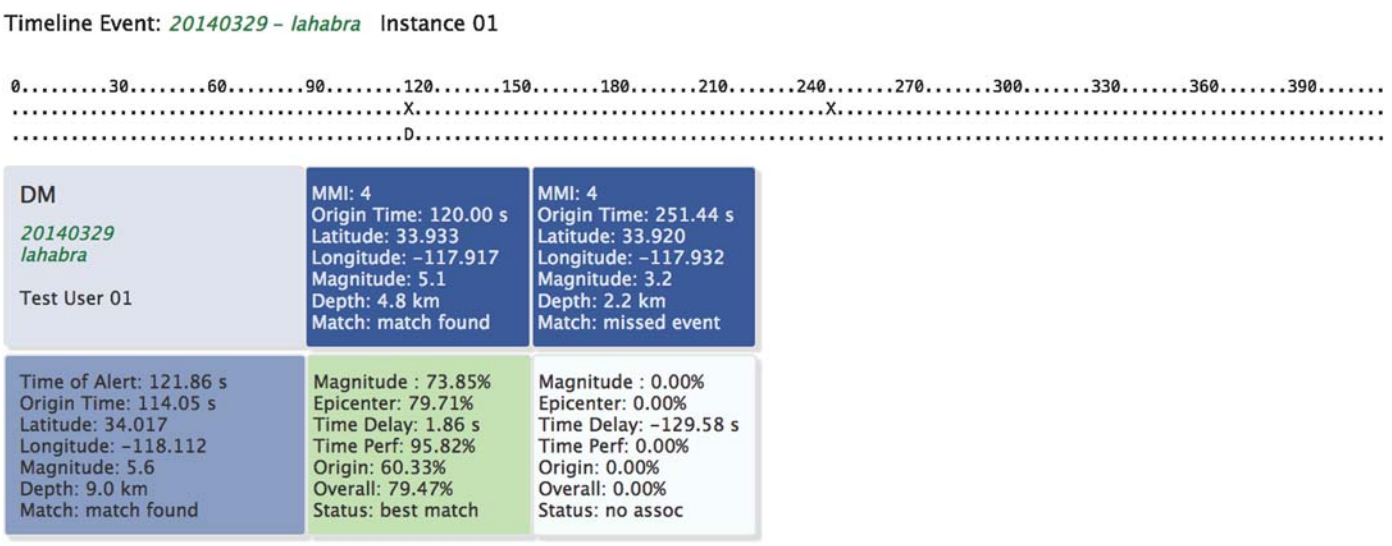
in which W_M is the weighting parameter for goodness of magnitude, W_E is the weighting parameter for goodness of epicenter, W_O is the weighting parameter for goodness of origin time, and W_T is the weighting parameter for goodness of time. Currently, $W_M = W_E = W_O = W_T = 1$, but this can be changed if higher or lower weighting of any parameter is desired.



▲ **Figure 2.** Schematic illustrating the parameters associated with alert time T_a, as described in the text. The dotted black circle is the distance corresponding to the *S*-wave travel time associated with modified Mercalli intensity (MMI) = IV ($T_{\max}$). The dashed black circle is the distance corresponding to the time for the *P* wave to reach the first four reporting stations ($T_{\min}$). The solid circle is the distance corresponding to the time that the alert is issued (T_a). White triangles show the hypothetical seismic stations. The orange star shows the hypothetical point-source epicenter. Background colors represent an example MMI distribution.

It is not always obvious which alert associates to which catalog event if there is more than one event within the 4-min window prior to the alert. To ensure that each alert is only associated with one event, we use the alert-event designation of *Best Match* for the largest A_g value among all possible alert-event pairs within the 4-min window and only if $M_g \neq 0\%$, $E_g \neq 0\%$, $O_g \neq 0\%$, and $T_g \neq 0\%$. If there is an event-alert association with $M_g \neq 0\%$, $E_g \neq 0\%$, and $O_g \neq 0\%$, but $T_g = 0\%$, then the alert is assigned as a *Best Match Not Useful* to designate that the alert was too late to be considered useful, even though the point-source parameter estimates were accurate. Possible causes for $T_g = 0\%$ are that the algorithm took too long to compute the alert or that the station distribution near the epicenter was too sparse to provide a timely alert.

An alert is defined as a *False Alert* if, within the 4-min window prior to the alert, there are no catalog events or if $A_g = 0\%$ for all alerts that are associated with an event. If two or more alerts are associated with a single catalog event, then only the alert with the largest A_g is assigned a *Best Match*, and all others are assigned as a *False Alert*. Additionally, if at least one of: $M_g = 0\%$, $E_g = 0\%$, or $O_g = 0\%$, even if $A_g \neq 0\%$, then this alert is designated as a *False Alert*. The alert may have associated with an event, but at least one of the point-source parameters was too poorly estimated to consider it a useful alert for end users. Finally, if the ANSS catalog shows that an earthquake occurred during the testing period, but the algorithm does not issue an alert, this is assigned as a *Missed Event*, and $A_g = 0\%$. Figure 3 shows an example of the point-source



▲ **Figure 3.** Sample test output from Testing and Certification Platform (TCP) analysis of system performance for the 2014 **M** 5.1 La Habra, California, earthquake. The timeline at the top indicates test duration time in seconds beginning 2 min before mainshock origin time; thus, the mainshock occurred at 120 s. The first row indicates times of all Advanced National Seismic System (ANSS) Comprehensive Catalog earthquakes (X symbols) within the western United States, and the second row indicates times of decision module (DM) alerts (D symbol) during test time period. The second X with no alerts below it indicates an aftershock for which no alert was issued, that is, a Missed event. In the boxes below the timeline, the top row (dark blue) indicates ANSS catalog source information for two western United States earthquakes that occurred during this test time period. The second row indicates DM alert parameter assessments resulting from event replay. The leftmost (medium blue) box shows DM alert parameters that are associated with the mainshock. The middle (green) box shows point-source assessment of the alert associated with the mainshock. The rightmost (white) box shows that the second ANSS event was not detected.

assessment applied to an alert generated during the 7-min waveforms from the 2014 **M** 5.1 La Habra, California, test event.

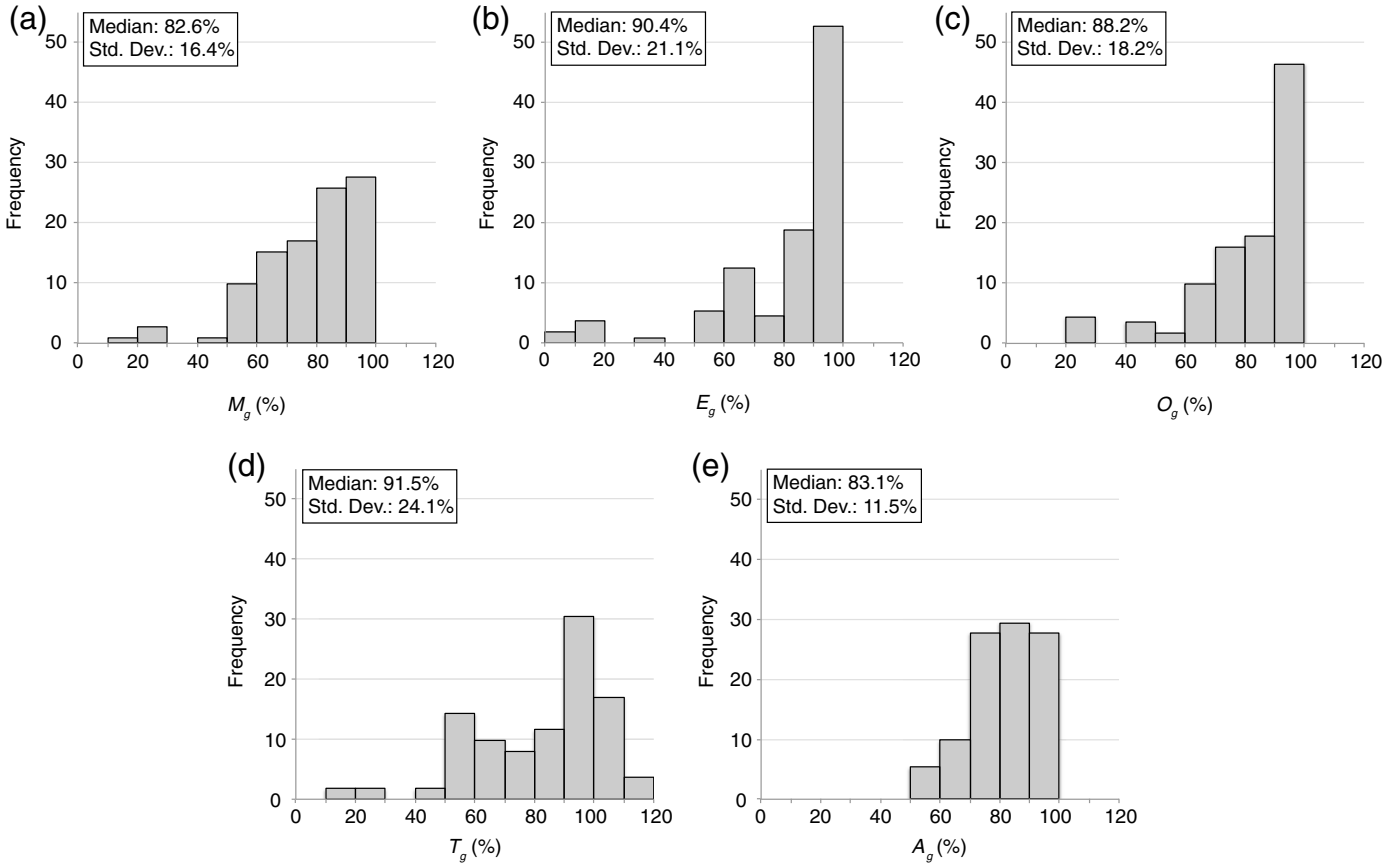
Using the performance assessment defined above, we describe the results of a recent TCP evaluation. In that test, we found that of the 40 historic mainshocks, 28 events were Best Match, 4 were Best Match Not Useful, and 8 events were Missed. In Figure 4, we show the M_g, E_g, O_g, T_g, and A_g for the 28 alerts designated as Best Match. For some alerts, O_g is greater than 100% because improvements made to the alert methods have allowed detections to be made with fewer than four stations. For this test, M_g has the lowest median value (82.6%), whereas O_g has the highest median value (91.5%). The combined individual alert metric (A_g) shows that the assessments tend to cluster between about 70% and 100%. The results presented here are assessments for one version of the alert software; the software is undergoing rapid development to improve performance, so these results are also expected to change (and further improve) through time.

We combine the results of individual events with a set of cumulative assessments. These cumulative performance measures are computed for magnitude bins 3.0–5.0, 5.0+, and 3.0+. Additionally, cumulative assessments are shown separately for local mainshock events (29 in California and 10 in Pacific Northwest) as well as for the entire test suite. There is flexibility in how (and which) events are used in the final assessments, depending on the goals of the test that are guided by what modifications were made to the code. For example, some code modifications may target certain types of events (e.g., local earthquakes with magnitude over 5.0) or suppression of teleseismic events, whereas others may be applicable to all waveform signals.

Cumulative assessments include the number of False Alerts, Missed Events, Best Match Alerts, Best Match Not Useful Alerts, Total Events, and Total Alerts. Additionally, we compute the following cumulative total definitions for all events across the four to eight instances of each event run:

- Overall Average Best Match Only = average of all A_g values for all events categorized as Best Match.
- Overall Average with Best Match Not Useful = average of all A_g values for all Best Match and Best Match Not Useful scores.
- Cumulative Average = average of all A_g values for all Best Match, Best Match Not Useful, and Missed Event scores. Missed events are given a score of 0%.
- Cumulative Average with False Alerts = average of all A_g values for all Best Match, Best Match Not Useful, and Missed Event scores, minus a 1% penalty for each False Alert. The False Alert penalty may lead to a negative percentage if there are numerous false alerts.

The parameters above provide an overall view of the algorithm performance. When these are combined with the detailed, event-specific results, they provide a valuable comparison between production and candidate versions of code and algorithm-based modules.



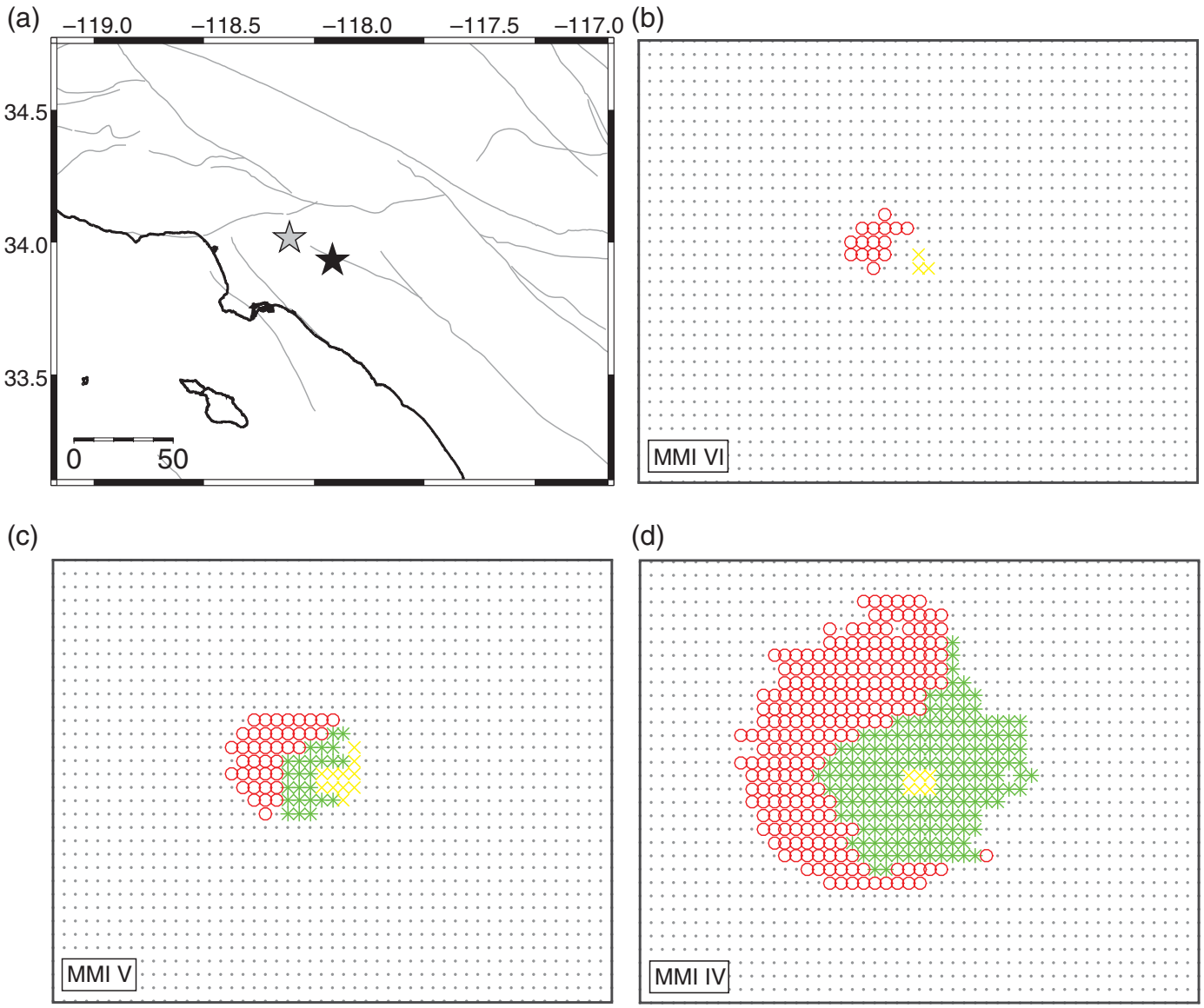
▲ **Figure 4.** Example of a suite of point-source assessments from replay of 40 historic mainshocks. Four instances of the alert algorithms were used in this run to account for system nondeterminism. In this run, there are 28 Best Match alerts providing a total of 112 alerts. Histograms are provided showing the distributions of (a) M_g, (b) E_g, (c) O_g, (d) T_g, and (e) A_g. The median and standard distributions are also provided in the upper left corner of each plot.

Ground-Shaking Intensity Predictions

Assessment of the accuracy of ground-motion predictions is an important additional evaluation tool within the TCP. The objective of an early warning system is to provide information about ground shaking at a user's site to mitigate damage, and ground motions are not controlled solely by the magnitude and location of the earthquake. Accurate prediction of ground motions at a site requires not only the hypocenter location and source magnitude but also the finite extent of the source fault, stress drop, path effects, and local site characteristics. We must assess how well a particular method is able to capture these effects to improve ground-motion predictions. Additionally, although the current ShakeAlert system only includes point-source-based alert methods, there are several methods under development that will predict additional source parameters, such as finite-fault length (e.g., Allen and Ziv, 2011; Böse *et al.*, 2012) or slip distribution (Grapenthin *et al.*, 2014; Minson *et al.*, 2014; Crowell *et al.*, 2016), as well as methods that predict ground motion directly from observed ground motion without requiring source information (e.g., Hoshiba, 2013; Hoshiba and Aoki, 2015; Kodera *et al.*, 2016). The point-source-only assessment tools described in the Point-Source Assessment section are insufficient for determining the performance of these new methods. However, all proposed methods ultimately can provide an estimate of predicted ground shaking across a region that can be assessed using a common approach.

We are developing a set of ground-motion-based assessments to augment the point-source-based metrics currently implemented in ShakeAlert TCP. The ground-motion-based assessments are only applied to the subset of alerts that are matched to a catalog event (Best Match and Best Match Not Useful). To assess the ground motions produced by ShakeAlert, we compare observed ShakeMaps, that is, near-real-time shaking intensity maps that are provided following significant earthquakes (Wald *et al.*, 1999), to shaking intensities predicted by the ShakeAlert system or an individual algorithm. The algorithms operating in the Production Prototype version of ShakeAlert currently only output point-source information products. Therefore, the TCP retrospectively produces maps of shaking intensity from alerts using the ShakeMap methodology (Worden and Wald, 2016). Ground-motion products are under development, and the ground-motion-based assessment tools described below will be used to assess these products directly in the future.

For each event, we reproduce the ground-truth ShakeMaps using archived ShakeMap input files, including finite-fault extent where available, on a specified grid. Similarly, we produce a



▲ **Figure 5.** Demonstration of the ground-motion assessment for an example test run of the 2014 **M** 5.1 La Habra, California, historic test suite event. (a) Location map showing the epicentral location of the La Habra earthquake (black star) and the location of the first alert issued by the DM (gray star). Performance is assessed at grid points with 0.05° spacing. Maps show the distribution of True Positive (green stars), False Positive (red circles), Missed (yellow crosses), and True Negative (gray dots) sites for ground-motion thresholds of (b) MMI = VI, (c) MMI = V, and (d) MMI = IV. Note that the latency does not consider communication delays between the stations and the central processing center (typically 1–3 s) or the time required to communicate messages to end users (unknown at present).

predicted ShakeMap using alert point-source information (magnitude and hypocentral location) on the same grid. The grid extent is chosen such that it covers all regions that are expected to exceed MMI II, for both the ground truth and ground motions predicted from alerts. At each grid point, we compare the MMI of the observed ShakeMap (MMI_{obs}) to the predicted MMI for the alert ShakeMap (MMI_{alert}). We compute a threshold-based assessment to determine if an end user would undertake a correct action given a particular MMI threshold for their site.

The methodology we apply is based on that developed by Meier (2017). The method compares observed with predicted ground motions at a site and classifies each site into one of four categories for a given MMI threshold as follows.

- True positive (TP): alert correctly predicts that ground motions will exceed a defined MMI threshold before the threshold is exceeded.
- False positive (FP): alert incorrectly predicts that ground motions will exceed a defined MMI threshold.
- True negative (TN): alert correctly predicts that ground motions will not exceed a defined MMI threshold.
- False negative (FN): alert incorrectly predicts that ground motions will not exceed a defined MMI threshold.

Here, we assume that the MMI threshold is exceeded at the time of the *S*-wave arrival, computed from the source-station hypocentral distance, and an *S*-wave velocity of 3.0 km/s. This assumption is generally adequate, but in a subset of cases, the MMI threshold could be exceeded before the *P*-wave arrival, particularly at close source-station distances and low MMI thresholds or after the *S*-wave arrival, particularly at greater distances (S. E. Minson *et al.*, unpublished manuscipt, 2017, see Data and Resources). Alternate ground-motion-based assessments could also be used, such as evaluation of the variance reduction between the predicted and observed ground motion or similar (e.g., Kodera *et al.*, 2016).

To undertake the threshold-based assessment, we compare each grid point between the ground truth and predicted ShakeMaps and classify each grid point using the definitions above. This is in contrast to Meier (2017), who compares predicted shaking to the observed ground-motion records only at locations with seismic stations. Here, we use the ShakeMap as ground truth for the events used in the test suite because available ground-motion observations may be limited, particularly for older historic events. More generally, the distribution of observations around an event could skew the interpretations if the stations are not uniformly distributed around the source, as is usually the case. However, we acknowledge that ground-truth ground motions estimated using the ShakeMap methodology may underestimate extreme ground motions at particular sites due to the smoothing applied. We calculate the ground truth and predicted ShakeMaps using the same GMPEs to reduce any bias that would occur from using different GMPEs. Figure 5 shows example threshold maps of TP, FP, TN, and FN sites for the 2014 La Habra earthquake for MMI IV–VI.

Following Meier (2017), we compute summary assessments from the suite of TP, FP, and FN values for each event and across the entire test suite. The summary assessments do not consider TN values because the number of TN sites depends on the distance over which sites are considered (in our case the size of the ShakeMap grid). We define the rate of true positive alerts as

$$\mathrm{TPRate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{8}$$

TPRate varies between 0 and 1. A TPRate of 1 indicates that all sites that exceed the MMI threshold were issued a timely and correct alert. A TPRate of 0 indicates that no alerts were provided for sites that exceeded the defined MMI threshold. Next, we define the rate of false positive alerts as

$$\mathrm{FPRate} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FN}}. \tag{9}$$

FPRate has a minimum value of 0, indicating that no sites with shaking levels below the MMI threshold were issued an alert. There is no absolute maximum value of FPRate, but this measure can be useful for quantitatively assessing the quality of an alert. For example, an FPRate of 2 means that an alert issued FP alerts at twice as many sites as the number of sites that experienced ground motions over a the defined MMI threshold.

Taken together, these two measures provide a comprehensive assessment of algorithm performance from the viewpoint of an end user by comparing with ideal performance defined by TPRate = 1 and FPRate = 0. The overall alert quality measured by determining the distance between the ideal case and the alert performance is given by

$$C_g = \sqrt{[(1 - \mathrm{TPRate})^2 + (\mathrm{FPRate})^2]}. \tag{10}$$

Again, these assessments can be applied to alerts generated for individual events or across the entire test suite to obtain a comprehensive assessment of system performance.

CONCLUSIONS

The purpose of the TCP is to provide a flexible platform for testing candidate algorithms, associated configuration files, modified operating system conditions, as well as new algorithms. The platform has been designed to be flexible so that components can be applied separately or together as in integrated test. The flexibility extends to the computing environment. The retrospective tests are conducted on a local server; however, it may be more desirable to undertake testing through a cloud instance because computing resources can be scaled as necessary, depending on the size of the test.

It is envisaged that new significant events within the SCSN, NCSS, and PNSN networks will be added to the database as they occur. Additional new test datasets that are being considered include datasets of locally recorded large earthquakes from around the world. For example, data from Japanese networks can be transformed to replicate the current ShakeAlert station configuration in California, Oregon, and/or Washington. In addition, synthetic waveforms from large-scenario events will be evaluated to determine if they can be used for testing. For example, the **M** 7.0 Hayward fault (Haywired) scenario earthquake models the impacts of an **M** 7.0 earthquake on the San Francisco Bay area (see Data and Resources); waveforms from this or other synthetic ruptures may be used in the future to test the sensitivity of algorithms to finite-fault behavior.

DATA AND RESOURCES

All test event data used by Testing and Certification Platform (TCP) are provided as miniSEED files and in tank file format at http://scedc.caltech.edu/research-tools/eewtesting.html (last accessed November 2017). The Earthworm open-source software is available at www.isti.com (last accessed November 2017). For Haywired website, see https://geography.wr.usgs.gov/science/mhdp/haywired.html (last accessed November 2017). The unpublished manuscript by S. E. Minson, M.-A. Meier, A. S. Baltay, T. C. Hanks, and E. S. Cochran, 2017, "The theoretical and observational limits of earthquake early warning: Timeliness of ground motion estimates", submitted to *Science Advances*.

ACKNOWLEDGMENTS

The authors wish to acknowledge the entire ShakeAlert research and development team, many of whom provided valuable discussions that helped inform the design of the system. The authors also thank internal U.S. Geological Survey (USGS) reviewers Sarah Minson and Glenn Biasi as well as two anonymous journal reviewers and Associate Editor Brendan Crowell for their helpful and insightful comments.

REFERENCES

Allen, R. M., and A. Ziv (2011). Application of real-time GPS to earthquake early warning, *Geophys. Res. Lett.* **38,** L16310, 1–7, doi: 10.1029/2011GL047947.

Böse, M., T. H. Heaton, and E. Hauksson (2012). Real-time finite fault rupture detector (FinDer) for large earthquakes, *Geophys. J. Int.* doi: 10.1111/j.1365-246X.2012.05657.x.

Crowell, B. W., D. A. Schmidt, P. Bodin, J. E. Vidale, J. Gomberg, J. R. Hartog, V. C. Kress, T. I. Melbourne, M. Santillan, S. E. Minson, *et al.* (2016). Demonstration of the Cascadia G-FAST geodetic earthquake early warning system for the Nisqually, Washington, earthquake, *Bull. Seismol. Soc. Am.* **87,** no. 4, doi: 10.1785/0220150255.

Cua, G., and T. Heaton (2007). The virtual seismologist (VS) method: A Bayesian approach to earthquake early warning, in *Earthquake Early Warning Systems*, P. Gasparini, G. Manfredi, and J. Zschau (Editors), Springer, Berlin, Germany, ISBN 13 978-3-540-72240-3.

Dewey, J. W., B. G. Reagor, L. Dengler, and K. Moley (1995). Intensity distribution and isoseismal maps for the Northridge, California, earthquake of January 17, 1994, *U.S. Geol. Surv. Open-File Rept. 95-92*, 35 pp.

Erdik, M., Y. Fahjan, O. Ozel, H. Alcik, A. Mert, and M. Gul (2003). Istanbul earthquake rapid response and early warning system, *Bull. Earthq. Eng.* **1,** no. 1, 157–163.

Espinosa Aranda, J. M., A. Jimenez, G. Ibarrola, F. Alcantar, A. Aguilar, M. Inostroza, and S. Maldonado (1995). Mexico City seismic alert system, *Seism. Res. Lett.* **66,** no. 6, 42–53, doi: 10.1785/gssrl.66.6.42.

Friberg, P., S. Lisowski, I. Dricker, and S. Hellman (2010), Earthworm in the 21st century, *Geophys. Res. Abstr.* **12,** Abstract, EGU 2010–12654.

Given, D., E. Cochran, T. Heaton, E. Hauksson, R. Allen, M. Hellweg, J. Vidale, and P. Bodin (2014). Technical implementation plan for the ShakeAlert Production Prototype system—An earthquake early warning system for the West Coast of the United States, *U.S. Geol. Surv. Open-File Rept. 2014–1097*, 25 pp.

Grapenthin, R., I. A. Johanson, and R. Allen (2014). Operational real-time GPS-enhanced earthquake early warning, *J. Geophys. Res.* **119,** doi: 10.1002/2014JB011400.

Hoshiba, M. (2013). Real-time prediction of ground motion by Kirchhoff-Fresnel boundary integral equation method: Extended front detection method for earthquake early warning, *J. Geophys. Res.* **118,** 1038–1050, doi: 10.1002/jgrb.50119.

Hoshiba, M., and S. Aoki (2015). Numerical shake prediction for earthquake early warning: Data assimilation, real-time shake mapping, and simulation of wave propagation, *Bull. Seismol. Soc. Am.* **105,** no. 3, 1324–1338, doi: 10.1785/0120140280.

Hoshiba, M., H. Negishi, K. Abe, A. Kaminuma, and Y. Fujinawa (2008). Earthquake early warning starts nationwide in Japan, *Eos Trans.* **89,** no. 8, 73–74.

Hsiao, N.-C., Y.-M. Wu, T.-C. Shin, L. Zhao, and T.-L. Teng (2009). Development of earthquake early warning system in Taiwan, *Geophys. Res. Lett.* **36,** L00B02, 1–5, doi: 10.1029/2008GL036596.

Johnson, C. E., A. Bittenbinder, B. Bogaert, L. Dietz, and W. Kohler (1995). Earthworm: A flexible approach to seismic network processing, *Incorporated Research Institutions for Seismology (IRIS) Newsletter* Vol. 14, 1–4.

Kennett, B. L. N., and E. R. Engdahl (1991). Traveltimes for global earthquake location and phase identification, *Geophys. J. Int.* **105,** 429–465.

Kodera, Y., J. Saitou, N. Hayashimoto, S. Adachi, M. Morimoto, Y. Nishimae, and M. Hoshiba (2016). Earthquake early warning for the 2016 Kumamoto earthquake: Performance evaluation of the current system and the next-generation methods of the Japan Meteorological Agency, *Earth Planets Space* **68,** no. 202, doi: 10.1186/s40623-016-0567-1.

Kohler, M. D., E. S. Cochran, D. Given, S. Guiwits, D. Neuhauser, I. Henson, R. Hartog, P. Bodin, V. Kress, S. Thompson, *et al.* (2017). Earthquake early warning ShakeAlert system: West coast wide Production Prototype, *Seismol. Res. Lett.* doi: 10.1785/0220170140.

Meier, M.-A. (2017). How "good" are real-time ground motion predictions from earthquake early warning systems, *J. Geophys. Res.* **122,** 5561–5577, doi: 10.1002/2017JB014025.

Minson, S. E., J. R. Murray, J. O. Langbein, and J. S. Gomberg (2014). Real-time inversions for finite fault slip models and rupture geometry based on high-rate GPS data, *J. Geophys. Res.* **119,** 3210–3231, doi: 10.1002/2013JB010622.

Richter, C. F. (1958). *Elementary Seismology*, W. H. freeman and Co., San Francisco, California.

Wald, D. J., V. Quitoriano, T. H. Heaton, H. Kanamori, C. W. Scrivner, and C. B. Worden (1999). TriNet "ShakeMaps": Rapid generation of peak ground motion and intensity maps for earthquakes in southern California, *Earthq. Spectra* **15,** no. 3, 537–555.

Wood, H. O., and F. Neumann (1931). Modified Mercalli intensity scale of 1931, *Bull. Seismol. Soc. Am.* **21,** 277–283.

Worden, C. B., and D. J. Wald (2016). ShakeMap manual online: Technical manual, user's guide, and software guide, *U.S. Geological Survey*, doi: 10.5066/F7D21VPQ.

Worden, C. B., M. C. Gerstenberger, D. A. Rhoades, and D. J. Wald (2012). Probabilistic relationships between ground-motion parameters and modified Mercalli intensity in California, *Bull. Seismol. Soc. Am.* **102,** no. 1, 204–221, doi: 10.1785/0120110156.

Elizabeth S. Cochran
Douglas D. Given
Stephen Guiwits
Mohammad Ahmad
Deborah Smith
U.S. Geological Survey
525 South Wilson Avenue
Pasadena, California 91106 U.S.A.
ecochran@usgs.gov

Monica D. Kohler
Jennifer Andrews
Men-Andrin Meier
California Institute of Technology
1200 East California Boulevard
Pasadena, California 91125 U.S.A.

Ivan Henson
University of California
McCone Hall, 215 Haviland, Path #4760
Berkeley, California 94720 U.S.A.

Renate Hartog
University of Washington
4000 15th Avenue NE
Seattle, Washington 98195 U.S.A.

Published Online 6 December 2017