AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Solid Earth

RESEARCH ARTICLE
10.1002/2017JB014025

Key Points:
- We suggest objective performance metrics for Earthquake Early Warning (EEW) systems that jointly quantify alert timeliness and correctness
- Metrics directly reflect usefulness of EEW alerts to end users and facilitate meaningful algorithm comparison
- They reveal both potential and limitations of algorithms and can show whether finite source algorithms are fast enough

Supporting Information:
- Supporting Information S1
- Data Set S1
- Data Set S2

Correspondence to:
M.-A. Meier,
mmeier@caltech.edu

Citation:
Meier, M.-A. (2017), How "good" are real-time ground motion predictions from Earthquake Early Warning systems?, *J. Geophys. Res. Solid Earth*, *122*, 5561–5577, doi:10.1002/2017JB014025.

Received 27 JAN 2017
Accepted 8 JUL 2017
Accepted article online 1 JUL 2017
Published online 31 JUL 2017

How "good" are real-time ground motion predictions from Earthquake Early Warning systems?

Men-Andrin Meier[1]

[1]Seismological Laboratory, California Institute of Technology, Pasadena, California, USA

Abstract Real-time ground motion alerts, as can be provided by Earthquake Early Warning (EEW) systems, need to be both timely and sufficiently accurate to be useful. Yet how timely and how accurate the alerts of existing EEW algorithms are is often poorly understood. In part, this is because EEW algorithm performance is usually evaluated not in terms of ground motion prediction accuracy and timeliness but in terms of other metrics (e.g., magnitude and location estimation errors), which do not directly reflect the usefulness of the alerts from an end user perspective. Here we attempt to identify a suite of metrics for EEW algorithm performance evaluation that directly quantify an algorithm's ability to identify target sites that will experience ground motion above a critical (user-defined) ground motion threshold. We process 15,553 recordings from 238 earthquakes with $M > 5$ (mostly from Japan and southern California) in a pseudo-real-time environment and investigate two end-member EEW methods. We use the metrics to highlight both the potential and limitations of the two algorithms and to show under which circumstances useful alerts can be provided. Such metrics could be used by EEW algorithm developers to convincingly demonstrate the added value of new algorithms or algorithm components. They can complement existing performance metrics that quantify other relevant aspects of EEW algorithms (e.g., false event detection rates) for a comprehensive and meaningful EEW performance analysis.

1. Introduction

The main goal of Earthquake Early Warning (EEW) systems is to detect ongoing earthquakes in real time and to provide information about the earthquake to target sites before they are hit by strong ground motion [*Heaton*, 1985]. If such real-time alerts are timely and sufficiently accurate, they allow EEW end users to trigger a wide range of protective emergency actions that can greatly reduce earthquake damage [e.g., *Strauss and Allen*, 2016]. Operating EEW systems have already provided useful public warnings during several large earthquakes, including the 2016 M_w 7.0 Kumamoto, Japan [*Kodera et al.*, 2016], the 2011 M_w 9.0 Tohoku, Japan [*Fujinawa and Noda*, 2013], and the 2012 M_w 7.4 Oxaca, Mexico, earthquakes [*Cuéllar et al.*, 2014].

At the same time it is clear that EEW systems have limitations: alerts may not always be fast enough, in particular at near-epicentral sites where ground motion is typically strongest and begins shortly after the event origin. Overpredictions of ground motion may lead to false alerts [e.g., *Yamasaki*, 2012], motivating end users to unnecessarily carry out emergency actions that can be costly and/or dangerous. Underpredictions, on the other hand, can lead to missed alerts, i.e., failing to notify end users to take action.

Unfortunately, we often know relatively little about how often and under what circumstances different algorithms succeed or fail to alert. In part, this is because the performance of EEW algorithms is not usually evaluated in terms of how accurately and how timely they can predict impending ground motion but in terms of how well they can characterize the earthquake source (magnitude, location, origin time, finite source, etc.) [e.g., *Meier et al.*, 2015; *Minson et al.*, 2014; *Kuyuk and Allen*, 2013]. While, e.g., magnitude estimates are indeed proxy metrics for ground motion strength (larger earthquakes produce stronger ground motion), it is not straightforward to estimate how errors in source characterization map into ground motion prediction errors.

Furthermore, studies that do evaluate the algorithms' ability to predict ground motion often neglect the crucial aspect of alert timeliness [e.g., *Xu et al.*, 2017; *Kodera et al.*, 2016; *Colombelli et al.*, 2012; *Allen*, 2007]. They thereby demonstrate that the predictions eventually match the observed ground motion levels, but not whether sufficiently accurate predictions became available before the strong shaking started. Studies that consider both accuracy and timeliness are rare, although there are notable exceptions [*Hsu et al.*, 2016; *Colombelli et al.*, 2015; *Hoshiba and Aoki*, 2015; *Böse et al.*, 2012].

With the numerous different approaches to EEW that have been proposed to date it would be desirable to have a set of meaningful and standardized performance metrics that bring out both the strengths and weaknesses of individual approaches. This would help to (i) inform both EEW end users and algorithm developers what EEW algorithms can and cannot do; (ii) meaningfully compare algorithms, to choose between algorithms and to combine algorithms with different strengths; and (iii) identify areas where algorithm improvements are most likely to pay off.

In this study we try to identify evaluation metrics that directly reflect the usefulness of EEW alerts to its users. We focus on the two attributes that are most important for users: alert timeliness and correctness. How often can timely alerts be provided and for what kind of ground motion? How often and under what circumstances do they come too late? How much warning time do the alerts typically come with? And how many false and missed alerts do end users have to put up with?

We use a large offline waveform data set to simulate a real-time environment. We quantify the ground motion prediction performance of two hypothetical EEW algorithms: a simple point source algorithm that represents established operational systems and an idealized finite source algorithm that can be considered an upper bound for the performance that EEW systems can possibly reach. In section 2 we describe how the two algorithms use the waveform data to make probabilistic real-time ground motion predictions. Section 3 describes the data set used for the analysis. In section 4 we evaluate the predictions by measuring (i) the prediction error as a function of warning time for individual sites and (ii) the resulting alerting performance for different ground motion thresholds in terms of warning times and ratios of correct and false alerts. In section 5 we discuss the implications of the insights we have gained through the employed end user perspective performance metrics.

2. Method

2.1. Probabilistic Real-Time Ground Motion Predictions

We evaluate the real-time ground motion predictions from two methods, using offline waveform data in a pseudo-real-time environment. We assume that an EEW system is constantly monitoring a region of interest, accurately detects incoming earthquake signals, and associates them to the correct event. As soon as an event is detected, i.e., as soon as the first station registers the *P* wave, the event can be characterized (albeit with initially large uncertainties) and ground motions can be predicted for the entire surrounding region. For each earthquake, we start making ground motion predictions for all sites within 200 km of the catalog hypocenter as soon as the first *P* wave is registered and update them in 0.5 s intervals (Figure 1a). For the sake of simplicity we neglect all delays (data transmission, processing, etc.) [*Behr et al.*, 2015].

The first method, in the following referred to as "simple point source" or "SPS" method, is similar to the methods implemented in the current ShakeAlert project [*Böse et al.*, 2014]. For each station at which the earthquake has been registered we make a magnitude estimate based on initial waveform observations. We use the regression relation of *Kuyuk and Allen* [2013] between peak-observed displacement at individual stations, $P_{d,i}(t)$, to estimate a probability density function for magnitude $p(\hat{m}_i, t)$.

$$p(\hat{m}_i, t) = 1.23 \log_{10}\left[P_{d,i}(t)\right] + 1.38\, p(\log_{10}[\hat{r}_i(t)]) + 5.39 + N\left(0, \sigma^2\right).$$

$P_{d,i}(t)$ is the peak absolute displacement amplitude in centimeters measured over the first 4 s since the *P* wave arrival at the *i*th station or over whatever waveform segment is available at point in time *t*. If the direct *S* phase arrives within 4 s of the *P* onset, we only measure $P_{d,i}(t)$ up to the *S* arrival minus 0.5 s. We use $\sigma^2 = 0.31^2$ from *Kuyuk and Allen* [2013], and we assume all magnitudes to be moment magnitudes throughout this paper. $p(\log_{10}[\hat{r}_i(t)])$ is a probabilistic source/station distance estimate. We simulate real-time location errors by perturbing the catalog epicenters with normally distributed errors, with standard deviations that decrease as the number of reporting stations increases (30, 20, 10, 8, 6, and 5 km for 1, 2, 3, 4, 5, and >6 stations, respectively). These values were chosen to approximately mimic the location errors of the operational ShakeAlert test system which has been performing real-time locations since the year 2012 [*Böse et al.*, 2014] and for which the real-time performance has been documented in log files. We assume a default earthquake depth of 8 km. For each station we map the location estimate $p\left(\hat{x}(t)\right)$ into a hypocentral distance estimate $p(\log_{10}\hat{r}_i(t))$.

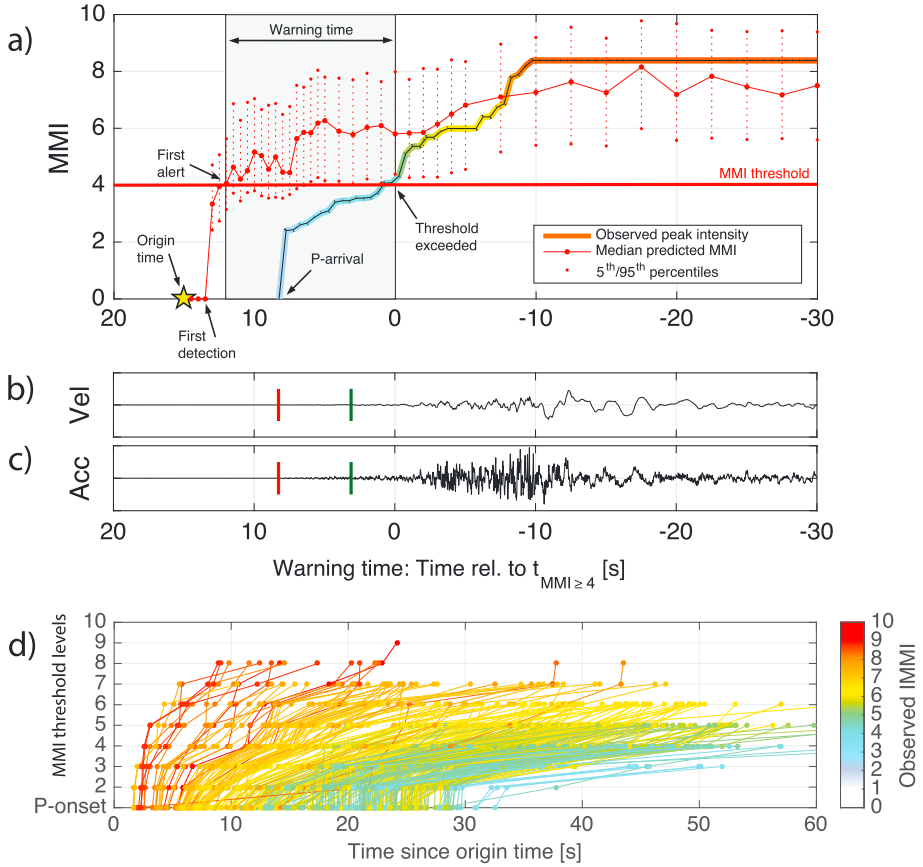

Figure 1. (a) Observed shaking intensity evolution and probabilistic ground motion prediction from the SPS method as a function of warning time for a site at 40 km from the hypocenter of the 1999 M_w 7.6 Chi Chi, Taiwan, earthquake. A few seconds after the origin time (yellow star), the *P* wave reaches the closest station to the epicenter, after which ground motion predictions are made for all sites within 200 km. An alert is declared when the probability of MMI $>$ 4 exceeds 50% (cf. section 4.3). This happens for the first time after three prediction updates, which are computed at intervals of 0.5 s. A few seconds later, the *P* wave reaches the site, but MMI 4 is not exceeded until 8 s later. Warning time is defined as the time from when the alert is declared until the ground motion threshold is reached; if an alert is declared after this, it is considered too late ("false negative"). Even though eventual peak ground motion is strongly underpredicted, a correct alert is given out with a warning time of ~12 s. (b) Corresponding vertical velocity and (c) acceleration waveforms with *P* (red line) and *S* (green line) phase arrivals. (d) Time of first exceedance of different MMI thresholds for sites around the Chi Chi earthquake. For each site the line shows how soon after origin time the site was reached by the direct *P* phase and how soon different intensity thresholds were crossed. Near-epicentral sites reach extreme ground shaking levels in a matter of seconds. How fast a given intensity threshold is reached directly affects the available warning time at a site.

We then compute a multistation magnitude probability by multiplying the $p(\hat{m}_i, t)$ from all stations that have recorded the earthquake at time *t*. Note that the magnitude estimates from different stations are not fully independent in a strict sense, since the records share a common source term and since the seismic waves at different stations may in part sample a common part of the seismic velocity structure. This may lead to a slight underestimation of magnitude uncertainties. However, previous work has shown that this effect is small and that combining probabilistic magnitude estimates in this way does not lead to a bias in magnitude estimates [*Meier et al*., 2015].

The second method, in the following referred to as "ideal finite source" or "IFS" method, represents the unrealistic ideal case in which the finite source is perfectly characterized as soon as the first station has detected the *P* wave; i.e., we assume to know catalog magnitude and finite source distances instantly upon detection of the earthquake.

For both methods we then compute probabilistic ground motion predictions with a Monte Carlo approach: for each site we compute distributions of peak ground velocity $p(\mathrm{PGV}_i, t)$ and peak ground acceleration

$p(\mathrm{PGA}_i, t)$ by computing 5e3 outcomes of a ground motion prediction equation (GMPE). We use a modified version of the ground motion prediction equation of *Cua and Heaton* [2009] (cf. supplementary information and comparison to equations of *Boore et al*. [2014]). For the IFS method we use the catalog magnitude and finite source distances as predictors; i.e., the spread of the peak amplitude distribution comes solely from the random error term of the GMPE. For the SPS method we randomly sample the predictors from their estimated distributions, $p(\hat{m}_i, t)$ and $p(\log_{10}\hat{r}_i(t))$, and thus propagate their uncertainties to the ground motion prediction.

We then transform the $p(\mathrm{PGV}_i, t)$ and $p(\mathrm{PGA}_i, t)$ distributions into distributions of predicted instrumental intensities $p(\mathrm{MMI}_i, t)$ using the relations of *Worden et al*. [2012]. Similarly, we translate the observed peak ground motion amplitudes from all waveforms into MMI observations. For each site we use the peak amplitude of the vector sum of the two horizontal components. The predictions are updated at 0.5 s intervals (Figures 1a–1c). For each waveform we also measure at what time different MMI levels are exceeded for the first time (Figure 1d).

The SPS and the IFS methods represent two end-members for real-time ground motion prediction: what can be done based on currently operational point source algorithms (SPS) and an ideal case with perfect (and unrealistic) knowledge of the source (IFS). Note, however, that the GMPs of the IFS method are not necessarily ideal because they could be further improved with more accurate and precise GMP models.

2.2. Performance Evaluation as Function of Warning Time Δt_w

The usefulness of an EEW alert at a target site depends on how many seconds the alert arrives before the onset of strong ground motion at that site. The absolute time of strong ground motion onset varies strongly with hypocentral distance. The time since origin time is therefore a poor measure for the "urgency" of an EEW alert. An alert that becomes available at, say, 10 s after origin time may already be too late for near-epicentral sites, while end users at more distant sites still have plenty of time to take action. In order to evaluate how useful a given ground motion prediction can be from an end user perspective, we therefore analyze the predictions as a function of warning time at each individual site, rather than as a function of time since origin time. We define warning time, $\Delta t_{w,i}$, as the time between when an alert is declared for a site and the time when a threshold MMI level is first exceeded (Figure 1d). This time may vary strongly depending on the chosen threshold MMI level and between individual earthquakes.

3. Data

We perform the real-time ground motion predictions for a set of 15,553 recordings from 238 earthquakes with magnitudes >5 and hypocentral distances <200 km (Figure 2). The data set is composed of (i) the waveforms of all shallow crustal onshore events with $M_{\mathrm{JMA}} \geq 5$ recorded by Japanese strong motion networks K-NET and KiK-net since their beginning in 1997 until June 2015, as well as those of the 2016 M_w 7.0 Kumamoto earthquake; (ii) all strong motion and broadband records from the Southern California Seismic Network with $M_L \geq 4$ ("SCSN") since 1990 until June 2015; and (iii) the four large earthquakes from the Next Generation Attenuation West 1 data set [*Chiou and Youngs*, 2008] for which *P* wave onsets are contained in the recordings and for which a sufficient number of local recordings are available: the 1995 M_w 6.9 Kobe, Japan (9 records), the 1999 M_w 7.6 Chi-Chi, Taiwan (355 records), the 1999 M_w 7.6 Kocaeli, Turkey (13 records), and the 1999 M_w 7.2, Duzce, Turkey (18 records), earthquakes. A list of all used records is provided as supporting information.

The goal is to get a detailed picture of the *typical* ground motion prediction performance of the two algorithms: What prediction accuracy can the algorithms provide at a certain warning time? And how does this translate into numbers of correct and incorrect alerts?

With the exception of the NGA records, the data set contains all earthquake records in the described magnitude, space, and time windows. These simple selection criteria lead to a natural and representative distribution of ground motion observations at randomly selected sites. In particular, the data set reflects both the fact that lower magnitude events are much more common than large ones [*Gutenberg and Richter*, 1956] and the fact that larger source-site distances are more common (for geometrical reasons). The EEW performance statistics inferred with this data set should therefore be representative of the performance an end user can expect, unless there is reason to believe that seismic hazard at the site of interest has nonaverage properties,

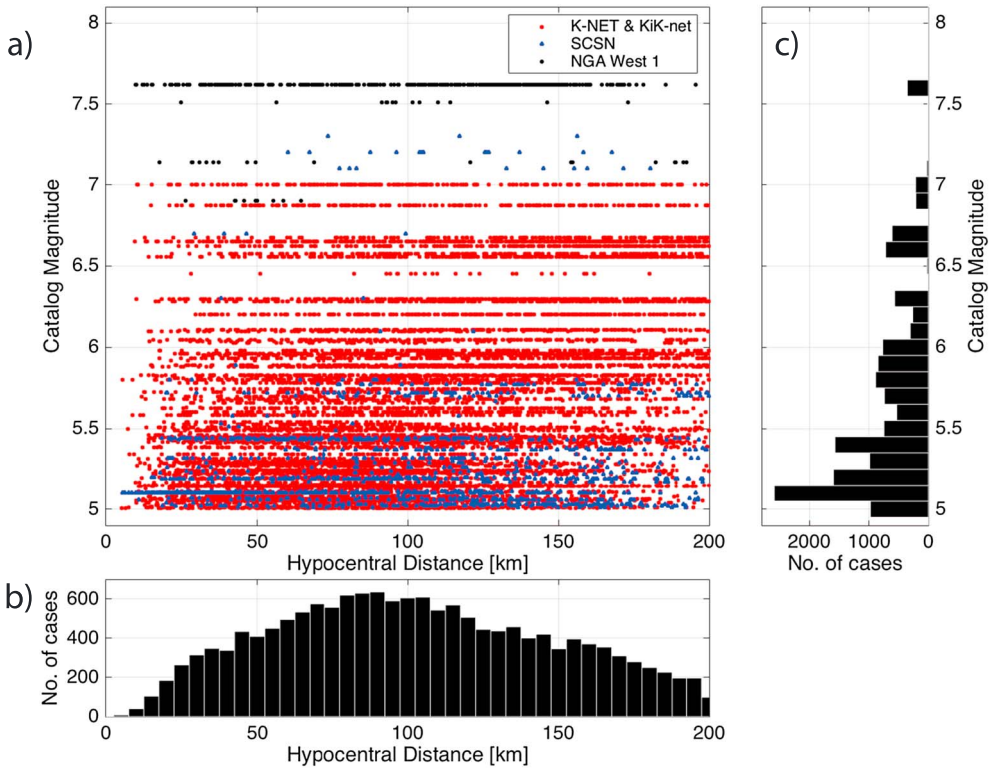

Figure 2. Data set overview. (a) Hypocentral distances and magnitudes of all records from the three data sources from Japan, southern California, and NGA West 1. Histograms of (b) hypocentral distances and (c) magnitudes.

e.g., because it is located close to a large fault with a distinct frequency magnitude distribution [e.g., *Ben-Zion*, 2008]. The contribution of the largest earthquakes may be somewhat misrepresented because the time window that the data set samples is short relative to the typical recurrence time of $M > 7$ events. This is counteracted, however, by including the four large events of the NGA West 1 data set.

4. Results

4.1. Ground Motion Prediction Error as a Function of Warning Time

We define the ground motion prediction error at the ith site as $\Delta\mathrm{MMI}_i(t) = \mathrm{MMI}_{i,\mathrm{obs}} - \widehat{\mathrm{MMI}}_i(t)$, the difference between the maximum observed intensity at the site, $\mathrm{MMI}_{i,\mathrm{obs}}$, and the most likely intensity prediction at each point in time, $\widehat{\mathrm{MMI}}_i(t) = \mathrm{maxarg}_p\, p(\mathrm{MMI}_i(t))$. Having measured the time at which the MMI threshold of interest is exceeded at each site, we can analyze $\Delta\mathrm{MMI}_i$ as a function of warning time, i.e., $\Delta\mathrm{MMI}_i(\Delta t_{w,i})$ (cf. Figure 1a).

Figure 3 shows $\Delta\mathrm{MMI}_i(\Delta t_{w,i})$ from the SPS and the IFS methods for the 15,553 records of the data set, using MMI 4 as a threshold to measure warning times. For the sake of clarity we subdivide the predictions into eight bins according to their observed final peak MMI. This gives a detailed impression of the algorithms' typical ability to predict different ground motion intensity levels. It also includes the information of the warning time with which a given alert becomes available: Recall that for each earthquake, the first ground motion predictions for all sites are made when the nearest station to the epicenter is reached by the *P* wave. This shows up in the form of near-vertical onsets of the curves in Figure 3 when the predictions go from 0 to a, hopefully, more realistic value. The near-vertical onsets therefore represent maximum possible warning times, i.e., the warning times that would be available if the first ground motion prediction led to an alert. To summarize these observations, and to directly compare the two algorithms, we can then analyze the median prediction error in each intensity bin (solid black lines and colored lines in bottom plots).

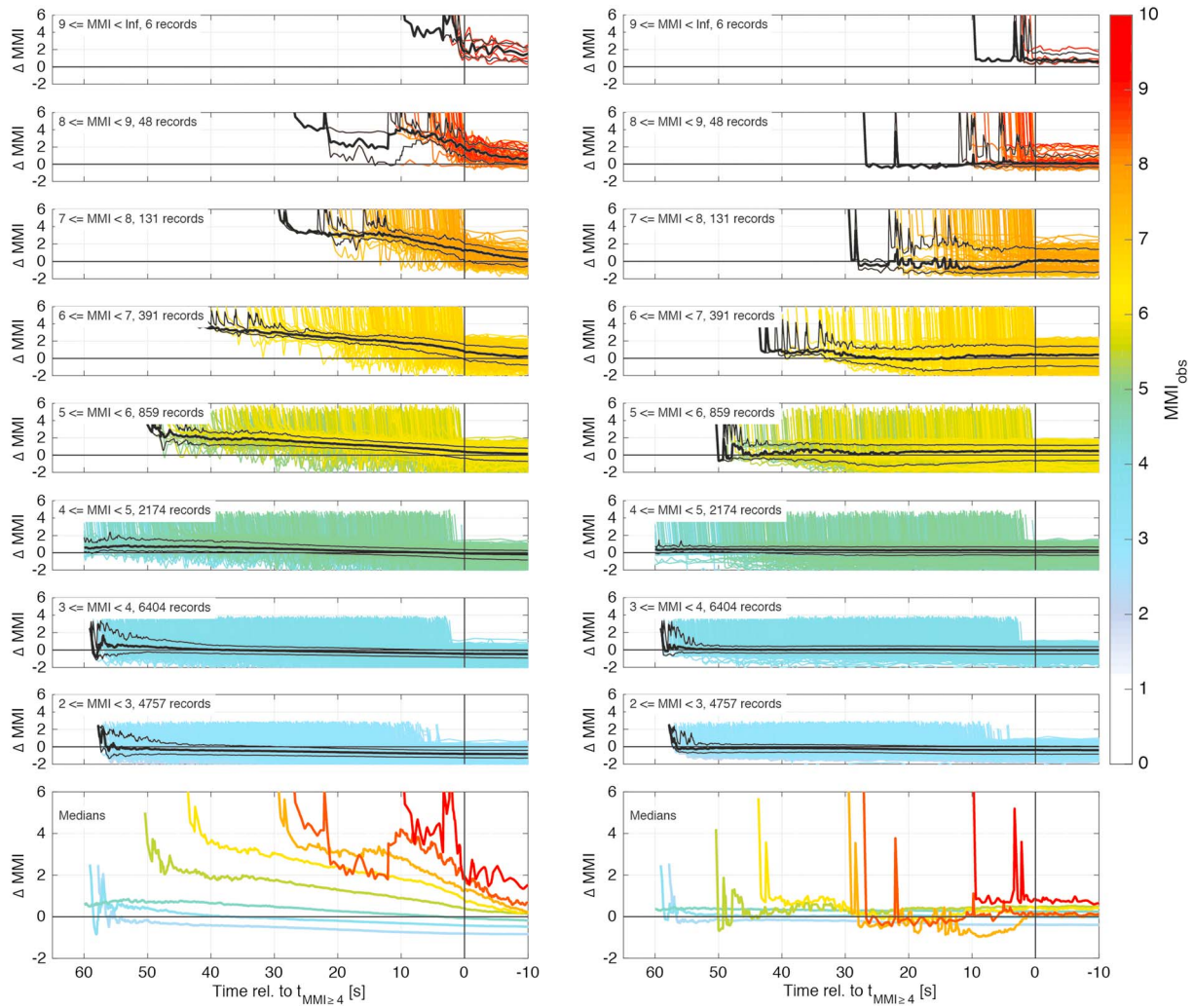

Figure 3. Ground motion prediction errors $\Delta MMI_i(\Delta t_{w,i})$ as a function of warning time for (left column) the SPS method and (right column) the IFS method for all records of the data set, divided into eight MMI bins. Warning times are relative to when MMI 4 is first exceeded at individual sites or for sites with MMI $<$ 4, relative to the theoretical *S* arrival. The lines are color coded with respect to the observed peak MMI value at each site. Black lines give the 16th, 50th, and 84th percentiles at each point in time. The medians (thick black lines) are shown again for direct comparison in the subfigures at the bottom.

These figures reveal that (i) the SPS method systematically and strongly underpredicts high-amplitude ground motion, (ii) ground motion could be, on average, accurately predicted if magnitude and distances were known (IFS method) albeit with large scatter, and (iii) warning times are invariably short for high-intensity ground motion sites, while lower intensity sites come with a wide range of warning times. Here "accurate" means that the predictions are not biased with respect to the observed ground motion, i.e., that the predictions are on average correct. The limited precision, on the other hand, i.e., the scatter around the average prediction, may lead to false and missed alerts.

Weak and light intensities (MMI 2–4) are on average accurately predicted from the first alert on to within ±0.5 MMI units, although overpredictions of 2 MMI units occur at times. Such overpredictions may cause false alerts at those sites. This general variability is observed with both SPS and IFS methods and corresponds to the aleatory variability of ground motion predictions. Warning times of up to a minute are possible in cases where large earthquakes cause perceivable ground motions even at large distances, because of the time it takes for the shaking to arrive at those sites. The same level of ground motion, however, can also be caused by smaller events, at shorter distances, in which case maximum warning times can be short.

Moderate and strong intensities (MMI 5–6), on the other hand, are systematically underpredicted by the SPS method, and the level of underprediction decreases toward shorter warning times. This reflects the fact that

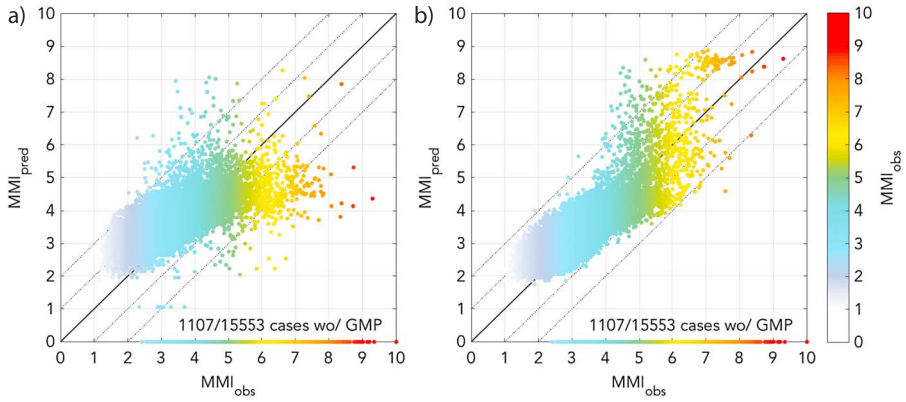

Figure 4. Predicted versus observed ground motion intensities for (a) the SPS and (b) the IFS methods at $\Delta t_{w,i} = 8$ s before MMI 4 is first exceeded at individual sites or for sites with MMI < 4, before the theoretical *S* arrival. Dotted lines delineate a prediction error of ±1 and ±2 MMI units. Data falling on the abscissa are sites from earthquakes that have not yet been detected at $\Delta t_{w,i} = 8$ s, and therefore, no ground motion prediction is available.

large magnitudes are initially underestimated, and it highlights the need to update the alerts from point source algorithms, since the first alerts may fail to alert many of the affected sites. The scatter around the median is on the order of 0.5–1 MMI units, as shown by the 16th and 84th percentiles in Figure 3.

For the sites with very strong to extreme ground motion (MMI ≥ 7) the underprediction is even stronger. Many of these sites have ground motion predictions of only MMI ~ 4–5 until shortly before the threshold of MMI 4 is reached. The proximity of these sites to the epicenter means that maximum possible warning times are invariably short. Real network-based EEW systems with detection, processing, and transmission delays may not be fast enough to provide alerts for many of these cases. For such sites, only on-site warning systems [e.g., *Hsu et al*., 2016] may be fast enough to provide alerts with short but positive warning times. There is a small number of high-intensity sites that have longer warning times; these are sites that locate in the direction of rupture propagation, close to the finite fault, but at considerable distance from the epicenter.

The GMPs from the IFS method are unbiased. This indicates that the systematic GM underprediction of the SPS method is a consequence of magnitude and location estimation errors. The scatter around the median is similar for both methods: ±0.5 MMI units for the sites with weak GM and ±1MMI units for sites with strong ground motion (16th to 84th percentiles).

A somewhat simpler, albeit less comprehensive, way of depicting an algorithm's ability to predict ground motion in real time is to compare predicted and observed ground motions at a specified warning time, e.g., $\Delta t_{w,i} = 8$ s (Figure 4). Considering an approximate delay of ~3 s (due to telemetry, processing, etc.), this would correspond to an actual warning time of ~5 s. For the SPS method this perspective shows a strong saturation of the ground motion predictions at MMI 4–5. This reflects that for large events, the method underestimates the final size of the earthquakes and, for smaller events, high intensities occur only near the epicenter, where warning times are invariably less than 8 s. The IFS method has, by construction, no such saturation, since we assume to know the true source size and location instantly upon event detection. However, the scatter of predicted ground motion strongly increases above MMI ~ 4. Points that plot on the abscissa are sites for which no alert is available at $\Delta t_{w,i} = 8$ s because they are located too close to the epicenter where ground motion arrives fast (1107 points in total). This highlights the important point that the majority of sites with very high ground motion intensities typically do not get 8 s of warning time, even under ideal conditions. If warning times are not considered in this comparison of predicted and observed intensities—as is common practice [e.g., *Xu et al*., 2017*; Kodera et al*., 2016*; Colombelli et al*., 2012*; Allen*, 2007]—this important fact may go unnoticed.

It is interesting to evaluate which magnitude and distances are dominant for different threshold levels. This is shown in Figure 5 where the data set is disaggregated with respect to magnitude and fault distance for different intensity levels and corresponding maximum possible warning times. More than 80% of cases with weak and light shaking intensities are caused by events with $M \leq 6.5$ (Figure 5a); the same is true for ~60%

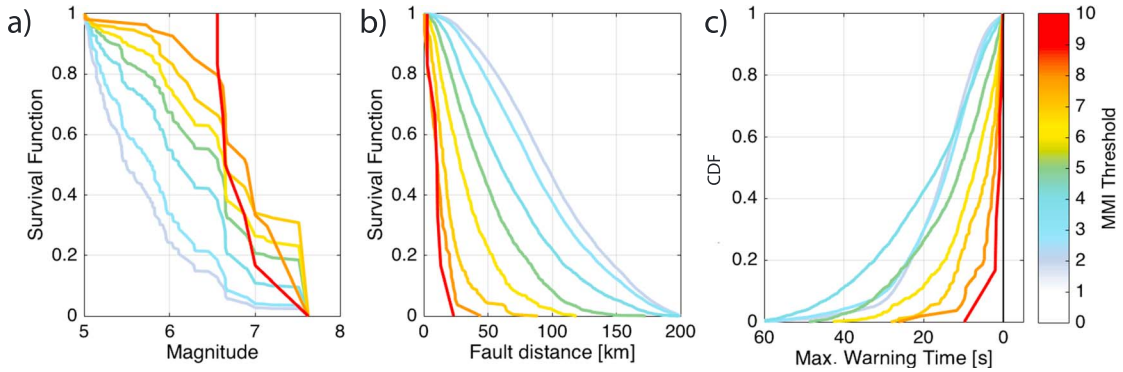

Figure 5. Empirical survival function for different MMI thresholds for (a) earthquake magnitude, (b) fault distance, and (c) empirical cumulative distribution function of maximum possible warning time. Each line represents the distribution of magnitude, distance, and warning time for all records with observed ground motion larger or equal to the threshold, which is given by the line color. Reading example: roughly 20% of ground motion records with MMI $\geq$ 5 (green) are from earthquakes with magnitudes $\geq$ 7. Maximum warning times are from when earthquake is first detected until MMI 4 is first exceeded or for sites with MMI $<$ 4, until theoretical *S* arrival.

of moderate to strong shaking intensities. The large earthquakes ($M > 6.5$) only dominate the very strong shaking classes (MMI $\geq$ 7). Such strong ground motion is mostly confined to fault distances shorter than ~30 km (Figure 5b). Figure 5c shows the resulting maximum possible warning times for the different ground motion thresholds. Note that all these aspects are properties of the data set and are independent of which EEW method is used.

4.2. Alert Triggering Criteria

After an EEW system has detected a big enough ongoing event it has to decide which sites should be alerted and which ones are far enough away from the earthquake, such that they should not be alerted. The obvious way to take this decision is via a *trigger criterion:* an alert can be issued for a site once a scalar ground motion prediction for the site exceeds a threshold level or once the probability of exceedance for a threshold ground motion exceeds a threshold probability level.

Since such criteria decide whether or not a site is alerted, the numeric ground motion prediction accuracy may not be directly relevant to EEW end users. Instead, what is most important is (i) whether or not the criterion is met in cases when an alert is indeed warranted (defines number of correct and missed alerts), (ii) how fast the criterion is met (defines the warning time), and (iii) how often the criterion is met when no alert should be sent out (defines number of false alerts).

These quantities are a direct consequence of the accuracy of the predictions of a given algorithm and of the chosen triggering criterion. Given the prediction errors, how well can an algorithm classify sites into above- or below-threshold sites in real time? In the following section we analyze how often sites are correctly classified and how often the classification fails, how this depends on the used triggering criterion, and how warning times are affected.

4.3. Classification Performance for a Chosen Trigger Criterion

We make use of the probabilistic nature of the ground motion predictions, $p\left(\widehat{\mathrm{MMI}}_i, \Delta t_{w,i}\right)$, and define a probabilistic triggering criterion that involves two different thresholds: (i) a ground motion amplitude threshold, here MMI', above which an end user would like to perform some kind of risk mitigation action, and (ii) a probability of exceedance threshold, p'_{ex}: action is triggered if $p(\mathrm{MMI}_i \geq \mathrm{MMI}') \geq p'_{\mathrm{ex}}$.

We assume that alerts are sent out to individual sites as soon as this triggering criterion is met for these sites. Because in real EEW systems alerts are difficult to cancel, we assume that we cannot cancel an alert once it has gone out. If an alert goes out before MMI' is reached, we count it as a true positive ("TP") and store the resulting warning time. The warning times at the TP sites therefore depend on how fast the predictions of a given algorithm meet the triggering criterion. If an alert goes out for a site at which ground motion never reaches MMI', it counts as a false positive ("FP"). Sites with final ground motion amplitudes above MMI' that are not alerted in time count as a false negative ("FN"). This can happen either because ground motion was

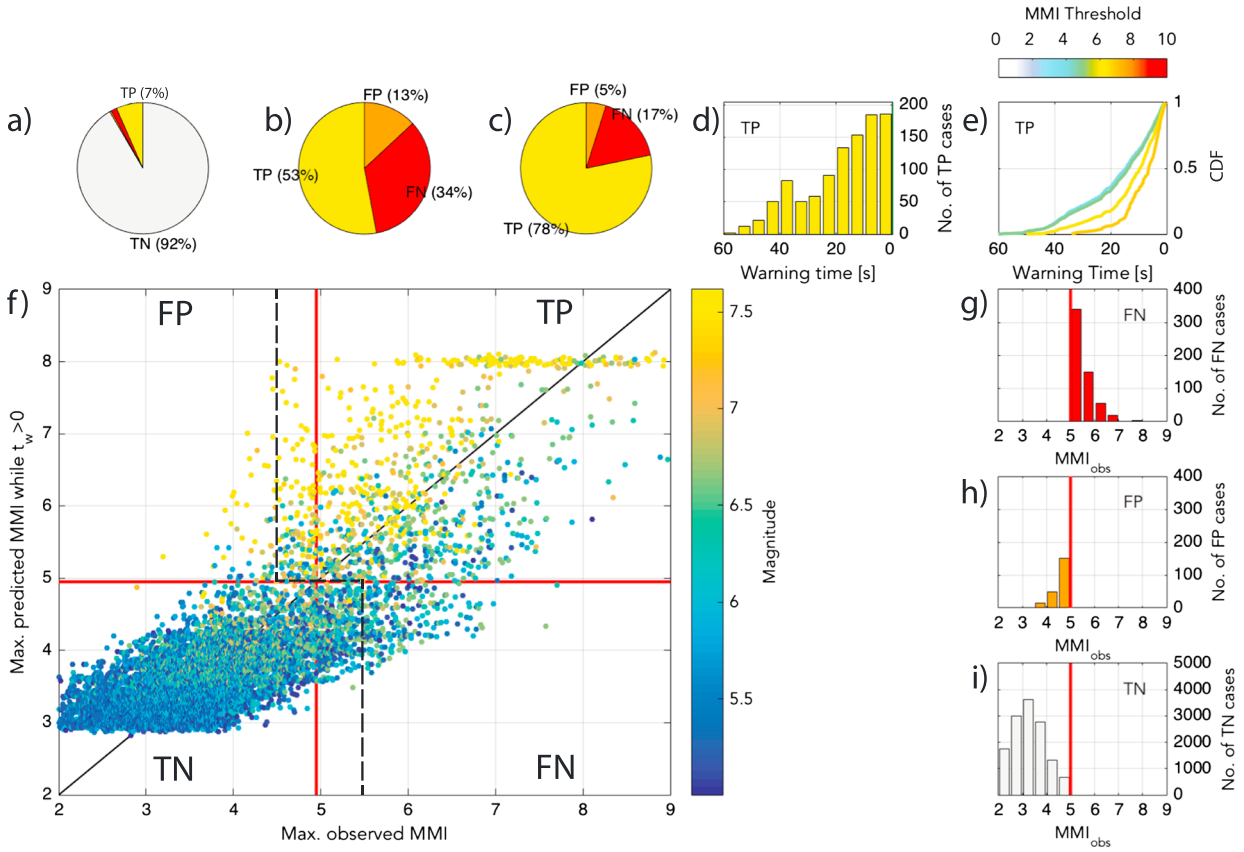

Figure 6. Overview of the real-time classification performance for the IFS Method with a trigger criterion of $p'_{ex} = 0.5$ for MMI = > 5. (a) Percentage of TP, TN, FP, and FN cases. (b) Same as Figure 6a without TN cases. (c) Same as Figure 6b with a tolerance range of ±0.5 MMI units. (d) Histogram of warning times for the TP cases and (e) empirical cumulative distribution functions (CDFs) for the warning times of sites with peak observed MMI above different MMI threshold levels (given by color of line); reading example: ~50% of sites with observed MMI ≥ 4 (cyan) have warning times longer than 18 s. (f) Predicted versus observed MMI levels, color coded by catalog magnitude. Predicted MMI values are the highest MMI level for which the probability of exceedance is larger than 50%. Predictions at negative warning times (i.e., that would be too late) are disregarded. Because p_{ex} are computed for discrete MMI values (from 3.0 to 8.0 in 0.2 increments), a small random fudge factor was added (with standard deviation 0.05 MMI units) so that data plots do not all plot on the same lines. A small number of records have peak predicted MMI above 8.0; these cases plot at MMI 8.0. (g) Observed peak MMI levels for FN cases; more than half of the FN cases had observed ground motion of less than 0.5 MMI units above the MMI' threshold. (h) Observed peak MMI levels for FP cases. (i) Observed peak MMI levels for TN cases.

underpredicted or because the alert went out too late, i.e., after MMI' was reached. Sites that are not alerted and do not exceed MMI' count as true negatives ("TNs"). In the following we use the term "classification performance" to describe the relative frequencies of TP, TN, FP, and FN cases, i.e., the ability of an algorithm to predict (or classify) which sites are about to experience above- and below-threshold ground motion.

In Figure 6 we summarize the classification performance for the IFS method using the triggering criterion $p_{ex} \geq 0.5$ for $MMI' = 5$ for all 15,553 records of the data set. Figure 6f shows observed ground motion intensities against the maximum predicted intensities. Here maximum predicted intensity means the highest MMI' for which the threshold p'_{ex} was exceeded. The four quadrants represent the TP, FN, TN, and FP cases (clockwise).

Despite the wide scatter between observed and predicted ground motion, the classification performance is relatively high. The 92% of the sites are correctly classified as MMI < 5 (TN), and 7% are correctly classified as MMI > = 5 (TP, Figure 6a). However, the fraction of TN cases is somewhat arbitrary: if we had included sites out to 1000 km hypocentral distance, TP would rise to >99.9% simply because we added a lot of "easy cases," but without any improvement in the actual algorithm. It may therefore be more meaningful to look only at the TP, FP, and FN cases (Figure 6b). This perspective shows that there is a substantial fraction of FN cases and that among all the alerts that have gone out (TP + FP) about 20% were false alerts (0.13/(0.53 + 0.13) ~ 0.2).

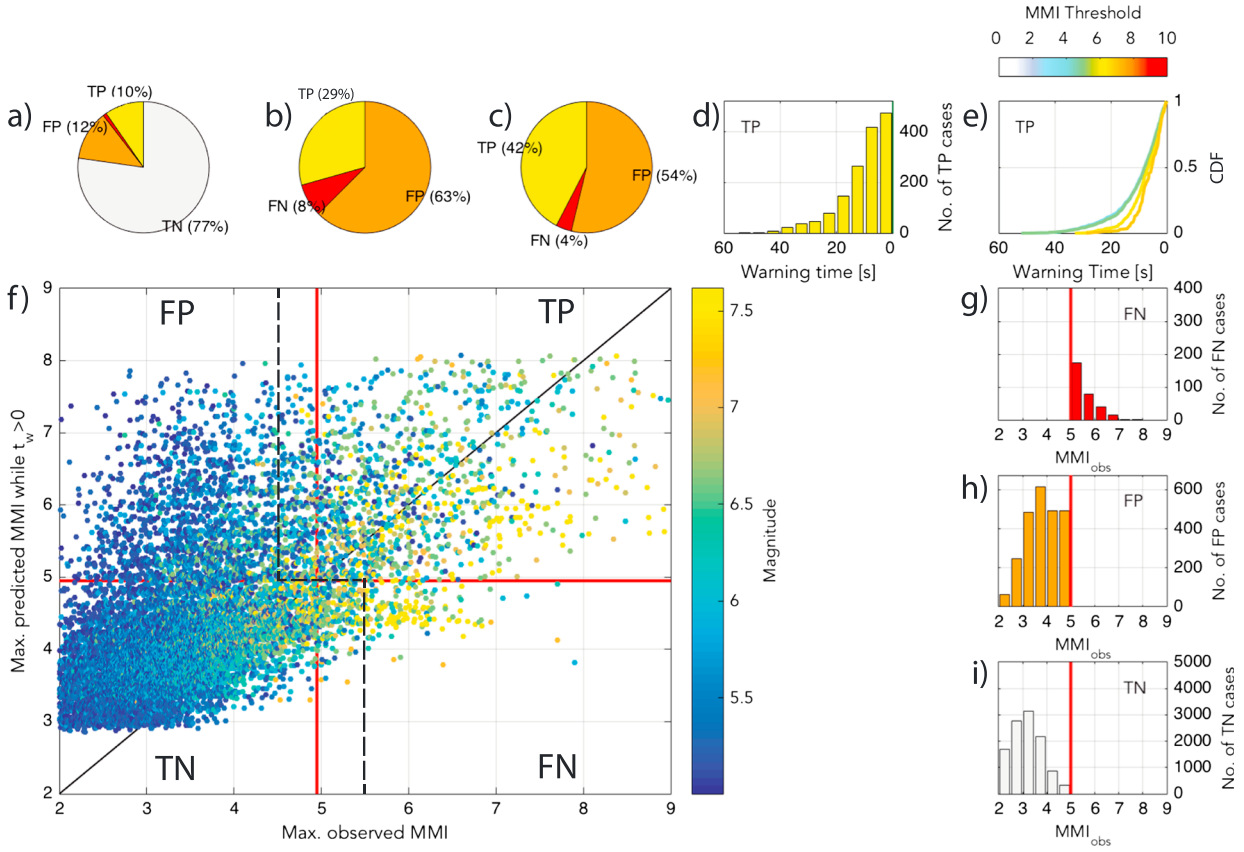

Figure 7. Same as Figure 6 but for SPS method. Occasional overestimation of the magnitude of smaller earthquakes leads to a large fraction of FP cases.

However, the sharp classification boundaries that we have used for this evaluation may result in an overly pessimistic picture. An end user that receives an early alert for ground motion with MMI $\geq$ 5.0, and then ends up experiencing MMI 4.8, may likely perceive the alert as correct and useful. For many end users there may be a certain tolerance level within which misclassifications are acceptable (cf. discussion in section 5). We may therefore introduce a tolerance range, chosen here as ±0.5 MMI units: if an alert has been issued for a site that experiences ground motion above $MMI' - 0.5$, we count it as a TP; only if the observed ground motion is more than 0.5 MMI units below the threshold do we count it as a FP. Likewise, if no alert has been given out, we count it as a TN if the observed ground motion does not exceed $MMI' + 0.5$. The 0.5 MMI units reflect an approximate lower bound for aleatory uncertainty in ground motion predictions (cf. section 4.1). Note that whether or not an alert is sent out is determined by a sharp boundary without tolerance range. The tolerance range only affects the subsequent classification of whether the alert was correct or not. The classification counts after introducing the tolerance range are shown in Figure 6c; the tolerance range changes the percentages substantially. This is because a majority of misclassified sites have observed ground motion amplitudes close to the threshold levels (Figures 6g–6i). However, even with the tolerance range, and despite the unrealistic assumptions of the IFS method, there is a substantial number of false classifications.

The warning times for the TP cases, i.e., for sites that are correctly classified as MMI $>$ 5, are defined by how quickly the triggering criterion is met after an event is detected and by how many seconds later the MMI threshold is actually exceeded. The warning times range from 0 to 60 s (Figure 6d). Figure 6e disaggregates the warning times for sites with observed MMI levels above a number of different MMI threshold levels. About 50% of the low-intensity sites have warning times $>$15 s, while most sites with MMI $\geq$ 7 have $<$10 s.

Figure 7 shows the same performance summary for the SPS method. Owing to magnitude and location estimation errors, the scatter between observed and predicted ground motion is significantly larger. In particular, the size of the relatively frequent *M*~5 earthquakes is sometimes overestimated, which leads to a large

number of ground motion overpredictions and, consequently, false positive cases. This strongly reduces the performance of the SPS method. Identical plots for both methods, and for different triggering criteria, are provided as supporting information.

4.4. Classification Performance as a Function of Trigger Criterion

The classification performance is a strong function of the chosen trigger criterion. So can we achieve significantly better classification performance with a different criterion? For a chosen criterion, how many false classifications does an end user of a particular algorithm have to accept over a certain time period, and what combinations of false and missed alert rates can the algorithm provide? To explore these relations, we compute the real-time classification statistics for all records of the data set, using a range of MMI' thresholds (i.e., scalar ground motion thresholds) and p'_{ex} thresholds (i.e., probability of exceedance thresholds), for both the SPS and the IFS method. Owing to the considerations above, we use a tolerance range of ±0.5 MMI units for the evaluation.

As expected, there is a clear trade-off between occurrence rates of missed and false alerts: Low p'_{ex} lead to a large number of false alert (FP) cases, while high p'_{ex} lead to a large number of missed alert (FN) cases. Figures 8a and 8b show normalized TP rates, TPR = TP/(TP + FN), versus normalized FP rates, FPR = FP/(TP + FN), for four different MMI' thresholds. Ideally, the curves would fall as close as possible to point [0, 1], i.e., have a maximum number of TP and a minimum of FP and FN cases. In reality, however, a certain number of misclassifications are inevitable with any algorithm.

This depiction is very similar, but not identical, to Receiver Operating Characteristic (ROC) plots [e.g., *Zechar*, 2010; *Spackman*, 1989]. The difference between these "pseudo-ROC" and the actual ROC plots is that here we do the normalization for both TPR and FPR with respect to TP + FN, i.e., the total number of cases in which the observed ground motion has exceeded the respective threshold. This is done to avoid using TN, which, as discussed in section 4.3, is strongly affected by arbitrary data selection practices.

These pseudo-ROC plots reveal that the classification works better for lower amplitude thresholds than for higher ones. Using the SPS method (Figure 8a), $p'_{ex} = 0.5$ and a MMI' of 4, the normalized TP rate is close to 1. This means normalized FN rates (= 1 − TPR) are close to zero: almost always when MMI 4 threshold is exceeded, timely and correct alerts are provided. Such high TP rates, however, come at the cost of relatively high normalized FP rates of >0.5. This means that on top of the correct alerts, false alerts are declared at slightly more than half the rate of the correct alerts. This is also expressed with the *correct alert rate*, CAR = TP/(TP + FP) (Figures 8c and 8d): given that an alert has been received, how likely is the alert correct? For $MMI' = 4$, this probability is ~0.6. Note that we may somewhat overestimate the TP rate for the lower thresholds because we do not consider earthquakes with $M < 5$, which can sometimes cause ground motion above such thresholds.

For high MMI' things are worse. To attain a normalized TP rate of >0.9 for $MMI' = 6$, a FP rate of almost 2 has to be accepted, i.e., a false alert rate that is twice as high as the rate at which MMI 6 is actually exceeded. For MMI 7, unless even higher FP rates are accepted, TP rates are only in the range of ~0.5–0.8, i.e., up to half the alerts are missed, which may be unacceptable for many end users.

If magnitude and distance were instantly known upon event detection (IFS method), normalized TP rates of 0.95 together with normalized FP rates as low as 0.1 would be possible for the lower MMI'. This suggests that a majority of the misclassifications we observe with the SPS method stem from errors in source characterization. In particular, overestimation of the size of medium size events may cause large numbers of FP cases (Figure 8f). Since we assume that we cannot cancel an alert once it has gone out, it is enough if the ground motion is overpredicted at *any* point during an ongoing earthquake to cause a false alert. Note that these classification statistics obtained with the IFS methods represent an unrealistic upper bound; real EEW algorithms will always need a certain amount of time to converge to sufficiently accurate source estimates.

For the higher MMI' levels, the classification performance improves dramatically from the SPS to the IFS method, highlighting the limitation of point source algorithms for such high-amplitude ground motion, which is often caused by large-magnitude events (cf. Figure 5a). This suggests that if more powerful methods than simple point source methods can provide better source characterizations with positive

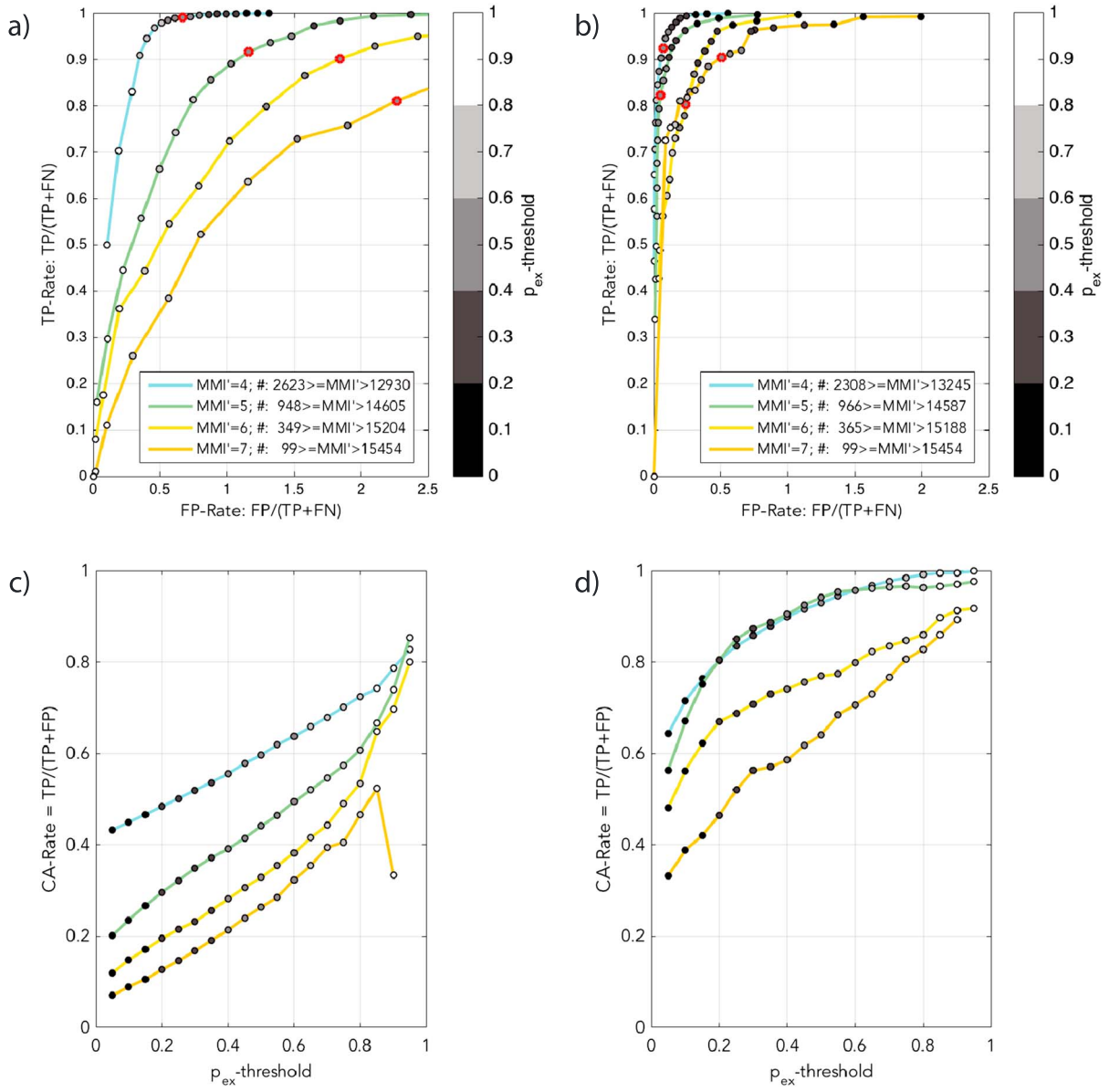

Figure 8. Real-time classification performance as a function of the trigger criterion [p'_{ex}, MMI$'$] (a and c) for SPS and (b and d) for IFS method. Figures 8a and 8b show pseudo-ROC plots, i.e., normalized true positive versus false positive rates. The normalization is with respect to the sum of TP and FN cases, i.e., the total number of cases in which a given MMI threshold was actually exceeded. The absolute numbers of how many records have observed MMI above and below each MMI threshold are given in the legend. Figures 8c and 8d show correct alert rates for the two methods, i.e., the fraction of correct alerts, relative to the total number of alerts, both correct and false, that are sent out (TP + FP). Red circles indicate points with $p'_{ex} = 0.5$.

warning times, high ground motion thresholds might become a viable option for EEW end users. The timeliness of such improved source characterizations, however, has not been convincingly demonstrated to date.

The low classification performance for high MMI$'$ may in part be a consequence of the fact that the probability of experiencing high-amplitude ground motion is itself relatively low. This is analogous to the often-invoked example to explain Bayes theorem, where the probability of actually having a rare disease is low even if a test with high sensitivity has given a positive result [*Mlodinow*, 2009]. The probability that ground motion will exceed the threshold (event "*X*"), given that an alert has been received (event "*A*") is $p(X \mid A) = p(A \mid X) \times p(X) / p(A)$. Since the probability of experiencing above-threshold ground motion, $p(X)$, is low for high thresholds, this directly reduces the probability that an alert is correct, $p(X \mid A)$. Unless the sensitivity, $p(A \mid X)$, is

Table 1. Calibrated Annual Occurrence Rates for TP, FP, FN, and TN Rates for the SPS and IFS Methods for Four Different MMI Thresholds and $p'_{ex} = 0.5$[a]

		TPR		FPR		FNR		TNR	
MMI	Rate [1/Year]	SPS	IFS	SPS	IFS	SPS	IFS	SPS	IFS
4	0.200	0.198	0.184	0.129	0.014	0.002	0.016	0.223	0.691
5	0.100	0.092	0.082	0.120	0.005	0.008	0.018	0.732	1.141
6	0.020	0.018	0.016	0.036	0.005	0.002	0.004	0.479	0.573
7	0.010	0.008	0.009	0.023	0.005	0.002	0.001	0.769	0.811

[a]The rates only quantify classification performance that results from source characterization and ground motion prediction. Other aspects of EEW systems that may be relevant for false alert rates, e.g., false event detections, are not considered here.

extremely high, this may lead to low values for $p(X \mid A)$, i.e., low probability of an alert being correct. This consideration goes to show that it is an intrinsically difficult task to correctly classify the rare occurrences of high-amplitude ground motion.

4.5. Yearly Rates of Correct and False Classifications

How often do correct and false classifications occur on an absolute scale? How many false alerts does an end user have to accept over a certain time interval? This is a difficult question because it involves estimating the rate at which a ground motion threshold will be exceeded over a certain time interval. Such rates are estimated by probabilistic seismic hazard studies [e.g., *Petersen et al*., 2015]. For the United States, the U.S. Geological Survey (USGS) Hazard Curve Application allows to estimate exceedance rates for different ground motion levels at any particular location. Although these kinds of estimates can only represent long-term averages of a highly variable quantity, and although their accuracy is debated [e.g., *Mulargia et al*., 2017*; Freedman and Stark*, 2003*; Heaton*, 2007], they may provide at least some rough guidance for what exceedance rates are to be expected. In the following we use such estimates to calibrate the relative rates of TP, FP, and FN cases that we infer from the data set.

According to the USGS Hazard Curve Application, for a site of soil class *BC* in downtown LA or San Francisco, MMI $\geq$ 7 is expected to occur on average once per 100 years, MMI $\geq$ 5 once per 30 years, and MMI $\geq$ 4 every five or so years. The number of times a ground motion threshold is exceeded corresponds to the sum of cases of TP and FN, which we have used to estimate the normalized TP rate, TPR = TP/(TP + FN) (Figure 8). At the same time, the rates provided by PSHA estimates correspond to yearly rates of exceedance, (TP + FN)/year. By multiplying the two quantities, we therefore obtain a yearly TP rate. If, for example, the normalized TP rate of the SPS method equals ~9/10 for MMI 5, and MMI 5 is estimated to occur 1/10 times per year, the absolute yearly rate of TP occurrences equals $\sim 9/10 \times 1/10 \cong 0{:}09$ times per year. Table 1 gives the corresponding calibrated annual rates for different MMI thresholds for the two methods, using $p'_{ex} = 0.5$.

We would like to caution at this point that the actual exceedance rates an end user will experience may be highly variable. They may strongly depend on whether or not the end user happens to live in a period with high seismicity, e.g., because of a very large earthquake on the San Andreas Fault.

Although the relative classification performance is much worse for higher MMI thresholds, the absolute rates of FN and FP cases can be similar or even better for higher MMI thresholds. This is because low-amplitude ground motion is far more common than high amplitudes, in nature as well as in the data set used here. Even if an individual low threshold alert has a higher probability of being correct, the cumulative number of false alerts can be larger, because the MMI threshold is exceeded more often.

This poses a dilemma for end users: if they choose a high MMI threshold, they generally get fewer alerts, both false and correct ones. But of the alerts that they do get, a majority may be false. End users that subscribe to an alert with low MMI thresholds receive alerts more often, and the probability that any individual alert is correct is much higher, but cumulatively, they obtain more false alerts. The optimal choice of a ground motion threshold for end users, furthermore, critically depends on both their costs of taking action and the costs that they avoid by successfully responding to correct EEW alerts.

Note that we here have only considered classification performance resulting from source characterization and ground motion prediction. Other aspects of EEW systems may cause false alerts, e.g., false event

detections. In order to provide end users with a comprehensive picture of false, missed, and correct alert probabilities, all other aspects of EEW systems that are relevant for classification will have to be added to the rates discussed here.

5. Discussion

Predicting earthquake ground motion in real time is a challenging task, and it will not always work well, even with sophisticated EEW algorithms. In order to understand the usefulness of EEW algorithms, and in order to improve them, we need to know how often and why different types of failures may occur with any given algorithm. Currently, we have a limited understanding of how often such failures are to be expected. In part, this is because we have so far focused on performance metrics that are not directly relevant to end users.

Here we have tried to identify metrics that directly quantify an algorithms' ability to provide what end users need: timely and accurate alerts of impending ground motion. The most useful metrics turn out to be fairly simple. The ground motion prediction error as a function of warning time, $\Delta\mathrm{MMI}_i(\Delta t_{w,i})$, shows how close the predicted ground motion comes to the observed one while there is still time for EEW users to take action. We have here chosen MMI, but any ground motion metric can be used instead. This prediction error directly determines whether or not a site is alerted (correctly or erroneously) and with how much warning time the alert was declared.

By studying multiple records, we can then compute statistics for these quantities, in particular (i) warning time distributions—the faster an algorithm reaches the triggering criterion, the longer the warning time—and (ii) classification statistics: true positive, false positive, and false negative rates, including their trade-offs, as well as correct alert rates (Figures 6–8); these rates provide an estimate for how often the provided alerts will be correct. In this evaluation it is crucial that alert timeliness is explicitly considered: alerts only count as TPs if they become available before the onset of significant ground motion; otherwise, they are FNs. Note that while in this study we have used a large data set, the metrics can equally be applied to smaller amounts of data, e.g., for all records of a single event.

In summary, the performance metrics proposed here inform both end users and algorithm developers of what performance can be expected from EEW algorithms, and they allow identifying which parts of an EEW algorithm have the largest potential for improvements. More sophisticated algorithms are only beneficial if they significantly enhance warning times and/or classification statistics.

A big open question in this context is to what extent more sophisticated algorithms can outperform SPS-type methods. While it is clear that finite source algorithms will eventually provide more accurate ground motion predictions, we do not know if such improved updates become available before strong ground motion starts at the affected sites. It would therefore be desirable that the performance of such algorithms be evaluated directly with such end user relevant metrics.

Furthermore, the metrics may facilitate direct comparisons of the performance of different algorithms in an objective and meaningful way. Both the ground motion prediction error and the classification statistics can be directly compared; for instance, the trade-off curves between TP and FP rates (Figures 8a and 8b) from different algorithms can be plotted on the same graph to identify which algorithm comes closer to a perfect classification. Because they are based on ground motion prediction rather than source characterization, they can be employed irrespective of how an EEW algorithm works. EEW algorithms that do not characterize the earthquake source at all [e.g., *Hoshiba and Aoki*, 2015] can be readily compared to more classical EEW methods.

Ideally, a testing and comparison platform would be established for EEW algorithms, similar to the Collaboratory for the Study of Earthquake Predictability [*Jordan*, 2006] where algorithms could be compared with identical input data sets and performance tests. Such a platform could potentially build on an already existing framework initially built for the ShakeAlert system [*Böse et al*., 2014], which, unfortunately, was discontinued. A new testing platform is currently being developed in Europe as part of the EPOS project, and it is a goal to include ground motion-based performance metrics such as the ones proposed here (J. Clinton, personal communication, 2017).

Note, however, that the metrics proposed here only cover the source characterization and ground motion prediction aspect of EEW. The overall EEW system performance is also affected by other challenging

aspects of the systems that we have neglected here, from event detection and association to technical details of data transmission and alert dissemination. The proposed metrics therefore constitute only one part of a comprehensive performance description of EEW systems.

For the analysis in sections 4.4 and 4.5 we have assumed a classification tolerance range of 0.5 MMI units. This corresponds to a lower level estimate of the aleatory variability in ground motion predictions. An alert (or nonalert) is considered correct if a real-time prediction is accurate to within this tolerance range. Whether or not using a tolerance range makes sense, and how large it should be, essentially depends on the cost and benefit profiles of individual end users. For real-time alert responses with low costs of false alerts (such as for individuals to seek cover under a sturdy table or slowing down trains), a tolerance range is appropriate since an alert may still be perceived as useful even if the observed ground motion only came close to the alerting threshold but did not quite cross it. End users with higher costs of false alerts (e.g., halting energy production and transmission systems), on the other hand, may want to consider sharper evaluation boundaries. To systematically determine the usefulness of EEW systems to different kinds of end users, future studies will need to consider their actual cost and benefit profiles as a function of different ground motion levels.

From the performance of the two considered hypothetical algorithms we can learn that simple point source algorithms tend to underpredict high-amplitude ground motion systematically. At the same time, the occasional overestimation of the magnitude of medium size earthquakes may lead to substantial numbers of false triggers if low-amplitude ground motion thresholds are considered. High-amplitude ground motions occur less often than low-amplitude ones, but they are generally more difficult to predict accurately. If finite source algorithms can indeed reach more accurate source characterizations while warning times are positive, real-time site classification could be strongly improved. Warning times are generally short for sites with high-intensity ground motion (MMI $>$ 7). Sites with lower amplitudes (MMI $<$ 5) can get much longer warning times, but for these sites the alerts are arguably less crucial. EEW may have the biggest potential for medium-intensity sites (MMI 5–7), where warning times can be long and ground motion can be strong.

These conclusions drawn from the two hypothetical EEW methods, however, are no final verdicts. The real potential and limitation of different kinds of EEW algorithms will become clear once these kinds of performance metrics are applied to the wide range of real algorithms. There is ample potential to improve upon the predictions from the SPS method, e.g., with better source characterization strategies, triggering criteria that are more robust to false alerts, or better ground motion prediction equations. Of the large variety of proposed EEW approaches, all have potential for increasing both warning times and/or classification statistics. The metrics should allow establishing the actual added value of each of the proposed methods.

The extent to which more sophisticated methods can outperform simple point source methods will also determine what kind of alerting strategy should be followed by EEW systems. If more elaborate approaches can indeed provide better site classification statistics, then it will be beneficial to include site-specific ground motion predictions in the alert information. End users can then adapt their emergency actions to the expected ground motion level for their specific site. If, on the other hand, the rates of misclassifications cannot be brought down substantially, a more simple alerting strategy may be more successful: The operating systems in Mexico (SESMEX) [*Espinosa-Aranda et al*., 2009] and Japan (M. Yamada, personal communication, 2016) both do not report the expected shaking strength to end users. They merely alert that a significant event is occurring, which implies that *some* level of ground motion is to be expected.

The simpler strategy that does not provide shaking predictions is sufficient for end user actions with low costs of false alerts. Trains, for example, can be stopped just in case ground motion turns out to be strong. But the situation is different for emergency actions with high costs of false alerts. Such actions can only be triggered by EEW alerts if the alerts are specific and reliable. An extreme end-member example may be nuclear power plants (NPPs): *Cauzzi et al*. [2016] report that an emergency shutdown of a NPP in Switzerland costs on the order of $250 M plus $1 M per day in lost revenue. Such an action can only be performed if it is near certain that the critical shaking level will actually be reached. There is a wide range of end users between the trains and the NPPs. More research is needed to evaluate for which kinds of users it pays off to subscribe to EEW alerts. This critically depends on their costs of taking action and on the damage reduction they can achieve. But it is clear that the more reliable we can make the real-time ground motion predictions, the more we can add expensive actions to the list of what can be triggered by EEW algorithms.

6. Conclusions

Providing timely alerts for impending ground motion is a challenging task and—because of the large uncertainties involved—will inevitably be inaccurate at times. While there is a growing number of different and powerful approaches to EEW, the absolute quality of the alerts they can provide is often unclear. In part, this is because the employed performance metrics often do not reflect the usefulness of the alerts to end users.

Here we have identified simple, objective, and meaningful EEW performance metrics that measure the ability of an EEW algorithm to provide timely and accurate ground motion alerts. Applying such metrics to the various proposed EEW algorithms could (i) inform both EEW end users and algorithm developers about the power and limitations of different algorithms, (ii) facilitate meaningful algorithm comparisons, and (iii) guide EEW system design by bringing forward which aspects of an algorithm have the highest potential for performance improvements.

The two hypothetical end-member algorithms studied here show that coarse ground motion amplitude classifications are possible with currently operational point source algorithms but that the rates of misclassifications are relatively high (e.g., less than 1% missed alerts but ~40% false alerts for the simple point source algorithm with a ground motion threshold of MMI 4).

It is clear that existing algorithms can already successfully trigger a wide range of emergency actions with low costs of false alerts. More expensive actions, on the other hand, require more accurate ground motion predictions. The proposed performance metrics have the potential to show whether the necessary level of accuracy can be reached fast enough so that there is sufficient time for carrying out the corresponding emergency actions.

Acknowledgments

The author would like to thank Sarah Minson, Elizabeth Cochran, Tom Hanks, Annemarie Baltay, Jennifer Andrews, Egill Hauksson, Tom Heaton, John Clinton, Jeremy Zechar, Stefan Wiemer, Monika Kohler, Zachary Ross, and the ShakeAlert group for discussions and comments. I appreciated the help of Annemarie Baltay and Han Yue with compiling the finite source model data. This research was supported by the Gordon and Betty Moore Foundation grants 3023 and 5229 and USGS/NEHRP Cooperative agreement G16AC00355 and the Swiss National Science Foundation. The Japanese waveform data can be downloaded from http://www.kik.bosai.go.jp/ (last accessed August 2015). We used the Seismic Transfer Program tool from http://scedc.caltech.edu/research-tools/stp-index.html (last accessed September 2014) to retrieve Southern California waveform, catalog, and arrival time data from the Caltech/USGS Southern California Seismic Network (SCSN, doi:10.7914/SN/CI), which is stored at the Southern California Earthquake Center (doi:10.7909/C3WD3xH1). The Next Generation Attenuation-West 1 waveform and metadata were obtained from http://peer.berkeley.edu (last accessed March 2014). The author does not have conflicts of interest of financial or other nature.

References

Allen, R. M. (2007), The ElarmS earthquake early warning methodology and application across California, in *Earthquake Early Warning Systems*, pp. 21–43, Springer, Berlin.

Behr, Y., J. Clinton, P. Kästli, C. Cauzzi, R. Racine, and M. A. Meier (2015), Anatomy of an earthquake early warning (EEW) alert: Predicting time delays for an end-to-end EEW system, *Seismol. Res. Lett.*, doi:10.1785/0220140179.

Ben-Zion, Y. (2008), Collective behavior of earthquakes and faults: Continuum-discrete transitions, progressive evolutionary changes, and different dynamic regimes, *Rev. Geophys.*, *46*, RG4006, doi:10.1029/2008RG000260.

Boore, D. M., J. P. Stewart, E. Seyhan, and G. M. Atkinson (2014), NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes, *Earthquake Spectra*, *30*(3), 1057–1085.

Böse, M., T. H. Heaton, and E. Hauksson (2012), Real-time finite fault rupture detector (FinDer) for large earthquakes, *Geophys. J. Int.*, *191*(2), 803–812.

Böse, M., et al. (2014), CISN ShakeAlert: An earthquake early warning demonstration system for California, in *Early Warning for Geological Disasters*, pp. 49–69, Springer, Berlin.

Cauzzi, C., Y. Behr, T. Le Guenan, J. Douglas, S. Auclair, J. Woessner, J. Clinton, and S. Wiemer (2016), Earthquake early warning and operational earthquake forecasting as real-time hazard information to mitigate seismic risk at nuclear facilities, *Bull. Earthquake Eng.*, 1–18.

Chiou, B. J., and R. R. Youngs (2008), An NGA model for the average horizontal component of peak ground motion and response spectra, *Earthquake Spectra*, *24*(1), 173–215.

Colombelli, S., O. Amoroso, A. Zollo, and H. Kanamori (2012), Test of a threshold-based earthquake early warning method using Japanese data, *Bull. Seismol. Soc. Am.*, *102*(3), 1266–1275.

Colombelli, S., A. Caruso, A. Zollo, G. Festa, and H. Kanamori (2015), AP wave-based, on-site method for earthquake early warning, *Geophys. Res. Lett.*, *42*, 1390–1398, doi:10.1002/2014GL063002.

Cua, G., and T. H. Heaton (2009), *Characterizing Average Properties of Southern California Ground Motion Amplitudes and Envelopes*, pp. 15–20, Earthquake Engineering Research Laboratory, Earthquake Engineering Research Laboratory, Pasadena, Calif.

Cuéllar, A., J. M. Espinosa-Aranda, R. Suárez, G. Ibarrola, A. Uribe, F. H. Rodríguez, R. Islas, G. M. Rodríguez, A. García, and B. Frontana (2014), The Mexican Seismic Alert System (SASMEX): Its alert signals, broadcast results and performance during the *M* 7.4 Punta Maldonado earthquake of March 20th, 2012, in *Early Warning for Geological Disasters*, pp. 71–87, Springer, Berlin.

Espinosa-Aranda, J. M., A. Cuellar, A. Garcia, G. Ibarrola, R. Islas, S. Maldonado, and F. H. Rodriguez (2009), Evolution of the Mexican Seismic Alert System (SASMEX), *Seismol. Res. Lett.*, *80*(5), 694–706.

Freedman, D. A., and P. B. Stark (2003), What is the chance of an earthquake?, *NATO Sci. Ser. IV: Earth Environ. Sci.*, *32*, 201–213.

Fujinawa, Y., and Y. Noda (2013), Japan's earthquake early warning system on 11 March 2011: Performance, shortcomings, and changes, *Earthquake Spectra*, *29*(s1), S341–S368.

Gutenberg, B., and C. F. Richter (1956), Earthquake magnitude, intensity, energy, and acceleration (second paper), *Bull. Seismol. Soc. Am.*, *46*(2), 105–145.

Heaton, T. H. (1985), A model for a seismic computerized alert network, *Science*, *228*(4702), 987–990.

Heaton, T. H. (2007), Will performance-based earthquake engineering break the power law?, *Seismol. Res. Lett.*, *78*(2), 183–185.

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*(3), 1324–1338.

Hsu, T. Y., H. H. Wang, P. Y. Lin, C. M. Lin, C. H. Kuo, and K. L. Wen (2016), Performance of the NCREE's on-site warning system during the 5 February 2016 M_w 6.53 Meinong earthquake, *Geophys. Res. Lett.*, *43*, 8954–8959, doi:10.1002/2016GL069372.

Jordan, T. H. (2006), Earthquake predictability, brick by brick, *Seismol. Res. Lett.*, *77*(1), 3–6.

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*(1), 202.

Kuyuk, H. S., and R. M. Allen (2013), A global approach to provide magnitude estimates for earthquake early warning alerts, *Geophys. Res. Lett.*, *40*, 6329–6333, doi:10.1002/2013GL058580.

Meier, M. A., T. Heaton, and J. Clinton (2015), The Gutenberg algorithm: Evolutionary Bayesian magnitude estimates for earthquake early warning with a filter bank, *Bull. Seismol. Soc. Am.*, *105*(5), 2774–2786.

Minson, S., 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. Solid Earth*, *119*, 3201–3231, doi:10.1002/2013JB010622.

Mlodinow, L. (2009), *The Drunkard's Walk: How Randomness Rules Our Lives*, pp. 114–123, Pantheon Books, New York.

Mulargia, F., P. B. Stark, and R. J. Geller (2017), Why is Probabilistic Seismic Hazard Analysis (PSHA) still used?, *Phys. Earth Planet. Inter.*, *264*, 63–75.

Petersen, M. D., et al. (2015), The 2014 United States national seismic hazard model, *Earthquake Spectra*, *31*(S1), S1–S30.

Spackman, K. A. (1989), Signal detection theory: Valuable tools for evaluating inductive learning, in *Proceedings of the Sixth International Workshop on Machine Learning*, pp. 160–163, Morgan Kaufmann, San Francisco, Calif.

Strauss, J. A., and R. M. Allen (2016), Benefits and costs of earthquake early warning, *Seismol. Res. Lett.*, doi:10.1785/0220150149.

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*(1), 204–221.

Xu, Y., J. P. Wang, Y. M. Wu, and H. Kuo-Chen (2017), Reliability assessment on earthquake early warning: A case study from Taiwan, *Soil Dyn. Earthq. Eng.*, *92*, 397–407.

Yamasaki, E. (2012), What we can learn from Japan's early earthquake warning system, *Momentum*, *1*(1), 2.

Zechar, J. D. (2010), Evaluating earthquake predictions and earthquake forecasts: A guide for students and new researchers, *Community Online Resource for Statistical Seismicity Analysis*, doi:10.5078/corssa-77337879.