AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER
10.1002/2016GL069887

Key Points:
- Earthquake early warning can be tackled probabilistically using "prior sampling"
- Full numerical wave propagation and crustal heterogeneity can be accommodated
- A quasi-real-time analysis of the 2008 M_w 5.4 Chino Hills event demonstrates viability of the approach

Supporting Information:
- Supporting Information S1
- Figure S1
- Figure S2
- Figure S3
- Figure S4
- Figure S5
- Figure S6
- Figure S7
- Figure S8
- Figure S9
- Figure S10

Correspondence to:
A. P. Valentine,
a.p.valentine@uu.nl

Citation:
Käufl, P., A. P. Valentine, and J. Trampert (2016), Probabilistic point source inversion of strong-motion data in 3-D media using pattern recognition: A case study for the 2008 M_w 5.4 Chino Hills earthquake, *Geophys. Res. Lett.*, *43*, doi:10.1002/2016GL069887.

Received 3 JUN 2016
Accepted 28 JUL 2016
Accepted article online 2 AUG 2016

©2016. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Probabilistic point source inversion of strong-motion data in 3-D media using pattern recognition: A case study for the 2008 M_w 5.4 Chino Hills earthquake

Paul Käufl[1], Andrew P. Valentine[1], and Jeannot Trampert[1]

[1]Department of Earth Sciences, Utrecht University, Utrecht, Netherlands

Abstract Despite the ever increasing availability of computational power, real-time source inversions based on physical modeling of wave propagation in realistic media remain challenging. We investigate how a nonlinear Bayesian approach based on pattern recognition and synthetic 3-D Green's functions can be used to rapidly invert strong-motion data for point source parameters by means of a case study for a fault system in the Los Angeles Basin. The probabilistic inverse mapping is represented in compact form by a neural network which yields probability distributions over source parameters. It can therefore be evaluated rapidly and with very moderate CPU and memory requirements. We present a simulated real-time inversion of data for the 2008 M_w 5.4 Chino Hills event. Initial estimates of epicentral location and magnitude are available ~14 s after origin time. The estimate can be refined as more data arrive: by ~40 s, fault strike and source depth can also be determined with relatively high certainty.

1. Introduction

The characterization of past and ongoing earthquakes by means of physical models of the rupture process is a fundamental task of seismology. A comprehensive understanding of any particular earthquake requires careful, detailed analysis and may entail months of research effort. Nevertheless, any information that can be obtained within the seconds and minutes immediately following rupture brings considerable value: it may enable actions to be taken that reduce the human and economic costs of the earthquake. This aspect of seismology is often referred to as "earthquake early warning" (EEW).

By its very nature, EEW operates in a regime where observed data may be limited in both quantity and quality, and results may therefore include significant ambiguities and uncertainties. Thus, in an EEW context it is especially desirable to work within a "probabilistic" framework, allowing results to encapsulate the full range of earthquake source models compatible with available observations. In practice, this typically entails some form of sampling-based approach: many candidate earthquake mechanisms must be tested against the data, mapping out the space of allowable models. To test each model, synthetic (predicted) seismograms must be computed; for most accurate results, it is desirable to make use of state-of-the-art numerical wavefield simulations that account for the complex effects of crustal heterogeneity and local topography. However, these are extraordinarily computationally expensive—each simulation may require tens or hundreds of CPU hours. Standard techniques for probabilistic inversion (e.g., Markov chain Monte Carlo algorithms) require the "real" observations to have been made before any simulations can be run and are therefore totally unsuited to EEW applications.

In *Käufl et al.* [2016], we set out an alternative framework for solving probabilistic inverse problems, which we call "prior sampling." In this approach, samples are not directed toward explaining any one data set, and all expensive computation can be performed in advance of data collection. A learning algorithm is used to encapsulate information about the relationship between observables and underlying models, permitting probabilistic inversion results within milliseconds once signals are available. With the present paper, we seek to demonstrate that this approach is viable—and powerful—for EEW. We use synthetic seismograms generated with the spectral element method [*Komatitsch et al.*, 2002] in conjunction with a corresponding high-quality regional crustal model [*Tape et al.*, 2009] and implement a system that enables real-time monitoring of a fault system (the Whittier and Chino faults) in Southern California. This allows recovery of a probabilistic point source moment tensor representation for any ongoing event, as we demonstrate using

strong-motion data recorded during a M_w 5.4 event occurring in 2008. We believe that this is the first demonstration of a probabilistic EEW system that utilizes physically complete waveform modeling including 3-D crustal structure.

2. Methodology

We represent earthquakes as deviatoric moment tensor point sources with triangular source time function, where the source geometry and radiation pattern is parameterized as suggested by *Tape and Tape* [2012]. A point source is thus described by the nine parameters nondouble-couple component γ, strike κ, rake σ, cosine of dip h, moment magnitude M_w, the centroid location (lat, lon, depth) and the source half-duration τ, which we combine into the source parameter vector **m**. The source parameters are also summarized in Table S1, available as supporting information to this article. Of course, this representation does not capture the full complexity of the rupture, and in the near-field (at epicentral distances of around one wavelength or less) waveforms derived from point sources may lack features observed in real seismograms. Nevertheless, a point source representation can provide a useful initial characterization of any seismic event.

In order to solve the probabilistic inverse problem, we closely follow the approach described in *Käufl et al.* [2014, 2015, 2016]. This makes use of an ensemble of neural networks to represent the mapping from observable data vectors to marginal posterior probability density functions (pdfs) over earthquake source parameters, which are parameterized by mixtures of Gaussians and are capable of accurately representing multimodal and non-Gaussian behavior [*Bishop*, 1995]. The individual neural networks are trained using a database of prior samples containing examples of synthetic data vectors and corresponding earthquake source parameter vectors. Here we generate neural network training sets efficiently by making use of the fact that the *l*th component of the synthetic seismogram at the *k*th receiver for a point source at location *j* is linearly related to the six independent components of the moment tensor **M** according to

$$u_{jkl}(\boldsymbol{m}, t) = \sum_{i=1}^{6} M_i G_{ijkl} * f(\tau, t), \tag{1}$$

where the temporal evolution of the source is described by the triangular source time function $f(\tau, t)$ with half-duration τ and $*$ denoting the convolution operator. We can therefore adopt a two-stage sampling procedure to obtain the necessary suite of training examples. This is described in detail in section S1 of the supporting information, but in overview: first, we generate a set of source locations, randomly distributed throughout the area to be monitored (and potentially taking into account prior knowledge about seismicity, fault locations, etc.). For each of these locations, and for all potential receiver locations, we compute the six unit Green's functions G_{ijkl}. Then, in a second stage of sampling, we generate a large suite of seismograms by repeatedly selecting one of these locations at random, generating a random moment tensor (potentially taking into account the local fault orientations) and source half duration, and applying equation (1).

This procedure allows us to build a large collection of source vectors and corresponding 3-D wavefield predictions, while keeping computational costs within manageable levels. We add noise to each waveform, to simulate observational and theoretical uncertainties—potentially, this can also include effects such as stations going "off-line." We extract short time windows from the synthetic seismograms, designed to mimic data collection following "triggering" of a detector in an EEW setting. We use waveforms from every station in the network: the fact that the *P* wave may not reach all stations within the time interval can help constrain event locations. The resulting time series are then used to construct a set of neural networks which take in seismic observations and output pdfs on event location, moment tensor components, and source half duration. A comprehensive discussion of this prior sampling, neural network methodology, its advantages and limitations can be found in *Käufl et al.* [2016].

3. Demonstration Using Strong-Motion Data in Southern California

The general framework described above can be adapted to a wide range of EEW tasks. In order to provide a specific example, we consider a test scenario in the Los Angeles Basin and construct a system focused on monitoring the Whittier-Chino fault system. We use this to first perform synthetic resolution tests and then present observed data for the 2008 M_w 5.4 Chino Hills earthquake to the trained networks in order to obtain posterior pdfs for each source parameter. The 3-D Green's functions database underlying the training set is obtained by simulating wave propagation using SPECFEM3D [*Tromp et al.*, 2008] in the tomographic Southern California

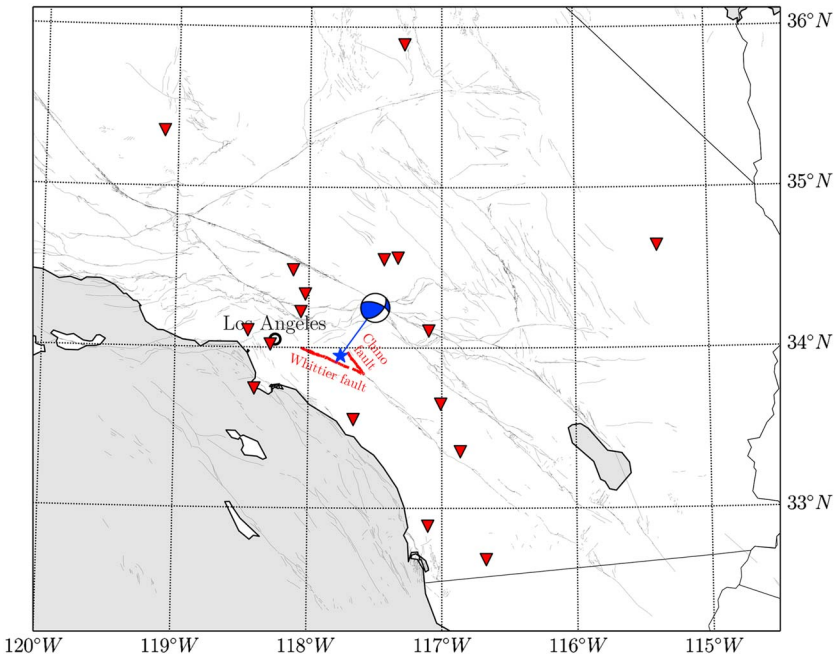

Figure 1. Surface traces of known faults in the study region (taken from http://earthquakes.usgs.gov/hazards/qfaults/). The locations of 17 strong-motion instruments of the SCSN and ANZA Regional Network are shown as red triangles, the epicentral location of the 2008 M_w 5.4 Chino Hills event, located between the Whittier and Chino fault sections, is shown as a blue star, where the beach ball corresponds to a point source moment tensor solution obtained by *Hauksson et al.* [2008].

model CVMH 11.9 [*Tape et al.*, 2009; C. Tape et al., personal communication, 2014]. See also *Taborda and Bielak* [2014] for a thorough comparison of models in the area. Figure 1 gives an overview over the model domain and shows surface traces of known faults in the region. A topographic map of the same area is provided in the supporting information (Figure S2) and illustrates the importance of 3-D waveform modeling in this region. Although this demonstration focuses solely on the Whittier-Chino fault system, the approach could readily be extended to encompass a wider source region covering multiple fault systems, given sufficient computing resources. A regional study on such a scale has been demonstrated using 1-D synthetic waveforms in *Käufl et al.* [2014] and *Käufl et al.* [2015].

We distribute a set of 150 source locations uniformly in a rectangular volume enclosing the Whittier and Chino fault sections (see Figure S3). In depth, point sources are distributed from the surface of the Earth down to 20 km. We calculate unit Green's functions of length 200s for a set of ~1300 existing receivers in Southern California and in addition for 600 virtual receivers located on a regular grid (see Figure S4). This database of waveforms is available for researchers to download: see the supporting information for a detailed description.

Source parameters are drawn from uniform distributions covering all possible orientations and double couple as well as nondouble-couple sources (see Table S1). Scalar magnitudes are distributed according to a log-uniform distribution corresponding to the range $5.0 \leq M_w \leq 8.0$ and half durations are correlated to scalar magnitudes by means of the relation

$$\tau = a \cdot 10^{-8} M_0^{1/3} \tag{2}$$

[*Ekström et al.*, 2012], where the proportionality factor a is drawn from a normal distribution with mean 1.05 and standard deviation 0.1 in order to account for deviations from this empirical relation. While from a technical perspective it would be straightforward to include geological prior information such as the fault geometry, we do not correlate source orientations to the known orientations of the Whittier and Chino fault sections, since here we are interested in exploring to what extent we can resolve the source orientation based on the data alone. Moreover, realistic uncertainty estimates on fault orientations taken from the literature are hard to obtain.

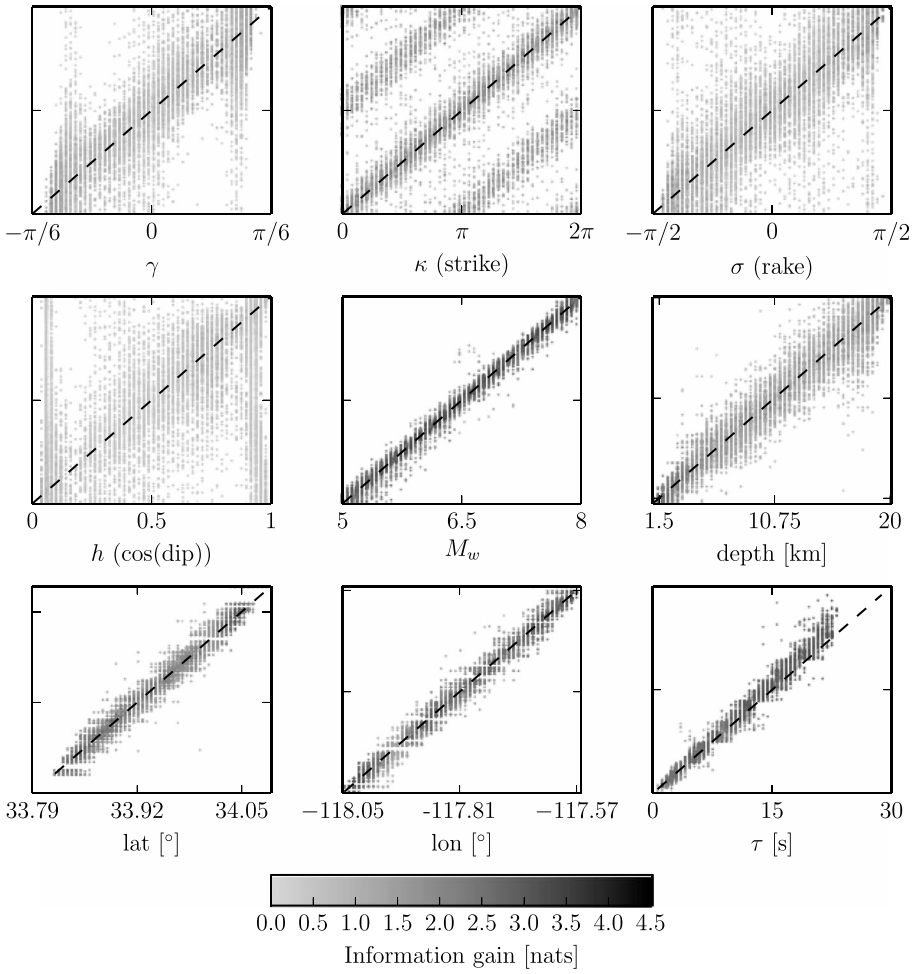

Figure 2. Prediction performance of network ensembles trained on the nine source parameters using 30 s of data. Each dot corresponds to a noisy synthetic test set example, the position of the posterior mode is plotted on the horizontal, and the target value on the vertical axis. The gray scale indicates the relative information gain with respect to the prior distribution.

3.1. Data, Observational Noise, and Frequency Content

We use observed and synthetic 3-C accelerograms recorded at 17 receiver locations of the Southern California Seismic Network (SCSN) and ANZA Regional Network, located within ~30 km and ~200 km distance from the fault sections. Based on epicentral distance and data availability for a 2008 M_w 5.4 event (star in Figure 1), we use 17 receivers in this study. We apply a band-pass filter to observed and synthetic waveforms between 6.25 and 100s. On one hand, this is intended to limit high-frequency contributions that might be contaminated by reverberations originating from the excitement of sedimentary basins, which may not be well explained by the tomographic model. On the other hand, the computational cost associated with neural network training scales with the number of input dimensions (in the worst case quadratically, see *Käufl et al.* [2015] and the supporting information) and the low-pass filter serves as a way to reduce the input dimensionality. All waveforms are thus downsampled to a 3 s sampling period.

We obtain a simplified observational noise model by analyzing 1800 s of data recorded at the 17 receivers during a period of no significant seismicity, which are preprocessed in the same way as data and synthetics. We assume that observational noise follows a Gaussian distribution and is temporally and spatially uncorrelated and obtain an estimated noise standard deviation of 10^{-5} m/s^2. Clearly, this is a simplifying assumption, and a more realistic noise model could potentially have an influence on the results. Therefore, in particular, if the methodology is to be used in the context of a system operating in an unsupervised real-time mode, a more thorough analysis might be required, taking into account known statistical properties of the ambient noise field in a particular area or local site conditions that may lead to a station-dependent noise estimate.

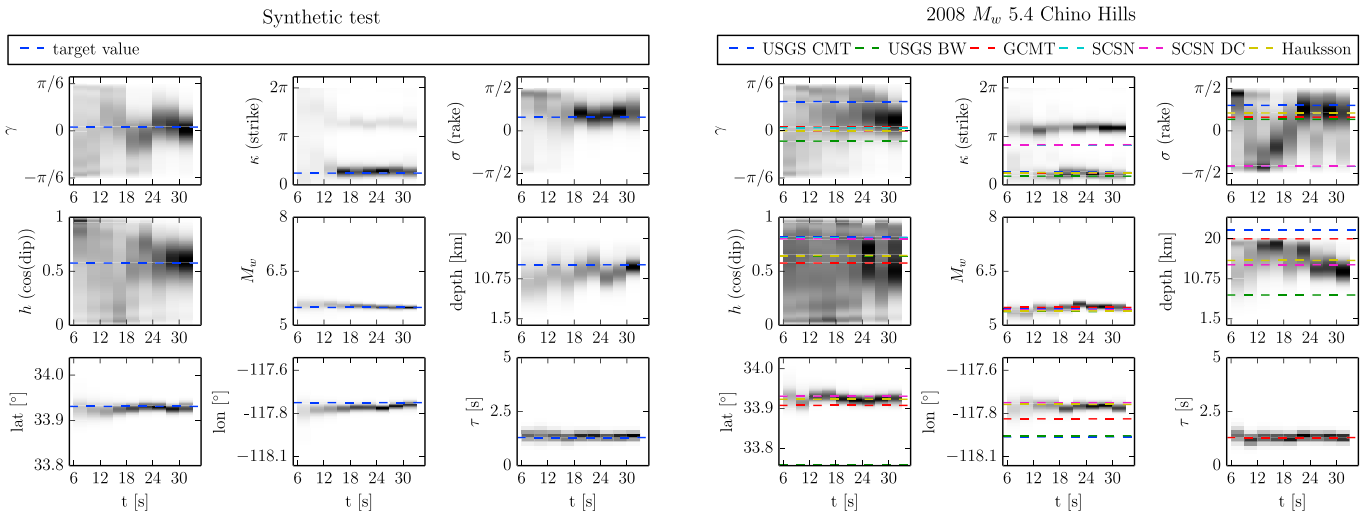

Figure 3. Marginal posterior distributions for the nine source parameters obtained by presenting 6 to 30 s of noisy (left) synthetic and (right) observed strong-motion data to a set of trained network ensembles. Darker colors correspond to higher-probability density. The horizontal lines on the left panels correspond to the known target value, the dashed, colored horizontal lines on the right panels to published point source solutions for this event: centroid moment tensor (CMT) and body wave (BW) determinations produced by the USGS, the Global CMT Catalog (GCMT) solution, unconstrained and double-couple (DC) sources obtained by the Southern California Seismic Network (SCSN) and the results published by *Hauksson et al.* [2008].

Note that the noise model can also incorporate effects such as failure of individual stations—for example, randomly chosen traces in the training data may be "zeroed out"—although this is not included in the examples presented here.

3.2. Resolution Tests

We perform two types of synthetic tests. First, a synthetic test data set containing 10,000 examples of source vectors and corresponding noisy synthetics is generated from the Green's functions database in the same way as the neural network training sets. The test set is subsequently presented to the trained neural network ensembles, and the obtained posterior pdfs are compared to the known target value. As a visual quality assessment the mode of the posterior marginal pdf is plotted against the target value in Figure 2. The gray scale corresponds to the information gain measured as the Kullback-Leibler divergence D_{KL}, which is a measure of posterior uncertainty reduction compared to the prior distribution (see *Käufl et al.* [2014], for details). A D_{KL} value close to zero indicates that the posterior closely resembles the prior distribution and nothing has been learned from the data, whereas relatively larger values correspond to more pronounced posterior pdfs. We therefore expect examples with a relatively high D_{KL} value (darker colors) to lie close to the diagonal in Figure 2, whereas the posterior mode has little or no significance for pdfs which are similar to the prior (lighter colors). Figure 2 gives an indication of the test set performance for networks trained using 30 s of waveform data. Other cases are provided in the supporting information to this article. Furthermore, histograms of D_{KL} for the test set are provided in Figure S8 for 6, 15, and 30 s of data, which give an impression to what extent each parameter can be resolved.

From Figures 2 and S8 we conclude that the parameters M_w, the source location, and the half duration are resolved best, κ can be generally resolved well up to a rotation of the source by 180°. The resolving power on nondouble component γ, rake σ, and cos(dip) h, while generally being lower, significantly improves as more data are used. This is partly due to the increased station coverage, since the longer the time windows become, the more stations have observed seismic arrivals and partly due to the fact that more phases have arrived at the stations closer to the source.

Moreover, as a second test, we calculate synthetics for a point source that is not located on one of the grid points of the synthetic Green's functions database by running an additional full moment tensor simulation with SPECFEM3D and subsequently presenting the data to the network ensembles. By that we intend to test the ability of the neural networks to interpolate between the fixed training set locations. We show 1-D posterior marginal distributions for the nine source parameters in Figure 3 (left) as a function of the window length and note that the posterior pdfs assign probability mass in the vicinity of the target value as expected. For all parameters uncertainties generally decrease as longer time windows are being used.

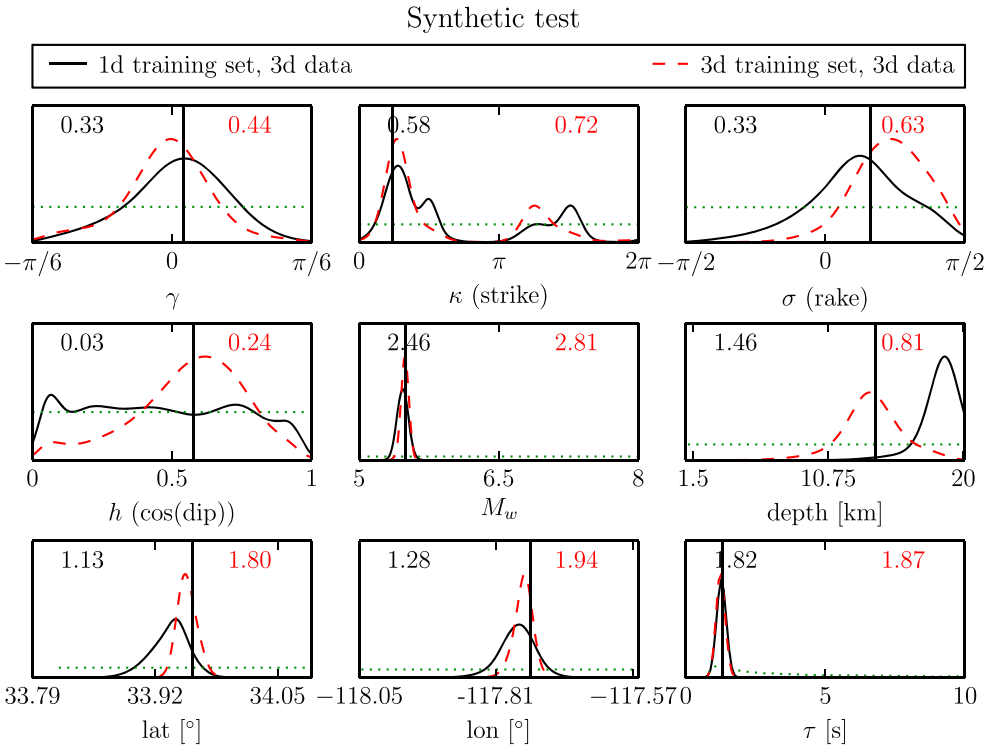

Figure 4. Synthetic 3-D waveform data presented to networks trained using a 1-D Green's functions database (black, solid line). For reference, also results obtained with networks trained using the 3-D Green's functions database are shown (red, dashed line). The target value for each parameter is denoted by the vertical line. The numbers plotted in each panel correspond to the information gain D_{KL}.

3.3. The 2008 M_w 5.4 Chino Hills Earthquake

The M_w 5.4 Chino Hills earthquake occurred on 29 July 2008 in the highly populated Los Angeles region. The main shock of the earthquake sequence is located between the Whittier and Chino sections of the Elsinore fault zone (see Figure 1) and has been detected by a dense strong-motion and broadband seismometer network. The SCSN released a first location, and local magnitude estimate 80 s and a first automatic moment tensor solution ~10 min after origin time [*Hauksson et al.*, 2008].

Here we present strong-motion waveform data recorded at 17 stations to a set of trained neural network ensembles to obtain a first robust, probabilistic estimate of moment tensor point source parameters based on accurate 3-D simulations. Using a 6 s data window allows us to obtain a first estimate of all source parameters ~14 s after origin time (o.t.), with good resolution on magnitude, epicentral location, and source half duration. If longer data windows are used, resolution of all parameters improves: around 15 s of data provides sufficient information to constrain the strike, and resolving the nondouble-couple component requires around 21s (i.e., this information is available 29s after o.t.). Larger time windows show a preference for only a weak nondouble-couple component, in accordance with most other published solutions for this event. The posterior mode for the rake angle appears somewhat unstable initially but converges to values around $\pi/4$ in accordance with other solutions, although subject to relatively large uncertainties. The dip angle cannot be resolved well even for large time windows, and depth can only be constrained if about 12s of data is used. Initially, a deeper centroid is preferred, while a window of above around 24s causes the depth estimate to converge to an intermediate depth.

Including even longer time windows did not—with the given set of 17 stations—significantly improve the result; longer epicentral distances would have to be included, which in turn would have increased the computational cost required for network training due to the increased input dimensionality. However, it is important to recognize that such costs are incurred before the monitoring system becomes operational. Once appropriate data windows have been collected for a given event, evaluating the neural network mapping to obtain pdfs on source parameters can be achieved within milliseconds on a standard desktop workstation. We envisage that a practical implementation might include multiple systems, acting on a range of different window

lengths, and allowing the need to obtain results rapidly to be balanced against the greater certainty possible when more data are used. The precise nature of such a setup, and the manner in which the various results are used operationally, is likely to be region and case dependent, and we do not discuss this further here.

3.4. 3-D Versus 1-D Green's Functions

It is instructive to compare results obtained using the training set based on 3-D Green's functions with predictions made using a 1-D training set. We therefore generate a set of 1-D Green's functions in a layered 1-D model (see Table S2) obtained from the 3-D model by averaging spatially, using a propagator matrix method [*O'Toole and Woodhouse*, 2011] for the same source locations as in the 3-D database and train neural network ensembles using the same setup as before. We subsequently present 30 s of synthetic waveform data, obtained in the full 3-D model, to networks trained on 1-D and 3-D synthetics. We find that in most cases the predictions obtained using the 1-D training set are more uncertain but compatible with the 3-D results (Figure 4). An exception is depth, for which the results obtained using the 1-D networks are misleading and the centroid is located significantly deeper than the target value. It appears that details present in the 3-D waveforms, which are not represented in the 1-D data set, can be used to constrain point source parameters to a greater extent, than when using a 1-D data set.

4. Conclusions

We have demonstrated how a probabilistic pattern-recognition-based framework can be used to invert near-source strong-motion observations for point source parameters based on simulation of wave propagation in a realistic 3-D heterogeneous medium in real-time. The key to our approach is that all expensive calculations are performed in advance of any data collection: real-time analysis can then be performed with minimal computational cost. We find that the incorporation of 3-D Green's functions in the period range 6–100 s leads to a reduction of posterior uncertainties on point source parameters compared to 1-D Green's functions. The inversion of observed strong-motion data for the 2008 M_w 5.4 Chino Hills earthquake yields results compatible to catalog solutions obtained using regional and teleseismic data sets. However, with the given training set, receiver configuration, and period range, we are not able to resolve dip and the posterior uncertainties for the parameters governing nondouble-couple component and rake are large.

Acknowledgments

We thank Carl Tape for his support with setting up the 3-D spectral element simulations. We have used waveform data collected and provided by the U.S. Geological Survey and California Geological Survey. Fault traces have been taken from the Quaternary fault and fold database for the United States, downloadable from the USGS web site: (http://earthquakes.usgs.gov/hazards/qfaults/, last accessed April 2015). The synthetic seismic data set generated for this study may be downloaded from http://www.geo.uu.nl/~jeannot/. P.K. and A.P.V. were supported by the Netherlands Organization for Scientific Research (NWO) under *Top-Subsidie* 854.10.002, and the use of the high-performance computer CARTESIUS was sponsored by the NWO under project SH-296-14.

References

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford Univ. Press, Oxford, U. K.

Crotwell, H. P., T. J. Owens, and J. Ritsema (1999), The TauP toolkit: Flexible seismic travel-time and ray-path utilities, *Seismol. Res. Lett.*, *70*(2), 154–160.

Ekström, G., M. Nettles, and A. M. Dziewonski (2012), The global CMT project 2004–2010: Centroid-moment tensors for 13,017 earthquakes, *Phys. Earth Planet. Inter.*, *200-201*, 1–9, doi:10.1016/j.pepi.2012.04.002.

Farr, T., et al. (2007), The shuttle radar topography mission, *Rev. Geophys.*, *45*, RG2004, doi:10.1029/2005RG000183.

Hauksson, E., K. Felzer, D. Given, M. Giveon, S. Hough, K. Hutton, H. Kanamori, V. Sevilgen, S. Wei, and A. Yong (2008), Preliminary report on the 29 July 2008 M_w 5.4 Chino Hills, Eastern Los Angeles Basin, California, earthquake sequence, *Seismol. Res. Lett.*, *79*(6), 855–866, doi:10.1785/gssrl.79.6.855.

Käufl, P., A. P. Valentine, T. B. O'Toole, and J. Trampert (2014), A framework for fast probabilistic centroid-moment-tensor determination-inversion of regional static displacement measurements, *Geophys. J. Int.*, *196*(3), 1676–1693.

Käufl, P., A. Valentine, R. de Wit, and J. Trampert (2015), Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition, *Bull. Seismol. Soc. Am.*, *105*(4), 2299–2312.

Käufl, P., A. Valentine, R. de Wit, and J. Trampert (2016), Solving probabilistic inverse problems rapidly with prior samples, *Geophys. J. Int.*, *205*(3), 1710–1728.

Komatitsch, D., J. Ritsema, and J. Tromp (2002), The spectral-element method, Beowulf computing, and global seismology, *Science*, *298*(5599), 1737–1742, doi:10.1126/science.1076024.

Laske, G., G. Masters, Z. Ma, and M. Pasyanos (2013), Update on CRUST1.0—A 1-degree global model of Earth's crust, *Geophys. Res. Abstracts, 15*, Abstract EGU2013–2658.

O'Toole, T. B., and J. H. Woodhouse (2011), Numerically stable computation of complete synthetic seismograms including the static displacement in plane layered media, *Geophys. J. Int.*, *187*(3), 1516–1536, doi:10.1111/j.1365-246X.2011.05210.x.

Plesch, A., et al. (2007), Community Fault Model (CFM) for Southern California, *Bull. Seismol. Soc. Am.*, *97*(6), 1793–1802, doi:10.1785/0120050211.

Taborda, R., and J. Bielak (2014), Ground-motion simulation and validation of the 2008 Chino Hills, California, earthquake using different velocity models, *Bull. Seismol. Soc. Am.*, *104*(4), 1876–1898, doi:10.1785/0120130266.

Tape, C., Q. Liu, A. Maggi, and J. Tromp (2009), Adjoint tomography of the Southern California crust, *Science*, *325*(August), 988–992.

Tape, W., and C. Tape (2012), A geometric setting for moment tensors, *Geophys. J. Int.*, *190*(1), 476–498, doi:10.1111/j.1365-246X.2012.05491.x.

Trnkoczy, A. (2002), Understanding and parameter setting of STA/LTA trigger algorithm, in *New Manual of Seismological Observatory Practice (NMSOP)*, vol. 2, pp. 1–19, GFZ, Potsdam.

Tromp, J., D. Komattisch, and Q. Liu (2008), Spectral-element and adjoint methods in seismology, *Commun. Comput. Phys.*, *3*(1), 1–32.