Holschneider, M., Zöller, G., Clements, R., Schorlemmer, D. (2014): Can we test for the maximum possible earthquake magnitude? - *Journal of Geophysical Research*, *119*, 3, p. 2019-2028

DOI: http://doi.org/10.1002/2013JB010319

AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Solid Earth

RESEARCH ARTICLE

10.1002/2013JB010319

Key Points:
- Mmax is poorly constrained by data
- Mmax is not testable on the basis of earthquake catalogs
- Long-term geological data can reduce uncertainties of Mmax

Correspondence to:
M. Holschneider,
hols@math.uni-potsdam.de

Citation:
Holschneider, M., G. Zöller, R. Clements, and D. Schorlemmer (2014), Can we test for the maximum possible earthquake magnitude?, *J. Geophys. Res. Solid Earth*, *119*, 2019–2028, doi:10.1002/2013JB010319.

Received 24 APR 2013
Accepted 14 JAN 2014
Accepted article online 21 JAN 2014
Published online 21 MAR 2014

Can we test for the maximum possible earthquake magnitude?

M. Holschneider[1], G. Zöller[1], R. Clements[2], and D. Schorlemmer[2]

[1]Institute of Mathematics, University of Potsdam, Potsdam, Germany, [2]Helmholtz Center Potsdam–GFZ German Research Center for Geosciences, Potsdam, Germany

Abstract We explore the concept of maximum possible earthquake magnitude, *M*, in a region represented by an earthquake catalog from the viewpoint of statistical testing. For this aim, we assume that earthquake magnitudes are independent events that follow a doubly truncated Gutenberg-Richter distribution and focus on the upper truncation *M*. In earlier work, it has been shown that the value of *M* cannot be well constrained from earthquake catalogs alone. However, for two hypothesized values *M* and *M*′, alternative statistical tests may address the question: Which value is more consistent with the data? In other words, is it possible to reject a magnitude within reasonable errors, i.e., the error of the first and the error of the second kind? The results for realistic settings indicate that either the error of the first kind or the error of the second kind is intolerably large. We conclude that it is essentially impossible to infer *M* in terms of alternative testing with sufficient confidence from an earthquake catalog alone, even in regions like Japan with excellent data availability. These findings are also valid for frequency-magnitude distributions with different tail behavior, e.g., exponential tapering. Finally, we emphasize that different data may only be useful to provide additional constraints for *M*, if they do not correlate with the earthquake catalog, i.e., if they have not been recorded in the same observational period. In particular, long-term geological assessments might be suitable to reduce the errors, while GPS measurements provide overall the same information as the catalogs.

1. Introduction

The title as it stands is provocative since tests always exist. The central question, however, concerns the quality of their performance. This is of uttermost importance in view of applications related to large-earthquake hazard, since decisions can be based only on sufficiently powerful tests with a tolerable error. The maximum possible magnitude *M* enters into probabilistic seismic hazard assessment, although it might be less influential than other components, e.g., the choice of the ground motion model. Nevertheless, the knowledge of *M* is crucial for various purposes related to worst case scenarios. In two previous publications, we have considered the problem of estimating *M* from earthquake catalogs in the context of a doubly truncated Gutenberg Richter law [*Holschneider et al.*, 2011; *Zöller et al.*, 2013]. It was shown that with high probability no finite confidence intervals exist. In a Bayesian setting it was shown that the posterior distribution is not normalizable unless strong prior assumptions are made.

In the present manuscript, we change the focus from the estimation of *M* in terms of a point estimator or a posterior distribution toward statistical testing to discriminate between alternatives. In particular, we hypothesize two values *M* and *M*′ and address the question, which value is more consistent with a given earthquake catalog. The precise meaning of this statement will be explained below. We do not claim that one of the values is "correct." The tests are characterized by two errors, the error of the first kind and the error of the second kind. If both errors are small in the context of a particular application, *M* and *M*′ can be distinguished by the data; otherwise, the earthquake catalog is too short. Therefore, we will also derive estimates for the required length of an earthquake catalog in order to perform tests with reasonable small errors.

This paper is structured as follows: In the next section, we describe the statistical model and the alternative statistical test in detail. Then we argue that this test is optimal; i.e., alternative tests leading to smaller errors do not exist. After discussing the trade-off between testing size and testing power (or between the error of the first and of the second kind) for seismicity in Japan, we present a case study for Switzerland. Finally, we explicitly compute confidence intervals and summarize our findings.

2. Alternative Statistical Testing of M Against M'

We assume independent random events for which magnitudes are drawn from the density of a doubly truncated Gutenberg-Richter (GR) distribution [*Gutenberg and Richter*, 1956]:

$$p_M(m) = \chi_{[m_0,M]}(m)\frac{\beta e^{-\beta m}}{e^{-\beta m_0} - e^{-\beta M}}, \tag{1}$$

where $\chi_{[m_0,M]}(m)$ is the indicator function that is 1 if $m \in [m_0, M]$ and 0 otherwise. The unknown parameter M is the maximum possible magnitude and is to be inferred from a catalog which consists of N events with magnitudes m_i, for $i = 1, \ldots N$. The magnitude of completeness, m_0, is supposed to be known. Here we set $\beta = b\log(10)$, where b is the Gutenberg-Richter b value, is also known.

We are concerned with the following pairs of hypotheses for which we will derive optimal testing strategies. The simplest one would be to test

$$H_1\text{: the maximum possible magnitude is } M \tag{2}$$

against the alternative

$$K_1\text{: the maximum possible magnitude is } M' \tag{3}$$

for fixed $M < M'$. We will build a test which allows $M' = \infty$. Surprisingly, this physically absurd hypothesis is not rejected by the catalogs we have at hand. Since the tests we develop are uniformly most powerful, the same procedures can be applied for the following compound hypothesis:

$$H_2\text{: the maximum possible magnitude} \leq M \tag{4}$$

against the alternative

$$K_2\text{: the maximum possible magnitude is} > M. \tag{5}$$

For convenience, we give here a very short testing primer, which allows us to introduce the necessary notation; more information can be found in *Lehmann and Romano* [2005]. A test is a mapping that takes the observed N magnitudes

$$\underline{\mathrm{m}} = [m_1, m_2 \ldots m_N] \tag{6}$$

and assigns to them a number $\phi = \phi(\underline{\mathrm{m}})$ in $[0, 1]$. This number is the probability with which we should reject H. Concretely, we need to draw a random number u uniformly in $[0, 1]$ and if, for an observed catalog, this number is $u > \phi$, we reject the hypothesis, otherwise we do not reject it. In the case that this mapping takes only the values 0 or 1, we have a nonrandomized test, since for every outcome $\underline{\mathrm{m}}$ we either reject H for $\phi(\underline{\mathrm{m}}) = 1$ or do not reject H for $\phi(\underline{\mathrm{m}}) = 0$. Psychologically, nonrandomized tests are preferable, although their performance in the long run is not better than randomized tests. A test has size α if the probability with which we reject the hypothesis H, even if it is true, is bounded by α:

$$\mathbb{E}(\phi(\underline{\mathrm{m}})|H) = \text{probability of erroneously rejecting } H \leq \alpha. \tag{7}$$

Thus, α is an upper bound of the error of the first kind, which means to erroneously reject the hypothesis H. The error of the second kind consists of failing to reject H even though it is false. We will rather use the equivalent notion which is the power of the test:

$$\kappa = \mathbb{E}(\phi(\underline{\mathrm{m}})|K) = \text{probability of correctly not rejecting } K, \tag{8}$$

so that $1 - \kappa$ measures the probability of failing to reject H even though K is true. The above definition holds for a simple alternative K. In case of a compound alternative, we obtain a power for each model in K. In the context of maximum possible magnitude both types of errors have obviously very different consequences. This can be illustrated in the framework of earthquake safety requirements for critical facilities. A large error of the first kind would make us spend unnecessary money on safety, whereas a large error of the second

kind bears the risk of underestimating devastating catastrophes. In a further step the mere probability considerations should be equipped with loss functions, which allow us to trade off between the two kinds of consequences in a systematic way. In this study, however, we will stick to the probability considerations only.

A test ϕ is said to be optimal if, for a size $\alpha \in [0, 1]$, it is as powerful as possible:

$$\mathbb{E}(\phi(\underline{\mathrm{m}})|K) \rightarrow \text{maximum} \quad \text{under constraint} \quad \mathbb{E}(\phi(\underline{\mathrm{m}})|H) \leq \alpha. \tag{9}$$

For a nonsimple hypothesis, the difficulty may arise that there is not a single strategy ϕ, which is optimal for all parameter values that constitute the hypothesis K. It may happen, however, that there is a test ϕ which is optimal for all possible members of the hypothesis K. In that case we speak of a uniformly most powerful test. The tests used in the remainder of this manuscript will be uniformly most powerful.

3. The Uniformly Most Powerful Tests

The key to testing is to consider the simple hypothesis H_1 against K_1. That is, we want to test, if the maximum magnitude is M, against the alternative that it is M', for fixed $m_0 < M < M' \leq \infty$. In this case, the likelihood ratio defines an ordering of the observations $\underline{\mathrm{m}}$. The likelihood of observing $\underline{\mathrm{m}}$ under the assumption of independent events is

$$p_M(\underline{\mathrm{m}}) = \prod_{i=1}^{N} p_M(m_i) = \chi_{[m_0,M]}(\mu) \frac{\beta^N e^{-\beta N \bar{m}}}{(e^{-\beta m_0} - e^{-\beta M})^N}, \tag{10}$$

with

$$\mu = \max\{m_i\}, \quad \bar{m} = \text{mean}\{m_i\}. \tag{11}$$

Here this ratio is given by

$$r(\underline{\mathrm{m}}) = \frac{p_{M'}(\underline{\mathrm{m}})}{p_M(\underline{\mathrm{m}})} = \frac{(e^{-\beta m_0} - e^{-\beta M})^N}{(e^{-\beta m_0} - e^{-\beta M'})^N} \frac{\chi_{[m_0,M']}(\mu)}{\chi_{[m_0,M]}(\mu)}. \tag{12}$$

It therefore takes on only two values, depending on whether $\mu \leq M$ or $M < \mu \leq M'$. According to the general theory of testing, an optimal test is then any test for which we always reject H_1 if $\mu > M$ and for which we fail to reject H_1 for $\mu < M$ for a fraction α of times. A randomized version of this test could therefore be designed by simply throwing a dice to decide if in the event of $\mu < M$ we reject, or we do not reject the hypothesis M. A nonrandomized version of the test can be achieved by fixing a threshold such that under H_1 this threshold is not reached with probability α. In formulas, we choose m_c defined through

$$\mathbb{P}(\mu < m_c|H_1) = 1 - \alpha. \tag{13}$$

An optimal nonrandomized testing procedure would then read

$$\text{optimal test: if } \mu \leq m_c \text{ do not reject } H_1, \text{ otherwise reject it.}$$

Since the cumulative density of a single event reads, for $m \geq m_0$,

$$F_M(m) = \min\left\{\frac{e^{-\beta m_0} - e^{-\beta m}}{e^{-\beta m_0} - e^{-\beta M}}, 1\right\}, \tag{14}$$

we need to solve

$$[F_M(m_c)]^N = 1 - \alpha. \tag{15}$$

This has the explicit solution

$$m_c = m_0 - \frac{1}{\beta} \log\left\{1 - (1 - \alpha)^{1/N}\left[1 - e^{-\beta(M - m_0)}\right]\right\}. \tag{16}$$

The power of the optimal tests (randomized or not) is then, for $0 < \alpha < 1$,

$$\kappa = 1 - [F_{M'}(m_c)]^N = 1 - [F_{M'}(M)]^N(1 - \alpha). \tag{17}$$

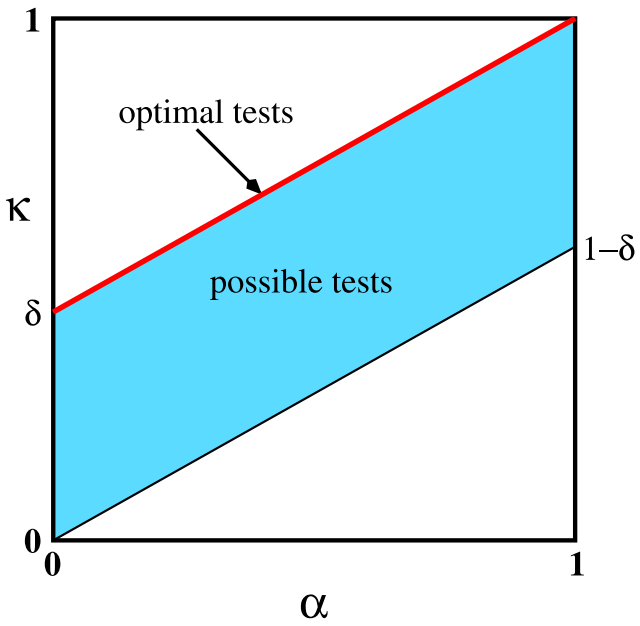

Figure 1. The performance rates of all possible test for testing M against M'. The optimal tests, i.e., the tests with highest power for a given size α are on the upper edge. The space of possibilities is parameterized by a parameter δ, which is the probability to observe at least one event with magnitude $> M$ among the N magnitudes in a catalog if M' is the true upper bound. Unless N is very large, this number δ is small and high power can only be achieved through high α.

In equation (17) we have used the fact that for $m_0 \le m \le M \le M'$,

$$\frac{F_{M'}(m)}{F_M(m)} = \text{constant} = F_{M'}(M). \quad (18)$$

In equation (17) the right-hand side is actually the power of the test, which can be understood as follows: The only possibility not to reject M under hypothesis M′ requires first to have all events below M, which has probability $[F'_M(M)]^N$, and second to draw a random number above α.

In the case of a compound hypothesis as H_2 against K_2 (maximum magnitude $\le M$ against $> M$), the same tests are optimal for all parameters in K_2. The optimal testing procedure would therefore consist in rejecting as soon as the maximum observed event is above m_c as defined above. The power, however, becomes now a function of the true magnitude.

4. Testing M' Against M: Trade-Off Between Testing Size and Power

In this section, we will explicitly discuss the trade-off between the testing size and power or between the error of the first kind and the error of the second kind when testing a maximum magnitude M against M'. In practical situations, it is required that both errors must not exceed a predefined size depending on the desired degree of safety. This requirement can, however, only be met for a large enough earthquake catalog in terms of event number N. For this aim, relations between α, κ, and N will be derived and applied to a hypothetical example.

To get an insight into the orders of magnitude involved, let us consider the α-κ diagram. The set of possible (α, κ) combinations is always a convex set which is point symmetric at the point $(1/2, 1/2)$. This follows by realizing that random guessing would produce the diagonal, any convex combination of two tests can be realized by suitable random choice between both decisions, and exchanging the rejection and the failure of rejection of the null hypothesis would produce the mirrored test [e.g., *Lehmann and Romano*, 2005]. The situation of our problem is depicted on Figure 1. The blue area consists of all possible realizable combinations of α and κ values. On the upper boundary (marked in red) are the optimal tests; i.e., the tests with the highest power for a given α, on the lower boundary are the worst performing tests. The area of possible tests is delimited by a parallelogram of height δ, where

$$\delta = 1 - [F_{M'}(M)]^N = \mathbb{P}(\mu > M | M'). \quad (19)$$

Therefore, we have

$$\delta \le 1 - [1 - e^{-\beta(M-m_0)}]^N. \quad (20)$$

As can be seen graphically on Figure 1, in order to achieve a high enough power, the value of α must be driven close to 1 unless N is extremely high. This is so, since only for large N we can get δ close to 1 in which case high power could be achieved with moderate values of α. Concrete numbers will be provided below.

On the other hand, the power of the test is measuring the probability that we do not erroneously fail to reject H although K is true. It is this error that should be very small, say $\epsilon = 10^{-4}$ requiring $\kappa \ge 1 - \epsilon$. The values α at which we need to run the test are therefore necessarily:

$$\alpha \ge 1 - \frac{\epsilon}{[F_{M'}(M)]^N} > 1 - \frac{\epsilon}{[1 - e^{-\beta(M-m_0)}]^N}. \quad (21)$$

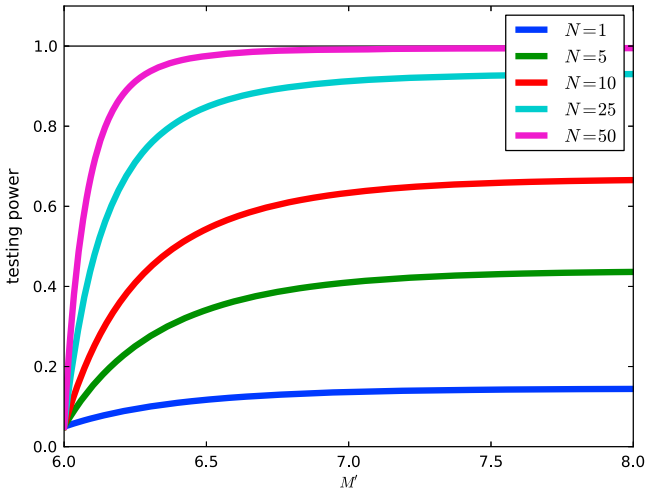

Figure 2. The power for testing against M' for $M = 6$. Here N refers to the number of events in the last magnitude bin [5, 6]. The size of the test has been fixed at $\alpha = 0.05$. Note that even for 25 events in the last bin essentially no power can be achieved not even when testing against $M' = \infty$.

The last inequality corresponds to the end-member case of testing against an unlimited GR law: $M' = \infty$. In order to achieve this high size of the test, the region where H is not rejected, as controlled by m_c, will be small since we have

$$m_c \leq m_0 - \frac{1}{\beta} \log \left[1 - \frac{\epsilon^{1/N}}{F_{M'}(M)} e^{-\beta(M-m_0)} \right]. \tag{22}$$

Again, for the most testable case where we only want to test against $M' = \infty$, we obtain the inequality

$$m_c \leq m_0 - \frac{1}{\beta} \log \left[1 - \frac{\epsilon^{1/N}}{e^{\beta(M-m_0)} - 1} \right]. \tag{23}$$

If we fix α (or m_c) and ϵ, we still have N, the size of the earthquake catalog. Solving for N yields, for the required number of observed events,

$$N \geq \frac{\log[\epsilon/(1-\alpha)]}{\log F_{M'}(M)} > \frac{\log[\epsilon/(1-\alpha)]}{\log[1 - e^{-\beta(M-m_0)}]}. \tag{24}$$

To further simplify the setting, we focus on the last magnitude bin, that is, we choose $m_0 = M - 1$. Then N is the number of events in the last magnitude bin before the limiting magnitude for which we want to test. Corresponding values for lower m_0 can be calculated by the GR law in straightforward manner. Under the optimistic assumption of $N = 10$, we have shown on Figure 2 the power of testing M against $M' > M$. The size of the test was fixed at $\alpha = 5\%$. As can be seen, no testing power is achieved, and the probability of erroneously rejecting the larger magnitude remains above 0.999. The only way to remedy this problem with as little as 10 events in the last magnitude bin is to accept α values which are close to 1. This means, however, we reject with very high probability the hypothesis M and take the alternative M'. In concrete numbers, in order to fail to correctly choose the larger magnitude in a fraction of cases of at most $\epsilon = 10^{-4}$, we must reject the lower magnitude with probability $\alpha > 0.999$.

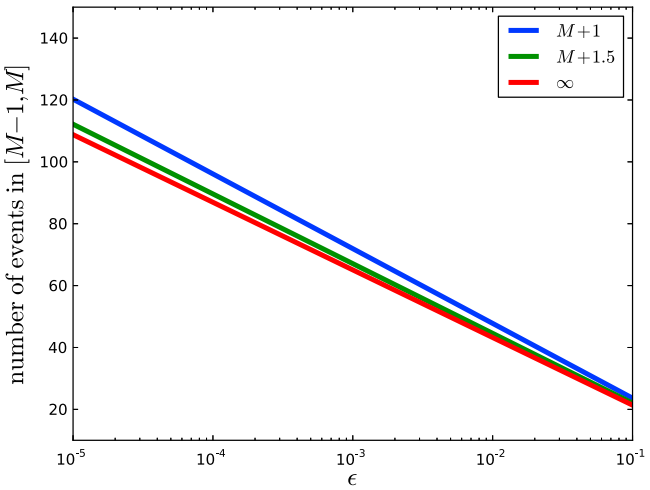

Figure 3. The number of events in the last magnitude bin as a function of the error of the second kind. The error of the first kind was fixed at $\alpha = 0.05$. For realistic acceptable values of ϵ we need a catalog that would have in the average about 90 events in the last magnitude bin. If we want to test for $M = 7$ and we have the magnitude of completeness $m_0 = 4$, the size of the catalog should be about $(1 + 10 + 100) \times 90 = 9990$ events with magnitude ≥ 4 (assuming $\epsilon = 10^{-4}$ and $b = 1$). If we want to test for $m = 8$, the size of the catalog has to be 99990 events of magnitude ≥ 4.

We now consider the dependency on N. Figure 3 shows the required number of events in the last magnitude bin for testing M against M' for $\alpha = 0.05$ and ϵ in the range from realistic 10^{-5} to careless 10^{-1}. The number in the last bin can be scaled to the size of the needed catalog in the case that a lower completeness cutoff is used in the catalog. As a result, the number of events, even for testing against $M' = \infty$ and for the highly insufficient value of $\epsilon = 0.01$, is at least 20. The best catalogs nowadays do not have as many large magnitude events close to the critical value. As one of the best

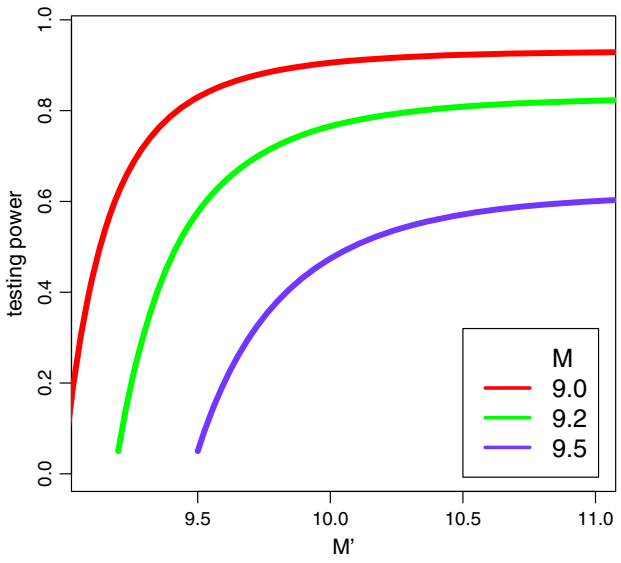

Figure 4. Case study Japan: Testing power of M against M' for three values of M. For the Gutenberg-Richter b value we have used $b = 0.92$ [*Toda and Enescu*, 2011].

study regions in terms of earthquake catalog quality, we refer to Japan. Following the NOAA catalog, 180 earthquakes with magnitudes $7 \leq m \leq 9$ (35 earthquakes with magnitudes $8 \leq m \leq 9$) occurred between the years 684 and 2012. Fixing the size of the test at 5% means that we reject erroneously the lower magnitude with a probability of 0.05. The largest observed event is the $m = 9.0$ Tohoku earthquake that occurred on 11 March 2011 [*Peng et al.*, 2012]. Results for the testing power are shown on Figure 4. The end-member, where M is tested against $M' = \infty$, results in the testing power $\kappa = 0.97$ for $M = 9$, $\kappa = 0.91$ for $M = 9.2$, and $\kappa = 0.74$ for $M = 9.5$. Even in the "best" situation ($M = 9$ and $M' = \infty$), the lower magnitude is erroneously not rejected in about 7% of cases, which is hardly tolerable. This value has been calculated with the Gutenberg-Richter b value of 0.92 [*Toda and Enescu*, 2011]. For $b = 0.92 \pm 0.05$ the lower magnitude is not rejected in 12% ($b = 0.97$) and 3% ($b = 0.87$) of cases, respectively. All of these numbers are far from being tolerable in practice. We emphasize that the data availability in Japan is probably the best in the world, at least in this magnitude range. In order to achieve an acceptable safety, i.e., for the power $\kappa = 1 - 10^{-4}$, we would need to adapt the size to values of $\alpha = 0.9968$. In other words, to achieve the required safety, we essentially always need to account for the larger magnitude. These numbers illustrate that even for excellent data and weak hypotheses ($M' = \infty$), there is no tractable strategy that can balance the error of the first and the second kind toward acceptable values.

5. Case Study: Switzerland

In the previous section, we focused on Japan as a region with excellent data coverage even for very large earthquakes. Now we study an area with moderate to low seismicity, Northwestern Switzerland. However, due to enormous loss potential including four nuclear power plants, this is clearly a high-risk area. In the framework of the project PEGASOS ("Probabilistische Erdbeben-Gefährdungs-Analyse für KKW-Standorte in der Schweiz"), a comprehensive seismic hazard study has been carried out for Northwestern Switzerland. *Burkhard and Grünthal* [2009] consider a large-scale seismic source zone (SSZ) model, which includes parts of adjacent countries, and estimate maximum magnitudes using the Electric Power Research Institute (EPRI) approach [*Johnston et al.*, 1997]. Although this is only one out of four models used in the PEGASOS project, it is suitable for testing purposes, because the maximum magnitude is provided by a sequence of values with individual probability weights for each SSZ. This allows us to test one value against another one. As an end-member, we also include $M = \infty$ in our analysis. The EPRI approach is a standard Bayesian approach as in *Holschneider et al.* [2011] with the difference that informative prior distributions containing various types of empirical knowledge are used. Based on crustal structure and tectonic arguments, the SSZs are associated with either a prior for the "stable continental" crust or a prior for the "extended continental crust." The result is a posterior probability distribution for the maximum magnitude M, which is truncated for the largest earthquakes in order to avoid unrealistically large events. The truncation point is subject to geological and "common sense" arguments. The posterior distribution is then discretized with respect to M: $M_l = M_1, M_2, \ldots M_n = M_u$, and the probability content of a bin $[M_{i+1}, M_i]$ is considered to be a "weighting factor" for the estimate M_i. The most important earthquake recurrence parameters which enter in the analysis are listed in Table 1. We note that the parameters of the Gutenberg-Richter recurrence relationship

$$\log [N(m)] = a - bm \tag{25}$$

Table 1. Results of Hypothesis Testing for Northwestern Switzerland[a]

Label	Name	m_0	$N(m \geq m_0)$	b	M_l	M_u	$T(M_l; M_u)$	$T(M_l; \infty)$
EF	Eastern France	2.3	7.0190	1.0470	6.0	7.2	3322	3138
RG	Rhine Graben	2.3	2.8950	0.8580	6.0	7.7	1575	1520
SG	South Germany	2.3	5.1890	0.7750	6.0	6.7	586	418
BG	Bresse Graben	2.3	0.8781	0.6730	5.5	7.5	498	476
AE	Alps External	2.3	4.4160	0.7720	7.0	7.8	3734	2833
AC	Alps Central	2.3	15.720	0.7720	6.5	7.0	556	327
AI	Alps Internal	3.3	1.3520	0.9170	6.0	7.9	662	650
PP	Po Plain	3.3	0.4511	1.0750	5.5	7.6	1518	1509

[a]The first two columns refer to the seismic source zone; m_0 is the magnitude of completeness, $N(m \geq m_0)$ is the estimated annual number of earthquakes with $m \geq m_0$, and b is the estimated Richter b value. Fur further details, e.g., uncertainties, see *Burkhard and Grünthal* [2009]. The magnitudes M_l and M_u refer to the lowest and the highest magnitude values in the posterior distribution as used by *Burkhard and Grünthal* [2009]. $T(M_l; M_u)$ and $T(M_l; \infty)$ are the respective number of years to achieve the testing power $\kappa = 0.95$ given $\alpha = 0.05$ for the alternative tests M_l against M_u and M_l against ∞.

are provided in *Burkhard and Grünthal* [2009]. Here $N(m)$ is the annual number of earthquakes with magnitude $\geq m$, b is the Gutenberg-Richter b value, and the Gutenberg-Richter a value is given by $a = bm_0 + \log[N(m_0)]$. We are thus able to estimate the time, in units of years, that is required to achieve a given testing power, say $\kappa \in [0.9; 1.0)$, when testing a magnitude M' against $M \leq M'$. For these calculations we impose the error of the first kind to be $\alpha = 0.05$. Results for three values of M are shown on Figure 5.

In the last two columns of Table 1, we provide the number of years which is required to achieve the testing power $\kappa = 0.95$, given the error of the first kind $\alpha = 0.05$. For each SSZ we consider two alternative tests based on the smallest (M_l) and the largest magnitude (M_u) values in *Burkhard and Grünthal* [2009]. First, we test $M = M_l$ against $M' = M_u$, and second $M = M_l$ against $M' = \infty$. We emphasize that because of the relatively high difference of M and M' and the low value imposed for the testing power, the given setting includes hypotheses which should be easily distinguishable. However, even in this situation, the number of years varies between hundreds to thousands of years. If we go to higher values of the testing power, which are more appropriate for engineering purposes ($\kappa > 0.99$), we observe a drastic increase of the required time to achieve this power; see Figure 5 for the Eastern France zone.

In summation, we find that rigorous statistical testing requires periods which are unacceptably long, even if "weak" hypotheses are considered. Facing the presence of nuclear power plants in the study region, acceptable errors will require α close to zero and κ close to unity, leading to essentially absurd testing periods.

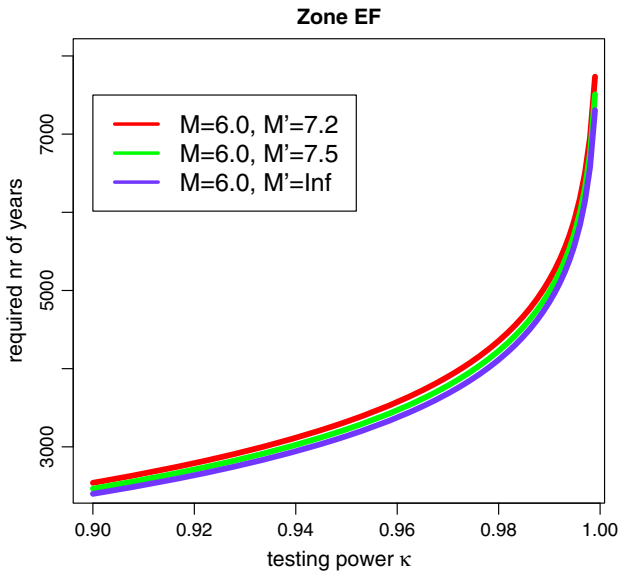

Figure 5. Seismic source zone Eastern France (EF): Number of years that is required to achieve the testing power κ for $\alpha = 0.05$ and three alternative tests labeled in the legend.

6. Confidence Intervals Revisited

Now that we have a uniformly most powerful test for M against $M' > M$ at hand, we can use a standard argument to obtain confidence intervals for a given level γ. For any M we construct, a region, where the null hypothesis H that the maximum magnitude is M, is not rejected. It is given by the following set of catalogs $\underline{m}$ of length N,

$$A_M = \{\mu(\underline{m}) \leq m_c(M, \gamma)\}, \qquad (26)$$

with $m_c(M, \gamma)$ the critical magnitude given by equation (16) with $\gamma = 1 - \alpha$.

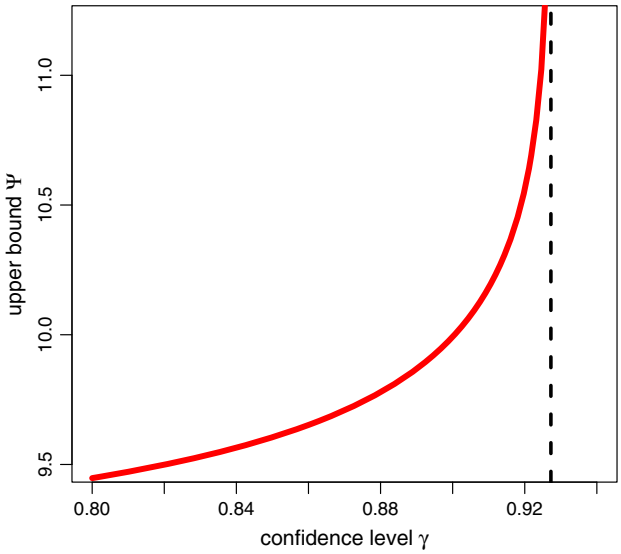

Figure 6. The upper bound of the optimal (i.e., smallest) confidence interval as a function of the confidence level γ for the parameters of Japan ($b = 0.92$, $N = 180$, $m_0 = 7$, $\mu = 9.0$). For confidence levels above $\gamma_c \simeq 0.93$ the upper bound becomes ∞ and no finite confidence interval exists anymore (dashed line).

We now define I_γ for a catalog $\underline{\mathrm{m}}$ to be the set of maximum magnitudes, for which we would not reject hypothesis H at the level γ

$$I_\gamma(\underline{\mathrm{m}}) = \{M | \underline{\mathrm{m}} \in A_M\} = \{M | \mu \leq m_c(M, \gamma)\}. \tag{27}$$

This is then a confidence set at level γ:

$$\mathbb{P}(M \in I_\gamma(\underline{\mathrm{m}}) | M) \geq \gamma. \tag{28}$$

Explicit computation yields the result published in *Holschneider et al.* [2011], namely,

$$I_\gamma(\underline{\mathrm{m}}) = [\mu, \psi], \quad \psi = \begin{cases} m_0 - \frac{1}{\beta} \log\left[\frac{\exp(-\beta(\mu - m_0)) - 1}{(1-\gamma)^{1/N}} + 1\right] & \mu < \mu_c \\ \infty & \mu \geq \mu_c \end{cases}. \tag{29}$$

The critical value μ_c is given by

$$\mu_c = m_0 - \beta^{-1} \log[1 - (1-\gamma)^{1/N}]. \tag{30}$$

If the maximum observed magnitude is above this value, no finite confidence interval for M at the level γ can be given. Since our tests are most powerful, these confidence intervals are optimal in the sense that any other confidence interval of the form $[\mu, \phi]$ will contain our confidence interval [see, e.g., *Lehmann and Romano*, 2005]. Note that in the present paper γ denotes the probability of finding M in the confidence interval, whereas in *Holschneider et al.* [2011] we used $\alpha = 1 - \gamma$. In the case of Japan, we see on Figure 6 that no confidence levels of higher probability than 0.93 exist.

7. Discussion and Conclusion

In this manuscript, we have performed alternative statistical tests for the maximum earthquake magnitude. Our findings are solely based on a statistical model for earthquake magnitudes (doubly truncated Gutenberg-Richter law) and on earthquake catalogs as "hard quantitative" data representing a seismically active region. Other information, e.g., from geology or paleoseismology, are not taken into account in this stadium of development. Future work will focus on the reduction of uncertainties from such long-term data, which are not correlated with the earthquake catalog.

As in previous publications on the maximum possible magnitude of earthquakes, we come to the conclusion that from earthquake catalogs alone, no useful information about the size of the maximum possible

magnitude can be gained, at least from the amount of data that is available in earthquake catalogs. If it is true that the largest events are independently drawn from a doubly truncated Gutenberg-Richter (GR) distribution, no other method of estimating the maximum magnitude, as complicated and sophisticated it might be, will perform better than our optimal tests. This, however, requires unrealistically many large events in order to allow differentiation between ultimate magnitudes that are at least one unit apart. Alternatively, in order to prevent the devastating impact of the error of the second kind, we need to accept very high rejection rates, which leads in practice always to the rejection the lower magnitude. This points to a serious problem in all fields dealing with maximum magnitudes, since this conclusion applies to all procedures, how complex they may be, which as the only "hard observational fact" have earthquake catalogs as input data. There are only four ways to "escape" from this conclusion:

1. A lot of data are available.
2. Earthquakes occur according to some other law [*Wesnousky*, 1994].
3. The maximum possible magnitude is a questionable quantity and may be replaced by the maximum magnitude in a predefined time interval for particular applications.
4. Other information, such as maximum rupture length, is independently available [*Wells and Coppersmith*, 1994].

In the first case, the modeling context of a doubly truncated GR law allows the inference of the maximum possible magnitude, only if unrealistically large catalogs are available. For the second road, this is clearly not covered by the results of this study. However, when it comes to such a serious question as evaluating the seismic risk for high-risk industrial plants, the community should be clear about what part of our models are really known and beyond the stadium of speculative or toy models. The third road suggests that we have to live with the risk of devastating earthquakes, but probably not during our lifetime. This has been elaborated in *Zöller et al.* [2013]. Here the maximum magnitude assessments rely on proper estimates of a and b values for a region.

The last escape road seems to be the most promising one, because recent studies and initiatives like the "Global Earthquake Model" (GEM) and the European project "Seismic Hazard Harmonization in Europe" provide new data and insights that might help to constrain M from a physical point of view [see, e.g., *Basili et al.*, 2013; *Haller and Basili*, 2011; *Holt et al.*, 2005; *Stirling et al.*, 2012, and references therein]. However, regarding rupture length, we must be aware that the actual length of the largest possible rupture is, in general, not available. Even if rupture lengths are recorded, the errors have considerable size. Moreover, information like average slip rates [*Kagan*, 2002] or estimated energy flux into a fault zone is often obtained from observations which are intimately related to the earthquakes that actually occurred, and we face the same problem of rare events again. In relation to such cases, our results can be considered as optimal, because we directly use earthquake information, while other quantities like slip rates are only for proxies for earthquakes and are therefore subject to additional uncertainties of unknown size. For this reason, new information together with uncertainties have to be examined carefully with respect to possible functional dependencies on earthquake catalogs, before using them as additional constraints for M.

At a first glance it seems that the difficulties to estimate M might be due to the sharp singular behavior of the distribution function at M. Therefore, tapered versions of the GR model have been proposed [see, e.g., *Kagan and Schoenberg*, 2001]. This is, however, misleading. The corner magnitude in such a tapered law, which plays the same role as M, bares exactly the same difficulties. The reason is that in the light of the observed magnitudes, the parameter M has only a very small influence on the probability density. Therefore, the likelihood function does essentially not depend on the parameter M. For that reason, the different values of the parameter become indistinguishable from the data alone. In other words, the parameter M that describes the shape near the upper cutoff is sensitive only to the large events, and there are only a few. This difficulty has already been mentioned in *Kagan and Schoenberg* [2001], where the authors question the usefulness of standard asymptotic arguments for the construction of confidence intervals. Besides this, it is not clear on which evidence we could base the choice of one of these tapered distribution. One might argue that for a given earthquake catalog, the best fitting distribution will be favorable. However, following *Schoenberg and Patel* [2012], different distributions can provide reasonable fits to the bulk of small earthquakes and are only distinguishable in the upper tail, where few, if any, earthquakes have been observed. The choice of the "best fitting model" becomes, therefore, to some extent arbitrary.

The results of this study complete the picture that it is impossible to get hold of the maximum possible magnitude *M* from earthquake catalogs using statistical methods. Moreover, it is questionable, which other hard quantitative information, except from earthquake catalogs, is available that might allow us to constrain the maximum possible magnitude. Of course, estimators of *M* can be constructed and will perform better with increasing data amount. If the imposed error one is willing to accept is not too small such estimators might be feasible for certain applications. On the other hand, for high levels of confidence, the range of values for possible values for *M* becomes unlimited in most cases [*Holschneider et al.*, 2011].

This prohibits to estimate *M* properly, even if time progresses within a human time scale and the amount of data grows. In contrast, alternative tests based on other models of seismicity and different observables as the total number of events and the maximum observed event as proposed in the present study are always possible, but reasonable error values may be achieved in hundreds to thousands of years at least. A careful analysis has to be carried out in these cases. In summary, all studies that estimate the maximum magnitude for all times from earthquake catalogs alone become highly questionable. These problems can be solved, if the maximum magnitude in a predefined finite time window is considered [*Zöller et al.*, 2013]. The introduction of such a time window leads of course to a different concept as it is used in probabilistic seismic hazard assessment. This allows at least to perform particular investigations, e.g., if the earthquake hazard is related to the lifetime of a specific building or infrastructure. In this case, the time window will be set to the lifetime of the infrastructure and the maximum magnitude for this time horizon can be estimated properly. As soon as the time horizon tends to infinity, the concept will fail again. We suggest that future work should focus on constraining the maximum magnitude using additional data and information that are independent of earthquake catalogs.

8. Data and Resources

NOAA earthquake catalog (684–2013) of Japan available via http://www.ngdc.noaa.gov, last accessed 7 January 2014.

Acknowledgments

This work was supported by the German Research Society (HA4789/2-1), the Potsdam Research Cluster for Geo-risk Analysis, Environmental Change and Sustainability (PROGRESS), and the Global Earthquake Model (GEM). We thank the Editor Robert Nowack, the Associate Editor, and two anonymous reviewers for providing helpful comments on our manuscript.

References

Basili, R., et al. (2013), The European Database of Seismogenic Faults (EDSF) compiled in the framework of the Project SHARE, http://diss.rm.ingv.it/share-edsf/, doi:10.6092/INGV.IT-SHARE-EDSF.

Burkhard, M., and G. Grünthal (2009), Seismic source zone characterization for the seismic hazard assessment project PEGASOS by the Expert Group 2 (EG1b), *Swiss. J. Geosci.*, *102*, 149–188.

Gutenberg, B., and C. F. Richter (1956), Earthquake magnitude, intensity, energy and acceleration, *Bull. Seismol. Soc. Am.*, *46*, 105–145.

Haller, K., and R. Basili (2011), Developing seismogenic source models based on geologic fault data, *Seismol. Res. Lett.*, *82*, 519–525, doi:10.1785/gssrl.82.4.519.

Holschneider, M., G. Zöller, and S. Hainzl (2011), Estimation of the maximum possible magnitude in the framework of the doubly truncated Gutenberg-Richter model, *Bull. Seismol. Soc. Am.*, *101*, 1649–1659.

Holt, W. E., C. Kreemer, A. J. Haines, L. Estey, C. Meertens, G. Blewitt, and D. Lavalleé (2005), Project helps constrain continental dynamics and seismic hazards, *Eos Trans. AGU*, *86*, 383–387, doi:10.1029/2005EO410002.

Johnston, C. L., K. J. Coppersmith, L. R. Kanter, and C. A. Cornell (1997), The earthquakes of stable continental regions—Assessment of large earthquake potential, *Electric Power Research Institute (EPRI)* TR-102261-V1,2, 1–98.

Kagan, Y. Y. (2002), Seismic moment distribution revisited: II. Moment conservation principle, *Geophys. J. Int.*, *149*, 731–754.

Kagan, Y. Y., and F. P. Schoenberg (2001), Estimation of the upper cutoff parameter for the tapered Pareto distribution, *J. Appl. Probab.*, *38A*, 901–918.

Lehmann, E. L., and J. P. Romano (2005), *Testing Statistical Hypotheses*, 3rd ed., Springer, New York.

Peng, Z., C. Aiken, D. Kilb, D. R. Shelly, and B. Enescu (2012), Listening to the 2011 magnitude 9.0 Tohoku-Oki, Japan, earthquake, *Seismol. Res. Lett.*, *83*, 287–293, doi:10.1785/gssrl.83.2.287.

Schoenberg, F. P., and R. D. Patel (2012), Comparison of Pareto and tapered Pareto distributions for environmental phenomena, *Eur. Phys. J.*, *205*, 159–166.

Stirling, M., et al. (2012), National seismic hazard model for New Zealand: 2010 update, *Bull. Seismol. Soc. Am.*, *102*, 1514–1542, doi:10.1785/0120110170.

Toda, S., and B. Enescu (2011), Rate/state Coulomb stress transfer model for the CSEP Japan seismicity forecast, *Earth Planets Space*, *63*, 171–185.

Wells, D. L., and K. J. Coppersmith (1994), New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, *Bull. Seismol. Soc. Am.*, *84*, 974–1002.

Wesnousky, S. G. (1994), The Gutenberg-Richter or characteristic earthquake distribution, which is it?, *Bull. Seismol. Soc. Am.*, *84*, 1940–1959.

Zöller, G., M. Holschneider, and S. Hainzl (2013), The maximum earthquake magnitude in a time horizon: Theory and case studies, *Bull. Seismol. Soc. Am.*, *101*, 860–875, doi:10.1785/0120120013.