JGR Solid Earth

RESEARCH ARTICLE
10.1029/2019JB017718

Key Points:

- We run three prominent Earthquake Early Warning algorithms on strong-motion data from 219 large earthquakes in Japan
- We evaluate how often, and under what circumstances, useful ground motion alerts can be provided, using a conservative warning time definition
- 40% to 60% of sites with strong to extreme ground motion get alerts with warning times ≥5 s

Supporting Information:

- Supporting Information S1

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

Citation:
Meier, M.-A., Kodera, Y., Böse, M., Chung, A., Hoshiba, M., Cochran, E., et al. (2020). How often can Earthquake Early Warning systems alert sites with high-intensity ground motion?. *Journal of Geophysical Research: Solid Earth*, *125*, e2019JB017718. https://doi.org/10.1029/2019JB017718

Received 25 MAR 2019
Accepted 7 JAN 2020
Accepted article online 10 JAN 2020

How Often Can Earthquake Early Warning Systems Alert Sites With High-Intensity Ground Motion?

Men-Andrin Meier[1], Yuki Kodera[2], Maren Böse[3], Angela Chung[4], Mitsuyuki Hoshiba[2], Elizabeth Cochran[5], Sarah Minson[6], Egill Hauksson[1], and Thomas Heaton[1]

[1]Seismological Laboratory, California Institute of Technology, Pasadena, CA, USA, [2]Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan, [3]Swiss Seismological Service (SED), ETH Zurich, Zurich, Switzerland, [4]U. C. Berkeley Seismological Laboratory, University of California, Berkeley, CA, USA, [5]U.S. Geological Survey, Pasadena, CA, USA, [6]U.S. Geological Survey, Menlo Park, CA, USA

Abstract Although numerous Earthquake Early Warning (EEW) algorithms have been developed to date, we lack a detailed understanding of how often and under what circumstances useful ground motion alerts can be provided to end users. In particular, it is unclear how often EEW systems can successfully alert sites with high ground motion intensities. These are the sites that arguably need EEW alerts the most, but they are also the most challenging ones to alert because they tend to be located close to the epicenter where the seismic waves arrive first. Here we analyze the alerting performance of the Propagation of Local Undamped Motion (PLUM), Earthquake Point-Source Integrated Code (EPIC), and Finite-Fault Rupture Detector (FinDer) algorithms by running them retrospectively on the seismic strong-motion data of the 219 earthquakes in Japan since 1996 that exceeded Modified Mercalli Intensity (MMI) of 4.5 on at least 10 sites (M_w 4.5–9.1). Our analysis suggests that, irrespective of the algorithm, EEW end users should expect that EEW can often but not always provide useful alerts. Using a conservative warning time (t_w) definition, we find that 40–60% of sites with strong to extreme shaking levels receive alerts with $t_w > 5$ s. If high-intensity shaking is caused by shallow crustal events, around 50% of sites with strong (MMI~6) and <20% of sites with severe and violent (MMI ≥ 8) shaking receive alerts with $t_w > 5$ s. Our results provide detailed quantitative insight into the expected alerting performance for EEW algorithms under realistic conditions. We also discuss how operational systems can achieve longer warning times with more precautionary alerting strategies.

Plain Language Summary Whether or not Earthquake Early Warning systems can provide useful (i.e., timely and correct) alerts before earthquake shaking begins at a site depends on numerous factors. The hardest hit sites tend to be the ones that are the most difficult to alert because they are typically located close to the epicenter, where the seismic waves arrive first. Sites that are farther away from the epicenter can potentially receive longer warning times, but in these places the alerts may be less important because the shaking is weaker. In this study we run three prominent Earthquake Early Warning algorithms retrospectively on 219 of the largest earthquakes in Japan. We analyze how often and under what circumstances the algorithms can provide useful warnings. We put special emphasis on the challenging sites that receive high shaking intensities. We find that all algorithms can alert a significant fraction even of these difficult cases.

1. Introduction

Under what circumstances can Earthquake Early Warning (EEW) algorithms provide useful alerts? The answer to this question is complex, because whether or not timely and correct alerts can be provided depends on a variety of factors, including the source/site geometry, the temporal evolution of the fault rupture, site conditions, and end user sensitivities (Meier, 2017; Minson et al., 2018). Experience in operational EEW systems during large earthquakes has revealed that providing timely warnings for sites with strong ground motion is possible but challenging and that near-epicentral sites may not always receive alerts in time (Böse et al., 2018; Cochran et al., 2018; Hoshiba et al., 2011; Kodera et al., 2016).

To quantify the expected theoretical performance of EEW systems, Heaton (1985) used a simple theoretical model and estimated possible warning times for sites with different peak ground motion amplitudes, assuming instantaneous detection of the earthquake and neglecting operational and wave propagation delays. Allen (2006) included realistic delays and estimated expected warning times for a set of 35 plausible scenario earthquakes in Northern California with randomized hypocenter locations across known faults. The author assumed that the final rupture size could be accurately characterized when 4 s of *P* wave data are available at the four closest stations. This assumption may not be realistic for large earthquakes, because they would still be growing at this point in time, and since rupture evolution may be only weakly predictable (Böse & Heaton, 2010; Meier et al., 2016; Meier et al., 2017; Melgar & Hayes, 2017).

Meier (2017) evaluated to what extent two hypothetical EEW algorithms can provide timely and correct ground motion alerts, using an extensive waveform data compilation. The author found that sites with strong ground motion are inherently more difficult to alert, in terms of both alert timeliness and accuracy, and suggested a series of EEW performance metrics that reflect the usefulness of the alerts from an end user perspective. Minson et al. (2018) evaluated the timeliness of ground motion alerts from an idealized network-type EEW system assuming a theoretically and empirically motivated moment rate evolution. They found that—if final magnitudes are only weakly predictable—warning times for sites with strong ground motion are necessarily short because such amplitudes are only predicted for a target site once the rupture has grown sufficiently large and close to the site, after which strong ground motion sets in fast. Minson et al. (2019) analyzed how the strong variability in ground motions limits the ability of EEW algorithms to provide correct ground motion alerts. They suggested a framework for estimating normalized cost reductions (CRs) from EEW-induced damage mitigation actions when predictions with limited accuracy can lead to false and missed ground motion alerts. They found that the strongest CRs are achieved if alerting thresholds are set at levels that are significantly lower than the threshold where damage is expected.

All three studies—Meier (2017), Minson et al. (2018), and Minson et al. (2019)—aimed to quantify the maximum potential of ideal or theoretical EEW algorithms. How much of this potential can actually be achieved with operational algorithms, however, is currently unclear. Here we evaluate and compare the performance of three operational EEW algorithms on a large set of strong-motion data: (1) The Earthquake Point-Source Integrated Code (EPIC) algorithm (Chung et al., 2019) estimates earthquake point-source parameters in real time from empirical early ground motion scaling relations; it is the point source algorithm in the operational ShakeAlert EEW system for the U.S. West Coast (Given et al., 2018). (2) The Finite-Fault Rupture Detector (FinDer) algorithm (Böse et al., 2018) uses precomputed binary ground motion templates to identify and track the line source that best explains the observed high frequency ground motion; it is the seismic finite-source algorithm in ShakeAlert. (3) The Propagation of Local Undamped Motion (PLUM) algorithm (Kodera et al., 2018) uses ground motion amplitude observations at nearby sites to predict impending ground motion at target sites without characterizing the earthquake source; it is one of the two operational algorithms in the Japanese public EEW system operated by the Japan Meteorological Agency (JMA, Doi, 2011; Kodera et al., 2018). The three algorithms are arguably among the most promising EEW algorithms, and they represent three different approaches to EEW. The study approach adopted here is similar to that of Ruhl et al. (2019) who have evaluated and compared the seismic point-source algorithm ElarmS and the geodetic finite fault algorithm G-larmS (Grapenthin et al., 2014) on a global data set of 32 earthquakes with $M \geq 6$.

In this study we run the EPIC, FinDer, and PLUM algorithms under maximally similar conditions on the seismic strong-motion waveform data from the 219 largest earthquakes in Japan from September 1996 to February 2017. This large data set allows us to develop realistic expectations for the alerting performance of each algorithm without making assumptions on earthquake or algorithm behavior. This approach allows us to address questions such as the following: How long are the warning times that the algorithms can provide? How often can sites with strong shaking intensities be warned successfully? How often do the algorithms produce false and missed ground motion alerts? How do these performance statistics vary for sites with different shaking intensities, and what are the main factors that determine them?

In section 2 we summarize the principles and implementations of the three EEW algorithms, the waveform processing, and the data set used. In section 3 we analyze how often and under what circumstances the algorithms can provide useful alerts to EEW end users. In section 4 we discuss the general implications of our findings for EEW systems.

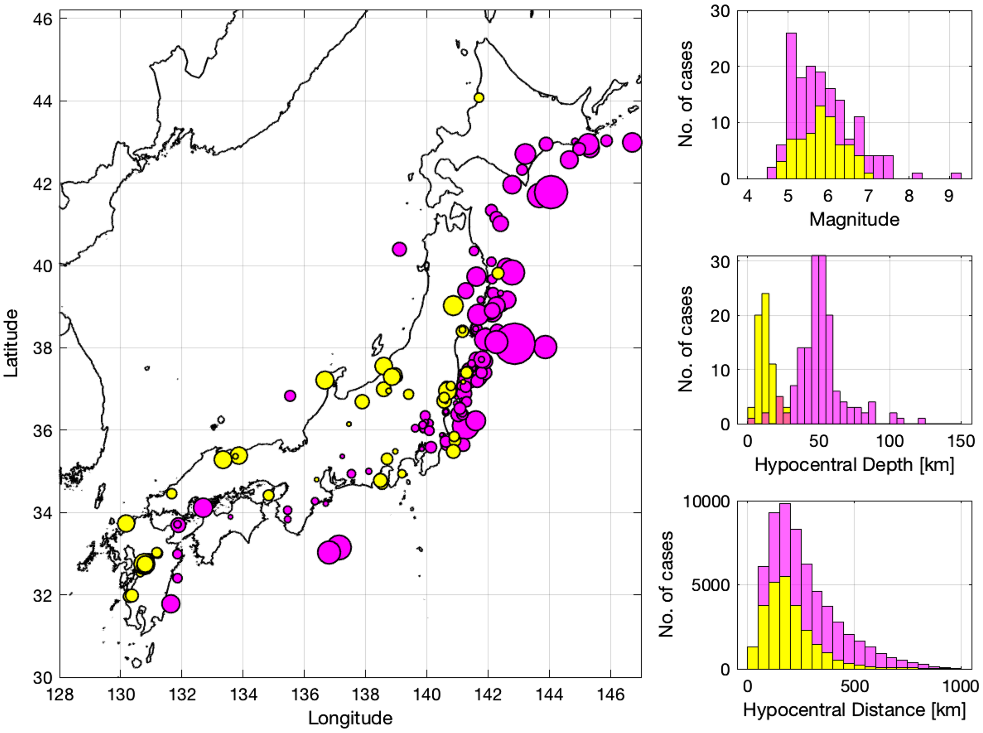

Figure 1. Data set overview. (a) Epicenter map for shallow crustal (yellow) and all other (magenta) events. Histograms of (b) catalog magnitudes, (c) hypocentral depths, and (d) hypocentral distances.

2. Data and Methods

We use a data set of 219 onshore and offshore earthquakes with moment magnitudes from 4.5 to 9.1 that were recorded at seismic strong-motion stations in Japan between September 1996 and February 2017. We selected all earthquakes for which the ground motion at 10 or more K-NET or KiK-net (Okada et al., 2004) sites exceeded Modified Mercalli Intensity (*MMI*) 4.5, corresponding to a JMA intensity of about $I_R = 3$ (Figure 1). The data set combines a total of 84,814 strong-motion records (see supporting information Figure S1) from subduction interface, shallow crustal, and deep events (Figure 1c), including the 2011 M_w 9.1 Tohoku-oki earthquake as the largest subduction event, and the 2016 M_w 7.0 Kumamoto earthquake as the largest shallow crustal event. We only use K-NET and KiK-net surface records, and we do not include any data from offshore networks. Out of the 84,814 records we remove 92 records for which our alerting threshold is already exceeded at the origin time because of waveform coda from an earlier event. The discarded records are mostly from the dense 2011 Tohoku-oki aftershock sequence. We roughly separate the shallow crustal events from all other events in our data set, by selecting all events with hypocentral depths <30 km for which the closest station has a hypocentral distance <40 km (yellow events in Figure 1). The other events (magenta) include subduction interface, outer rise, and deep events. The data set shows that a large fraction of the highest shaking intensity cases came from large subduction interface events (Figure S5).

We replay the waveform data from all events and compare the observed and predicted ground motions at each K-NET and KiK-net site, for each of the three algorithms. Each algorithm frequently updates its estimates and ground motion predictions as more data become available over time. Throughout the article we use peak JMA seismic intensity, I_R, as the ground motion metric (Kunugi et al., 2013), which is defined and described below, and we also refer to corresponding *MMI* values. For the algorithms, we aimed at staying as close as possible to their operational versions. The algorithms, along with modifications that were necessary for this offline study, are described in the following sections.

2.1. The EPIC Algorithm

The EPIC point source algorithm is a modified version of the ElarmS EEW algorithm that estimates the origin time, hypocentral location, and magnitude of an earthquake (Chung et al., 2019). Origin time and location are estimated using trilateration and a grid search. Magnitude is calculated using an empirical scaling relationship between catalog magnitudes and early peak absolute displacement amplitudes (Kuyuk et al., 2014; Wurman et al., 2007). We used the operational code itself with its native waveform processing system, but, because ElarmS was originally designed for shallow strike-slip earthquakes in California, some adaptions to the Japanese data set were necessary (Brown et al., 2009). Unlike the version of EPIC running on the ShakeAlert production system, which uses a fixed depth of 8 km for all events, the version for this study performs a grid search over a number of allowable depths (2, 8, 20, 26, 30, 36, and 42 km) to optimize *P* wave phase arrival times. We also tested larger allowable depths, but the overall performance did not improve significantly with the addition of depths beyond 42 km.

In order to prevent false alerts from teleseismic signals, EPIC employs two teleseismic checks. One of these, the filter-bank teleseismic filter (Chung et al., 2019), requires 30 s of data prior to the trigger to be able to distinguish a teleseismic signal from a local earthquake. As the Japanese data set consists of triggered data, with many stations not containing the required preevent data of 30 s, the filter was turned off for these replays. In addition, as many of the events were far offshore, the maximum station-to-event distance used to estimate hypocentral location was increased from 200 to 600 km and the maximum station-to-event distance used to calculate the magnitude was increased from 200 to 500 km. While triggers from more distant stations can be associated with an event in order to prevent those triggers from being used to create a separate split event, they are only used in the location and magnitude estimation if they are within the previously mentioned maximum distances. Finally, the Short Term Average / Long Term Average (STA/LTA) trigger threshold used (see Wurman et al., 2007, for details) was decreased from 20 to 12 and the horizontal-to-vertical check (Chung et al., 2019) was turned off. These are quite significant alterations to the EPIC code, and if they were implemented in the ShakeAlert version of EPIC it is possible that they could result in more false alerts. They were modified for this study, however, in order to allow EPIC to trigger for earthquakes further than 200 km away, and during intense earthquake sequences with numerous events in close succession.

The outputs of EPIC are source parameter estimates (location, origin time, and magnitude) that are updated as more data become available over time. To calculate ground motion predictions, we calculate predicted peak ground velocity (PGV) values for each site as per Si and Midorikawa (1999), using empirical site corrections that we have computed from this data set (Supporting Information S1). I_R was then calculated from the corrected PGV values using an empirically derived relationship following Midorikawa et al. (1999).

2.2. The FinDer Algorithm

The FinDer algorithm (Böse et al., 2012; Böse et al., 2015; Böse et al., 2018) rapidly determines line source models of small (*M*3) to large (*M*9) earthquakes by matching the spatial distribution of observed ground motion amplitudes in a seismic network with theoretical template maps. These templates are precalculated from empirical ground-motion prediction equations (GMPEs) for line sources of different lengths and magnitudes, and they are rotated to constrain the strike of the earthquake fault rupture. The template with the highest correlation with the observed ground motion pattern is found from a combined grid-search and divide-and-conquer approach (Böse et al., 2018). The resulting finite-source model, characterized by the line source centroid, length, strike, and corresponding likelihood functions, is updated every second until peak shaking is reached.

In this study, we use FinDer to predict PGV, which we later convert to I_R with an empirical relationship (Figure 2). This is different from the ShakeAlert implementation of FinDer, which uses Peak Ground Acceleration (PGA) observations and templates. We construct FinDer templates from the GMPE for PGV after Si and Midorikawa (1999): one set of symmetric templates for onshore events assuming an average source depth of $D = 10$ km, and one set of asymmetric templates for offshore events assuming $D = 20$ km (see Böse et al., 2015, for details). For offshore earthquakes we constrain the strike search to 90–270° on the east coast, and to 0–90° on the west coast. Which template set is used for any given event is currently set manually. In a separate project, we are working on how to automate this selection using misfit values between predicted and observed ground motions and initial centroid locations. We use the following PGV thresholds for binary template matching (Böse et al., 2018): 0.16, 0.28, 0.51, 0.91, 1.63, 2.92, 5.21, 9.33, 16.68, 29.84, 53.37, 95.45, 170.74 cm/s, corresponding to $I_R = 1$–7. FinDer triggers once PGV exceeds 0.16 cm/s at two or more stations that are within 50 km of each other. In the forward prediction of PGV (at later times and/or at larger distances), we apply the same equation, while using the shortest distance to the line

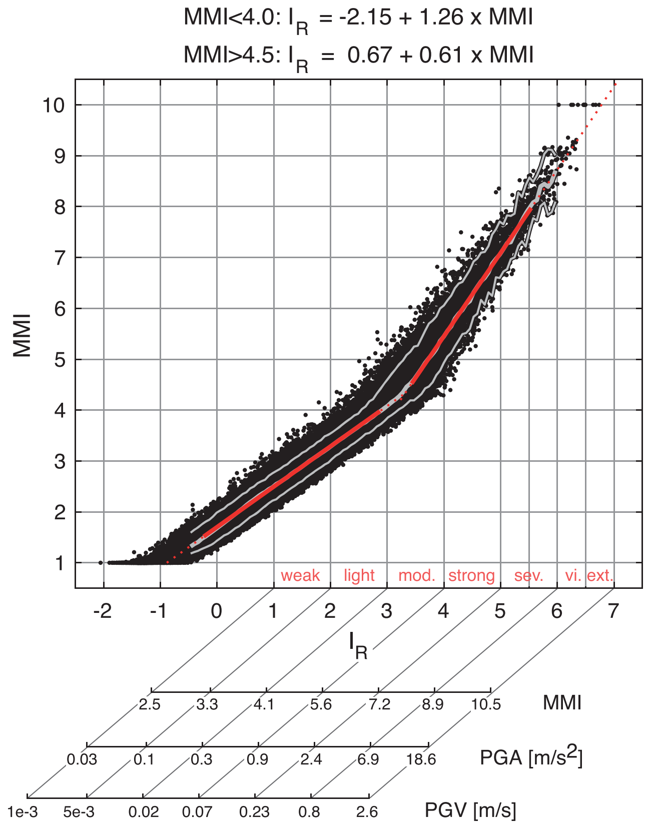

Figure 2. Empirical relation between JMA instrumental intensities I_R and Modified Mercalli Intensities *MMI*, as computed with independently measured PGA and PGV amplitudes (Worden et al., 2012, see Supporting Information S1 for details). The kink at *MMI*~4 comes from the kink in the *MMI* relation of Worden et al. (2012). Gray lines give fifth, fiftieth, and 95th percentiles in narrow ground motion amplitude bins. Red lines show a least squares regression performed on the fiftieth percentiles, so that the much more numerous low-amplitude records do not dominate the regression estimate.

source determined by FinDer as the distance metric and applying individual station corrections (Supporting Information S1).

This study is the first systematic test of FinDer using PGV instead of PGA templates. To ensure a similar magnitude convergence of FinDer for the 2016 Kumamoto earthquake as in Böse et al. (2018), which was based on PGA, we apply a new scheme for selecting the optimal PGV threshold in real time. We compute the area over which the predicted PGV exceeds each of the candidate PGV thresholds, using the FinDer real-time magnitude estimate. We compare these areas to the areas over which the observed PGV exceeds the same thresholds. The PGV threshold with the most similar exceedance area is used as the minimum threshold for FinDer to search over in the next iteration. This approach ensures that the PGV threshold for template matching increases as the event grows, which usually leads to a faster magnitude convergence as documented in Böse et al. (2018).

2.3. The PLUM Algorithm

The PLUM method (Kodera et al., 2018) is a simplified version of the wavefield-based method proposed by Hoshiba (2013) and Hoshiba and Aoki (2015). It predicts shaking intensities at all sites in a network without characterizing the seismic source, but by extrapolating the observed shaking intensities at nearby sites. Specifically, it calculates $Ir_{\text{pred}}^{(k)}$, a predicted seismic intensity at target site k, by

$$Ir_{\text{pred}}^{(k)} = \max_{i \in C_R}\left\{Ir_{\text{obs}}^{(i)} - F_{\text{Ir}}^{(i)}\right\} + F_{Ir}^{(k)},$$

where index i represents spatial position, C_R denotes the circular region centered at k with radius R (set to 30 km in this study), $Ir_{\text{obs}}^{(i)}$ represents the real-time JMA seismic intensity observed at i, and $F_{Ir}^{(i)}$ and $F_{Ir}^{(k)}$ indicate site factors at i and k, converted into intensity differences relative to bedrock. That is, PLUM estimates the shaking intensity at a target site as the maximum of observed shaking intensities within a distance of 30 km from that target site, and corrects only for differences in site amplification. $F_{Ir}^{(i)}$ was obtained using $\text{ARV}_{700}^{(i)}$ (the amplitude ratio of PGV relative to bedrock with 700 m/s at target site i) by

$$F_{Ir}^{(i)} = 1.72\log_{10}\left(\text{ARV}_{700}^{(i)}\right)$$

based on Midorikawa et al. (1999). With $R = 30$ km, PLUM can provide warning times of up to ~10s relative to the *S* wave arrival.

2.4. Ground Motion Metric

We use the JMA seismic intensity, I_R, (Kunugi et al., 2013) to evaluate the performance of the three algorithms. This ground motion metric combines acceleration and velocity ground motion via a half-integration in the frequency domain. The I_R statistic requires that the peak observed amplitude is exceeded for at least 0.3 s, which makes it largely independent of very high frequency oscillations. I_R is computed from the three-component vector sum. Figures 2 and S2 provide a detailed comparison of I_R with *MMI*, PGA, PGV, and Peak Ground displacement amplitudes.

While PLUM operates in I_R space directly, EPIC and FinDer use the GMPE of Si and Midorikawa (1999) for PGV (see Supporting Information S1), and then convert the PGV values to I_R with the empirical relationship $I_R = 2.58 + 1.98\ \log_{10}(\text{PGV})$, where PGV is in centimeters per second. We determined these regression coefficients with a least squares regression from the PGV and I_R amplitudes measured on all waveforms of our data set. They are slightly different from those suggested by Midorikawa et al. (1999), because (i) our data set is significantly larger, (ii) it extends down to lower intensities, and (iii) we measured PGV from the vector sum of all three components, while Midorikawa et al. (1999) used only the two horizontals.

2.5. Event Data Replay

All three EEW algorithms use the same 100 Hz acceleration records from K-NET and KiK-net stations (Okada et al., 2004) as input data. Original records with a 200 Hz sampling rate were low-pass filtered at 20 Hz and down-sampled to 100 Hz. While EPIC works on the acceleration waveforms directly, the other two algorithms compute envelope time series of PGV (FinDer) and of I_R (PLUM). For EPIC and PLUM we use their native waveform processors, whereas for FinDer we use an offline version that has been used to develop and apply FinDer in various places around the world, including Chile, Central America, and Switzerland.

For acceleration we use the raw (unfiltered) strong-motion records. For velocity and displacement time series we use time domain recursive filters to remove the instrument response, and to do the ground motion unit conversions (single integration for velocity, double integration for displacement). The recursive filter to obtain velocity uses an eigenperiod $T = 1$ s and a damping factor $h = 0.50$. The filter for displacement uses $T = 6$ s and $h = 0.55$. The recursive filters act as high-pass filters with corner frequencies of 1/T, that is, with 1 and 1/6 Hz, respectively.

For each record we process each component individually, and then compute the envelopes and peak amplitude values from the vector sum of the three components. The peak values were obtained every second by taking the maximum within the latest 1 s time window. The time stamp of the peak values was defined by assigning the initial time of the window; for example, if a time window started at 12:00:00.00 and ended at 12:00:00.99, the time stamp was set to 12:00:00.00. This maximum shift of 1 s does not systematically affect the estimated warning times because it affects the observed and the predicted ground motion envelopes in the same way.

2.6. Performance Analysis

For each earthquake we run all algorithms on the same data and produce pseudo-real-time ground motion predictions that evolve as the rupture and the observed ground motions develop. We assume that the end user alerting is based on a ground motion alerting threshold $I_R{}'$. Throughout this study we use $I_R{}' = 2.0$, which corresponds to MMI' ~3.3, that is, weak shaking, at the lower limit of what is noticeable by humans. When the predicted shaking intensity for a site exceeds $I_R{}'$ we assume an alert is immediately issued to that site. An alert is considered correct (true positive, TP) if the observed ground motion subsequently exceeds the alerting threshold. If an alert is issued, but the observed ground motion remains below $I_R{}'$, it counts as a false ground motion alert (false positive, FP). If an algorithm fails to issue an alert, either because of ground motion underprediction or because the prediction exceeds $I_R{}'$ too late (when the observed I_R already exceeded $I_R{}'$), we count it as a missed ground motion alert (false negative, FN). We then analyze (i) how often and under what circumstances can the algorithms provide useful alerts (i.e., alerts that are both correct and timely) and (ii) how long are the warning times for the correctly alerted sites.

For the TP cases we measure warning time as

$$t_w = t\left(I_R^{obs} \geq I_R^{'}\right) - t\left(I_R^{pred} \geq I_R^{'}\right),$$

that is, the time difference between when the observed (I_R^{obs}) and the predicted (I_R^{pred}) ground motion intensities exceed the same alerting threshold, $I_R^{'}$ (see Figure 2 in Meier, 2017). We include processing delays but neglect network transmission delays of data (from stations to data center) and alerts (from data center to end user). Because we use a low alerting threshold ($I_R{}' = 2.0$), our warning time definition is a conservative one. From the time when the observed ground motion exceeds $I_R{}'$, it usually takes some time until the high-intensity ground motion starts that requires precautionary actions. Also, note that we analyze the performance on a site-by-site basis, that is, at locations of the K-NET and KiK-net stations from which we use data. This is somewhat different from the regional alerting strategies that are employed in practice, where entire forecast regions (Japan) or regions spanned by an alerting polygon (United States) are alerted at once, which may increase warning times. For these reasons, our warning time estimates are conservative estimates and could be longer in practice (see section 4).

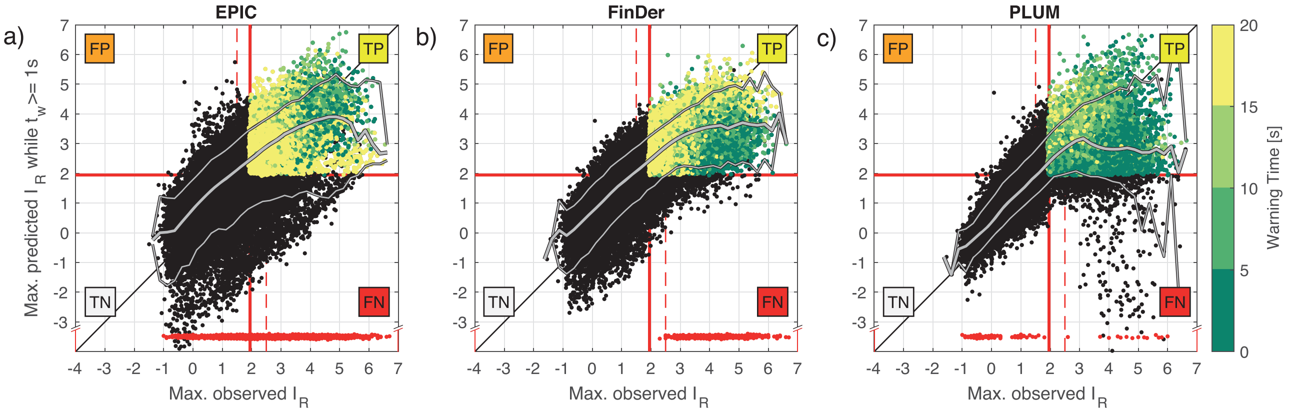

Figure 3. Predicted versus observed final peak I_R amplitudes from the EPIC (a), FinDer (b) and PLUM (c) algorithms. The predicted I_R are the maximum predicted amplitudes while there was at least 1 s of warning time; that is, at least 1 s before the alerting threshold of $I_R' = 2.0$ was actually exceeded. For the true positive cases (TP, correct alerts, top right quadrant) the color indicates the warning time, that is, the time from when the prediction first exceeded I_R' until the observed ground motion did. The other quadrants show true negative (TN), false positive (FP), and false negative (FN) classifications. Gray lines give fifth, fiftieth, and 95th percentiles in bins of observed I_R. The majority of misclassifications occur for sites with observed intensities close to the alerting threshold I_R'. Red dots show cases where no prediction has been made at all at 1 s before I_R' is exceeded because the events have not been detected at that point. These values are not used for computing the percentile curves. The dashed red lines indicate a tolerance level of 0.5 I_R units (see text for details).

3. Results

First, we evaluate how many useful alerts the different algorithms can provide for the ground motion alerting threshold of $I_R' = 2.0$. In a later section we also discuss the effect of higher alerting thresholds. Figure 3 compares the peak observed and peak predicted ground motions from each algorithm. The peak observed ground motions refer to the maximum absolute amplitude taken over the entire record. The peak predictions, on the other hand, have to be made before the observed ground motion exceeds the alerting threshold, since afterward an alert would be considered too late. The predicted amplitudes in Figure 3 are the maximum predictions made at least 1 s before the observed ground motion exceeds the alerting threshold. Equivalent figures with $I_R' = 4.0$ are provided in the supplementary material (Figure S3).

The predictions of all algorithms exhibit the correct first-order trend, but have considerable scatter. The fiftieth percentile curves saturate at $I_R = 3$–4. The saturation reflects (i) that at the time when the observed ground motion exceeds the alerting threshold, the final level of ground motion is not yet clear, for example, because a rupture may still be growing, and (ii) that high amplitude cases are often above-average amplitude cases, that is, cases with higher amplitudes than the median amplitudes predicted by GMPEs. For cases of strong to extreme ground motion it is rather rare that we can accurately predict the final amplitude ahead of time.

The FN cases are cases in which below-threshold amplitudes were predicted even at 1 s before the alerting threshold was actually exceeded. This happens particularly at near-epicentral sites where impulsive ground motion amplitudes can jump above the alerting level almost instantaneously. The red dots in in Figure 3 show cases for which no ground motion predictions have been made at all at this point in time. These are the sites closest to the epicenter, that is, the sites that are first reached by the wave fronts before the event is detected by the algorithms. These cases constitute 4.4%, 1.9%, and <0.1% of cases with above-threshold observed peak amplitudes for the EPIC, FinDer, and PLUM algorithms, respectively.

All three algorithms exhibit a slight tendency to overpredict the lower intensities. Since the alerts are based on the maximum intensity prediction, a temporary overestimation of magnitude will lead to a FP even if later, more accurate magnitude estimates are lower. Also, both EPIC and FinDer tend to slightly overestimate the magnitudes of the smaller events in the data set. Since PLUM by design assumes "undamped" ground motion it is expected to generally overpredict ground motions, but this tendency turns out to only be weak.

For all algorithms, roughly half the FP and FN misclassifications are from sites that have observed peak amplitudes close to the alerting threshold. Owing to the scatter in the observed and predicted ground motion amplitudes, such classification errors near the alerting threshold are unavoidable. These cases are arguably

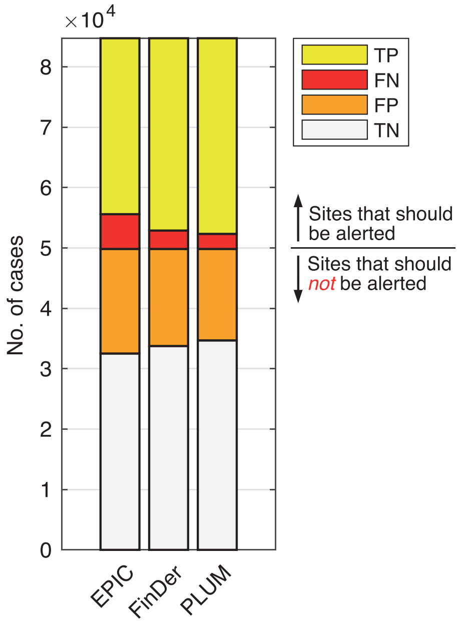

Figure 4. Classification performance for the three algorithms with an alerting threshold of $I_R' = 2$ and without considering warning times. Sites with observed ground motion intensities above the alerting threshold should be alerted (TP and FN cases). Sites with $I_R < I_R'$ should not be alerted (TN and FP cases). Roughly half of misclassifications are cases with $I_R \sim I_R'$ (not shown here).

less problematic than more severe misclassifications, for example, when an algorithm predicts below-threshold amplitudes but the shaking reaches very high intensities.

3.1. Classification Performance Without Considering Warning Times

Taken at face value, the EPIC, FinDer, and PLUM algorithms accurately classify 73%, 77%, and 79% of sites (Figure 4), for the particular alerting threshold of $I_R' = 2.0$. We compute accuracy as (TP + TN)/(TP + TN + FP + FN). Here we at first neglect the effect of warning times: a correct alert with, say, 2 s of warning time counts as a success (TP). In reality, such an alert might actually be rather useless because in a real system the alerts would be further delayed, for example, by alert transmission latencies (Behr et al., 2015). We first study these face value classification statistics to get a first-order overview (this section), and then study the effect of warning times in detail in the following sections (Figures 5–7).

The face value classification statistics are remarkably similar for all three algorithms (Figure 4). The tendency to somewhat overpredict low-amplitude ground motion causes a significant number of FP cases. The absolute number of TN cases is somewhat arbitrary since this number depends on the maximum distance out to which ground motion records are considered. At large distances there are large numbers of trivial TN cases where both the observed and the predicted ground motions are much lower than the alerting threshold. Because all algorithms use the same data set, however, the number of TN cases is meaningful in a relative sense. If classification accuracies are computed without the TN cases, that is, as TP/(TP + FP + FN), they are 56%, 63%, and 65% for EPIC, FinDer, and PLUM, respectively.

As mentioned earlier, most misclassifications occur for sites with peak observed amplitudes near the alerting threshold (Figure 3). If we instead allow for some classification tolerance and consider false classifications that are within 0.5 I_R units of the alerting threshold to be correct instead (red dashed lines in Figure 3), the fractions of correct classifications from EPIC, FinDer, and PLUM increase to 86%, 90%, and 92% including TN cases, and to 77%, 83%, and 86% without the TN cases, respectively.

3.2. What Fraction of Sites can be Successfully Alerted?

Whether or not a site can get a useful alert critically depends on the warning time. Alerts with very short warning times, for example, $t_w < 5$ s, may nominally count as TP cases but in reality they may be useless, if data and alert transmission delays are considered. In this section we therefore analyze the warning times in detail. What fraction of sites can get alerts with sufficient warning times? How many of the sites with high

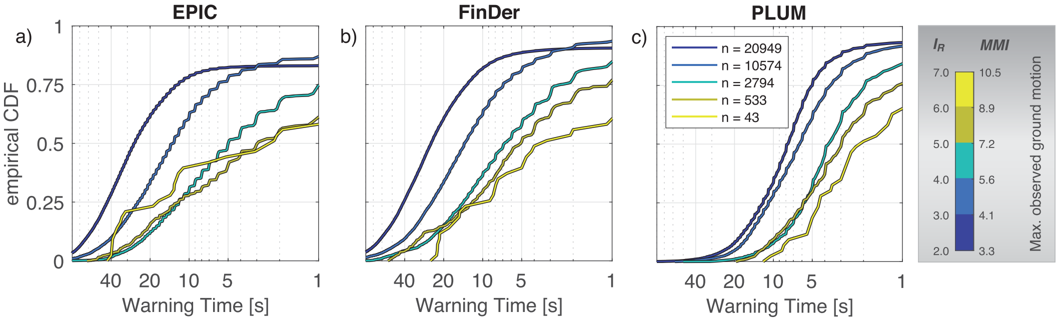

Figure 5. Cumulative distribution functions (CDFs) of warning times from EPIC (a), FinDer (b) and PLUM (c) for sites that should have been alerted, that is, sites with observed $I_R >= 2.0$, in different ground motion intensity bins. Reading example: 50% of sites with observed peak intensities between I_R 4.0 and 5.0 have warning times >5 s from EPIC. For the same ground motion bin, the CDF value at 1 s is ~0.75, showing that 1–0.75 = 25% of such sites did not get alerted at all (e.g., because of ground motion underprediction) or had warning times <1 s. The legend in (c) gives the number of records in each bin.

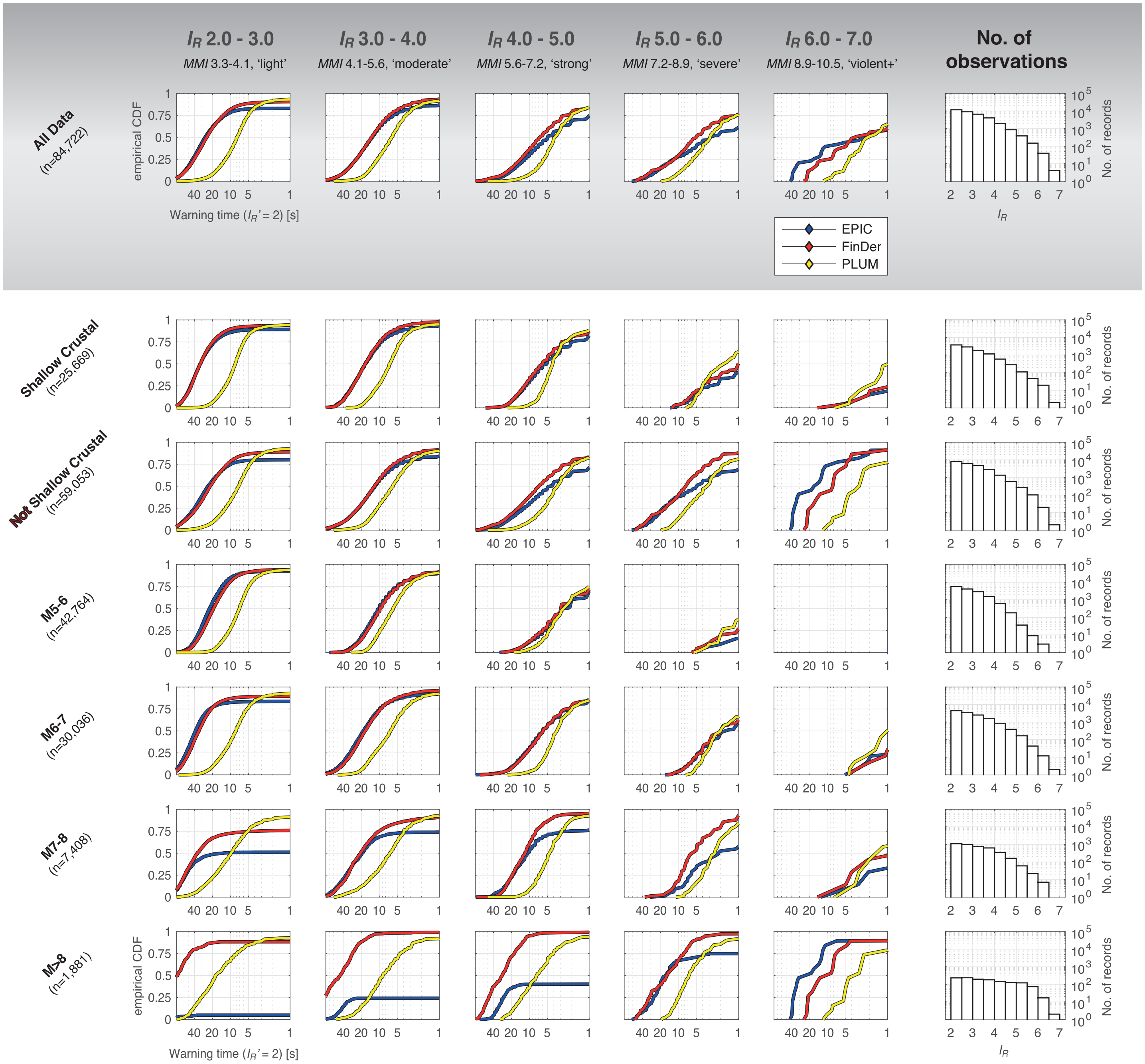

Figure 6. Warning time distributions like in Figure 5, but for different data subsets (rows) and for different peak observed ground motion intensity bins (columns). The last column gives the number of data points in each I_R bin. The data subsets are shallow crustal events, all events that are not shallow crustal events, and four different magnitude bins (M4–5, M5–6, M6–7, and $M > 7$).

ground motion intensities can we successfully alert? And what are the most important factors that determine the warning times? For this we focus on the subset of sites that should have been alerted, that is, sites with observed peak ground motions above the alerting level of $I_R' = 2.0$. We analyze what fraction of sites were successfully alerted, using the conservative warning time definition.

The distribution of warning times for all sites that should have been alerted is a strong function of peak ground motion intensity (Figure 5). Sites with high peak observed ground motions tend to have shorter warning times, because they tend to locate closer to the epicenter where the seismic waves arrive soon

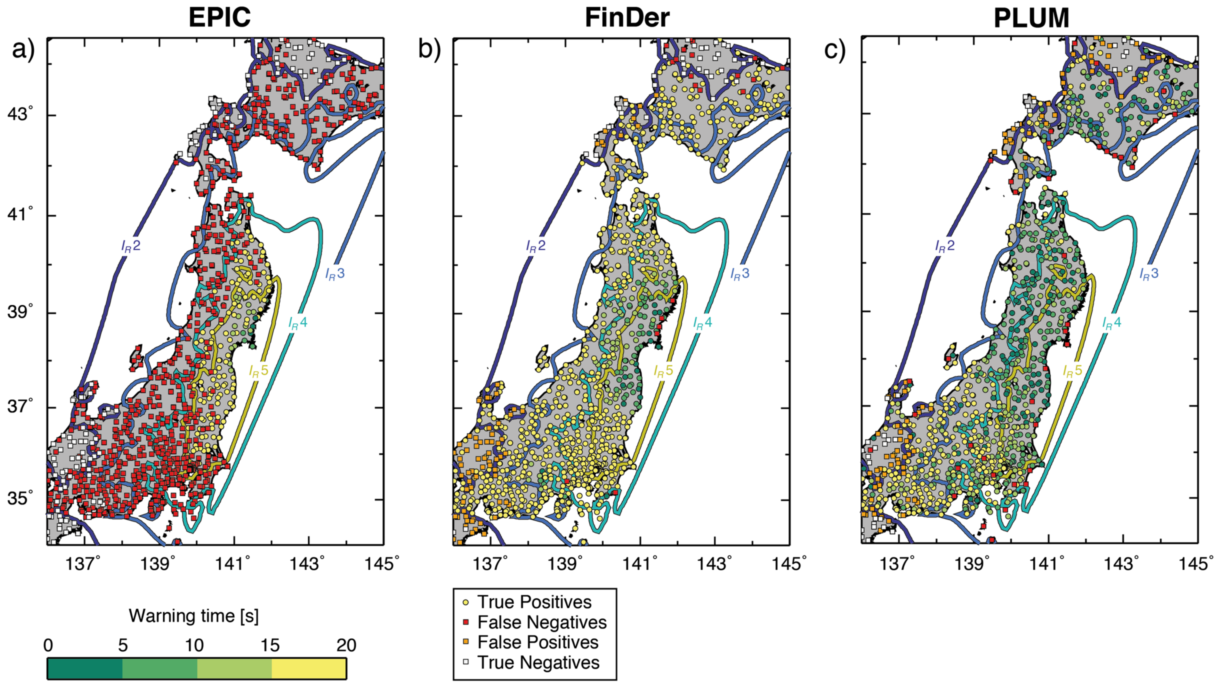

Figure 7. Warning time maps for EPIC (a), FinDer (b) and PLUM (c) for the 2011 M_w 9.1 Tohoku-oki earthquake. Warning times are relative to the alerting threshold $I_R' = 2.0$. The contour lines show peak observed ground motion levels, as computed with the natural neighbor interpolation in Matlab.

after the origin time. Of the sites with strong to extreme shaking levels ($I_R \geq 4.0$, $MMI \geq 6$) about 40–60% get >5 s of warning time from the EPIC and FinDer algorithms. For PLUM, this fraction is on the order of 25%. These are the high amplitude cases that would be successfully alerted in an operational real-time system if the additional transmission delays can be kept on the order of a few seconds (Behr et al., 2015).

Another 20–45% of these high amplitude sites may not get any alert, as suggested by the values of the cumulative distribution function at $t_w = 1$ s in Figure 5. This includes all FN cases, that is, (i) cases where the above-threshold prediction was made too late and (ii) cases where ground motion was erroneously predicted to remain below the alerting threshold. These fractions are similar for all three algorithms.

For moderate and lower amplitude shaking ($I_R < 4.0$, $MMI \leq 5$) the typical warning times are much longer, and a higher fraction of sites is successfully alerted. Around 65% of EPIC alerts for sites with I_R 3.0–4.0 have warning times >10 s and 30% get more than 20 s. The PLUM warning times are, by construction of the algorithm, limited to ~10 s because it uses observed intensities from sites with a maximum distance of 30 km for the predictions at a target site. Longer warning times for PLUM are possible, for example, if at the observing site the alerting threshold is exceeded by the P wave and/or if at the target site the threshold is exceeded later than by the direct S-phases. For EPIC and FinDer, the fraction of successfully alerted sites is slightly lower in the lowest amplitude bin (I_R 2.0–3.0) than in the next higher bin. This is because of the misclassifications of sites near the alerting threshold of $I_R' = 2.0$ (see text above).

In general, the warning time distributions are relatively broad for all ground motion bins because the same shaking intensity can be caused by a smaller nearby earthquake (short t_w), or by a larger more distant event (potentially longer t_w). If only alerts above a minimum warning time are considered (e.g., 5 s), the fraction of sites that are successfully alerted is similar for all three algorithms.

3.3. Disaggregation of Warning Times

The warning times are strongly affected by the source/receiver geometry and by the earthquake magnitude. Large subduction events, for example, tend to have more potential for long warning times because it takes more time for the strong ground motion to travel to the onshore sites. Smaller shallow crustal events, on the other hand, can produce similarly strong ground motion near the epicenter, but here the strong motion can occur within seconds of the event origin. In the following we disaggregate the warning time distributions with respect to tectonic regimes and magnitude bins (Figure 6).

Across the entire data set (top row) all algorithms provide at least 5 s warning time for a majority of sites with moderate and smaller ground motion intensities (I_R < 4.0). For sites with strong ground motion intensities (I_R 4.0–5.0) this fraction is ~50% for EPIC, ~60% for FinDer, and ~25% for PLUM. EPIC and FinDer can provide long warning times of ≥10 s for ~30% of sites with strong to extreme ground motion. At all intensity levels, the algorithms can only alert a certain fraction of sites. This fraction is significant, but it is generally lower for sites with higher intensities.

For shallow crustal events (hypocentral depth <30 km and closest station has hypocentral distance <40 km; second row) more than 50% of sites with strong ground motion (I_R 4.0–5.0) get at least 5 s of warning times from EPIC and FinDer. This fraction reduces to ~20% for sites with severe (I_R 5.0–6.0), and close to 0% for sites with violent and extreme ground motion (I_R > 6.0). For this data subset EPIC and FinDer perform remarkably similar. PLUM can generally provide alerts with t_w > 5 s only for a smaller fraction of sites than the other two algorithms. It does reach higher fractions of sites for very short warning times t_w = 1–5 s.

For the events that are not shallow crustal events (i.e., mostly subduction interface, but also outer rise and deep events; third row), successful alerts are often possible even for sites with severe, violent, and extreme ground motion (I_R > 5.0), and with long warning times. Over 75% of sites with violent and extreme ground motion (I_R > 6.0) from this data subset receive >5 s warning time from EPIC and FinDer.

The magnitude disaggregation shows that it is extremely difficult to alert high-intensity sites (I_R > 5.0) for moderate size events *M*5.0–6.0 (fourth row). For such events high intensities are observed only near the epicenter where strong ground motion arrives quickly, leaving little time to alert. Only for sites with up to and including strong ground motion ($I_R \leq$ 5.0) can a significant fraction be alerted. For larger size events (*M*6.0–7.0), on the other hand, a majority of sites with strong (I_R 4.0–5.0) and about 20% of sites with severe (I_R 5.0–6.0) ground motion have warning times >5 s. In larger events (*M*7.0–8.0), even sites with severe ground motion can often be alerted successfully, with FinDer providing t_w > 5 s for over 50% of such sites. Here the limitations of the point-source algorithms start to become important: EPIC can provide 5 s warning times only to ~30% of such sites.

For the two *M* > 8.0 events (2011 M_w 9.1 Tohoku-oki and 2003 M_w 8.3 Tokachi-oki) all three algorithms provide at least 5 s of warning time for ~75% of sites with severe ground motion intensities. Remarkably, this includes the EPIC point-source algorithm. Although EPIC drastically underestimates the magnitude of the Tohoku-oki earthquake, and overestimates the rupture distances, it predicts above-threshold ground motion amplitudes for the proximal sites on the Sendai coast, alerts a significant fraction of them, and can provide long warning times (Figure 7). The point-source limitations only affect the more distant sites with light, moderate, and strong shaking intensities. At the crucial near-source sites that EPIC successfully alerts the EPIC warning times are longer than those of the other algorithms, in part because the EPIC magnitude estimate reaches *M* > 6 faster than the FinDer estimate.

3.4. What Factors Determine Warning Times?

We next examine the prime factors that determine the warning times. Figure 8 shows the warning times from the three algorithms against hypocentral distances, magnitudes, and observed peak I_R intensities. The warning times strongly increase with recording distance, and to a lesser extent also with magnitudes. Beyond ~80 km EPIC and FinDer have median warning times of >10 s, including for the cases with high ground motion intensities.

A majority of high-intensity sites, however, tend to be at shorter hypocentral distances, and hence there is a distinct inverse correlation between ground motion intensity and warning time. The cases where a site has a large hypocentral distance but is close to the finite rupture extent (and, hence, high shaking intensities and potentially long warning times) are comparatively rare.

All algorithms, including PLUM, have a small late alert zone (i.e., a zone where the alert arrives after the observed ground motion exceeds the alerting threshold) because for an alert to count as a TP, we require the warning time to be at least 1 s. The fraction of sites that should be alerted gradually decreases with increasing hypocentral distance. At larger distances there are relatively few cases of short warning times

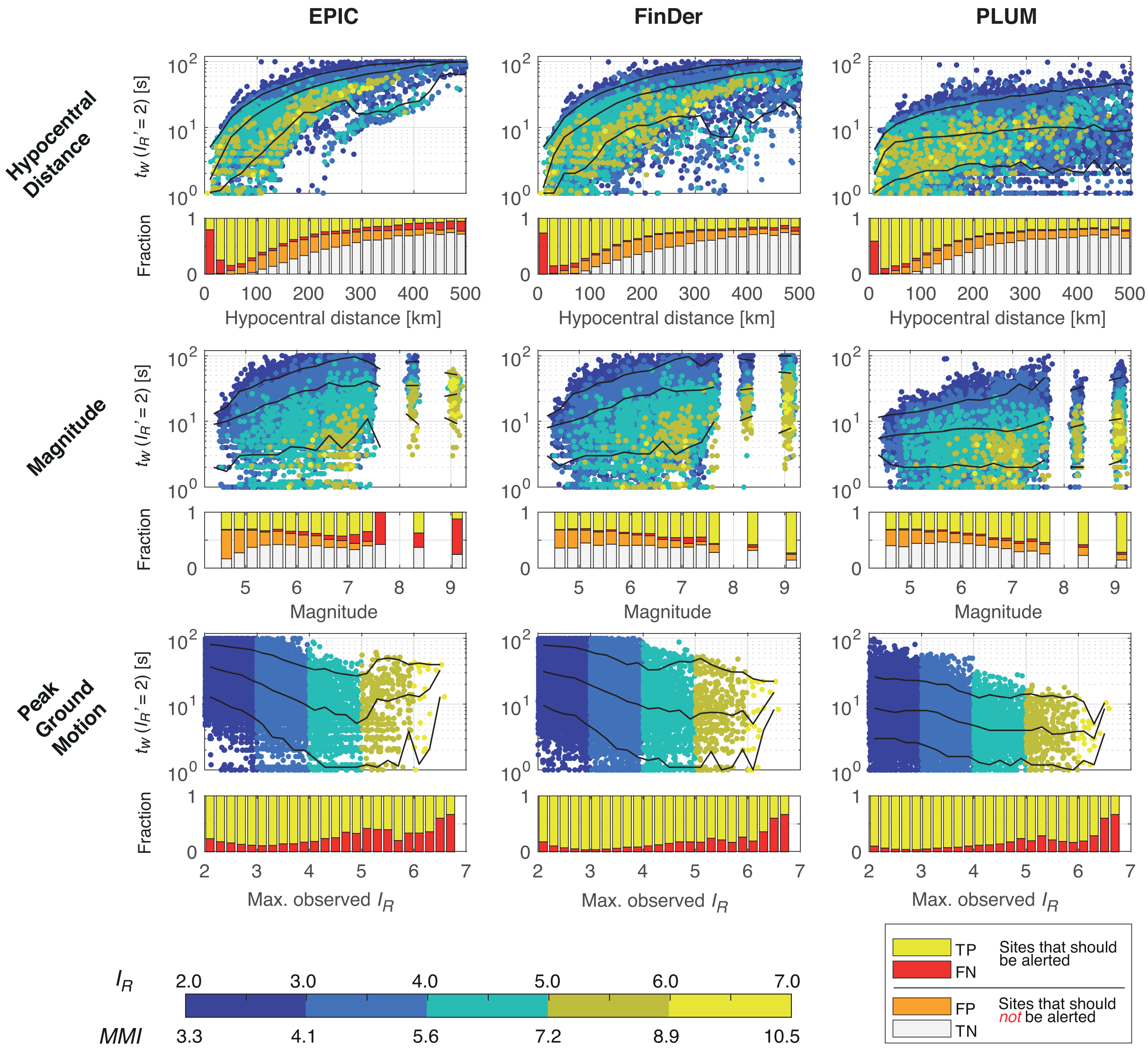

Figure 8. Warning times as a function of hypocentral distance (top), magnitude (middle), and observed peak intensities (bottom), along with classification statistics for EPIC (left), FinDer (middle), and PLUM (right). Black lines give fifth, fiftieth, and 95th percentiles of *y* values in narrow bins. The bar plots below each figure give relative fractions of true positives (yellow), false negatives (red), false positives (orange), and true negatives (white) in each bin. The magnitude values in the scatter plots were perturbed by a random value between ±0.05 to increase visibility and the data were sorted such that the highest observed intensities plot on top.

from EPIC and FinDer. This suggests that the missed alerts in this domain stem from shaking underpredictions, rather than from the alerts not being fast enough. FinDer and PLUM perform well all the way up to the largest events. EPIC expectedly fails to alert a majority of sites for very large magnitude events, although it often does successfully alert the crucial most proximal sites, as the M_w 9.1 Tohoku-oki example shows (Figure 7).

3.5. Dependence on Alerting Threshold I_R'

The binary classification performance shown in Figures 3, 4, and 8 is a strong function of the alerting threshold I_R', with generally lower performance for higher thresholds. This is important, since certain end users may want to take action only if the shaking is expected to be strong and potentially damaging. However, previous studies have found that such an alerting strategy leads to a much larger fraction of missed and false alerts, especially for sites with very high shaking amplitudes (Meier, 2017; Minson et al., 2018; Minson et al., 2019; Ruhl et al., 2019). Our results support this observation: Figure 9 shows precision, TP/(TP + FP), and recall, TP/(TP + FN) for six different alerting thresholds, requiring a minimum

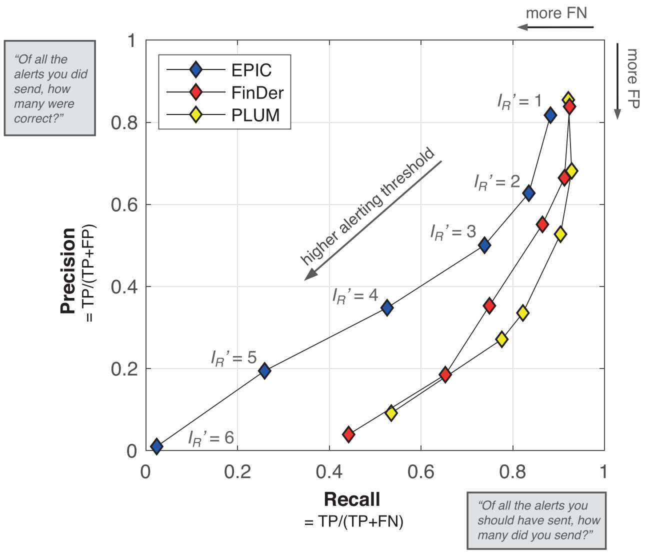

Figure 9. Precision/recall plot for different alerting thresholds, IR', for the three algorithms. The higher the threshold the more false classifications (FN and FP).

warning time of 1 s. Precision quantifies the fraction of correct alerts, among all the alerts that were sent out. Recall quantifies how many alerts were sent out, relative to the number of alerts that should have been sent.

All algorithms have the best classification performance for the lowest thresholds. In ground motion prediction, as in other statistical domains, it is inherently more difficult to accurately predict rare outcomes (here, high-amplitude ground motion) than more frequently observed ones (e.g., Sivia & Skilling, 2006), a phenomenon sometimes referred to as the false-positive paradox. Interestingly, as I_R' is increased, PLUM starts producing more FP cases, while FinDer tends to produce more FN cases. The behavior of EPIC lies somewhere between the two.

3.6. Cost Reduction Through EEW

Minson et al. (2019) introduced a framework for estimating the usefulness of an EEW algorithm by considering the CR that an algorithm can provide, relative to the case where no EEW alerts are used. The normalized CR depends on the relative frequencies of correct, false, and missed alerts, and on the ratio r between the cost of preventable damage, C_{damage}, and the cost of taking a damage mitigation action, C_{action}. The higher the C_{damage}, relative to the cost of taking action, the larger the achieved CR. In other words, end users with low costs of false alerts, and high savings from taking action, are most likely to profit from using EEW alerts. Because CR only depends on the ratio of these two costs it can be computed without knowing or estimating the absolute costs or seismicity rates. A normalized CR of 100% means that an algorithm has achieved the maximum possible CR, that is, it has facilitated the mitigation of all preventable damage, whereas a negative CR means that EEW caused more costs than it mitigated (see Minson et al., 2019, for details).

Figure 10 shows CR as a function of the cost ratio r. PLUM and FinDer generally achieve higher CRs than EPIC, because they have fewer false negative cases (cf. Figure 4). Damage mitigation actions only make sense if the preventable damage is larger than the cost of taking action ($C_{\text{damage}} > C_{\text{action}}$), that is, $r > 1$. Consequently, in a cost-reduction framework missed alerts are inherently worse than false alerts, since failing to mitigate C_{damage} is more expensive than unnecessarily spending C_{action} (Minson et al., 2019).

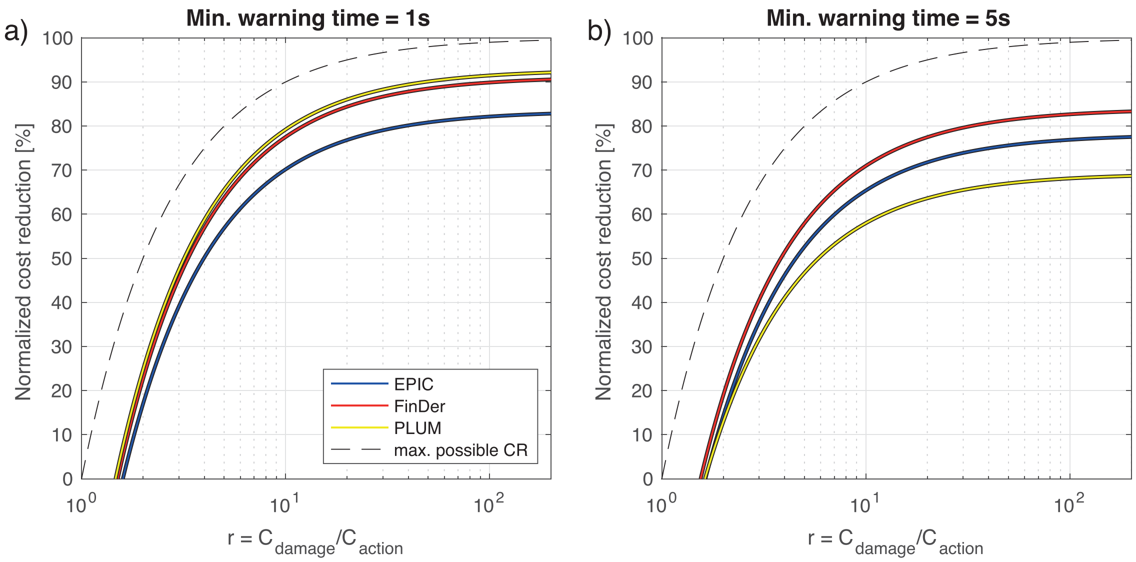

Figure 10. Normalized CRs for the three algorithms with a minimum warning time of (a) 1 s and (b) 5 s. The higher the cost/damage ratio, r, the larger the CR that can be achieved. The dashed line gives the maximum possible CR, which is always <100% because some costs incur even if mitigation actions are successfully taken.

Note that the CR statistic uses the face value alert classifications, without considering the warning times. An alert with only 2 s of warning times counts as a successful alert (TP). As discussed above, the alerting delays of a real operational EEW system may in practice reduce such a small warning time to near or below zero, in which case the alert should be counted as a missed alert (FN case). If we only consider alerts to be successful if the warning time is at least 5 s and count cases with shorter warning times as missed alerts, the CR is ~65% for EPIC, ~70% for FinDer, and ~60% for PLUM for cost ratios of r~10. This shows that most cases in which PLUM can provide an advance alert, but EPIC and FinDer cannot, are alerts with warning times of <5 s.

4. Discussion

The goal of this analysis is to develop realistic expectations for the alerting performance that EEW algorithms can provide under realistic conditions, and with minimum assumptions. How often can they provide accurate and timely alerts at different ground motion levels, and how long are the warning times of these alerts?

For this purpose, we have run offline implementations of the EPIC, FinDer, and PLUM algorithms on the data from 219 significant earthquakes in Japan. In practice, many aspects of how these algorithms are implemented can be changed and optimized. For example, different GMPE choices and site correction parameterizations may enhance the alerting performance. Innovative strategies, such as the real-time estimation of GMPE event-terms (e.g., De Matteis & Convertito, 2015) may bring about further improvements. Here we have attempted to adopt a simple and realistic implementation that allows us to study the first-order tendencies of EEW alerting performance.

With the chosen implementation, the alerting performance can be summarized as follows. Of the sites with strong (I_R 4.0–5.0, *MMI* 6), severe (I_R 5.0–6.0, *MMI* 8), and violent (I_R 6.0–7.0, *MMI* 9) ground motion levels, about 50%, 45%, and 40% of sites are successfully alerted with warning times >5 s. If we only consider shallow crustal earthquakes, these fractions are 50%, 20%, and <10%, respectively. For sites with lower peak ground motion intensities, much larger fractions of sites can be successfully alerted, and warning times are often much longer. For each ground motion level, we can only alert a certain fraction of sites, but this fraction is substantial even for the most difficult high-intensity cases.

How realistic are the warning time estimates? Real warning times from operational EEW systems may differ from our estimates in several ways. For one, we did not consider data and alert transmission delays. The effect of such latencies can be readily included by subtracting expected latencies for a given network (or region) from the warning times presented in Figures 5–8. A more important effect on warning times comes from the adopted alerting strategy. Here we have evaluated the performance on a site-to-site basis. That is, we have measured how soon we would issue alerts for each site where the K-NET and KiK-net strong-motion seismometers are located. This leads to conservative warning time estimates, and it is somewhat different from the alerting strategies of operational EEW systems. The ShakeAlert system (Given et al., 2018) computes polygons within which the expected ground motion is above an alerting level. The operational JMA EEW system (Kodera et al., 2018) alerts 188 “forecast regions” (Doi, 2011). If any of the numerous alerting points inside a forecast region exceeds the alerting threshold, the entire region is alerted, including the parts of the region that are located at larger distances. The warning times for these sites would be longer than the ones we find with the site-by-site alerting strategy studied here. Since the forecast regions can be large (>100 km), this can make a large difference for warning times. With such more “precautionary” alerts, longer warning times can be achieved, at the cost of also alerting more distant sites for which peak ground motion may or may not stay below the alerting threshold.

Furthermore, our definition of warning time is itself conservative. We measure it from the time the alert is generated until the observed ground motion exceeds the alerting threshold. Because we use low alerting thresholds, a substantial amount of time may pass from when the threshold is exceeded until damaging high-intensity ground motion begins at the same site. This adds valuable seconds for taking emergency measures. For these two reasons, the warning time estimates from this study should be considered conservative estimates.

This retrospective study was conducted with triggered waveform data, that is, we have implicitly assumed that all waveforms were accurately detected and associated to the correct event. In practice, phase detection and association are difficult and error-prone tasks, especially during intense aftershock sequences (Cochran et al., 2018; Hoshiba et al., 2011; Meier et al., 2019). Here the PLUM and FinDer algorithms have a key advantage over standard point-source algorithms in that they do not strongly rely on phase associations. PLUM's ground motion predictions can be made entirely independent of event associations (Kodera et al., 2018). Similarly, FinDer performs a continuous ground motion association, by finding the best single model that best explains the observed ground motion field, independent of pick-associations. A possible error source for FinDer, however, is that it might connect simultaneously occurring smaller events into a single larger event.

The EPIC point-source algorithm uses a maximum of 4 s of data from each site for estimating magnitudes and shows good performance up to *M*7 events, after which the magnitude estimates saturate. This is consistent with expectations from rupture evolution models with weak rupture predictability. If the source time function typically reaches the peak amplitude between one third and one half of the full rupture duration as suggested by Meier et al. (2017), 4 s would be sufficient to accurately estimate the final magnitude of a rupture with a full duration of ~12 s, which is the typical duration of a M ~ 7.0 event (e.g., Hanks & Thatcher, 1972). For larger magnitudes, longer time windows need to be considered.

Despite this expected saturation, point-source algorithms can be useful even during the largest earthquakes, as the Tohoku example shows. The EPIC magnitude estimate never exceeded 6.7 but this was enough to alert the coastal sites in Sendai, which were among the most heavily affected. The magnitude saturation only affected the alerting of more distant sites. The sites that EPIC did alert it alerted faster than the other two algorithms. FinDer, on the other hand, whose alerts were too slow for some of these coastal sites (Figure 7), correctly alerted a vast majority of the more distant sites, because it accurately characterized the growing rupture in real time. This shows that the algorithms may have complementary strengths, and that they may best be used in conjunction (e.g., Iervolino et al., 2006; Given et al., 2018; Kodera et al., 2018), for example, using the probabilistic framework suggested by Minson et al. (2017).

With this experiment we have gained some detailed quantitative insight into the practical capabilities of three important types of operational EEW algorithms. However, there are a few noteworthy caveats. The data on large shallow crustal events is limited to the 2016 M_w 7.0 Kumamoto and the 2008 M_w 6.9 Iwate-Miyagi earthquakes, that is, there are no very large strike slip earthquakes. Using additional data from other regions, for example, from the 2008 M_w 7.9 Wenchuan, China, or from the 1999 M_w 7.7 Chi-Chi, Taiwan, earthquakes, or from simulations, could provide additional insights. Furthermore, our ground motion predictions are not true out-of-sample estimates in that we have used the same data at least partly to constrain the site correction factors and the coefficients of the intensity conversion equation. Splitting the data set into training and validation subsets would thin out the crucial but rare high ground motion cases even more. Finally, we have used the dense and homogeneous strong-motion network from Japan (Okada et al., 2004). Station network properties are certain to affect the EEW alerting performance, and understanding the relation between network properties and EEW performance is an important subject for future studies.

5. Conclusions

We have applied the EPIC, FinDer, and PLUM EEW algorithms retrospectively to a suite of data from 219 real earthquakes under maximally realistic conditions. We evaluated the alerting performance of the three algorithms from an end user perspective, and on a site-by-site basis. Despite using a conservative warning time definition, we find that existing, operational EEW algorithms can alert a large fraction of affected sites, including those with high-intensity ground motion.

While the overall performance of the algorithms is surprisingly similar, their relative strengths and weaknesses may be complimentary. PLUM is simpler and likely more stable during intense earthquake sequences. It has the smallest number of false and missed alerts, but it generally provides shorter warning times than EPIC and FinDer. EPIC performs as well as FinDer up to M ~ 7, despite the point-source approximation. The alerts of FinDer are typically as fast as those of EPIC. Our warning time estimates are conservative estimates; more precautionary alerting strategies can potentially lead to longer warning times and better overall alerting performance.

With the site-by-site alerting strategy adopted here, all algorithms can successfully alert a substantial fraction, but never all, of affected sites. This fraction ranges from near 100% in large subduction zone earthquakes or for sites with light to moderate shaking, to ~50% for sites with strong to severe shaking, and to <20% for near-epicentral sites with extreme shaking levels. Since any given site may experience a range of ground motion intensities from various source types over time, end users should be prepared that EEW can often, but not always, provide correct and timely ground motion alerts.

Acknowledgments
The authors would like to thank Jeff McGuire, Annemarie Baltay, Stephen Hickman, the Associate Editor, and an anonymous reviewer for constructive comments that helped improve the manuscript. We also thank Deborah Smith, Jennifer Andrews, and Claude Felizardo and the ShakeAlert team for valuable support and inputs. The Japanese waveform data can be downloaded online (http://www.kik.bosai.go.jp/) (last accessed October 2017). This research was supported by the Gordon and Betty Moore Foundation Grant 5229 to Caltech, USGS Cooperative Agreements G17AC00347, G19AS00034 and G19AC00252 and by a Swiss National Science Foundation fellowship to M.-A. Meier.

References

Allen, R. M. (2006). Probabilistic warning times for earthquake ground shaking in the San Francisco Bay Area. *Seismological Research Letters*, *77*(3), 371–376.

Behr, Y., Clinton, J., Kästli, P., Cauzzi, C., Racine, R., & Meier, M. A. (2015). Anatomy of an earthquake early warning (EEW) alert: Predicting time delays for an end-to-end EEW system. *Seismological Research Letters*, *86*(3), 830–840.

Böse, M., Felizardo, C., & Heaton, T. H. (2015). Finite-Fault Rupture Detector (FinDer): Going real-time in Californian ShakeAlert warning system. *Seismological Research Letters*, *86*(6), 1692–1704. https://doi.org/10.1785/0220150154

Böse, M., & Heaton, T. H. (2010). Probabilistic prediction of rupture length, slip and seismic ground motions for an ongoing rupture: Implications for early warning for large earthquakes. *Geophysical Journal International*, *183*(2), 1014–1030.

Böse, M., Heaton, T. H., & Hauksson, E. (2012). Real-time Finite Fault Rupture Detector (FinDer) for large earthquakes. *Geophysical Journal International*, *191*(2), 803–812. https://doi.org/10.1111/j.1365-246X.2012.05657.x

Böse, M., Smith, D. E., Felizardo, C., Meier, M.-A., Heaton, T. H., & Clinton, J. F. (2018). FinDer v.2: Improved real-time ground-motion predictions for *M*2–*M*9 with seismic finite-source characterization. *Geophysical Journal International*, *212*, 725–742. https://doi.org/10.1093/gji/ggx430

Brown, H. M., Allen, R. M., & Grasso, V. F. (2009). Testing ElarmS in Japan. *Seismological Research Letters*, *80*(5), 727–739.

Chung, A. I., Henson, I., & Allen, R. M. (2019). Optimizing earthquake early warning performance: ElarmS-3. *Seismological Research Letters*. https://doi.org/10.1785/0220180192

Cochran, E. S., Kohler, M. D., Given, D. D., Guiwits, S., Andrews, J., Meier, M. A., et al. (2018). Earthquake early warning ShakeAlert system: Testing and certification platform. *Seismological Research Letters*, *89*(1), 108–117.

De Matteis, R., & Convertito, V. (2015). Near-real-time ground-motion updating for earthquake shaking prediction. *Bulletin of the Seismological Society of America*, *105*(1), 400–408.

Doi, K. (2011). The operation and performance of earthquake early warnings by the Japan Meteorological Agency. *Soil Dynamics and Earthquake Engineering*, *31*(2), 119–126.

Given, D. D., Allen, R. M., Baltay, A. S., Bodin, P., Cochran, E. S., Creager, K., et al. (2018). *Revised technical implementation plan for the ShakeAlert system—An earthquake early warning system for the West Coast of the United States* (No. 2018-1155). US Geological Survey.

Grapenthin, R., Johanson, I. A., & Allen, R. M. (2014). Operational real-time GPS-enhanced earthquake early warning. *Journal of Geophysical Research: Solid Earth*, *119*, 7944–7965. https://doi.org/10.1002/2014JB011400

Hanks, T. C., & Thatcher, W. (1972). A graphical representation of seismic source parameters. *Journal of Geophysical Research*, *77*(23), 4393–4405.

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

Hoshiba, M. (2013). Real-time prediction of ground motion by Kirchhoff-Fresnel boundary integral equation method: Extended front detection method for earthquake early warning. *Journal of Geophysical Research: Solid Earth*, *118*(3), 1038–1050. https://doi.org/10.1002/jgrb.50119

Hoshiba, M., & Aoki, S. (2015). Numerical shake prediction for earthquake early warning: Data assimilation, real-time shake mapping, and simulation of wave propagation. *Bulletin of the Seismological Society of America*, *105*(3), 1324–1338.

Hoshiba, M., Iwakiri, K., Hayashimoto, N., Shimoyama, T., Hirano, K., Yamada, Y., et al. (2011). Outline of the 2011 off the Pacific coast of Tohoku Earthquake (M w 9.0)—Earthquake Early Warning and observed seismic intensity. *Earth, Planets and Space*, *63*(7), 7.

Iervolino, I., Convertito, V., Giorgio, M., Manfredi, G., & Zollo, A. (2006). Real-time risk analysis for hybrid earthquake early warning systems. *Journal of Earthquake Engineering*, *10*(06), 867–885.

Kodera, Y., Saitou, J., Hayashimoto, N., Adachi, S., Morimoto, M., Nishimae, Y., & Hoshiba, M. (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 and Space*, *68*(1), 202.

Kodera, Y., Yamada, Y., Hirano, K., Tamaribuchi, K., Adachi, S., Hayashimoto, N., et al. (2018). The propagation of local undamped motion (PLUM) method: A simple and robust seismic wavefield estimation approach for earthquake early warning. *Bulletin of the Seismological Society of America*, *108*(2), 983–1003. https://doi.org/10.1785/0120170085

Kunugi, T., Aoi, S., Nakamura, H., Suzuki, W., Morikawa, N., & Fujiwara, H. (2013). An improved approximating filter for real-time calculation of seismic intensity. *Zisin 2*, *65*(3), 223–230. https://doi.org/10.4294/zisin.65.223 (in Japanese)

Kuyuk, H. S., Allen, R. M., Brown, H., Hellweg, M., Henson, I., & Neuhauser, D. (2014). Designing a network-based earthquake early warning algorithm for California: ElarmS-2. *Bulletin of the Seismological Society of America*, *104*(1), 162–173. https://doi.org/10.1785/0120130146

Meier, M. A. (2017). How "good" are real-time ground motion predictions from earthquake early warning systems? *Journal of Geophysical Research: Solid Earth*, *122*(7), 5561–5577. https://doi.org/10.1002/2017JB014025

Meier, M. A., Ampuero, J. P., & Heaton, T. H. (2017). The hidden simplicity of subduction megathrust earthquakes. *Science*, *357*(6357), 1277–1281.

Meier, M. A., Heaton, T., & Clinton, J. (2016). Evidence for universal earthquake rupture initiation behavior. *Geophysical Research Letters*, *43*(15), 7991–7996. https://doi.org/10.1002/2016GL070081

Meier, M. A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., et al. (2019). Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning. *Journal of Geophysical Research: Solid Earth*, *124*, 788–800.

Melgar, D., & Hayes, G. P. (2017). Systematic observations of the slip pulse properties of large earthquake ruptures. *Geophysical Research Letters*, *44*(19), 9691–9698. https://doi.org/10.1002/2017GL074916

Midorikawa, S., Fujimoto, K., & Muramatsu, I. (1999). Correlation of new J. M. A. instrumental seismic intensity with former J. M. A. seismic intensity and ground motion parameters. *Journal of Social Safety Science*, *1*, 51–56. (in Japanese)

Minson, S. E., Baltay, A. S., Cochran, E. S., Hanks, T. C., Page, M. T., McBride, S. K., et al. (2019). The limits of earthquake early warning accuracy and best alerting strategy. *Scientific Reports*, *9*(1), 2478.

Minson, S. E., Meier, M. A., Baltay, A. S., Hanks, T. C., & Cochran, E. S. (2018). The limits of earthquake early warning: Timeliness of ground motion estimates. *Science Advances*, *4*(3), eaaq0504.

Minson, S. E., Wu, S., Beck, J. L., & Heaton, T. H. (2017). Combining multiple earthquake models in real time for earthquake early warning. *Bulletin of the Seismological Society of America*, *107*(4), 1868–1882.

Okada, Y., Kasahara, K., Hori, S., Obara, K., Sekiguchi, S., Fujiwara, H., & Yamamoto, A. (2004). Recent progress of seismic observation networks in Japan—Hi-net, F-net, K-NET and KiK-net. *Earth, Planets and Space*, *56*, 15–28. https://doi.org/10.1186/BF03353076

Ruhl, C. J., Melgar, D., Chung, A. I., Grapenthin, R., & Allen, R. M. (2019). Quantifying the value of real-time geodetic constraints on earthquake early warning using a global seismic and geodetic dataset. *Journal of Geophysical Research: Solid Earth*, *124*(4), 3819–3837. https://doi.org/10.1029/2018JB016935

Si, H., & Midorikawa, S. (1999). New attenuation relationships for peak ground acceleration and velocity considering effects of fault type and site condition. *Journal of Structural and Construction Engineering*, *523*(64), 63–70. https://doi.org/10.3130/aijs.64.63_2 (in Japanese)

Sivia, D., & Skilling, J. (2006). *Data analysis: A Bayesian tutorial* (p. 5). Oxford: OUP.

Worden, C. B., Gerstenberger, M. C., Rhoades, D. A., & Wald, D. J. (2012). Probabilistic relationships between ground-motion parameters and modified Mercalli intensity in California. *Bulletin of the Seismological Society of America*, *102*(1), 204–221.

Wurman, G., Allen, R. M., & Lombard, P. (2007). Toward earthquake early warning in Northern California. *Journal of Geophysical Research*, *112*, B08311. https://doi.org/10.1029/2006JB004830