EARTHQUAKE ENGINEERING AND STRUCTURAL DYNAMICS
Earthquake Engng Struct. Dyn. 2006; **35**:969–987
Published online 10 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/eqe.565

Modelling peak accelerations from earthquakes

Debbie J. Dupuis[1,*,†] and Joanna Mills Flemming[2,‡]

[1]*Department of Management Sciences, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal, Que., Canada H3T 2A7*
[2]*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1*

SUMMARY

This paper deals with the prediction of peak horizontal accelerations with emphasis on seismic risk and insurance concerns. Non-linear mixed effects models are used to analyse well-known earthquake data and the consequences of mis-specifying assumptions on the error term are quantified. A robust fit of the usual model, using recently developed robust weighted maximum likelihood estimators, is presented. Outlying data are automatically identified and subsequently investigated. A more appropriate model accounting for the extreme value nature of the responses, is also developed and implemented. The implication on acceleration predictions is demonstrated. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: earthquake; peak horizontal accelerations; seismic risk assessment; non-linear mixed-effects models; partially non-linear mixed-effects models; weighted maximum likelihood; robust estimation; generalized extreme value distribution

1. INTRODUCTION

Most advanced countries in the world now have specifications on which buildings are constructed such that public safety is not anticipated to be an issue in the case of an earthquake. The Uniform Building Code is used across the United States to regulate construction of buildings. It categorizes locations into one of six seismic zones, based on the probability of expected intensity of ground shaking resulting from earthquakes. For instance, the maximum seismic risk zone (Zone 4) corresponds to regions where expected peak accelerations (as

*Correspondence to: Debbie J. Dupuis, Department of Management Sciences, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal, Que., Canada H3T 2A7.
†E-mail: debbie.dupuis@hec.ca
‡E-mail: Joanna.Flemming@dal.ca

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada
Contract/grant sponsor: Swiss National Science Foundation; contract/grant number: 1214-66989

Received 17 November 2004
Revised 20 October 2005
Accepted 19 December 2005

a fraction of the acceleration of gravity g) are greater than $0.3g$. Expected peak accelerations at a given location are not known exactly, rather they are estimated from seismological studies. In this paper we investigate some important issues surrounding this estimation.

Large earthquakes are small probability events with large negative consequences. Numerous organizations and people are at risk of exposure to earthquakes: insurance and real estate industries, mortgage lenders, investors, governments, relief and emergency management organizations, city planners, property and business owners, and of course residents at large. And, risk is ever increasing as we live and build closer to earthquake faults. It is therefore essential to appropriately model earthquake response for risk management and insurance applications. A general discussion on seismic risk and insurance, and some of their statistical considerations, is found in Reference [1].

Most current seismic risk analyses are based on acceleration values. Accordingly, in this paper, we focus on the specific problem of intensity of motion as measured by peak horizontal accelerations. Much effort has been directed towards obtaining an expression to predict maximum horizontal accelerations at a specified location during a large earthquake in terms of quake magnitude and distance from the epicentre. A variety of specific functional forms, involving a finite number of real-valued parameters, have been set down. The forms have come from blends of physical and empirical analyses.

Data for 23 large earthquakes in western North America between 1940 and 1980 are considered in Reference [2]. One hundred and eighty two maximum horizontal acceleration values were recorded at available seismometer locations. Based on physical and empirical considerations

$$A = \frac{1}{d} e^{\beta M} e^{\gamma d}$$

was proposed for the maximum peak horizontal acceleration A at distance d in kilometers from an event of moment magnitude M. The relationship

$$\log_{10} A = -1.02 + 0.249M - \log_{10} \sqrt{d^2 + 7.3^2} - 0.00255\sqrt{d^2 + 7.3^2} \tag{1}$$

was developed for the particular data of interest. To prevent earthquakes with many recordings from dominating the estimates, fitting was carried out in two stages in Reference [2]. First magnitude was not included in the model, but rather an event constant was. Then the event constant estimates were regressed on magnitude to obtain the term $-1.02 + 0.249M$.

Improvements for (1) were sought in Reference [3] where a random effect was used as a mechanism for incorporating data from parallel but formally distinct circumstances, i.e. earthquakes. In addition the use of the log and square root transformations in such a relationship has been questioned, as some theory had suggested, but was not definitive about. By employing a non-parametric approach the relationship of A with both M and d was tested. The analysis initially suggested that a cube root transform was more appropriate for the response. However, this bemused seismologists (see Reference [4]) due in part to the fact that this form was based strictly on empirical considerations. Furthermore, in cases of interest results did not differ dramatically. Finally, Brillinger proposed a variant of (1) in Reference [5] by fitting the model

$$\log_{10} A_{ij} = \alpha_i + \beta_i M_i - \log_{10} \sqrt{d_{ij}^2 + \delta_i^2} - \gamma_i \sqrt{d_{ij}^2 + \delta_i^2} + \varepsilon_{ij} \tag{2}$$

to data (d_{ij}, M_i, A_{ij}), $i = 1, \ldots, I$ and $j = 1, \ldots, J$ where i indexes the earthquake and j indexes the record within the earthquake. The $\alpha_i, \beta_i, \gamma_i, \delta_i$ are independent, normally distributed, random effects for the ith earthquake and the ε_{ij} are independent normal error. The model ties together the earthquakes but each earthquake has its own α, β, γ and δ. Desirably this form was based on both empirical and functional considerations. Implications of this model are that records for the same event are correlated and that the disparate number of records for the event are handled automatically. The general effect was desirable as it was observed that one was better able to fit particular earthquakes. As expected, in cases where one has many observations for a particular earthquake the distinction between the two approaches is not as apparent.

The same data were reconsidered in References [6, 7] and are now a resident data set in Splus (a popular statistical analysis software developed by Insightful). These later analyses determined that $\alpha_i, \beta_i, \delta_i$ could be fixed effects and only γ_i need be a random effect. That is

$$\alpha_i = \alpha, \quad \beta_i = \beta, \quad \delta_i = \delta, \quad \gamma_i = \gamma + b_i b_i \sim \mathrm{N}(0, \sigma_\gamma^2) \tag{3}$$

The new analyses were similar to the original analysis in Reference [5] in all other respects, e.g. the error was assumed to be normal.

It is with the assumption of normal error that we take issue here. We observe that

$$A_{ij} = \max\{A_{ij1}, \ldots, A_{ijn_{ij}}\}$$

We do not know the n_{ij} observations on the right-hand side but rather only that the one observation (A_{ij}) that was permanently reported is the maximum of n_{ij} observations taken by the seismometer at site j during earthquake i. That is, an A_{ij} is the *maximum* recorded acceleration at record j of earthquake i. The maximum comes from an accelogram, a non-stationary time series where marginal values will not be identically distributed. The usual assumption is that the distribution of the $\log(A_{ij})$ at individual sites about a regression line that estimates the average value of $\log(A_{ij})$ over all sites is normal. We feel that since the recorded value is the maximum from an accelogram its distribution may be better modelled by something other than a normal. The intermediate step of taking the maximum of a series of varying length induces changes in the error distribution. The idea is perhaps most easily demonstrated by Figure 1 where the density of the maximum of two standard normal random variables is shown along with that of the standard normal. The first density is skewed and shows different tail behaviour. Modelling our data using an error distribution with such properties may lead to a better fit.

According to extreme value theory, it is well known that the only non-degenerate distribution for maxima is the generalized extreme value (GEV) distribution, see References [8, 9]. The GEV is the asymptotic, in this case as $n_{ij} \to \infty$, distribution for maxima but it is standard to fit the model to observed maxima to draw inferences, see Reference [9] for standard fitting techniques. Some consequences of mis-specifying assumptions in non-linear mixed effects models were investigated in Reference [10], but only mis-specification of the random effects distribution was considered, not the error distribution. A general implication of the study in Reference [10] is that very little bias is introduced by deviation from normality of the random effects, but that the variability of the parameter estimates is more profoundly affected. In Section 2 of the paper, our simulation study will show that mis-specification of the error distribution has the same effect.

The remainder of the paper is organized as follows. In Section 3, maximum likelihood and robust estimation of the non-linear mixed-effects model is presented and the earthquake data

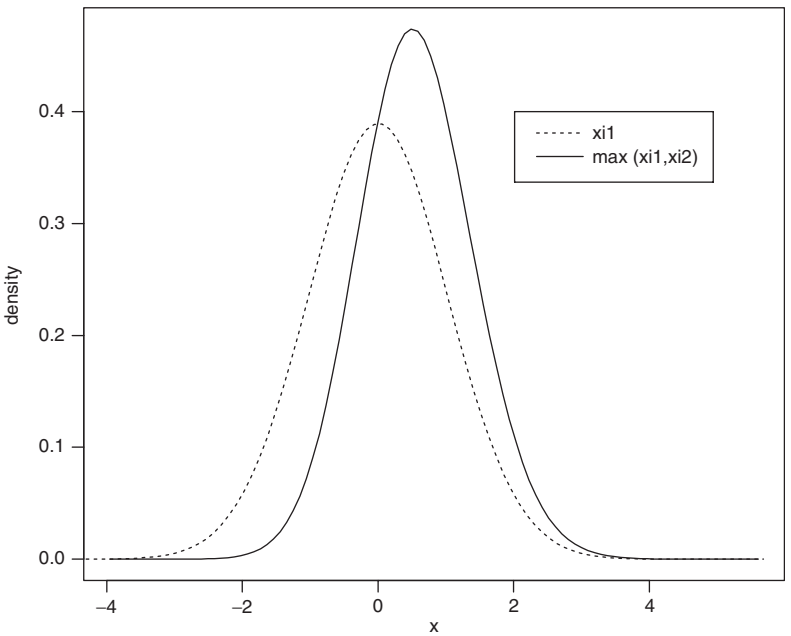

Figure 1. Standard normal density and density of the maximum of two standard normal random variables.

are analysed. In Section 4, a new model with GEV errors is presented and fit to the data. A discussion of implications to acceleration predictions is present in Section 5. Finally, some concluding remarks regarding seismic risk and insurance are included in Section 6.

2. MIS-SPECIFICATION OF THE ERROR DISTRIBUTION

Here we present results of a simulation study designed to investigate consequences of mis-specification of the error distribution. As our interest is specifically in the model for peak horizontal accelerations as given by (2) with parameters as in (3), we will only generate data according to this model. It is however expected that the general conclusions drawn here will apply to other non-linear mixed effects models.

We consider two distributions for the error ε_{ij}: Normal and $\log_{10}$(GEV) (see Section 4 for details). True values of the parameters are set close to the estimated values for the western North America data described in Section 1 with the assumption of normal error. We choose $\alpha = -1$, $\beta = 0.2$, $\delta = 8$, $\gamma = 0.005$, and $\sigma_\gamma = 0.005$. Error distributions listed in Table I are considered. GEV parameter settings were set such that the expected value and standard deviation of the $\log_{10}$(GEV) were also approximately 0 and 0.2, respectively. The three distributions are plotted in Figure 2 for reference and can be classified with respect to their tail index

Table I. Error distributions considered.

Model	Error	Parameter values
1	Normal	$\mu = 0,\ \sigma = 0.2$
2	$\log_{10}$(GEV)	$\mu = 0.8169, \eta = 0.3352, \xi = 0.2327$
3	$\log_{10}$(GEV)	$\mu = 0.8064, \eta = 0.2694, \xi = 0.4186$

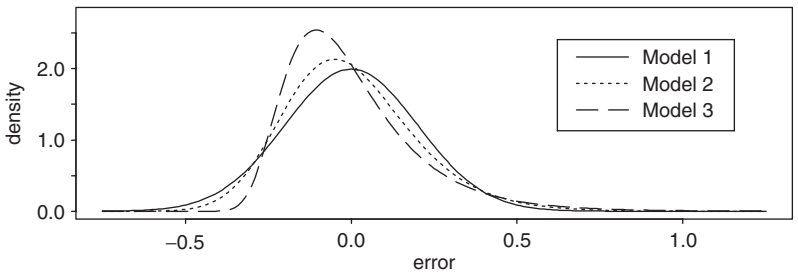

Figure 2. Error distributions.

defined by

$$\tau(F) = \frac{F^{-1}(0.99) - F^{-1}(0.5)}{F^{-1}(0.75) - F^{-1}(0.5)} \Big/ \frac{\Phi^{-1}(0.99) - \Phi^{-1}(0.5)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.5)}$$

where Φ is the standard normal distribution (see Reference [11]). Tail indices are equal to 1, 2.14, and 3.18, respectively.

Data were simulated to emulate those observed in the western North America data set. Each simulated data set had $I = 23$ earthquakes. Earthquake magnitudes were sampled with replacement from those in the western North America data set. The number of readings taken for each earthquake was random and also sampled with replacement from the observed lengths in the western North America data set. Finally, these readings were taken at random distances, the latter also sampled with replacement from the western North America data set. We carried out 1000 simulations. Model (2) with parameters as in (3), along with normally distributed error, was fitted using `nlme` (fits both linear and non-linear normal mixed effects models) in Splus. We tried both the maximum likelihood (`ML`) and the restricted maximum likelihood (`REML`) `method`. The two methods differ only in their approach to estimating variances and covariances among the observations, and yield almost identical results. Boxplots of the sampling distributions of the estimators of each model parameter under the unbalanced design are shown in Figure 3. A horizontal line indicates the true value of the parameter in each case. As the centres of the boxes (which indicate the median of the distributions) for five of the six parameters lie close to the respective lines, we conclude that there is very little bias introduced by deviation from normality of the error distribution for these parameters. We note that σ_γ is underestimated in all cases. The structure of the data, which includes several earthquakes with very few readings, makes the estimation of σ_γ difficult. Repeating the simulation study

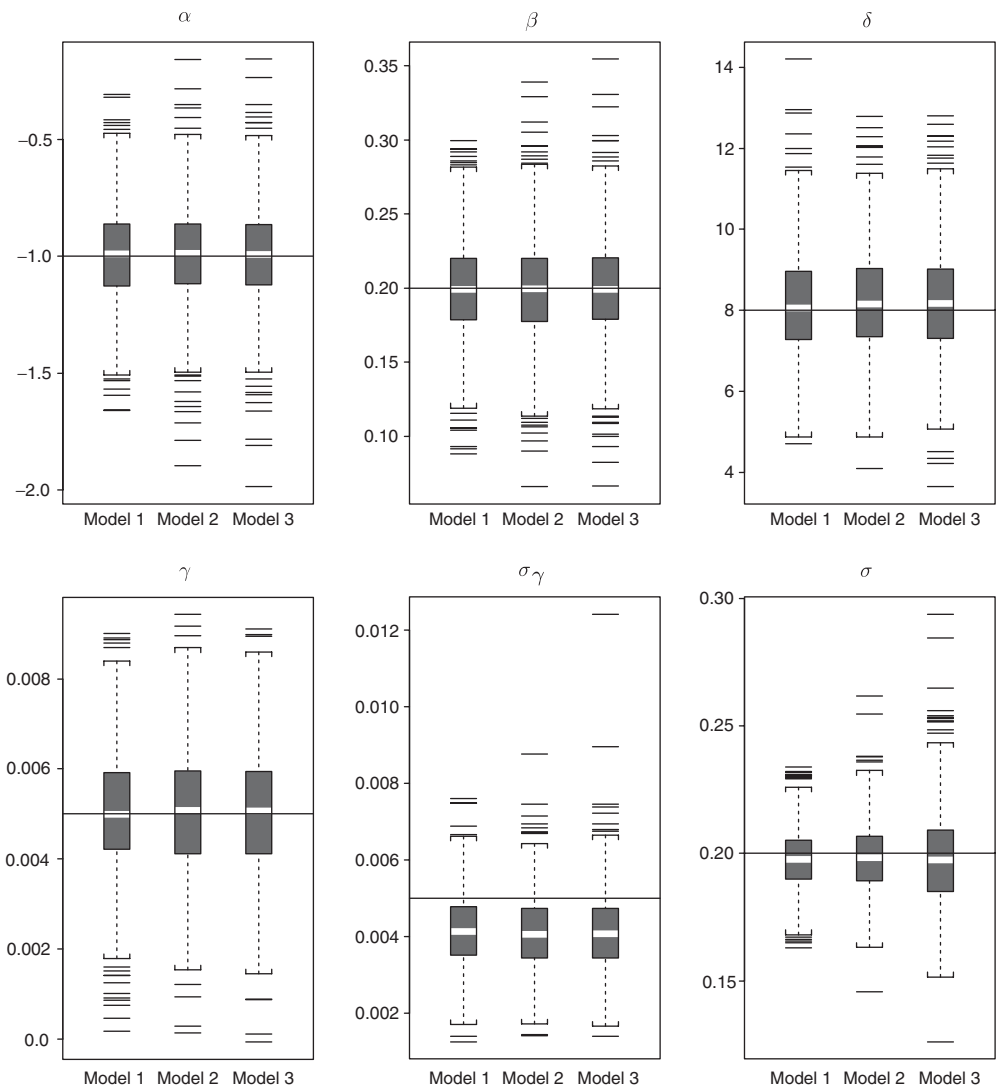

Figure 3. Boxplots of estimates of α, β, δ, γ, σ_γ, and σ for each error distribution. Method of fitting is `ML`. Horizontal bar is set to true value of the parameter. Each box shows the middle 50% of the data with the white bar indicating the median, the square brackets are 1.5 times the interquartile range from the median, and the dashes above and below these the potential outliers. The interquartile range is the difference between the upper and lower quartiles.

for a balanced design, or even an unbalanced one with fewer small clusters, we find these estimates to be unbiased (results not shown). The height of the boxes and the distance of the whiskers from the median line are measures of variability. We see that the variability of the estimators is only mildly affected for the first five parameters, however the variability of the scale parameter σ of the fitted normal error distribution is profoundly affected. The

predicted values, shown to be (10), which involves (9), a function of (6) and thus σ, are directly affected. Thus any subsequent inference is negatively impacted by an inflated estimate of σ. Implications could be more negative than seen here as we were looking at $\log_{10}(A_{ij})$ and some of the negative effect of the thicker than expected upper-tail was reduced by taking logs. Direct estimation of a model for maxima or minima would likely be further affected. A typical scenario, inspired by the real data described in Section 1, was considered in our simulation study. While other scenarios could lead to quantitative differences in the results, the essence of the study remains: the scale parameter is not estimated efficiently and subsequent inference is negatively impacted.

3. A ROBUST MODEL FOR ACCELERATIONS

Having determined in the previous section (using simulated data) that assuming normal errors when the true error distribution has a tail index larger than 1 causes detrimental effects on estimation, we focus here on devising an efficient means of fitting our model while at the same time developing robust enhancements so that we can readily identify parts of the data which are outlying. Consider our non-linear mixed-effects model (2) but with parameters as in (3). Note that individual (aka earthquake) coefficients of the response function enter linearly, while other non-linear parameters do not vary across earthquakes. That is, we have a *Partially Non-linear Mixed Model*. Estimation of this model is only a little more complicated than that of a fully linear mixed model.

3.1. A maximum likelihood approach

Because the model is linear in the random effects, maximum likelihood estimates (MLE) can be obtained in a straightforward manner. We have

$$Y_{ij} = \alpha + \beta M_i - \log_{10}\sqrt{d_{ij}^2 + \delta^2} - (\gamma + b_i)\sqrt{d_{ij}^2 + \delta^2} + \varepsilon_{ij} \tag{4}$$

with $Y_{ij} = \log_{10} A_{ij}$. Let $X_{ij} = (1, M_i, -\log_{10}\sqrt{d_{ij}^2 + \delta^2}, -\sqrt{d_{ij}^2 + \delta^2})$, $Z_{ij} = -\sqrt{d_{ij}^2 + \delta^2}$, and $u = (\alpha, \beta, 1, \gamma)'$. There are two types of random variables in this model. The intra-earthquake model, concerns the residuals, ε_{ij}. Conditional on the random effects b_i, we have $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ where $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$ or $\varepsilon_i \sim \mathrm{N}(0, \sigma^2 I_{n_i})$ where I_{n_i} is the $n_i \times n_i$ identity matrix. The inter-earthquake model relates to b_i over the population of earthquakes. It is assumed that $b_i \sim \mathrm{N}(0, \sigma_\gamma^2)$. From these assumptions, the marginal mean and covariance matrix of Y_i are, respectively,

$$m_i = E(Y_i) = X_i u \tag{5}$$

that is, $E(Y_{ij}) = \alpha + \beta M_i - \log_{10}\sqrt{d_{ij}^2 + \delta^2} - \gamma\sqrt{d_{ij}^2 + \delta^2}$, and

$$S_i = \mathrm{cov}(Y_i) = Z_i \sigma_\gamma^2 Z_i' + \sigma^2 I_{n_i} \tag{6}$$

Model (4) is non-linear in δ, but linear in $u=(\alpha,\beta,1,\gamma)'$ (where we include 1 for notational simplicity) and the individual coefficients b_i. Let $q=(\delta,\sigma_\gamma,\sigma)$ denote all non-linear parameters of the model. The expected value and covariance matrix of Y_i are $m_i=m_i(u,q)$ and $S_i=S_i(q)$, respectively. The log-likelihood for a sample of N observations (i.e. earthquakes) is

$$L(u,q)=-\frac{1}{2}\sum_{i=1}^{N}[n_i\log(2\pi)+\log|S_i(q)|+Q_i(u,q)] \tag{7}$$

where

$$Q_i(u,q)=(Y_i-m_i(u,q))'S_i(q)^{-1}(Y_i-m_i(u,q))$$

To obtain MLE, any of several numerical methods can be used to maximize $L(u,q)$. We proceed as follows:

Step 1: Examination of (7) reveals that α, β and γ only enter the log-likelihood through the $Q_i(u,q)$ terms. Commencing with reasonable starting values for all of the parameters, hold those values for δ, σ_γ, and σ fixed, and use Splus' `nlminb` (a non-linear minimizer) to minimize the term $\sum_{i=1}^{N}Q_i(u,q)$ to arrive at estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$.

Step 2: Now, minimize (again using `nlminb`) the negative of the log-likelihood as given in (7) with values of α, β, and γ fixed at $\hat{\alpha}_1$, $\hat{\beta}$ and $\hat{\gamma}$ obtained in Step 1. This minimization yields estimates $\hat{\delta}$, $\hat{\sigma}_\gamma$, and $\hat{\sigma}$.

The above two steps are then iterated until convergence. A relative tolerance of 10^{-6} is used. Final estimates are $\hat{\alpha}=-0.802,\hat{\beta}=0.222,\hat{\delta}=8.012,\hat{\gamma}=0.0053,\hat{\sigma}_\gamma=0.00418,\hat{\sigma}=0.217$, with a log-likelihood of 2.14. Note that these values are a little different than those reported in Reference [6]. The above algorithm performs better than Splus' `nlme` in seeking the maximum log-likelihood in this case (2.12 for previously reported estimates versus 2.14 for ours). Figure 4 shows a quantile–quantile plot of the standardized residuals for the MLE of model (4) assuming $\varepsilon_{ij}\sim N(0,\sigma^2)$. Standardized residuals $(y_{ij}-\hat{y}_{ij})/\hat{\sigma}$ are calculated as explained in Section 3.2. It is clear that the distributional assumption is not adequate as both very large negative and very large positive residuals are poorly explained by the tails of the Normal distribution. Note that large (small) predicted accelerations can correspond to any size residual so that discrepancies in the upper (lower) tail of Figure 4 do not indicate that large (small) accelerations are poorly fitted by the model.

3.2. Robust enhancements

As argued in Section 1 and clearly seen in Figure 4, the assumption of normal errors is questionable for these maximum acceleration data. A robust fit of the normal model would identify any points possibly in violation of that assumption. Robust estimation of parametric models through the use of weighted maximum likelihood techniques is achieved in Reference [12]. Their estimator downweights with respect to the model and can be used for complicated likelihoods. Basically, a weighted negative log-likelihood of the form $-\log[f_Y^w]$ is minimized using fixed weights w based on the last parameter estimates. This procedure is repeated until convergence, at which point a correction term must be computed to get approximate Fisher consistency. Fisher consistency, for a statistical estimation procedure, describes the desirable situation in which the estimate produced coincides with what we wish to estimate, that is, the

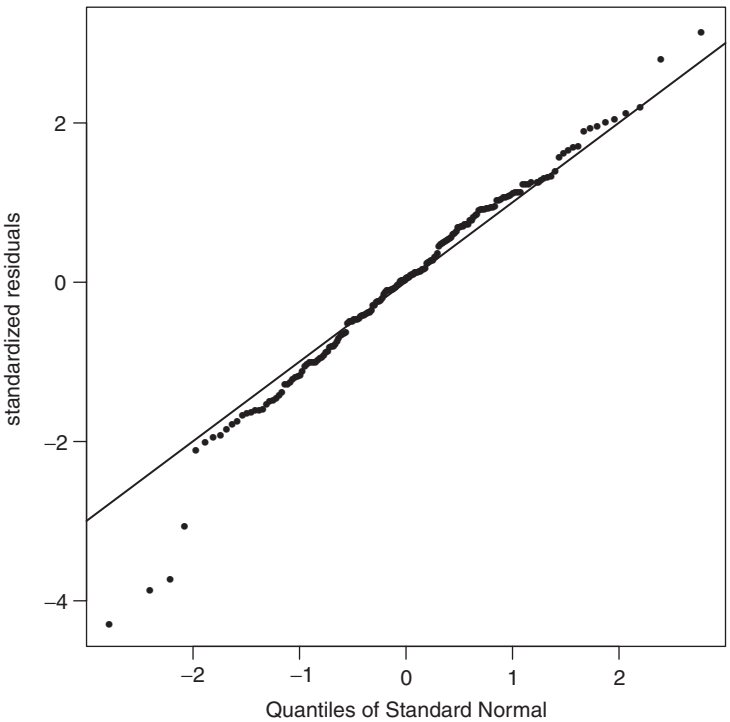

Figure 4. Quantile–quantile plot of standardized residuals from MLE fit to model (4) assuming normal errors.

procedure has no estimation bias in theory. Here, in the spirit of Reference [13], we use the probabilistic weighting functions

$$w(y_{ij}) = \begin{cases} \Phi\left(\dfrac{y_{ij}-\hat{y}_{ij}}{\hat{\sigma}}\right)\Big/p_1 & \text{if } \Phi\left(\dfrac{y_{ij}-\hat{y}_{ij}}{\hat{\sigma}}\right) < p_1 \\ 1 & \text{if } p_1 \leqslant \Phi\left(\dfrac{y_{ij}-\hat{y}_{ij}}{\hat{\sigma}}\right) \leqslant 1-p_2 \\ \left[1-\Phi\left(\dfrac{y_{ij}-\hat{y}_{ij}}{\hat{\sigma}}\right)\right]\Big/p_2 & \text{if } \Phi\left(\dfrac{y_{ij}-\hat{y}_{ij}}{\hat{\sigma}}\right) > 1-p_2 \end{cases} \tag{8}$$

where p_1 and p_2 are the robustness constants and $\hat{y}_{ij}$ are the fitted values. Larger values of p_1 and p_2 produce a more robust, and less efficient, estimator at the true model. In dealing with a mixed-effects model fitted values may be obtained at different levels, see Reference [7] for further details. Our weights, and subsequently our robust estimates, depend on our choice of fitted values. Consider fitted values given the random effects. Although technically the random effects b_i are not parameters of our model, they do behave in some ways like parameters and often we want to *estimate* their values. *Best Linear Unbiased Predictors* or

Table II. Parameter estimates for robust fits.

	$p_1 = p_2 = 0.005$	$p_1 = p_2 = 0.0075$	$p_1 = p_2 = 0.01$
$\hat{\alpha}$	−0.685	−0.686	−0.686
$\hat{\beta}$	0.205	0.205	0.205
$\hat{\gamma}$	0.005	0.005	0.005
$\hat{\delta}$	7.959	7.950	7.947
$\hat{\sigma}_\gamma$	0.0038	0.0038	0.0038
$\hat{\sigma}$	0.209	0.209	0.209

Table III. Downweighted observations and weights for robust fits. Results for $p_1 = p_2 = 0.005$, $p_1 = p_2 = 0.0075$, and $p_1 = p_2 = 0.01$, respectively, are listed.

Code	Quake	Richter	Distance	Weight
3	14	5.2	17.0	3.7e − 02
13	18	5.8	5.3	0.39
13	18	5.8	7.4	0.11
13	18	5.8	23.4	6.1e − 04
13	18	5.8	30.0	0.12
13	18	5.8	38.9	8.3e − 03
3	14	5.2	17.0	2.5e − 03
13	18	5.8	5.3	0.26
13	18	5.8	7.4	7.9e − 02
13	18	5.8	23.4	4.1e − 04
13	18	5.8	30.0	8.3e − 02
13	18	5.8	38.9	5.5e − 03
3	14	5.2	17.0	1.8e − 03
13	18	5.8	5.3	0.19
13	18	5.8	7.4	5.9e − 02
13	18	5.8	23.4	3.1e − 04
13	18	5.8	30.0	6.2e − 02
13	18	5.8	38.9	4.1e − 03

BLUPS replace the random effects b_i by their conditional means $\hat{b}_i$ given the data and then make predictions using those values. That is $\hat{b}_i$ is the conditional expectation

$$\hat{b}_i = E(b_i|Y_i) = \sigma_\gamma^2 Z_i' S_i^{-1}(Y_i - m_i) \tag{9}$$

evaluated at the MLE of u and q. The corresponding BLUPS which we take to be our fitted values are

$$\hat{y}_i = \hat{m}_i + \hat{Z}_i \hat{b}_i \tag{10}$$

where $\hat{m}_i$ and $\hat{Z}_i$ are m_i and Z_i, evaluated at the current WMLE of u and q. Parameter estimates for three different levels of robustness appear in Table II and downweighted observations are listed in Table III. The robust procedure identifies a few observations that have a large impact

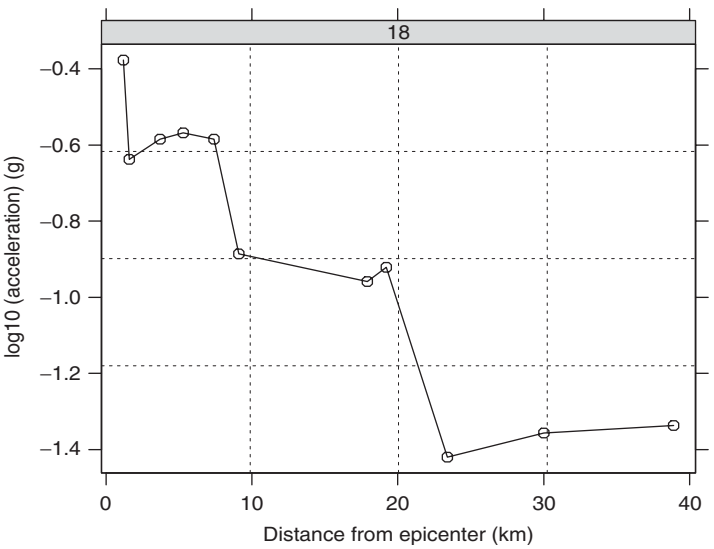

Figure 5. $\log_{10}$ accelerations for Quake 18 (Coyote Lake, CA 1979).

on the fitted values under the assumed model. One observation from Quake 14 (Hollister, CA 1974) is identified and significantly downweighted as well as several observations from Quake 18 (Coyote Lake, CA 1979). Note that the consistency of the message across different levels of robustness leads us to believe that any outlying observations have been properly identified. Figure 5 shows the observed accelerations for Quake 18. From left to right, the 4th, 5th, 9th, 10th, and 11th observations are being downweighted and this is certainly reasonable given that the first two observations appear too large while the latter three appear too small. We note that the same observation from Quake 14 is also identified as outlying in Reference [3] where residual and probability plots from a non-parametric model are used.

Figure 6 shows observed and predicted values, along with standardized residuals, for both maximum likelihood and robust estimates. With the presence of larger errors than expected under normality, MLE are biased in an attempt to accommodate the *outliers*. The robust fit better represents the bulk of the data.

4. A MORE APPROPRIATE MODEL

In model (4) normal errors are usually assumed for simplicity since, along with normally distributed linear random effects, the responses are then also normally distributed. However, as argued in Section 1, a distribution that takes into account that the data are maxima might be more appropriate for the errors. That is, normal errors make for lognormally distributed accelerations while we would prefer that these accelerations be modelled by a GEV distribution. We will now choose errors to enable the latter.

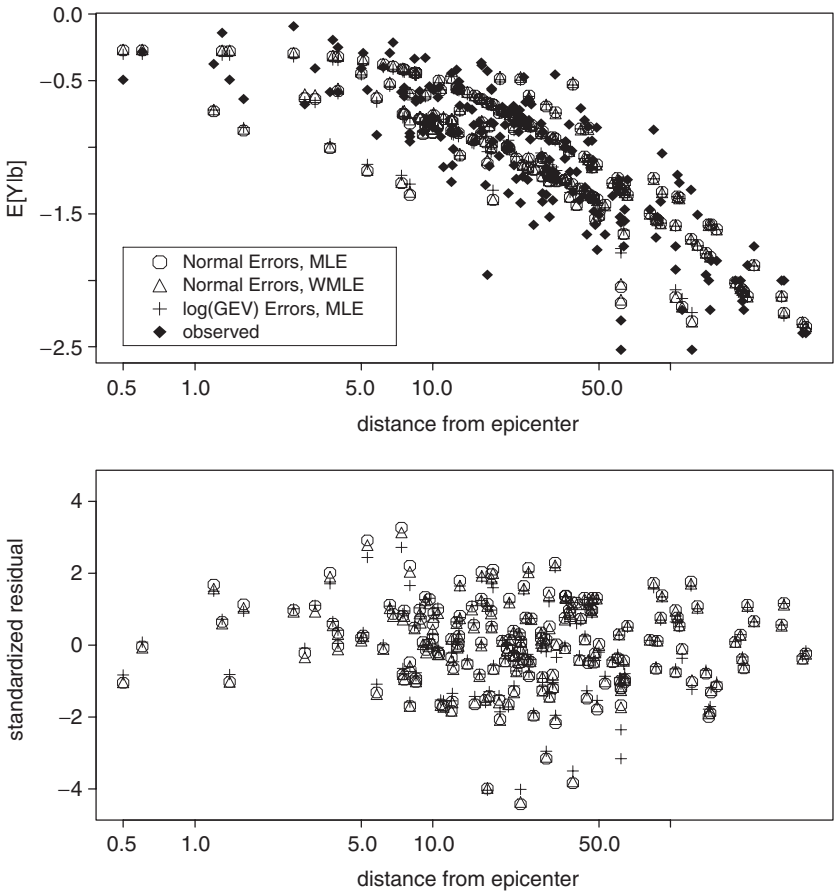

Figure 6. Predicted $\log_{10}$ accelerations and residuals; under normal errors, MLE and WMLE with $p_1 = p_2 = 0.005$, and $\log_{10}(\text{GEV})$ errors.

Suppose model (4) where conditional on the random effects b_i, we now have that $\varepsilon_{ij} \sim \log_{10}(\text{GEV})$ and $\text{cov}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$. The GEV distribution is

$$\begin{aligned} F(x;\mu,\eta,\xi) &= \exp\left[-\left\{1+\frac{\xi(x-\mu)}{\eta}\right\}_+^{-1/\xi}\right], \quad \xi \neq 0 \\ &= \exp\left[-\exp\left\{-\frac{(x-\mu)}{\eta}\right\}\right], \quad \xi = 0 \end{aligned} \tag{11}$$

with parameters satisfying $\eta > 0$ and $\mu, \xi \in \Re$. We have used the usual notation $\{s\}_+ = \max(s, 0)$. Further note that x is bounded by $\mu - \eta/\xi$ from above if $\xi < 0$ and from below if $\xi > 0$. It is unbounded if $\xi = 0$. The choice of the GEV is consistent with extreme value modelling as this parametric family is motivated by limiting distributions in extreme value theory. We let $\theta = (\mu, \eta, \xi)$ and f_θ denote the density corresponding to (11). A transformation

of variable approach yields

$$g_\theta(x) = \ln(10)\exp(x\ln(10))f_\theta(\exp(x\ln(10)))$$

as the density of ε_{ij}. For notational simplicity, define

$$V_{ij} = \alpha + \beta M_i - \log_{10}\sqrt{d_{ij}^2 + \delta^2} - (\gamma + b_i)\sqrt{d_{ij}^2 + \delta^2}$$

so that $Y_{ij} = V_{ij} + \varepsilon_{ij}$ and

$$Y_{ij}|b_i = E[V_{ij}] - b_i\sqrt{d_{ij}^2 + \delta^2} + \varepsilon_{ij}$$

so that, conditional on the random effect, observations within an earthquake are independent. Thus, the joint density of observations within an earthquake is

$$f(y_{i1},\ldots,y_{in_i}) = \int \prod_j g_\theta(y_{ij} - E[V_{ij}] + x\sqrt{d_{ij}^2 + \delta^2})\frac{1}{\sigma_\gamma}\phi\left(\frac{x}{\sigma_\gamma}\right)\mathrm{d}x \tag{12}$$

where ϕ is the standard normal density. The log-likelihood for a sample of I earthquakes is

$$L(u,\delta,\sigma_\gamma,\theta) = \sum_{i=1}^{I}\log\left(\int \prod_j g_\theta(y_{ij} - E[V_{ij}] + x\sqrt{d_{ij}^2 + \delta^2})\frac{1}{\sigma_\gamma}\phi\left(\frac{x}{\sigma_\gamma}\right)\mathrm{d}x\right) \tag{13}$$

Note that we now have eight unknown parameters to estimate: $\alpha, \beta, \gamma, \delta, \sigma_\gamma, \mu, \eta$, and ξ. We constrain the mean of the error distribution to be 0, determining the required value of μ given η and ξ, so maximization is over 7 parameters. Numerical integration is used to evaluate (12) and Splus' `nlmin` (a non-linear minimizer) is used for the optimization. Final estimates are $\hat{\alpha} = -0.835, \hat{\beta} = 0.226, \hat{\delta} = 8.430, \hat{\gamma} = 0.0036, \hat{\sigma}_\gamma = 0.00205, \hat{\mu} = 0.880, \hat{\eta} = 0.437$, and $\hat{\xi} = 0.0026$, with a log-likelihood of 4.77. We compute the fitted values, conditional on the random effects, under the new model as in (10) with the $\hat{b}_i$ given by (9) where S_i is as in (6) where σ^2 is now the variance of a random variable with density $g_\theta(x)$ with $\theta = (\hat{\mu}, \hat{\eta}, \hat{\xi})$. Here, $\hat{\sigma} = 0.230$. These fitted values appear in Figure 6. One can see that the fitted values corresponding to the model with $\log_{10}$(GEV) errors are generally closer to those observed than with MLE and WMLE under normal errors, especially for shorter distances from the epicentre. Furthermore, the residuals for the model with $\log_{10}$(GEV) errors are less extreme yet still well scattered. Figure 7 shows a quantile–quantile plot of the standardized residuals for the MLE for log-likelihood (13). We have better reconciliation in the tails. While goodness-of-fit tests for such complex models are not available, any reasonable measure of fit will favour Figure 7 over Figure 4 as the lower-left point in the latter is so far from the diagonal. It is worth emphasizing that the q–q plots are those of the standardized residuals and not the accelerations themselves. It is only through an examination of the residuals that we can assess the validity of the model assumptions. The upper-tail of the q–q plot represents large positive standardized residuals, i.e. differences between observed accelerations and predicted accelerations are (comparatively) large positive values, and these do not correspond to the largest accelerations. Since good estimation of the larger accelerations may be more important for hazard analysis, we examine more closely the observed large accelerations and their fitted values under both models. The sum of the squared differences between the largest

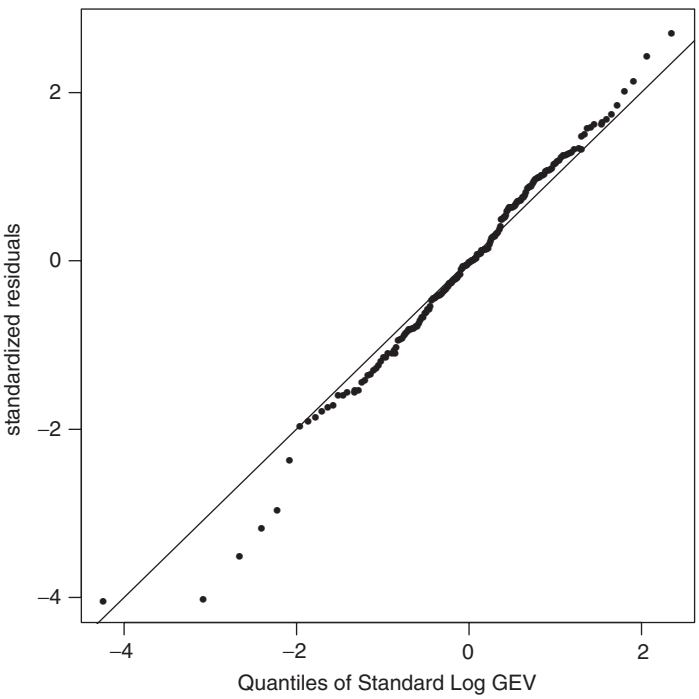

Figure 7. Quantile–quantile plot of standardized residuals from MLE fit to log-likelihood (13).

half of the observed log(accelerations) and their fitted values is 4.65 and 4.72 under the GEV and normal models, respectively. The sum of the squared differences between the largest 25% of the observed log(accelerations) and their fitted values is 3.12 and 3.15 under the GEV and normal models, respectively. Given all these considerations, we conclude that the GEV distributional assumptions provide a better fit.

Finally, the results of a small simulation study are shown in Figure 8. Data were generated as described in Section 2 with error distribution following Model 3 of Table I. Parameter estimates when maximizing log-likelihoods (7) and (13), respectively, are shown. It is clear that both the bias (approximately equal to the distance from the centre of the box to the horizontal line set at the true parameter value) and variance (for which the height of the boxes is a proxy) of all estimates are reduced when the correct model (i.e. the one with log-likelihood (13) as it is the true model for the generated data) is fitted. Fitting models with normal errors when errors are $\log_{10}(\text{GEV})$ is simply not appropriate.

5. DISCUSSION

Optimally, our new model (hereafter referred to as the GEV model) should be fitted robustly. Unfortunately, the form of (13) does not allow for downweighting observations individually given presently available robust methodology. For interest sake, we refit the GEV model without the outlying observations as identified in Table III. Differences in parameter estimates

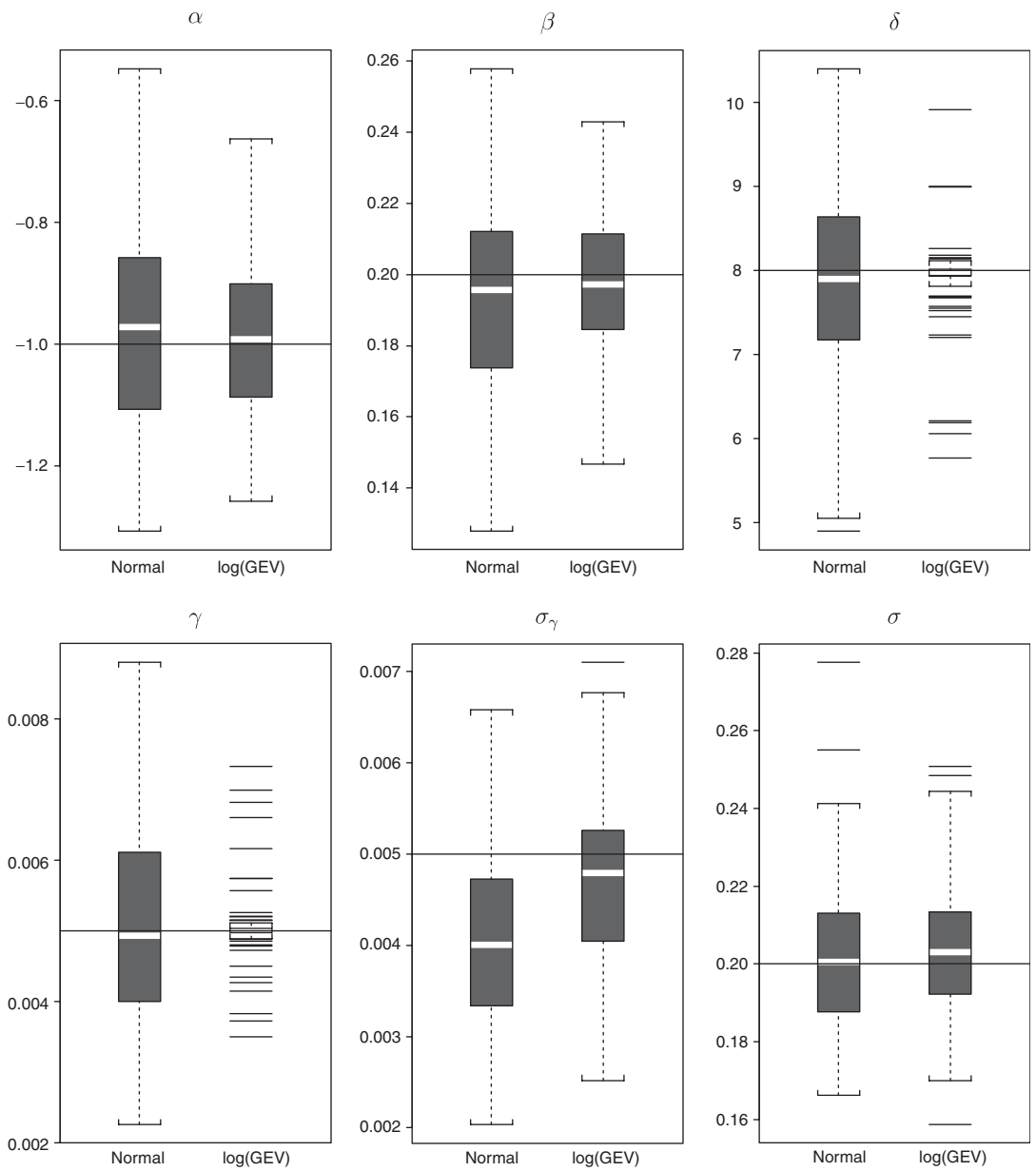

Figure 8. Boxplots of estimates of α, β, δ, γ, σ_γ, and σ maximizing log-likelihoods (7) and (13), respectively, when true error distribution is Model 3. Horizontal bar is set to true value of the parameter. For log-likelihood (13), σ is the standard deviation of the fitted error distribution.

were found to be quite small, and the results best summarized by stating that the average predicted accelerations were $\approx 1\%$ larger when the outlying observations were excluded. Furthermore, over distances from 100 to 400 miles from the epicentre, the average predicted

accelerations did not differ. Similar comparisons between MLE and WMLE for the model with normal errors yields an increase of $\approx 2\%$ with the WMLE fit. Over 100–400 miles, the average estimates based on WMLE are $\approx 0.3\%$ lower than those based on the MLE. One can therefore conclude that these particular observations do not have as profound an effect on the new analysis as they did with the original one granting us additional confidence that our new model tends to fit the data better in general.

New attenuation models continue to be developed for use in seismic risk studies, see e.g. References [14, 15]. In all such cases that we have encountered the error terms are assumed to be normally distributed. Our analyses suggest that these assumptions are questionable and that improvements can be made by using a GEV model like that proposed in Section 4. Additionally, the robust weights resulting from the robust fit of the model with normal errors can be used as helpful diagnostics as was illustrated in Section 3.2.

5.1. Probabilistic seismic hazard analysis

To assess the impact of the new model, we carry out a simplified probabilistic seismic hazard analysis. Figure 9 shows the probability of exceedance of certain peak accelerations under the normal and GEV models when an earthquake of magnitude 6 is assumed. Differences in the two estimates are more pronounced for smaller peak accelerations, with the GEV model yielding larger probabilities of exceedance over most distances of interest. This paper seeks only to address the ground motion model (attenuation relationship), so we will consider a fictitious seismic-hazard source model to complete the comparison. Suppose a source model composed of two scenarios. The first event is an earthquake of magnitude 6 that occurs every 22 years, and the second is an earthquake of magnitude 7.8 that occurs every 300 years. Both are strike slip events located 10 km from the site of interest. The Poissonian probability of exceeding each ground-motion level over a 50-year period is shown in Figure 10. The two attenuation models only differ in their predictions of the more unlikely events. The normal model yields $1.41g$ as the peak acceleration with 1% probability of exceedance in 50 years while the GEV model yields a smaller $1.34g$ as that value.

6. CONCLUSIONS

Our analysis reveals that, for a particular model of accelerations, assumptions on the distribution of the errors can have a large effect on subsequent conclusions. Predicted accelerations are used by engineers for design considerations, and by insurers for loss modelling. By underestimating accelerations, we are then in a situation where engineers are without adequate design specifications and insurers financially ill-prepared for impending losses. That is, the severity of the claims could be larger with greater accelerations and this should be reflected in the setting of insurance premiums. By overestimating, engineers and insurers are overly cautious, and money and resources are wasted. Our new analysis, more appropriately modelling the errors as a GEV distribution, helps correct this problem by providing a better fit to the data and in turn more accurate acceleration estimates.

Partial non-linear mixed-effects models may be used to model many phenomena that exhibit degradation in time or space and the methodology developed here could also be applied more

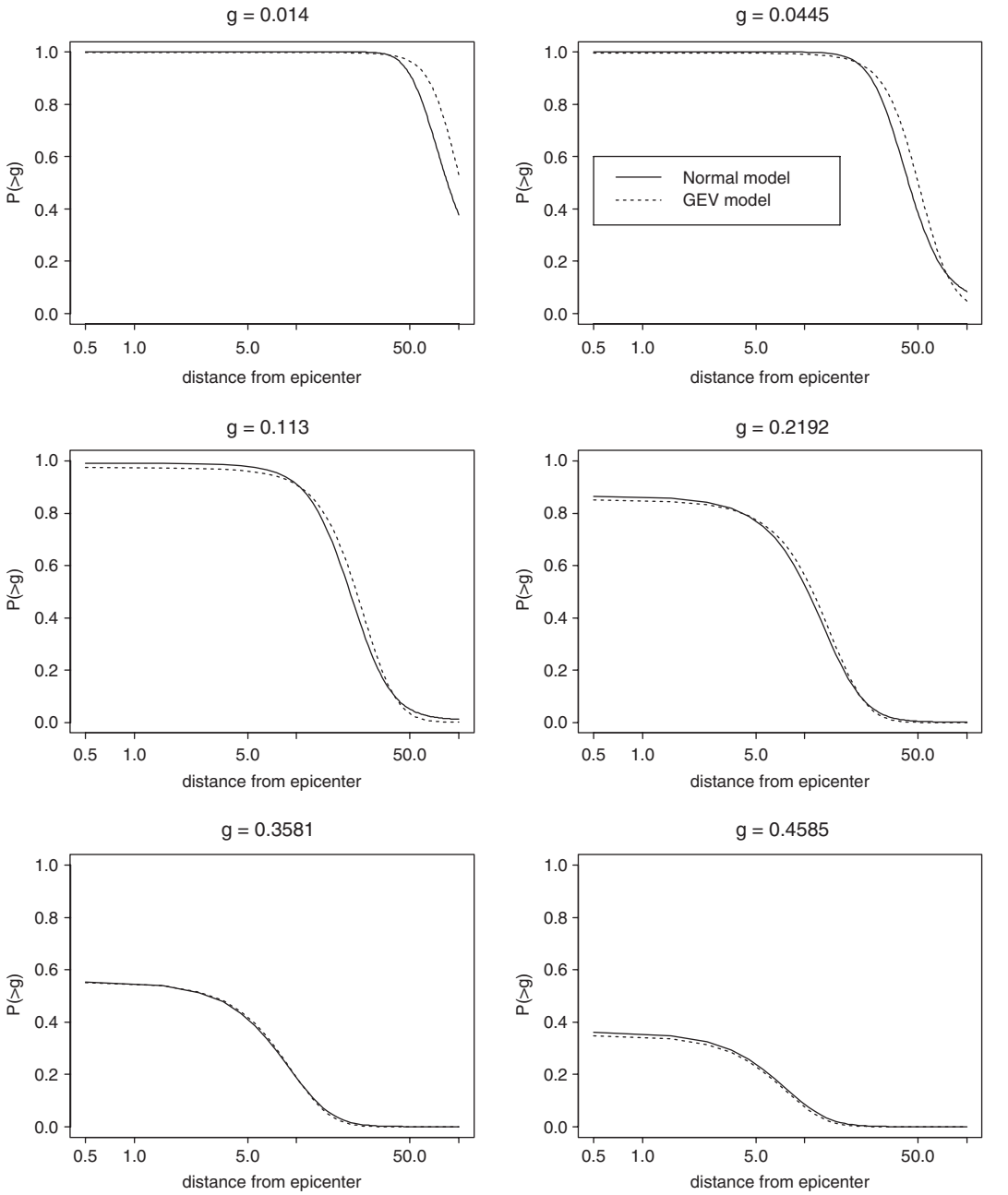

Figure 9. Probability of exceeding peak acceleration g in the event of an earthquake of magnitude $M_i = 6$. Plots are shown for g equal to the $10, 25, 50, 75, 90$, and 95% quantiles among the observed accelerations (182 data points described in Section 1).

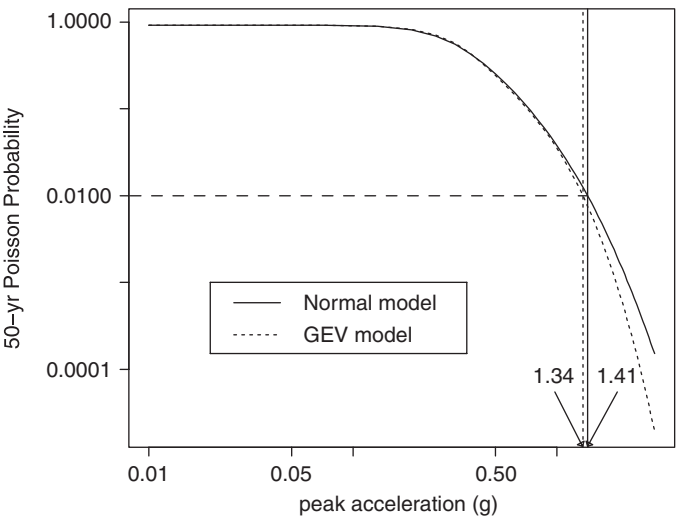

Figure 10. The hazard curve (50-year Poissonian probability of exceedance) for the two-scenario model described in the text. The peak accelerations with 1% chance of exceedance are 1.34 and 1.41g for the GEV and Normal model, respectively.

generally. The error distribution need not be GEV but rather something more appropriate to the situation, and estimation could be carried out as outlined in Section 4.

ACKNOWLEDGEMENTS

Both authors acknowledge ongoing grant support from the Natural Sciences and Engineering Research Council of Canada. The second author also acknowledges the financial support of the Swiss National Science Foundation Project Number 1214-66989. The authors wish to thank two anonymous referees for helpful comments which improved the presentation.

REFERENCES

1. Brillinger DR. Earthquake risk and insurance. *Environmetrics* 1993; **4**(1):1–21.
2. Joyner WB, Boore DM. Peak horizontal accelerations and velocity from strong-motion records including records from the 1979 Imperial Valley, California, earthquake. *Bulletin of the Seismological Society of America* 1981; **71**:2011–2038.
3. Brillinger DR, Preisler HK. An exploratory analysis of the Joyner–Boore attenuation data. *Bulletin of the Seismological Society of America* 1984; **74**(4):1441–1450.
4. Brillinger DR, Preisler HK. Further Analysis of the Joyner–Boore attenuation data. *Bulletin of the Seismological Society of America* 1985; **75**(2):611–614.
5. Brillinger DR. Some examples of the statistical analysis of seismological data. In *Observatory Seismology*, Litehiser JJ (ed.). University of California Press: Berkeley, 1989; 266–278.
6. Davidian M, Giltinan DM. *Non-linear Models for Repeated Measurement Data*. Chapman & Hall: London, 1995.
7. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and Splus*. Springer: New York, 2000.
8. Jenkinson AF. The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quarterly Journal of the Royal Meteorological Society* 1955; **81**:158–172.
9. Coles S. *An Introduction to the Statistical Modeling of Extreme Values*. Springer: Berlin, 2001.
10. Hartford A, Davidian M. Consequences of misspecifying assumptions in non-linear mixed effects models. *Computational Statistics and Data Analysis* 2000; **34**:139–164.

11. Hoaglin DC, Mosteller F, Tukey JW (eds.). *Understanding Robust and Exploratory Data Analysis*. Wiley: New York, 1983.
12. Dupuis DJ, Morgenthaler S. Robust weighted likelihood estimators with an application to bivariate extreme value problems. *Canadian Journal of Statistics* 2002; **30**(1):17–36.
13. Field C, Smith B. Robust estimation—a weighted maximum likelihood approach. *International Statistical Review* 1994; **62**(3):405–424.
14. Ozbey C, Sari A, Manuel L, Mustafa E, Fahjan Y. An empirical attenuation relationship for Northwestern Turkey ground motion using a random effects approach. *Soil Dynamics and Earthquake Engineering* 2004; **24**:115–125.
15. Rhoades DA. Estimation of attenuation relations for strong-motion data allowing for individual earthquake magnitude uncertainties. *Bulletin of the Seismological Society of America* 1997; **87**(6):1674–1678.

