Theme I – Introductory Material

CORSSA: the Community Online Resource for Statistical Seismicity Analysis

Andrew J. Michael[1] · Stefan Wiemer[2]

1. United States Geological Survey
2. Swiss Seismological Service, ETH Zurich

How to cite this article:

Michael, A.J., and S. Wiemer (2010), CORSSA: the Community Online Resource for Statistical Seismicity Analysis, Community Online Resource for Statistical Seismicity Analysis, doi:10.5078/corssa-39071657. Available at http://www.corssa.org.

Document Information:

Issue date: 1 September 2010 Version: 1.0

Contents

Abstract Statistical seismology is the application of rigorous statistical methods to earthquake science with the goal of improving our knowledge of how the earth works. Within statistical seismology there is a strong emphasis on the analysis of seismicity data in order to improve our scientific understanding of earthquakes and to improve the evaluation and testing of earthquake forecasts, earthquake early warning, and seismic hazards assessments. Given the societal importance of these applications, statistical seismology must be done well. Unfortunately, a lack of educational resources and available software tools make it difficult for students and new practitioners to learn about this discipline. The goal of the Community Online Resource for Statistical Seismicity Analysis (CORSSA) is to promote excellence in statistical seismology by providing the knowledge and resources necessary to understand and implement the best practices, so that the reader can apply these methods to their own research. This introduction describes the motivation for and vision of CORRSA. It also describes its structure and contents.

1 The Difficulties of Statistical Seismology

To illustrate the need and the woes of statistical seismology, we explore a very simple application: counting earthquakes. The first step toward understanding the hazards posed by earthquakes might be to simply ask, "How many earthquakes have occurred in this region in the past?" And indeed such counts are fundamentally important to seismic hazards assessment and evaluating earthquake forecasts, and so it is important that we be able to make these counts accurately. It may seem like counting earthquakes would be a simple enough exercise to be done by a first grader; however, in practice, counting earthquakes in a meaningful way, and accounting for the uncertainty and imperfection of the underlying data, requires making a series of critical decisions that demand understanding both the ways that we record earthquakes using seismographic networks and the statistical properties of earthquakes.

The first challenge to be tackled is to find out how well we have recorded earthquakes in the past. A seismographic network cannot record every earthquake because as earthquakes get smaller they are well recorded only on stations at shorter and shorter distances. The principal behind this is basic physics: the amplitude of a seismic wave decays with distance. Consequently, at some magnitude level an earthquake will not be recorded on enough stations to be analyzed and some earthquakes are not detected by any stations at all. To produce a meaningful result we must only count the earthquakes above some minimum magnitude set by us, and this minimum should be greater than the so called magnitude of completeness. Picking a minimum magnitude that is lower than the magnitude of completeness can lead to incorrect conclusions. Because seismic networks change with time, the count will only be valid for a certain period.

To be safe rather than sorry, we might be tempted to pick a minimum magnitude for a period that is clearly greater than the magnitude of completeness. For instance, the magnitude of completeness for global earthquake networks over the past 20 or so years varies from M 5.3 to M 6 depending on the region (*Woessner*

and Wiemer 2005). Certainly, we can be confident that global networks can easily detect and analyze all earthquakes over magnitude 7 (unless we are in the seconds after a very large earthquake – but that is yet another story). However, if we choose to count only earthquakes of magnitude 7 or greater, we find that there are only about 15 events per year around the globe (*USGS* 2010). If we want to count earthquakes in either smaller units of time or smaller areas, then any statistical analysis will become impossible due to the small numbers of events.

To have larger numbers of events in our counts we can reduce the magnitude threshold. This is very effective because every time we go down one unit of magnitude there are approximately 10 times as many earthquakes; this empirical observation is described by the frequency-magnitude relationship known as the Gutenberg-Richter relationship (*Gutenberg and Richter* 1944). Therefore, we would expect to count about 150 earthquakes above magnitude 6 per year and that is the value modern seismographic networks currently record around the globe. But as we get closer to the magnitude of completeness we need to be careful. For instance, if we count the number of magnitude 6 or greater earthquakes in the early 1900's there are only 10 to 50 events per year through 1923 (*International Seismological Centre* 2010). This suggests that the number of earthquakes increased at some point and that would be a very interesting observation, if it were true. However, this observation can be explained by the fact that in the early1900's the global earthquake catalogs were complete only down to about magnitude 7 (*Engdahl and Villaseñor* 2002). As the networks improved, they recorded more earthquakes in the magnitude 6 to 7 range, and, thus, the observed count of earthquakes above magnitude 6 increases even though there is no evidence that the actual number of magnitude 6 or greater earthquakes actually changed. Further complications in counting earthquakes include that the magnitude of each earthquake is uncertain, that different magnitude scales exist, and, if we are trying to count the events in a given region, that the location and specifically depths of the events are uncertain. So in summary, the magnitude of completeness that we need to know for our count is a complex and uncertain function of space and time.

The problems only get worse if we try to now do something useful with these counts. A standard step in seismic hazards assessment would be to determine the probability of a damaging earthquake occurring in a given region during the next years or decades. A simple approach would be to pick a minimum magnitude that produces damage, then count the number of earthquakes over that magnitude during some past period of time, convert that count into a rate by dividing it by the length of the past time period, and then compute the probability of an event during the future time period using the Poisson model.

The Poisson model is a simple statistical model that assumes that each earthquake is an independent event that occurs with equal probability (randomly) at any point in time. A serious problem with the Poisson model is that it does not describe the actual occurrence of earthquakes unless we average over a large region and/or long time period. On shorter time scales, and for more local estimates, earthquakes are not independent events but instead cluster together in time and space. In other

words, earthquakes interact with each other. The most well known evidence of this interaction, generally called clustering, are the prominent aftershock sequences that follow large earthquakes. The process of clustering also includes the well-known phenomenon of foreshocks–earthquakes which have an 'aftershock' that is larger than the initiating event, as well as earthquake swarms–earthquake sequences with many events of about the same magnitude. Clustering takes place on a wide variety of spatial and temporal scales even after small earthquakes and its physical mechanism is only partially understood today.

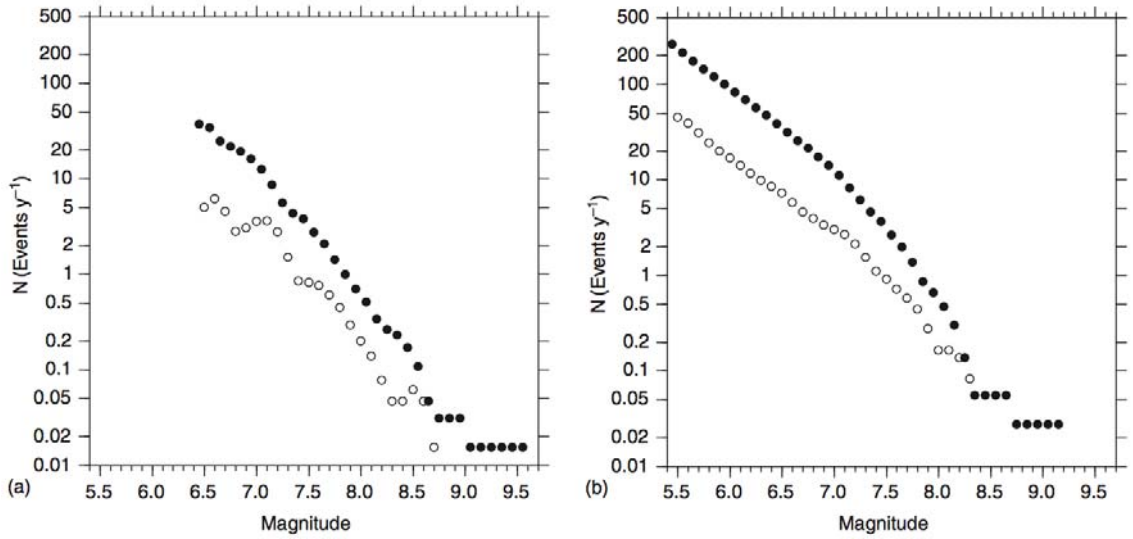

Fig. 1 Frequency magnitude (Gutenberg and Richter) relations for the Centennial catalog. Open circles represent single frequencies (incremental number of earthquakes with magnitudes in M±0.05) and filled circles represent cumulative frequencies (total number of earthquakes with magnitudes ≥M). The single and cumulative frequencies are normalized to events per year, and the magnitudes have been adjusted to M_S: (a) historical seismicity (1900 - 1963), and (b) recent seismicity (1964 - 1999). From *Engdahl and Villaseñor* (2002).

Applying the Poisson model, which is heavily used in probabilistic seismic hazard assessment, to clustered data is invalid and can produce meaningless probabilities. If we try to determine the average, long-term background rate but base our estimate on a period that happens to include a rich aftershocks sequence, we will estimate long-term rates that are too high. Aftershock sequences to large earthquakes in well-monitored regions can often contain thousands or even tens of thousands of events, so the clustered part can be more than half of the observed seismicity in a region. One solution to this problem is to decluster the catalog of earthquakes by removing aftershocks and other non-independent events from it. While this is a standard approach, there is no perfect declustering method because we have no unique and universally accepted description of the process that creates clustering in the first place, and so the results depend on which method is chosen. Therefore, declustering introduces additional uncertainty into the results. Another

approach is to use a statistical model that includes clustering in its distributions. However, doing that requires knowing the form of these more complicated distributions. These models, just like declustering schemes, are controversial and choosing one introduces similar uncertainties.

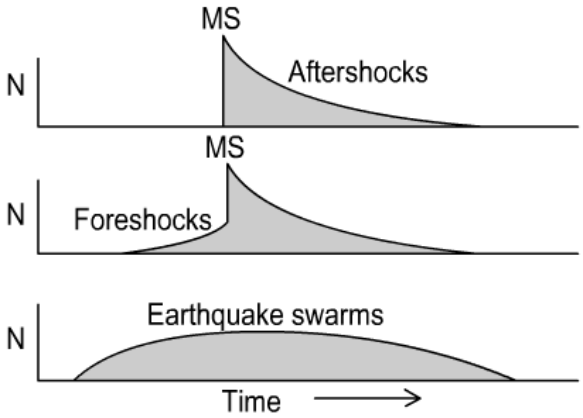

Fig. 2 Schematic view of the three types of time-dependent earthquake occurrence. The number of events is plotted as a function of time. MS indicates the mainshock.

Clustering also strongly affects evaluations of earthquake prediction and forecasting methods. To evaluate a proposed prediction method, one can compare its success against a simpler and widely accepted approach known as a null hypothesis. Tests of earthquake prediction methods often fail to include earthquake clustering in the null hypothesis despite that this is a well-known feature of earthquake catalogs. This is unfortunate because it has been demonstrated that clustering is an important factor in testing earthquake predictions. For instance, when studying a proposed earthquake precursor based on propagation anomalies in very low frequency magnetic waves, *Gokhberg et al.* (1989) and *Marenko* (1989) ignored earthquake clustering in their null hypothesis and reached the conclusion that the proposed VLF precursor was statistically significant. In discussions, they agreed that ignoring earthquake clustering was likely to be a problem, but they didn't know how to address it. Using relatively simple, albeit incomplete, models of temporal earthquake clustering, *Michael* (1997) was able to demonstrate that their positive result was an artifact of ignoring clustering and that the proposed precursor was not statistically significant. The basic issue is that when a proposed precursor has free-parameters (e.g. the length of an alarm window) that are optimized by fitting the data, we are essentially looking for the most extreme behavior. When the data includes clustering, extreme behaviors become more extreme and this must also be accounted for in the null hypothesis. *Stark* (1996) reached a similar conclusion from a theoretical perspective while examining the proposed VAN earthquake prediction method (*Varotsos et al.* 1996) and how it had been tested.

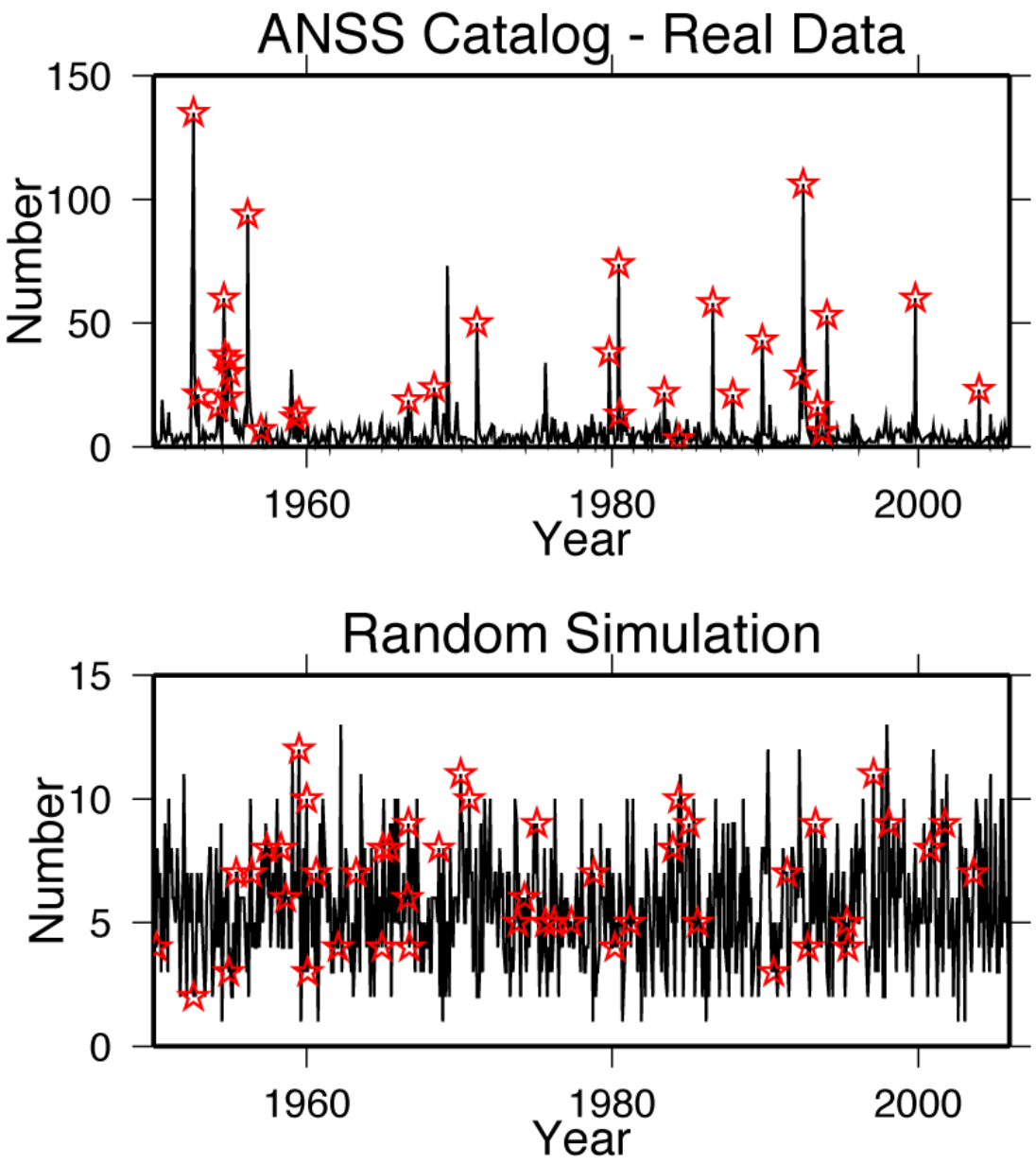

Fig. 3 The number of magnitude 4 or greater events, per year, in California and Nevada from 1950 through 2005 with the years containing magnitude 6 or greater events shown as red stars. Note the high degree of variability in the real data where many events cluster with the larger earthquakes. The bottom panel shows a random simulation of the same data using a Poisson model. Note the much smaller range of numbers of events per year and that the large earthquakes have no relationship to the number of events. After *Hardebeck et al.* (2008).

2 The Statistical Education of Seismologists and the Seismological Education of Statisticians

Now that the field of statistical seismology "knows" about the importance of clustering it would be good if including this effect became standard practice. Perhaps because the seismological literature is so large, and therefore useful ideas can be missed, this is not the case. For instance, a front-page article in *EOS* (*Kamogawa* 2006) recently stated that very low frequency electromagnetic precursors "do exist." This was based not only on the works mentioned above, but on a recently published result (*Liu et al.* 2006). Unfortunately, the authors of this new paper were not aware of the importance of including clustering in their null hypothesis and so appear to have repeated the earlier mistake. We cite this case not to criticize these authors but only to illustrate the general problem that advances in statistical seismology are not sufficiently utilized.

It is a regrettable fact that, despite its critical importance to our science, and especially to parts of our science with great impact on public policy, statistical seismology is almost ignored in the education of seismologists. Given that seismology is a field of applied physics, it is reasonable that starting with only Hooke's Law we are taught to derive the wave equation, Snell's Law, reflection and refraction coefficients, and the behavior of surface waves. But it is also a field of applied statistics and few of us are taught even the most basic methodologies.

For instance, while most seismology texts mention the Gutenberg-Richter magnitude-frequency relationship, few include *Aki*'s (1965) demonstration that the optimal way to estimate the parameters of this relationship is by using maximum likelihood. Even his own classic textbook (*Aki and Richards* 1980) fails to mention this despite having a section on inverse theory, in which they discuss maximum likelihood. Two exceptions are *Bullen and Bolt*'s (1985) seismology text and *Reiter*'s (1990) text on hazard analysis. But the latter mentions using least-squares as an alternative method without discussing *Bender*'s (1983) paper that demonstrates that this approach results in a bias of the parameters which depends on the number of data points. Thus it is not surprising that many researchers are poorly informed and may fall prey to the traps of using inadequate methods.

Similarly, the operation of seismic networks is a specialized subdiscipline of seismology in which waveform data from a network of seismometers is converted into a list of earthquakes known as a catalog. Online data centers have made it very easy to access these catalogs that appear to be relatively simple data sets. However, there are a host of issues that can create uneven data quality within a catalog. For instance, in urban areas, the level of noise in the waveforms is lower at night when people are not working or driving, and this effect can create an apparent diurnal signal in the number of earthquakes. The addition of seismic stations as a network grows, the temporary removal of stations when they are not working, and changes in processing techniques can create longer period apparent variations in earthquake catalogs. Unfortunately, few seismologists, and perhaps

even fewer statisticians, fully appreciate the wide variety of issues that impact the quality of earthquake catalogs. But without understanding the issues of how one gets from the basic data, which are waveforms, to earthquake catalogs it is difficult to properly apply statistical methods when doing research based on the catalogs.

The fact that many researchers fail to appreciate the intricacies of earthquake catalogs and fail to use state-of-the-art statistical seismology methods degrades the quality of seismological research and slows our progress. While a large fraction of these questionable studies do not make it into the peer-reviewed literature, because reviewers and editors of journals are on the lookout, many do end up published. Educating reviewers is thus also needed.

A final observation is that with increasing amounts of data available for analysis due to improved seismic networks and processing techniques, and with increasing computer power, it is now possible to analyze and simulate the space-time evolution of seismicity in every increasing detail. Modern earthquake networks record upwards of 100,000 earthquakes every year. The statistical seismology community, therefore, has seen a rapid growth in the past decade, as more and more scientists realize that in order to exploit this wealth of data, they need the very basic as well as quite sophisticated approaches of statistical seismology.

3 CORSSA – A New Educational Vision

The goal of the Community Resource for Online Statistical Seismicity Analysis, a.k.a. CORSSA, is to promote best practices in statistical seismology by providing the relevant knowledge and resources necessary to understand the best practices so that the reader can apply them to their own research needs.

CORSSA covers a wide variety of ***themes***:

I. Introductory Material
II. Introduction to Basic Features of Seismicity
III. Basic Features of Statistics Applicable to Seismicity
IV. Understanding Seismicity Catalogs and Their Features
V. Basic Techniques for Analyzing and Modeling Seismicity
VI. Methods for Testing Earthquake Predictability and Other Hypotheses
VII. Data Standards

Each of these themes includes a series of articles that are listed in the CORSSA Table of Contents.

The series of themes was devised to make it easy for the reader to focus on their personal requirements to get an introduction to statistical seismology (Theme I), or to learn about the basics of earthquakes (Theme II), statistics (Theme III), and/or the intricacies of seismicity catalogs (Theme IV) before moving onto applications found in Themes V and Theme VI. Theme VII provides information about data formats and standardized data sets that can be used for testing computer codes.

CORSSA is an open community of authors and readers. It is a community of authors because it requires many people to cover the breadth of expertise necessary to authoritatively address these complex problems. Authority in science stems from authors with known expertise and from implementing a peer review system for CORSSA articles. Thus, identified experts write articles for CORSSA, these articles are peer reviewed and eventually are approved or rejected by an editorial board. CORSSA expands this community through an online forum that allows readers and authors to discuss the articles and issues. These comments will be used to improve the articles.

CORSSA is an educational resource and will contain only methods that have already been published in established peer-reviewed journals. CORSSA will not contain new scientific results. Such advances should be published through the traditional scientific journals before being included in CORSSA.

CORSSA is a living online resource so that it is open-access, in order to take advantage of new publishing approaches that are not possible on the printed page, to develop a dialogue throughout the CORSSA community by including forums in the resource, so that it can go online when the first sections are completed, and so that it can be frequently expanded and updated.

4 Expectations of a CORSSA Article

The goal of each article is to provide a tutorial that relies on the published, peer-reviewed literature. Each article covers a specific task or topic discussing why the topic is useful for research, a brief referenced review of theory, a list of methods and software that address this topic, a discussion of tradeoffs in analysis choices, pitfalls to be aware of, example results from applying the method to one of the CORSSA standard data sets, examples of excellent applications in the scientific literature, pointers on further reading, and next steps for the reader to take.

The audience or readers that we envision is quite varied. A CORSSA article should serve undergraduate students as a starting point to understand the issues, it should serve graduate students as a resource for their own research. Last but not least, it should serve experienced researchers from outside the statistical seismology group, and even from within that group, as a point of reference and resource to enhance the quality of their research. The Discussion Forum we envision likewise can address a wide range of issues, from the seemingly simple to expert discussions.

CORSSA also provides a glossary that we hope will provide a growing resource and build the basis for a common ontology of terms used in statistical seismology.

5 Expectations of software cited by CORSSA

CORSSA articles seek to provide links to software that can carry out the analysis steps described in the articles. The open exchange of software between researchers is a key part of improving research in statistical seismology. Standard data sets are included that can serve as a starting point for understanding the method and software. The software cited by CORSSA articles and listed in the CORSSA table of software is largely written by individual researchers for their own purposes and much of it is made available directly by them. CORSSA seeks to cite software that is widely used in statistical seismology but does not guarantee the accuracy of any software.

6 How to Cite a CORSSA Article

The authors of CORSSA articles donate their time to this effort, because they are convinced that this resource will benefit others and also themselves. Readers can support and acknowledge their efforts by citing the resource and these articles in their research papers. Citations of CORSSA articles should refer to the CORSSA web site and the DOI included on the title page of each article.

7 Invitation to Contribute

CORSSA invites you to contribute new articles to the resource, to become a co-author of an existing article, or to provide software for use by CORSSA's readers. To contribute a new article, please send a one-page proposal outlining the contents of the article and where it fits into CORSSA's Table of Contents to the Executive Committee at contributions@corssa.org. To contribute to an existing article please contact the lead author of that article. If you have written software that should be mentioned in one of the articles, please contact the authors of that article. To have software listed in the CORSSA table of software, please contact software@corssa.org. Please remember that CORSSA does not publish new scientific results and that the basis for all methods must have been previously published in the peer-reviewed scientific literature.

CORSSA likewise is keen to know your opinion of the articles and its content and the discussion forum provided with each article. Is the article clear, are elements missing, are there additional resource that should be pointed out to other readers? Our hope is that the discussion forum will become an important asset of its own and a place to exchange ideas and resources.

8 CORSSA Publication Process

After approval of a proposal, all submissions to CORSSA must be prepared using the CORSSA LaTeX or Microsoft Word templates available on the web site. The submission will undergo peer review by one or more referees under the supervision of a member of the Executive Committee. The Executive Committee makes all acceptance and rejection decisions.

9 CORSSA Organizational Structure

CORSSA is overseen by an executive committee of seven people who are responsible for the operation and promotion of CORSSA and for overseeing the contents of the resource. The executive committee acts as an editorial board and is responsible for peer review of contributed articles and for further oversight of the contents of CORSSA. All members of the CORSSA community are invited to shape the resource by sharing ideas for improvement, new articles, and by contributing comments to the forum.

References

Aki, K. (1965), Maximum-likelihood estimate of b in the formula N=a 10 ** -bM and its confidence limits, *Bull. Earth. Res. Inst.*, v. 45, 237-239.

Aki, K., and P.G. Richards (1980), Quantitative Seismology, San Francisco, Freeman, 932 p.

Bender, B. (1983), Maximum likelihood estimation of b values for magnitude grouped data, *Bull. Seismol. Soc. Am.*, 73, p. 831-851.

Bullen, K.E., and B.A. Bolt (1985), An Introduction to the Theory of Seismology, Cambridge, Great Britain, University Press, 499 p.

Engdahl, E.R., and A. Villaseñor (2002), Global Seismicity: 1900-1999, in Lee, W.H.K., P.C. Jennings, C. Kisslinger, and H. Kanamori, eds., International Handbook of Earthquake and Engineering Seismology: Academic Press, p. 665-690.

Gokhberg, M.B., I.I Gufeld, A.A. Rozhnoy, V.F. Marenko, V.S. Yamplosky, and E.A. Ponomarev, (1989), Study of seismic influence on the ionosphere by super long-wave probing of the Earth-ionosphere waveguide, *Phys. Earth Planet. Inter.*, 57, 64-67.

Gutenberg, B., and C.F. Richter (1944), Frequency of earthquakes in California, *Bull. Seismol. Soc. Am.*, 34, 185-188.

Hardebeck, J.L., K.R. Felzer, and A.J. Michael (2008), Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant, *J. Geophys. Res.*, 113, B08310.

International Seismological Centre (2010), Online-Bulletin, www.isc.ac.uk.

Kamogawa, M. (2006), Preseismic lithospheric-atmospheric coupling, *EOS*, 87 (40), 417.

Liu, J.Y., Y.I. Chen, Y.J. Chuo, and C.S. Chen (2006), A statistical investigation of preearthquake ionospheric anomaly, *J. Geophys. Res.*, 111, A05304.

Marenko, V.F. (1989), Investigation of the relationship between seismic processes and disturbances to the lower ionosphere by means of VLF radio transmissions, Ph. D. Dissertation, Irkutsk, USSR Academy of Sciences, Siberian Department.

Michael, A.J. (1997), Testing prediction methods; earthquake clustering versus the Poisson model, *Geophys. Res. Lett.*, 24 (15), 1891-1894.

Reiter, L. (1990), Earthquake Hazard Analysis, Issues and Insights, New York, Columbia University Press, 254 pp.

Stark, P.B. (1996), A few considerations for ascribing statistical significance to earthquake predictions, *Geophys. Res. Lett.*, 23, 1399-1402.

USGS (2010), Earthquake Facts and Statistics, USGS.

Varotsos, P., K. Eftazias, F. Valianatos, and M. Lazaridou (1996), Basic principles for evaluating an earthquake prediction method, *Geophys. Res. Lett.*, 23, 1295-1298.

Woessner, J., and S. Wiemer (2005), Assessing the quality of earthquake catalogues: estimating the magnitude of completeness and its uncertainty, *Bull. Seismol. Soc. Am.*, 95, 684-698.