Geophysical Journal International

Geophys. J. Int. (2014) **198,** 1438–1457 doi: 10.1093/gji/ggu198

GJI Seismology

CyberShake-derived ground-motion prediction models for the Los Angeles region with application to earthquake early warning

Maren Böse,[1,2] Robert W. Graves,[3] David Gill,[4] Scott Callaghan[4] and Philip J. Maechling[4]

[1] *Seismological Laboratory, California Institute of Technology (Caltech), 1200 E. California Blvd., Mail Code 252–21, Pasadena, CA 91125, USA. E-mail: mboese@sed.ethz.ch*
[2] *Institute of Geophysics, Swiss Federal Institute of Technology, ETH Zurich, CH-8092, Switzerland*
[3] *Earthquake Hazards Program, US Geological Survey, Pasadena, CA, USA*
[4] *Southern California Earthquake Center (SCEC), University of Southern California (USC), CA, USA*

Accepted 2014 May 28. Received 2014 May 26; in original form 2013 August 28

SUMMARY

Real-time applications such as earthquake early warning (EEW) typically use empirical ground-motion prediction equations (GMPEs) along with event magnitude and source-to-site distances to estimate expected shaking levels. In this simplified approach, effects due to finite-fault geometry, directivity and site and basin response are often generalized, which may lead to a significant under- or overestimation of shaking from large earthquakes ($M > 6.5$) in some locations. For enhanced site-specific ground-motion predictions considering 3-D wave-propagation effects, we develop support vector regression (SVR) models from the SCEC CyberShake low-frequency (<0.5 Hz) and broad-band (0–10 Hz) data sets. CyberShake encompasses 3-D wave-propagation simulations of >415 000 finite-fault rupture scenarios ($6.5 \leq M \leq 8.5$) for southern California defined in UCERF 2.0. We use CyberShake to demonstrate the application of synthetic waveform data to EEW as a 'proof of concept', being aware that these simulations are not yet fully validated and might not appropriately sample the range of rupture uncertainty. Our regression models predict the maximum and the temporal evolution of instrumental intensity (MMI) at 71 selected test sites using only the hypocentre, magnitude and rupture ratio, which characterizes uni- and bilateral rupture propagation. Our regression approach is completely data-driven (where here the CyberShake simulations are considered data) and does not enforce pre-defined functional forms or dependencies among input parameters. The models were established from a subset (~20 per cent) of CyberShake simulations, but can explain MMI values of all >400 k rupture scenarios with a standard deviation of about 0.4 intensity units. We apply our models to determine threshold magnitudes (and warning times) for various active faults in southern California that earthquakes need to exceed to cause at least 'moderate', 'strong' or 'very strong' shaking in the Los Angeles (LA) basin. These thresholds are used to construct a simple and robust EEW algorithm: to declare a warning, the algorithm only needs to locate the earthquake and to verify that the corresponding magnitude threshold is exceeded. The models predict that a relatively moderate M6.5–7 earthquake along the Palos Verdes, Newport-Inglewood/Rose Canyon, Elsinore or San Jacinto faults with a rupture propagating towards LA could cause 'very strong' to 'severe' shaking in the LA basin; however, warning times for these events could exceed 30 s.

Key words: Spatial analysis; Earthquake ground motions; Site effects; Wave propagation; Early warning; North America.

1 INTRODUCTION

Predicting site-specific ground-motion parameters for large earthquakes ($M > 6.5$) is a major challenge in seismic hazard assessment and engineering applications. This is particularly true for systems designed to provide earthquake early warning (EEW) to locations outside of the epicentral area, a few seconds to tens of seconds before potentially destructive waves arrive (Allen *et al.* 2009a).

Commonly, ground-motion parameters, such as peak values or instrumental intensity (e.g. modified Mercalli intensity, MMI), are predicted from magnitude and source-to-site distance using empirical ground-motion prediction equations (GMPEs). This approach, however, is problematic for real-time applications such as EEW because the fault rupture extent is often unknown. The earthquake typically is approximated as a point-source with usage of hypocentral, instead of rupture-to-site distances in GMPEs. Also, site/basin and directional effects, caused by rupture directivity and seismic radiation patterns, are generally simplified or neglected. In some cases this can lead to a serious under- or overestimation of ground-motion parameters. In this paper we focus on ground-motion predictions for EEW. However, results of this study are also relevant for other applications, including for example GMPE-based probabilistic seismic hazard analysis (PSHA).

Ignoring finite-fault and 3-D wave-propagation effects can mean that EEW alerts are, in some cases, not issued. An instructive example is the *M*7.8 ShakeOut scenario earthquake in southern California with a 300-km-long rupture along the San Andreas Fault starting at Bombay Beach (Graves *et al.* 2008; Jones *et al.* 2008). If source dimensions are neglected (as is typically done in EEW; Böse *et al.* 2013), shaking intensities in the 240-km-distant Los Angeles (LA) basin are predicted as 'light' to 'moderate' (MMI = IV–V; Table 1). Under an operational warning system, no warning would likely be issued. This predicted MMI intensity, however, is four to five units less than what seismic 3-D wave-propagation simulations by Graves *et al.* (2008) suggest. Assuming that these simulations are closer to true shaking, a warning should be declared.

The situation can be improved with the application of the *Fin*ite Fault Rupture *De*tecto*r* algorithm 'FinDer' (Böse *et al.* 2012a). This algorithm is based on the observation that high acceleration values are typically observed at seismic stations close to the rupturing fault, and thus allowing the estimation of finite source dimensions and rupture-to-site distances while the fault rupture is still in progress. With this enhancement, ground-motions in the LA basin for the ShakeOut scenario earthquake are estimated as MMI = VIII ('severe' shaking), which is in much better agreement with the wave-propagation simulations. However, even with FinDer it takes about half a minute after nucleation before the rupture has reached a critical length that suggests that a warning in LA is needed. Thus the warning times, which in this particular scenario could exceed 60 s, are significantly shortened. Likewise, effects caused by fault geometry, site and basin response and directional effects, particularly important for large earthquakes ($M > 6.5$) and at longer shaking periods ($T > 1$ s), are not considered.

For a user to receive maximum value from a received warning, it is crucial for the warning to accurately specify when significant shaking is expected to arrive at the user site. Different response actions based on the expected severity of ground shaking, in particular those performed by automated control systems, may require different time duration for execution and completion. Also, if shaking is expected to occur *late*, for instance in more than 20 or 30 s, additional redundancy tests (e.g. do additional sensors also record strong shaking?) can be performed to reduce the risk of false alarms and associated costs. If, on the other hand, strong shaking is expected to occur within a few seconds, a user needs to respond quickly and no verification tests can be performed. Typically, warning times are estimated from the expected arrival of the direct *S* wave at the user location (Böse *et al.* 2013). However, as will be shown in this paper, depending on the fault rupture-user geometry and effects of wave-propagation, strong shaking in a large earthquake can be caused by later arriving phases and warning times to significant shaking may be much longer.

The aim of this study is to develop models to predict the temporal evolution of instrumental intensity MMI (strictly speaking,

Table 1. ShakeMaps intensity scale (Worden *et al.* 2012) and distribution of maximum intensity $\mathrm{MMI}_{\mathrm{max}}$ derived from the CyberShake BB (0–10 Hz) and LF (<0.5 Hz) simulations at stations ALIS (rock) and DLA (deep LA basin) with $N = 432\,480$ and $416\,970$ ruptures, respectively.

	Instrumental intensity	I	II	III	IV	V	VI	VII	VIII	IX	X+
	Perceived shaking	Not felt	Weak	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
	Potential damage	None	None	None	None	Very light	Light	Moderate	Mod./heavy	Heavy	Very heavy
BB	ALIS rock	0 (0.0 per cent)	0 (0.0 per cent)	624 (0.1 per cent)	44 630 (10.3 per cent)	67 404 (15.6 per cent)	61 345 (14.2 per cent)	53 449 (12.4 per cent)	126 248 (29.2 per cent)	68 427 (15.8 per cent)	10 353 (2.4 per cent)
	DLA deep basin	0 (0.0 per cent)	0 (0.0 per cent)	0 (0.0 per cent)	2882 (0.7 per cent)	14 984 (3.6 per cent)	35 474 (8.5 per cent)	80 370 (19.3 per cent)	161 940 (38.8 per cent)	94 057 (22.6 per cent)	27 261 (6.5 per cent)
LF	ALIS rock	0 (0.0 per cent)	102 (0.0 per cent)	10 254 (2.4 per cent)	74 665 (17.3 per cent)	60 954 (14.1 per cent)	52 529 (12.1 per cent)	89 392 (20.7 per cent)	101 254 (23.4 per cent)	36 765 (8.5 per cent)	6567 (1.5 per cent)
	DLA deep basin	0 (0.0 per cent)	0 (0.0 per cent)	5 (0.0 per cent)	4949 (1.2 per cent)	17 270 (4.1 per cent)	38 936 (9.3 per cent)	90 154 (21.6 per cent)	153 628 (36.8 per cent)	86 320 (20.7 per cent)	25 708 (6.2 per cent)

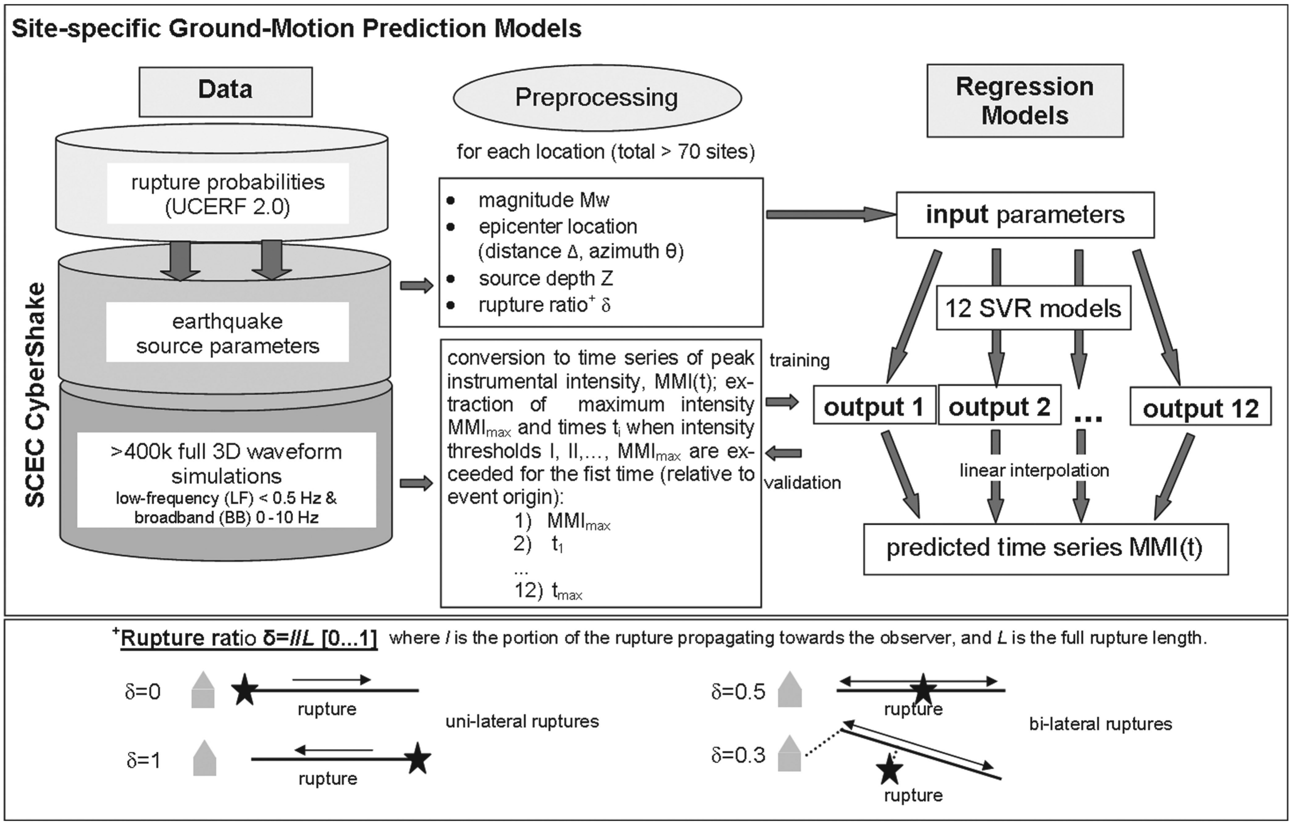

Site-specific Ground-Motion Prediction Models
Data
Preprocessing
Regression Models
SCEC CyberShake
rupture probabilities (UCERF 2.0)
earthquake source parameters
>400k full 3D waveform simulations
low-frequency (LF) < 0.5 Hz & broadband (BB) 0 -10 Hz
for each location (total > 70 sites)
• magnitude Mw
• epicenter location (distance Δ, azimuth θ)
• source depth Z
• rupture ratio+ δ
conversion to time series of peak instrumental intensity, MMI(t); extraction of maximum intensity MMImax and times ti when intensity thresholds I, II,..., MMImax are exceeded for the fist time (relative to event origin):
1) MMImax
2) t1
...
12) tmax
training
validation
input parameters
12 SVR models
output 1
output 2
...
output 12
linear interpolation
predicted time series MMI(t)
+Rupture ratio δ=l/L [0...1] where l is the portion of the rupture propagating towards the observer, and L is the full rupture length.
δ=0
rupture
uni-lateral ruptures
δ=1
rupture
δ=0.5
rupture
bi-lateral ruptures
δ=0.3
rupture

Figure 1. Top panel: scheme of our proposed approach for site-specific ground-motion predictions. The regression models are derived and validated with earthquake source and ground-motion parameters of the pre-processed SCEC CyberShake data set with >400 k full 3-D waveform simulations at 15 BB (0–10 Hz) and 56 LF (<0.5 Hz) sites. Goal is to predict the temporal evolution of instrumental intensity MMI at the user sites (strictly speaking the times when intensity thresholds $\mathrm{MMI}_{\mathrm{thres}} = [\mathrm{I, II}, \ldots, \mathrm{X+}]$ are first exceeded) for large earthquakes in southern California. Bottom panel: one of the input parameters for the regression models is the rupture ratio $\delta = l/L$, where L is the length of the surface projected 2-D rupture and l the portion of the rupture propagating towards the observer. The parameter is a measure of whether the rupture propagates uni- ($\delta = 0$ if the rupture propagates away from the user; $\delta = 1$ if the rupture propagates towards the user) or bilateral ($0 < \delta < 1$). Stars show the points of rupture nucleation, the solid lines the surface projected 2-D line ruptures; the house marks the user location.

the times when intensity thresholds $\mathrm{MMI}_{\mathrm{thres}} = [\mathrm{I, II}, \ldots, \mathrm{X+}]$ are first exceeded) at ∼70 selected test sites in and around the LA basin for large earthquakes in southern California. We seek a simple and robust model that requires a minimum amount of information about the earthquake (hypocentre, magnitude and rupture ratio δ, which characterizes the direction of rupture propagation) and still is capable of providing fast and reliable ground-motion estimates as needed for EEW and other applications.

Correcting ground-motion predictions for 3-D wave-propagation effects, as well as estimating warning times, requires having a large data set of seismic observations (including those of earthquakes with $M > 6.5$) from which these relations can be derived. This requirement is usually not fulfilled, due to the infrequency of $M \geq 6.5$ earthquakes. However, using high-quality 3-D wave-propagation simulations instead is an attractive alternative. We will establish ε-support vector regression (ε-SVR) models from source and ground-motion parameters in the Southern California Earthquake Center (SCEC) CyberShake 1.4 data set with >415 000 rupture scenarios along active faults in southern California (Graves *et al.* 2010). While nowadays synthetic seismograms have limitations, as will be discussed, we use the CyberShake data set in this study to demonstrate their application to EEW as a 'proof of concept'. Clearly, as the simulation and rupture models evolve and improve within the next years, they can be incorporated into updates to our regression models.

Complexities in rupture directivity, site/basin response and 3-D wave-propagation, as well as coupling of these effects, are usually not included in GMPEs. We do not anticipate that these relations can be expressed through simple equations. The aim of this study is rather to develop models that are capable of predicting ground-motions and warning times from a few input parameters without enforcing predefined functional forms or dependencies among these parameters. We rather follow a completely data-driven regression approach.

Even though the relationship between earthquake source parameters (magnitude, location, rupture ratio) and ground-motions is expected to be complex and non-linear, it is of deterministic nature (at least for the long-period motions) and should be well approximated by statistical models. One of the major challenges for the models is to *learn* the relationship between the earthquake magnitude and fault rupture length, as well as the relationship between the finite-fault geometry relative to the point of observation and the level of ground shaking at this site (Fig. 1).

2 DATA

2.1 CyberShake waveform simulations

The SCEC CyberShake data set encompasses around 415 000 3-D wave-propagation simulations for large earthquakes ($6.5 \leq M \leq$

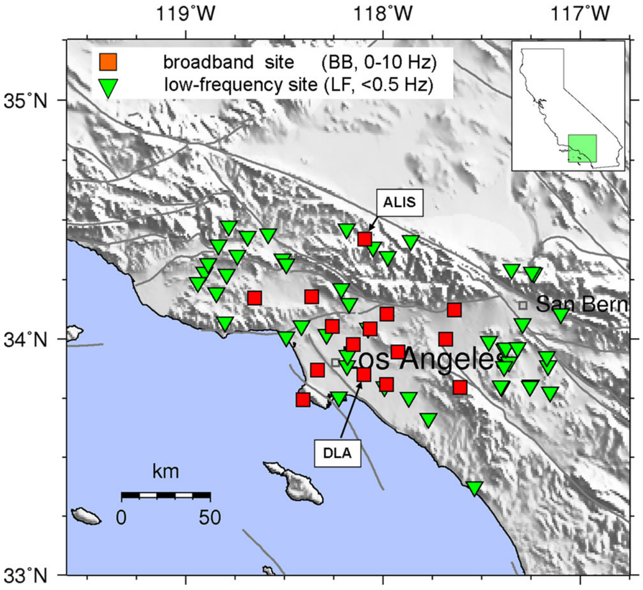

119˚W
118˚W
117˚W
35˚N
34˚N
33˚N
🟥 broadband site (BB, 0-10 Hz)
▼ low-frequency site (LF, <0.5 Hz)
ALIS
DLA
San Bern
Los Angeles
km
0
50

Figure 2. Distribution of 15 BB (red squares) and 56 LF (green triangles) CyberShake sites in the greater LA area as used in this study. Additional data analyses are made at stations ALIS and DLA that are assumed to be representative for rock and the deep LA basin, respectively.

8.5) at around 200 locations in and around the LA basin (Graves *et al.* 2010). One of the goals of the CyberShake project is to develop a physics-based computational approach to PSHA, by using reciprocity to simulate synthetic seismograms for a suite of rupture realizations obtained from the Unified California Earthquake Rupture Forecast, version 2.0 (UCERF 2.0; Working Group on California Earthquake Probabilities 2007). The large repository of simulated waveforms generated by CyberShake also allows for a systematic investigation of other effects, including rupture directivity, ground motion limits and basin response (e.g. Donovan *et al.* 2012; Wang & Jordan 2012; Denolle *et al.* 2014).

The original set of CyberShake simulations consists of low-frequency (LF, <0.5 Hz) 3-D finite-difference simulations computed in the SCEC CVM-4.0 seismic velocity model (Magistrale *et al.* 2000; Kohler *et al.* 2003). For a typical site in the LA region, UCERF 2.0 identifies more than 7000 earthquake ruptures (i.e. 'faults') with moment magnitudes $M \geq 6.5$ that might affect this site. For each of these ruptures, it is important to also capture the possible variability in the earthquake rupture process. To do this, CyberShake creates a variety of hypocentre and slip distributions for each rupture yielding a total of over 415 000 rupture variations, each representing a potential earthquake. The rupture variations are generated using the method of Graves & Pitarka (2005), which produces a detailed kinematic description of slip evolution across the prescribed fault for each scenario earthquake.

Callaghan *et al.* (2011) have recently extended the LF CyberShake results to broadband (BB, 0–10 Hz) for a subset of the originally considered sites (Fig. 2). The BB results were computed by adding 1-D semi-stochastic high-frequency (HF) components to the existing 3-D deterministic LF results using the hybrid simulation methodology of Graves & Pitarka (2010). The transition from the semi-stochastic to deterministic frequency bands is at 0.5 Hz. In this paper we consider both the LF and BB simulations at 56 and 15 selected test sites, respectively (Fig. 2). Additional data analyses in this study are made at stations ALIS (Aliso; 34.42°, −118.09°; $Vs30 = 724$ m s^{-1}) and DLA (Del Amo; 33.848°, −118.096°; $Vs30 = 301$ m s^{-1}), which are considered representative for sites located on rock and for sites where sediments in the LA basin reach maximum thickness, respectively. In the remainder of this paper, we will refer to these sites as 'rock' and 'deep basin', respectively. For these two selected sites there are both BB and LF simulations available.

2.2 Comparison of CyberShake simulations with GMPEs

Unlike most current GMPEs, 3-D wave-propagation simulations directly incorporate directivity and basin response effects. For comparison of the two approaches we analyse in Fig. 3 the residuals of pseudo-spectral acceleration (PSA), $\log_{10}[\mathrm{PSA}_{\mathrm{CyberShake}}(T)]-\log_{10}[\mathrm{PSA}_{\mathrm{GMPE}}(T)]$, at two selected sites, ALIS and DLA, at various periods ($T = 0.1$–10 s) and across the entire set of >415 k CyberShake ruptures. Here, PSA refers to the geometric mean of the two horizontal components. For the GMPEs we select relations by Boore & Atkinson (2008; BA08) and Campbell & Bozorgnia (2008; CB08) as references; the latter accounts for basin effects through an additional term that specifies the depth beneath the site to a shear wave velocity of 2.5 km s^{-1} (Z2.5). As noted earlier ALIS and DLA are considered representative for rock and the deep LA basin, respectively; the basin depth at DLA is measured from CVM-4.0 to be $\mathrm{Z2.5} = 5.3$ km.

For the rock site ALIS there is good agreement between the wave-propagation simulations and the two GMPEs at all periods T (mean $\bar{E} = 0.0$ and standard deviation $\sigma = 0.23$; Fig. 3). For the deep basin site DLA, however, the simulations predict higher PSA levels than the GMPEs for $T > 1$ s ($\bar{E} = 0.09$, $\sigma = 0.22$ using CB08; $\bar{E} = 0.31$, $\sigma = 0.31$ using BA08). For BA08, we suspect the errors get larger for increasing periods because basin effects are not explicitly considered in this relation. This is expected because there are smaller basin response effects for the shorter periods, and because the HF simulation in CyberShake does not explicitly include the 3-D basin.

A systematic comparison of CyberShake simulations and GMPEs was performed by Wang & Jordan (2012), who used an averaging-based factorization scheme to facilitate a geographically explicit comparison of seismic hazard models derived from the two approaches. Generally, the GMPEs tend to predict lower long-period ground-motions in the LA basin compared with the waveform simulations. This shows the need for models that include 3-D wave-propagation effects.

3 PRE-PROCESSING

3.1 Estimating MMI from CyberShake waveforms

We apply empirical relations by Worden *et al.* (2012) to convert the CyberShake velocity waveforms into time-series of peak instrumental intensities, MMI(t). To avoid differentiation of the CyberShake velocity waveforms and to keep computational efforts small, we apply the peak ground velocity (PGV)-to-MMI relations for all intensities, even though smaller intensities (MMI ≤ V) tend to correlate slightly better with HF parameter peak ground acceleration (PGA; Worden *et al.* 2012).

First we convert each velocity amplitude to the corresponding intensity value; we do this for each of the >400 k rupture scenarios and for each of the 71 selected test sites (Fig. 2). Then for each time t we determine if the intensity has increased or decreased compared to the previous time (t–dt) and keep the larger of both values: $\mathrm{MMI}(t) = \max[\mathrm{MMI}(t), \mathrm{MMI}(t\text{–d}t)]$. We only consider the

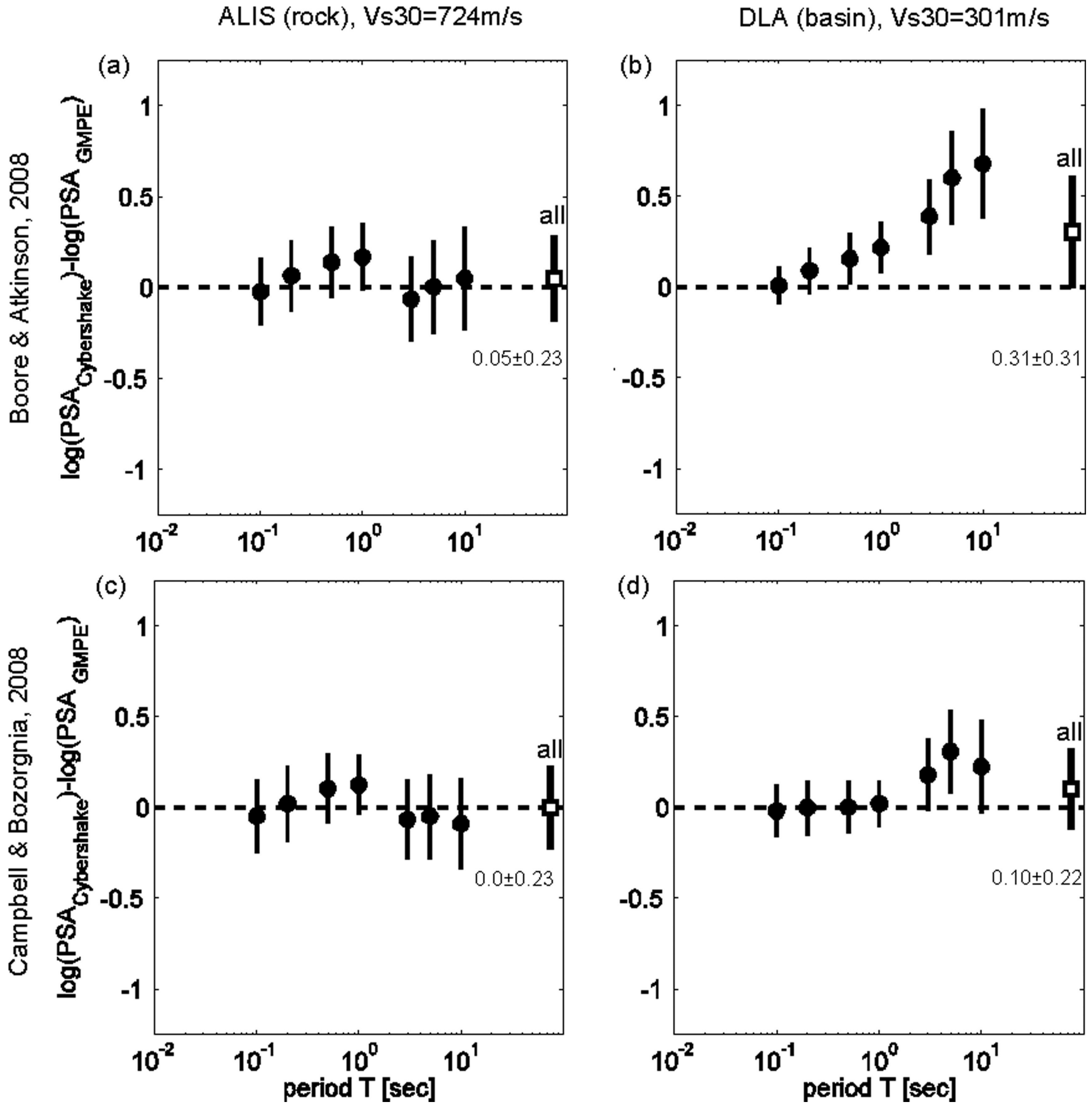


Figure 3. Comparison of pseudo-spectral acceleration PSA at different periods T determined from the CyberShake BB simulations and GMPEs by Boore & Atkinson (2008; BA08) and Campbell & Bozorgnia (2008; CB08) using the geometric mean of the horizontal components for all >400 k rupture scenarios in the CyberShake data set. Dots and bars show the mean $\bar{E}$, and standard deviation σ of the error distributions; the open squares and thick lines show the corresponding values taken over all seven periods. For the rock site ALIS there is good agreement between the wave-propagation simulations and the two GMPEs for all periods (a and c). For the deep basin site DLA (b and d), however, the simulations predict for $T > 1$ s higher PSA levels than the GMPEs; the residuals get larger for increasing periods for BA08 (b), since basin effects are not explicitly considered in this relation.

larger value determined from the waveforms of both horizontal components. This gives us, for each user site and for each rupture scenario, a time-series describing the temporal evolution of peak intensity with intensity values either increasing or being constant over time (Fig. 4).

Next we determine the times when intensity thresholds $\mathrm{MMI}_{\mathrm{thres}}$ = [I, II, . . . , X+] are exceeded for the first time in the obtained time-series, which gives us $t_1, t_2, \ldots, t_{10}$. Note that all CyberShake waveforms start with the earthquake origin time (Fig. 4). If a certain intensity threshold n is not exceeded in the entire time-series, the corresponding time t_n is not considered. The maximum intensity value $\mathrm{MMI}_{\mathrm{max}}$, the time when this value is reached t_{max} and times $t_1, t_2, \ldots, t_{10}$ are the 12 target values for our regression models (Fig. 1).

Table 1 compares the distributions of $\mathrm{MMI}_{\mathrm{max}}$ at ALIS (rock) and DLA (deep basin). As expected, we observe stronger shaking on average in the basin than on rock. While there is significant range in the maximum intensity for the set of ruptures at both sites, the distributions of intensity values are peaked at $\mathrm{MMI}_{\mathrm{max}}$ = VIII, which corresponds to ‘severe’ shaking. This is to be expected because there are a significant number of large magnitude ruptures on major faults in the CyberShake data set (e.g. various San Andreas scenarios) that will produce strong intensities. However, many of these scenarios have low probability of occurrence, which is not accounted for in the results in Table 1. Since our models assume that all the ruptures are equally likely, changing the rupture rates or probabilities will not affect the results.

3.2 Comparison of MMI for LF and BB CyberShake simulations

All 71 selected test sites consist of LF (<0.5 Hz) waveforms, with just a subset of sites having full BB (0–10 Hz) results currently available (Fig. 2). Clearly, the BB results provide a more

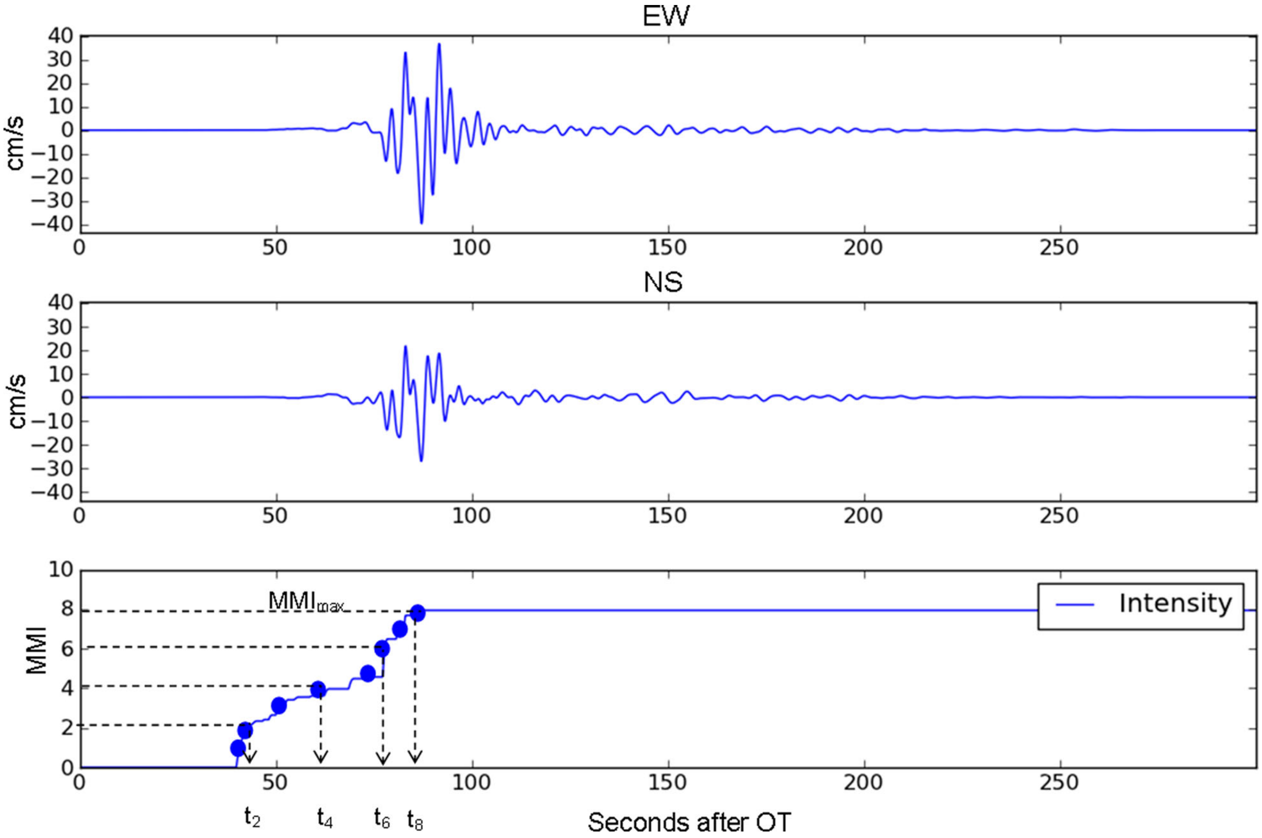


Figure 4. Instrumental intensity (MMI) is estimated from the CyberShake velocity waveforms using relations by Worden *et al.* (2012). We determine the times when intensity thresholds MMI$_{\mathrm{thres}}$ = [I, II, . . . , X+] are exceeded for the first time in the obtained time-series. These times, t_1, t_2, . . . , t_{10}, along with the maximum intensity MMI$_{\max}$ and the time when this value is reached, $t_{\max}$, are the target values for our regression models (Fig. 1).

comprehensive representation of the ground motion response compared to the LF-only results. However, it would be beneficial if we could supplement the CyberShake BB simulations with the LF-only simulations for establishing the regression models in this study, because the LF simulations provide a broader and more regular geographic distribution of sites throughout the LA basin region.

To assess the adequacy of using the LF-only simulations in developing our models, we compare MMI values derived from the BB and LF waveforms at the two selected sites, ALIS and DLA (Table 1 and Fig. 5). In general, both MMI$_{\max}$ and the arrival times of incremental MMI are in good agreement for the BB and LF simulations. For MMI$_{\max}$, there is a tendency for the LF data to predict slightly lower values than the BB, particularly for the rock site ALIS, although the majority of the values agree within 0.5 MMI units. For the arrival times, there is a tendency for the BB results to predict slightly shorter times than the LF results, particularly for the smaller MMI values at the basin site DLA.

In our current analysis, we view the level of misfit between the BB and LF results to be of minor significance, especially at the larger MMI values for which ground motion predictions and early warning are most needed. Also, as was pointed out earlier, this study is mainly intended to give a 'proof-of-concept', independent from the characteristics of the underlying data. In the following analyses we will establish BB models for sites at which BB data is available, for the remaining sites we derive LF-only models. Since the differences between the predictions of the two model types are generally small (Fig. 5), we will plot them in some of the following figures together (Figs 7 and 12) using distinct symbols.

4 REGRESSION MODELS

4.1 ε-SVR

The aim of this study is to develop models that are capable of predicting ground-motions and warning times from a few input parameters without enforcing predefined functional forms or dependencies among these parameters. SVR is one of the most popular approaches in machine learning (Smola & Schölkopf 2004) that is suited for multidimensional non-linear regression. The desired mapping relations are determined from sets of example or training patterns. SVR favours smooth models that are not overfitted to the training data which is prerequisite for a high generalization capability towards unseen data. Using kernel functions the input parameters are implicitly mapped into a higher (infinite) dimensional feature space where linear regression can be performed. In this study, we apply a special type of SVR, called ε-SVR (Vapnik 1995; Smola & Schölkopf 2004). See the Appendix for details.

4.2 Model parameters and training

For a given earthquake and user location our goal is to predict the maximum instrumental intensity MMI$_{\max}$, the time $t_{\max}$ when this value will be reached (relative to rupture nucleation), as well as when each of the intensity thresholds MMI$_{\mathrm{thres}}$ = [I, . . . , X+] will be exceeded for the first time, that is times $t_1, \ldots, t_{10}$. These are the 12 target metrics z_i, $i = 1, \ldots, n$, in our regression models (see the Appendix), where n is the number of training patterns.

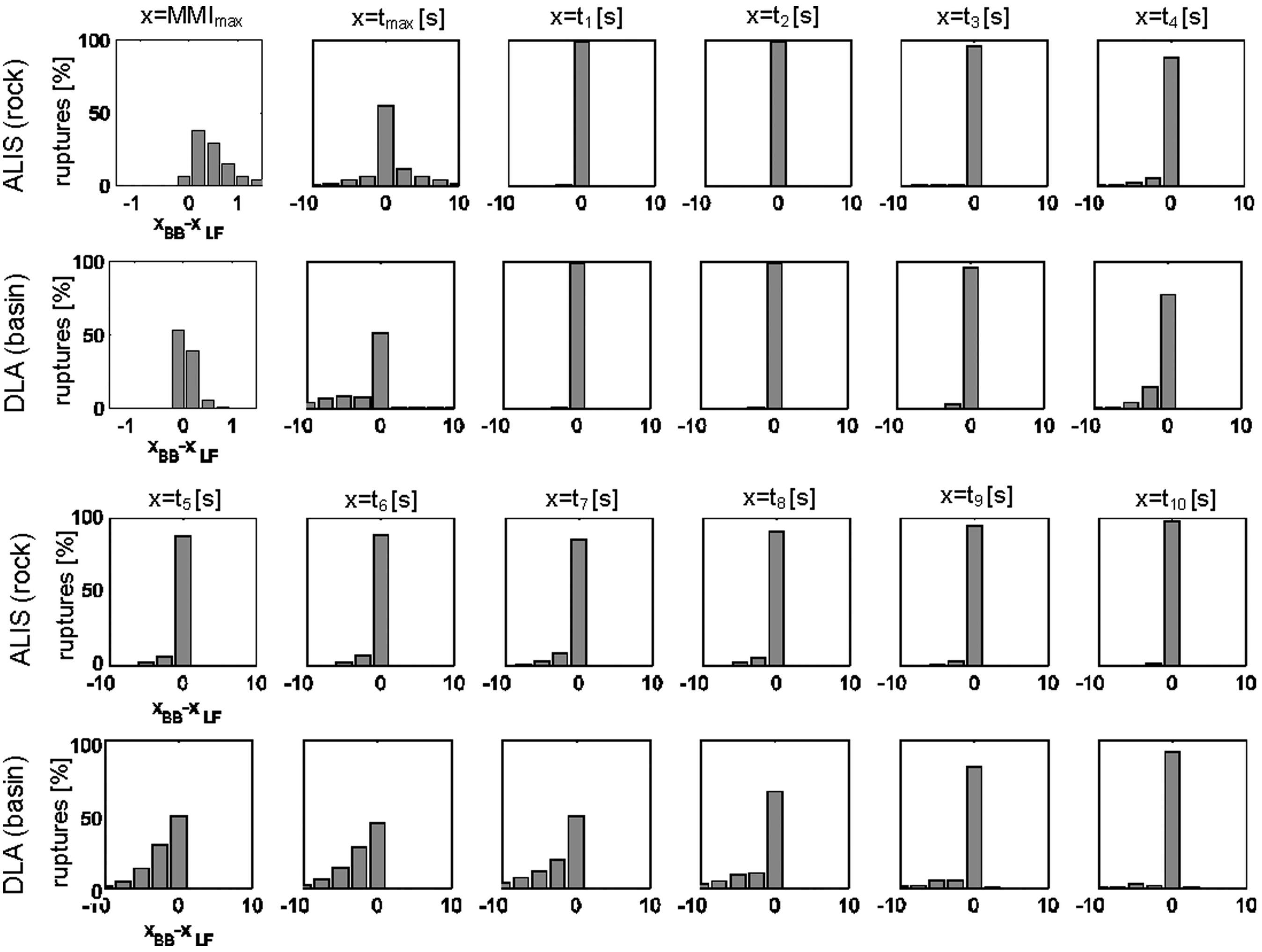


Figure 5. Comparison of the 12 metrics (maximum intensity MMI$_{\max}$, and times $t_1, \ldots, t_{\max}$) characterizing the temporal evolution of instrumental intensity MMI(t) for the Cybershake BB and LF simulations at ALIS (rock, top panel) and DLA (deep LA basin, bottom panel). Shown are the histograms of the corresponding residuals $x_{\mathrm{BB}} - x_{\mathrm{LF}}$. For the times $t_1, \ldots, t_{10}$ we consider only events that reach the corresponding MMI level for both BB and LF. The metrics largely agree for the majority of BB and LF-only simulations, in particular for the larger MMI values for which ground-motion predictions and early warning are most needed.

We select simple and easily calculable features of the earthquake to make our approach applicable to real-time procedures such as EEW. We expect the target output values to depend on (moment) magnitude M, (epicentral) distance to the user site R, source depth Z, (back-)azimuth between the user site and the earthquake θ and on the rupture ratio δ; the latter characterizes whether the rupture propagates mainly unilateral (towards or away from the user) or bilateral (Fig. 1, bottom panel).

How quickly these parameters can be determined in an operational system, depends mainly on the station density and data latencies in the seismic network where the EEW algorithms are applied. For instance, real-time tests of the *CISN ShakeAlert* system in California have shown that event magnitudes and locations can be determined within 5 s from event origin with uncertainties of ±0.6 magnitude units and ±15 km, respectively, if seismic sensors are located within 10–15 km from the epicentre (Böse *et al.* 2013). These estimates can be updated as more data is received, and errors usually decrease within a couple of seconds to ±0.4 magnitude units and <4 km, respectively. Very large earthquakes, such as the 2011 M9.0 Tohoku earthquake in Japan, can be difficult to be recognized quickly from the observations of the initial shaking at the close-by sensors (Hoshiba *et al.* 2011). First magnitude predictions still tend to provide an estimate of the lower bound of earthquake magnitudes (Kanamori 2005).

The azimuth θ between the user and earthquake location has periodicity every 360º. For instance, earthquakes at $\theta = 359^\circ$ and 1° are very close to each other in space for the same R (for instance, they are less than 5 km apart if $R < 150$ km). To account for a smooth transition every 360°, we use a simple trick by defining two separate feature parameters, $\cos(\theta)$ and $\sin(\theta)$; each of these trigonometric functions produces a smooth output with values $[-1\ldots+1]$, and, at the same time, allow reconstructing the original azimuth θ from Euler's formula $\exp(i\theta) = \cos(\theta) + i\sin(\theta)$. Note that θ is the azimuth between the user and the earthquake epicentre, not the strike of the rupturing fault. Instead of using the polar coordinates R and θ to characterize the epicentre, we could use the geographic longitude and latitude.

The rupture ratio parameter δ is calculated as $\delta = l/L$, where l is the rupture length from the epicentre to the rupture end that is closest to the user; L is the total length of the surface projected 2-D line rupture (Fig. 1, bottom panel). The parameter δ can take values $[0\ldots1]$, where $\delta = 0$ indicates that the rupture propagates away and

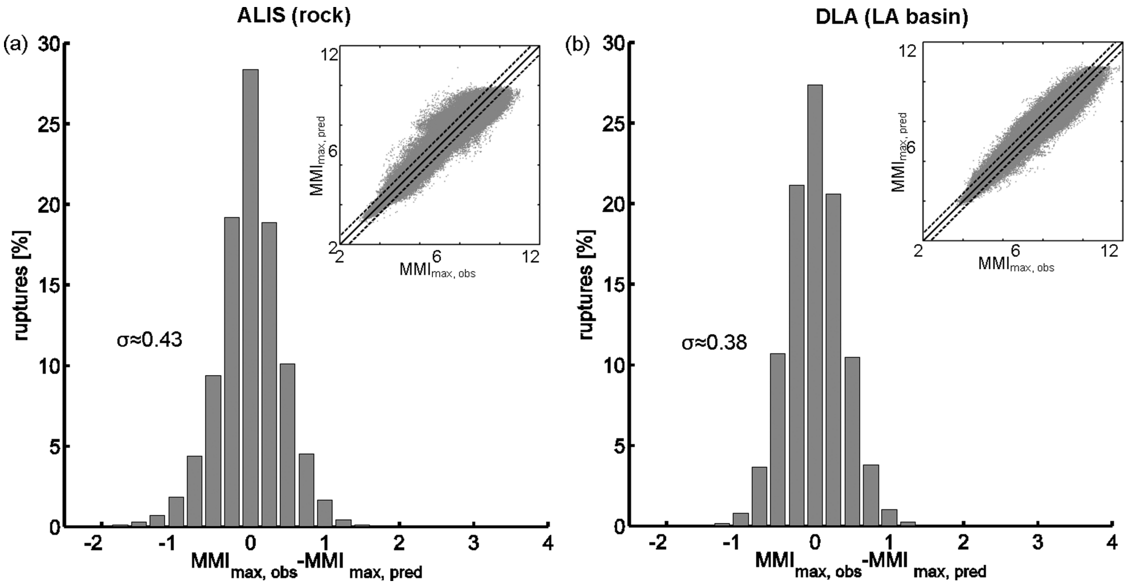


Figure 6. Predicted versus observed (CyberShake simulated) maximum instrumental intensity $\mathrm{MMI}_{\mathrm{max}}$ at (a) ALIS (rock) and (b) DLA (deep LA basin). Although only 20 per cent of the CyberShake BB waveforms at the two sites were used to establish the models, they can explain $\mathrm{MMI}_{\mathrm{max}}$ of all >400 k rupture scenarios with $\sigma \approx 0.43$ (ALIS) and $\sigma \approx 0.38$ (DLA), respectively.

$\delta = 1$ towards the user; $0 < \delta < 1$ characterizes bilateral rupture propagation.

We define the input feature vector $x_j \in R^6$ (see the Appendix) as

$$x_i = \{\tilde{M}_i, \tilde{R}_i, \tilde{Z}, \cos(\theta_i), \sin(\theta_i), \tilde{\delta}_i\}, \qquad i = 1, \ldots, n, \qquad (1)$$

where ~ denotes that M, R, Z and δ were linearly scaled to fall into the range $[-1 \ldots +1]$.

In this study we use LIBSVM (Chang & Lin 2011), a free library for support vector machines and regression (http://www.csie.ntu.edu.tw/~cjlin/libsvm/), to determine and test our SVR models. For each user site and each target value z_i characterizing MMI(t), we determine a separate SVR model. We randomly select 20 per cent of the CyberShake rupture scenarios to establish our regression models. For each site, we train a total of 12 finite-fault models to predict the maximum intensity ($\mathrm{MMI}_{\mathrm{max}}$) and the temporal evolution ($t_1, t_2, \ldots, t_{\mathrm{max}}$) using the input vector in eq. (1) (Fig. 1). We use only a small subset of the whole CyberShake data set to establish our prediction models (1) to demonstrate that >415 k are not required to obtained stable models (20 per cent appears to be sufficient), (2) to prove the high generalization capability of our models by using a large test set (80 per cent) that is unknown to the models but for which accurate predictions of MMI can be achieved as will be shown and (3) to keep the computational efforts small when deriving the models.

5 RESULTS

In the following subsections, we will assess the accuracy of our ground-motion prediction models through comparison with parameters derived from the CyberShake waveform simulations. For model validation we compare our predictions with seismic observations during the 2008 M5.4 Chino Hills earthquake in southern California. Furthermore, we show that there are fundamental differences in the predicted ground motions depending on whether the earthquake fault rupture propagates towards or away from the observer. Finally, we develop the concept of a simple and robust magnitude-threshold based early warning algorithm for southern California.

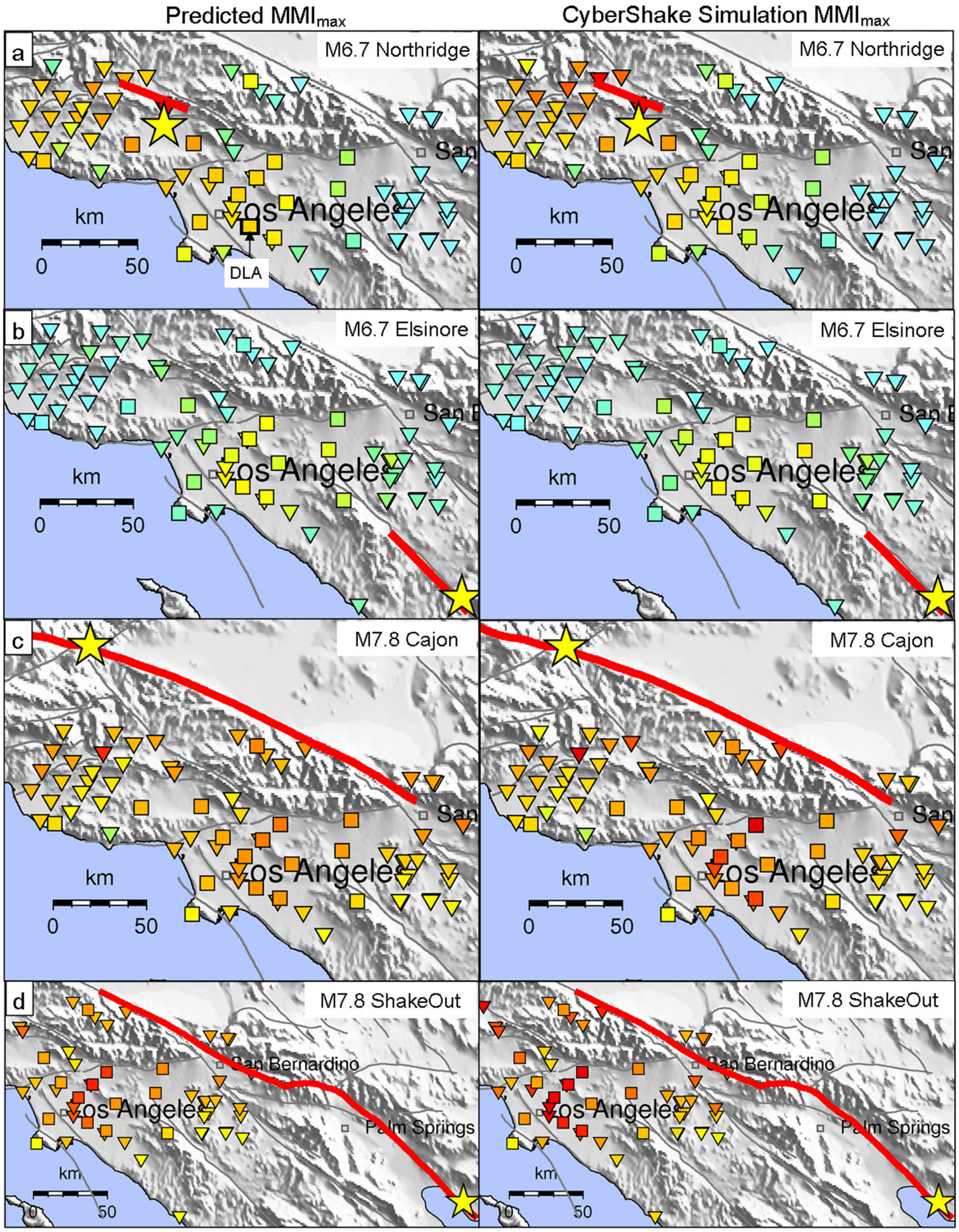
5.1 Prediction accuracy

To assess the prediction accuracy of our finite-fault regression models we analyse their performance for the two selected BB sites, ALIS and DLA, that are assumed to be representative for rock and the deep LA basin, respectively. Though only 20 per cent of the CyberShake simulations were used for training, the regression models can explain the maximum intensity values $\mathrm{MMI}_{\mathrm{max}}$ of all >415 000 rupture scenarios with a standard deviation of $\sigma \approx 0.43$ at ALIS (rock) and $\sigma \approx 0.38$ at DLA (deep basin), respectively (Fig. 6). The standard deviations remain the same if the training and test data sets are analysed separately from each other. We perform a fivefold cross-validation to confirm these values. The order of these intensity prediction errors is representative for all 71 test sites.

We believe that the slightly larger errors at ALIS were caused by the 2-D line-source approximation of fault ruptures as needed for the calculation of rupture ratio δ (Fig. 1, bottom panel); this simplification may be problematic for earthquakes with ruptures through the 'Big Bend' section of the San Andreas Fault with a strong divergence from a line source. Since ALIS is located closer to the fault (Fig. 2), $\mathrm{MMI}_{\mathrm{max}}$ predictions at this site are more strongly affected by this shortcoming than at DLA (Fig. 6).

5.2 Model application: examples

In the following, we apply our models to predict the shaking intensities at the 71 selected sites (15 BB and 56 LF) for four rupture scenarios that are part of the CyberShake data set (Figs 7 and 8): (i) along the Pico thrust fault with nucleation point close to the 1994 M6.7 Northridge earthquake, (ii) along the Elsinore strike-slip fault with nucleation close to Temecula, (iii) along the San Andreas strike-slip fault with a bilateral rupture starting 45 km northwest of Lake Hughes and terminating at the Cajon pass, close to the assumed rupture termination of the 1857 M7.8 Fort Tejon



PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	none	none	none	Very light	Light	Moderate	Mod./Heavy	Heavy	Very Heavy
PEAK ACC.(%g)	<0.05	0.3	2.8	6.2	12	22	40	75	>139
PEAK VEL.(cm/s)	*<0.02*	*0.1*	*1.4*	*4.7*	*9.6*	*20*	*41*	*86*	*>178*
INSTRUMENTAL INTENSITY	I	II–III	IV	V	VI	VII	VIII	IX	X+

Figure 7. Predicted (left-hand panel) and CyberShake simulated (right-hand panel) maximum intensities MMI$_{max}$ in the larger LA area for four scenario earthquakes; yellow stars mark the points of rupture nucleation (epicentre), red lines the surface-projected 2-D fault ruptures. Each square and each triangle represents the target site of an individual BB (square) or LF (triangle) model; the colour codes MMI$_{max}$ at this site. Even though the estimates at different sites are independent from each other, the maps provide highly consistent pictures of ground-motions variations in and around the LA basin and show an excellent agreement with the simulated intensities.

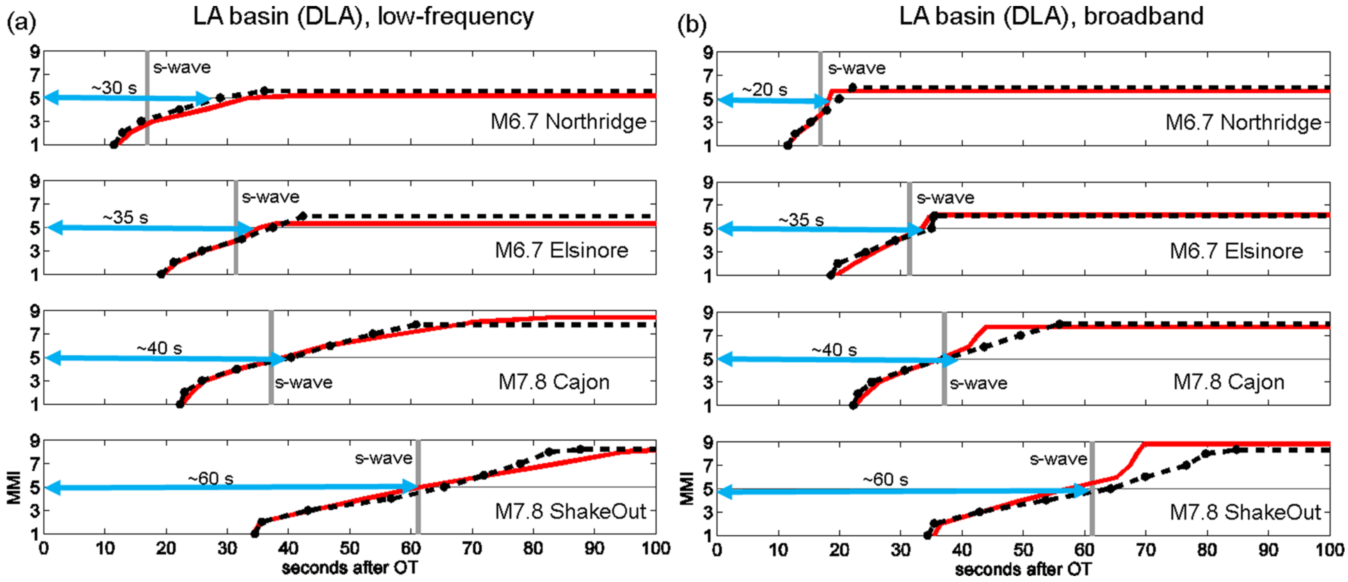


Figure 8. Our models predict the temporal evolution of instrumental intensity, MMI(t), within a fraction of a second using only information on the location of the earthquake, its magnitude, and rupture ratio δ. Red lines show MMI(t) derived from the CyberShake simulations at basin site DLA for the four scenario earthquakes in Fig. 7; the dashed black lines show the corresponding predictions from our finite-fault models. The time-series are obtained from linear interpolation between the predicted times $t_1, t_2, \ldots, t_{\max}$ when MMI thresholds I to MMI$_{\max}$ are exceeded for the first time. Both (a) LF and (b) BB results are shown. The theoretical arrival of the S wave assuming a constant shear wave velocity of 3.55 km s^{-1} is marked by grey bars. The blue arrows show the time to MMI = V.

earthquake and (iv) the earlier described *M*7.8 ShakeOut scenario earthquake (Jones *et al.* 2008) along the southern portion of the San Andreas Fault with nucleation point close to Bombay Beach. Since 20 per cent (randomly selected) of the CyberShake simulations were used to establish the regression models, there is some chance that single data points were used for training; the majority of points, though, are new, that is unknown, to the models.

Ground-motion estimates at the 71 test sites come from different models and are thus independent from each other. Still the maps in Fig. 7 provide highly consistent pictures of ground-motion variations in and around the LA basin and show an excellent agreement with the simulated intensities from the CyberShake data set. In all four scenarios ground-motions in the LA basin are strongly amplified relative to the surrounding rock.

In addition to maximum intensity MMI$_{\max}$ the regression models also predict the temporal evolution of instrumental intensity, MMI(t), at the test sites. Strictly speaking, they predict when intensity thresholds MMI$_{\text{thres}}$ = [I, II, ..., MMI$_{\max}$] are exceeded for the first time relative to the event origin time (Fig. 3). The temporal evolution MMI(t) is obtained from linear interpolation between the times $t_1, t_2, \ldots, t_{\max}$ (Fig. 1, right-hand panel). Since these estimated times are independent from each other, it can happen in some cases that $t_n < t_{n-1}$ if t_n and t_{n-1} are very close to each other, that is if the change from one intensity level to next higher is very quick. Therefore we apply a moving average procedure to smooth the predicted time values and thus to obtain an even intensity evolution.

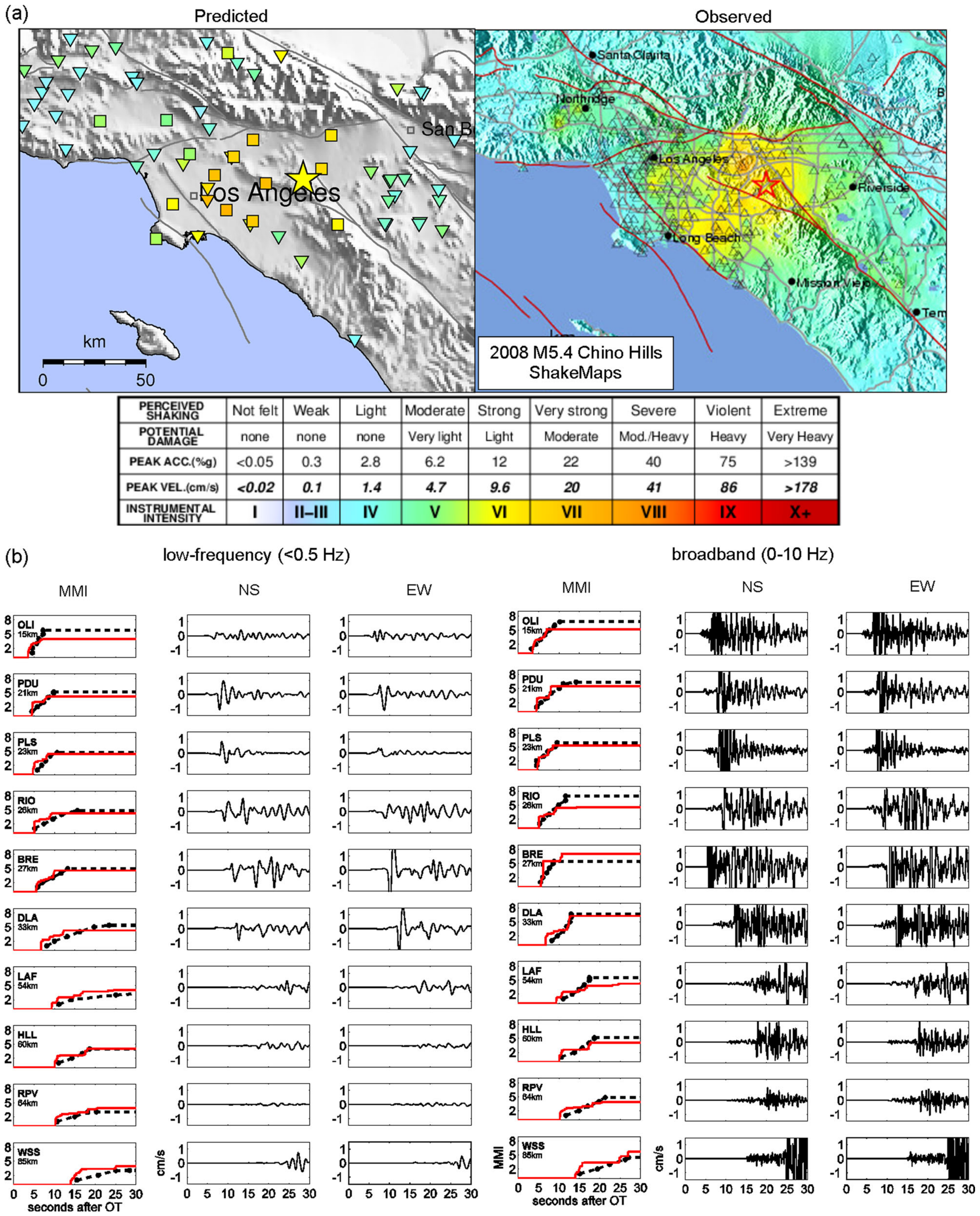
Fig. 8 shows the predicted and observed (CyberShake simulated) temporal evolution of MMI at basin site DLA for the four scenario earthquakes in Fig. 7. As expected the models are more accurate in predicting the long-period deterministic than the semi-stochastic HF motions (Fig. 8b). In all four scenarios intensity level MMI = V (‘moderate’ shaking) is exceeded once the direct *S*-wave arrives, which occurs between 20 s (for the Northridge earthquake) and 60 s (for the ShakeOut scenario) after rupture nucleation; larger intensity values are reached 10 or more seconds later.

As described earlier, the maximum intensity in the LA basin for the ShakeOut scenario is simulated as MMI = VIII-IX (‘severe’ to ‘violent’ shaking). Using a point-source approximation of this earthquake with GMPEs by Cua (2005) predicts ‘light’ to ‘moderate’ shaking (MMI = IV–V) only. Using rupture-to-site instead of hypocentral distances through the application of the FinDer algorithm (Böse *et al.* 2012a) predicts ‘severe’ shaking (MMI = VIII). However, it takes around 30 s after rupture nucleation before the rupture in this scenario has reached a critical length that suggests that a warning in LA is needed. Using the finite-fault regression models from this study (where FinDer is used only to determine the normalized rupture ratio parameter δ rather than the full rupture length), MMI is estimated as MMI = VIII–IX (‘severe to violent’ shaking; Fig. 8), which agrees well with the wave-propagation simulations by Graves *et al.* (2008). We will see later that an accurate magnitude determination is actually not needed in this scenario. Assuming that the magnitude is determined within 5–10 s from rupture nucleation, warning times for users in the LA basin for the ShakeOut scenario earthquake could be on the order of 60 s.

Maps and time-series as shown in Figs 7 and 8 can be calculated from our finite-fault regression models within a fraction of a second (as needed for EEW). They only require estimates of event magnitude, the rupture nucleation point and whether the rupture propagates uni- or bilaterally (as quantified through rupture ratio parameter δ). The fault rupture extent (Fig. 7, red lines) is plotted for visualization only.

5.3 Model validation

For model validation we compare the predictions of MMI$_{\max}$ and MMI(t) with seismic observations during the 2008 *M*5.4 Chino Hills earthquake (Fig. 9), caused by oblique slip faulting along the Yorba Linda fault. While finite source effects, for which our models were established, are generally small for an earthquake of this size, the Chino Hills earthquake is the best-recorded



PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	none	none	none	Very light	Light	Moderate	Mod./Heavy	Heavy	Very Heavy
PEAK ACC.(%g)	<0.05	0.3	2.8	6.2	12	22	40	75	>139
PEAK VEL.(cm/s)	*<0.02*	*0.1*	*1.4*	*4.7*	*9.6*	*20*	*41*	*86*	*>178*
INSTRUMENTAL INTENSITY	I	II–III	IV	V	VI	VII	VIII	IX	X+

Figure 9. Model verification using the 2008 *M*5.4 Chino Hills earthquake. (a) Comparison of predicted instrumental intensities $\mathrm{MMI}_{\mathrm{max}}$ at the 71 test sites, and observed intensities in the USGS ShakeMap (http://earthquake.usgs.gov). On average the models predict slightly higher intensity values, but otherwise agree well with the seismic observations during the earthquake. (b) Comparison of the predicted (dashed black line, left-hand column) and observed (red line) temporal evolution of seismic intensity MMI at 10 SCSN stations (<100 km) for the LF (<0.5 Hz, left-hand panel) and BB (0–10 Hz, right-hand panel) models. Accelerograms were integrated and filtered (for the LF models) to make them comparable to the CyberShake data sets (middle and right-hand column). The event magnitude of the Chino Hills earthquake is assumed to be *M*6.25, because *M*5.4 is outside of the training range of the regression models and the prediction results become otherwise unstable; this explains why $\mathrm{MMI}_{\mathrm{max}}$ is slightly overestimated. In any case, there is good overall agreement between the observed and predicted temporal evolution of MMI.

moderate-sized earthquake in the LA area with reliable time information (Hauksson *et al.* 2008). Other LA area events such as the 1987 *M*5.9 Whittier-Narrows and 1994 *M*6.7 Northridge earthquakes were considered for this validation, but nearly all of the ground motion observations for these events lack absolute timing information, which makes them unsuitable for testing of our models. The recent 2010 *M*7.2 El Mayor-Cucapah earthquake was also considered for this test, but it is so distant from the LA region that the observed intensities were quite low, plus it is located beyond the 200 km source-to-site limit used in generating the CyberShake data set (Graves *et al.* 2010). In contrast to other significant earthquakes in southern California, waveform simulations of the Chino Hills earthquake are not included in the CyberShake data set. The Chino Hills earthquake is thus considered as an independent event that is suited for model validation.

A drawback of this choice, however, is that the magnitude of the Chino Hills earthquake is around one full unit smaller than the minimum magnitude in the CyberShake data set ($6.5 \leq M \leq 8.5$), and is thus outside of the training range of our regression models. Therefore, for this test we increase the event magnitude in this example to *M*6.25; this is the minimum magnitude our models start to produce stable results. Our finite fault models should not be applied to smaller earthquakes. We suspect that the higher magnitude explains why the predicted $\mathrm{MMI}_{\max}$ values are on average slightly higher than in ShakeMap, which shows the interpolated observed instrumental intensities for the Chino Hills earthquake (Fig. 9a).

The temporal evolution MMI(*t*) at 10 randomly selected seismic stations (<100 km) for the LF (<0.5 Hz) and BB (0–10 Hz) models is shown in Fig. 9(b). The corresponding accelerograms were downloaded from the Southern California Seismic Network (SCSN, www.scsn.org), integrated and low-pass filtered (for the LF models) to allow comparison with the CyberShake simulations. The conversion of the velocity waveforms to time-series of instrumental intensities, MMI(*t*), is analogous to the earlier described preprocessing of the CyberShake data set using relations by Worden *et al.* (2012). Again, because we had to assume a higher magnitude (*M*6.25), $\mathrm{MMI}_{\max}$ is slightly overestimated by our models; however, overall the observed and predicted temporal evolution of MMI for the Chino Hills earthquake agree well, verifying the first-order applicability of the regression models to real earthquakes (Fig. 9b).

5.4 The role of rupture ratio δ

Fig. 10 shows an example of the dependency of MMI on magnitude and rupture ratio δ for a fixed epicentre-observer location pair as determined from the regression models. The observer in this example is located in the LA basin at site DLA, the rupture nucleation point is on the San Jacinto Fault at 160 km distance. Overall, the isoseismals in Fig. 10 show that, as desired, support vector regression produces smooth ground-motion prediction models that are not overfitted to the training data. As expected small rupture ratios (that is the rupture propagates mainly away from the observer) require larger magnitudes to cause the same shaking as a rupture that propagates towards the observer. For instance, to cause 'strong' shaking (MMI = VI) at DLA, the magnitude for $\delta = 0$ needs to be almost one unit larger (*M*7.4) than for $\delta = 1$ (*M*6.55).

In the following, we restrict potential earthquake locations to those from the CyberShake data set (as defined in the UCERF 2.0 earthquake rupture forecast) and use our regression models to predict $\mathrm{MMI}_{\max}$ in the LA basin at site DLA for a magnitude *M*6.5 and *M*7 earthquake at depth $Z = 10$ km (Fig. 11); the corresponding fault ruptures are assumed to propagate either away ($\delta = 0$) or towards the user ($\delta = 1$). Typical rupture lengths of magnitude *M*6.5 and *M*7 earthquakes are ∼10 and ∼60 km, respectively (Wells & Coppersmith 1994).

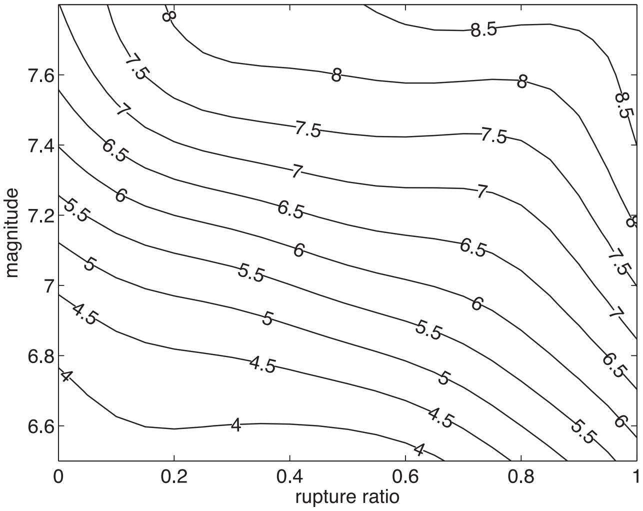
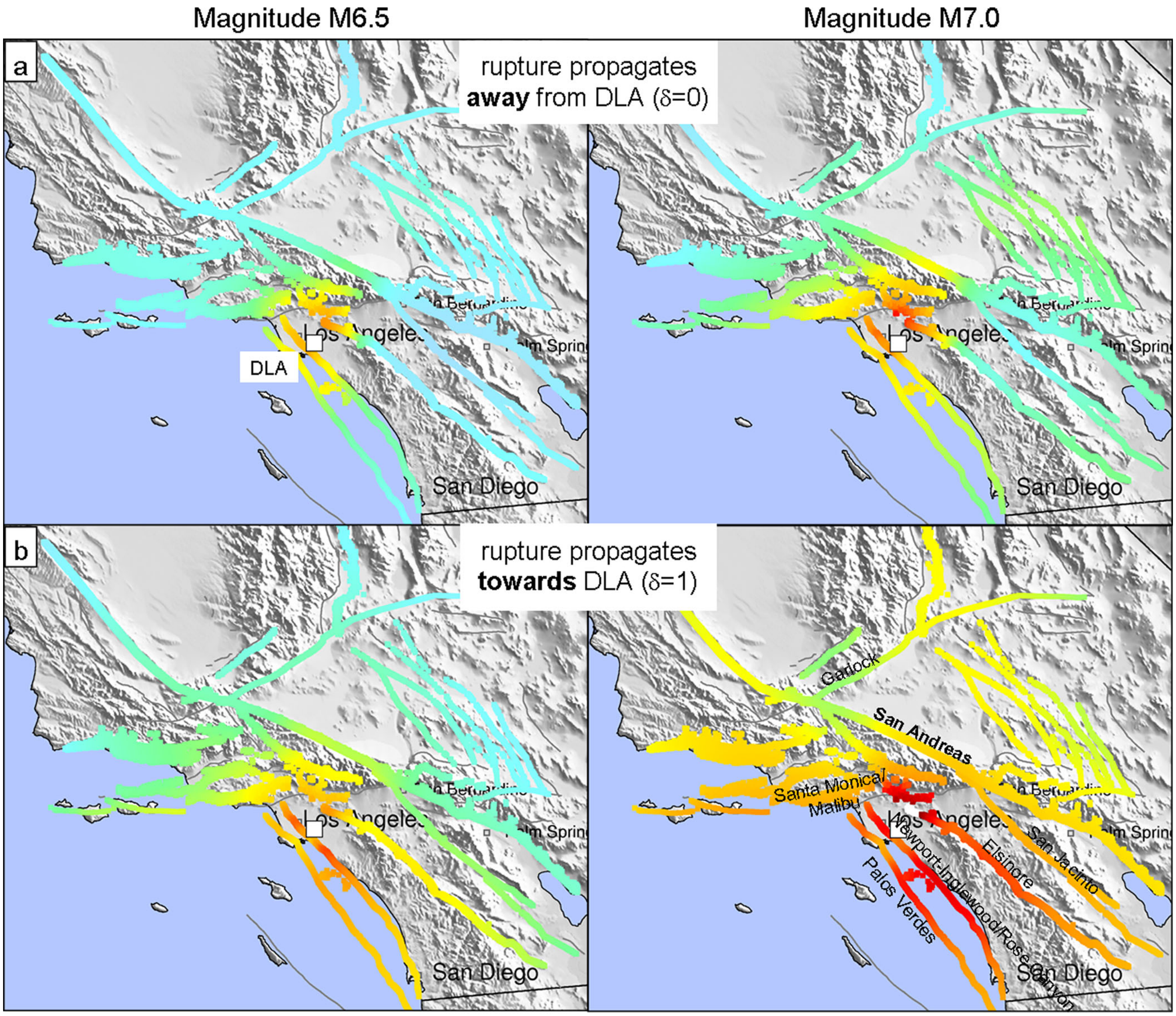


Figure 10. MMI as a function of magnitude and rupture ratio δ for a given observer-earthquake location pair as predicted by the regression models. The user in this example is located in the LA basin (DLA), the rupture nucleation point is 160 km southeast along the San Jacinto Fault. As expected small rupture ratios (that is the rupture propagates mainly away from the observer) require larger magnitudes to cause the same shaking as a rupture propagating to the opposite direction. Overall, support vector regression produces smooth ground-motion prediction models (as preferred) with quite simple isoseismals.

As expected, close earthquakes (e.g. along the Sierra Madre, Raymond, northern Newport-Inglewood faults) tend to cause stronger shaking than those at larger distances (Fig. 11a). However, this is only true if $\delta < 0.5$; if the rupture propagates towards the observer, earthquakes at larger distances pose a significant and in some case an even larger threat than close events (Fig. 11b). Obviously, it does not need to be a very large earthquake for a user in the LA basin to experience significant shaking: a relatively distant moderate *M*6.5–7 earthquake along the Palos Verdes, Newport-Inglewood/Rose Canyon, Elsinore or San Jacinto faults with a rupture propagating towards LA have the potential to cause 'very strong' (MMI = VII) to 'severe' shaking (MMI = VIII) in the LA basin (Fig. 11).

5.5 When do we need to issue an alert?

Current algorithms for EEW, such as 'τ_c–P_d Onsite' (Kanamori 2005), 'Virtual Seismologist' (Cua *et al.* 2009), 'ElarmS' (Allen *et al.* 2009b), 'PRESTo' (Zollo *et al.* 2009) or 'PreSEIS/PreSEIS Onsite' (Böse *et al.* 2008; Böse *et al.* 2012b), aim to provide estimates of the earthquake hypocentre and magnitude as quickly and as accurately as possible. However, it has not been established yet how quickly these algorithms can determine, for instance, whether an earthquake is *M*7 or *M*7.5. Several research papers address the achievable accuracy in the predicted magnitudes, in particular in EEW algorithms that only use information from the first few seconds of the seismic *P* wave (e.g. Kanamori 2005). There is an ongoing debate on whether the predicted magnitudes will saturate for large magnitude earthquakes (Rydelek & Horiuchi 2006; Rydelek *et al.* 2007).

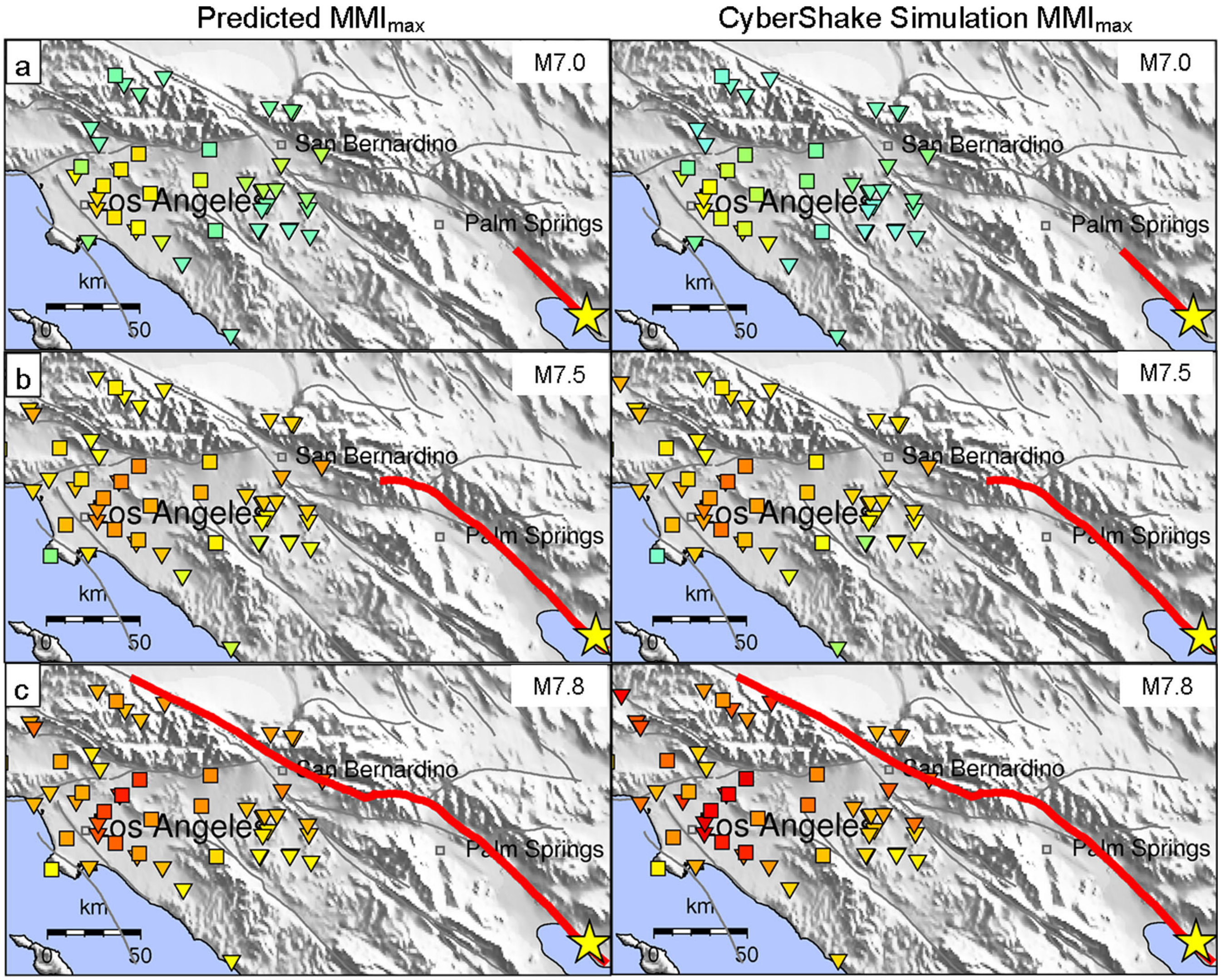


PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	none	none	none	Very light	Light	Moderate	Mod./Heavy	Heavy	Very Heavy
PEAK ACC.(%g)	<0.05	0.3	2.8	6.2	12	22	40	75	>139
PEAK VEL.(cm/s)	*<0.02*	*0.1*	*1.4*	*4.7*	*9.6*	*20*	*41*	*86*	*>178*
INSTRUMENTAL INTENSITY	I	II–III	IV	V	VI	VII	VIII	IX	X+

Figure 11. Predicted maximum intensity $\mathrm{MMI}_{\max}$ for a user located in the deep LA basin for a magnitude *M*6.5 (left-hand panel) and *M*7.0 (right-hand panel) earthquake at various locations with a rupture (a) propagating away ($\delta = 0$) or (b) towards the user ($\delta = 1$). Each coloured pixel represents the nucleation point (epicentre) of an individual earthquake rupture; the colour quantifies the shaking intensity at basin site DLA (white square). Even a relatively distant moderate *M*6.5–7 earthquake along the Palos Verdes, Newport-Inglewood/Rose Canyon, Elsinore or San Jacinto faults with a rupture propagating towards LA has the potential to cause 'very strong' (MMI = VII) to 'severe' shaking (MMI = VIII) in the LA basin.

Our results suggest that high accuracy in the estimated magnitudes is not essential for many EEW applications. Fig. 12 shows the predicted and observed (CyberShake simulated) maximum intensities $\mathrm{MMI}_{\max}$ for the earlier described ShakeOut scenario earthquake along the southern San Andreas Fault (Jones *et al.* 2008), assuming that the magnitudes were estimated as *M*7, *M*7.5 and *M*7.8, respectively. In all three cases a warning should be issued immediately, because shaking in the LA basin is expected to be at least 'strong' [MMI> = VI; we assume that the general public will care about events causing at least 'moderate' shaking (MMI> = V) at their site]. More accurate magnitude estimation is not needed in this (and other) examples.

Another way to view this problem is as follows: the final goal of EEW is to predict if shaking at a given user site is expected to exceed a pre-defined maximum level which requires taking protective actions to reduce expected damage by the approaching seismic waves.



PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	none	none	none	Very light	Light	Moderate	Mod./Heavy	Heavy	Very Heavy
PEAK ACC.(%g)	<0.05	0.3	2.8	6.2	12	22	40	75	>139
PEAK VEL.(cm/s)	*<0.02*	*0.1*	*1.4*	*4.7*	*9.6*	*20*	*41*	*86*	*>178*
INSTRUMENTAL INTENSITY	I	II–III	IV	V	VI	VII	VIII	IX	X+

Figure 12. Predicted (left-hand panel) and simulated (right-hand panel) maximum intensities MMI$_{\mathrm{max}}$ in the larger LA area for the ShakeOut scenario earthquake along the San Andreas Fault with epicentre (yellow star) at Bombay Beach (Jones *et al.* 2008). Magnitudes are estimated as (a) *M*7, (b) *M*7.5 and (c) *M*7.8. Red lines show the surface-projected 2-D fault ruptures. Each square and each triangle represents the target location for an individual BB (squares) or LF (triangles) model; the colour quantifies MMI$_{\mathrm{max}}$ at this site. Ground shaking in the LA basin is 'strong' to 'violent' (MMI = VI–IX) in all three cases, that is a warning needs to be issued as soon as the estimated magnitude is $M \geq 7.0$. Accurate magnitude estimation is not necessary in this example.

That is, for a given user and earthquake location, we mainly seek to determine a critical magnitude M_{critical}; if the predicted magnitude is expected to exceed this threshold, $M_{\mathrm{pred}} \geq M_{\mathrm{critical}}$, a warning needs to be issued, otherwise not. Note that a user, in particular when operating automated control systems, may want to set multiple ground motion thresholds and thus multiple M_{critical} for different response actions.

Fig. 13 shows M_{critical} at basin station DLA for three warning levels: (a) MMI ≥ V ('moderate' shaking), (b) MMI ≥ VI ('strong' shaking) and (c) MMI ≥ VII ('very strong' shaking). Each coloured pixel represents the nucleation point of an individual earthquake rupture; the colour codes M_{critical}. We compare M_{critical} determined from a point-source (Fig. 13, left-hand panel) and a finite-fault model (middle panel), where MMI$_{\mathrm{max}}$ is estimated from empirical GMPEs (Cua 2005; Worden *et al.* 2012) using the hypocentral distance or closest rupture-to-site distance (as determined from FinDer; Böse *et al.* 2012a), respectively. On the right-hand panel we show the corresponding results from our finite-fault models developed in this study. We assume that the fault ruptures propagate towards the LA basin ($\delta = 1$), that is these are the worst case scenarios with maximum effects of rupture directivity coupled with basin response. M_{critical} is determined from solving the relations for the minimum required magnitude to cause at least the respective MMI level in the LA basin.

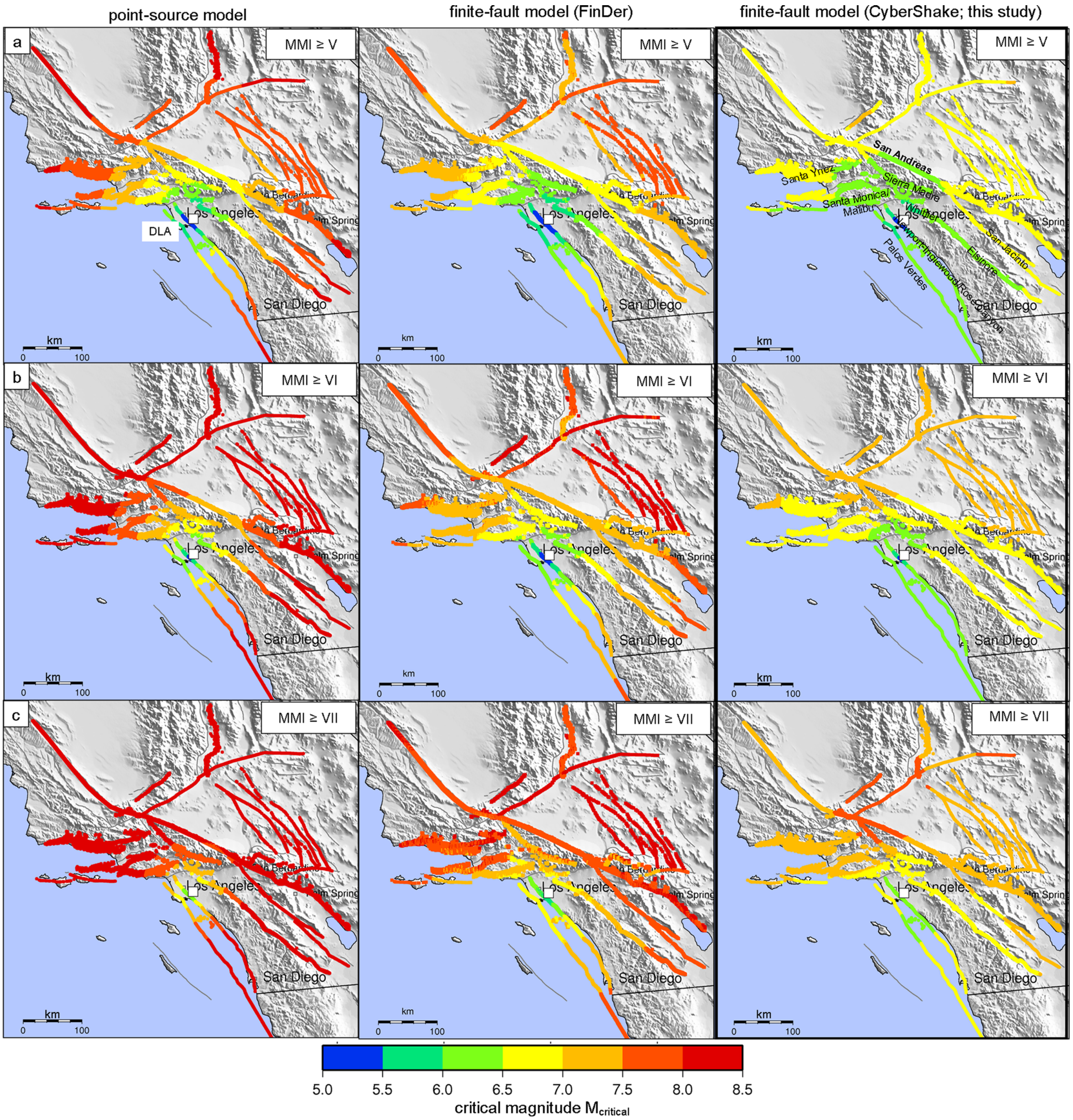


Figure 13. Threshold (or critical) magnitudes M_{critical} of earthquakes along various active faults in southern California that need to be exceeded to cause at least (a) MMI ≥ V ('moderate' shaking), (b) MMI ≥ VI ('strong' shaking) or (c) MMI ≥ VII ('very strong' shaking) in the LA basin. Each coloured pixel marks the nucleation point of an individual earthquake rupture, the colour codes the corresponding M_{critical}. Shown are the results from three models. Left-hand panel: results for a point-source model with usage of hypocentral distances and GMPEs by Cua (2005) as is currently implemented in CISN ShakeAlert (Böse *et al.* 2013). Middle panel: results for the FinDer finite-fault model (Böse *et al.* 2012a) with usage of rupture-to-site distances and GMPEs by Cua (2005). Right-hand panel: results for the finite-fault model developed in this study, which considers effects of rupture directivity and basin response. M_{critical} as determined from our finite-fault models (right-hand panel) is up to 2.5 magnitude units smaller than if determined from a point-source model (left-hand panel) and up to 1 magnitude unit smaller than if determined from alternative finite-fault models such as FinDer (middle panel). Note that all three models can predict M_{critical} at a higher magnitude resolution than displayed here. The corresponding warning times are shown in Fig. 14.

While all three models predict similar M_{critical} within ∼50 km from the user location (e.g. a close-by $M5$ to $M5.5$ has the potential to cause MMI ≥ V ('moderate' shaking) at DLA), there are strong differences at larger distances (Fig. 13). For instance, at ∼100 km distance the point-source model predicts $M_{\text{critical}} = 7$–7.5 for MMI ≥ V; that is only large earthquakes are considered as posing a threat. The finite-fault model derived from the CyberShake data set, however, predicts M_{critical} as being up to 2.5 magnitude units smaller.

M_{critical} also varies significantly depending on the fault where rupture is occurring. For instance, M_{critical} is up to 0.5 magnitude units smaller if the rupture propagates along a south-north trending fault pointing towards the LA basin, including for instance the Palos Verdes, Newport-Inglewood/Rose-Canyon, Elsinore or San Jacinto faults; a moderate-sized or larger earthquake ($M > 6.5$) along any of these faults poses a significant seismic threat to users located in the LA basin and requires a warning as soon as it has been detected.

The corresponding warning times for these scenarios are shown in Fig. 14. These are the maximum warning times an EEW system could provide, assuming the hypocentre and magnitude of the earthquake were immediately known. For the point-source model (Fig. 14, left-hand panel) we assume that strong shaking starts with the *S*-wave arrival (constant *S*-wave speed: $3.55\,\mathrm{km\ s^{-1}}$). For the finite-fault model using FinDer (Böse *et al.* 2012a) warning times are calculated from the time difference between the onset of shaking of a particular MMI level and the time required for the rupture to reach the rupture point that is closest to the user; we are assuming constant rupture velocities of $2.8\,\mathrm{km\ s^{-1}}$. In all three models, the maximum achievable warning time for close earthquakes ($\leq$50 km) is $\leq$10 s. However, for many other rupture scenarios, including those of more distant moderate-sized earthquakes along the Palos Verdes, Newport-Inglewood/Rose-Canyon, Elsinore or San Jacinto faults, for which strong shaking in the LA basin is expected (Fig. 13), warning times as determined from our regression models (right-hand panel) could exceed 30 s. Warning times are significantly shortened if alternative finite-fault models such as FinDer are used (Fig. 14, middle panel), because the fault rupture needs to have reached a critical length before strong shaking is expected.

6 DISCUSSION

In 2007 the development and implementation of an EEW demonstration system for California, named CISN ShakeAlert, was started (Böse *et al.* 2013). This hybrid system combines the outputs from three algorithms implemented in parallel, τ_c–P_d Onsite (Kanamori 2005; Wu *et al.* 2007; Böse *et al.* 2009), Virtual Seismologist (Cua *et al.* 2009) and ElarmS (Allen *et al.* 2009b), to calculate real-time the most probable earthquake magnitude and location. A UserDisplay receives these alert messages in real-time, calculates for a given user the expected local shaking intensity (MMI), and displays the information on a map (Böse *et al.* 2013).

The UserDisplay software estimates MMI from generic empirical GMPEs by Cua (2005) using the ShakeAlert magnitude and source-to-site distance. Recently, Böse *et al.* (2012a) developed the finite fault rupture detector algorithm FinDer for the real-time estimation of 2-D source dimensions and rupture-to-site distances while the fault rupture is still in progress. With this enhancement, ground-motion estimates in the ShakeAlert system are expected to become more accurate for moderate to large earthquakes ($M > 6.5$) once the algorithm is fully implemented and integrated in the system. However, so far directivity and basin response effects have been neglected, which can lead to an underestimation of shaking in large earthquakes.

In this study, we used the SCEC CyberShake data set with full 3-D wave-propagation simulations for >400 k rupture scenarios ($6.5 \leq M \leq 8.5$) in southern California to develop ε-SVR models for enhanced ground-motion predictions that consider effects of 3-D wave-propagation. Our models allow prediction of the temporal evolution of MMI at a given user-site within a fraction of a second using only information on the hypocentre, magnitude and rupture ratio (uni- or bilateral propagation). These parameters can be easily provided by the existing CISN ShakeAlert EEW demonstration system with finite fault extension (Böse *et al.* 2013). Even if the rupture ratio parameter was unknown, δ could be set to '1' to simulate the worst case scenario with the rupture propagating towards the user. In this paper we are not designing an entire EEW algorithm, but rather a shaking intensity estimator given input parameters from an assumed pre-existing EEW module, such as CISN ShakeAlert. We are not conducting a detailed analysis of false and missed alarms due to the lack of a clear definition of these terms that was applicable to users in general. Communicating uncertainties in the estimated parameters to end-users, however, will remain one of the major challenges in future EEW applications.

Certainly, the performance of our regression models depends on the quality and completeness of the training data set. The CyberShake waveforms were calibrated and validated with numerous seismic observations, including, for example those of the recent 2010 *M*7.2 El Mayor Cucapah earthquake in Baja California (Graves & Aagaard 2011), and the 2008 *M*5.4 Chino Hills earthquake in this study. Each rupture scenario in the set was simulated multiple times sampling a range of potential slip distributions and hypocentres.

However, physics-based simulation models, dynamic and kinematic, are in a state of ongoing modification, verification, testing and improvement. An important issue with kinematic models (hence CyberShake results) is the assumptions regarding the development and assignment of numerical values to the kinematic parameters, for example slip vector, rise time, rupture velocity (and their correlations) to which model results are sensitive. Such parameters are selected, in part by being guided by the more realistic, though more complicated and perhaps even less developed, dynamic models. Even though much progress has been made within the past decade, and particularly within the past few years, in developing objective verification methodologies for reviewing and comparing individual simulation models and validating them against observed data, a lot remains to be done in order to make them sufficiently reliable as practical tools, especially for EEW applications.

Even though CyberShake clearly is a high-quality data set that applies state-of-the-art scientific and computational knowledge and resources, it is affected by these limitations. Furthermore, it is likely that not all possible rupture scenarios were considered, particularly those involving the rupture along multiple faults. That is, our use of the CyberShake data set is not meant as a claim that these simulations are fully validated and appropriately sample the range of rupture uncertainty, but rather we are using these data to demonstrate their application to EEW as a 'proof of concept'. Our method is not limited to the CyberShake data set; once more accurate BB ground-motion simulations or new rupture scenarios (such as in UCERF 3.0, Field *et al.* 2013) become available, our models can be easily updated.

Another uncertainty in our current regression models comes from the parametrization of earthquake ruptures. To keep the approach as simple and robust as possible, we restricted our models to requiring only a few input parameters (magnitude, location, rupture ratio). Additional input parameters might help to make the ground-motion predictions more accurate, but the real-time application will likely get more challenging and possibly unrealistic. From the analyses in this paper we saw that the rupture ratio δ, which is a normalized parameter that characterizes if the rupture propagates mainly towards ($\delta = 1$) or away ($\delta = 0$) from the user site, can have important implications on the predicted level of shaking (Figs 10 and 11). Determining δ in real-time is thus critical for ground-motion predictions and EEW applications. While MMI could be over- or underpredicted

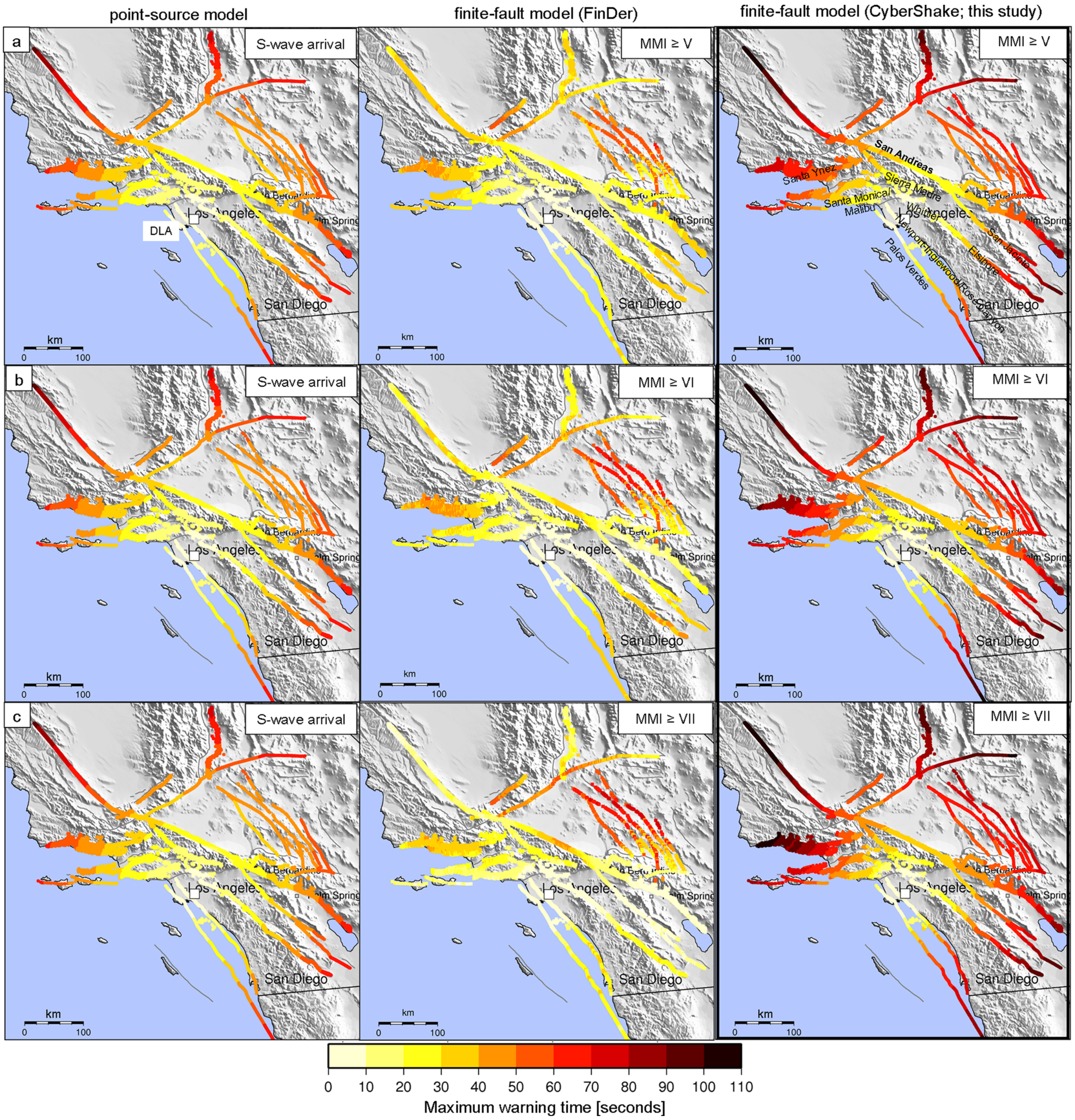


Figure 14. Maximum warning times (in seconds after rupture nucleation) until a user in the LA basin will experience (a) MMI ≥ V ('moderate' shaking), (b) MMI ≥ VI ('strong' shaking) or (c) MMI ≥ VII ('very strong' shaking) for the scenarios shown in Fig. 13. Left-hand panel: results for a point-source model assuming that peak shaking starts with the *S*-wave arrival (*S*-wave speed: 3.55 km s^{-1}); a distinction of the three intensity levels cannot be made, that is all plots from top to bottom are the same. Middle panel: results for the FinDer finite-fault model (Böse *et al.* 2012a) as calculated from the time difference between the rupture reaching the point that is closest to the user at DLA (rupture speed: 2.8 km s^{-1}) and the onset shaking corresponding to the three MMI levels. Right-hand panel: results for the finite-fault models developed in this study. Our finite-fault models predict that for many dangerous rupture scenarios, including moderate- to large-sized earthquakes along the Palos Verdes, Newport-Inglewood/Rose-Canyon, Elsinore or San Jacinto faults for which shaking in the LA basin can be strong (Fig. 13), warning times of >30 s are a realistic expectation (right-hand panel).Warning times are significantly shortened if alternative finite-fault models are used (middle panel), because the ruptures need to have reached a critical length before strong shaking is expected.

by two or more intensity units (depending on the location of the earthquake and its magnitude, Fig. 10) if the rupture propagates in the opposite direction from what was assumed or estimated, tests of the FinDer algorithm have shown that these extreme cases are not likely for large earthquakes ($M > 7$) if applied in dense seismic networks ($\Delta \approx 30$ km). A limitation of FinDer, however, arises from the simplified 2-D representation of ruptures which appears to be suited to describe long and narrow ruptures (low W/L aspect ratios), but less suited to describe relatively short and wide ruptures ($W/L \sim 1$, e.g. reverse faults) such as the Northridge earthquake.

We selected the instrumental intensity (MMI) at the user site as model output, because this parameter is well understood by the general public; MMI uses a descriptive scale (see Table 1) making the predicted intensities easily comprehendible by untrained individuals. The conversion of the simulated ground motions to MMI or any other ground-motion parameter using statistically based models (e.g. Worden *et al.* 2012), as done in this paper, however, is an additional source of uncertainty in our regression models.

In this study we developed ground-motion prediction models for ∼70 selected test sites in and around the LA basin. However, the real-time application of our approach within CISN ShakeAlert (Böse *et al.* 2013) will require ground-motion estimates at a much denser grid of possible user locations. This would possibly allow for the rapid prediction of ShakeMap-like ground-motion maps (Wald *et al.* 1999). Aside from the site-specific ground-motion models as presented in this study, we should develop generic models using various site characteristics, such as *Vs*30 or thickness of sediments in the basin, as additional input parameters. Since these generic models will have to average over multiple site-rupture geometries, we expect the prediction results to be less accurate compared to the site-specific prediction models developed in this study. Nevertheless generic models might be useful for spatial interpolation at sites, where no ground-motion simulations and, consequently, no regression models are available.

An earthquake similar to the 2008 ShakeOut scenario (Jones *et al.* 2008) on the southern portion of the San Andreas Fault starting at Bombay Beach could offer ∼60 s of warning to recipients in the LA basin. The most recent large earthquake on this section of the fault occurred over 300 yr ago, although the average recurrence rate of large earthquakes on the southern San Andreas is only about 150 yr (e.g. Jones *et al.* 2008). Moreover, rupture simulations by Böse & Heaton (2010) suggest that the probability for a rupture along a smooth fault such as the San Andreas Fault to evolve into a major event is high. Still we should be careful not to neglect the threat from other faults (Palos Verdes, Newport-Inglewood/Rose-Canyon, Elsinore or San Jacinto) where moderate-sized magnitude earthquakes could cause significant shaking and damage in the LA basin, but that are still far enough away to provide reasonable long warning times (>30 s) for early warning. Our finite-fault models could be useful for designing or optimizing seismic and geodetic networks for EEW (e.g. Oth *et al.* 2010) by predicting shaking levels and available warning times for various rupture scenarios and station distributions.

7 CONCLUSIONS

EEW systems need a reliable and accurate method to quickly determine expected shaking and arrival time of this shaking at a user site once an earthquake has been detected. In this paper we developed and demonstrated a method to achieve this goal by quantifying the relationship between earthquake parameters (hypocentre, magnitude, rupture ratio) and shaking for specific sites in and around LA. Regression models of the earthquake parameters and shaking intensities were derived from the SCEC CyberShake database with around 415 000 finite-fault rupture scenarios ($6.5 \leq M \leq 8.5$) along active faults in southern California. The results were shown to be a substantial improvement over existing GMPEs, because basin response and directivity effects are directly considered. Our models predict that a relatively distant moderate *M*6.5–7 earthquake along the Palos Verdes, Newport-Inglewood/Rose Canyon, Elsinore or San Jacinto faults with a rupture propagating towards LA could cause 'very strong' to 'severe' shaking in the LA basin. However, warning times for these events could be >30 s. Finally, we want to point out that our use of the CyberShake data set is not meant as a validation of these simulations, but rather we are using these data to demonstrate their application to EEW as a 'proof of concept'.

ACKNOWLEDGEMENTS

To train and test SVR models in this study, we used LIBSVM (Chang & Lin 2011), a free library for Support Vector Machines and Regression. Maps in this paper were made with Generic Mapping Tools version 4.2.1 (www.soest.hawaii.edu/gmt; Wessel & Smith 1998). We would like to thank J. Hardebeck, E. Hauksson, T. Heaton, two anonymous reviewers and editor E. Fukuyama for constructive comments that helped to improve this paper. This work is funded through contract G09AC00258 from USGS/ANSS to the California Institute of Technology (Caltech). Funding was also provided by the Gordon and Betty Moore Foundation through Grant GBMF #3023 to Caltech. This is contribution #10094 of the Seismological Laboratory, Geological and Planetary Sciences at Caltech. This research was also supported by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0529922 and USGS Cooperative Agreement 07HQAG0008. The SCEC contribution number for this paper is 1799.

REFERENCES

Allen, R.M., Gasparini, P., Kamigaichi, O. & Böse, M., 2009a. The status of earthquake early warning around the world: an introductory overview, *Seismol. Res. Lett.,* **80**(5), 682–693.

Allen, R.M., Brown, H., Hellweg, M., Khainovski, O., Lombard, P. & Neuhauser, D., 2009b. Real-time earthquake detection and hazard assessment by ElarmS across California, *Geophys. Res. Lett.,* **36,** L00B08, doi:10.1029/2008GL036766.

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, *Geophys. J. Int.,* **183**(2), 1014–1030.

Böse, M., Wenzel, F. & Erdik, M., 2008. PreSEIS: a neural network based approach to earthquake early warning for finite faults, *Bull. seism. Soc. Am.,* **98**(1), 366–382.

Böse, M., Hauksson, E., Solanki, K., Kanamori, H. & Heaton, T.H., 2009. Real-time testing of the on-site warning algorithm in southern California and its performance during the July 29 2008 Mw5.4 Chino Hills earthquake, *Geophys. Res. Lett.,* **36,** L00B03, doi:10.1029/2008GL036366.

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

Böse, M., Heaton, T. & Hauksson, E., 2012b. Rapid estimation of earthquake source and ground-motion parameters for earthquake early warning using data from single three-component broadband or strong-motion sensor, *Bull. seism. Soc. Am.,* **102**(2), 738–750.

Böse, M. *et al.*, 2013. CISN ShakeAlert—an earthquake early warning demonstration system for California, in *Early Warning for Geological Disasters—Scientific Methods and Current Practice,* pp. 49–69, eds Wenzel, F. & Zschau, J., Springer Berlin.

Boore, D.M. & Atkinson, G.M., 2008. Ground motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 and 10.0 s, *Earthq Spectra,* **24**(S1), 99–138.

Callaghan, S., Maechling, P.J., Small, P., Milner, K., Graves, R.W. & Jordan, T.H., 2011. Broadband CyberShake platform: seismogram synthesis for broadband physics-based probabilistic seismic hazard analysis

CyberShake collaboration, in *Proceedings of the American Geophysical Union, Fall Meeting 2011*, San Francisco, abstract #NH51B-1696.

Campbell, K.W. & Bozorgnia, Y., 2008. NGA ground motion model for the geometric mean horizontal component of PGA, PGV, PGD, and 5%-damped linear elastic response spectra for periods ranging from 0.01 to 10 s, *Earthq. Spectra,* **24**(S1), 139–172.

Chang, C.C. & Lin, C.-J., 2011. LIBSVM: a library for support vector machines, in *ACM Transactions on Intelligent Systems and Technology,* 2:27:1–27:27. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed January 2013).

Cua, G., 2005. Creating the Virtual seismologist: developments in earthquake early warning and ground motion characterization, *PhD thesis,* Department of Civil Engineering, California Institute of Technology.

Cua, G., Fischer, M., Heaton, T., Wiemer, S. & Giardini, D., 2009. Real-time performance of the Virtual Seismologist method in southern California, *Seismol. Res. Lett.,* **80,** 740–748.

Denolle, M.A., Dunham, E.M., Prieto, G.A. & Beroza, G.C., 2014. Strong ground motion prediction using virtual earthquakes, *Science,* **343**(6169), 399–403.

Donovan, J., Jordan, T.H. & Brune, J.N., 2012. Testing CyberShake using precariously balanced rocks, in *Proceedings of the SCEC Annual Meeting,* Palm Springs, CA, September, 2012.

Field, E.H. *et al.*, 2013. Uniform California earthquake rupture forecast, version 3 (UCERF3): the time-independent model, USGS Open-File Report: 2013–1165.

Graves, R. & Pitarka, A., 2005. Kinematic rupture model generator, in *Proceedings of the SCEC Annual Meeting,* Palm Springs, CA, September, 2005.

Graves, R. & Pitarka, A., 2010. Broadband ground motion simulation using hybrid approach, *Bull. seism. Soc. Am.,* **100**(5A), 2095–2123.

Graves, R. *et al.*, 2010. CyberShake: a physics-based seismic hazard model for Southern California, *Pure appl. Geophys.,* **168,** 367–381.

Graves, R.W. & Aagaard, B.T., 2011. Testing long-period ground-motion simulations of scenario earthquakes using the Mw 7.2 El Mayor-Cucapah mainshock: evaluation of finite-fault rupture characterization and 3D seismic velocity models, *Bull. seism. Soc. Am.,* **101**(2), 895–907.

Graves, R.W., Aagaard, B.T., Hudnut, K.W., Star, L.M., Stewart, J.P. & Jordan, T.H., 2008. Broadband simulations for Mw 7.8 southern San Andreas earthquakes: ground motion sensitivity to rupture speed, *Geophys. Res. Lett.,* **35,** L22302, doi:10.1029/2008GL035750.

Hauksson, E. *et al.*, 2008. Preliminary Report on the 29 July 2008 Mw5.4 Chino Hills, Eastern Los Angeles Basin, California, Earthquake Sequence, *Seismol. Res. Lett.,* **79,** 855–868.

Hoshiba, M., Iwakiri, K., Hayashimoto, N., Shimoyama, T., Hirano, K., Yamada, Y., Ishigaki, Y. & Kikuta, H., 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 Space,* **63**(7), 547–551.

Jones, L.M. *et al.*, 2008. The ShakeOut Scenario: U.S. Geological Survey Open File Report 2008-1150, and California Geological Survey Preliminary Report 25. Available at: http://pubs.usgs.gov/of/2008/1150/ (last accessed June 2014).

Kanamori, H., 2005. Real-time seismology and earthquake damage mitigation, *Annu. Rev. Earth planet. Sci.,* **33,** 195–214.

Kohler, M., Magistrale, H. & Clayton, R., 2003. Mantle heterogeneities and the SCEC three-dimensional seismic velocity model version 3, *Bull. seism. Soc. Am.,* **93,** 757–774.

Magistrale, H., Day, S., Clayton, R.W. & Graves, R., 2000. The SCEC Southern California reference three-dimensional velocity model version 2, *Bull. seism. Soc. Am.,* **90**(6B), S65–S76.

Oth, A., Böse, M., Wenzel, F., Köhler, N. & Erdik, M., 2010. Evaluation and Optimization of Seismic Networks and Algorithms for Earthquake Early Warning: the case of Istanbul (Turkey), *J. geophys. Res.,* **115,** B10311, doi:10.1029/2010JB007447.

Rydelek, P. & Horiuchi, S., 2006. Is the earthquake rupture deterministic? *Nature,* **442,** doi:10.1038/nature04963.

Rydelek, P., Wu, C. & Horiuchi, S., 2007. Comment: peak ground displacement and earthquake magnitude, *Geophys. Res. Lett.,* **34,** L20302, doi:10.1029/2007GL029387.

Smola, A.J. & Schölkopf, B., 2004. A tutorial on support vector regression, in *Statistics and Computing,* Vol. 14, pp. 199–222, Kluwer Academic Publishers.

Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory,* Springer.

Wald, D., Quitoriano, V., Heaton, T., Kanamori, H., Scrivner, C. & Worden, C., 1999. TriNet ShakeMaps: rapid generation of instrumental ground motion and intensity maps for earthquakes in southern California, *Earthq. Spectra,* **15,** 537–556.

Wang, F. & Jordan, T.H., 2012. Using averaging-based factorization to compare seismic hazard models derived from 3D earthquake simulations with NGA ground motion prediction equations, in *Abstract S51A-2386 presented at 2012 Fall Meeting, AGU*, San Francisco, CA, 3–7 December.

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

Wessel, P. & Smith, W.H.F., 1998. New, improved version of the Generic Mapping Tools Released, *EOS, Trans. Am. geophys. Un.,* **79,** 579.

Worden, C.B., Gerstenberger, M.C., Rhoades, D.A. & Wald, D.J., 2012. Probabilistic relationships between peak ground motion and Modified Mercalli intensity, *Bull. seism. Soc. Am.,* **102**(1), 204–221.

Working Group on California Earthquake Probabilities, 2007. The Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2), USGS Open File Report 2007–1437.

Wu, Y.-M., Kanamori, H., Allen, R.M. & Hauksson, E., 2007. Determination of earthquake early warning parameters, τ_c and P_d, for southern California, *Geophys. J. Int.,* **170,** 711–717.

Zollo, A. *et al.*, 2009. The Earthquake early warning system in Southern Italy: methodologies and performance evaluation, *Geophys. Res. Lett.,* **36,** L00B07, doi:10.1029/2008GL036689.

APPENDIX: ε-SUPPORT VECTOR REGRESSION (ε-SVR)

For a given set of training data points with input feature vector x and target output z_i, $\{(x_1; z_1), \ldots, (x_n; z_n)\} \subset \chi \times \mathbb{R}$, where χ denotes the space of input patterns (e.g. $\chi = \mathbb{R}^d$) and n is the number of example patterns, the goal of ε-SVR is to find a function f, that has at most ε deviation from the actually obtained targets (ε-insensitive loss function, Fig. A1b), and that at the same time is a flat as possible (Vapnik 1995; Smola & Schölkopf 2004). In the simplest case, we seek a linear function

$$f(x) = \langle w, x \rangle + b, \quad \text{with } w \in \chi, b \in \mathbb{R}, \tag{A1}$$

where $\langle \cdot,\cdot \rangle$ denotes the dot product in χ. Flatness of the regression function in (A1) means to find a small w, which can be achieved from minimization of the norm $\|w\|^2 = \langle w, w \rangle$. To cope with otherwise infeasible constraints in the optimization problem, slack variables ξ_i and ξ_i^* are introduced (Fig. A1). The optimization problem in ε-SVR can thus be written as

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{A2.1}$$

subject to

$$z_i - \langle w, \phi(x_i) \rangle - b \le \varepsilon + \xi_i \tag{A2.2}$$

$$\langle w, \phi(x_i) \rangle + b - z_i \le \varepsilon + \xi_i^* \tag{A2.3}$$

$$\xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots n, \tag{A2.4}$$

where $\phi(x_i)$ is a non-linear mapping function to transform the input vector into a higher (infinite) feature space where linear regression can be performed; b is the bias term. The regularization parameter $C > 0$ in eq. (A2.1) controls the trade-off between the flatness of

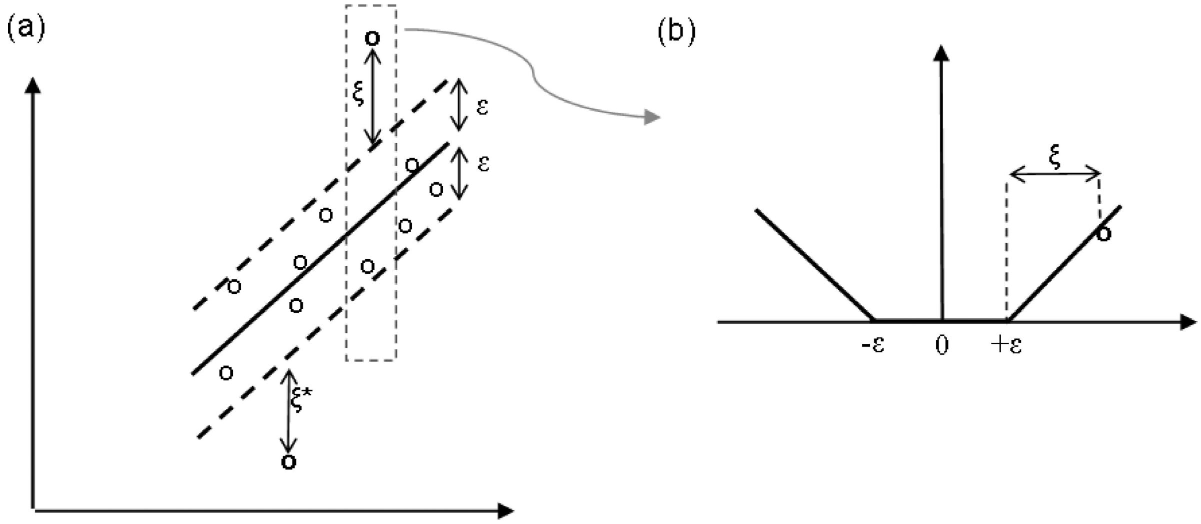


Figure A1. Principle of ε-SVR. (a) The input data is transformed to a higher (infinite) dimensional feature space using kernel functions (here: radial basis function, RBF) where linear regression can be performed. (b) The error function in ε-SVR is ε-insensitive, that is errors $<|\varepsilon|$ are neglected. Goal of ε-SVR is to find a function *f*, that has at most ε deviation from the actually obtained targets, and that at the same time is as flat as possible.

the regression function and the amount to which errors $>|\varepsilon|$ are tolerated (Smola & Schölkopf 2004).

A solution to the optimization problem in eqs (A2.1)–(A2.4) is found by solving the corresponding dual problem (e.g. Smola & Schölkopf 2004) using quadratic programming (e.g. Chang & Lin 2011). Feature vectors x_i, for which the corresponding Lagrangian multipliers α_i and α_i^* are positive, are the so-called 'support vectors' of the regression function (Fig. A1)

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad \text{(A3)}$$

with kernel $k(x, x') := \langle \phi(x), \phi(x') \rangle$. Note that *f* only depends on dot products between the input data, which is computationally effective. Here we select the radial basis function (RBF)

$$k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \quad \gamma > 0 \qquad \text{(A4)}$$

which is a commonly used kernel in SVR. The parameter γ in eq. (A4) defines the width of the (Gaussian) kernels. This leaves us with two critical model parameters, *C* and γ, (eqs A2.1 and A4), which are generally not straightforward to select. We determine the optimum values in this study from the application of the Direct Search algorithm. The advantage of SVR compared to other regression techniques is that the obtained models are very flat which is prerequisite for a high generalization capability towards unseen data.