Bulletin of the Seismological Society of America, Vol. 94, No. 6, pp. 2164–2185, December 2004

On the Use of Response Spectral-Reference Data for the Selection and Ranking of Ground-Motion Models for Seismic-Hazard Analysis in Regions of Moderate Seismicity: The Case of Rock Motion

by Frank Scherbaum, Fabrice Cotton, and Patrick Smit

Abstract The use of ground-motion-prediction equations to estimate ground shaking has become a very popular approach for seismic-hazard assessment, especially in the framework of a logic-tree approach. Owing to the large number of existing published ground-motion models, however, the selection and ranking of appropriate models for a particular target area often pose serious practical problems. Here we show how observed ground-motion records can help to guide this process in a systematic and comprehensible way. A key element in this context is a new, likelihood based, goodness-of-fit measure that has the property not only to quantify the model fit but also to measure in some degree how well the underlying statistical model assumptions are met. By design, this measure naturally scales between 0 and 1, with a value of 0.5 for a situation in which the model perfectly matches the sample distribution both in terms of mean and standard deviation. We have used it in combination with other goodness-of-fit measures to derive a simple classification scheme to quantify how well a candidate ground-motion-prediction equation models a particular set of observed-response spectra. This scheme is demonstrated to perform well in recognizing a number of popular ground-motion models from their rock-site-recording subsets. This indicates its potential for aiding the assignment of logic-tree weights in a consistent and reproducible way. We have applied our scheme to the border region of France, Germany, and Switzerland where the M_w 4.8 St. Dié earthquake of 22 February 2003 in eastern France recently provided a small set of observed-response spectra. These records are best modeled by the ground-motion-prediction equation of Berge-Thierry *et al.* (2003), which is based on the analysis of predominantly European data. The fact that the Swiss model of Bay *et al.* (2003) is not able to model the observed records in an acceptable way may indicate general problems arising from the use of weak-motion data for strong-motion prediction.

Introduction

The issue adressed in this paper is the selection and ranking of appropriate ground-motion models for seismic-hazard assessment in an arbitrary target region to obtain transparent, data-driven criteria for the assignment of logic-tree weights. Because of the improvement of seismological networks, the number of proposed ground-motion models has strongly increased in the last decade. A recent review (Douglas, 2003) summarizes more than 120 studies that have derived equations for the estimation of peak ground acceleration and 80 studies that derived equations for the estimation of response spectral ordinates. Therefore, there is a need for quick and efficient testing to determine if a model is appropriate for a particular target region. Unless the prerequisites for appropriateness are defined very carefully, and the reasons for the selection process are fully documented step by step, the selection of candidate models, and especially the assignments of logic-tree weights, easily become a highly subjective and non-transparent process. Possible selection criteria, such as tectonic environment, stress regime, and/or propagation properties in the target region, are often hard to quantify, and there is no common understanding about the relative importance of individual criteria. As a consequence, the influence of personal preferences for particular regression schemes, for particular sets of independent parameters, and/or for the degree of simplicity or sophistication of a model are difficult to avoid. In the context of a logic tree with more than a few alternative ground-motion-model branches, it will therefore be hard if not impossible to keep the judgment of the complete set of candidate models internally consistent and the verdict on particular models repro-

ducible. There is an additional related problem, which is easy to be overlooked. The definition of control parameters in ground-motion models such as magnitude and distance definitions are usually different between different models, which implies that users will correct the proposed models with their own distance metrics or magnitude definitions. There is therefore a need to judge not only the original models but also the "corrected" ones in a consistent way. A more detailed discussion of these issues is given in Bommer *et al.* (2004b).

Here we propose a transparent and reproducible mechanism to guide the selection and ranking process of ground-motion models that minimizes the risk of inconsistent judgment by using recorded-response spectra to judge the match or mismatch of ground-motion models to a particular target region. Commonly, the main purpose of ground-motion-prediction equations in the context of seismic-hazard assessment (SHA) is to estimate the properties of earthquake records that have not yet been observed in the target region. On the other hand, a good model should naturally also describe the properties of records that have been observed already. On the basis of this notion, our approach to model selection/ranking consists of the evaluation of how well a particular ("corrected") ground-motion model "predicts" observed reference data in a quantitative way. In addition to the regional aspect, there is also a need for evaluation methods if the ground-motion from strong earthquakes can be predicted well by models derived from weak-motion data, (e.g., that of Raoof *et al.*, 1997; Malagnini *et al.*, 1999; Bay *et al.*, 2003). At first glance, these models are attractive because they can be developed by using records of the target region even if this region is characterized by low to moderate seismicity.

This paper will focus on rock ground-motion, which is often used as reference motion in seismic-hazard projects. However, the geotechnical or geophysical characterization of the so-called rock-site stations is usually rather poor, and a geologically defined rock site can be affected by weathering (Steidl *et al.*, 1996; Boore and Joyner, 1997). In addition, most accelerometric station sites are located on sediments. Hence, the number of rock-site records in datasets that have been used to generate ground-motion models is usually rather small. For rock conditions, it is therefore especially important to have a tool to quantitatively evaluate whether the average site conditions for the host region of the ground-motion model seem to be similar to the conditions for the target region.

Following a brief description of candidate goodness-of-fit measures, we evaluate their performance on synthetic data. Subsequently, they are applied to rock-site subsets of the generating datasets of several ground-motion-prediction equations (Ambraseys and Simpson, 1996; Ambraseys *et al.*, 1996; Sabetta and Pugliese, 1996; Abrahamson and Silva, 1997; Ambraseys and Douglas, 2003; Berge-Thierry *et al.*, 2003) to check their success in identifying the corresponding ground-motion models. Finally, using the records of the M_w 4.8 St. Dié earthquake of 22 February 2003, we will show that for the border region of eastern France, southwestern Germany, and northern Switzerland, a few observed-response spectra of high quality can already provide considerable constraints on the selection of ground-motion models for seismic-hazard analysis.

Goodness-of-Fit Measures

Given a set of observed ground-motion data for the target region and a candidate ground-motion model, there are several approaches for evaluating whether the observed data confirm the model prediction. We find it convenient to group these approaches into two different classes, depending on strategy. The first group deals with test statistics for the properties of residual distributions, whereas the second group deals with tests for sets of individual data-value predictions.

Testing Residual Distribution

In order to simplify this test, we first normalize the differences between data and model predictions (residuals) in the following way. Because ground-motion models are commonly expressed as logarithmic quantities, we subtract the logarithmic-model predictions from the logarithms of the data values and subsequently divide the results by the corresponding standard deviations of the logarithmic model. Ideally, this should result in residuals that are normally distributed with zero mean and unit variance.

Central Tendency and Variance Tests. The question addressed with this type of test is whether the target-region dataset (data distribution) and the ground-motion model (model distribution) have the same central tendency (mean or median) and/or variance. As a most direct approach, we calculate the *mean, median, and standard deviation of the normalized data residuals.* The deviation of the mean and the median from zero, and the deviation of the standard deviation from 1, help to detect weak models. Large differences between mean and median should help to identify models for which the residual distribution is skewed. Because we have scaled the data by the model mean and variance, we additionally evaluate the hypothesis that the mean of the normalized distribution is zero. Since the variance is assumed to be known, this test is based on the normal distribution instead of the Student's distribution, which is usually used for unknown variances (Wolfram, 1996). The test quantity that we evaluate is the *mean test* p-*value*, which is the probability of the estimated mean being as large as it is just by chance, given that the hypothesized population parameters are true. A small numerical *p*-value (0.05 or 0.01) means that the observed difference between the estimated mean and the model mean is "very significant" and thus it is very unlikely that the observations have been produced by the candidate model, whereas a large *p*-value enhances our confidence in the model (Press *et al.*, 2001). In a similar way, we have calculated the *p*-value to test the normalized

residuals for unit variance. Again, this *variance test* p-*value* is a number between 0 and 1, with a small numerical value implying a strong rejection of the hypothesis (Wolfram, 1996).

Testing the Shape of the Residual Distribution. In order to test if the data distribution differs from the model distribution not only in mean and/or variance but also in shape, we generalize the question asked in the previous paragraph. Two different approaches were used to test the shape of the normalized residual distribution. As one approach, we have employed the *chi-square test* (Press *et al.*, 2001). The idea of this test is to compare the frequencies of the normalized residuals with the frequencies that would be expected, in our case assuming a normal distribution of unit variance. The significance is given by the probability that the observed chi-square value will be exceeded by chance, given that the model is correct. This computed probability gives a quantitative measure of the goodness-of-fit of the model. If it is very small for some particular dataset, then the apparent discrepancies are unlikely to be chance fluctuations. As a second test in this category, we have evaluated the *Kolmogorov–Smirnov test* to check whether the normalized residual distribution is not significantly different from a zero mean normal distribution with unit variance. Differently from the chi-square test, the Kolmogorov–Smirnov test checks the deviations from the model distribution not only in a general sense but also for the most deviant values of the criterion variable. Owing to the normalization of the data, the test quantity to determine for the Kolmogorov–Smirnov test in our case is the maximum value of the absolute difference between the normalized-data residuals and the cumulative normal distribution for zero mean and unit variance. An advantage of the Kolmogorov–Smirnov test is that it is an exact test that does not require binning. One of its known disadvantages is that it is more sensitive near the centers of the distribution than at the tails (Press *et al.*, 2001).

Testing Sets of Individual Observations

A slightly different look at the goodness-of-fit problem is taken in this group of tests, which compare sets of individual data values with their corresponding model predictions without particular assumptions about the special type of the residual distribution.

Correlation between Individual Model Predictions and Data. For a good model, even if it does not predict individual data values exactly, it should still be expected that the general tendency of the predictions would be similar to the observations. This correlation can be expressed by the *Pearson correlation coefficient* (Wolfram, 1996), which takes values between 1 and -1 for complete correlation and anti-correlation, respectively. A value near zero indicates that the model prediction and the data are uncorrelated. The correlation coefficient is, however, a rather poor statistic for deciding whether an observed correlation is statistically significant and/or whether one observed correlation is stronger than another. The reason is that it is ignorant of the individual distributions, and therefore it is not possible to compute its distribution in the case of uncorrelated data (null hypothesis).

Chi-square Misfit. In the context of model fitting, the chi-square statistic is often used in a slightly different context than in the previous paragraph. A misfit value, which if minimized results in a maximum likelihood estimate of the model often also referred to as chi-square, is defined as (Press *et al.*, 2001)

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{d_i - m_i}{\sigma_i} \right)^2. \quad (1)$$

Here d_i and m_i are the elements of the data and the model-predictions vector, respectively, and σ_i is the standard deviation for sample i. For a moderately good fit, an χ^2 value on the order of the number of degrees of freedom is typical (Press *et al*., 2001). Since in the present context we have to compare different sample sizes, for practical applications we consider equation (1), normalized by N.

Variance Reduction. In contrast to the chi-square misfit above, which normalizes the differences between model and data vector by the standard deviation for the sample, the variance reduction often used in seismogram modeling (Cohee and Beroza, 1994; Cotton and Campillo, 1995) is a goodness-of-fit value where the residuals are additionally normalized by data amplitudes:

$$\Delta\sigma^2 = \left(1 - \frac{[\vec{d} - \vec{g}(\vec{m})]^T \cdot \mathbf{C}_d^{-1} \cdot [\vec{d} - \vec{g}\vec{m})]}{\vec{d}^T \cdot \mathbf{C}_d^{-1} \cdot \vec{d}} \right) \cdot 100\%, \quad (2)$$

where $\mathbf{C}_d^{-1}$ represents the weighting by the inverse of the data covariance matrix, $\vec{d}$ is the data vector, and $\vec{g}(\vec{m})$ is the model prediction for the model $\vec{m}$. For cases in which the covariance matrix is a constant times the identity matrix (constant standard deviation for all samples), equation (2) reduces to (in the notation of equation 1)

$$\Delta\sigma^2 = \left(1 - \sum_{i=1}^{N} \left(\frac{d_i - m_i}{d_i} \right)^2 \right) \cdot 100\% \text{ for } d_i > 0. \quad (3)$$

In other words, all residuals (whatever their absolute amplitude and therefore whatever the distance and magnitude scenarios) are weighted equally. Here again, for practical applications we consider the sum in equation (3) normalized by N.

A New Goodness-of-Fit Measure In yet another approach, we define a goodness-of-fit measure especially suited for the

present context, based on the concept of likelihood (Edwards, 1992). As before, we assume that each ground-motion model can be described by a lognormal distribution—in other words, a normal distribution for $\ln(Y)$. If the model is perfect, any observation can be treated as a random sample drawn from this distribution.

If for a given magnitude M, distance R, and frequency f, $\mu(M, R, f)$ is the predicted mean value for $\ln(Y)$, the probability of a single observation $x = \ln(Y_x)$ to fall into the interval $(x, x + dx)$ is

$$dF = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right) \cdot dx. \quad (4)$$

Here, $\sigma(M, R, f)$ is the standard error of the ground-motion model, which in general depends on magnitude, distance, and frequency. If we standardize the sample by model mean μ and standard deviation σ, we obtain the normalized residual $z = \dfrac{x - \mu}{\sigma}$ for which the probability density function would be

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-z^2}{2}\right). \quad (5)$$

Because of its convenient scaling, outlined in more detail as follows, a good measure for the goodness-of-fit of a ground-motion model is the probability for the absolute value of a random sample from the normalized distribution to fall into the interval between the modulus of a particular observation $|z_0|$ (expressed as normalized variable) and ∞. For a positive z_0, this is

$$\begin{aligned} u(z_0) &= \frac{1}{\sqrt{2\pi}} \int_{z_0}^{\infty} \exp\left(\frac{-z^2}{2}\right) \cdot dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \exp\left(\frac{-z^2}{2}\right) \cdot dz \\ &\quad - \frac{1}{\sqrt{2\pi}} \int_{0}^{z_0} \exp\left(\frac{-z^2}{2}\right) \cdot dz \\ &= \frac{1}{2}\left(Erf(\infty) - Erf\left(\frac{z_0}{\sqrt{2}}\right)\right). \end{aligned} \quad (6)$$

Here $Erf(z)$ is the error function $\dfrac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) \cdot dt$. Using the generalized form

$$Erf(z_0, z_1) \equiv \frac{2}{\sqrt{\pi}} \cdot \int_{z_0}^{z_1} e^{-t^2} \cdot dt = Erf(z_1) - Erf(z_0)$$

(e.g., Wolfram, 1996). (7)

$u(z)_0$ can be expressed as

$$u(z_0) = \frac{1}{2} Erf\left(\frac{z_0}{\sqrt{2}}, \infty\right). \quad (8)$$

$u(z)_0$ is the likelihood of an observation (in this case, the normalized residual) to be equal to or larger than z_0. Considering both tails of the distribution, we define

$$LH(|z_0|) \equiv 2 \cdot u(|z_0|) = Erf\left(\frac{|z_0|}{\sqrt{2}}, \infty\right) \quad (9)$$

and refer to it simply as the *LH* value. In the context of quantifying goodness-of-fit, *LH* values have some interesting properties:

- *LH* reaches its maximum value of 1 for $|z_0| = 0$, in other words, for an observation that coincides with the mean value of the model.
- The *LH* value decreases with increasing distance from the mean (decreasing quality of the fit). For $|z_0| = \infty$ we obtain $LH = 0$.
- If the model assumptions are matched exactly, in other words, for samples drawn from a normal distribution with unit variance, the samples of the random variable *LH* are evenly distributed between 0 and 1. The proof for this is given in the Appendix. This allows the goodness-of-fit to be conveniently quantified from easily determined properties of the distribution of *LH* values.

These properties are further illustrated in Figure 1. We are using the median to quantify the properties of the distribution of *LH* values in a single number, mainly because of its stability regarding outliers (which will be more our concern with real data than with the simulated data here). Other parameters, such as moments of the distribution, are also conceivable, but their analysis was beyond the scope of this article.

Figure 1a shows a case in which a synthetic residual model matches the data exactly in terms of both mean and variance. Here, *LH* is evenly distributed between 0 and 1, and the median of *LH* is about 0.5. In case the model is unbiased in terms of the mean, but the sample variance is smaller than the model variance as in Figure 1b, the distribution of *LH* becomes asymmetric, and the median of *LH* increases to a value above 0.5. In case the sample variance becomes larger than the model variance—still with an unbiased model as in Figure 1c—the frequency of low *LH* values increases and the median of the *LH* distribution decreases below 0.5. The decrease of the median of the *LH* distribution will be especially strong for the simultaneous increase in sample variance and a shift of the mean value (Fig. 1d–f). Because of these properties, the distribution of *LH* values seems to be a good indicator for (1) the goodness-of-fit of ground-motion models to observed response spectral values as well as for (2) how well the model assumptions are met.

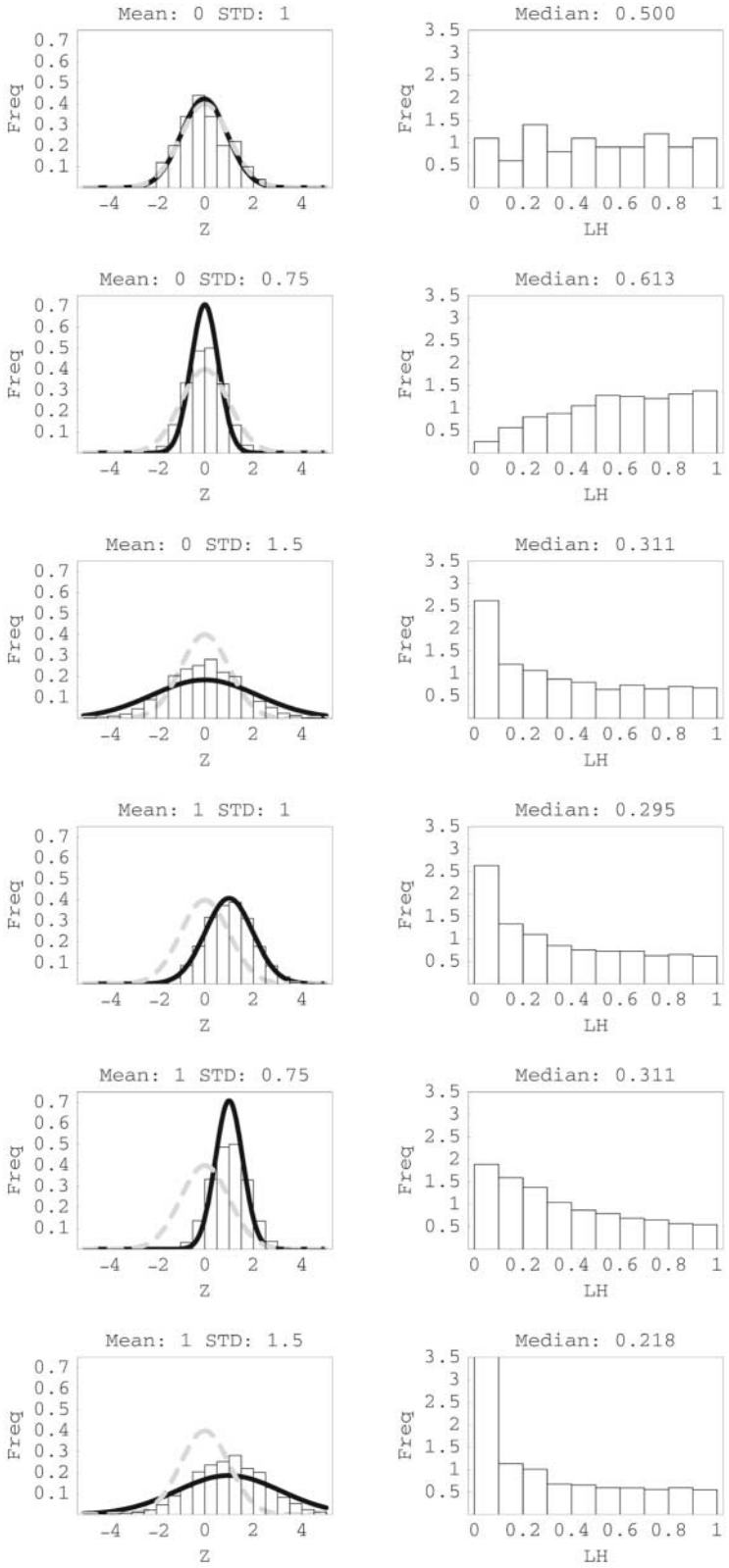

Figure 1. Distribution of residuals (left panels) and corresponding *LH* values (right panels) for different simulated distributions. Mean values and standard deviations for the residual distributions are indicated on tops of the left panels. The two distribution functions in the left panel indicate the unit variance normal distribution and the actual residual distribution, respectively. On top of the right panels the median values of the resulting *LH*-value distributions are displayed.

Furthermore, the median of the *LH* distribution seems to conveniently describe both of these by a single number.

Application to Generating Datasets of Existing Ground-Motion Models

In order to test the individual goodness-of-fit measures in practice, we tested their performance on the rock site subsets of the generating datasets of several ground-motion prediction equations. To match the properties of a particular data set, distance metrics, magnitudes, and components had to be converted for some of the ground-motion models. The distance conversion was performed following the relationships of Scherbaum *et al*. (2004). Because all the ground-motion models that require a conversion from moment magnitude to surface wave magnitudes are "European," we used the Ambraseys and Free (1997) relation (without depth dependence). For conversion to Japanese Meteorological Agency (JMA) magnitude, which is used by Lussou *et al*. (2001), we assumed a one-to-one relationship to M_w as suggested by Heaton *et al*. (1986). The same was done for the local magnitudes of Sabetta and Pugliese (1996), which, according to F. Sabetta (personal comm., 2002), does not require any conversion. The component conversions were based on the empirical relationships determined by Proseis (Proseis, 2002).

What Is a Practical Measure for Goodness-of-Fit?

In order to understand and compare the performance of the different goodness-of-fit measures in a fairly realistic but still well controlled situation, we made our first test on a set of purely synthetic-spectra. For this purpose, we calculated synthetic response spectra for the ground-motion model of Abrahamson and Silva (1997), using the same magnitudes and distances as those of the real rock site subset of the generating dataset. For each frequency sample, the log of spectral values was subsequently perturbed by a random value drawn from a normal distribution with zero mean and the variance given by the Abrahamson and Silva (1997) model. This dataset was compared with a variety of different candidate models (Ambraseys and Simpson, 1996; Ambraseys *et al*., 1996; Sabetta and Pugliese, 1996; Abrahamson and Silva, 1997; Atkinson and Boore, 1997; Boore *et al*., 1997; Toro *et al*., 1997; Spudich *et al*., 1999; Lussou *et al*., 2001; Somerville *et al*., 2001; Ambraseys and Douglas, 2003; Bay *et al*., 2003; Berge-Thierry *et al*., 2003; Abrahamson and Silva, 1997). Their basic characteristics in terms of magnitude, distance, frequency coverage, and assumed site conditions are given in Table 1. Only those records that fully fall into the validity range of a model in terms of magnitude, distance, and frequency coverage have been used for the comparison.

The resulting goodness-of-fit measures are displayed in

Table 1
Ground-Motion Models Used in This Study and Their Assumed Validity Ranges

Model Name	*M* type	*M* Range	Frequency [Hz]	R* Type	R* Range	Site Condition	#Rock (#Total)
Abrahamson and Silva (1997)	M_w	4.4–7.4	0.2–100	RUP‡	3–150	Class 0	80 (655)
Ambraseys and Douglas (2003)	M_S	4.8–7.8	0.2–25	JB§	≤15	Rock	—
Ambraseys and Simpson (1996)	M_S	4.0–7.9	0.5–10	EPI‖ ($MS^{\#} \leq 6$)/ JB§ ($MS^{\#} > 6$)	≤200	Rock	72 (422)
Ambraseys *et al.* (1996)	M_S	4.0–7.9	0.5–10	EPI‖ ($MS^{\#} \leq 6$)/ JB§ ($MS^{\#} > 6$)	≤200	Rock	—
Atkinson and Boore (1997)	M_w	4.0–7.25	0.5–20	HYP**	10–500	Rock	—
Bay *et al.* (2003)	M_w	2.0–6.5	0.5–20	HYP**	10–300		—
Berge-Thierry *et al.* (2003)	$M_S \leq 6$/ $M_w > 6$	4.0–7.3	0.1–34	HYP**	5–100	Rock	122 (965)
Boore *et al.* (1997)	M_w	5.5–7.5	0.5–10	JB§	≤80	620 m/sec	—
Campbell and Bozorgnia (2003a,b)	M_w	4.7–7.7	0.25–20	SEIS‡‡	3–100	Soft rock	—
Lussou *et al.* (2001)	JMA†	3.5–6.3	0.1–50	HYP**	10–200	—	–
Sabetta and Pugliese (1996)	$M_L \leq 5.5$/ $M_S > 5.5$	4.6–6.8	0.25–25	JB§	≤100	Stiff	29 (95)
Somerville *et al.* (2001)	M_w	6.0–7.5	0.25–100	JB§	≤500	Non-rifted	—
Spudich *et al.* (1999)	M_w	5.0–7.0	0.5–10	JB§	≤100	Rock	—
Toro *et al.* (1997)	M_w	5.0–8.0	1–35	JB§	≤500	Mid-continent	—

Last column indicates the number of available rock records for those ground-motion models used to test the proposed selection and ranking scheme. Number in parentheses gives the total number of records used to generate the complete model.

*R = distance

†JMA = Japanese Meteorological Agency magnitude

‡RUP = rupture distance

§JB = Joyner-Boore distance

‖EPI = epicentral distance

**HYP = hypocentral distance

‡‡SEIS = distance to seismogenic part of the rupture

#MS = surface wave magnitude

Table 2. The corresponding distributions of data residuals and *LH* values are shown in Figures 2 and 3, respectively. Because these spectra are lognormally distributed random samples from the Abrahamson and Silva (1997) model, it is not surprising that all goodness-of-fit measures in Table 2 give high scores to the Abrahamson and Silva (1997) model. In terms of ranking the remaining models, however, they provide fairly different answers. The model of Berge-Thierry *et al*. (2003) consistently receives fairly high scores on most of the goodness-of-fit measures (Figs. 2 and 3). This result can be partly explained by the fact that the database used by Berge-Thierry *et al*. (2003) includes 17% of North American records that are common to the Abrahamson and Silva (1997) dataset. None of the statistical-significance tests, however, provide ranking results that are easily comprehensible on the basis of the visual interpretation of the residuals or the *LH* values shown in Figures 2 and 3, respectively. Both the chi-square (CHISQ) and the Kolmogorov–Smirnov (KS) test reject most ground-motion models as completely unacceptable, and even reject the Berge-Thierry *et al*. (2003) model.

In the case of real data, for example, for the subset of rock site records that were used to generate the Abrahamson and Silva (1997) model shown in Table 3 and Figures 4 and 5, the situation becomes even more problematic because it is no longer guaranteed that the data residuals are normally distributed at all. Therefore, measures that test the significance of either shape, mean, or variance deviations will often provide very low probabilities for given observations. This forces experimenters to become very tolerant in practice in accepting statistically weak models (Press *et al*., 2001). This makes it difficult to interpret the corresponding significance values (which will easily cover several orders of magnitude) in the context of selecting and ranking ground-motion models and assigning logic-tree weights. Weights on logic-tree branches naturally scale between 0 and 1, with the lowest and highest chosen weights generally about 0.1 and 0.9 (e.g., Reiter, 1990). A goodness-of-fit measure that scales between 0 and 1, and which naturally provides the same numerical scaling between poor and good models, would therefore greatly facilitate the decision process of a seismic hazard analyst.

A different scaling issue appears for the variance-reduction (VARRED) values for the synthetic as well as for the real data example. Here the values are similar from one model to another, which also does not allow good discrimination among different models. Yet another problem occurs for the Pearson correlation coefficient (PCC). This quantity measures the degree of correlation between model prediction and data, which should be high for a good model. According to Table 3, the highest value for the real-data example is obtained for the Sabetta and Pugliese (1996) model (PCC = 0.752). In contrast, however, from the median and the standard deviation of the corresponding normalized residual distribution of 0.25 and 1.26, respectively, this model would receive a much different rating. We conclude that high correlation coefficients do not sufficiently characterize good models. Therefore, none of the statistical tests for significance (KS, CHISQ, V-Test, and M-Test), or the VARRED or the PCCs seem to provide practically useful constraints on the model choices and so were not considered further.

On the other hand, the distribution of *LH* values natu-

Table 2

Comparison of Different Ground-Motion-Prediction Relations to Model a Synthetic Dataset Produced to Mimic the Generating Dataset of Rock-Site Records for the Abrahamson and Silva (1997) Ground-Motion Model as Described in the Text

Model Name	MEDLH	VARRED	CHISQ MF	PCC	KS	CHISQ	V-Test	M-Test	MEDNR	MEANNR	STDNR	No. of Records
Abrahamson and Silva (1997)	0.494	0.999	1.03	0.725	0.286	0.305	0.643	0.338	0.0814	0.0458	1.01	73
Berge-Thierry *et al.* (2003)	0.464	0.999	1.1	0.709	0	0	0.0302	0	0.473	0.5	0.921	62
Lussou *et al.* (2001)	0.453	0.997	1.22	0.533	0.00166	2.16e-10	0.33	0.0000505	0.36	0.338	1.06	24
Campbell and Bozorgnia (2003)	0.447	0.999	1.49	0.705	0.103	0	2.92e-10	0.251	0.0214	0.0552	1.22	72
Ambraseys and Douglas (2003)	0.38	0.999	1.68	0.685	0.0206	0	1.13e-7	0.15	0.171	0.109	1.3	29
Somerville *et al.* (2001)	0.378	0.998	1.6	0.663	0.00247	0	2.13e-9	0.0926	0.0761	0.101	1.26	46
Sabetta and Pugliese (1996)	0.323	0.999	2.02	0.758	1.7e-8	0	0	3.01e-6	0.391	0.248	1.4	59
Atkinson and Boore (1997)	0.153	0.997	3.82	0.583	0	0	0	1.12e-10	0.281	0.355	1.93	55
Bay *et al.* (2003)	0.016	0.993	7.1	0.675	0	0	0.0000789	0	2.41	2.4	1.17	48

Goodness-of-fit measures used are: median *LH* value (MEDLH), variance reduction (VARRED), chi-square misfit (CHISQMF), Pearson correlation coefficient (PCC), Kolmogorov–Smirnov statistic (KS), chi-square statistic (CHISQ), *p*-value of the variance test (V-Test), *p*-value of the mean test (M-Test), and median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively).

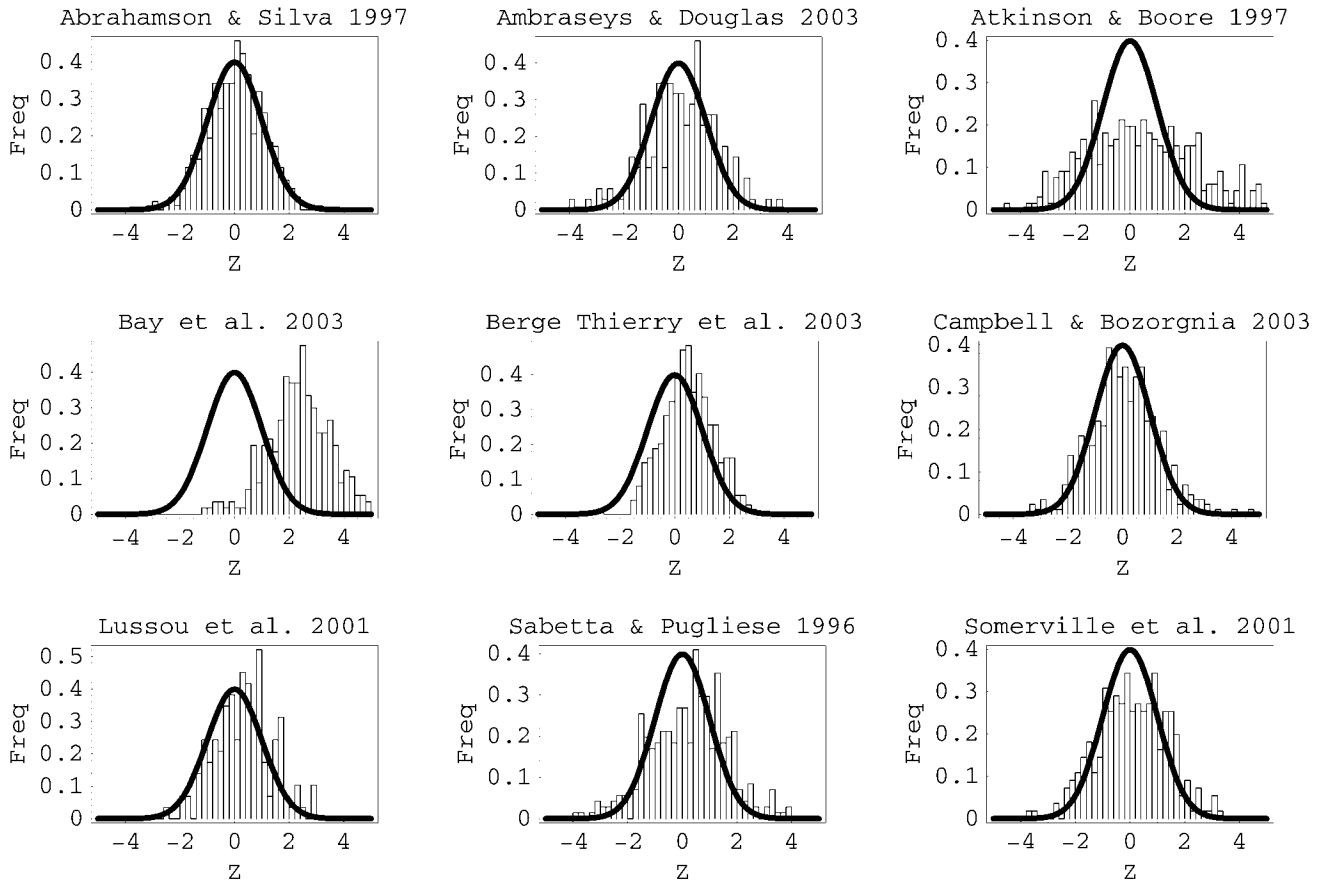

Figure 2. Residual distribution (normalized by model standard deviation) of a synthetic dataset to mimic the rock-site-record subset of the generating dataset for the Abrahamson and Silva (1997) ground-motion model with respect to different ground-motion-prediction equations. Solid line shows the expected distribution function for a standard normal distribution.

rally leads to a transparent and robust ranking scheme that is not affected by the scale problems previously discussed. For the Abrahamson and Silva (1997), Berge-Thierry *et al*., (2003), and Campbell and Bozorgnia (2003) ground-motion models, the shape of the *LH*-value distributions (Fig. 5) visually resemble an even distribution between 0 and 1, as would be expected for good models. For the remaining models, there is an increased frequency of small *LH* values, indicating a large number of poorly predicted data points. The median values of the *LH* distributions for the individual ground-motion-prediction equations to match the Abrahamson and Silva (1997) data set are displayed on top of each panel and are additionally given in the first column of Table 3. From the foregoing discussion, one would expect a median *LH* value close to 0.5 for the model-generating dataset. As can be seen in Table 3 this is actually the case for the Abrahamson and Silva (1997) model and also for the Berge-Thierry *et al*. (2003) model. The fact that the median *LH* value for the Berge-Thierry *et al*. (2003) model is larger than for the Abrahamson and Silva (1997) model probably reflects the fact that the variance of the Berge-Thierry *et al*. (2003) model is also larger than for the Abrahamson and Silva (1997) model. Tables 2 and 3 show that the ranking sequence obtained from the median of the *LH*-value distribution is fairly similar to the one obtained by the normalized chi-square misfit (CHISQMF) value. In contrast to the *LH* value, the absolute values of CHISQMF are not easily interpreted because this would require knowledge of an unknown number of degrees of freedom fo the observed-response spectra. Therefore, this parameter was dropped from further consideration.

Because for practical applications we are interested in robust measures that can be interpreted in terms of absolute numbers, we decided to use, in addition to the *LH* values, the mean, median, and standard deviation of the normalized residual distribution (Mean-NRES, Med-NRES, and Std-NRES) to characterize the goodness of fit of ground-motion models. These parameters make it easy to interpret characterizations of the central tendency as well as those of the spread of the distribution. Table 3 as well as Figure 4, for example, shows that the models of Lussou *et al*. (2001) and Bay *et al*. (2003) are under-predicting the target region da-

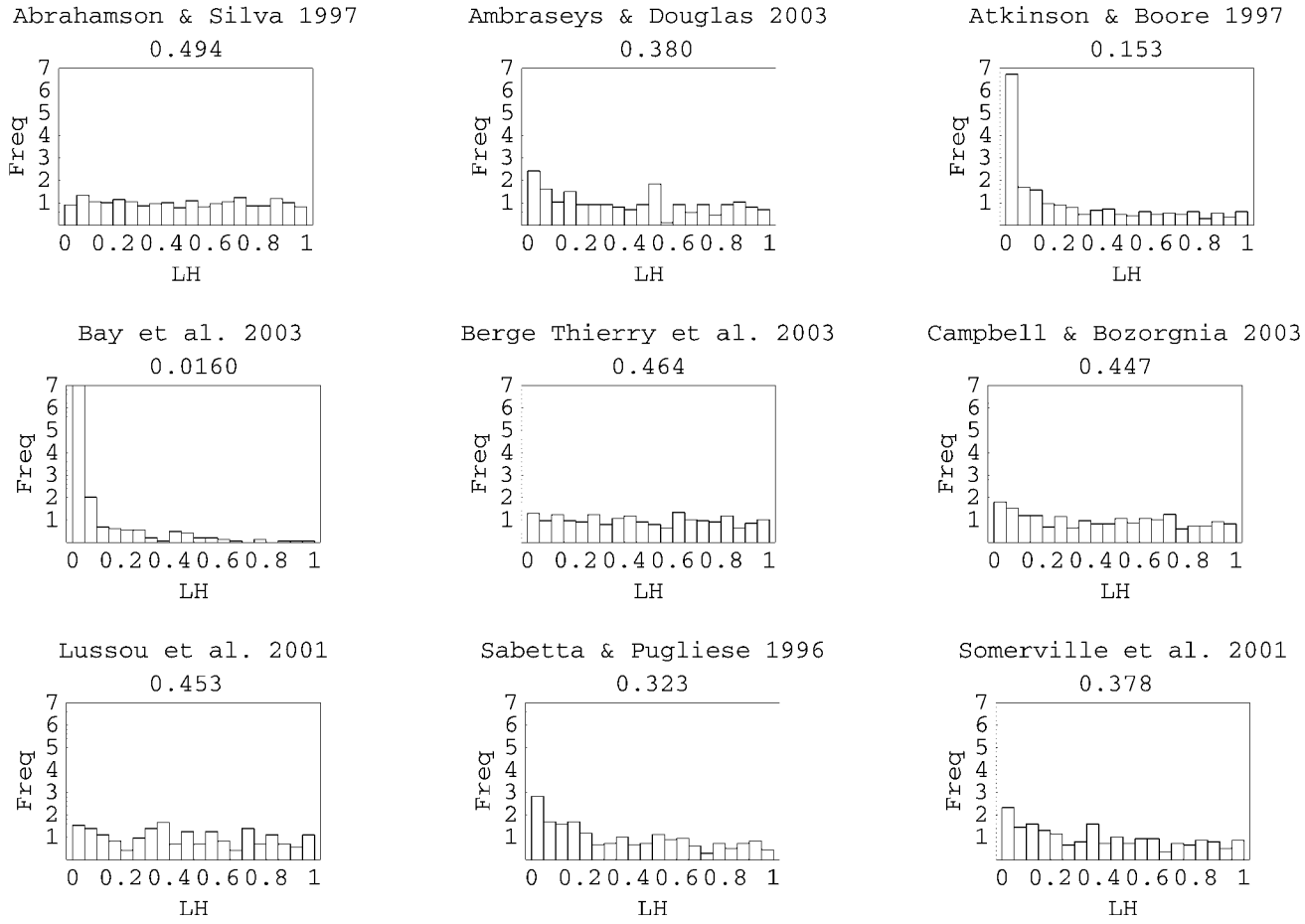

Figure 3. Distribution of *LH* values for a synthetic dataset to mimic the rock-site-record subset of the generating dataset for the Abrahamson and Silva (1997) ground-motion model with respect to different ground-motion-prediction equations. All panels are scaled to the same maximum value.

taset and that Ambraseys and Douglas (2003) give a mean prediction (and median) value that is higher than the observed one. Mean and median are similar in that they both quantify the central tendency of the distribution, but they are known to be different in the norm, they minimize (mean, L2; median, L1), and therefore in their robustness with respect to outliers. The Std-NRES clearly shows that some of the models (e.g., Sabetta and Pugliese, 1996) that perform well in terms of the central tendency of the distribution do not perform equally well in terms of the spread of the residuals.

So, in conclusion, we decided to consider further only those goodness-of-fit measures that could be easily interpreted in an absolute sense and that, in combination, would allow the characterization of a good model. These measures are indicated by the gray-shaded columns in Table 3.

Goodness-of-Fit-Measure Variance. An additional practical issue is the determination of goodness-of-fit-measure variance. The accuracy with which a particular goodness-of-fit measure can be determined in practice will be different for each dataset, depending on the distances, magnitudes, and frequencies for which observed data will be available. In order to quantify the corresponding variance, we used "delete-1" jackknife resampling (e.g., Wu, 1986) on both individual frequency values and complete spectra. The square root of the sum of both variances is assumed to be an estimate of the overall goodness-of-fit-measure standard deviation. For the subset of rock-site records that was used to generate the Abrahamson and Silva (1997) model, the complete set of used goodness-of-fit values, together with their "delete-1" jackknife estimators of standard deviation, is shown in Table 4.

What Is a Well-Matching Model?

In order to develop a categorization scheme to quantify overall model capability for predicting observed-response spectra, we calculated the selected set of goodness-of-fit measures discussed previously, using the rock-site-record subsets for generating a number of popular ground-motion models (Ambraseys and Simpson, 1996; Sabetta and Pugliese, 1996; Berge-Thierry *et al*., 2003). These records are

Table 3

Comparison of Different Ground-Motion-Prediction Relations to Model the Generating Dataset of Rock-Site Records for the Abrahamson and Silva (1997) Ground-Motion Model

Model Name	MEDLH	VARRED	CHISQ MF	PCC	KS	CHISQ	V-Test	M-Test	MEDNR	MEANNR	STDNR	No. of Records
Berge-Thierry *et al.* (2003)	0.496	0.999	0.928	0.727	3.42e-9	9.63e-9	0.00292	0	0.397	0.366	0.892	62
Abrahamson and Silva (1997)	0.466	0.999	1.09	0.725	0.000221	0	0.568	1.48e-6	0.218	−0.23	1.02	73
Campbell and Bozorgnia (2003)	0.42	0.999	1.6	0.715	0.000121	0	0	1.43e-6	0.163	−0.232	1.24	72
Sabetta and Pugliese (1996)	0.389	0.999	1.62	0.752	0.0000166	0	0	0.00272	0.25	0.159	1.26	59
Somerville *et al.* (2001)	0.368	0.998	1.65	0.525	0.000433	0	3.4e-10	0.0131	−0.104	−0.149	1.28	46
Lussou *et al.* (2001)	0.355	0.996	1.38	0.589	0	0	0.714	0	0.691	0.658	0.976	24
Ambraseys and Douglas (2003)	0.317	0.998	1.67	0.643	0.000099	0	9.15e-7	0.000969	−0.396	−0.25	1.27	29
Atkinson and Boore (1997)	0.178	0.998	3.41	0.45	0	0	0	0.0000561	0.241	0.222	1.84	55
Bay *et al.* (2003)	0.0224	0.994	5.89	0.69	0	0	0.228	0	2.28	2.19	1.05	48

Goodness-of-fit measures used are: median *LH* value (MEDLH), variance reduction (VARRED), chi-square misfit (CHISQMF), Pearson correlation coefficient (PCC), Kolmogorov–Smirnov statistic (KS), chi-square statistic (CHISQ), *p*-value of the variance test (V-Test), *p*-value of the mean test (M-Test), and the median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively).

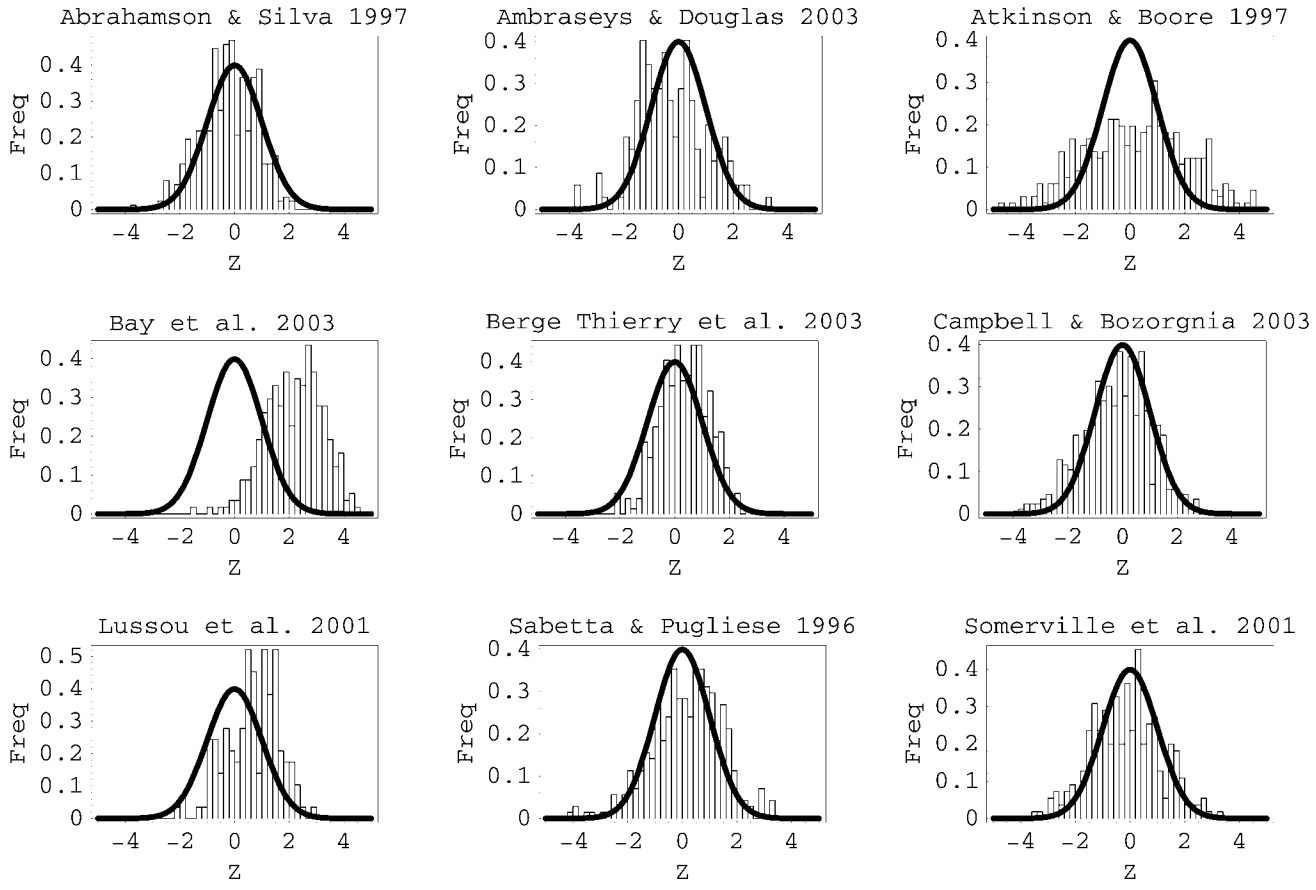

Figure 4. Residual distribution (normalized by model standard deviation) of the rock-site-record subset of the generating dataset for the Abrahamson and Silva (1997) ground-motion model with respect to different ground-motion prediction equations. Solid line shows the expected distribution function for a standard normal distribution.

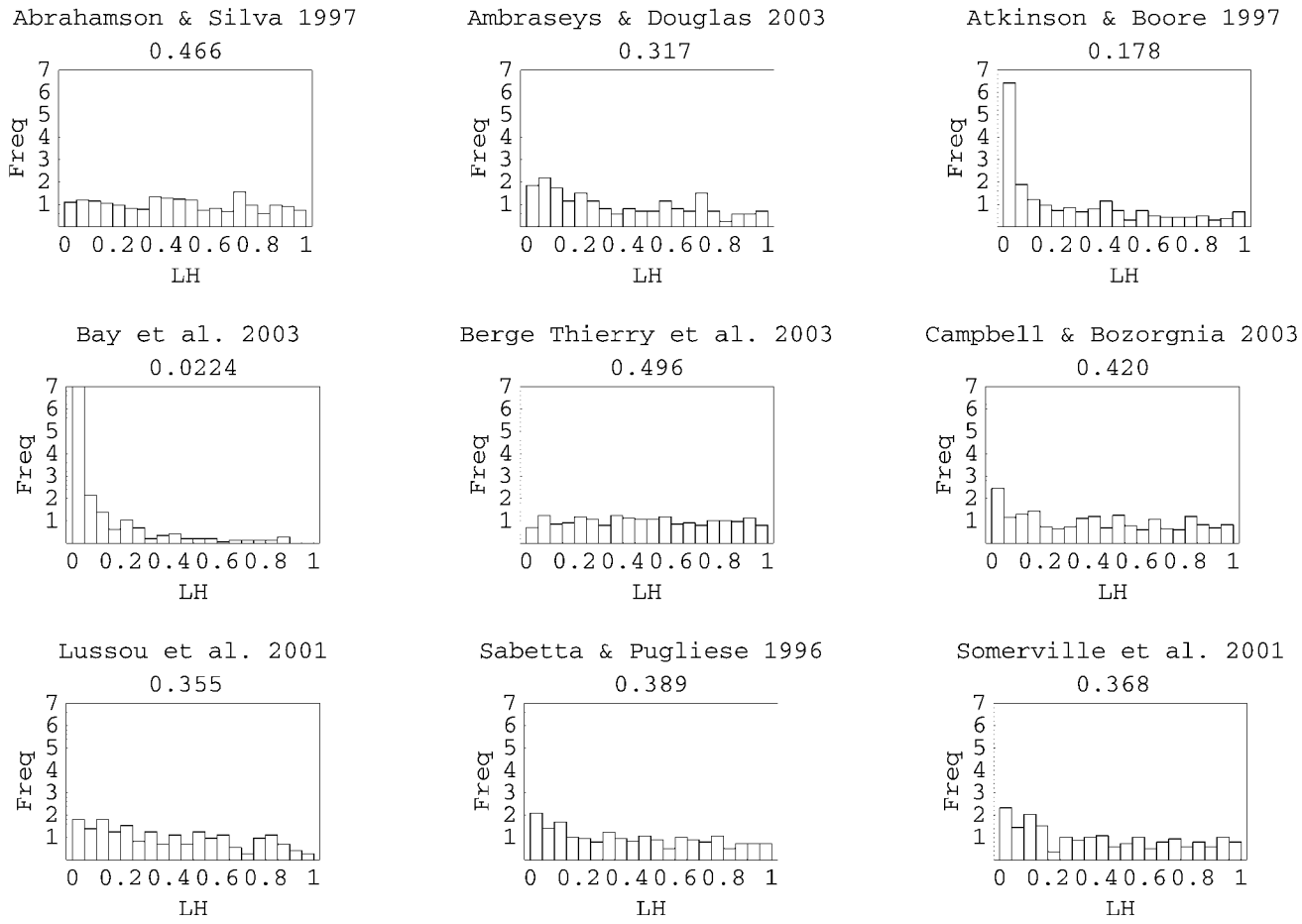

Figure 5. Distribution of *LH* values for the rock-site-record subset of the generating dataset for the Abrahamson and Silva (1997) ground-motion model with respect to different ground-motion prediction equations. All panels are scaled to the same maximum value.

available within the European database of strong-motion records (Ambraseys *et al*., 2004; Ambraseys *et al*., 2000) and were specially flagged for easy extraction. Tables 5–7 show the goodness-of-fit measures together with their standard deviations, and Figures 6–9 display the corresponding distributions of residuals and *LH* values. For some of the ground-motion models, their limited frequency coverage prevented their use for some of the datasets—for example, Ambraseys *et al*. (1996), and Boore *et al*., (1997), and Spudich *et al*. (1999) for the Berge-Thierry *et al*. (2003) rock-site subset. What looks surprising at first glance in Table 5 is that only 102 out of the 122 available rock-site records for the Berge-Thierry *et al*. (2003) model passed the validity range test for the model that was generated from them, whereas for the Abrahamson and Silva (1997) model 114 records were used. However, this result is caused by the fact that Berge-Thierry *et al*. (2003) recommended their model for use only for distances below 100 km, although the distance range covered by the generating dataset was actually larger, as is the distance range covered by the Abrahamson and Silva (1997) model. As can be seen from Table 5 and Figures 6, and 7, the rock site subset of the Berge-Thierry *et al*. (2003) model is very well predicted by the model of Lussou *et al*. (2001) and also by the Abrahamson and Silva (1997) model. The Lussou *et al*. (2001) model performs very well on the median of the *LH* values and the spread of the distribution (indicated by Std-NRES = 1.01), whereas it under-predicts the central tendency. The Abrahamson and Silva (1997) model, on the other hand, has both a stronger bias and a larger spread of the residuals. The Berge-Thierry *et al*. (2003) model is unbiased, has a fairly high median of *LH* values, and matches the spread of the residual distribution very well. Hence, the combination of the interpretation of *LH* values and the parameters describing central tendency and spread of the NRES seems to provide a comprehensible ranking. For the rock-site-record subset of the Sabetta and Pugliese (1996) model, the results are somewhat counter-intuitive at first glance. Here the median *LH* value for the original model comes out only third in rank. However, at closer look this is not too surprising, because the small number of carefully selected records from a single region results in a small model variance, which in turn is penalized by the *LH* values as dis-

Table 4
Rankings of Different Ground-Motion-Prediction Relations to Model a Subset of the Generating Dataset of the Abrahamson and Silva (1997) Ground-Motion Model

Model Name	Rank	MEDLH	σ	MEDNR	σ	MEANNR	σ	STDNR	σ	No. of Records
Berge-Thierry *et al.* (2003)	B	0.496	0.0249	0.397	0.0982	0.366	0.101	0.892	0.0696	62
Abrahamson and Silva (1997)	A	0.466	0.0146	−0.218	0.152	−0.23	0.0984	1.02	0.0315	73
Campbell and Bozorgnia (2003)	B	0.42	0.0209	−0.163	0.138	−0.232	0.156	1.24	0.0876	72
Sabetta and Pugliese (1996)	C	0.389	0.00816	0.25	0.0983	0.159	0.0802	1.26	0.0476	59
Somerville *et al.* (2001)	C	0.368	0.0189	−0.104	0.279	−0.149	0.207	1.28	0.0533	46
Lussou *et al.* (2001)	C	0.355	0.136	0.691	0.17	0.658	0.197	0.976	0.0466	24
Ambraseys and Douglas (2003)	C	0.317	0.029	−0.396	0.114	−0.25	0.163	1.27	0.0715	29
Atkinson and Boore (1997)	D	0.178	0.0722	0.241	0.491	0.222	0.441	1.84	0.204	55
Bay *et al.* (2003)	D	0.0224	0.0136	2.28	0.218	2.19	0.168	1.05	0.024	48

Goodness-of-fit measures used are: median *LH* values (MEDLH) and the median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively). Ranking scheme and calculation of standard deviation σ are explained in the text.

Table 5
Rankings of Different Ground-Motion-Prediction Relations to Model a Subset of the Generating Dataset of the Berge-Thierry *et al.* (2003) Ground-Motion Model

Model Name	Rank	MEDLH	σ	MEDNR	σ	MEANNR	σ	STDNR	σ	No. of Records
Lussou *et al.* (2001)	A	0.481	0.00113	0.186	0.0447	0.248	0.0371	1.01	0.00179	76
Berge-Thierry *et al.* (2003)	A	0.474	0.00264	0.0597	0.00346	0.092	0.00523	1.07	0.0105	102
Abrahamson and Silva (1997)	B	0.397	0.00544	−0.219	0.0303	−0.261	0.0392	1.24	0.0098	114
Somerville *et al.* (2001)	D	0.255	0.00729	−0.0689	0.132	−0.162	0.0811	1.62	0.00987	41
Sabetta and Pugliese (1996)	D	0.223	0.0131	−0.386	0.0524	−0.365	0.0504	1.69	0.0313	105
Ambraseys and Douglas (2003)	D	0.2	0.0167	−0.58	0.0646	−0.78	0.0183	1.7	0.0301	18

Goodness-of-fit measures used are: median *LH* values (MEDLH) and the median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively). Ranking scheme and calculation of standard deviation σ are explained in the text.

cussed previously. Because the model standard deviations of the Abrahamson and Silva (1997) and the Berge-Thierry *et al*. (2003) models are larger than for the Sabetta and Pugliese (1996) model, this also explains why, in terms of *LH* values, these models perform well on the data subset of the Sabetta and Pugliese (1996) model but not vice versa. Note that the standard deviations given in Table 6 are normalized with respect to the model standard deviations of each of the candidate models. Hence the value of 1.24 for the Sabetta and Pugliese (1996) model means that the rock-site-subset sample standard deviation is actually about 25% larger than the one for the full model. In the context of a Probabalistic Seismic Hazard Analysis (PSHA), this could potentially cause problems because at low frequencies of exceedance the hazard curve is strongly affected by the scatter of the ground-motion model (Restrepo-Velez and Bommer, 2003). A similar effect (1.21) is also visible for the Ambraseys and Simpson (1996) model, the only vertical-component model that we considered (Table 7).

On the basis of the analysis of the datasets discussed previously, we feel that currently only a combination of different goodness-of-fit measures seems to sufficiently well describe the overall capability of a model to match an existing dataset. For practical applications, we have used the following scheme, which consists of three categories:

- For a model to be ranked in the lowest accepted capability

Table 6
Rankings of Different Ground-Motion-Prediction Relations to Model a Subset of the Generating Dataset of the Sabetta and Pugliese (1996) Ground-Motion Model

Model Name	Rank	MEDLH	σ	MEDNR	σ	MEANNR	σ	STDNR	σ	No. of Records
Berge-Thierry *et al.* (2003)	B	0.513	0.00315	0.311	0.0153	0.314	0.00603	0.816	0.0118	20
Abrahamson and Silva (1997)	B	0.491	4.97e-16	−0.273	0.0346	−0.376	0.0311	1.12	0.023	24
Sabetta and Pugliese (1996)	B	0.416	0.0106	0.0618	0.0126	0.00882	0.0287	1.24	0.0114	21
Ambraseys and Douglas (2003)	C	0.395	0.0174	−0.362	0.0688	−0.367	0.135	1.31	0.086	8
Lussou *et al.* (2001)	C	0.39	0.0143	0.679	0.00152	0.505	0.04	0.882	0.0273	12
Somerville *et al.* (2001)	C	0.339	0.0115	0.155	0.116	0.0778	0.0661	1.37	0.0166	13

Goodness-of-fit measures used are: median *LH* values (MEDLH) and the median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively). Ranking scheme and calculation of standard deviation σ are explained in the text.

Table 7
Rankings of Different Ground-Motion-Prediction Relations to Model a Subset of the Generating Dataset of the Ambraseys and Simpson (1996) Ground-Motion Model

Model Name	Rank	MEDLH	σ	MEDNR	σ	MEANNR	σ	STDNR	σ	No. of Records
Berge-Thierry *et al.* (2003)	A	0.479	0.0246	−0.0606	0.0302	−0.0279	0.0247	1.06	0.00739	57
Ambraseys and Simpson (1996)	B	0.425	0.011	−0.0411	0.0217	−0.117	0.0174	1.21	0.0103	68
Abrahamson and Silva (1997)	C	0.401	0.011	−0.317	0.0358	−0.269	0.0226	1.26	0.0217	65
Lussou *et al.* (2001)	C	0.382	0.0122	0.568	0.0398	0.615	0.0106	1.12	0.00777	42
Bay *et al.* (2003)	D	0.339	0.0154	0.809	0.141	0.93	0.0601	1.22	0.013	60
Ambraseys and Douglas (2003)	D	0.298	0.0196	−0.982	0.0884	−0.911	0.0711	1.1	0.0379	11
Sabetta and Pugliese (1996)	D	0.271	0.0167	−0.175	0.0448	−0.143	0.0345	1.57	0.0168	62
Somerville *et al.* (2001)	C	0.27	0.0312	−0.647	0.19	−0.736	0.0877	1.44	0.0171	27
Campbell and Bozorgnia (2003)	D	0.26	0.0195	−0.922	0.0232	−0.831	0.0432	1.31	0.0222	62

Goodness-of-fit measures used are: median *LH* values (MEDLH) and the median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively). Ranking scheme and calculation of standard deviation σ are explained in the text.

class (C), we require a median *LH* value of at least 0.2, the absolute value of mean and median of the normalized residuals, and their standard deviations to be smaller than 0.75. In addition, the normalized sample standard deviation is required to be smaller than 1.5.

- For a model to be ranked in the intermediate capability class (B), we require a median *LH* value of at least 0.3, the absolute value of mean and median of the normalized residuals, their standard deviations to be smaller than 0.5, and the normalized sample standard deviation to be smaller than 1.25.
- For a model to be ranked in the highest capability class (A), we require a median *LH* value of at least 0.4, the absolute value of both measures of central tendency of the normalized residual distribution, and their standard deviations not to deviate more than 0.25 from zero. In addition, the normalized sample standard deviation is required to be smaller than 1.125.

A model that does not meet the criteria for any of these categories is ranked unacceptable or class D. It should be emphasized again that the purpose of this scheme is to provide a data-driven selection and ranking of ground-motion models for seismic-hazard assessment. Therefore, it does not provide judgments about particular models being "good" or "bad." It only measures, within the limits of the available

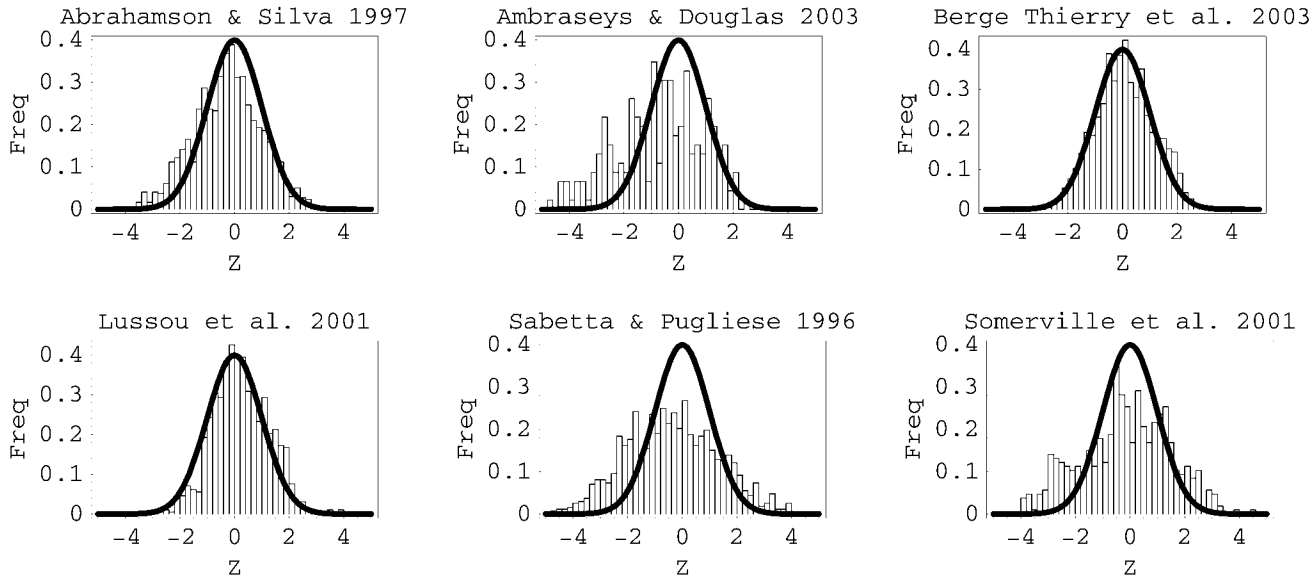

Figure 6. Residual distribution (normalized by model standard deviation) of the rock-site-record subset of the generating dataset for the Berge-Thierry *et al.* (2003) ground-motion model with respect to different ground-motion prediction equations. Solid line shows the expected distribution function for a standard normal distribution.

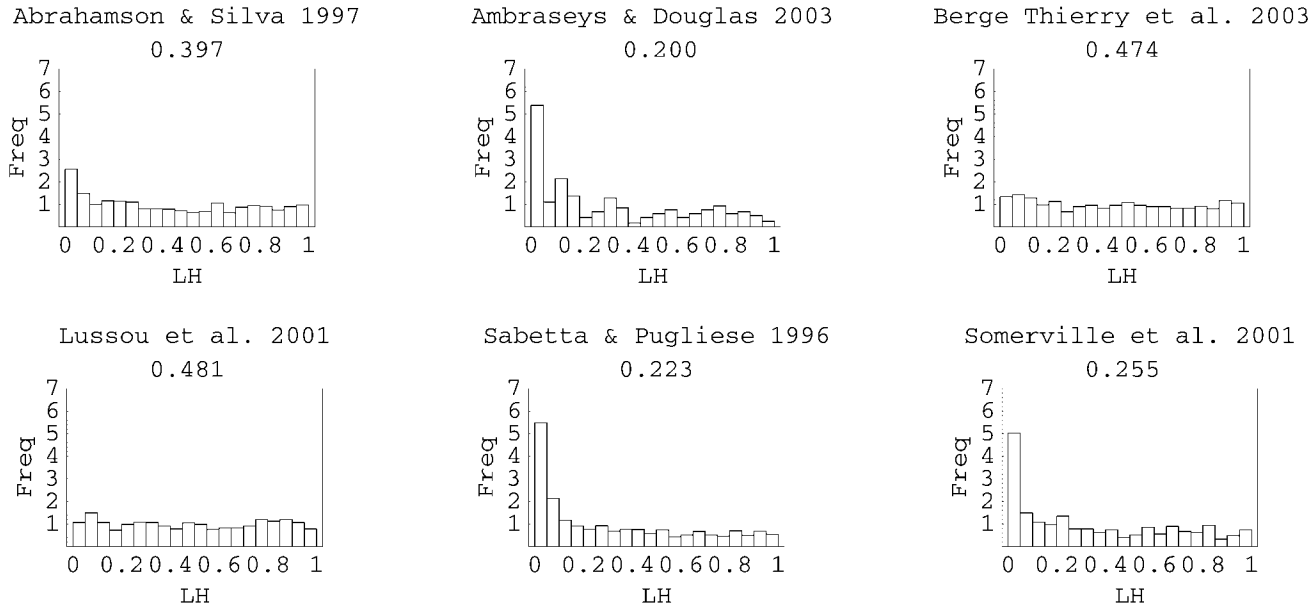

Figure 7. Distribution of *LH* values for the rock-site-record subset of the generating dataset for the Berge-Thierry *et al.* (2003) ground-motion model with respect to different ground-motion-prediction equations. All panels are scaled to the same maximum value.

data, how well each model seems to appropriately represent a given magnitude, distance, and frequency range for a particular region. The ranking results for the rock-site subsets of the generating datasets already discussed are given in Tables 4–7. Next we discuss the application of this scheme to the recordings of the M_w 4.8 St. Dié earthquake of 22 February 2003 in eastern France. Despite its moderate magnitude, this event is important for SHA in the border region of France, Germany, and Switzerland, especially for critical installations.

St. Dié Earthquake

The Saint Dié earthquake occurred in eastern France at 21:41 p.m. local time on 22 February 2003. According to

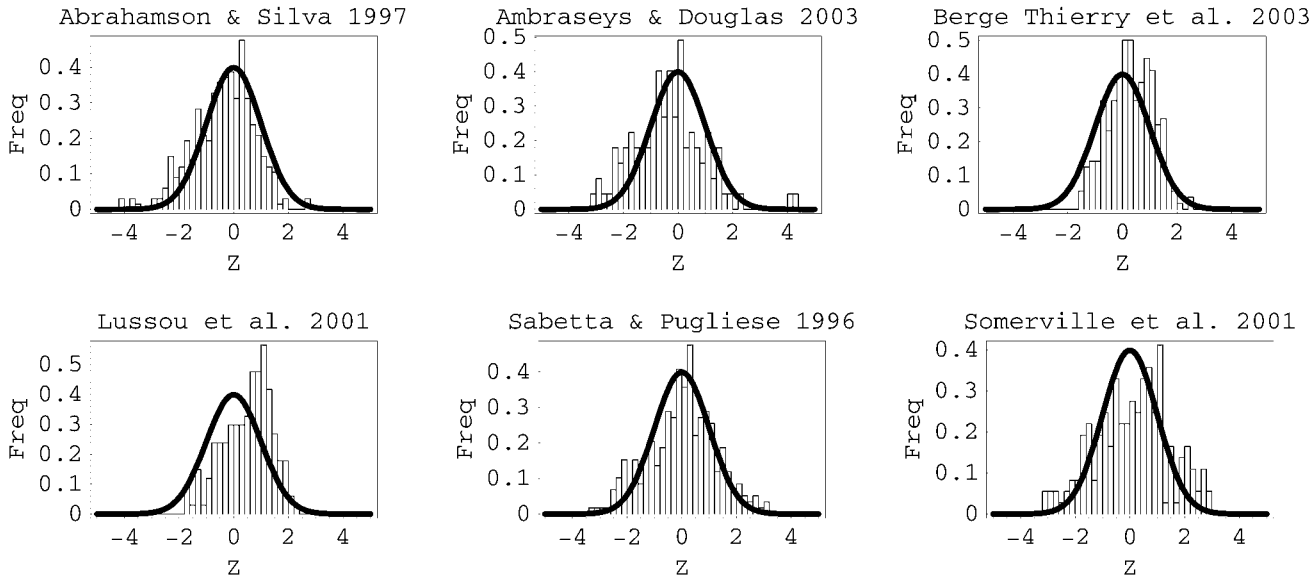

Figure 8. Residual distribution (normalized by model standard deviation) of the rock-site-record subset of the generating dataset for the Sabetta and Pugliese (1996) ground-motion model with respect to different ground-motion-prediction equations. Solid line shows the expected distribution function for a standard normal distribution.

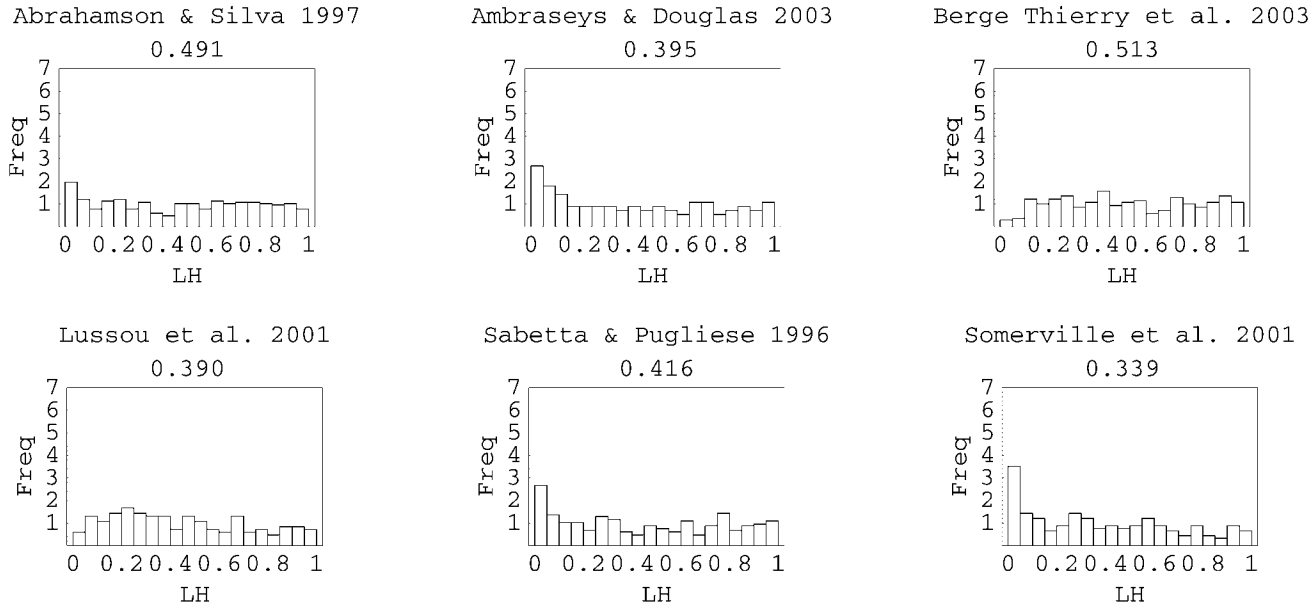

Figure 9. Distribution of *LH* values for the rock-site-record subset of the generating dataset for the Sabetta and Pugliese (1996) ground-motion model with respect to different ground-motion-prediction equations. All panels are scaled to the same maximum value.

the Bureau Central Seismologique Français (BCSF), the maximum intensity reached values of VI–VII on the EMS98 scale. No injury was reported, although several chimneys did collapse in the epicentral region. The hypocenter was located at latitude 48.34° N, longitude 6.66° E, with a focal depth of about 10 km, according to the Renass French seismological network. Magnitude estimates vary significantly according to national or international agencies (Table 8) and range between 4.3 and 5.8. However, M_w and M_b magnitude determinations show similar values, close to 4.8. This earthquake did occur north of a zone where several earthquakes (1682: I_0 = VIII) and swarms (1984, 1971–1974) were observed (Haessler and Hoang-Trong, 1985). The mainshock had a predominantly normal mechanism with a strike-slip

Table 8
Description of the St. Dié Earthquake

Agency	Magnitude Type	Magnitude	Lat. (°N)	Long. (°E)	Depth (km)
LDG (France)	M_L	5.8	48.34	6.66	10
RENass (France)	M_L	5.4	48.34	6.66	10
EMSC (Mix, EC)	M_b	4.7	48.3	6.6	10
EMSC (EC)	M_L	5.8	48.3	6.8	10
SED (Switzerland)	M_w	4.78	48.35	6.8	9
SED (Switzerland)	M_L	5.5	48.4	6.5	10
INGV (Italy)	M_w	4.7	—	—	—
GSRC (Russia)	M_s	4.3	48.9	6.9	—
LED (Germany)	M_L	5.4	48.3	6.7	10

Because this earthquake occurred near the boundaries of three countries (France, Germany, and Switzerland), the magnitude and location of the earthquake were provided by several European agencies.

component. The first nodal plane has a strike of 298°, a rake of −131°, and a dip of 62°. The second nodal plane has a strike of 179°, a rake of −39°, and a dip of 48° (Eidgenössische Technische Hochschule Zürich [ETHZ], http://seismo.ethz.ch). About 100 have been recorded. According to preliminary locations, these aftershocks are located at a depth between 12 and 13 km and favor the second nodal plane (BCSF, http://eost.u-strasbg.fr/bcsf).

Within a distance of 200 km, the mainshock was recorded by a total number of 13 free-field accelerograph stations on rock. Their spatial distribution is displayed in Figure 10. The station coordinates are given in Table 9. To test the constraints on the selection of ground-motion models for this region provided by this dataset, we have compared the observed rock-site-response spectra with the same set of ground-motion models used previously (Ambraseys *et al*., 1996; Sabetta and Pugliese, 1996; Abrahamson and Silva, 1997; Atkinson and Boore, 1997; Boore *et al*., 1997; Toro *et al*., 1997; Spudich *et al*., 1999; Lussou *et al*., 2001; Somerville *et al*., 2001; Bay *et al*., 2003; Berge-Thierry *et al*., 2003; Campbell and Bozorgnia, 2003a, b). In this context we ignored the fact that some of these models do not fully cover the frequency range between 0.5 and 25 Hz, the magnitude range down to 4.8 (mostly down to 5.0), and the distance range up to 200 km. Applying the classification scheme proposed previously results in the quality assignments displayed in Table 10. The corresponding normalized residuals and *LH* values are shown in Figures 11 and 12. From the 12 candidate models, only 7 pass as acceptable at all. Out of those, only the model of Berge-Thierry *et al*. (2003) is assigned rank A, whereas four models (Ambraseys *et al*., 1996; Abrahamson and Silva, 1997; Spudich *et al*., 1999; Lussou *et al*., 2001) are categorized as class B because of their larger bias. Some models fail completely on the basis of their large bias (Sabetta and Pugliese, 1996; Toro *et al*., 1997) and others on both bias and low *LH* values (Atkinson and Boore, 1997; Boore *et al*., 1997; Bay *et al*., 2003). Figure 13 shows a comparison between the observed spectral values (1 Hz and 10 Hz) and the class A and B model predictions. Only a qualitative visual evaluation of the fit between data and model predictions is provided by such classical comparison. The *LH* method ranks the Berge-Thierry *et al*. (2003) model, with objective and automatic numerical criteria, as the best model, which is confirmed by careful visual but naturally incomplete (because only two frequencies are examined) analysis of Figure 13.

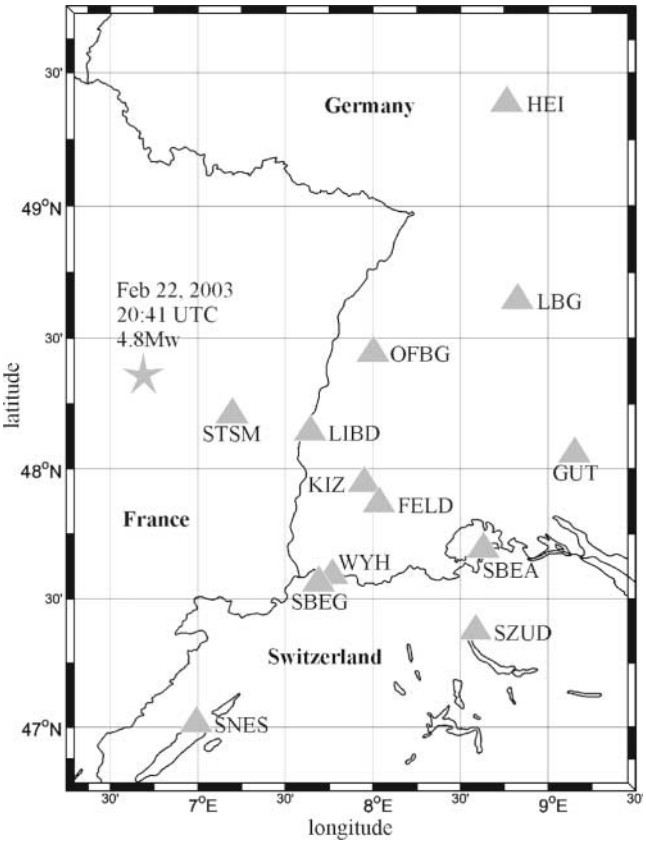

Figure 10. Location of the M_w = 4.8 St. Dié earthquake of 22 February 2003 (star). Triangles show free-field rock sites with accelerometer recordings within a distance range of 200 km.

Of special relevance for the area of the St. Dié earthquake is the model of Bay *et al*. (2003), which was derived from the analysis of weak-motion data from Switzerland. Although generated from a dataset regionally very close to

Table 9
Station Names and Coordinates Used for Analysis of the St. Dié Earthquake

Station	Agency	Lat. (°N)	Long. (°E)	Altitude (m)	Epicentral Distance (km)
WYH	LEDBW	47.5508	7.7018	310	131
OFBG	LEDBW	48.4453	7.9770	240	106
LIBD	LEDBW	48.1505	7.6030	210	85
HEI	LEDBW	49.3991	8.7274	562	193
FELD	LEDBW	47.8763	8.0040	1480	125
LBG	LEDBW	48.6639	8.7945	583	168
KIZ	LEDBW	47.9562	7.9182	499	115
GUT	LEDBW	48.0709	9.1153	647	195
STSM	RAP	48.22	7.16	580	51
SZUD	SSMNET	47.371	8.581	615	193
SNES	SSMNET	47.001	6.954	481	162
SBEG	SSMNET	47.572	7.666	370	127
SBEA	SSMNET	47.700	8.597	490	173

Recording agencies are Landeserdbebendienst Baden-Württemberg, Germany (LEDBW); Réseau Accélerométrique Permanent, France (RAP); and Swiss Strong Motion Network, Switzerland (SSMNET).

Table 10
Rankings of Different Ground-Motion-Prediction Relations for Modeling the Dataset of the M_w 4.8 St. Dié Earthquake of 22 February 2003

Model Name	Rank	MEDLH	σ	MEDNR	σ	MEANNR	σ	STDNR	σ	No. of Records
Lussou *et al.* (2001)	B	0.579	0.0849	0.479	0.17	0.454	0.175	0.716	0.122	13
Berge-Thierry *et al.* (2003)	A	0.575	0.0478	0.114	0.198	0.075	0.19	0.793	0.0909	13
Abrahamson and Silva (1997)	B	0.558	0.0899	0.383	0.157	0.336	0.186	0.8	0.098	13
Ambraseys *et al.* (1996)	B	0.508	0.0857	0.00176	0.29	−0.0682	0.221	1.02	0.0732	13
Somerville *et al.* (2001)	C	0.435	0.0584	0.678	0.179	0.663	0.242	1.08	0.179	13
SEA99	B	0.434	0.0952	0.0505	0.418	−0.00254	0.227	1.06	0.0874	13
Campbell and Bozorgnia (2003)	C	0.43	0.115	−0.585	0.402	−0.63	0.22	0.975	0.092	13
Toro *et al.* (1997)	D	0.404	0.0892	−0.74	0.232	−0.808	0.236	0.98	0.144	13
Sabetta and Pugliese (1996)	D	0.228	0.0731	−1.11	0.256	−1.2	0.3	1.35	0.102	13
Boore *et al.* (1997)	D	0.161	0.0683	−1.35	0.292	−1.37	0.272	1.34	0.1	13
Atkinson and Boore (1997)	D	0.147	0.034	−1.41	0.197	−1.28	0.223	1.27	0.0905	13
Bay *et al.* (2003)	D	0.0116	0.0126	2.52	0.34	2.51	0.207	0.935	0.129	13
Bay *et al.* (2003) (HIGH STRESSDROP)	A	0.572	0.0572	−0.04	0.183	0.0199	0.208	0.835	0.148	13

Goodness-of-fit measures used are: median *LH* values (MEDLH) and the median, mean, and standard deviation of the normalized residuals (MEDNR, MEANNR, and STDNR, respectively). Ranking scheme and calculation of standard deviation σ are explained in the text. Rejected models are class D, and the parameter values causing the rejections are shaded gray.

SEA99 (in first column) is the ground-motion-prediction relation by Spudich *et al.* (1999).

the St. Dié earthquake, it shows a strong under-prediction of the rock-site records considered here. This is indicated by the strong bias of the central-tendency measures (Table 10) and the residual distribution shown in Figure 11. To understand the reasons for this discrepancy, we have performed a number of permutations to their set of suggested model parameters, subsequently testing whether the modified model spectra would match the observed ones. The modifications included the stress drop, the kappa values, the geometrical spreading factors, and the damping values. Although discussion of further details of this separate study are beyond the scope of this article, for the present context it is inter-

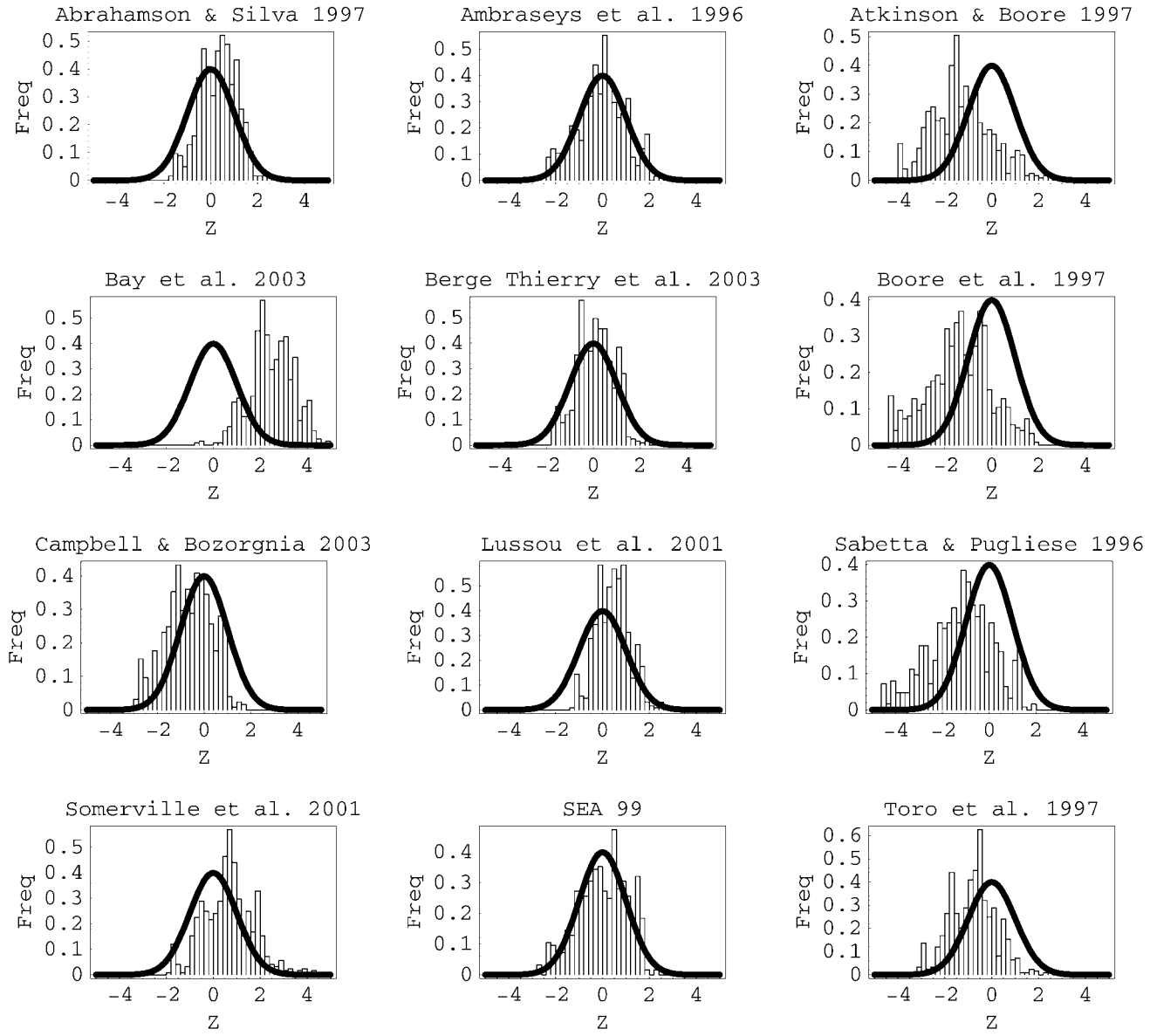

Figure 11. Residual distribution (normalized by model standard deviation) of the rock-site-record subset of the $M_w = 4.8$ St. Dié earthquake of 22 February 2003 with respect to different ground-motion-prediction equations. Solid line shows the expected distribution function for a standard normal distribution. SEA99 is the ground-motion-prediction relation by Spudich *et al.* (1999).

esting to note that all the successful models suggest a stress drop for the St. Dié earthquake of at least 80 bars. Independent evidence for a high stress drop of this event can also be seen in the relatively high M_L values shown in Table 8. With the modified stress drop (referred to as Bay *et al*., 2003 HIGH STRESSDROP in Table 10), the Bay *et al*. (2003) model would become the second best model to explain the St. Dié response spectra and would be ranked A (last row in Table 10). This dramatic improvement can also be seen in the corresponding distribution of residuals and *LH* values shown in Figure 14. Although at first glance this simply seems to suggest a modification of the Bay *et al*. (2003) model in terms of the stress parameter (stress drop), we actually do not make this proposition. We believe that the underlying problem may be of a more general nature, which, for larger earthquakes, may also require additional modifications (e.g., geometrical spreading), and that we are currently still far from understanding to what degree weak motion can be used to predict strong motion.

Discussion and Conclusions

With the growing number of ground-motion-prediction equations being published in recent years (Douglas, 2003),

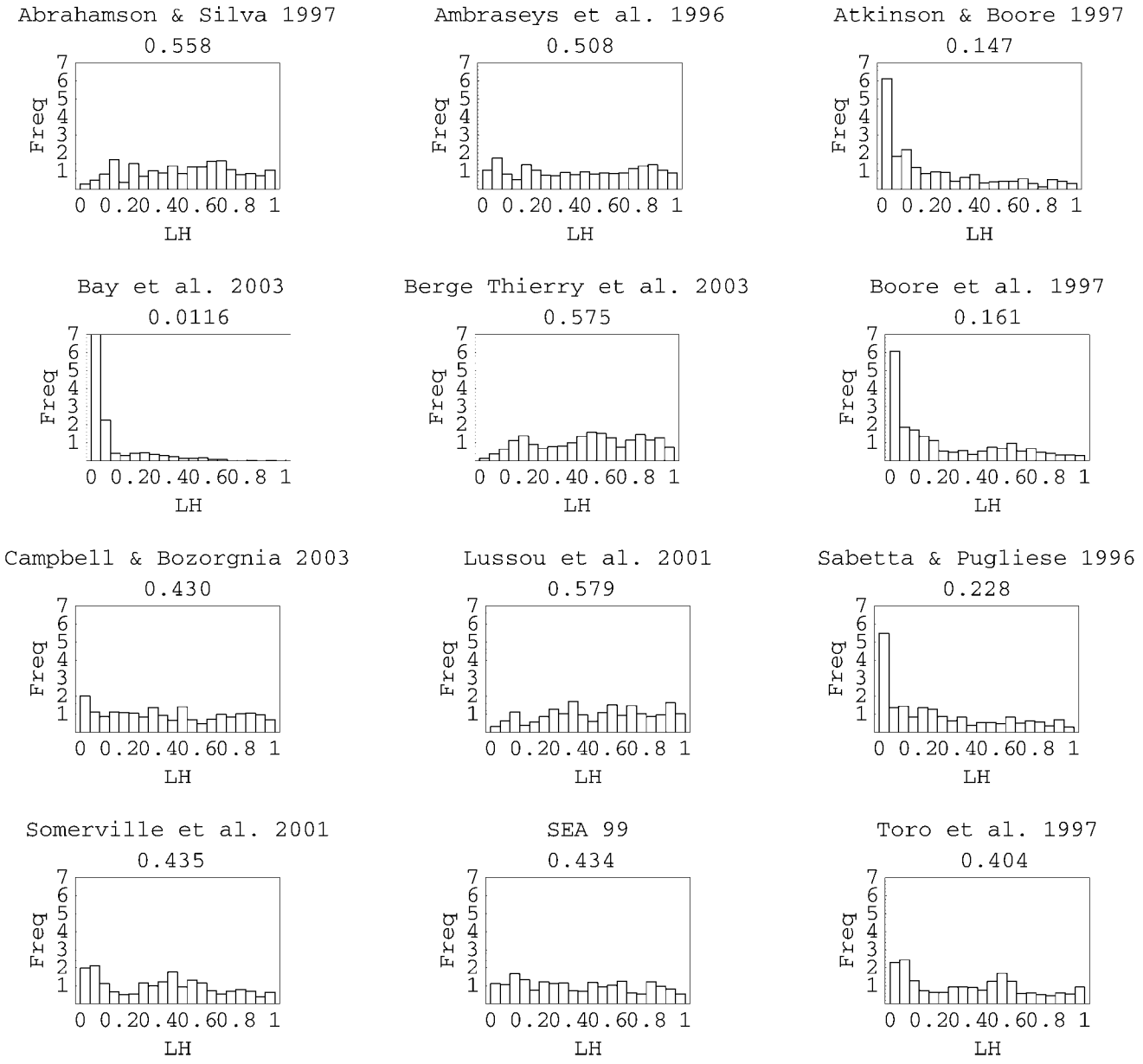

Figure 12. Distribution of *LH* values for the rock-site-record subset of the M_w = 4.8 St. Dié earthquake of 22 February 2003 with respect to different ground-motion-prediction equations. All panels are scaled to the same maximum value.

the selection of ground-motion models has become a serious practical problem in the context of designing logic-trees for seismic-hazard assessment. Logic-trees are widely used as a tool to capture the epistemic uncertainty associated with input parameters related to seismogenic sources and the ground-motion-prediction model (Bommer *et al*., 2004b). The final stage of generating a logic-tree is to assign weights to each of the selected ground-motion models. These weights reflect the degree to which each equation is judged to be the best estimate of earthquake ground-motion in that particular region. The aim of the method proposed here is to include observed data in this judgment. On the basis of on the analysis of a number of different goodness-of-fit measures, we have identified a set of measures that allow a consistent and comprehensible ranking of candidate ground-motion models according to their overall capability to predict observed strong-motion records (response spectra). For this purpose, we have developed a new, likelihood-based, goodness-of-fit measure that has the property not only to quantify the model fit but also to measure in some degree how well the underlying model assumptions are met. By design it naturally scales between 0 and 1, which we find a convenient aspect in the context of assigning logic-tree-branch weights. However, since goodness-of-fit to observed-

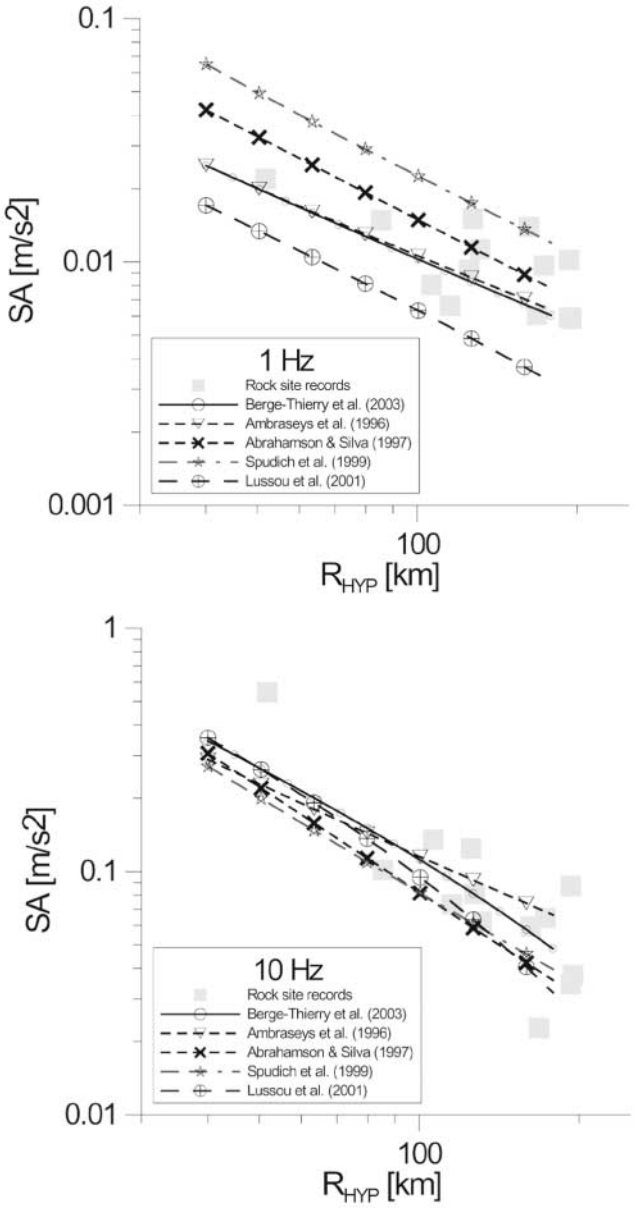

Figure 13. Comparison of observed rock-site-spectral values (1 Hz and 10 Hz) of the St. Dié earthquake with the values predicted by class A and B empirical models.

response spectra may not be the only relevant aspect influencing the assignment of branch weights, we are not proposing a fully automated weighting scheme. Instead, on the basis of tests to recognize a number of popular ground-motion models from the rock-site subsets of the datasets used to generate them in the first place, we have derived a simple categorization scheme to rank candidate ground-motion models into different classes. Such a scheme is intended to assist the seismic-hazard analyst in judging the appropriateness of ground-motion models for a particular target area in a data-driven, consistent, and reproducible way.

During our tests we recognized that the variability of the rock-site samples of the generating datasets for some of the empirical models is larger than the one for the full model. This result could affect seismic-hazard estimates for rock sites, because the hazard curve is strongly affected by the scatter of strong motion models for low frequencies of exceedance. Finally, using the records of the M_w 4.8 St. Dié earthquake of 22 February 2003, we find that a few observed response spectra of high quality can already provide considerable constraints on the selection of ground-motion models for seismic-hazard analysis, because 5 out of 12 candidate models were ranked inappropriate for that region. In this context, we found a strong under-prediction of the St. Dié dataset by the only regional candidate model that, however, was entirely derived from weak-motion data (Bay *et al*., 2003). This points to a more general question: to what degree are weak-motion data useful for strong motion prediction, which in our opinion is currently not understood. As the discussion of the recent literature demonstrates (e.g., Ide and Beroza, 2001), several factors such as magnitude-dependent stress drop and geometrical spreading variation, could explain why ground-motion models based on weak motion records cannot be used directly for strong ground-motion modeling.

We envision several potential areas of application for our classification scheme. Because it allows a consistent and transparent categorization of ground-motion models, one of its main applications is seen in connection with generating weights for logic-tree branches occupied by different candidate models. In regions with few or no indigenous equations, our scheme can also help to select or reject published empirical models for application. As shown with the example of the St. Dié earthquake, this can even be performed with a rather small dataset collected during a single earthquake. In such a case, however, the inter-event variability will not be captured within the observed data, and the total variability will be underestimated. Therefore, it is desirable to include records from various earthquakes, if available. Such an analysis would be particularly useful and often possible in regions where modern digital strong-motion networks have recently been installed but the databases are still insufficient for deriving new native empirical models (e.g., Algeria, Iran). In such a situation, our scheme could help to select appropriate foreign ground-motion models until native ones become available.

In combination with the hybrid approach (Campbell, 2003) our classification scheme allows us to measure the performance of particular host-to-target regional conversions. As a consequence, in connection with simple classifications to describe the data coverage of a ground-motion model in terms of magnitude, distance, and frequencies, composite ground-motion models for arbitrary target areas with a small number of observed-response spectra can be constructed in a completely transparent and reproducible way. This is the subject of current work.

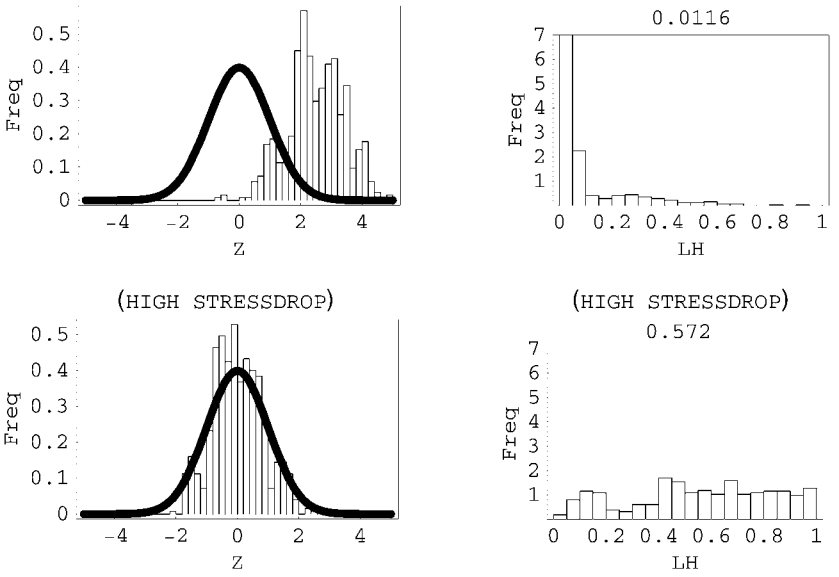

Figure 14. Comparison of the model of Bay *et al.* (2003) (upper panel) with a modified version in which the stress drop was set to 90 bars (lower panel). Left column shows the residuals, and the right column displays the corresponding distribution of *LH* values.

Acknowledgments

This paper is contribution EG2/DT-03 from a series of studies inspired by participation in the Pegasos project (Abrahamson *et al*., 2002). We thank the following people for providing a stimulating environment and continuous support for the development of the ideas presented herein: Philip Birkhäuser, Jim Farrington, Andreas Hölker, Philippe Roth, and Christian Sprecher. Frank Scherbaum and Fabrice Cotton are especially thankful for the numerous discussions with Julian Bommer, Hilmar Bungum, Fabio Sabetta, and Norm Abrahamson. We also thank Norm Abrahamson for generously providing us with the dataset used to generate the Abrahamson and Silva (1997) ground-motion model, Francesca Bay for her cooperation and patience with our questions regarding her ground-motion model, Matthias Ohrnberger for his help in checking the response spectral calculations, Ken Campbell for his support in properly implementing the Campbell and Bozorgnia (2003a) ground-motion model, and Julian Bommer, Dave Boore, and two anonymous reviewers for their constructive criticism of the manuscript. We are grateful to the Landesamt fuer Geologie, Rohstoffe und Bergbau Baden-Wuerttemberg in Freiburg i. Br., Landeserdbebendienst (LED), the Réseau Accélerométrique Permanent, and the Swiss strong-motion network, Switzerland, for providing the data on the St. Dié earthquake.

References

Abrahamson, N. A., and W. J. Silva (1997). Empirical response spectral attenuation relations for shallow crustal earthquakes, *Seism. Res. Lett.* **68,** 94–127.

Abrahamson, N. A., P. Birkhäuser, M. Koller, D. Mayer-Rosa, P. M. Smit, C. Sprecher, S. Tinic, and R. Graf (2002). *PEGASOS—A comprehensive probabilistic seismic hazard assessment for nuclear power plants in Switzerland*. 12 ECEE, Elsevier, London.

Ambraseys, N. N., and A. Douglas (2003). Near-field horizontal and vertical ground motions, *Soil Dyn. Earthquake Eng.* **23,** 1–18.

Ambraseys, N. N., and M. W. Free (1997). Surface-wave magnitude calibration for European region earthquakes, *J. Earthquake Eng.* **1,** 1–22.

Ambraseys, N. N., and K. A. Simpson (1996). Prediction of vertical response spectra in Europe, *Int. J. Earthquake Eng. Struct. Dyn.* **25,** 401–412.

Ambraseys, N. N., K. A. Simpson, and J. J. Bommer (1996). Prediction of horizontal response spectra in Europe, *Int. J. Earthquake Eng. Struct. Dyn.* **25,** 371–400.

Ambraseys, N. N., P. M. Smit, R. Berardi, D. Rinaldis, F. Cotton, and C. Berge-Thierry (2000). Dissemination of European strong-motion data. CD-ROM collection. European Council, Directorate General XII, Science, Research and Development, Environment and Climate Programme, Brussels.

Ambraseys, N., P. Smit, R. Sigbjornsson, P. Suhadolc, and B. Margaris (2004). Internet-Site for European Strong-Motion Data, www.isesd.hi.is/ESD_Local/home.htm (last accessed October 2004). European Commission, DGXII, Science, Research, and Development, Brussels.

Atkinson, G. M., and D. M. Boore (1997). Some comparisons between recent ground-motion relations, *Seism. Res. Lett.* **68,** 24–40.

Bay, F., D. Fäh, L. Malagnini, and D. Giardini (2003). Spectral shear-wave ground-motion scaling in Switzerland, *Bull. Seism. Soc. Am.* **93,** 414–429.

Berge-Thierry, C., F. Cotton, O. Scotti, D. A. Griot-Pommera, and Y. Fukushima (2003). New empirical response spectral attenuation laws for moderate European earthquakes, *J. Earthquake Eng.* **7,** 193–222.

Bommer, J. J., F. Scherbaum, H. Bungum, F. Cotton, F. Sabetta, and N. A. Abrahamson (2004). On the use of logic trees for ground-motion prediction equations in PSHA, submitted to *Bull. Seism. Soc. Am.*

Boore, D. M., and W. B. Joyner (1997). Site amplifications for generic rock sites, *Bull. Seism. Soc. Am.* **87,** 327–341.

Boore, D. M., W. B. Joyner, and T. E. Fumal (1997). Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes: a summary of recent work, *Seism. Res. Lett.* **68,** 128–153.

Campbell, K. W. (2003). Prediction of strong ground motion using the hybrid empirical method: example application to eastern North America, *Bull. Seism. Soc. Am.* **93,** 1012–1033

Campbell, K. W., and Y. Bozorgnia (2003a). Updated near source ground motion relations for horizontal and vertical components of peak ground acceleration, peak ground velocity and pseudo-absolute acceleration response spectra, *Bull Seism. Soc. Am.* **93,** 314–331.

Campbell, K. W., and Y. Bozorgnia (2003b). Erratum: Updated near-source ground-motion (attenuation) relations for the horizontal and vertical components of peak ground acceleration and acceleration response spectra, *Bull. Seism. Soc. Am.* **93,** 1413.

Cohee, B. P., and G. C. Beroza (1994). Slip distribution of the 1992 Landers earthquake and its implication for earthquake source mechanics, *Bull. Seism. Soc. Am.* **84,** 692–712.

Cotton, F., and M. Campillo (1995). Frequency domain inversion of strong motions: application to the 1992 Landers earthquake, *J. Geophys. Res.* **100,** no. B3, 3961–3975.

Douglas, J. (2003). Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and spectral ordinates, *Earth Sci. Rev.* **61,** 43–104.

Edwards, A. W. F. (1992). *Likelihood*, Expanded ed., Johns Hopkins Univ. Press.

Haessler, H., and P. Hoang-Trong (1985). La crise sismique de Remiremont (Vosges) de décembre 1984: implications tectonique régionales, *Compte-Rendus de l'Académie des Sciences Paris* **300,** 14.

Heaton, T., F. Tajima, and A. Wildenstein Mori (1986). Estimating ground motions using recorded accelerograms, *Surveys in Geophys.* **8,** 25–83.

Ide, S., and G. C. Beroza (2001). Does apparent stress vary with earthquake size?, *Geophys. Res. Lett.* **28,** 3349–3352.

Lussou, P., P. Y. Bard, and F. Cotton (2001). Seismic design regulation codes: contribution of K-Net data to site effect evaluation, *J. Earthquake Eng.* **5,** 13–33.

Malagnini, L., R. B. Herrmann, B. M. Di, and K. Koch (1999). Ground motion attenuation at regional distance in Italy and Germany, *Seism. Res. Lett.* **70,** 214.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2001). *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, Second Ed., Cambridge University Press, Cambridge.

Proseis (2002). Note on the statistical analysis of the ratios of different definitions of the horizontal component, Pegasos Technical Note TP2-TN-0269, Baden, Switzerland.

Raoof, M., R. B. Herrmann, and L. Malagnini (1997). Attenuation and source spectral modeling of three-component ground motion in southern California, *EOS* **46,** Suppl. 434.

Reiter, L. (1990). *Earthquake hazard analysis: Issues and insights*, Columbia University Press, New York.

Restrepo-Velez, L. F., and J. J. Bommer (2003). An exploration of the nature of the scatter in ground-motion prediction equations and the implications for seismic hazard assessment, *J. Earthquake Eng.* **7,** no. 1, 171–199.

Rose, C., and M. D. Smith (2002). *Mathematical Statistics with Mathematica*, Springer, New York.

Sabetta, F., and A. Pugliese (1996). Estimation of response spectra and simulation of nonstationary earthquake ground motion, *Bull. Seism. Soc. Am.* **86,** 337–352.

Scherbaum, F., J. Schmedes, and F. Cotton (2004). On the conversion of source-to-site distance measures for extended earthquake source models, *Bull. Seism. Soc. Am.* **94(3),** 1053–1069.

Somerville, P., N. Collins, N. A. Abrahamson, M. R. Graves, and C. Saikia (2001). Ground Motion Attenuation Relations for the Central and Eastern United States, Final Report, U.S. Geol. Surv. Award Number 99HQGR0098.

Spudich, P., W. B. Joyner, A. G. Lindh, D. M. Boore, B. M. Margaris, and J. B. Fletcher (1999). SEA99: a revised ground motion prediction relation for use in extensional tectonic regimes, *Bull. Seism. Soc. Am.* **89,** 1156–1170.

Steidl, J. H., A. G. Tumarkin, and R. Archuleta (1996). What is a reference site?, *Bull. Seism. Soc. Am.* **86,** 1733–1748.

Toro, G. R., N. A. Abrahamson, and J. F. Schneider (1997). Model of strong ground motions from earthquakes in central and eastern North America: best estimates and uncertainties, *Seism. Res. Lett.* **68,** 41–57.

Wolfram, S. (1996). *The Mathematica Book*, Wolfram Media, Champaign, Illinois.

Wu, C. F. (1986). Jackknife, bootstrap, and other resampling methods in regression analysis, *Ann. Math. Statist.* **14,** 1261–1295.

Appendix

Let x be a continuous random variable with pdf $f(x)$, and let $y = u(x)$ define a one-to-one transformation between x and the transformed random variable y. Then the probability density function of y, say $g(y)$, is $g(y) = f[u^{-1}(y)] \cdot |J|$, where $x = u^{-1}(y)$ is the inverse of $y = u(x)$, and $J = du^{-1}(y)/dy$ denotes the Jacobian of the transformation (Rose and Smith, 2002). In our case

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-x^2}{2}\right) \tag{A1}$$

and

$$u(x) = Erf\left(\frac{x}{\sqrt{2}}, \infty\right). \tag{A2}$$

We temporarily ignore the modulus sign (for z_0) in (equation (9) in the text) and consider only positive values of x. Solving $y \equiv u(x)$ for y (e.g., using Mathematica [Wolfram, 1996]), we obtain the inverse transformation as

$$u^{-1}(y) = InverseErf(\infty, -y) \cdot \sqrt{2}. \tag{A3}$$

The Jacobian becomes

$$\begin{aligned} J &= \frac{\mathrm{d}}{\mathrm{d}y}\left[InverseErf(\infty, -y) \cdot \sqrt{2}\right] \\ &= -e^{InverseErf(\infty, -y)^2} \cdot \sqrt{\frac{\pi}{2}}. \end{aligned} \tag{A4}$$

Finally, with

$$f[u^{-1}(y)] = \frac{e^{-InverseErf(\infty, -y)^2}}{\sqrt{2\pi}} \tag{A5}$$

we obtain

$$\begin{aligned} g(y) &= f[u^{-1}(y)] \cdot |J| \\ &= \frac{e^{-InverseErf(\infty, -y)^2}}{\sqrt{2\pi}} \cdot e^{InverseErf(\infty, -y)^2} \cdot \sqrt{\frac{\pi}{2}} \\ &= \frac{1}{2}. \end{aligned} \tag{A6}$$

Considering both tails of the distribution results in an even distribution for y with a constant pdf of 1.

Institute of Geosciences
University of Potsdam
POB601553, 14415 Potsdam
Germany
(F.S.)

LGIT
University Joseph Fourier
Grenoble, France
(F.C.)

National Emergency Operations Centre
Zurich, Switzerland
(P.S.)

Manuscript received 23 July 2003.