On the Forecasting of Ground-Motion Parameters for Probabilistic Seismic Hazard Analysis

Danny Arroyo[a] **and Mario Ordaz**[b]

It is well understood that the range of application for an empirical ground-motion prediction model is constrained by the range of predictor variables covered in the data used in the analysis. However, in probabilistic seismic hazard analysis (PSHA), the limits in the application of ground-motion prediction models (GMPMs) are often ignored, and the empirical relationships are extrapolated. In this paper, we show that this extrapolation leads to a quantifiable increment in the uncertainty of a GMPM when it is used to forecast a future value of a given intensity parameter. This increment, which is clearly of epistemic nature, depends on the adopted functional form, on the covariance matrix of the regression coefficients, on the used regression technique, and on the quality of the data set. In addition, through some examples using the database of the Next Generation of Ground-Motion Attenuation Models project and some currently favored functional forms we study the increment in the seismic hazard produced by the extrapolation of GMPMs. [DOI: 10.1193/1.3525379]

INTRODUCTION

Since the occurrence of small- and moderate-magnitude earthquakes is more frequent than the occurrence of large seismic events, most of ground-motion databases used in the development of GMPMs are primarily comprised of accelerograms produced by small and moderate earthquakes. Hence, as magnitude increases, the sets of ground motions become sparse. For instance, in the database used in the Next Generation of Ground-Motion Attenuation Models project (NGA; Power et al. 2008 and Chiou et al. 2008) there are only five earthquakes with $M_w > 7.5$ and two of them yielded only two recordings per earthquake. For events with $M_w < 6$, the range of distances covered by the data is between 30 km and 200 km, while for large magnitude events the range of distances covered by the data is between 1 and 300 km, although the set is clearly sparse in this range of magnitudes (see Figure 2 in Chiou et al. 2008). Moreover, the database becomes even sparser for frequencies smaller than 0.33 Hz. The same lack of data is observed in other seismic regions. The ground-motion database for the Mexican subduction zone includes only two events with $M_w > 7.5$ and two events with magnitudes between 7 and 7.5. Furthermore, there are no data for M_w between 6.1 and 6.5 and the data become sparse for distances greater than 200 Km (Arroyo et al. 2010).

[a] Universidad Autónoma Metropolitana, Avenida San Pablo #180, Colonia Reynosa Tamaulipas, Azcapotzalco, Mexico City, email: aresda@correo.azc.uam.mx
[b] Instituto de Ingeniería, UNAM, Torre de Ingeniería, Segundo piso, Coyoacan 04510, Mexico City

This lack of data has led to concerns about whether or not the available data are able to constrain the coefficients of the functional form during the regression analysis, concerns about the possible extrapolation of empirical GMPMs in PSHA (Abrahamson and Silva 2008 and Power et al. 2008) and concerns about the application of GMPMs right down to the lower magnitude limit in the datasets (Bommer et al. 2007). Several studies have been made in the past regarding the variability and uncertainty associated to GMPMs (Abrahamson et al. 1991, Atkinson 2006, Bommer and Abrahamson 2006, Purvance et al. 2008, Strasser et al. 2009). Abrahamson et al. (1991) presented a method to model the uncertainty in finite fault simulations. Atkinson (2006) showed that the standard deviation of multiple recordings at single stations is smaller (roughly 10%) than the standard deviation computed from regional recordings; Bommer and Abrahamson (2006) presented a discussion about the correct way to include the variability of GMPMs in PSHA; Purvance et al. (2008) suggested that the inconsistency observed in PSHA computations for precariously balanced rocks at some sites of California may be due to the ergodic assumption, or due to the fact that GMPMs previous to the NGA project yielded too high estimations of median ground motions; and Strasser et al. (2009) presented a review of the current state of knowledge regarding the estimation of the variability in PSHA.

These studies, however, dealt with uncertainty issues that are different from the one we study here. In this paper, we develop a way to quantify how the uncertainty in a GMPM increases due to the fact that its coefficients are not known numbers, but statistical estimates, that might be good or poor depending on the quality of the data. We show that this increase in uncertainty can be particularly large in magnitude-distance regions poorly sampled by the ground-motion data, that is, when a GMPM is extrapolated. But we also show that this uncertainty increase can take place even in relatively well sampled magnitude-distance regions. This additional uncertainty is clearly of epistemic nature, since it would vanish if the ground-motion sample was infinitely large. As we will show later, this issue has been studied in the past and ways have been found to include this extra uncertainty in PSHA. In our view, however, the approach we propose allows for two useful things: 1) taking notice of situations in which GMPMs are not well constrained by data even at magnitude-distance regions for which data are not particularly scarce; and 2) having more solid grounds to find ways to include the extra uncertainty in PSHA.

METHOD

Consider the linear regression model defined in Equation 1:

$$\boldsymbol{Y} = \boldsymbol{X}\alpha + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{Y}$ is a known $n_o \times 1$ matrix which includes n_o observations of a certain measure of seismic intensity (typically the logarithm of a spectral acceleration), $\boldsymbol{X}$ is a known $n_o \times n_p$ matrix which comprises n_o observations of n_p parameters considered in the model, $\boldsymbol{\alpha}$ is an unknown $n_p \times 1$ matrix which comprises the coefficients to be determined by regression analysis, and $\boldsymbol{E}$ is a unknown $n_o \times 1$ matrix which comprises the regression residuals.

Nowadays, the common regression methods used in the development of GMPMs are the one-stage maximum-likelihood method and the two-stage method. The one-stage

maximum-likelihood approach was introduced by Brillinger and Priesler (1984), and later, Abrahamson and Youngs (1992) and Joyner and Boore (1993, 1994) proposed computational algorithms to implement it. In addition, Joyner and Boore (1993, 1994) studied how the one-stage maximum-likelihood method and the two-stage method are related; they found that both methods lead essentially to the same results. In the case of the one-stage maximum-likelihood method, which is the method we will use in the present paper, it is assumed that the elements of $\boldsymbol{E}$ are correlated, normally distributed random variables with zero mean. The correlation between elements of $\boldsymbol{E}$ is defined through an unknown $n_o \times n_o$ matrix $\boldsymbol{\Omega}$, which is defined in Equation 2:

$$\boldsymbol{\Omega} = \boldsymbol{\Phi}\boldsymbol{\Sigma} \tag{2}$$

where $\boldsymbol{\Phi}$ is an unknown $n_o \times n_o$ matrix which accounts for the correlation between the rows of $\boldsymbol{Y}$, while the scalar Σ is the variance of the residuals.

For this model the likelihood of $\boldsymbol{Y}$ is defined in Equation 3:

$$L(\boldsymbol{Y}|\boldsymbol{\alpha}, \Sigma, \boldsymbol{\Phi}, \boldsymbol{X}) \propto \Sigma^{-n_0/2}|\boldsymbol{\Phi}|^{-1/2} \exp\left\{-\frac{1}{2}[\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^T \boldsymbol{\Phi}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})]\right\} \tag{3}$$

where the symbol $\propto$ stands for proportionality, since we have omitted the normalization constant.

Following Joyner and Boore (1993, 1994) we considered that the elements of $\boldsymbol{E}$, say ε_{ij}, can be expressed as the sum of earthquake-to-earthquake variability (ε_e) and record-to-record variability (ε_r). In addition, the following considerations were made:

a) For a given earthquake, the coefficient of correlation between residuals at different sites is equal to γ_e.
b) Residuals related to different earthquakes are independent.

According to these assumptions, the matrix $\boldsymbol{\Phi}$ is a block diagonal matrix:

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\varphi}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\varphi}_{n_e} \end{bmatrix} \tag{4}$$

where n_e is the number of earthquakes and the square submatrix $\boldsymbol{\varphi}_i$ related to earthquake i is given by

$$\boldsymbol{\varphi}_i = \begin{bmatrix} 1 & \gamma_e & \cdots & \gamma_e \\ \gamma_e & 1 & \cdots & \gamma_e \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_e & \gamma_e & \cdots & 1 \end{bmatrix} \tag{5}$$

The rank of $\boldsymbol{\varphi}_i$ is equal to the number of records for earthquake i. Note that γ_e is equal to parameter γ in Joyner and Boore (1993 and 1994).

For a given γ_e the values of $\boldsymbol{\alpha}$ and Σ which maximize the likelihood are the well known weighted least-squares estimators, given by Equations 6 and 7 (Searle 1971, Drapper and Smith 1981, Rowe 2002):

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\boldsymbol{\Phi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Phi}^{-1}\boldsymbol{Y} \tag{6}$$

$$\hat{\Sigma} = \frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}})^T\boldsymbol{\Phi}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}})}{n_0} \tag{7}$$

Furthermore, the variance of the inter-event residual (σ_e^2) is equal to $\gamma_e\hat{\Sigma}$, while the variance of the record-to-record residual (σ_r^2) can be computed from $\hat{\Sigma} = \sigma^2 = \sigma_e^2 + \sigma_r^2$.

In the maximum likelihood method, the value of γ_e which maximizes the likelihood is found iteratively and the final values of $\boldsymbol{\alpha}$ and Σ are the ones related to γ_e of maximum likelihood.

Based on these considerations, the covariance matrix of $\hat{\boldsymbol{\alpha}}$ for a given $\boldsymbol{\Phi}$ is defined by (Searle 1971, Drapper and Smith 1981, Broemeling 1984, Rowe 2002):

$$COV(\hat{\boldsymbol{\alpha}}) = \frac{1}{n_o - n_p - 2}\left[(\boldsymbol{X}^T\boldsymbol{\Phi}^{-1}\boldsymbol{X})^{-1}((\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}})^T\boldsymbol{\Phi}^{-1}(\mathbf{Y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}}))\right] \tag{8}$$

Note that the least squares method is a particular case of the maximum likelihood method. The well-known least squares estimators can be found setting $\gamma_e = 0.0$ (i.e., $\boldsymbol{\Phi} = \mathrm{I}$) in Equations 3 to 8.

Normally, $\hat{\boldsymbol{\alpha}}$ and $\hat{\Sigma}$ related to the maximum likelihood are used to forecast future observations of $\boldsymbol{Y}$ for a given future value of $\boldsymbol{X}$. However, we note that $\hat{\boldsymbol{\alpha}}$ and $\hat{\Sigma}$ are conditioned to the data employed in the analysis (i.e., $\boldsymbol{Y}$ and $\boldsymbol{X}$) and may not be valid to forecast a future value of $\boldsymbol{Y}$ for any given value of $\boldsymbol{X}$. For instance, suppose that the regression analysis is performed over the linear model defined in Equation 1 with only two data points. According to Equation 7, $\hat{\Sigma}$ would be equal to zero. Is this value a rational estimation of the variability of the model even if the model is extrapolated? Of course the answer is no. Certainly, this is an extreme case of application of regression analysis; nevertheless it is useful to qualitatively illustrate that the variability related to a given set of data is not necessarily representative of the variance of a future observation.

A better way to forecast the variance of a future value of $\boldsymbol{Y}$ is through the predictive variance of a future observation, Σ_p (Searle 1971, Drapper and Smith 1981, Rowe 2002):

$$\Sigma_p = \hat{\Sigma} + \mathbf{Z}\,COV(\hat{\boldsymbol{\alpha}})\mathbf{Z}^T \tag{9}$$

where $\hat{\Sigma}$ and $COV(\hat{\boldsymbol{\alpha}})$are defined in Equations 7 and 8, respectively, and $\boldsymbol{Z}$ is a n_p row vector comprised of parameters for which a certain value, w, is being forecasted.

Interestingly, the predictive variance of a forecasted value depends on the variability contained in the database (given by $\hat{\Sigma}$) and on the uncertainty in the regression coefficients

(given by $\mathbf{Z}COV(\hat{\boldsymbol{\alpha}})\mathbf{Z}^T$). If the data are well sampled, then the variance of the regression coefficients will be small and the variance of w, Σ_p, will tend to $\hat{\Sigma}$. On the other hand, for a poorly sampled dataset the variance of the regression coefficients will be large and the variance of w will be larger than $\hat{\Sigma}$.

DATA SET SELECTION

In order to illustrate the implications of the results discussed in the previous section in the forecasting of ground-motion parameters, we present examples using a dataset which includes 906 accelerograms recorded at rock sites during 44 earthquakes of the NGA database. We considered 874 accelerograms recorded at sites with average shear-wave velocity in the upper 30 meters of sediments (V_{S30}) between 450 m/s and 900 m/s and 32 records from sites with V_{S30} between 900 m/s and 1428 m/s. We included those 32 records in view of the fact that they were recorded at sites with NEHRP B classification. We only used records at free field stations and in first floor of buildings with no more than two stories. This dataset is very similar to the one utilized by Idriss (2008) in the NGA project, although ours is quite smaller since we have excluded events that yielded only one record.

We considered two different intensity measures for the regression analysis: PGA and the spectral elastic ordinate at a period of 3 seconds ($SA(T=3)$). For PGA we used 906 accelerograms while for $SA(T=3)$ we only used 458 accelerograms recorded during 28 events since we discarded some recordings based on the minimum useful frequency reported in the NGA database.

FUNCTIONAL FORMS

We selected three functional forms of the NGA project. Firstly, we worked with the functional form adopted by Boore and Atkinson (2008; hereafter referred to as FF1), which is shown in Equation 10:

$$y = F_M(M_w) + F_D(R_{RUP}, M_w) \tag{10}$$

where F_M and F_D are the magnitude scaling and the distance function, respectively, and y is the natural logarithm of the GMRotI50 (Boore et al. 2006) of SA in g units. Functions F_D and F_M are given by:

$$F_D(R_{JB}, M_w) = c_1 \ln\left(\frac{R}{R_{ref}}\right) + c_2\left(M_w - M_{ref}\right)\ln\left(\frac{R}{R_{ref}}\right) + c_3\left(R - R_{ref}\right) \tag{11}$$

$$F_M(M_w) = \begin{cases} e_2 SS + e_3 NS + e_4 RS + e_5(M_w - M_h) + e_6(M_w - M_h)^2 & \text{if } M_w \le M_h \\ e_2 SS + e_3 NS + e_4 RS + e_7(M_w - M_h) & \text{otherwise} \end{cases} \tag{12}$$

where

$$R = \sqrt{R_{RUP}^2 + h^2} \tag{13}$$

and c_1, c_2, c_3, e_2, e_3, e_4, e_5, e_6, e_7, and h are free coefficients to be defined by regression analysis. SS, NS, and RS are dummy variables used to denote strike-slip, normal-slip, and reverse-slip fault type, and M_{ref}, R_{ref}, and M_h are coefficients to be set in the analysis

Secondly, we considered the functional form adopted by Abrahamson and Silva (2008) (hereafter referred to as FF2) which is defined in Equation 14:

$$y = f_1(M_w, R_{RUP}) + a_{12}F_{RV} + a_{13}F_{NM} + f_8(R_{RUP}, M_w) \tag{14}$$

where $f_1(M_w, R)$ is a modeling term for the magnitude and distance dependence for strike-slip events, defined in Equation 15, while F_{RV} and F_{NM} are dummy variables used to denote reverse-slip and normal-slip fault type, respectively. $f_8(R_{RUP}, M_w)$ is a large-distance attenuation term given by Equations 16 and 17, and a_{12}, a_{13}, a_1, a_4, a_8, a_2, a_3, c_1, and a_{18} are free coefficients to be determined by regression analysis.

$$f_1(M_w, R_{RUP}) = \begin{cases} a_1 + a_4(M_w - c_1) + a_8(8.5 - M_w)^2 + [a_2 + a_3(M - c_1)]\ln(R) & \text{If } M_w \le c_1 \\ a_1 + a_5(M_w - c_1) + a_8(8.5 - M_w)^2 + [a_2 + a_3(M - c_1)]\ln(R) & \text{otherwise} \end{cases} \tag{15}$$

$$f_8(R_{RUP}, M_w) = \begin{cases} 0 & \text{If } R_{RUP} < 100 \\ a_{18}(R_{RUP} - 100)T_6(M_w) & \text{otherwise} \end{cases} \tag{16}$$

$$T_6(M_w) = \begin{cases} 1 & \text{If } M_w < 5.5 \\ 0.5(6.5 - M_w) + 0.5 & \text{If } 5.5 \le M_w \le 6.5 \\ 0.5 & \text{otherwise} \end{cases} \tag{17}$$

Finally, we used the functional form adopted by Campbell and Bozorgnia (2008) (hereafter referred to as FF3) for the NGA project, which is defined in Equation 18:

$$y = f_{mag} + f_{dis} + f_{flt} \tag{18}$$

where f_{mag}, f_{dis}, and f_{flt} are the magnitude, the distance and the fault type terms, respectively, defined in Equations 19 to 21:

$$f_{mag} = \begin{cases} c_0 + c_1 M_w & \text{if } M_w \le 5.5 \\ c_0 + c_1 M_w + c_2(M_w - 5.5) & \text{if } 5.5 < M_w \le 6.5 \\ c_0 + c_1 M_w + c_2(M_w - 5.5) + c_3(M_w - 6.5) & \text{otherwise} \end{cases} \tag{19}$$

$$f_{dis} = (c_4 + c_5 M_w)\ln\left(\sqrt{R_{RUP}^2 + c_6^2}\right) \tag{20}$$

$$f_{flt} = c_7 F_{RV} + c_8 F_{NM} \tag{21}$$

In Equations 19 to 21 F_{RV} and F_{NM} are dummy variables used to denote reverse-slip events and normal-slip events, respectively, while c_0 to c_8 are free coefficients to be defined by regression analysis.

In order to facilitate the comparisons, we performed the analysis only for rock sites. Therefore, we removed from the original functional forms the terms related to site amplification, hanging-wall effect, aftershock events and basin response effect. In addition, in the

case of the Boore and Atkinson (2008) model we used as distance parameter R_{RUP} instead of R_{JB}. As a result of the modifications, the predictor variables are the same for the three functional forms.

REGRESSION ANALYSIS

For each functional form we performed the regression analysis using the one-stage maximum-likelihood method (Joyner and Boore 1993 and 1994). In principle, FF1 has 11 free coefficients to be determined by regression analysis. However, the information contained in the dataset was not enough to properly constrain all coefficients (Boore and Atkinson 2008). Hence, in all the analysis presented, coefficients c_3 and h were fixed to the values of the Boore and Atkinson (2008) model. In this model those values were fixed using data recorded during three small California events included in the NGA database and additional data from broadband accelerometers (further details can be found in Boore and Atkinson 2008). Therefore, we set $M_{ref} = 4.5$, $R_{ref} = 1$, and $M_h = 6.75$ for FF1. Also, in the case of PGA we set $c_3 = -0.01151$, $e_7 = 0$, $h = 1.35$ while for $SA(T = 3)$ we set $c_3 = -0.00191$, and $h = 2.83$. Hence, there are seven and eight free coefficients for PGA and $SA(T = 3)$, respectively.

In order to stabilize the regression analysis, several regression coefficients were also fixed in Abrahamson and Silva's (2008) model. Accordingly, for FF2 we set $c_1 = 6.75$, $c_4 = 4.5$, and $a_5 = -0.398$. Also, for PGA we set $a_{18} = -0.0067$ while for $SA(T = 3)$ we set $a_{18} = 0$, and the values of a_3 and a_4 were constrained to those values obtained during the regression analysis for PGA. Thus, for FF2 there are seven and five free coefficients for PGA and $SA(T = 3)$, respectively.

The functional form adopted by Campbell and Bozorgnia (2008) has nine regression coefficients. In this case, we fixed the value of c_6 to 5.60 and 4 for PGA and $SA(T = 3)$, in order to use a magnitude term similar to the one used in the Campbell and Bozorgnia (2008) model. Hence, for FF3 there are eight free coefficients.

Results of the regression analyses are presented in Table 1. In general, the level of accuracy of the three GMPMs is similar, judging from the computed σ values, which are in the range from 0.659 to 0.728. In Figure 1 we compare the predicted median values related to each functional form for strike-slip events. The larger differences between GMPMs are observed for large values of M_w and for short and large R_{Rup} values, which is not surprising since the datasets tend to become sparse in these magnitude-distance regions. Although not shown, similar trends were observed for other fault types.

However, according to Equation 9, the uncertainty in the regression coefficients increases the overall uncertainty in a GMPM. This increase can be estimated using Equation 9 and the covariance matrix of the regression coefficients, reported in Tables 2 to 4. As has already been stated, this extra uncertainty is of epistemic nature. Its size can be assessed comparing Σ_p and $\hat{\Sigma}$ through parameter s, defined as follows:

$$s = \sqrt{\frac{\Sigma_p}{\hat{\Sigma}}} \tag{22}$$

In Figures 2 and 3 we present comparisons of s contours for each functional form, for strike-slip events, and PGA and $SA(T = 3)$, respectively. For reference, in these figures we

Table 1. Results of the regression analyses

FF1			FF2			FF3		
	PGA	$SA(T=3)$		PGA	$SA(T=3)$		PGA	$SA(T=3)$
c_1	−9.748E-01	−1.085E + 00	a_1	1.095E + 00	−4.303E-01	c_0	2.472E + 00	−1.179E + 01
c_2	1.859E-01	1.596E-01	a_4	2.127E-01	–	c_1	−4.780E-02	1.692E + 00
e_2	−3.387E-02	−1.126E + 00	a_8	1.023E-01	−1.424E-01	c_2	−5.039E-01	4.453E-03
e_3	−2.159E-01	−1.235E + 00	a_2	−1.100E + 00	−8.713E-01	c_3	−4.726E-02	−1.930E + 00
e_4	1.074E-01	−1.029E + 00	a_3	2.428E-01	–	c_4	−2.856E + 00	−1.931E + 00
e_5	−2.528E-01	4.690E-01	a_{12}	2.523E-01	3.070E-01	c_5	2.502E-01	1.631E-01
e_6	−8.017E-02	−5.256E-01	a_{13}	−1.478E-01	5.929E-02	c_7	2.468E-01	1.551E-01
e_7	–	−9.664E-02	σ_e	0.384	0.426	c_8	−1.343E-01	−2.748E-02
σ_e	0.428	0.436	σ_r	0.536	0.503	σ_e	0.384	0.424
σ_r	0.547	0.595	σ	0.659	0.659	σ_r	0.535	0.591
σ	0.695	0.737				σ	0.659	0.728

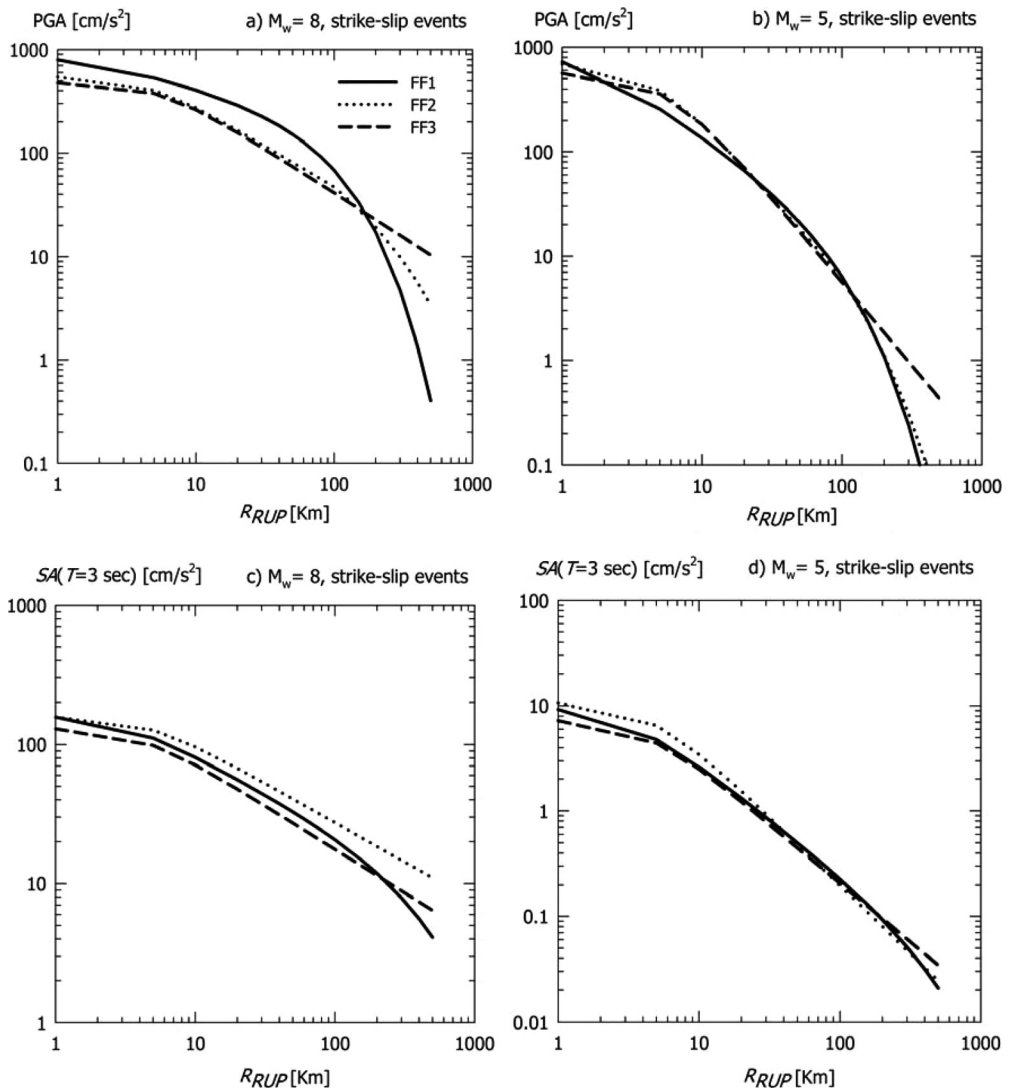

Figure 1. Comparison of different GMPMs.

have plotted with black circles the data used in the regression analysis. We observe similar s contours for the three GMPMs. An increment of s is observed for M_w values that are outside the range covered by the dataset, especially in the small magnitude range. On the other hand, the value of s remains constant for R_{RUP} values that are outside the range covered by the database. Therefore, for the presented examples, the extra uncertainty associated to distance coefficients is smaller than the extra uncertainty related to the magnitude coefficients.

For PGA, the smallest s values are observed for FF2, while FF1 and FF3 lead to similar s contours. Values of s larger than 1.1 are observed for $M_w < 4.5$ and $M_w > 8$. The peak

Table 2. Covariance matrixes associated to FF1

PGA	c_1	c_2	e_2	e_3	e_4	e_5	e_6	
c_1	6.216E-03	−2.403E-03	3.456E-04	1.504E-05	−1.439E-04	1.460E-02	2.430E-03	
c_2	−2.403E-03	1.059E-03	−1.440E-03	−1.168E-03	−1.097E-03	−6.315E-03	−9.939E-04	
e_2	3.456E-04	−1.440E-03	3.990E-02	2.652E-02	2.155E-02	3.279E-02	8.583E-03	
e_3	1.504E-05	−1.168E-03	2.652E-02	6.437E-02	2.450E-02	4.337E-02	1.303E-02	
e_4	−1.439E-04	−1.097E-03	2.155E-02	2.450E-02	3.474E-02	4.572E-02	1.838E-02	
e_5	1.460E-02	−6.315E-03	3.279E-02	4.337E-02	4.572E-02	1.872E-01	8.173E-02	
e_6	2.430E-03	−9.939E-04	8.583E-03	1.303E-02	1.838E-02	8.173E-02	4.305E-02	
$SA(T=3)$	c_1	c_2	e_2	e_3	e_4	e_5	e_6	e_7
c_1	2.634E-02	−9.430E-03	−1.968E-02	−1.703E-02	−1.858E-02	3.755E-02	1.839E-04	3.730E-02
c_2	−9.430E-03	3.575E-03	5.346E-03	4.699E-03	5.226E-03	−1.370E-02	1.683E-04	−1.433E-02
e_2	−1.968E-02	5.346E-03	9.416E-02	6.320E-02	5.658E-02	5.104E-02	2.254E-02	−9.049E-02
e_3	−1.703E-02	4.699E-03	6.320E-02	1.287E-01	5.470E-02	7.788E-02	3.628E-02	−7.276E-02
e_4	−1.858E-02	5.226E-03	5.658E-02	5.470E-02	7.154E-02	6.661E-02	3.727E-02	−7.472E-02
e_5	3.755E-02	−1.370E-02	5.104E-02	7.788E-02	6.661E-02	3.887E-01	1.568E-01	−6.683E-02
e_6	1.839E-04	1.683E-04	2.254E-02	3.628E-02	3.727E-02	1.568E-01	8.071E-02	−4.626E-02
e_7	3.730E-02	−1.433E-02	−9.049E-02	−7.276E-02	−7.472E-02	−6.683E-02	−4.626E-02	2.297E-01

Table 3. Covariance matrixes associated to FF2

PGA	a_1	a_4	a_8	a_2	a_3	a_{12}	a_{13}
a_1	1.284E-01	−2.255E-01	−4.395E-02	−3.529E-03	−4.912E-03	−1.731E-02	−9.076E-03
a_4	−2.255E-01	6.577E-01	1.143E-01	5.637E-04	4.459E-03	−5.642E-03	−7.456E-03
a_8	−4.395E-02	1.143E-01	2.083E-02	1.377E-04	1.640E-03	−8.591E-05	−1.616E-03
a_2	−3.529E-03	5.637E-04	1.377E-04	8.127E-04	4.169E-05	2.778E-04	2.907E-04
a_3	−4.912E-03	4.459E-03	1.640E-03	4.169E-05	1.180E-03	5.685E-04	2.782E-04
a_{12}	−1.731E-02	−5.642E-03	−8.591E-05	2.778E-04	5.685E-04	2.459E-02	1.299E-02
a_{13}	−9.076E-03	−7.456E-03	−1.616E-03	2.907E-04	2.782E-04	1.299E-02	4.315E-02
$SA(T=3)$	a_1	a_8	a_2	a_{12}	a_{13}		
a_1	8.385E-02	−4.590E-03	−6.320E-03	−4.236E-02	−2.900E-02		
a_8	−4.590E-03	7.344E-04	−1.860E-05	1.594E-03	−5.917E-04		
a_2	−6.320E-03	−1.860E-05	1.588E-03	7.789E-04	9.079E-04		
a_{12}	−4.236E-02	1.594E-03	7.789E-04	5.689E-02	2.825E-02		
a_{13}	−2.900E-02	−5.917E-04	9.079E-04	2.825E-02	1.116E-01		

Table 4. Covariance matrixes associated to FF3

PGA	c_0	c_1	c_2	c_3	c_4	c_5	c_7	c_8
c_0	6.009E + 00	−1.133E + 00	1.438E + 00	−5.492E-01	−2.293E-01	3.380E-02	9.328E-02	1.153E-02
c_1	−1.133E + 00	2.163E-01	−2.840E-01	1.095E-01	3.425E-02	−5.125E-03	−2.063E-02	−5.624E-03
c_2	1.438E + 00	−2.840E-01	4.486E-01	−2.306E-01	−4.495E-03	6.692E-04	2.218E-02	1.080E-02
c_3	−5.492E-01	1.095E-01	−2.306E-01	2.379E-01	7.269E-03	−1.057E-03	3.119E-03	−5.061E-04
c_4	−2.293E-01	3.425E-02	−4.495E-03	7.269E-03	6.106E-02	−8.980E-03	−8.314E-04	−2.066E-04
c_5	3.380E-02	−5.125E-03	6.692E-04	−1.057E-03	−8.980E-03	1.339E-03	1.669E-04	7.568E-05
c_7	9.328E-02	−2.063E-02	2.218E-02	3.119E-03	−8.314E-04	1.669E-04	2.674E-02	1.412E-02
c_8	1.153E-02	−5.624E-03	1.080E-02	−5.061E-04	−2.066E-04	7.568E-05	1.412E-02	4.371E-02
$SA(T=3)$	c_0	c_1	c_2	c_3	c_4	c_5	c_7	c_8
c_0	1.509E + 01	−2.847E + 00	3.756E + 00	−1.570E + 00	−7.335E-01	1.028E-01	2.116E-01	2.244E-01
c_1	−2.847E + 00	5.480E-01	−7.641E-01	3.255E-01	1.022E-01	−1.448E-02	−4.624E-02	−5.249E-02
c_2	3.756E + 00	−7.641E-01	1.293E + 00	−6.739E-01	5.819E-04	1.815E-04	4.905E-02	8.546E-02
c_3	−1.570E + 00	3.255E-01	−6.739E-01	5.453E-01	−1.072E-04	−1.110E-04	1.116E-02	−2.199E-02
c_4	−7.335E-01	1.022E-01	5.819E-04	−1.072E-04	1.868E-01	−2.600E-02	1.543E-03	5.538E-03
c_5	1.028E-01	−1.448E-02	1.815E-04	−1.110E-04	−2.600E-02	3.648E-03	−1.011E-04	−6.312E-04
c_7	2.116E-01	−4.624E-02	4.905E-02	1.116E-02	1.543E-03	−1.011E-04	5.151E-02	2.946E-02
c_8	2.244E-01	−5.249E-02	8.546E-02	−2.199E-02	5.538E-03	−6.312E-04	2.946E-02	9.697E-02

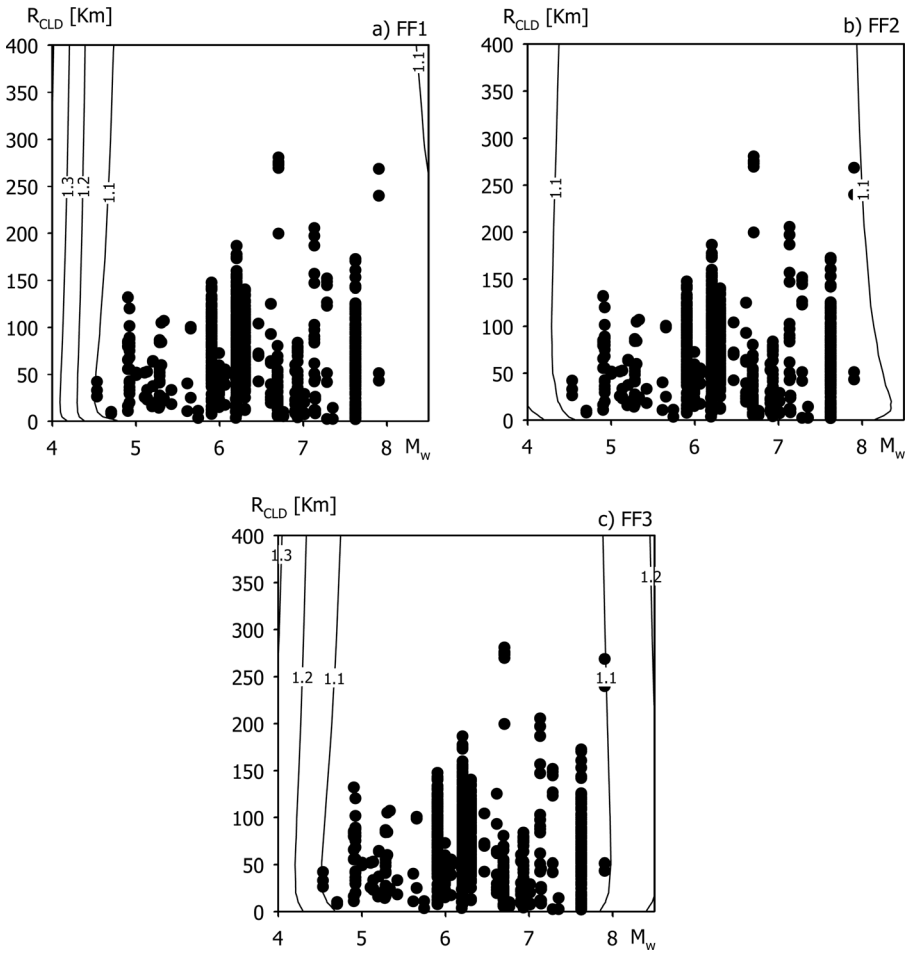

Figure 2. Comparison of s contours for PGA.

s values are roughly 1.3 for M_w close to 4. Hence, we conclude that data were able to properly constrain the regression analysis with FF2 for the magnitude-distance regions shown in Figure 2, and, in the case of FF1 and FF3, for the range between $4.5 < M_w < 8$. In the NGA database, data become sparse as period increases, since the processing of recordings affects their usable bandwidth. Therefore, the number of data points decreases for $SA(T = 3)$ and larger s values are observed than those related to PGA for FF1 and FF3. For FF2, s values are nearly the same for both intensity measures since for $SA(T = 3)$ the number of free coefficients was reduced. Values of s larger than 1.1 are observed for $M_w < 5$ and $M_w > 7.5$. The peak s values are roughly 1.5 for M_w close to 4. Hence, we conclude that data were enough to properly constrain the regression analysis with FF2 for the ranges of M_w and R_{RUP} shown in Figure 3 and, in the case of FF1 and FF3, for the range $5 < M_w < 7.5$ and $R_{RUP} > 25$ km. The increase in the uncertainty when the GMPMs are extrapolated is evident, particularly

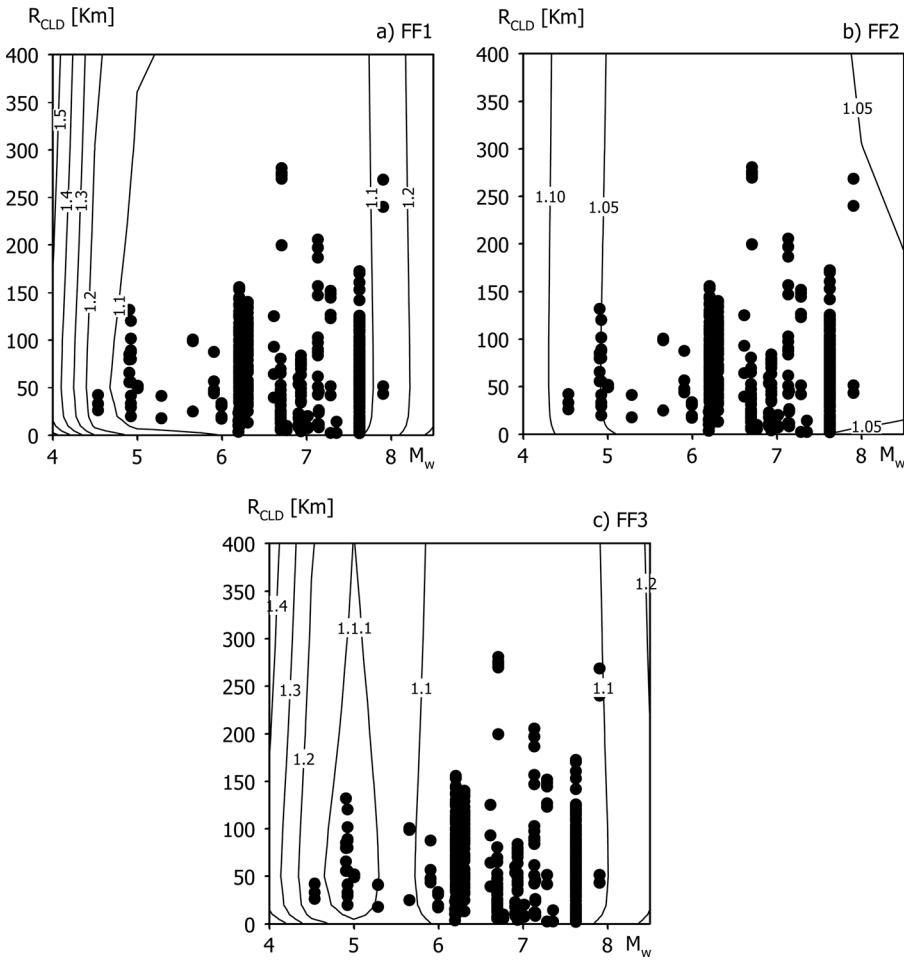

Figure 3. Comparison of s contours for $SA(T = 3)$.

for FF1 and FF3. Interestingly, this increment is observed even in magnitude-distance ranges covered relatively well by the dataset (see Figure 3c).

PSHA COMPUTATIONS

So far, we have followed the standard procedure in regression analysis, and normally the results presented in Table 1 would be used in PSHA computations. However, judging from the s contours, it is clear that there are magnitude-distance regions for which the models are not as reliable as they are for other regions, since the predictive variance Σ_p is greater than the common variance $\hat{\Sigma}$. For these regions, the effect of the extra epistemic uncertainty should be introduced, especially since, in general, an increment in the variance of a GMPM leads to an increment in the seismic hazard (i.e., a larger rate of exceedance for a given

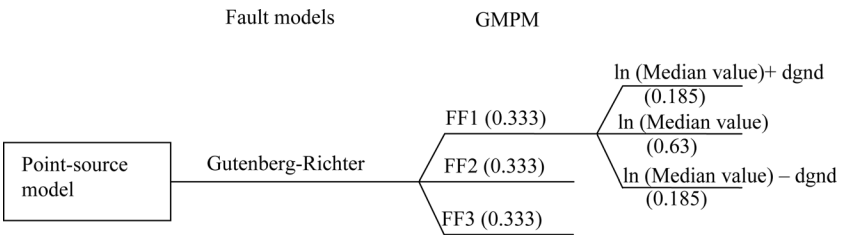

Figure 4. Logic tree for the PSHA procedure used by Petersen et al. (2008).

intensity value). This increment in the hazard level may be viewed as a consequence of the extrapolation of the ground-motion model and of the quality of the data used in the regression analysis.

In the past, this extra uncertainty has been included in PSHA using various approaches. An example is the procedure used in the development of the 2008 United States National Seismic Hazard Maps (Petersen et al. 2008), which is sketched in Figure 4. In this procedure, the additional epistemic uncertainty was included in PSHA in order to take into account the data limitations for large earthquakes. The additional epistemic uncertainty was modeled with the inclusion of extra branches for a given GMPM in the logic tree used for PSHA, as shown in Figure 4. For a given GMPM, three branches were added to the logic tree. The first branch, to which a weight $w_1 = 0.633$ was assigned, used as median value the one predicted by the GMPM. The second and third branches were assigned weights $w_2 = w_3 = 0.185$, and they used as logarithm of the median value the one predicted by the GMPM plus/minus a factor (*dgnd*). The value of *dgnd* was set depending on the magnitude-distance bin, and it was assumed to be equal to 0.4 (based on a 90 percent confidence limits) for the $M_w \geq 7$, $R_{RUP} < 10$ km- bin (Petersen et al. 2008). For other bins, *dgnd* was computed according to the square root of the ratio between the number of records in the considered bin and the number of records in the $M_w \geq 7$, $R_{RUP} < 10$ km- bin (Petersen et al. 2008). Note that the structure of the logic tree in Figure 4 is very simple since the present study deals only with the uncertainty in the GMPMs.

However, it can be shown that constructing an N-branch logic tree for a given GMPM, as was done by Petersen et al. (2008), amounts to using a final probability density function for the intensities equal to the weighted sum of N lognormal probability density functions:

$$p(SA) = \sum_{i=1}^{N} \frac{w_i}{\sqrt{2\pi}\sigma SA} e^{-\frac{1}{2}\frac{(\ln(SA)-\ln(m_i))^2}{\sigma^2}} \tag{23}$$

where w_i and m_i are the weighting factor and the median related to the ith branch, respectively. We note that, as done by Petersen et al. (2008), we have used in Equation 23 the same σ value in all branches. Interestingly, it can be demonstrated that, provided that the weights w_i assigned to the different values of m_i are roughly proportional to a lognormal density function, and N is large, then $p(SA)$ in Equation 23 would itself be also a lognormal probability density function.

In other words, the inclusion of the extra epistemic uncertainty has transformed the probability density function of a given intensity from a simple lognormal function to the density given in Equation 23. This means that, regarding only this particular source of uncertainty, hazard could have been computed either with a three-branch logic tree, as done by Petersen et al. (2008), or with a single branch in which the probability density function is given by Equation 23 with the appropriate values for w_i and m_i. Even if the later procedure is unusual and most PSHA computer codes are not prepared to handle it, results would have been exactly the same.

So we have shown that the effect of epistemic uncertainty, as treated by Petersen et al. (2008), is to change the original probability distribution of a given intensity to a new one that, in general, has larger uncertainties associated that are quantified through parameters w_i and m_i. It can be demonstrated that the first and the second moments of the final probability density function given in Equation 23 are:

$$M_1 = \int_0^\infty SA\, p(SA) dSA = e^{\frac{1}{2}\sigma^2} \sum_{i=1}^{N} w_i m_i \tag{24}$$

$$M_2 = \int_0^\infty SA^2\, p(SA) dSA = e^{2\sigma^2} \sum_{i=1}^{N} w_i m_i^2 \tag{25}$$

It is possible to find a lognormal probability density function that is equivalent–in the second-moment sense- to the one given in Equation 23, by way of making its first two moments equal to the moments defined in Equations 24 and 25. The moments of an equivalent lognormal density function with median m_e and logarithmic standard deviation σ_e, would be given by:

$$\hat{M}_1 = m_e e^{\frac{1}{2}\sigma_e^2} \tag{26}$$

$$\hat{M}_2 = m_e^2 e^{2\sigma_e^2} \tag{27}$$

Making $M_1 = \hat{M}_1$ and $M_2 = \hat{M}_2$ and solving for σ_e and m_e, we find that:

$$\sigma_e^2 = \sigma^2 + \ln\left[\frac{\sum_{i=1}^{N} w_i m_i^2}{\left(\sum_{i=1}^{N} w_i m_i\right)^2}\right] \tag{28}$$

$$m_e = \frac{\left(\sum_{i=1}^{N} w_i m_i\right)}{\sqrt{\sum_{i=1}^{N} w_i m_i}} \tag{29}$$

Table 5. Inferred s values in Petersen et al. (2008)

M_w and R_{RUP} range	*dgnd*	S
$5 \leq M_w < 6$, $R_{RUP} < 10$	0.375	1.06
$5 \leq M_w < 6$, $10 \leq R_{RUP} < 30$	0.21	1.02
$5 \leq M_w < 6$, $R_{RUP} \geq 30$	0.245	1.02
$6 \leq M_w < 7$, $R_{RUP} < 10$	0.23	1.02
$6 \leq M_w < 7$, $10 \leq R_{RUP} < 30$	0.225	1.02
$6 \leq M_w < 7$, $R_{RUP} \geq 30$	0.23	1.02
$M_w >= 7$, $R_{RUP} < 10$	0.4	1.06
$M_w >= 7$, $10 \leq R_{RUP} < 30$	0.36	1.05
$M_w >= 7$, $R_{RUP} \geq 30$	0.31	1.04

Note that Equation 28 links the final uncertainty resulting from the procedure used by Petersen et al. (2008), that is, σ_e, with Σ_p. Using Equation 28, we inferred the s values implicitly used by Petersen et al. (2008) for PGA. Results go from 1.02 to 1.06, depending on the magnitude-distance region. As shown in Table 5, these values are similar to s values shown in Figure 2.

Note also the similarity between Equations 28 and 9. In Equation 9, the uncertainty in σ^2 values is accounted for using the covariance matrix of the regression parameters, while in Equation 28, the weights and median values assigned to the branches produce an augmented σ value. In practice, thus, construction of an N-branch logic tree to account for the extra uncertainty we have discussed in this paper is, at least at the second-moment level, equivalent to using a single probability density function for the intensities that has a larger σ which can be computed from Equation 9.

Based on these observations, we propose another method to include the epistemic uncertainty derived from uncertainty in the estimation of the regression coefficients. This method consists, simply, in using Σ_p instead of the common variance $\hat{\Sigma}$ in PSHA computations. As we have seen, the use of a single probability distribution for the intensity instead of constructing branches of a logic tree is justified by the fact that it is always possible to find a single distribution that, in the second-moment sense, is equivalent to the one implicitly used in the logic tree.

In order to illustrate the effect of the use of Σ_p in PSHA, we computed seismic hazard curves considering a point-source model for fixed values of R_{RUP}. For the computations we assumed that the distribution of M_w can be described by a modified Gutenberg-Richter curve with parameters $\beta = 2$, $M_{wmin} = 4$, $M_{wmax} = 8$, and $\lambda_0 = 1$. We considered the three GMPMs together with the logic tree shown in Figure 5. In Figures 6 and 7, we plotted with continuous line the mean hazard curves computed with the value of Σ_p for each M_w and R_{RUP} combination (i.e., considering the uncertainty in the regression coefficients) and with dashed line the mean hazard curves computed with $\hat{\Sigma}$ (i.e., disregarding the uncertainty in the regression coefficients).

As expected, the hazard level related to Σ_p is larger than that computed using $\hat{\Sigma}$. However, the increment in the hazard level is small except for PGA and $R_{\mathrm{RUP}} = 10$ km where

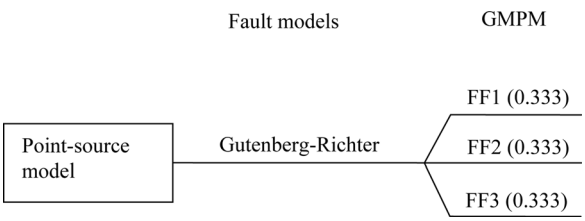

Figure 5. Logic tree used in the PSHA computations for hazard curves shown in Figures 6 and 7.

increments of 18%, 21%, 23%, 25%, and 30% in PGA for rates of exceedance of 0.01, 0.005, 0.002, 0.001, and 0.0001, respectively, are observed. The increment in the hazard curve observed in Figure 6a is produced by the additional epistemic uncertainty for small magnitude events at short distances. Conversely, at large distances this increment is not observed since the hazard level is mainly controlled by large events, so the extra uncertainty for small magnitudes has no significant effect in the hazard level. In spite of the fact that larger values of s were observed for $SA(T=3)$ than for PGA, the effect of the use of Σ_P is more pronounced in PGA. For a given rate of exceedance, peak increments of roughly 10% were observed in the case of $SA(T=3)$ while peak increments of 30% were observed in PGA. The reason of this trend can be explained as follows: large s values for $SA(T=3)$ were observed for small magnitude events, which have a relatively small contribution to the hazard level for low frequency oscillators. On the other hand, the small magnitude events contribute more to the hazard level in the high frequency range.

The examples presented show that it is difficult to define *a priori* the effect on the hazard curves of the additional uncertainty we have discussed here. Hence, the trends observed are valid only for these examples. In practice, similar analysis can be performed in order to assess the extra uncertainty related to the estimation of regression coefficients.

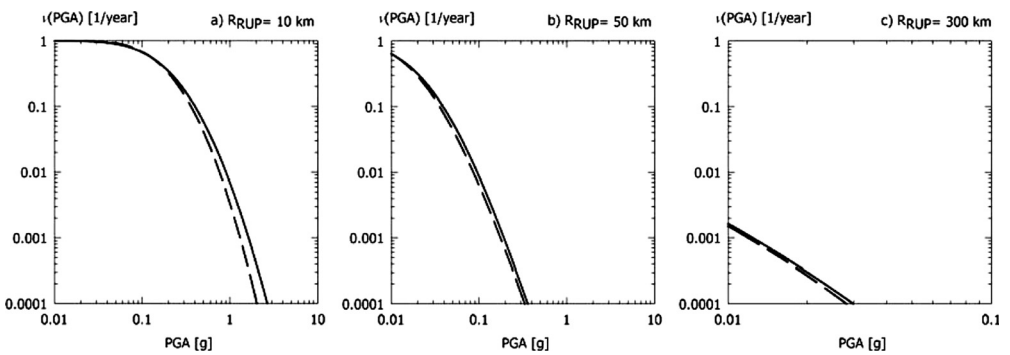

Figure 6. Comparison of mean hazard curves for PGA. The continuous line is the mean hazard curve computed considering the uncertainty in the regression coefficients while the dashed line is the mean hazard curve computed disregarding this uncertainty.

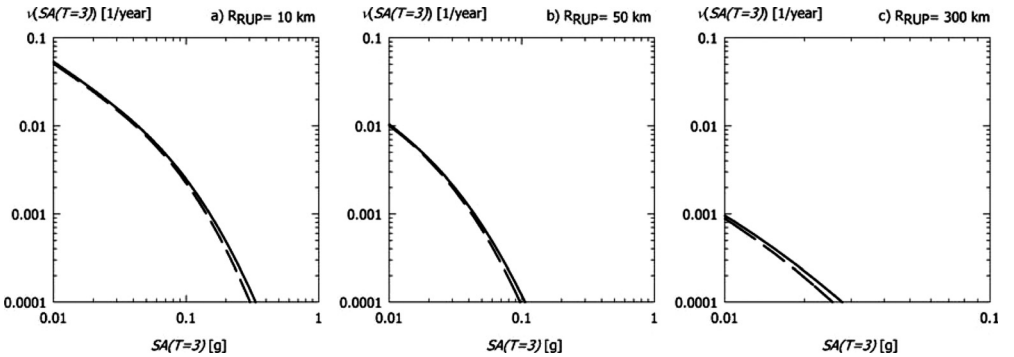

Figure 7. Comparison of mean hazard curves for $SA(T=3)$. The continuous line is the mean hazard curve computed considering the uncertainty in the regression coefficients while the dashed line is the mean hazard curve computed disregarding this uncertainty.

CONCLUSIONS

We have shown that the use of the predictive variance associated to a GMPM allows for the quantitative assessment, in the framework of PSHA, of the possible consequences of uncertainties in the estimation of the coefficients that constitute the GMPM.

These consequences are normally more important when extrapolating the GMPM or when using it for magnitude-distance ranges poorly sampled by the databases used to construct the model. In these cases, there is an additional epistemic uncertainty that must be accounted for, and that yields, frequently, to an increment in the hazard level for a given intensity value.

This additional uncertainty can be evaluated using the predictive variance, Σ_p (see Equation 9), which depends on the sample, on the regression technique, and on the adopted functional form. In the current practice of PSHA, the second term in Equation 9 has frequently been ignored because its size is assumed to be small, although it is rarely, if ever, measured. However, in this paper we have presented a real example (Petersen et al. 2008) in which this uncertainty has been accounted for, as well as numerical examples that show the consequences of ignoring it in PSHA.

We have proposed a method to account for the extra epistemic uncertainty derived from poor coefficient estimation. The method consists, simply, in using Σ_p instead of the common variance $\hat{\Sigma}$ (see Equation 9) in PSHA computations. As we have shown, the use of a single probability distribution for the intensity instead of constructing branches of a logic tree is justified by the fact that it is always possible to find a single distribution that, in the second-moment sense, is equivalent to the one implicitly used in the logic tree.

Although the evaluation of current GMPM has been mostly based on the variance of the residuals and on the predicted median values, we believe, based on the results presented, that the evaluation should also take into account the covariance matrix of the regression coefficients. Unfortunately, this information is not available for most existing GMPMs.

Also, plots of *s* contours, whose construction presents no particular numerical problems, could become a standard practice when deriving GMPMs, in order to give users quantitative indications as to the range of applicability of the model and the consequences of its extrapolation.

ACKNOWLEDGMENTS

Thorough and clever reviews by Professor Julian Bommer and two anonymous reviewers were very helpful to greatly improve the original version of this manuscript.

REFERENCES

Abrahamson, N. A.,Somerville, P. G., and Cornell, A. C., 1991. Uncertainty in numerical strong motion predictions, *Proc. of the 4th US National Conference on Earthquake Engineering* **1**, 407–416.

Abrahamson, N. A., and Youngs, R. R., 1992. A stable algorithm for regression analysis using the random effects model, *Bull. Seism. Soc. Am.* **82**, 505–510.

Abrahamson, N. A., and Silva, W., 2008. Summary of the Abrahamson & Silva NGA ground-motion relations, *Earthquake Spectra* **24**, 67–97.

Atkinson, G. A., 2006. Single-station sigma, *Bull. Seism. Soc. Am.* **96**, 446–455.

Arroyo, D., Ordaz, M., García, D., Mora, M., and Singh, S. K., 2010. Strong ground-motion relations for Mexican interplate earthquakes, *J. Seismol*. **14**, 769–785.

Bommer, J., and Abrahamson, N. A., 2006. Why do modern probabilistic seismic-hazard analyses often lead to increased hazard estimates? *Bull. Seism. Soc. Am.* **96**, 1967–1977.

Bommer, J., Stafford, P. J., Alarcón J. E., and Akkar, S., 2007. The influence of magnitude range on empirical ground-motion prediction, *Bull. Seism. Soc. Am.* **97**, 2152–2170.

Boore, D. M., and Atkinson, G. M., 2008. Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10 s, *Earthquake Spectra* **24**, 99–138.

Boore, D. M., Watson-Lamprey, J., and Abrahamson, N. A., 2006. Orientation-independent measures of ground motion, *Bull. Seism. Soc. Am.* **96**, 1502–1511.

Brillinger, D. R., and Preisler, H. K., 1984. An exploratory analysis of the Joyner-Boore attenuation data, *Bull. Seism. Soc. Am.* **74**, 1441–1450.

Broemeling, L. D., 1984. *Bayesian Analysis of Linear Models*, Marcel Dekker, Inc., New York, 454 pp.

Campbell, K. W., and Bozorgnia, Y., 2008. NGA ground-motion model for the geometric mean horizontal component of PGA, PGV, PGD, and 5% damped linear elastic response spectra for periods ranging from 0.01 s and 10 s, *Earthquake Spectra* **24**, 139–171.

Chiou, B., Darragh, R., Gregor, N., and Silva, W., 2008. NGA Project strong-motion database, *Earthquake Spectra* **24**, 23–44.

Drapper, N.R., and Smith, H., 1981. *Applied Regression Analysis*, 2nd Edition, Wiley, New York, 709 pp.

Idriss, I. M., 2008. An NGA empirical model for estimating the horizontal spectral values generated by shallow crustal earthquakes, *Earthquake Spectra* **24**, 217–242.

Joyner, W. B., and Boore, D. M., 1993. Methods for regression analysis of strong-motion data, *Bull. Seism. Soc. Am.* **83**, 469–487.

Joyner, W. B., and Boore, D. M., 1994. Errata: Methods for regression analysis of strong-motion data, *Bull. Seism. Soc. Am.* **84**, 955–956.

Petersen, M. D., Frankel, A. D., Harmsen, S. C., Mueller, C. S., Haller, K. M., Wheeler, R. L., Wesson, R. L., Zeng, Y., Boyd, O. S., Perkins, D. M., Luco, N., Field, E. H., Wills, C. J., and Rukstales, K. S., 2008. *Documentation for the 2008 Update of the United States National Seismic Hazard Maps*, Open-File Report 2008-1128, U.S. Geological Survey.

Power, M., Chiou, B., Abrahamson, N., Bozorgnia, Y., Shantz, T., and Roblee, C., 2008. An overview of the NGA project, *Earthquake Spectra* **24**, 3–21.

Purvance, M. D., Brune J. N., Abrahamson, N. A., and Anderson G. A., 2008. Consitency of precariously balanced rocks with probabilistic seismic hazard estimates in Southern California, *Bull. Seism. Soc. Am.* **98**, 2629–2640.

Rowe, D. B., 2002. *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*, Chapman & Hall/CRC, New York, 329 pp.

Searle, S. R., 1971. *Linear Models*, Wiley, New York, 532 pp.

Strasser, F. O., Abrahamson, N. A., and Bommer, J., 2009. Sigma: Issues, insights, and challenges, *Seismological Research Letters* **80**, 40–56.

(Received 12 March 2009; accepted 7 June 2010)