JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 101, NO. B9, PAGES 20,083–20,090, SEPTEMBER 10, 1996

Peak factors of Mexican accelerograms: Evidence of a non-Gaussian amplitude distribution

A. A. Gusev

Instituto de Geofisica, Universidad Nacional Autónoma de México, Mexico City
Institute of Volcanic Geology and Geochemistry, Russian Academy of Science, Petropavlovsk-Kamchatsky

Abstract. To study statistical properties of the near-source radiation field, 35 horizontal accelerograms were selected, recorded at distances $R < 110$ km from five $M > 7$ Mexican earthquakes. For each one, the maximum-amplitude segment, assumedly containing direct S waves, was selected by an automatic procedure. Then two parameters were determined for this segment: the peak to rms amplitude ratio, or peak factor PF, and the slope K of the dependence of PF on log duration. These two values were compared to the corresponding values expected for a simplest stochastic record model, of a segment of the stationary Gaussian process. For about 40% of the records, both parameters show clear (above 2σ) deviations from this model. Another record model was also tested, of quasi-stationary Gaussian process (with time-varying rms amplitude). With this end, the original records were "stationarized," that is, modified to attain nearly constant rms amplitude, and then analyzed in the same manner as the original ones. However, for a distinct fraction of records, deviations persisted. The type of the deviations is characteristic for a "heavy-tailed" amplitude distribution, with the enhanced probability of large peaks. The degree of expression of this phenomenon seems to decay with distance. This observation suggests that source radiation, as it is generated at the fault, often is inherently non-Gaussian (heavy-tailed), but because of the scattering and multipathing, it quickly loses this property. One can expect this picture not to be specific for Mexican earthquakes only. From the viewpoint of stochastic strong motion simulation, the results imply that the PF value calculated from the stationary Gaussian model can underestimate average peak acceleration, on the average, by some 25%. In an analysis aimed at the recurrence of rare peaks, the underestimation may be even larger: among 35 records studied, five have peak factors more than 50% above the values expected for the stationary Gaussian model.

Introduction

The statistical properties of high-frequency radiation from earthquake sources were studied during the last decades mainly from the practical viewpoint, in order to provide a sound basis for estimates of future strong motion. *Hanks and McGuire* [1981] demonstrated that the amplitude distribution of many accelerograms is close to Gaussian, and this gave certain grounds to the accelerogram model as a segment of the stationary Gaussian process (already used in mostly a priori manner in earlier studies). This model was widely applied to relate peak acceleration to rms acceleration [e.g., *Hanks and McGuire,* 1981; *Gusev,* 1983; *Boore,* 1983]. Although this approach may be generally acceptable for engineering purposes, and even adequate for a large percent of accelerograms, some near-source records clearly show more complicated behavior. First, the acceleration amplitudes vary during the "source-related" maximum phase of an accelerogram, which, for large earthquakes with sufficiently long source duration, may usually be associated with the direct (or somewhat distorted by forward-scattering) S waves, radiated from the source. The pattern in point is not the evident random fluctuations, rather it is the gradual change in average power, seen, for example, as subevents of considerable duration (5-10 s or more). Second, one can sometimes note short, powerful individual acceleration spikes. Both these conspicuous features of real accelerograms may bear important information on the fine structure of a rupturing fault. In particular, statistics of acceleration maxima can reflect statistics of local stress drop, and each prominent "direct-S-wave" acceleration peak may be formed by the failure of a single, strong asperity [*Gusev*, 1989].

We have already noted that the statistics of acceleration amplitudes are also relevant for the strong motion prediction. When a calculation scheme for such a prediction uses the spectral representation of a signal, it usually employs a theoretical value of the peak-to-rms acceleration ratio, that is, of the peak factor PF, in order to pass from the rms to the peak value. Usually, PF values used are based on a certain statistical model of a record, so that an inadequate statistical model may bias the estimates of peak acceleration even if all previous steps in its calculation are correct. Therefore the clear understanding of real acceleration amplitude statistics has obvious practical meaning as well.

Paper number 96JB00810
0148-0227/96/96JB-00810$09.00

On the basis of the results of the visual inspection noted above, I anticipate two kinds of deviations from the simplest stationary Gaussian model. The first is the modulated or quasi-stationary Gaussian process, with time-varying variance (or rms acceleration). Such a process is locally Gaussian, but stacking of amplitudes from subsegments of different variance produces a non-Gaussian distribution for the entire analyzed segment. Another, more interesting alternative is an inherently non-Gaussian process. Specifically, the presence of intense spikes suggests that the appropriate generic alternative model must have a "heavy-tailed" amplitude distribution, with higher-than-Gaussian probability of large deviations. (An example is the pulse process with the power law pulse amplitude distribution proposed by *Gusev* [1989]. To discriminate between these two alternatives is an important issue. To test the model of the quasi-stationary Gaussian process, I will try to reduce observed signal to time-independent variance/rms amplitude, "stationarize" it. If the result is near to the Gaussian process, this would indicate the quasi-stationary model as the preferable one. In the opposite case, one should resort to an intrinsically non--Gaussian (heavy-tailed) model. I will not go into any particular non-Gaussian models here and will restrict the discussion only to the preliminary stage of data analysis. My main aim is to show that the simple Gaussian model is often inadequate and, generally, cannot be saved even if extended to allow time-varying rms amplitude.

I shall analyze only the maximum segment of an accelerogram, because (1) it contains direct source radiation probably reflecting the rupture process and (2) it is the most energetic part, important for applications. To obtain the record durations sufficiently long for any meaningful statistical analysis, I use records of large-magnitude earthquakes, with several seconds or more of source-related ground motion present.

Theory

Initially, I introduce the technique of the statistical analysis which then will be applied to the accelerogram data. The whole approach is based on extreme values; this kind of technique is well adjusted to our case: checking the Gaussian model against a heavy-tailed alternative. The parameters and tests applied are based on the probabilistic background due to *Cartwright and Longuet-Higgins* [1956] and *Gumbel* [1958]; see also [*Solnes*, 1992], whom I follow in the subsequent presentation. For a general stationary Gaussian process with the power spectrum $S(f)$, they give the following expressions for the important time-domain parameters of the process, namely, the frequency of maxima μ and the bandwidth parameter ϵ^2 ($0 < \epsilon^2 < 1$):

$$\mu = m_4/m_2; \qquad \varepsilon^2 = 1 - m_2^2/m_0 m_4 \tag{1}$$

where m_i, $i = 0$, 2 and 4 are the spectral moments:

$$m_i = \int_{-\infty}^{\infty} f^i S(f)\,df. \tag{2}$$

The μ parameter is the mean number of maxima per unit time, so that the value of μ^{-1} is the characteristic time interval between adjacent maxima of the process (note that "maxima" here are local maxima, never assumed to be positive). The ϵ^2 value characterizes the relative bandwidth of the process: it is near zero for a narrow-band signal and approaches unity if the bandwidth is very large. To estimate the empirical μ and ϵ^2 values, I replace the power spectrum in (2) by squared modulus of the Fourier spectrum of a record. The evident alternative is the direct time domain determination of μ; however, trying this approach, I met with large underestimation of μ: a significant percent of maxima is lost. This loss is the result of working with digitized data (sampling frequency of 100 Hz was used), whereas the theory used has been derived for continuous signal. Using interpolated signal with artificially increased sample rate, I could improve the agreement between the result of integration in frequency domain (which I consider correct and use in the following) and of the direct count in the time domain but considered this unnecessary.

Now let the process variance σ^2 (or $a_{rms} = \sigma$) be unity, then the probability density of maxima is given by

$$p(x) = \frac{\varepsilon}{\sqrt{(2\pi)}}\left[\exp\left(\frac{-x^2}{2\varepsilon^2}\right) + \varphi\exp\left(\frac{-x^2}{2}\right)\int_{-\infty}^{\varphi}\exp\left(\frac{-u^2}{2}\right)du\right] \tag{3}$$

where $\varphi = (x/\varepsilon)(1-\varepsilon^2)^{1/2}$ (see Figure 1a). Notice that at $\epsilon^2 = 0$ (infinitely narrowband), $p(x)$ degenerates to the Rayleigh density and at $\epsilon^2 = 1$ (infinitely wideband) degenerates to the Gaussian one. To facilitate the comparison of various distributions by eye introduce the complementary cumulative distribution function (CCDF) $Q(x) = 1 - P(x) = \int_x^{\infty} p(u)\,du$, and take log$Q$ and x^2 as the new variables (see Fig. 1b). At $\epsilon^2 = 0$ (Rayleigh-law case), the log$Q$ versus x^2 relationship becomes exactly linear in this scale (which is, essentially, the "exponential law probability paper").

At unit variance, or unit rms amplitude, the largest absolute extremum over a segment is identical to peak factor PF or the peak to rms amplitude ratio. To estimate PF, I must consider the set of extrema over a segment of duration T. In our actual case of sufficiently wide bandwidth (ϵ^2 is estimated to be near 0.8 below), I can assume the independence of extrema, because the correlation time of the sequence of extrema is of the order of (bandwidth)$^{-1}$. (However, it must be accounted for in the narrow band case, for example in the estimation of response spectrum, when the number of independent extrema is about $2T\times$(bandwidth) rather than $2T\times$(central frequency) [*Gusev*, 1990].) Thus I ignore the statistical dependence between successive extrema, and assume $N = 2\mu T$ to be the number of independent extrema and also assume N to be large. Then, for squared PF, the (cumulative) distribution function mean and variance are

$$P(PF^2) = \exp(-N_{\text{eff}}\exp(-PF^2/2)) \tag{4}$$

$$\mathrm{E}(PF^2) = 2(\ln N_{\text{eff}} + \gamma) \tag{5}$$

$$\sigma^2(PF^2) = 2\pi^2/3 \tag{6}$$

whereas for PF proper

$$\mathrm{E}(PF) = (2\ln N_{\text{eff}})^{1/2} + \frac{\gamma}{(2\ln N_{\text{eff}})^{1/2}}; \quad \sigma^2(PF) = \frac{\pi^2}{12\ln N_{\text{eff}}} \tag{7}$$

where $\gamma \approx 0.5772$ is the Euler's constant and $N_{\text{eff}} = (1-\varepsilon^2)^{1/2}N$. The effect of the correction factor $(1-\varepsilon^2)^{1/2}$ can be seen as the asymptotic relative shift to the left of curves of Figure 1b at small Q, with respect to the straight line for $\epsilon^2 = 0$.

The above expressions suggest several ways to compare observations with theory. In particular, one can correlate the

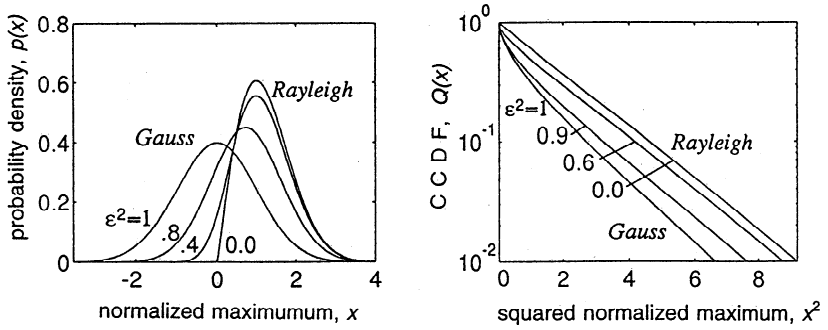

Figure 1. (a) Distribution density $p(x)$ (3), for values of maxima of a realization of a Gaussian process with zero mean and unit variance, for various values of the bandwidth parameter ϵ^2. The case $\epsilon^2 \to 0$ is the narrowband one; then all (local) maxima are positive and the probability of a negative maximum approaches zero. The case $\epsilon^2 \to 1$ is that of increasingly broad band. In this case, the high-frequency component produces very frequent local maxima, whereas the low-frequency component produces long-period oscillation that makes these maxima to be positive or negative with equal (asymptotically at $\epsilon^2 \to 1$) probability, as shown by the symmetrical bell curve for $\epsilon^2=0$. (b) Corresponding complementary distribution functions (CCDF) $Q(x)$. Note the change of the abscissa scale.

observed PF^2, denoted PF_o^2, with its theoretical mean value, PF_t^2; and one also can compare the PF_o^2 versus ln$N$ trend with the almost linear PF_t^2 versus ln$N$ relationship (5) as well. With this latter end, one can split the analyzed segment into, for example, 2, 4, 8 and 16 equal parts, determine PF_o^2 for each subsegment, and then estimate $K = d(PF^2)/d\ln N_{\text{eff}}$ by linear regression. The theoretical estimate of K, denoted K_t, can be obtained by linear approximation to (5) treated as a PF^2 versus ln$N$ relation; practically, K_t is very near to 1.95. Use of two independent parameters simultaneously enables one to improve the reliability of results.

To construct normalized values based on PF and K, I reduce them to zero mean and unit variance obtaining the new parameters δPF and δK:

$$\delta PF = (PF_o^2 - PF_t^2)/\sigma(PF_t^2) \quad (8)$$

$$\delta K = (K_o - K_t)/\sigma(K_t). \quad (9)$$

For the calculation of δPF, (5) and (6) immediately provide the mean and variance of PF^2 for the null hypothesis. For δK, the variance of K can easily be estimated during the regression procedure described above, if one replaces the value of the residual variance in this procedure by its theoretical value (6) for the null hypothesis; the result is $\sigma(K) = 0.6151$ for my actual choice of $1+2+4+8+16 = 31$ PF values. The way to find the mean value K_t has already been explained.

The normalized parameters δPF and δK are convenient for the preliminary step of data analysis but cannot give accurate significance levels because the distribution defined by equation (4) is different from Gaussian. If such significance value is needed for an individual anomalous PF value, one can directly use (4); for K, however, the accurate distribution is not available. The particular case considered below is when n out of m PF values are large and corresponding significance values are known (based on evaluation of (4)). As usual in such a case, one can employ the binomial distribution to find the joint significance.

The two introduced parameters will be inevitably strongly intercorrelated because a large PF value will make K also large; whereas I would prefer the parameters to be as independent as possible. To reduce this unwanted dependence, I substitute the PF value with its estimate from data. Note that in calculation of K, the PF value would appear several times: for the full segment and for all subsegments covering the time moment of the peak. In all these positions, I replace the true PF value by its estimate through the second largest peak, with the corresponding correction of $+2\ln 2$, equal to the mean difference between the rms-normalized largest and the rms-normalized second largest. This modification, applied for all analyzed records, markedly reduces the correlation between δPF and δK.

As I assume the data to be nonstationary, I can expect that correlation takes place between relatively distant extrema, reflecting the temporal structure of "modulation." This correlation must disappear if stationarity is genuinely reconstructed; this gives us a means to qualitatively check the efficiency of the "stationarization" procedure.

Data Selection and Processing

Near-source records of large-magnitude earthquakes accumulated by Mexican accelerograph networks [see, e.g., *Singh et al.* 1990] provide an appropriate data set for this study. I used the CD-ROM [*Seekins et al.* 1992] as the immediate data source for acceleration data files. In the data selection (Table 1), I chose all reliable horizontal components for all five large ($M > 7$) earthquakes covered by the data set, at distances up to 110 km. The distances given are from the source center, in the case of the September 19, 1985, event - from the center of the nearest subevent.

The data processing consisted of the following stages:

1. Apply high-pass filter with the cutoff of 0.4 Hz. This is the practical means to secure zero mean of the analyzed data. The choice of the cutoff value, be it sufficiently small, is more or less arbitrary.

2. Determine the data window. The aim of this stage is to isolate probably source-related, maximum segment of the record. After several trial and error steps, a fully automatic windowing procedure was designed, giving results which look reasonable, producing limited duration differences be-

Table 1. Strong Earthquakes and Recording Stations

Event Date	Epicenter		Depth,	M_s	Station
	°N	°W	km		(Distance, km)
March 14, 1979	17.46	101.46	20	7.6	SICC(95), VILB(98)
Oct. 25, 1981	17.75	102.25	16	7.3	SICC(21)
Sept. 19, 1985	18.14	102.71	16	8.1	CALE(17), VILE(44), ZACA(44), IN12(64), INMD(64), UNIO(68), PAPN(102)
Sept. 21, 1985	17.62	101.82	16	7.6	AZIH(36), PAPN(63), UNIO(73), ZACA(95), INMD(101), SUCH(108)
April 30, 1986	18.40	102.97	21	7.0	CALE(48), ARTG(75)

Data listed are taken mostly from *Singh et al.*[1990] in whose work additional details can be found.

tween components of the same record, and making both component durations not too different from the inverse-corner-frequency durations used by *Singh et al.* [1990]. The procedure consists of two steps: (1) take the absolute value of the record and smooth it by the low-pass filter with the cutoff of $f_{dw} = 0.15$ Hz and (2) define the data window as a segment where the above function exceeds 40% of its own maximum. Such an operation excludes manual adjustment which can unintendedly bias results. (Using the threshold value of 30% changes the results only marginally.)

3. Cut out the segment within the data window; normalize it to produce unit average variance. Determine observed and theoretical μ, μ_o, and μ_t.

4. Make a sequence of squared extrema. Find their correlation function using the sequential number as argument.

5. Form CCDF of data. Find PF_o, PF_t and δPF.

6. Subdivide the initial segment into 2, 4, 8 and 16 subsegments and find the *PF* value for each one. For the full segment, and for all subsegments covering the time point where the full-segment *PF* is situated, replace their *PF* value by its estimate through the second largest peak. In calculating subsegment *PF*, normalize each peak by the rms value determined for this particular subsegment. (Alternative normalization, by the whole-segment rms value used on each segment, does not change the results significantly.) Find K_o, K_t, and δK.

After application of the above procedure to the original records, they were "stationarized," and the procedure was applied again to the thus modified data (less step 2, the data window of the original record was preserved). The stationarization procedure applied is essentially that of the automatic gain control. I determine the signal modulus, smooth it in the frequency domain using a certain upper cutoff f_{sm}, and then divide the input record by its smoothed modulus. The critical issue here is the choice of f_{sm}. Use of large f_{sm} (short time constant) results in "over-stationary" output; its amplitude distribution has unrealistic, over-suppressed tails, which are "lighter" than in the Gaussian case. In the opposite case, with too small f_{sm} (too long smoothing time), I simply do not attain the goal of stationarization.

Unfortunately, no ready theoretical means are known to me to formalize this choice (whereas its effect on the conclusions of the analysis may be radical). Below I applied a new, specially designed approach, based on the idea that, as is well known, the estimates of the "center" of distribution that are based on the median are not very sensitive to heavy-tailed data contamination. Thus, if, in the case of the genuine quasi-stationary process, I would tune the f_{sm} value in such a manner that the median δPF value over the whole data set would be equal to zero, I could believe that this f_{sm} value is the adequate one for performing "true" stationarization. Thus the checks applied to thus modified data and aimed at finding a non-Gaussian behavior would give a negative result for this case, and if I find the opposite, I can consider the non-Gaussian behavior to be successfully identified. Such an identification technique is rather reliable, because the procedure described is in fact notably conservative. To make this clear, notice that if the heavy-tailed component is present, its effect on the calculation on the median δPF of modified (or raw) data is to shift this median to positive values. To suppress this effect, a higher-than-adequate value of f_{sm} will be found and applied in the process of the tuning explained above. As a result, too much stationarization will take place, and "non-Gaussian" peaks will be additionally suppressed. Thus the hypothesis testing will take place in the conditions of this additional suppression of peaks. If, despite this bias, the null hypothesis will be rejected, this will take place on the lower significance level (i.e., with more certainty) than the level calculated formally.

Below I shall indeed reject the hypothesis of quasi-stationary process using the described procedure. After rejection of the null hypothesis, it will be reasonable to decrease somewhat the f_{sm} value used, as definitively overestimated. However, it was difficult to determine the accurate value of this decrease. Based on informal criteria, I reduced the f_{sm}, from the value determined in the "tuning process", to 60% of it. This second f_{sm} value was considered as more adequate for determination of real relative proportion of non-Gaussian records.

Results of analysis

I begin with two detailed examples of the data analysis. The first one is a prominent case of non-Gaussian behavior. Figures 2 and 3 show details of the analysis procedure applied in the identical manner to the original record and to its modified/stationarized version, correspondingly . Figures 2a and 3a show input accelerograms (original and modified). In Figure 2a, one can also see two smooth curves showing low-pass filtered modulus of signal, multiplied by arbitrary positive or negative factor. The top curve corresponds to the modulus obtained using the cutoff frequency $f_{dw} = 0.15$ Hz; it was used to define the data window, also shown as vertical dashed lines. The lower curve corresponds to $f_{sm} = 0.06$ Hz and was used within the stationarization procedure. Figures 2c and 3c are the autocorrelation plots for subsequent squared signal peaks treated as a time series. On Figure 2c one can see, in addition to the trivial central peak, a wide positive rise; this rise is practically suppressed for the modified record, showing the decorrelative effect of stationarization. Figures 2b and 3b show the theoretical and the empirical CCDF. In Figure 2b, one can see very prominent deviation of the extremal part of distribution from the expectations of stationary Gaussian case. In Figure 3b, this deviation is somewhat reduced but still is prominent. Figures 2d and 3d illustrate the relationship between the PF^2 value and the number of extrema on a subsegment. For example, the rightmost point corresponds to the whole segment analyzed,

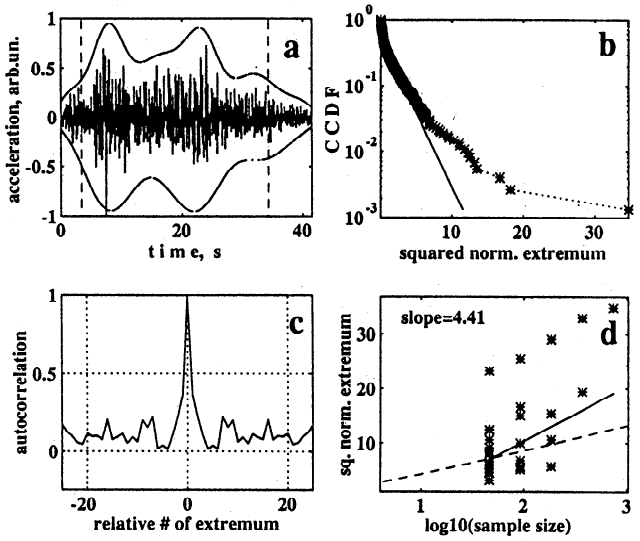

Figure 2. Statistical analysis of the original accelerogram of the September 21, 1985, event at the station UNIO, component N-S. (a) Time history. (b) CCDF of its squared extrema. (c) Autocorrelation of extrema. (d) Squared extremum versus sample size, that is the number of subsequent extrema on a subsegment. See text for details.

containing about 750 extrema. In the leftmost group of 16 points sitting on the same vertical line, each point corresponds to one of 16 equal subsegments. Each of the subsegments contains an approximately equal number of extrema, near to 750/16 $\approx$ 47. As the normalization by rms amplitude has been done on each segment separately, the largest (over subsegments of a similar size) maximal squared extremum somewhat varies for different sizes of a subsegment. The empirical linear fit (giving K_o, solid line) does not immediately corresponds to plotted data, it is practically always more gradual, because, as explained above, in calculation of this line, I substituted the full-segment-maximum value by its estimate through the second maximum. This

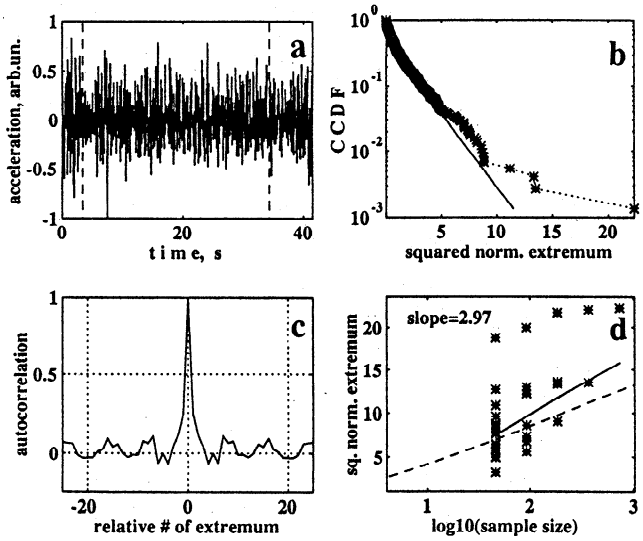

Figure 3. Same as Figure 2, except modified ("stationarized") version of the same accelerogram. The value f_{sm} = 0.06 Hz was used in the stationarization procedure.

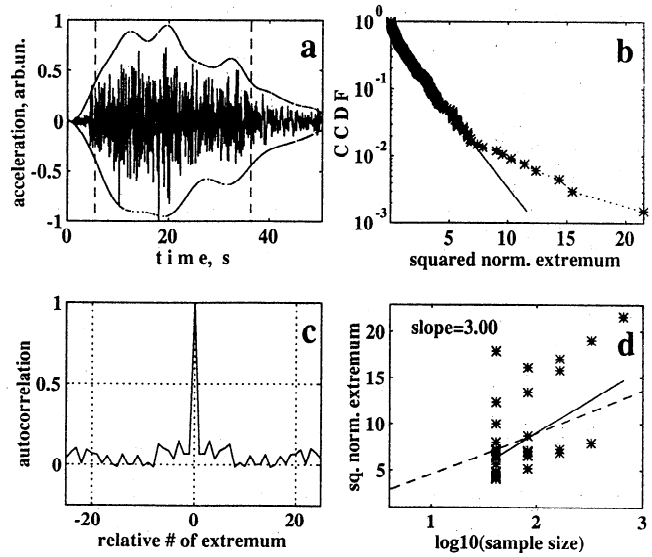

Figure 4. Same as Figure 2, for the September 19, 1985, event at the station CALE, component E-W.

empirical fit can be compared to the expected theoretical dependence (dashed line). Here again one can see clear deviations from the null hypothesis (when $K \approx 2$). Another example (Figures 4 and 5) is the same data for an assumedly quasi-stationary Gaussian case. One can see that, for the original record (Figure 4), extremal peak statistics and K value are again far from expected for the stationary Gaussian case, but now these deviations are almost completely suppressed by stationarization (Figure 5).

Now consider the whole data set. For the basic parameters of the signal, average values and standard deviations over the data set are $\epsilon = 0.81 \pm 0.09$ and $\mu = 14.2 \pm 5.2$ Hz. The key data pertinent to the testing of the null hypothesis for individual records are given in Table 2, both for original data and for modified ones for f_{sm} = 0.10 Hz. In Figure 6, I show the δPF-δK scatter. One can see, primarily, the radical deviation of δPF and δK values of original

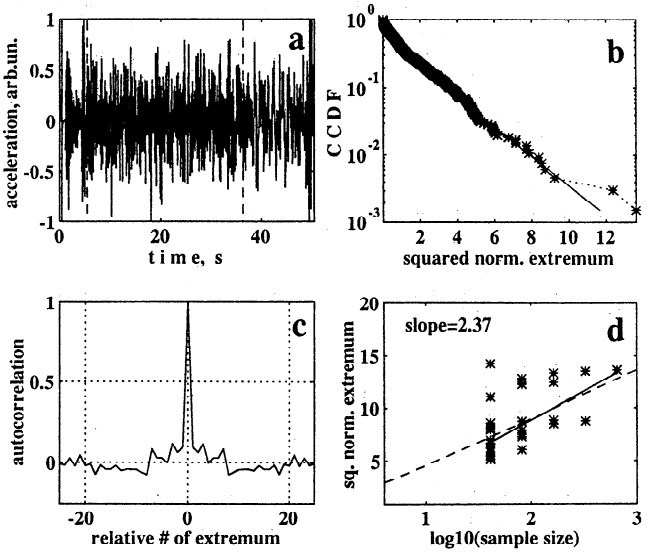

Figure 5. Same as Figure 3, for the September 19, 1985, event at the station CALE, component E-W.

Table 2. Parameters of Original and Modified (f_{sm} = 0.10 Hz) Records

Station Component	Duration, s	*PF* Theo.	*PF* Orig.	δPF Orig.	*K* Theo.	*K* Orig.	δK Orig.	*PF* Mod.	δPF Mod.	*K* Mod.	δK Mod.
				March 14, 1979							
SICC/N	8.7	3.00	3.28	0.69	1.84	2.05	0.34	2.84	-0.39	1.69	-0.24
SICC/E	8.9	3.04	3.44	1.00	1.85	2.49	1.04	3.00	-0.12	2.00	0.27
VILB/NE	9.1	3.15	3.10	-0.12	1.91	1.38	-0.87	2.76	-0.91	1.84	-0.11
VILB/NW	6.8	3.01	3.05	0.10	1.87	1.92	0.08	3.28	0.67	2.21	0.56
				October 25, 1981							
SICC/N	13.8	3.22	3.33	0.27	1.92	2.38	0.74	2.88	-0.87	2.03	0.18
SICC/E	16.1	3.25	3.02	-0.55	1.90	1.66	-0.40	2.83	-1.00	1.49	-0.67
				September 19, 1985							
CALE/N	25.3	3.56	3.77	0.60	1.98	2.42	0.71	3.04	-1.35	1.23	-1.23
CALE/E	30.8	3.61	4.51	2.84	1.99	3.00	1.94	3.58	-0.09	2.52	0.86
IN12/NE	42.2	3.72	5.01	4.40	1.99	3.90	3.11	4.04	0.99	2.59	0.99
IN12/NW	41.7	3.69	4.10	1.26	1.98	2.57	0.95	3.46	-0.64	2.16	0.30
INMD/NE	34.9	3.84	4.87	3.50	2.00	4.52	4.09	4.18	1.09	2.04	0.07
INMD/NW	36.5	3.84	4.07	0.71	2.00	2.22	0.37	3.30	-1.47	1.14	-1.40
PAPN/N	13.8	3.46	4.69	3.91	1.97	2.70	1.18	3.75	0.80	1.65	-0.52
PAPN/E	11.3	3.41	3.43	0.05	1.97	2.90	1.52	3.21	-0.51	1.98	0.02
UNIO/N	25.3	3.50	3.77	0.77	1.97	2.46	0.78	3.74	0.70	1.63	-0.56
UNIO/E	24.6	3.51	3.18	-0.85	1.97	1.81	-0.26	3.09	-1.09	1.39	-0.95
VILE/N	46.5	3.55	4.92	4.52	1.98	3.55	2.56	3.74	0.48	1.36	-1.00
VILE/E	39.5	3.58	6.02	9.12	1.98	4.60	4.26	3.94	1.00	1.69	-0.48
ZACA/N	39.4	3.58	4.30	2.22	1.98	4.09	3.41	3.13	-1.24	1.65	-0.55
ZACA/E	41.8	3.62	3.88	0.76	1.99	2.55	0.92	4.11	1.44	1.13	-1.39
				September 21, 1985							
AZIH/N	14.7	3.38	4.14	2.22	1.96	3.49	2.48	3.37	-0.04	2.23	0.43
INMD/NE	30.3	3.79	4.37	1.86	1.99	3.12	1.82	4.28	1.52	2.37	0.60
INMD/NW	32.3	3.81	4.75	3.14	1.99	4.40	3.91	3.85	0.11	2.01	0.02
PAPN/N	7.4	3.35	3.95	1.73	1.96	2.99	1.67	3.61	0.71	2.37	0.67
PAPN/E	7.2	3.38	4.07	2.01	1.96	2.19	0.37	3.73	0.96	1.95	-0.02
SUCH/N	6.2	2.88	3.15	0.64	1.85	2.48	1.02	2.90	0.05	2.25	0.68
SUCH/E	9.5	3.14	6.03	10.33	1.92	3.41	2.43	5.31	7.14	2.37	0.76
UNIO/N	31.1	3.62	6.02	9.03	1.99	4.41	4.07	4.85	4.04	2.92	1.52
UNIO/E	28.0	3.53	6.64	12.35	1.98	4.17	3.56	4.42	2.73	1.94	-0.06
ZACA/N	21.5	3.34	5.21	6.22	1.95	2.95	1.62	3.55	0.59	2.37	0.69
ZACA/E	18.4	3.32	5.07	5.71	1.94	5.04	5.04	3.89	1.57	3.56	2.64
				April 30, 1986							
ARTG/N	8.8	3.26	3.06	-0.48	1.95	1.87	-0.13	2.73	-1.26	1.37	-0.94
ARTG/E	11.6	3.34	3.85	1.43	1.96	2.71	1.22	3.30	-0.08	1.60	-0.57
CALE/N	11.3	3.22	3.46	0.63	1.94	2.69	1.22	3.14	-0.22	2.32	0.62
CALE/E	10.8	3.27	4.96	5.42	1.95	1.99	0.06	4.04	2.17	1.29	-1.06
Average		3.42	4.24	2.78	1.95	2.95	1.62	3.57	0.50	1.95	0.00
S. d.		0.25	0.97	3.29	0.04	0.95	1.52	0.60	1.68	0.53	0.87

Theo., theoretical value; Orig., from original record; Mod., from stationarized record.

records (crosses) from what could be expected with the null hypothesis. The square plotted on the graph represents $\pm 2\sigma$ boundaries, approximately corresponding to 2.3% quantiles if one assumes the normal distribution. Acccurate quantiles, based on (4), are about 3% for δPF and between 2.3% and 3.1% for δK. In the following, I will consider a point anomalous if it is outside these boundaries. Sixteen of 35 δPF values are above 2σ, against expected about 1 of 35. Therefore the null hypothesis (the model of a segment of the stationary Gaussian process) can be rejected with confidence. The ratio, (16-1)/35 $\approx$ 43% is the rough but rather stable estimate of the fraction of anomalous component. For example, shifting the threshold value to 1.5σ gives 19 anomalous cases against about 3 expected for the null hypothesis, making practically the same estimate, (19-3)/35 = 47% of anomalous component. For δK values, I have likewise 11 points above 2σ against 1 expected, and for both criteria combined, again 16 (now against 2), so that δK parameter seems less powerful in the case studied; it merely confirms the conclusions based on δPF by an independent check.

Now I can test the alternative null hypothesis of the quasi-stationary Gaussian process. As explained above, to be on the conservative side, I put this hypothesis in a favorable position, and specially chose the value of cutoff frequency f_{sm}(=0.10 Hz) in such a manner that medians of the resulting δPF and δK distributions are near to zero: they are equal to 0.11 and 0.02, correspondingly. Even in so tough a test, five points are outside the "2σ square" on the plot. Furthermore, two δPF values are above 4σ. In terms of the significance value, the three most prominent *PF* outliers have probability values 5.5×10^{-5}, 3.2×10^{-3}, and 1.8×10^{-2}. If one considers one, two or three most prominent of these outliers as anomalous, the significance values associated with their occurrence within the sample of the size 35 are about 0.002, 0.007, and 0.03, correspondingly. Therefore the strongly non-Gaussian component of data still reveals itself sufficiently clearly even in a deliberately unfavorable test, and I can now reliably reject the second null hypothesis, that of the quasi-stationary Gaussian process. (To avoid confusion, notice that $\approx$55% of data, which are "stationary

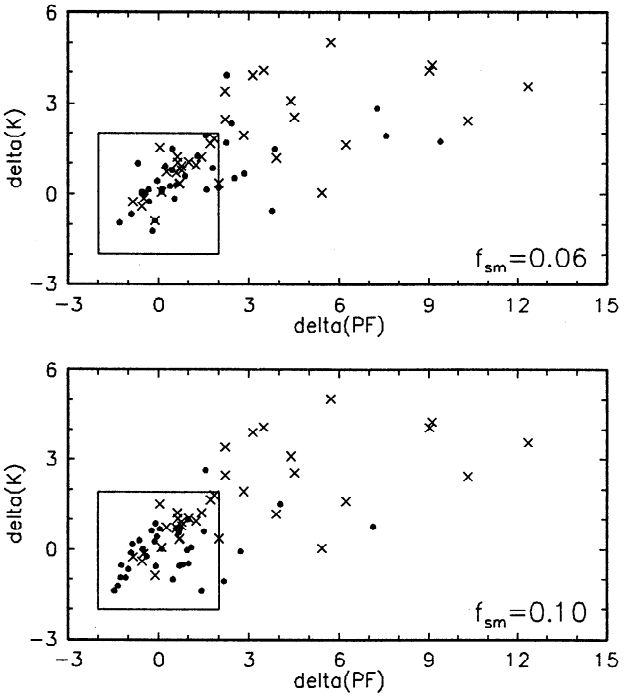

Figure 6. Scattergram of two signal parameters: δPF versus δK. Crosses indicate original records; dots indicate stationarized records for two values of the filter cutoff: (top) f_{sm} = 0.06 Hz and (bottom) f_{sm} = 0.10 Hz. The square is the $\pm 2\sigma$ field for the null hypothesis.

Gaussian" based on the first test, are at the same time "quasi-stationary Gaussian" because constant rms amplitude is a particular case of variable one.)

Again, I can assume the data to be a mix, now of quasi-stationary Gaussian and non-Gaussian components and try to estimate, at least very crudely, the actual fraction of the non-Gaussian component. The statistics for f_{sm} = 0.10 given above leads to an estimate of (5-2)/35 $\approx$ 10%, based on the number data outside the 2σ square. This is essentially the minimal estimate because of strict conditions used to select the f_{sm} value. If I decrease f_{sm} down to 0.06 Hz, I obtain a quite different version of a δPF-δK scatter-plot for stationarized signal (Figure 4, bottom). Here, 10 points are outside the 2σ square, and this may indicate the non-Gaussian (heavy-tailed) contribution of 25-30%. Although subjectively I consider this version as more realistic, I do not have at hand an adequate technique to select optimal f_{sm} value and can treat this estimate only as tentative. In the following, I will consider the fraction value of 10-25% as a preliminary order-of-magnitude estimate.

I can now sum up the above observations as follows:

1. More than 40% of original accelerograms clearly deviate from the model of the stationary Gaussian process.

2. Among them, a significant part deviate from the model of the locally/intrinsically Gaussian process with time-varying variance ("quasi-stationary Gaussian process") as well.

3. If the data set is considered as a mix of records with different distribution functions, then about 50-60% are near to stationary Gaussian, and among the other 40-50%, some are near to quasi-stationary Gaussian, and the rest are heavy-tailed non-Gaussian.

4. The fraction of heavy-tailed records can be tentatively estimated as 10-25%.

Note, however, that the three-type grouping given above is somewhat arbitrary; another approach to interpretation of the same data can be imagined, when all the three data types are generated by a certain more general statistical model.

Discussion

The revealed deviations from the stationary Gaussian case can be compared with the results of *Hanks and McGuire* [1981], who give abundant data on accelerogram *PF*. In their Figure 6b, the data on the 1971 San Fernando earthquake are given. If one uses (6) to determine the expected data scatter and compares it with the plotted data, one can find that more than 20% of data points on this plot are above the upper 2σ boundary. Furthermore, the eight nearest points, that correspond to the hypocentral distances below 28 km, give the average *PF* of about 4.7, (against theoretical 3.4$\pm$0.3), suggesting a rather clear deviation from the stationary Gaussian case. This observation is supported by the data on other California earthquakes given by *Hanks and McGuire* [1981] Figure 6c, where again about 20% of the data are above the expected upper 2σ level. Therefore, the conclusions of *Hanks and McGuire* [1981] regarding the applicability of the Gaussian model cannot be considered as universal: the data presented here as well as their own indicate that, roughly, 20-40% of strong-motion accelerograms at distances below 100-120 km can be expected to clearly deviate from the stationary Gaussian model. In numerical terms, however, the deviation is not dramatic, empirical peak factor being, on the average, about 25% above the one expected in the Gaussian case. Of 35 cases studied here, deviations above +50% are found in seven.

All this means that in calculations for the stochastic strong motion prediction, the use of peak factor estimates based on the stationary Gaussian model may underestimate peak acceleration value, by something like 25% on the average. Much larger underestimation can be expected if this model is used to estimate a rare event, such as peak acceleration with the repeat time of, say, 1000 years. Note, however, that the possible effect of such a bias on response spectrum or on nonlinear response of a building or a soft ground layer may be somewhat smaller than that on peak acceleration.

An important feature of the obtained results is that "Gaussian" and non-Gaussian records can be produced by the same source. It seems unlikely that the statistical properties of radiation depend on the propagation direction or instrument orientation. The more probable explanation of the lack of a systematic picture is related to the properties of propagation path. At the hypocentral distances 30-100 km, small earthquake records often do not show short body wave pulses whose duration can be related to the source process. Instead, they often have long *S* group whose probable origin is related to the (forward) scattering and/or multipathing and/or site resonance. The record of a local earthquake can be considered as an empirical medium response. Now assume that a high-frequency source generates a heavy-tailed random source function. Convolution of such a function with such a medium response will tend to suppress isolated strong spikes and may well result in "normalization," that is, in transformation of the signal distribution function in

such a manner that it approaches the Gaussian one. This point of view is supported by decrease of "*PF* anomaly" with distance for the San Fernando earthquake data of *Hanks and McGuire* [1981] as well as by quite similar decrease (evident but not studied in any detail here) in Mexican data viewed in wider distance range as compared to the one studied here, for example, 0-300 km. The explanation can be formulated in another way: to form a non-Gaussian record, some of the rays connecting bright spots of a source with the station must be "weak-scattering"; if they are all "strong-scattering," a Gaussian record arises. With increasing overall distance, the probability of a "weak-scattering ray" decreases, resulting in the observed distance dependence of *PF*. The stochastic character of this picture explains why one can find high δPF values even at relatively large distances despite the overall trend.

In parallel with $K = dPF^2/d\ln N$, it was interesting to determine values of $\alpha^{-1} = d\ln PF/d\ln N$. As was shown by *Gusev* [1989], α gives the estimate of the exponent of the power law distribution for amplitudes of pulses which are supposedly generated by small asperity subsources and add together to form an accelerogram. For original (not stationarized) records, $\alpha^{-1} = 0.21 \pm 0.05$. This value characterizes a typical record at 30-100-km distance. It will probably decrease nearer to the source, so that the corresponding α value of 5 can be considered as an upper bound.

Conclusion

A simple statistical technique was designed and applied to the maximum segment of near-source accelerograms of large-magnitude Mexican earthquakes. For 40-50% of them, the null hypothesis of the stationary Gaussian process is rejected based on statistics of extrema. Another record model, of the quasi-stationary Gaussian process, agrees with a larger part of the data (roughly 75-90%) yet cannot explain the whole data set: for 10-25% of it, a heavy-tailed non-Gaussian distribution model is preferable. Thus each of the mentioned models reflects properties of some fraction of data. From the engineering point of view, the deviations of the observed peak factor values from ones expected in a stationary Gaussian case are moderate to large (25% on the average and above 50% in 7 of 35 cases). The non-Gaussian behavior of accelerogram amplitudes suggests the analogous properties of source radiation, generally supporting *Gusev*'s [1989] concept of heavy-tailed distribution for radiated acceleration peaks.

Appendix: Derivation of Key Formulas for the Narrowband Case

The derivation of the cited equations (3)-(7) is lengthy and cannot be repeated here even briefly, but for the narrowband case, the derivation is much shorter and can be reproduced here. Let $x(t)$ be a realization of a narrowband Gaussian process, then its Hilbert transform $y(t) = H[x(t)]$ is uncorrelated with it and is also Gaussian with the same variance parameter. Analytical signal $a(t) = x(t)+iy(t) = A(t)\exp(i\phi(t))$ contains nonnegative instant amplitude $A(t)$ and phase $\phi(t)$, mutually independent. For the narrowband case, extrema of $x(t)$ are near to moments $\phi(t) = k\pi$, $k = \ldots -2, -1, 0, 1, 2, \ldots$. Then, the distribution of extrema coincides with that of $A(t)$ and the distribution of squared extrema - with that of $A^2(t) = x^2(t)+y^2(t)$. (Because of lacking correlation between $A(t)$ and $\phi(t)$, extrema can be treated as if they occur at random time moments.) Now the sum of two squared normal variates of identical variance $\sigma^2 = x^2_{rms}$ is χ^2_2 or exponentially distributed with the parameter $2\sigma^2$, so that $z = x^2_{peak}/2\sigma^2$ follows the standard exponential cumulative distribution $P(z) = 1-\exp(-z)$ (and $z^{1/2}$ has the Rayleigh distribution). To derive *PF* distribution, assume N independent identically distributed exponential variates z_1, z_2, ..., z_k, ..., z_N. The distribution function of the maximum z_M among z_k is $P_N(z_M) = (P(z_M))^N = (1-\exp(-z_M))^N$. At large z_M, let $\exp(-z_M) = 1/q$, then $P_N(z_M) = [(1-1/q)^q]^{N/q} \approx \exp(-N\exp(-z_M))$. Now note that $PF^2 = 2z_M$ and obtain an analog of (4). To obtain the analog of (5) for median, set $P_M(z_{M,50\%}) = 0.50$ and find $z_{M,50\%} = \ln N - \ln\ln 2$ or $(PF^2)_{50\%} = 2\ln N - 2\ln\ln 2$.

Acknowledgments. The work was stimulated by a discussion with T. Hanks. I am indebted to P. Spudich, T. Sato, and D. Perkins for suggestions that helped to improve the manuscript.

References

Boore, D., Stochastic simulation of high-frequency ground motions based on seismological models of the radiated spectra, *Bull. Seismol. Soc. Am., 73*, 1865-1894, 1983.

Cartwright, D. E., and M. S. Longuet-Higgins, The statistical distribution of maxima of a random function, *Proc. R. Soc. London A, 237*, 212-223, 1956.

Gumbel, E. J., *Statistics of extremes*. Columbia Univ. Press, New York, 1958.

Gusev, A. A., Descriptive statistical model of earthquake source radiation and its application to an estimation of short-period strong motion, *Geophys. J. R. Astron. Soc., 74*, 787-808, 1983.

Gusev, A. A., Multiasperity fault model and the nature of short-period subsources, *Pure Appl. Geophys. 130*, 635-660, 1989.

Gusev, A. A., A preliminary version of design seismic load for Petropavlovsk-Kamchatsky (in Russian), *Voprosy Inzhenernoi Seismologii*, iss. 31, pp. 67-85, Nauka, Moscow, 1990.

Hanks, T. C., and R. K. McGuire., The character of high-frequency strong motion, *Bull. Seismol. Soc. Am., 71*, 2071-2095, 1981.

Seekins, L. C., A. G. Brady, C. Carpenter and N. Brown, *Digitized strong-motion accelerograms of North and Central American earthquakes 1933-1986* [CD-ROM], U.S. Geol. Surv. Digital Data Series DDS-7, 1992.

Singh, S. K., E. Mena, J. G. Anderson, R. Quaas, and J. Lermo. Source spectra and RMS accelerations of Mexican subduction zone earthquakes, *Pure Appl. Geophys., 133*, 447-474, 1990.

Solnes, J., *Theory of stochastic processes and random vibration with application to geophysics and engineering*, Inst. Geofis., Univ. Nac. Autón. Méx., Mexico City, 339 pp., 1992.

A. A. Gusev, Institute of Volcanic Geology and Geochemistry, Russian Academy of Science, 9 Piip Boulevard, 683006 Petropavlovsk-Kamchatsky, Russia. (e-mail: seis@volgeo.kamchatka.su)

(Received August 29,1995; revised February 29, 1996; accepted March 6, 1996.)