Bulletin of the Seismological Society of America, Vol. 98, No. 5, pp. 2262–2277, October 2008, doi: 10.1785/0120080033

Probability of Occurrence of Velocity Pulses in Near-Source Ground Motions

by Iunio Iervolino and C. Allin Cornell

Abstract Near-source ground-motion records affected by directivity may show unusual features in the signal resulting in low-frequency cycle pulses in the velocity time history, especially in the fault-normal component. Such an effect causes the seismic demand for structures to deviate from that of so-called ordinary records. This circumstance may be particularly hazardous for structural engineering applications if it is not properly accounted for. In fact, current attenuation laws are not able to capture such effects well, if at all, and therefore current probabilistic seismic hazard analysis (PSHA) is not able to predict this peculiar spectral shape. This failure may possibly lead to an underestimation of, in particular, the nonlinear demand. Accounting for pulse-type records in earthquake engineering practice should be reflected both in the PSHA and in the record selection for seismic assessment of structures. These applications require a model for the probability of occurrence of pulselike records. Herein such a model is proposed on an empirical basis. A set of pulselike fault-normal ground motions from the Next Generation Attenuation of Ground Motions (NGA) Project dataset, as systematically identified by Baker (2007), is used. The independent variables studied are chosen from those considered by seismologists to affect the amplitude of directivity pulses. Issues related to the dataset and the explanatory power of the proposed models are also discussed.

Introduction

A site located close to the source of a seismic event may be in a geometrical configuration with respect to the propagating rupture that favors the constructive interference of the approaching waves (i.e., a synchronization of phases causing a buildup of energy) resulting in a large velocity pulse. This phenomenon requires the rupture propagating toward the site and the alignment of the site with the slip of the fault. If these two conditions are met, the ground motion at the site may show forward directivity effects. In fact, directivity causes, in theory, full-cycle velocity pulses while the fling step, which is related to the permanent tectonic deformation at the site, is believed to cause half-cycle pulses (Bolt and Abrahamson, 2003).

Parameters believed to affect the amplitude of the pulse are related to the aforementioned rupture-to-site geometry, while empirical models positively correlating the earthquake's magnitude to the period of the pulse (T_p) have been proposed (e.g., Somerville [2003]); global geophysics-based directivity predictors are also available (e.g., Spudich *et al.* [2004]).

Pulse-type records are of interest to structural engineers because (1) they may induce unexpected demand in structures having a fundamental period equal to a certain fraction of the pulse period and (2) such a demand may not be adequately captured by the current best-practice ground-motion intensity measures such as first-mode spectral acceleration (Howard *et al.*, 2005; Tothong and Luco, 2007). An effective way to visualize the hazardous features of pulselike records is a plot of the inelastic ($S_{d,i}[T]$) to elastic ($S_{d,e}[T]$) displacement ratio versus the oscillation period (T) of a bilinear single degree of freedom system (SDOF) with a 5% hardening stiffness and damping ratio. In Figure 1, where such a plot is given for an SDOF with a strength reduction factor (R_s) equal to 4, the abscissa is T normalized by the pulse period of the record T_p (Tothong and Cornell, 2006).

The panels in Figure 1 are given, as an example, grouping pulselike records by the pulse period for three T_p values. The plots show a bump in the ratio $S_{d,i}/S_{d,e}$ between inelastic and elastic demand at $T/T_p \cong 0.5$, indicating a comparatively large inelastic demand of this kind of near-source ground motions, which may not be similar to that of ordinary records and, therefore, calling for a specific investigation about its occurrence. This is particularly clear in Figure 1b,c, while in Figure 1a, where T_p is short, the pulse effect is partially overwhelmed by the high-frequency content of the ground motions.

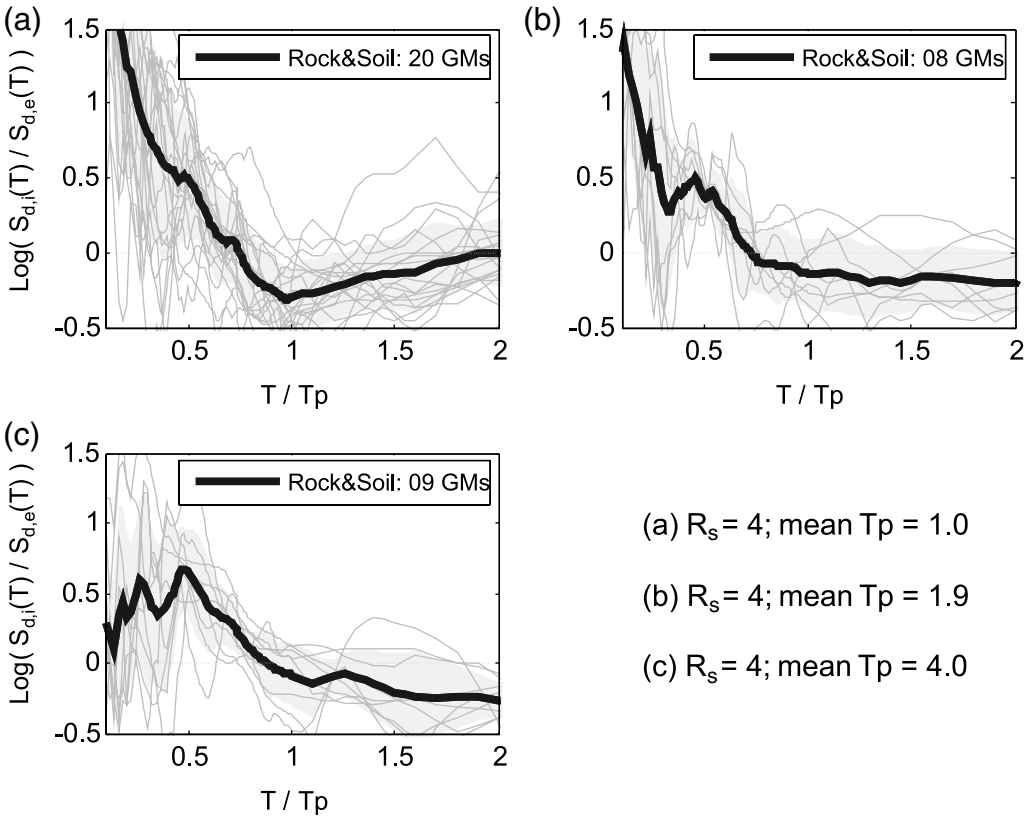

Figure 1. Empirical pulselike records' inelastic to elastic displacement demand ratios (Tothong and Cornell, 2006).

Hazard Analysis in the Near Source and the Need for a Pulse Occurrence Probability Model

Because not all near-source ground-motion records show a pulse in the velocity time history, it may be argued that near-source records do not always induce nonordinary seismic demand for structures. Near-source records that do not contain a pulse display virtually the same response behavior as far-field records (e.g., Tothong and Cornell [2006]). Therefore, the current distinction of far-field and near-source records may not be the most practical; it should be replaced by ordinary versus pulselike ground motions.

It is clear that it is not possible to apply the current earthquake engineering practice to the near source, and the procedures have to be reviewed and adjusted consistently. A rational approach to the seismic risk analysis requires a probabilistic model for the occurrence of directivity effects in ground motions. The systematic deviations of pulselike signals with respect to the ordinary imply that, in the probabilistic assessment of structures, a pulse occurrence model is required to incorporate such effects accurately in the probabilistic seismic hazard analysis (PSHA). The phenomena should also be reflected in the record selection, because the latter should be related with the disaggregation of seismic hazard (Cornell, 2004). This issue is briefly reviewed in the following, although for a more comprehensive review, the reader should refer to the article by Tothong *et al.* (2007).

Assuming that all seismic sources are within 30 km from a certain site of interest and given that, as discussed, not all near-source (NS) ground motions are pulselike, the PSHA expressed as the mean annual frequency ($\lambda_{S_a,\mathrm{NS}}$) of the spectral acceleration (S_a) exceeding a certain value (x) should be separated into two terms:

$$\lambda_{S_a,\mathrm{NS}}(x) = \lambda_{S_a,\mathrm{NS\ \&\ pulse}}(x) + \lambda_{S_a,\mathrm{NS\ \&\ no\ pulse}}(x). \qquad (1)$$

One of the two, the near-source nonpulselike ($\lambda_{S_a,\mathrm{NS\ \&\ no\ pulse}}$), should be from, say, ordinary PSHA, which requires a near-source attenuation law computed with records not showing pulses but still coming from short source-to-site distances (e.g., within 30 km). The second part should be the near-source term ($\lambda_{S_a,\mathrm{NS\ \&\ pulse}}$) due to pulselike records. This requires ground-motion prediction relationships able to capture the peculiar spectral shape driven by the pulses. These so-called narrowband attenuation laws are currently under the attention of seismologists, for example, the Next Generation Attenuation of Ground Motions (NGA) Project (see Data and Resources section). In this case, the attenuation law will not only depend on magnitude and distance but also on a vector of other parameters ($\underline{Z}$) that are assumed to be meaningful to predict directivity effects. The total hazard is the linear combination of the two hazard curves weighted by the pulse occurrence probability as in equations (2) and (3), in which a single fault is assumed:

$$\lambda_{S_a,\text{NS \& pulse}}(x) = \nu \int_m \int_r \int_{\underline{z}} \int_{t_p} P[\text{pulse}|m, r, \underline{z}] \times G_{S_a|\text{pulse},M,R,\underline{Z},T_p}(x|m, r, \underline{z}, t_p) \times f_{T_p|\underline{Z},M,R} f_{\underline{Z}|M,R} f_{M,R}\, dt_p\, d\underline{z}\, dm\, dr, \quad (2)$$

$$\lambda_{S_a,\text{NS \& no pulse}}(x) = \nu \int_m \int_r \int_{\underline{z}} (1 - P[\text{pulse}|m, r, \underline{z}]) \times G_{S_a|\text{no pulse},M,R}(x|m, r) \times f_{\underline{Z}|M,R} f_{M,R}\, d\underline{z}\, dm\, dr. \quad (3)$$

In equation (2), ν is the mean rate of events on the fault, M is the magnitude of the event, and R is the source-to-site distance. dt_p and $d\underline{z}$ are the integration intervals of the variables pulse period, T_p, and $\underline{Z}$, respectively. $G_{S_a|\text{pulse},M,R,\underline{Z},T_p}$ is the complementary cumulative distribution function of S_a conditional on M, R, $\underline{Z}$, and T_p; $f_{T_p|\underline{Z},M,R}$ is the probability density function (PDF) of T_p given M, R, and $\underline{Z}$. Similarly, $f_{\underline{Z}|M,R}$ is the conditional distribution of $\underline{Z}$ given M and R, while $f_{M,R}$ is the joint PDF of M and R. The same meaning of the symbols applies to equation (3).

The conditional probability of having a pulse is needed to evaluate equations (2) and (3). In the following, empirical pulse probability models, based on logistic regression for strike-slip (SS) and non-strike-slip (NSS) rupture data, are proposed and results discussed.

The Issue of Identifying a Pulse and the Dataset Used

It is a nonstraightforward task to ascertain whether a record shows directivity effects, for example, a pulse in the velocity time history, and its properties such as the period T_p. Many seismologists and other earthquake science experts have engaged in this exercise but no widely accepted method is readily available. The bulk of the difficulties in identifying a pulse in the ground motion are related to the wave propagation effects and to the higher frequency content that may give an unclear picture of the directivity features. A common option is to visually analyze the waveform looking for pulses, but this method requires strong expertise in the field and may be not very efficient for short-period pulses or for small or moderate magnitude events, where the pulse may be lost in the high frequency. Above all, this method does not allow one to investigate large datasets looking for the fraction of signals showing directivity effects.

Baker (2007) analyzed extensively the NGA database, and he is the only researcher we are aware of who has looked systematically at all records in the database. Therefore, we know which records are the pulses and also which records are the nonpulses, which is crucial to develop any pulse occurrence probability model. Baker (2007) developed a method based on wavelets to assign a score, a real number between 0 and 1, to each analyzed record and to determine the pulse period. The larger the score determined the more likely the record was to show a pulse. In this way, J. W. Baker (personal comm., 2006) has found pulselike records in both fault-normal and fault-parallel components of the ground motions investigated. Herein only those in the fault-normal component have been considered; in particular, those ground motions that have a pulse score larger or equal to 0.85 have been, arbitrarily, counted as pulse-type records. Of these records, 98 are classified as within 30 km of the fault by the NGA flat file (see Data and Resources section). Six of them of them do not have a measure of the closest distance to fault rupture but their epicentral distance is within 30 km; however, they still have not been included herein because in the NGA they lack information about geometry of the fault/site that is useful for predicting directivity. This set has also to be cleared of those considered as late pulses, for example, occurring at the end of the records and, therefore, too late to be directivity caused; there are 19 of these records (J. W. Baker, personal comm., 2006). The resulting dataset, given in Table 1, consists of 73 records from 23 events;[1] 12 of those are SS. The events' magnitude ranges from M 5.2 to 7.5.

This study proposes models for the estimation of the pulse occurrence probability based on empirical evidence (i.e., relative frequency). Therefore, the complementary set of identified nonpulse records is needed. The total database considered in the following is made of records within 30 km (in terms of closest distance to fault rupture) coming from the NGA catalog[2] and whose characteristics have been determined via the NGA flat file and related documentation (see Data and Resources section).

In the flat file, the total number of events matching the selection requirements discussed previously and featuring records within 30 km in terms of closest distance to fault rupture is 45 (this, again, excludes Chi-Chi and aftershocks); 22 of them are SS. The number of records from these SS events is 133. Because the records identified as pulses in the given dataset are 34, the marginal SS pulse occurrence probability is 34/133 or 26%. For the NSS[3] case, the identified pulse records are 39 among 229 within 30 km; therefore, the marginal probability of a pulse within 30 km is 39/229 or 17%. Pooling together SS and NSS data, the overall pulse probability in records within 30 km of the fault is about 20%. The modeling in the following will describe how this probability depends on geometry and magnitude.

[1]Note that the Chi-Chi pulse records within 30 km (about 80 in number) were excluded. This is primarily because we did not want to have to define whether Chi-Chi is SS or NSS. In fact, it is not straightforward as the very long rupture did not slip in a prevalent direction during the event.

[2]Also, those records not having a fault-normal component in the NGA database at time of access are not considered herein.

[3]Records with unknown mechanism have been included in non-SS events.

Table 1

Pulses' Dataset

Earthquake Name	Date (mm/dd/yyyy)	Magnitude	Mechanism	Station Name
San Fernando	02/09/1971	6.6	Reverse	LA—Hollywood Stor FF
San Fernando	02/09/1971	6.6	Reverse	Lake Hughes #1
San Fernando	02/09/1971	6.6	Reverse	Lake Hughes #4
San Fernando	02/09/1971	6.6	Reverse	Pacoima Dam (upper left abut)
Friuli, Italy—02	09/15/1976	5.9	Reverse	Buia
Santa Barbara	08/13/1978	5.9	Reverse-Oblique	Santa Barbara Courthouse
Coyote Lake	08/06/1979	5.7	Strike-Slip	Coyote Lake Dam (SW Abut)
Coyote Lake	08/06/1979	5.7	Strike-Slip	Gilroy Array #6
Coyote Lake	08/06/1979	5.7	Strike-Slip	SJB Overpass, Bent 3 g.l.
Coyote Lake	08/06/1979	5.7	Strike-Slip	SJB Overpass, Bent 5 g.l.
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	Aeropuerto Mexicali
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	Agrarias
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	Brawley Airport
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	EC County Center FF
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	EC Meloland Overpass FF
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #10
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #11
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #3
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #4
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #5
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #6
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #7
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Array #8
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	El Centro Differential Array
Imperial Valley—06	10/15/1979	6.5	Strike-Slip	Holtville Post Office
Mammoth Lakes—02	05/25/1980	5.7	Strike-Slip	Convict Creek
Irpinia, Italy—01	11/23/1980	6.9	Normal	Bagnoli Irpino
Irpinia, Italy—01	11/23/1980	6.9	Normal	Sturno
Westmorland	04/26/1981	5.9	Strike-Slip	Parachute Test Site
Morgan Hill	04/24/1984	6.2	Strike-Slip	Coyote Lake Dam (SW Abut)
Morgan Hill	04/24/1984	6.2	Strike-Slip	Gilroy Array #6
Morgan Hill	04/24/1984	6.2	Strike-Slip	Hollister Diff Array #1
Drama, Greece	11/09/1985	5.2	Normal-Oblique	Drama (basement)
North Palm Springs	07/08/1986	6.1	Reverse-Oblique	North Palm Springs
San Salvador	10/10/1986	5.8	Strike-Slip	Geotech Investigation Center
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Bell Gardens—Jaboneria
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Compton—Castlegate Street
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Downey—Co Maintenance Building
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Glendale—Las Palmas
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	LA—West 70th Street
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	LB—Orange Avenue
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	LB—Rancho Los Cerritos
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Lakewood—Del Amo Boulevard
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Norwalk—Imp Highway, South Grnd
Whittier Narrows—01	10/01/1987	6.0	Reverse-Oblique	Santa Fe Springs—East Joslin
Superstition Hills—02	11/24/1987	6.5	Strike-Slip	Parachute Test Site
Loma Prieta	10/18/1989	6.9	Reverse-Oblique	Gilroy Array #2
Loma Prieta	10/18/1989	6.9	Reverse-Oblique	Saratoga—Aloha Avenue
Erzican, Turkey	03/13/1992	6.7	Strike-Slip	Erzincan
Cape Mendocino	04/25/1992	7.0	Reverse	Fortuna—Fortuna Boulevard
Cape Mendocino	04/25/1992	7.0	Reverse	Petrolia
Landers	06/28/1992	7.3	Strike-Slip	Lucerne
Landers	06/28/1992	7.3	Strike-Slip	Yermo Fire Station
Northridge—01	01/17/1994	6.7	Reverse	Jensen Filter Plant
Northridge—01	01/17/1994	6.7	Reverse	Jensen Filter Plant Generator
Northridge—01	01/17/1994	6.7	Reverse	LA—Century City CC North
Northridge—01	01/17/1994	6.7	Reverse	LA—Wadsworth Virginia Hospital North
Northridge—01	01/17/1994	6.7	Reverse	LA Dam
Northridge—01	01/17/1994	6.7	Reverse	Lake Hughes #9
Northridge—01	01/17/1994	6.7	Reverse	Newhall—West Pico Canyon Road
Northridge—01	01/17/1994	6.7	Reverse	Pacoima Dam (downstream)

(*continued*)

Table 1 (*Continued*)

Earthquake Name	Date (mm/dd/yyyy)	Magnitude	Mechanism	Station Name
Northridge—01	01/17/1994	6.7	Reverse	Pacoima Dam (upper left)
Northridge—01	01/17/1994	6.7	Reverse	Rinaldi Receiving Station
Northridge—01	01/17/1994	6.7	Reverse	Sylmar—Converter Station
Northridge—01	01/17/1994	6.7	Reverse	Sylmar—Converter Station East
Northridge—01	01/17/1994	6.7	Reverse	Sylmar—Olive View Med FF
Kobe, Japan	01/16/1995	6.9	Strike-Slip	Takarazuka
Kobe, Japan	01/16/1995	6.9	Strike-Slip	Takatori
Kocaeli, Turkey	08/17/1999	7.5	Strike-Slip	Arcelik
Kocaeli, Turkey	08/17/1999	7.5	Strike-Slip	Gebze
Duzce, Turkey	11/12/1999	7.1	Strike-Slip	Lamont 1060
Sierra Madre	06/28/1991	5.6	Reverse	LA—City Terrace
Sierra Madre	06/28/1991	5.6	Reverse	San Marino—SW Academy

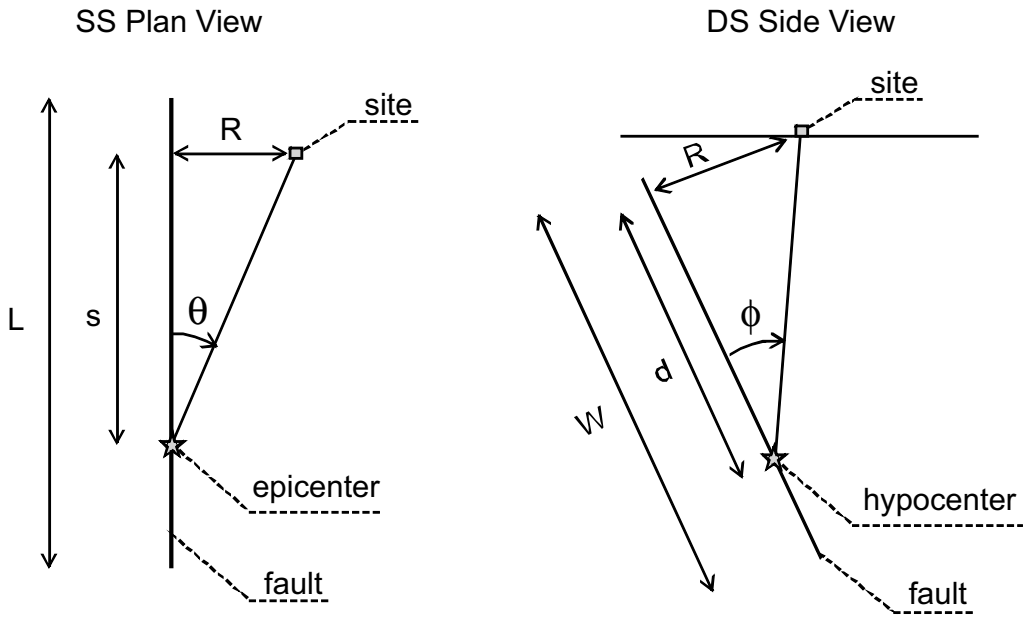

Figure 2. SS and DS directivity schematics (Somerville *et al.*, 1997).

Directivity Factors and Pulse Occurrence Covariates

In Somerville *et al.* (1997), for SS events, the amplitude of spectral modification of ordinary attenuation laws due to directivity in ground motions depends on $X\cos(\theta)$, where $X = s/L$ is the ratio of the distance from the epicenter to the site (measured along the rupture direction) and the fault length; θ is the angle between the fault strike and the path to the site with respect to the rupture (measured in degrees herein). For dip-slip (DS) events, the analogous parameter is $Y\cos(\phi)$, where $Y = d/W$ and ϕ have similar meaning of X and θ, respectively, if the hypocenter and the plane of the rupture (W is the fault width) are considered in place of the epicenter and the fault direction (Fig. 2).

Other factors that may explain directivity effects are the event's magnitude, which is correlated with the pulse period (Somerville, 2003) and the source-to-site distance. Neither of them appears explicitly in $X\cos(\theta)$ and $Y\cos(\phi)$, although the rupture length is related to magnitude and source-to-site distance is not independent from the geometrical configuration. Recently, also the s distance (d distance) alone has been considered as a meaningful predictor of directivity (N. A. Abrahamson, personal comm., 2005) and it is confirmed in the following.

A rough explanation is that for large s (or d) the chance that the rupture evolves yielding directivity effects increases independently of the fault length. Therefore, the variables considered herein as possible covariates in the model to predict pulse occurrence are (1) the closest distance to fault rupture (R), (2) the event's magnitude (M), (3) the length ratio X (Y for NSS events), (4) the θ angle[4] (ϕ for NSS events), (5) the s distance (d for NSS events), and the Somerville *et al.* (1997) parameter $X\cos(\theta)$ ($Y\cos[\phi]$ for NSS events).

[4]From Figure 2, it is possible to derive a geometrical relationship between the θ (ϕ) angle and the R and s (d) distances. However, because ruptures may not always be represented by straight lines, such relationships may be lost.

In Figures 3–5, the dataset, separated into pulselike and nonpulselike records is represented in terms of several of the covariates listed previously. These plots show, although weakly, empirical trends that the probability models are expected to reproduce. For example, it is possible to observe in Figure 3 that, at least for the SS case, the pulse occurrence likelihood decreases with R and increases with s. Similarly Figure 4 suggests that there is a negative trend of pulse occurrence probability with respect to θ and ϕ, as the fraction of pulselike records seems larger for low values of these angles. At the same time, it should be noted that some pulses correspond to values of the covariates that are unfavorable according to the directivity prediction models discussed previously.

Such figures also give a picture of the covariate ranges and, therefore, of the applicability of the models. In particular, it seems that there is no practical information in the data on pulse occurrence beyond about 40 km in terms of s for the SS case and beyond 20 km in terms of d for the NSS case. Furthermore, insufficient pulse data are available for R below 5 km in the NSS case.

Pulse Occurrence Probability Models Based on Logistic Regression

The occurrence of a pulse in a near-source ground motion may be represented as an indicator variable (I) that can assume the two values: 1 if there is a pulse in the record or 0 if the record does not show a pulse. The probability of the occurrence of the pulse is $p = P[I = 1]$; the probability of the record not showing a pulse is $1 - p = P[I = 0]$. To link this categorical response to a specific variable believed to have some prediction power, the most used model is the logistic regression (Agresti, 2002). Logistic regression assumes the log of the odds ratio to be a linear function of the explanatory variable. This means $\log[p/(1-p)] = \alpha + \beta z$, where p is the occurrence probability given z and $\{\alpha, \beta\}$ are the coefficients to be determined. In general, multivariate logistic regressions are of the type in equation (4), which is written in the case of k predictor variables:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k. \quad (4)$$

Several simple and multiple regression models for pulse occurrence in SS and NSS cases have been investigated in this exploratory study using the *glmfit* tool, which serves to fit generalized linear models, in MATHWORKS—MATLAB® software. Unfortunately, it is not easy to determine the prediction power and to compare logistic models. One qualitative way to determine whether the logistic distribution is a good approximation of the data is to group the sample in z bins and then to estimate p as the ratio of occurrences over

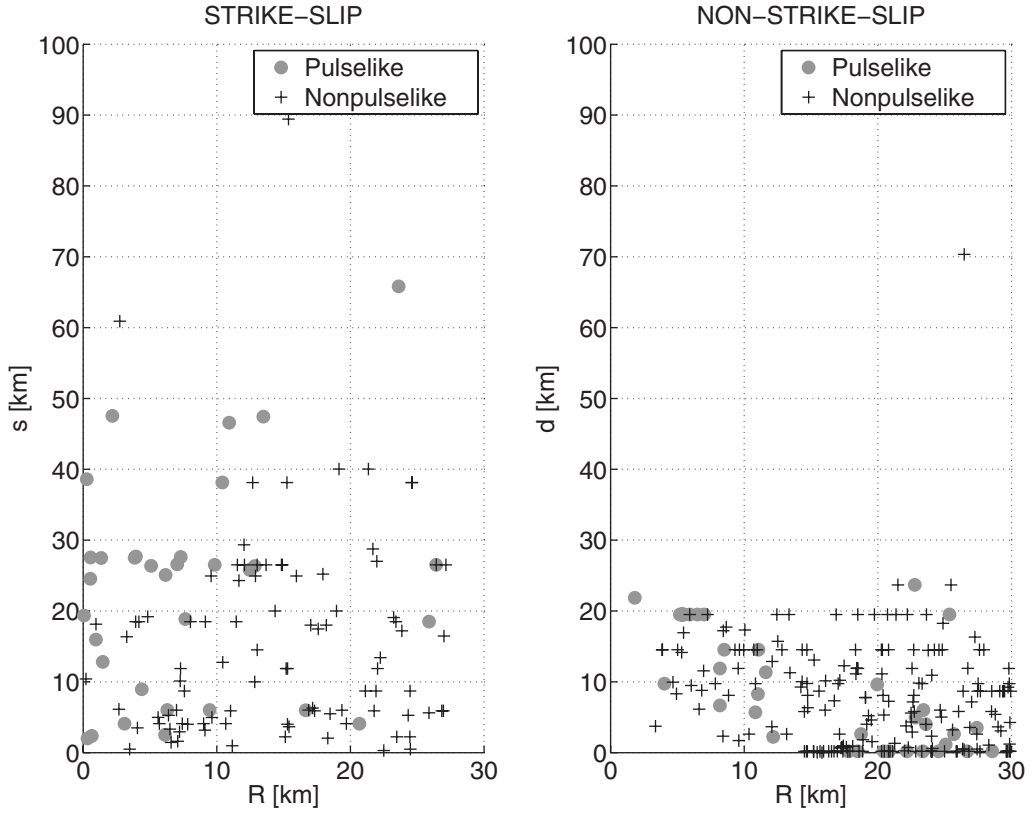

Figure 3. Dataset used in terms of projected distance (along the rupture plane) from the origin of the rupture toward the site versus the closest distance to fault rupture.

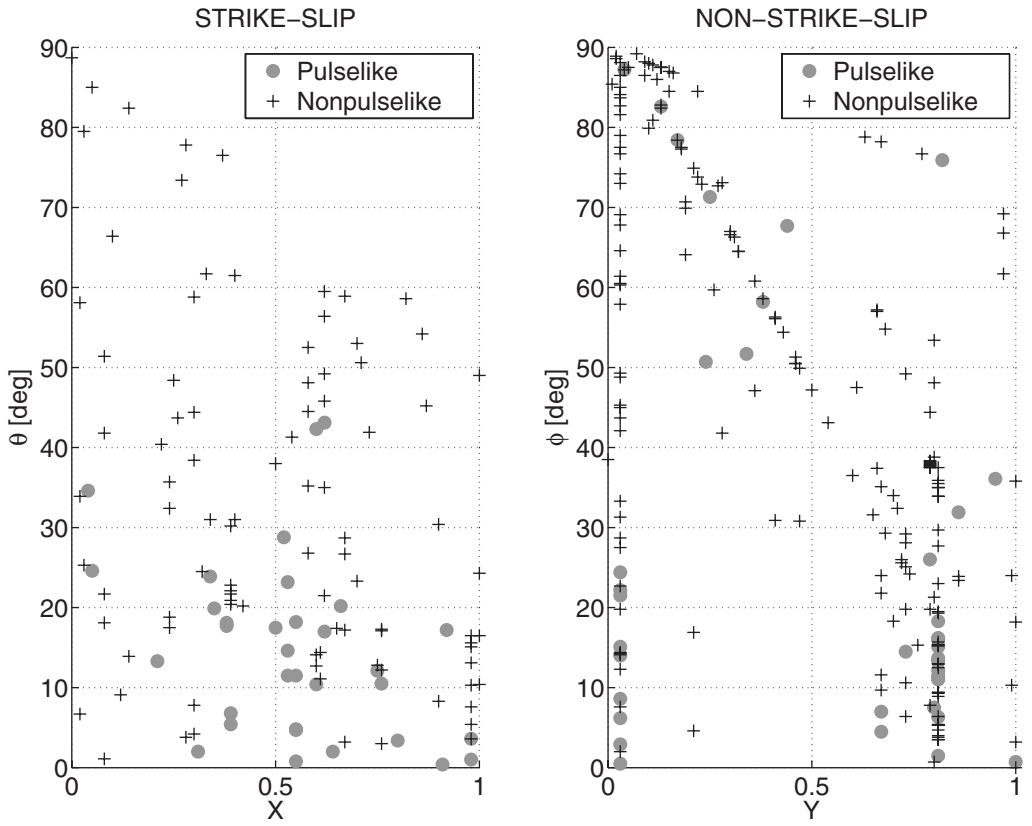

Figure 4. Dataset used in terms of the angle between the site and the fault plane versus the normalized distance.

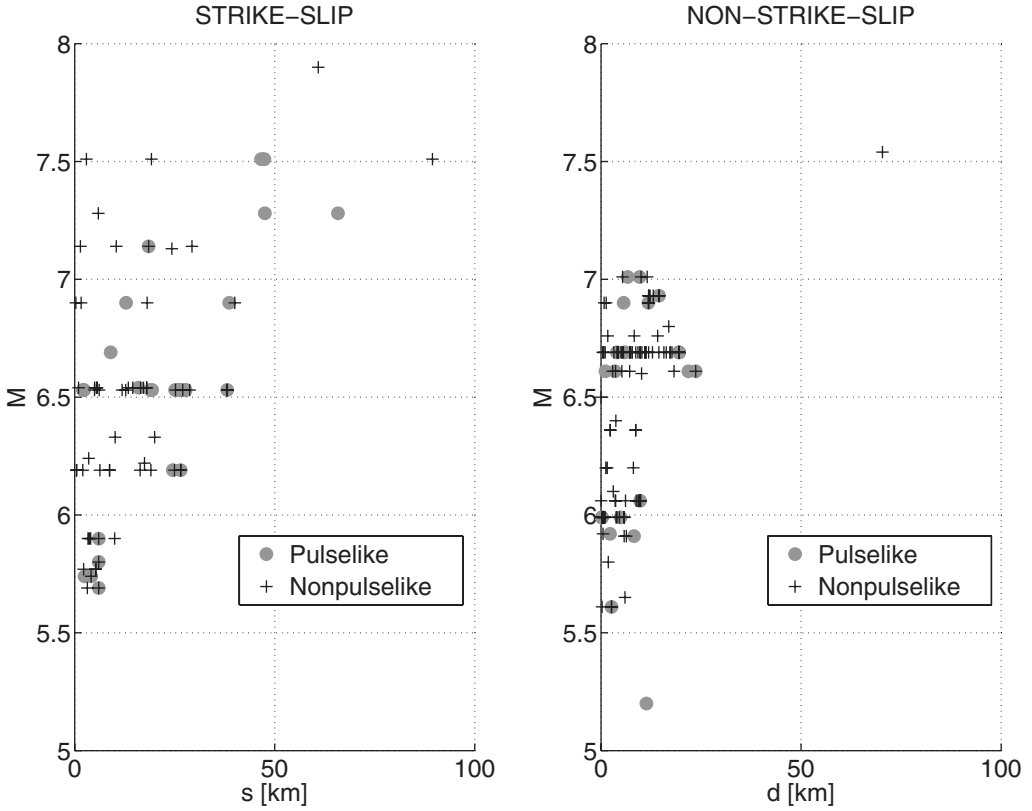

Figure 5. Dataset used in terms magnitude versus the projected distance (along the rupture plane) from the origin of the rupture toward the site.

the number of data within the bin; plotting these frequencies versus the fit gives a picture of the adequacy of the model.

Furthermore, there is no widely accepted direct analog to R^2 as defined for ordinary least-squares regressions. Nonetheless, a number of logistic R^2 measures have been proposed. These approximations of R^2 are not actual percents of variance explained by the model, but rather attempts to measure strength of association. In equations (5) (Efron, 1978) and (6) (McFadden, 1974), two of these measures are given,[5] where n is the sample size, $\hat{p}_i$ is the probability estimated by the model, y_i is the ith binary response in the sample, and $\overline{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$:

$$R_E^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{p})^2}, \tag{5}$$

$$R_{MF}^2 = 1 - \frac{\sum_{i=1}^{n} y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)}{n[\overline{p}\log(\overline{p}) + (1 - \overline{p})\log(1 - \overline{p})]}. \tag{6}$$

To choose the best functional form, the Akaike information criterion (AIC) may be used; it allows one to compare models with a different number of terms counterbalancing the improvement of fit with the manageability of the equation. AIC is defined in equation (7), where the first term in the parentheses is the maximized log likelihood (for n independent Bernoulli trials) and q is the number of parameters (this penalizes the model for having many parameters); it may be stated that the lower the AIC is, the better is the model:

$$\text{AIC} = -2\left\{\left[\sum_{i=1}^{n} y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right] - q\right\}. \tag{7}$$

Univariate Probability Models

Strike-Slip. Simple (univariate) logistic regression allows one to determine the probability trends with respect to those parameters previously discussed. Consider the SS records first: in Figure 6, the continuous estimated conditional occurrence probability is plotted versus each of the predictors along with the pulse observations (as coded by the values of the indicator variable defined). The parameters showing the largest explanatory power with respect to the pulse occurrence are the closest distance to fault rupture, the θ angle, and the distance measured along the rupture. In Figure 6a, it is possible to see the clear decreasing trend of pulse occurrence probability with R. Such a trend is confirmatory with respect to what is qualitatively observed in the left panel of Figure 3. The occurrence probability at zero distance is 0.58 and drops to 0.03 at 30 km. The s distance, Figure 6b, also shows some predictive power with an expected trend. The plot refers to the 0–90-km range, which is the data availability interval. Between these limits, the probability of observing a pulse increases from 0.16 to 0.75; however, as discussed, the actual upper bound for the applicability of the model should be around 40 km, where the occurrence has 0.4 probability. The θ angle, Figure 6c, seems also significant for pulse occurrence, estimating an occurrence probability of 0.54 for a site that is sitting on the line of the rupture (in the most favorable condition to see a pulse) and drops to 0.01 for a site orthogonally placed with respect to the fault in a way that the rupture proceeds beyond it.

Other candidates, to be directivity-related parameters, all seem to have small, if any, predictive power with respect to pulse occurrence probability. It can be observed in Figure 6d, e,f that the occurrence probability does not vary much in the data intervals for the length ratio X and magnitude; a slightly greater trend is shown for $X\cos(\theta)$. (Note that given these results, magnitude can, in principle, be dropped from the conditional pulse occurrence probability in equations 2 and 3.)

The coefficients, $\{\alpha, \beta\}$, relative to the univariate regressions shown in Figure 6, along with pseudo-R^2 measures and AIC scores, are given in Table 2; their values confirm the discussion given previously.

Non-Strike-Slip. The same analyses just described have been repeated for the NSS sample with respect to the directivity predictors that apply to the DS case. Qualitative trends obtained (Fig. 7) are generally the same as the SS case if instead of s, θ, and X, their DS analogs are considered. Moreover, the ranking of the variables in terms of predictive power for the pulse occurrence probability of the SS case also holds for NSS. R seems to significantly affect pulse occurrence as do the ϕ angle and d distance, while $Y\cos(\phi)$ and Y show lower predictive power. The regression on the event's magnitude is still nonsignificant and for the NSS case turns out in a slightly negative trend.

Generally, the predicted probability is lower with respect to the SS case as anticipated by the marginal frequencies observed in the input dataset. In fact, the NSS data are more heterogeneous than SS because they do not focus on a specific fault mechanism (i.e., being generic NSS).

Although it may be problematic to compare directly R^2 for different logistic regressions, the larger values for the SS solutions with respect to those for NSS (Table 3) seem to suggest a lower efficiency of the models for the latter case with respect to the former. This may reflect the fact that the geometry and the physics in the SS case are, in principle, cleaner than in the NSS. Note that the AIC is larger for lower R^2 and for larger sample size.

Multivariate Logistic Regression Models

For a given site and rupture schematic such that near-source conditions occur, specific values for the predictors are available (see Fig. 2); therefore, to perform a PSHA as sug-

[5]These tend to be smaller than ordinary R^2, and values of 0.2–0.4 are considered highly satisfactory.

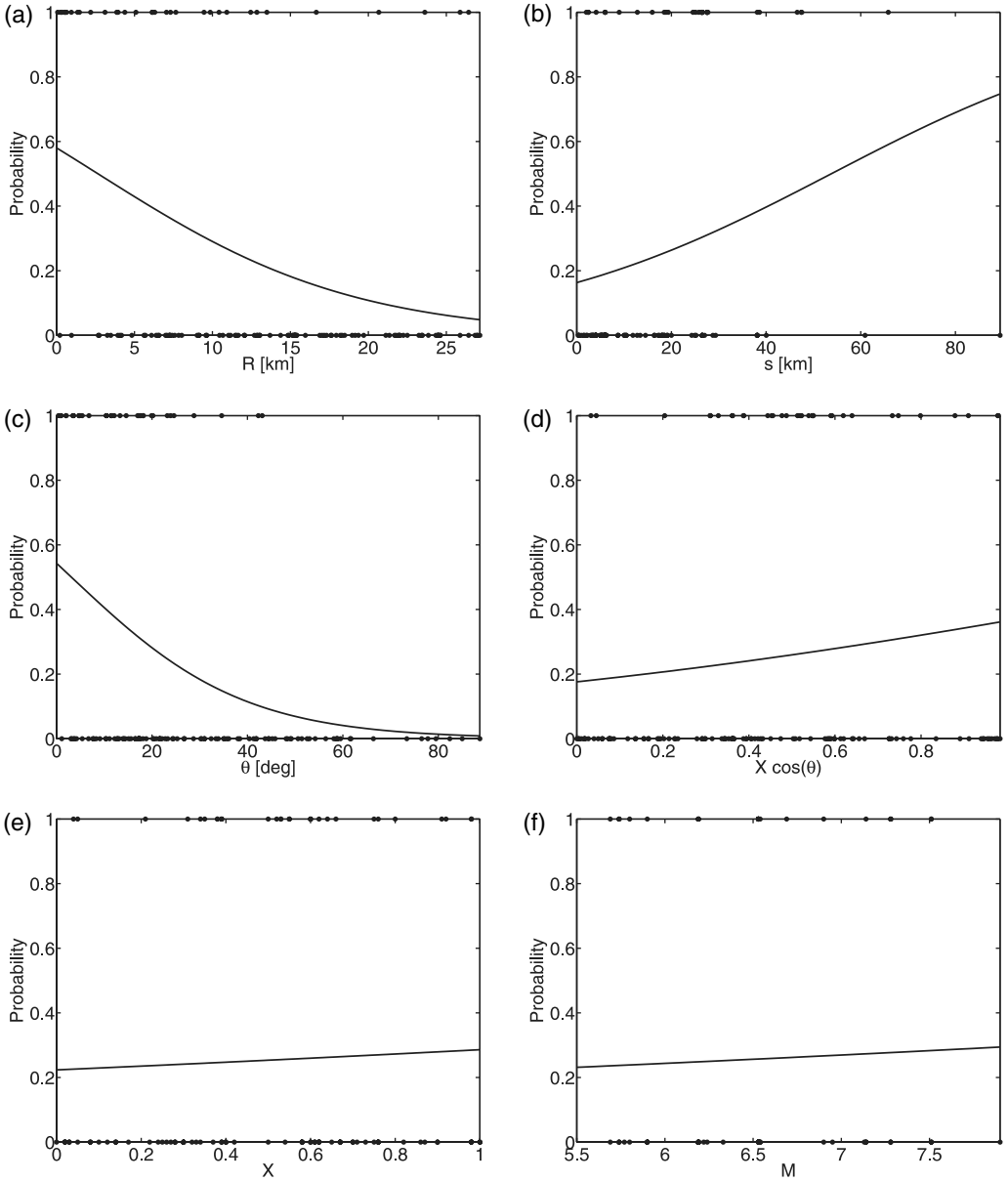

Figure 6. Univariate logistic regressions for the SS case.

gested by equations (2) and (3), the pulse occurrence probability has to be conditioned on a vector of covariates treated jointly. To this aim, multivariate logistic regressions, equation (4), have been investigated to build up pulse occurrence probability models.

Strike-Slip. When constructing a multivariate regression model, choosing the appropriate covariates and terms is not a straightforward task. Including only few covariates may lead to a lower prediction power than including many terms and interactions. On the other hand, a complex model, al-

Table 2
Univariate Logistic Regression Coefficients and Scores for the SS Case

Covariate	Constant Term	Linear Term	R^2_E	R^2_{MF}	AIC
R (km)	0.32347	−0.12169	0.16051	0.12467	136.3561
s (km)	−1.6349	0.030403	0.040238	0.034712	149.958
θ (deg)	0.17263	−0.05545	0.12381	0.12873	135.7425
M	−1.9439	0.13523	0.000848	0.000775	155.0894
X	−1.2452	0.32971	0.001278	0.001522	154.9764
$X\cos(\theta)$	−1.5435	0.99033	0.011531	0.013597	153.1507

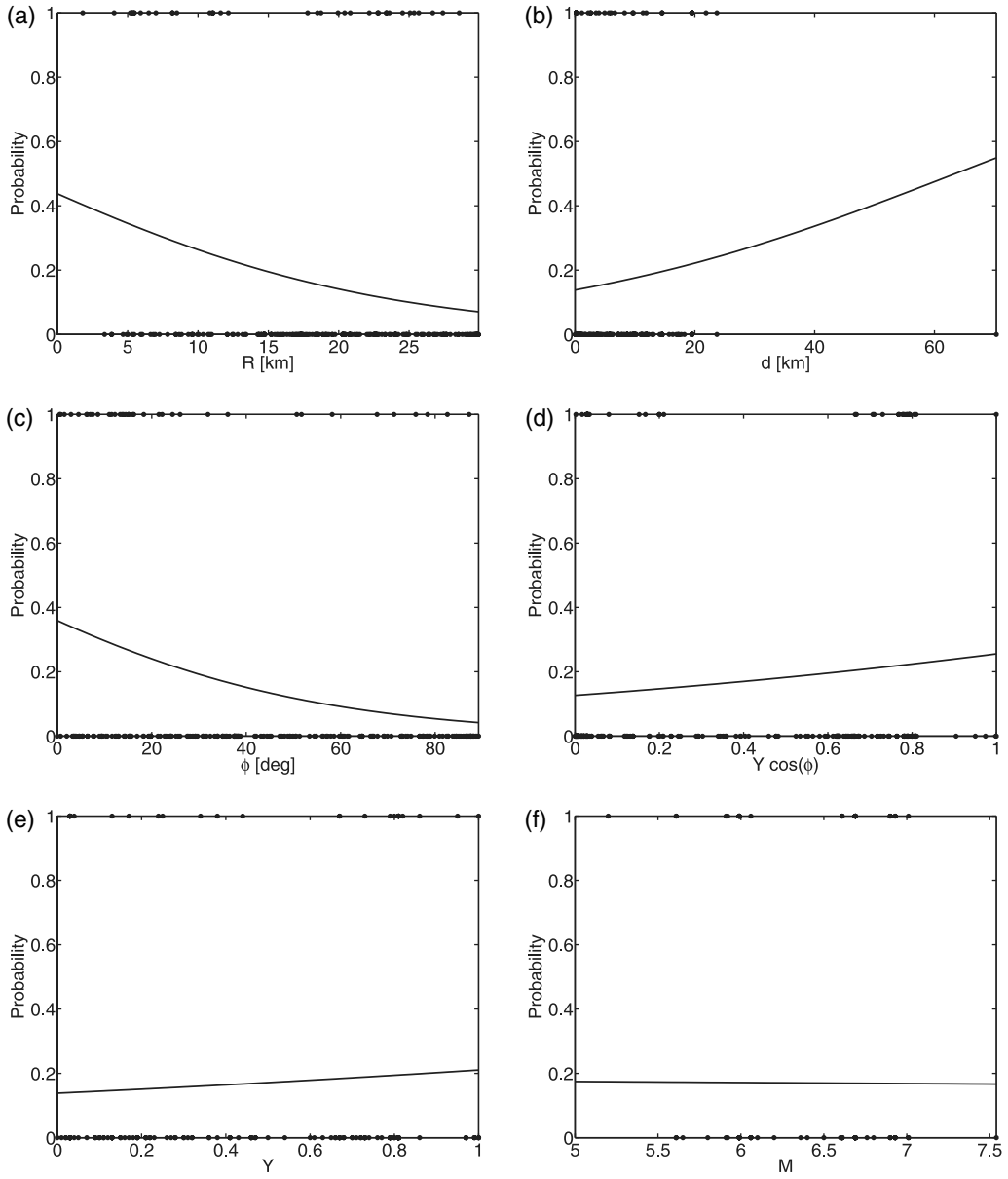

Figure 7. Univariate logistic regressions for the NSS case.

though better representing the input sample, may be less manageable and may lose in generality. Several approaches exist to build up multiple ordinary regression models once a basic set of covariates has been established, such as forward selection and backward elimination (Agresti, 2002), which also apply to the logistic case. A full quadratic model, which would include all of the five basic predictor candidates for the SS case, $\{R, s, \theta, X, M\}$, should have 20 terms because of the interactions and the squared variables. Because the number of pulselike records for the SS case is only slightly

Table 3

Univariate Logistic Regression Coefficients and Scores for the NSS Case

Covariate	Constant Term	Linear Term	R_E^2	R_{MF}^2	AIC
R (km)	−0.25026	−0.07785	0.058661	0.050954	202.3675
d (km)	−1.8323	0.028856	0.007687	0.009624	211.0061
ϕ (deg)	−0.57869	−0.0286	0.082195	0.080198	196.255
M	−1.4399	−0.02245	1.42×10^{-05}	1.36×10^{-05}	213.0149
Y	−1.8301	0.50793	0.004479	0.004608	212.0546
$Y\cos(\phi)$	−1.9319	0.86251	0.013224	0.012865	210.3287

Table 4
Multivariate Regression Models for the SS Case

Covariates	α	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	R^2_{MF}	R^2_E	AIC
$\{R, s\}$	−0.27006	−0.13331	0.038801								0.22514	0.17175	131.2366
$\{R, \theta\}$	1.2809	−0.1041	−0.05083								0.25698	0.21624	124.5098
$\{s, \theta\}$	−0.05116	0.0085	−0.05287								0.12816	0.13093	137.4095
$\{R, s, R \cdot s\}$	−0.33844	−0.12695	0.042577	−0.00032							0.22516	0.17189	133.2153
$\{R, \theta, R \cdot \theta\}$	1.79891	−0.14938	−0.08167	0.002579							0.26221	0.22699	124.8835
$\{s, \theta, s \cdot \theta\}$	−0.10312	0.011646	−0.05036	−0.00019							0.12768	0.13104	139.3921
$\{R, s, \theta\}$	0.85925	−0.11137	0.018704	−0.04441							0.2708	0.22511	125.1679
$\{R, s, \theta, R \cdot s\}$	0.923176	−0.11676	0.015526	−0.04465	0.000261						0.27147	0.22521	127.1535
$\{R, s, \theta, R \cdot \theta\}$	1.38158	−0.15484	0.017515	−0.07415	0.002482						0.27482	0.23457	125.7386
$\{R, s, \theta, s \cdot \theta\}$	1.02575	−0.11315	0.008837	−0.05156	0.000584						0.27505	0.22602	127.0316
$\{R, s, \theta, R \cdot \theta, s \cdot \theta\}$	1.49624	−0.15572	0.01026	−0.07907	0.002473	0.000418					0.27718	0.23504	127.6677
$\{R, s, \theta, R \cdot s, R \cdot \theta\}$	1.76409	−0.18723	0.002134	−0.07965	0.001225	0.00289					0.27721	0.23649	127.4479
$\{R, s, \theta, s \cdot \theta, R \cdot s\}$	1.04591	−0.11518	0.007895	−0.05146	0.000568	9.99×10^{-05}					0.27526	0.22603	129.0296
$\{R, s, \theta, R \cdot s, R \cdot \theta, s \cdot \theta\}$	1.79208	−0.18554	−0.00016	−0.08157	0.001143	0.002859	0.000193				0.27836	0.23658	129.4335
$\{R, s, \theta, R \cdot s, R \cdot \theta, s \cdot \theta, R^2, s^2, \theta^2\}$	1.319972	−0.38593	0.045024	0.008349	0.001275	0.003687	−0.00104	0.007086	−0.00037	−0.00147	0.30062	0.26244	131.5234

larger (34) than the number of terms in the model, there is significant risk of data overfitting. Therefore, based on the results of the previous section, the variables showing the lowest marginal predictive power, $\{X, M\}$, have been excluded from candidates for the multivariate regressions for pulse occurrence. This does not imply that magnitude or fault length do not affect hazard analysis in the near source, but rather, that they are more related to the amplitude of the directivity effects once the geometry has been determined.

For the $\{R, s, \theta\}$ set of covariates, several multiple regression models have been fitted. They are linear and quadratic. Results in terms of coefficients are given in Table 4. Among those computed, the linear combination of the covariates is reported here, equation (8), as it is the best performing model that includes all three covariates known given the rupture and the site. There is no support from the analyses for the inclusion of interaction or squared terms. In fact, while a two-variable model (with only R and θ) has a slightly lower (better) AIC, we prefer a model that makes use of the maximum level of information about the source-to-site configuration,

$$P[\text{pulse}|R, s, \theta] = \frac{e^{\alpha+\beta_1 R+\beta_2 s+\beta_3 \theta}}{1 + e^{\alpha+\beta_1 R+\beta_2 s+\beta_3 \theta}}. \qquad (8)$$

Because, for the rupture's schematic geometry, θ is known given R and s, it is possible to represent this model in a three-dimensional plot, which is given in Figure 8 up to the bounds of the covariates determined by data availability. Assuming the epicenter of the event at the origin and the rupture direction being coincident with the s axis, any point in the $\{R, s\}$ plane may be considered as a site for which the θ angle is also known ($\theta = \arctan[R/s]$). Therefore, for that site, the pulse occurrence probability predicted according to the proposed model may be read on the vertical axis.

From the plot, it is possible to observe the expected marginal trends of pulse occurrence with respect to the three covariates. The probability generally decreases with R and increases with s. Because the univariate regression suggested a low probability for large θ, it should be reflected in the left corner of the model, as it actually is.

Non-Strike-Slip. For the NNS case, an analogous set of covariates, $\{R, d, \phi\}$, and analogous functional forms of the SS case have been investigated and results are reported in Table 5. The corresponding model of equation (8) is given in Figure 9 for the NSS case. Again, the shape of the probability surface is similar between SS and NSS. In comparing Figure 9 to Figure 8, it has to be recalled that the former refers to a restricted covariates' domain with respect to the latter because of the applicability boundaries of NSS data. The NSS multivariate model, as also observed for the univariate regressions, generally predicts lower conditional probabilities with respect to the SS case.

It may be observed from the results reported in Table 5 that this model is not the best performing in terms of global

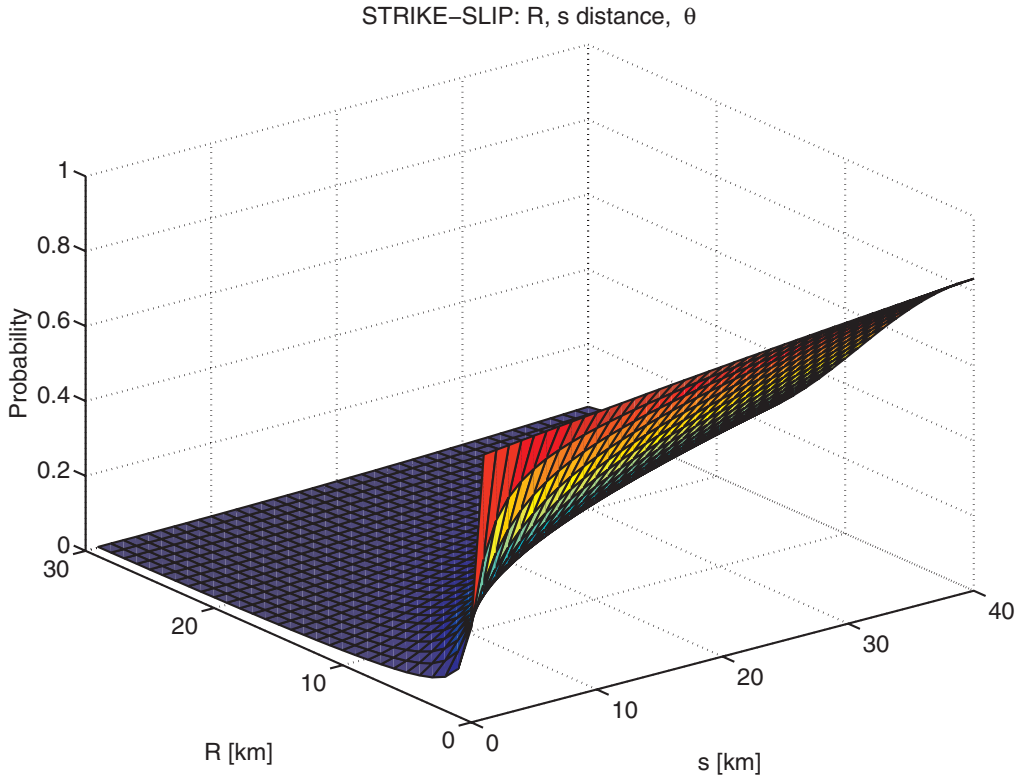

Figure 8. Selected multivariate logistic regression model for the SS case.

Table 5

Multivariate Regression Models for the NSS Case

Covariates	α	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	R^2_{MF}	R^2_E	AIC
$\{R, d\}$	−0.4355	−0.07339	0.012651								0.064345	0.052306	204.0847
$\{R, \phi\}$	0.098269	−0.04864	−0.02397								0.10374	0.099313	194.2595
$\{d, \phi\}$	−0.41364	−0.01199	−0.03024								0.085203	0.08153	197.9765
$\{R, d, R \cdot d\}$	−3.29478	0.074365	0.255602	−0.01477							0.15439	0.12236	191.4417
$\{R, \phi, R \cdot \phi\}$	0.29809	−0.06075	−0.03401	0.000545							0.10185	0.10033	196.0464
$\{d, \phi, d \cdot \phi\}$	−0.41263	−0.0123	−0.03035	2.45×10^{-05}							0.085204	0.081532	199.9761
$\{R, d, \phi\}$	0.55278	−0.0551	−0.02669	−0.0271							0.10726	0.10404	195.2706
$\{R, d, \phi, R \cdot d\}$	−2.12929	0.069785	0.191008	−0.02149	−0.01278						0.18724	0.15551	186.5141
$\{R, d, \phi, R \cdot \phi\}$	0.894841	−0.07305	−0.03048	−0.04132	0.000754						0.10377	0.10594	196.8751
$\{R, d, \phi, d \cdot \phi\}$	0.623423	−0.05682	−0.03502	−0.02909	0.000465						0.10712	0.10456	197.1626
$\{R, d, \phi, R \cdot \phi, d \cdot \phi\}$	1.13514	−0.08163	−0.04732	−0.04882	0.000962	0.000815					0.10247	0.10726	198.599
$\{R, d, \phi, R \cdot d, R \cdot \phi\}$	−3.29865	0.124788	0.238805	0.005782	−0.01493	−0.00131					0.19828	0.1601	187.5541
$\{R, d, \phi, d \cdot \phi, R \cdot d\}$	−2.16045	0.082317	0.182319	−0.03072	0.002653	−0.0159					0.20045	0.16652	186.2113
$\{R, d, \phi, R \cdot d, R \cdot \phi, d \cdot \phi\}$	−2.68782	0.105662	0.20442	−0.01665	−0.01641	−0.00061	0.002301				0.20266	0.16735	188.0384
$\{R, d, \phi, R \cdot d, R \cdot \phi, d \cdot \phi, R^2, d^2, \phi^2\}$	−0.84201	0.13504	0.004417	−0.13063	−0.01686	−0.00011	0.007143	−0.00201	0.00556	0.001022	0.23177	0.20505	186.1593

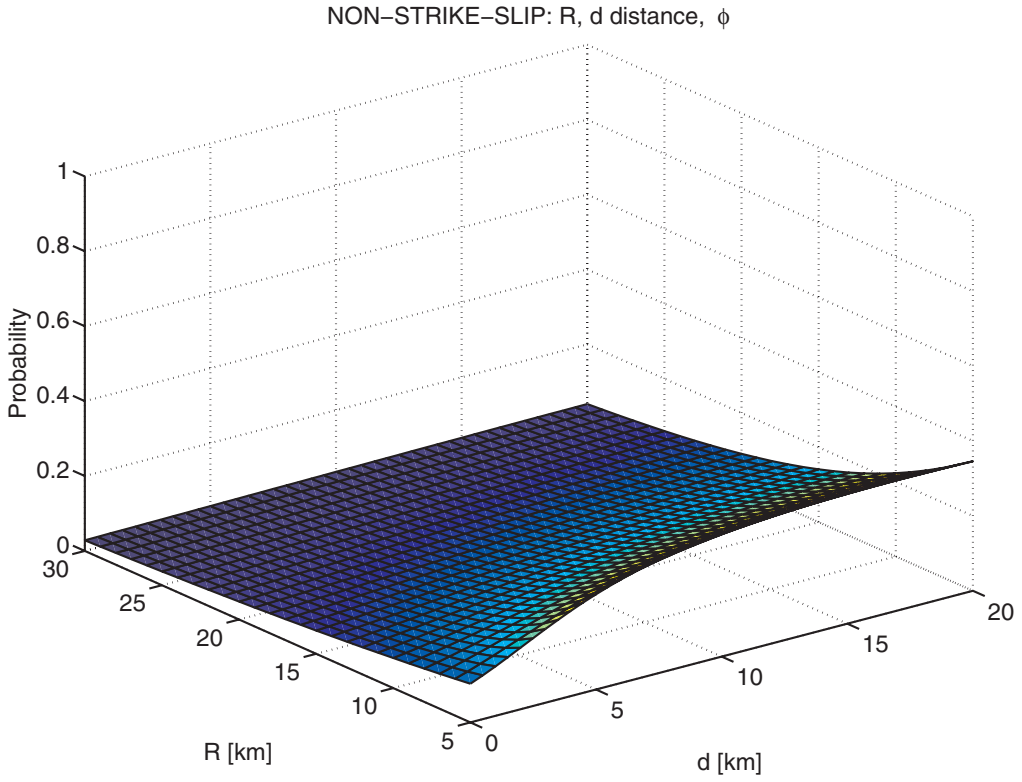

Figure 9. Selected multivariate logistic regression model for the NSS case.

scores. In fact, other functional forms scoring better AICs can lead to significantly different conditional probabilities, which may approach 1 for some values of the input parameters, although this behavior seems to be not justified by the input data. In other words, the multivariate models for the NSS case lack robustness if compared to the SS case. This is mainly due to the aforementioned heterogeneity of the covariate data. In fact, the NSS dataset includes several fault mechanisms while the set of predictors has been calibrated for DS ruptures. Therefore, the presented model has been chosen for consistency with the SS case, as the same points supporting its preference in the SS case still hold.

Discussion and Conclusions

Near-source issues in earthquake engineering are of concern for nonlinear assessment of structures. Because of the peculiar spectral features that ground motion may experience, the PSHA at the site requires appropriate procedures. Then, record selection cannot follow the current far-field practice and should reflect the near-source pulse and nonpulse hazards. Both of these issues call for a pulse occurrence probabilistic model. The study presented attempted to build and propose such models empirically. The fundamentals of the analyses are related to the choice of the covariates (i.e., independent variables) and the determination of the response sample (i.e., the dataset).

Because PSHA refers to a specific site, the occurrence of pulses should be conditional on some parameters, available for the source-to-site configuration, which are believed to predict directivity effects. To this aim, covariates were chosen among factors identifying near-source conditions and, according to seismologists, affecting the amplitude of pulses specifically for SS and DS events.

As directivity effects are generally observed most strongly in the velocity signals recorded in the direction orthogonal to the strike, the empirical dataset was made up of fault-normal rotated records. All of those records within 30 km in terms of closest distance to fault rupture (arbitrarily considered as a practical upper bound for near-source conditions) reported by the Next Generation Attenuation database were used (except the Chi-Chi related records). Pulse-like velocity ground motions have been identified by the rational method, based on wavelets, proposed by Baker (2007). Some judgment was used to identify (and classify as nonpulselike) those velocity recordings showing multiple low-frequency cycles that are likely not related to directivity. Finally, consistent with the seismological parameterization of directivity, the obtained dataset was split into two parts featuring SS and NSS records, in compliance with the different physics of directivity in the two cases. The dataset was purged of those records for which the information regarding the covariates was not available. This alone allows one to evaluate the marginal pulse occurrence frequencies,

which are not larger than 26%, a number found for the SS sample.

Simple and multiple logistic regression models have been investigated to associate pulse occurrence in the dataset to the covariates, for both SS and NSS samples. General findings hold for SS and NSS. Pulse occurrence probability has shown, as expected, significant dependence on distance to the rupture, R, along the rupture, s, and also on the θ angle (which, in principle, is a deterministic nonlinear function of the other two parameters). Less explanatory power, if at all, for pulselike records occurrence, was found for the event's magnitude and other pulse-amplitude-related factors. Although these results may sound intuitively unexpected, it has to be recalled here that this study dealt with pulse occurrence probability alone, rather than on the prediction of the amplitude of such pulses, for which the excluded parameters do play a role.

The strength of the association between the response variables and the covariates, in simple logistic models, has been found to be systematically weaker in the NSS case than in the SS. This may be because, for NSS, the set of physical conditions to determine a pulselike record seems more difficult to realize, as suggested by the marginal pulse frequency, which is 17%. Moreover, because the NSS dataset by definition includes different mechanisms, while the chosen predictors refer specifically to DS events, some predictors may show comparatively less explanatory power.

Multivariate logistic regression models were also investigated. To avoid data overfitting, only the covariates shown to be the best predictors in the simple univariate models were considered in the multiple regressions, which have been investigated up to complete quadratic functional forms. For the SS case, the proposed model is the linear combination of the geometrical predictors, as there is no empirical support to use models that include interaction or quadratic terms. In the NSS case, the discussed intrinsic features of the sample weakened the robustness of the regressions. Therefore, the proposed NSS multivariate model has been arbitrarily selected by analogy with the SS case.

Finally, it is worth mentioning that because it is a general opinion that strong directivity is associated with a strong asperity and with a smooth fault plane, there may be a case for a common event term in the regression models. On the other hand, one can argue that in the same event, the geometry varies from site to site so the event term might be small. Therefore, selected univariate (depending on R) and multivariate (depending on the linear combination of the five basic covariates) logistic regression models, for both SS and NSS cases, have been tested for random effect (RE). The RE considered is the event, that is, the specific earthquake for each sample. The XTLOGIT routine of STATACORP—STATA™ software (version 8.0) software was used to test the logistic models. Such analyses for those models investigated led to the conclusion that the proportion of the total variance contributed by the event variance component is small and, at least assuming an 0.05 significance level, the null hypothesis that the random effect is negligible cannot be rejected. Therefore, based on these results, there is no compelling need to use logistic regression models that include specific event terms.

Data and Resources

The records' pulse scores used in the analyses herein have been kindly provided by J. W. Baker. Other information about the considered ground motions is obtained via the Next Generation Attenuation of Ground Motions Project (http://peer.berkeley.edu/products/nga_project.html) and, in particular, from its online published flat file (http://peer.berkeley.edu/assets/NGA_Flatfile.xls) and documentation (http://peer.berkeley.edu/nga/NGA_Documentation.xls), both of them last accessed in August 2006.

Acknowledgments

The authors would like to thank Jack W. Baker of Stanford University and Polsak Tothong of AIR Worldwide Corporation. I. Iervolino would also like to thank Massimiliano Giorgio of the Second University of Naples and Paolo Viarengo of the University of Naples Federico II. C. Allin Cornell was supported in part by the Earthquake Engineering Research Centers Program of the National Science Foundation under Award Number EEC-9701568 through the Pacific Earthquake Engineering Research Center (PEER). The support from the Department of Structural Engineering of the University of Naples Federico II to I. Iervolino is also gratefully acknowledged.

This work started in 2005, and at the end of October 2007, we submitted the manuscript to *BSSA*. A few weeks later, on 14 December 2007, C. Allin Cornell passed away. I. Iervolino wants to recall here the unique personal and scientific education he received from such a great person.

References

Agresti, A. (2002). *Categorical Data Analysis*, Second Ed., Wiley and Sons, New York.

Baker, J. W. (2007). Quantitative classification of near-fault ground motions using wavelet analysis, *Bull. Seismol. Soc. Am.* **97,** 1486–1501.

Bolt, B. A., and N. A. Abrahamson (2003). Estimation of strong seismic ground motions, in *International Handbook of Earthquake and Engineering Seismology*, Part B, W. H. K. Lee, H. Kanamori, P. C. Jennings and C. Kisslinger (Editors), Academic Press, New York, 983–1001.

Cornell, C. A. (2004). Hazard, ground motions and probabilistic assessment for PBSD, in *Performance Based Seismic Design Concepts and Implementation*, PEER Report 2004/05, Pacific Earthquake Engineering Research Center, Berkeley, California, 39–52.

Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation, *J. Am. Stat. Assoc.* **73,** 113–121.

Howard, K., C. A. Tracy, and R. G. Burns (2005). Comparing observed and predicted directivity in near-source ground motion, *Earthq. Spectra* **21,** 1063–1092.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in *Frontiers in Economics*, P. Zarembka (Editor), Academic Press, New York, 105–142.

Somerville, P. G. (2003). Magnitude scaling of the near fault rupture directivity pulse, *Phys. Earth Planet. Interiors* **137,** 201–212.

Somerville, P. G., N. F. Smith, R. W. Graves, and N. A. Abrahamson (1997). Modification of empirical strong ground motion attenuation relations to include the amplitude and duration effects of rupture directivity, *Seism. Res. Lett.* **68,** 199–222.

Spudich, P., B. S. J. Chiou, R. Graves, N. Collins, and P. G. Somerville (2004). A formulation of directivity for earthquake sources using isochrone theory, *U.S. Geol. Surv. Open-File Rept. 2004-1268*.

Tothong, P., and C. A. Cornell (2006). Probabilistic seismic demand analysis using advanced ground motion intensity measures, attenuation relationships, and near-fault effects, PEER Report 2006/11, Pacific Earthquake Engineering Research Center, Berkeley, California.

Tothong, P., and N. Luco (2007). Probabilistic seismic demand analysis using advanced ground motion intensity measures, *Earthq. Eng. Struct. Dyn.* **36,** 1837–1860.

Tothong, P., C. A. Cornell, and J. W. Baker (2007). Explicit directivity-pulse inclusion in probabilistic seismic hazard analysis, *Earthq. Spectra* **23,** 867–891.

Dipartimento di Ingegneria Strutturale
Università degli Studi di Napoli Federico II
Via Claudio 21, 80125
Naples, Italy
iunio.iervolino@unina.it
(I.I.)

Department of Civil and Environmental Engineering
Stanford University
Stanford, California 94305
(C.A.C.)

Manuscript received 1 November 2007