Pure Appl. Geophys.
© 2015 Springer International Publishing
DOI 10.1007/s00024-015-1194-7

Pure and Applied Geophysics

Reply to: "Comment on Pisarenko et al., 'Characterization of the Tail of the Distribution of Earthquake Magnitudes by Combining the GEV and GPD Descriptions of Extreme Value Theory'" by Mathias Raschke in Pure Appl Geophys (2015)

V. F. Pisarenko,[1] A. Sornette,[2] D. Sornette,[3,4] and M. V. Rodkin[1,5]

1. Raschke writes (Conclusions, line 4–10):
   In summary, I advise against using the procedures as applied by Pisarenko et al. GPD and GEVD work well only for the extremes of TED or GTED when the block size is very large and/or the threshold is very close to the upper bound magnitude. The crucial point of earthquake magnitudes is the poor convergence of their upper tail to the GPD.

As a matter of fact, we did not recommend in our paper to apply our procedures to the TED (truncated exponential distribution) model, which is advocated by Raschke. We applied our method to the Harvard catalog 1977–2006 and to the Fennoscandia catalog 1900–2005, and obtained a quite satisfactory result, in our view. Below, we are providing detailed arguments that support this statement.

We use the same notations as in (Raschke 2016). The data are the same as in (Pisarenko *et al.* 2014): the lower threshold $m_{\min} = 6.8$; sample size (number of main shocks exceeding $m_{\min}$) $n = 261$. We have fitted the GPD (Generalized Pareto Distribution) to this sample using the Moment Method (MM) recommended in our paper and obtain the following estimates:

$$\text{GPD}: \gamma = -0.214; \quad \sigma = 0.548.$$

In order to quantify the goodness of fit, we use two characteristics: Chi-square sum S and the Kolmogorov distance KD.

$$S = \sum_{k=1}^{11} \frac{(n_k - np_k)^2}{np_k};$$

$$p_k = F_{\text{GPD}}(\mathrm{d}_{k+1}) - F_{\text{GPD}}(\mathrm{d}_k);$$

$$F_{\text{GPD}}(x) = 1 - \left[1 + \frac{\gamma(x - m_{\min})}{\sigma}\right]^{-1/\gamma}, \quad (1)$$
$$m_{\min} \le x \le m_{\min} - \sigma/\gamma;$$

$$d_1 \ldots d_{12} = 6.80, 6.85, 6.95, 7.05, 7.15, 7.30, 7.50, 7.60, 7.70, 7.80, 8.00, \infty;$$

$$n_k = \#\{\mathrm{d}_k \le x_j < \mathrm{d}_{k+1}\}; \quad k = 1, \ldots, 11;$$

$x_1, \ldots, x_n$ are m_w magnitudes of the Harvard earthquake catalog 1977–2006; $m_w \ge 6.8$; $n_1 \ldots n_{11} = 16\ 45\ 43\ 26\ 33\ 34\ 15\ 13\ 8\ 17\ 11$.

Since we estimate two parameters, the number of degrees of freedom for S is equal to $11-3 = 8$.

$\text{KD} = \sqrt{n}|F_{\text{GPD}}(x|\gamma,\sigma) - F_n(x)|$, $F(x|\gamma,\sigma)$ is GPD with parameters $m_{\min}$, γ, σ; $F_n(x)$ is sample distribution function.

We obtain the following results:

We would like to thank Mathias Raschke for his detailed analysis of our paper and for his valuable remarks. At the same time, we disagree with some of his remarks and comments and would like to explain our point of view in the present note.

[1] Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences, Profsoyuznaya 84/32, Moscow 117997, Russia. E-mail: pisarenko@yasenevo.ru; rodkin@mitp.ru

[2] ETH Zurich, Swiss Seismological Service, 8092 Zurich, Switzerland.

[3] ETH Zurich, D-MTEC, Scheuchzerstrasse 7, 8092 Zurich, Switzerland. E-mail: dsornette@ethz.ch

[4] University of California, Los Angeles, CA 90095, USA.

[5] Institute of Marine Geology and Geophysics, Far East Branch, Russian Academy of Sciences, Yuzhno-Sakhalinsk 693022, Russia.

Published online: 23 December 2015

$S_{GPD} = 8.46$; p value (probability of exceeding S_{GPD}) = 0.392; $KD_{GPD} = 0.570$; p value = 0.569.

These values give us a solid foundation to accept as satisfactory the performed GPD fitting. Usually, a hypothesis is rejected when the p value is less than 0.1 or 0.05.

Although we did not refer in our paper to the use of the Truncated Exponential Distribution (TED) as suggested by Raschke, we decided to apply it here to the same sample for the sake of comparison. The obtained Maximum Likelihood (ML) estimates are

$$\text{TED}: \beta = 2.117; \quad m_{max} = 9.00.$$

The Chi-square sum S was calculated with the same $\{d_k\}$ and with p_k corresponding to TED:

$$p_k = \frac{\exp[-\beta(d_k - m_{min})] - \exp[-\beta(d_{k+1} - m_{min})]}{1 - \exp[-\beta(m_{max} - m_{min})]};$$
$$k = 1 \ldots 11.$$

We obtain the following results:

$S_{TED} = 14.28$ (p value = 0.0748);

$KD_{TED} = 0.751$ (p value = 0.351).

One can see that the TED provides a poorer goodness of fit compared to the GPD, especially in the Chi-square test. The exceedance probability of 7.5 % suggests that it is reasonable to reject the TED.

In Fig. 1, we compare the tail fitting both by the GPD (thick line) and the TED (thin line).

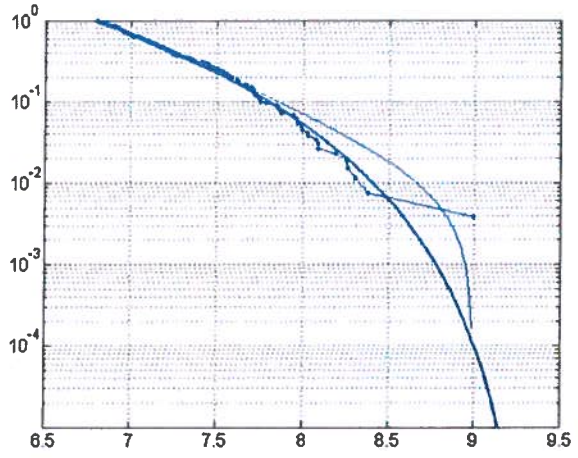

Figure 1
Empirical tail $1 - F_n(x)$ (*dots*), GPD tail (*thick line*), TED tail (*thin line*)

We see that the GPD gives a better fit to our data sample compared to TED, in particular in the range of extreme events.

We tried several other lower threshold values $m_{min} = 6.6$ ($n = 392$); 7.0 ($n = 174$); 7.2 ($n = 114$). In all cases, the results were very close to the estimates obtained for $m_{min} = 6.8$.

2. The main point of Raschke's comment is formulated below (Conclusion, lines 10–11):
   The crucial point of earthquake magnitudes is the poor convergence of their upper tail to the GPD.

This assertion is illustrated in Fig. 1 in (RASCHKE 2016) where several TED curves are compared with the uniform distributions (UD), which represent the limit case of the extreme value distributions for the corresponding TEDs. In statistics, the terms "convergence" and "convergence speed" typically refer to the analysis where the sample size n tends to infinity. Instead, Raschke arbitrarily chooses TED as the most adequate law for fitting and then compares TED versus UD. We would like to stress once more that we did not recommend in our paper to use TED as an approximation of the distribution of maxima. Comparing any fitting distribution with a distribution based on the extreme value index is a rather meaningless exercise, in our view, since that index is very unstable and depends on the smallest details of the asymptotic of the distribution function. For instance, one can perturb the original distribution by adding a very light tail that contains the total probability of, say, 10^{-10}. That tiny addition would practically not affect any statistical characteristic of TED, while it could dramatically change the extreme value index which depends on the decay rate. Thus, one cannot judge about the quality of fitting by a given model based on its limit extreme value index. For example, the TED approximates the exponential distribution (ED) with increasing accuracy as $m_{max} \to \infty$, whereas the extreme value indices of these distributions are -1 and 0, respectively.

As a suitable goodness-of-fit metric, we suggest the classical Chi-square sum and the Kolmogorov distance. These estimates are stable and robust in contrast with the extreme value index. The example with TED and ED clearly shows that two distributions can be very close (in any reasonable statistical

metric), while their extreme value indices are significantly different.

Remark Qualitatively speaking, the extreme value index for a bounded distribution is determined by the index of the power function that approximates the density function in the vicinity of the upper bound, say, (m_{max}–ε; m_{max}), where ε is some small quantity. Since the graph of the TED density in this region is close to a constant, its extreme value index γ equals -1. So in order to efficiently use the limit extreme value distribution as an approximation for the distribution of maxima, one needs to have a sample with sufficient number of data points in the interval (m_{max}–ε; m_{max}) which is equivalent to having an extremely large sample size n.

However, the Peak Over Threshold (POT) distribution of TED can be well approximated by a GPD with a small negative γ even for a moderate sample size n. The TED is very close to ED when (m_{max}–m_{min}) is large. But the GPD with small γ is close to ED ($\gamma = 0$) too. Thus, GPD and TED can be very close for small γ and large (m_{max}–m_{min}). We did observe such closeness in our simulations where we have generated 10,000 TED samples and obtained $\langle\gamma\rangle = -0.1001$, ($m_{max}$–$m_{min}$) = 2.2, β = 2.12 (see point 4 below). Thus, the GPD approximation works well in the case of TED too!

3. Raschke writes in his Introduction:

   As mentioned previously, Pisarenko et al. suggest applying the GEVD and the GPD for the estimation of the upper bound magnitude, and explain the link between GEVD and GPD. I present this link in a more straightforward and transparent manner in the following section. I also explain why these models and methods of extreme value statistics do not work well in the case of the TED and other truncated exponential distributions.

Again, we did not suggest in our paper to apply the GEVD and the GPD for the estimation of the upper bound of the magnitude. We proposed to apply the GPD to real data samples for estimating two parameters, namely γ and σ. In the case when $\gamma < 0$, GPD has the upper bound ($m_{min} - \sigma/\gamma$) that we denoted as M_{max}. This parameter is estimated with a large uncertainty, especially when γ is small. It plays only a secondary role in our method. We discussed in detail the instability and other deficiencies of this parameter (see Pisarenko *et al.* 2014, Introduction, lines 45–71; Section 2.6 A Remark on the instability of M_{max} versus quantiles $Q_\tau(q)$). We introduced a random value M_τ—the maximum magnitude in future τ years. It cardinally differs from M_{max} in spite of their resemblance. The former is a random value, whereas the latter is a non-random parameter. Perhaps, this resemblance between M_τ and M_{max} led Raschke to a misunderstanding of our concept.

In our work, we use M_τ since it makes it possible to obtain statistical characteristics of seismicity (namely, quantiles of M_τ for arbitrary τ) starting from a catalog of any fixed duration T. That feature makes our approach particularly useful and applicable to practical situations.

4. We thus disagree with Raschke's assertion that:

   "... models and methods of extreme value statistics do not work well in the case of the TED and other truncated exponential distributions".

We have generated 10,000 independent TED samples of size $n = 261$ with parameters that were obtained for the real Harvard sample used above, namely

$\beta = 2.117$; $m_{min} = 6.8$; $m_{max} = 9.00$. We applied our GPD procedure to these samples and obtained the following results:

$\langle KD_{GPD}\rangle = 0.687$; p-v(KD_GPD) = 0.617; $\langle S_{GPD}\rangle = 9.01$; p-v(S_GPD) = 0.42; $\langle\gamma\rangle = -0.1001$.

Here, the brackets $\langle\rangle$ mean averaging over 10,000 realizations in our simulations. We see that the results of the GPD calibration are quite satisfactory. The Maximum Likelihood procedure based on TED gave

$\langle KD_{TED}\rangle = 0.708$; p-v(KD_TED) = 0.35.

$\langle S_{TED}\rangle = 9.17$; p-v(S_TED) = 0.41.

We see that GPD and TED methods produced practically the same quality of fit in spite of the fact that the samples had been generated by TED!

5. In Section 3, Raschke writes:

   For example, for the GEVD of the Harvard catalog, we have an MSE(γ) = 0.047 for $T = 80$ days, which means a standard deviation Std(γ) = 0.21. The standard deviation in the authors' corresponding Fig. 2, however, has the value Std(γ) = 0.02–0.025 for $T = 75$ days, which is only a small fraction of 0.21. I strongly

reject their interpretation that "this is not surprising since the latter gives only the scatter conditional to the same unique data sample", as the bootstrap method is specifically used for quantifying the error distribution and its corresponding standard error (Das Gupta 2008). Such a sizable difference may be an indicator that approximating the distribution of extreme magnitudes by GEVD and GPD does not work well.

There is a confusion here that we need to clarify. By an MSE of a statistical estimate $\hat{\alpha}$ of parameter α, we meant the value $\sqrt{E(\hat{\alpha}-\alpha)^2}$ which has the same dimensionality as α and as the bias $(E\hat{\alpha}-\alpha)$. That makes it very convenient to compare the MSE relative to α (E denotes here the mathematical expectation of its argument). Such a definition is rather popular in the statistical literature [see e.g., The probability and the Mathematical Statistics, Encyclopedia, M:, "Great Russian Encyclopedia", 1999, p. 222 (in Russian)]. Sometimes, however, the Mean Square Error is defined as $E(\hat{\alpha}-\alpha)^2$ (without square root). Apparently, Raschke used the latter definition of MSE which caused the confusion in his comparison of the estimates with ours. Probably, we should have given the explicit expression of the MSE that we used in our work.

As to Raschke's strong rejection of our explanation of the fact that the bootstrap estimate of the variance can be smaller than the simulation estimate of variance, we think that it is not justified. Any bootstrap estimate is a conditional estimate under a fixed sample $(x_1 \ldots x_n)$. Such a conditional estimate can in some cases be smaller and in some other cases be greater than the unconditional (ensemble) estimate of the parameter. So, for some samples, the bootstrap estimate of the variance can be smaller than the ensemble estimate that we obtained through simulation with a very large number of realizations (10,000).

Let us consider a simple example in order to illustrate the above-described situation. A sample consists of three independent Bernoulli random values X_1, X_2, X_3 which take the values ± 1 with the probability □. Suppose that we ignore the underlying distribution law of the observed sample. Let us order this sample and denote the ordered sample as $Y_1 \le Y_2 \le Y_3$. We wish to estimate the variance VAR of the sample median Y_2 by the bootstrap method (see EFRON 1973). We have

$$\Pr\{Y_2 = +1\} = 1/2; \quad \Pr\{Y_2 = -1\} = 1/2.$$

$$\mathrm{VAR} = 1.$$

We denote our particular sample as (x_1, x_2, x_3) and the corresponding ordered sample as $(y_1 \le y_2 \le y_3)$. The bootstrap procedure consists in forming random triples (z_{1k}, z_{2k}, z_{3k}):

$$\Pr\{z_{1k} = x_1\} = 1/3; \Pr\{z_{1k} = x_2\} = 1/3;$$
$$\Pr\{z_{1k} = x_3\} = 1/3; \quad k = 1, \ldots N;$$

N is the very large number of bootstrap samples. This is the so-called sampling with replacement. Then we order each triplet (z_{1k}, z_{2k}, z_{3k}) and get for each the ordered triplet $(w_{1k} \le w_{2k} \le w_{3k})$.

$$\Pr\{w_{2k} = +1\} = (\text{numberof}(+1)\text{among}\,(x_1, x_2, x_3))/3;$$

$$\Pr\{w_{2k} = -1\} = (\text{numberof}(-1)\text{among}(x_1, x_2, x_3))/3.$$

Finally, we take the bootstrap sample $(w_{21}, w_{22}, \ldots, w_{2N})$ and calculate its sample variance $\mathrm{VAR}_{\mathrm{boot}}$

$$\mathrm{VAR}_{\mathrm{boot}} = \frac{1}{N}\sum_{k=1}^{N}(w_{2k} - \bar{w}_2)^2; \quad \bar{w}_2 = \frac{1}{N}\sum_{k=1}^{N} w_{2k}.$$

These estimates depend on the realization (x_1, x_2, x_3).

$$\bar{w}_2 \to \Pr\{w_{2k} = +1\} - \Pr\{w_{2k} = -1\}.$$

$$\mathrm{VAR}_{\mathrm{boot}} \to 4 \cdot [\Pr\{w_{2k} = +1\}] \cdot [\Pr\{w_{2k} = -1\}];$$
$$\text{as } N \to \infty.$$

If, say, $x_1 = x_2 = 1$; $x_3 = -1$, then the conditional probabilities are $\Pr\{w_{2k} = +1\} = 2/3$; $\Pr\{w_{2k} = -1\} = 1/3$;

$$\mathrm{VAR}_{\mathrm{boot}} \to 4 \cdot [\Pr\{w_{2k} = +1\}] \cdot [\Pr\{w_{2k} = -1\}] = 8/9.$$

Thus, $\mathrm{VAR} > \lim(\mathrm{VAR}_{\mathrm{boot}})$ for this sample. Of course, this example is oversimplified (sample size $n = 3$), but it just demonstrates the possible existence of a situation when the bootstrap estimate is smaller than the ensemble estimate. Only when the sample size n tends to infinity does the bootstrap estimate converge to the ensemble estimate.

Now, some explanations are needed regarding the standard deviations (std) in figure captions in

(PISARENKO *et al.* 2014). Unfortunately, there are misprints in the notations used in these captions, although the notations in the text are correct, see (PISARENKO *et al.* 2014), equations (22), (28), (35), (38). Since random deviations have often large skewness, it is convenient and appropriate to indicate such deviations not by the usual ± std, but by the upper 84 % and the lower 16 % confidence limits. In the case of the Gauss distribution, these limits coincide with ± std. We have widely used this probabilistically equivalent change in (PISARENKO and RODKIN 2010). This explains the difference of the upper and lower deviations in Fig. 2–14. In these figures, the probabilistic equivalents of ± std are shown in the form of the upper 84 % and lower 16 %.

We apologize for these misprints and are grateful to Raschke for noting them.

In our paper, the bootstrap scatter of parameter γ, mentioned by Raschke, had the following values: upper 84 %-bound = −0.152; lower 16 %-bound = −0.221. If we equate their difference to 2 std, we get std = 0.0345, which is somewhat less than the simulation's MSE = 0.047 (500 random samples) that we took as an ensemble estimate. The bias could be neglected, as we previously explained. The obtained values are comparable, and the fact that the latter is a bit larger is quite possible, as discussed above.

6. In Section 6 Raschke writes:
   Finally, I want to point out that earthquake data do not need to occur as a homogeneous Poisson process in order to apply extreme value theory and statistics.

That is true, but Raschke overlooks one important point in our procedure, namely the possibility to obtain statistical characteristics of the maximum distribution for an arbitrary future time interval, starting with a catalog that covers some fixed time interval. This property is based on the fundamental Theorem 3.4.13, point (d) (see EMBRECHTS *et al.* 1997, p. 165). This Theorem establishes a direct connection between the GPD and the GEV for observations in the form of time series (such as an earthquake catalog), and requires the Poissonian property of the seismic process in question. Some generalizations of this Theorem to non-stationary (non-homogeneous) Poisson processes were obtained in (PISARENKO and RODKIN 2014). It is questionable whether a statistical inference can be done for future times unless we know in detail the properties of a non-Poissonian seismic process. That is why, before using the seismic catalog, we cleaned it from aftershocks and swarms, making it close to a Poisson flow of events, and then chose a sufficiently high magnitude threshold to ensure the stationarity of the seismic flow.

We hope that we have provided well-founded answers to all essential claims and remarks by Raschke. Some of them were based on assertions mistakenly attributed to us. Others were due to some misunderstandings that we tried to clarify. We agree with Raschke that applications of the extreme value methods should be done with care and detailed verification of the statistical aspects of the studied data. That being said, we are convinced to have offered in (PISARENKO *et al.* 2014) a good perspective of these methods for the statistical analysis of earthquake catalogs and in other fields.

REFERENCES

B. EFRON (1973), Bootstrap methods: Another look at the Jackknife, Ann. Statist., vol. *8*, #1, 1–26.

P. EMBRECHTS, C. KLUPPELBERG, T. MIKOSCH (1997) Modelling Extremal Events, Springer, Berlin, 1997.

V.F. PISARENKO, M.V. RODKIN (2010) Heavy-Tailed Distributions in Disaster Analysis, Springer, Heidelberg.

V.F. PISARENKO, M.V. RODKIN (2014) Statistical Analysis of Natural Disasters and Related Losses, Springer, Heidelberg.

PISARENKO V. F., SORNETTE A., SORNETTE D., and M. V. RODKIN (2014), Characterization of the Tail of the Distribution of Earthquake Magnitudes by Combining the GEV and GPD Descriptions of Extreme Value Theory, Pure Appl. Geophys. *171*, 1599–1624.

RASCHKE, M. (2016) Comment on Pisarenko et al., "Characterization of the Tail of the Distribution of Earthquake Magnitudes by Combining the GEV and GPD Descriptions of Extreme Value Theory", Pure Appl. Geophys.

(Received September 25, 2015, accepted October 12, 2015)