

Award Number: 07HQGR0034

**Probabilistic Estimates of Network Completeness:
Creating a living SCEC/ANSS Resource**

Thomas H. Jordan and Danijel Schorlemmer

Southern California Earthquake Center
University of Southern California
Department of Earth Sciences
Los Angeles, CA 90089-1147
Office: (213) 740-5843
Fax: (213) 740-0011
tjordan@usc.edu
ds@usc.edu

Abstract

During the award period, we were investigating multiple seismological networks and addressed the scientific questions posed in the award proposal. To account for incomplete raw data sets, we investigated the correlation of detection-probabilities with attenuation curves used at a particular network. For the Southern California Seismic Network we were able to show basic correlation, however, comparison with geological site conditions needs further to be performed. The present results are very promising and helped identifying part of the detection-probability distribution which should be used with care.

We analyzed the impact of earthquake clusters on the method. This investigation was carried out during an initial completeness computation for the Northern California Seismic Network. We identified two large clusters, the Geysers cluster and the aftershock sequence of the San Simeon earthquake. We found significant changes in the detection-probability distributions used for computing the probability-based completeness magnitude. We compared computations with unchanged catalogs to computations using catalogs from which we removed the clusters. The detection-probability distributions from the corrected catalogs were more regular and our preferred choice for all subsequent computations.

During a study of the Swiss Seismological Networks, we developed a method to estimate the uncertainty of the computed probabilities and completeness magnitude. Because no analytical solution exists that describes the propagation of uncertainties of the raw data into uncertainties of completeness magnitudes, we have chosen to employ a bootstrap approach to estimate uncertainties of the detection-probability distributions. These uncertainties are then propagated through a Monte Carlo approach to uncertainties of completeness magnitudes. This approach is computationally intensive but is providing important insight of reliability of the detection-probability distributions. These results helped us understanding possible parameterization scenarios of these distributions as well as the quality of correlations of attenuation curves.

For the computation of completeness magnitudes for the Northern and Southern California Seismic Network, we completely developed new codes for the PMC method using the programming language Python. The codes are now included in the QuakePy project, a seismicity analysis toolkit that is fully based on the QuakeML data model. This approach will facilitate further use of the codes as Python is fully open source and license free, and the Advanced National Seismic Network is in the process of supporting of QuakeML as new data standard. The new codes are fully object-oriented and allow for easy adaption of the codes to new networks. The data structures are completely defined in XML and the result structures can be directly fed into a webservice. This webservice, which we also developed, can be used by everybody for data and image retrieval. The webservice is designed to host completeness estimates of multiple networks.

This work has result in the following publications:

- Nanjo, K. Z., J. Woessner, S. Wiemer, D. Schorlemmer, and D. Giardini, Earthquake detection capability and its uncertainties of seismic networks in Switzerland, in preparation.
- Bachmann, C. E., D. Schorlemmer, and E. Kissling, Probabilistic Magnitude of Completeness of Northern California, in preparation.
- Schorlemmer, D. F. Euchner, A Completeness Webservice for California, in preparation.

Summary of Results During this Award Period

During the award period we were able to achieve multiple results which are the basis for three publications. Besides scientific achievements, we also developed a completely new implementation of the PMC codes which are freely accessible and published under an open-source license.

Parameterization of Probability Distributions

The detection-probability distributions per station show in general a relatively regular behavior. With increasing magnitude, the probabilities of detection are becoming larger, and with increasing distance, they are becoming smaller. However, despite this systematic changes, irregularities in the distributions exist. One example can be seen in the right frame of Figure 1. Detection probabilities for events of magnitudes 3–3.5 at distances to the station of less than 30 km are much lower (red patch) than for lower magnitudes at the same distance. Such a distribution violates basic principles as higher-magnitude events should be easier to detect.

We introduced during the review process of [Schorlemmer and Woessner, in print] a smoothing algorithm to the detection-probability distributions of stations. This smoothing algorithm applies basic physical principles as it simply does not allow probabilities for a given magnitude to decrease with decreasing distance and for a given distance to decrease with increasing magnitude. The top frame in Figure 1 shows the smoothed version of the detection-probability distribution of the one in the right frame. The smoothed distribution gives a much clearer picture of the detection probabilities. When tracing the yellowish color, which corresponds to $P_D \approx 0.8$, one can see a gradual increase. Such an increase is expected, however, for higher magnitudes ($M \geq 3.5$), a larger step is visible. We investigated the influence of these remaining irregularities for higher magnitudes and found that they can be neglected as many stations contribute to the detection probabilities of a network for such events ($M \geq 3.5$). Only at the edges of a network, such a data flaw may have a slight influence. We decided that it is beneficial to accept such a minor influence than to try to correct the detection-probability distributions, which may introduce much larger errors.

We inspected detection-probability distributions of all stations contributing to the Southern California Seismic Network to understand the probabilities for low-magnitude events ($M \leq 1$). The example distribution in the top frame of Figure 1 shows a problem: Detection probabilities for events of magnitude 0–1 at distance of 0–20 km are consistently high ($P \approx 1$). It is also noticeable from the distribution that the probabilities for detecting $M = 0$ events are higher than expected. We investigated this effect and found that the relatively high probabilities for very low magnitude (here $M = 0$) stem from small raw data samples that contain only information about picked events. This has to be expected as for events that are so small that only four stations can detect them, missed events are not in the record. As soon as more than four stations can pick an event, a miss at one station will still keep this event in the records and the miss at the particular station is recorded. In the case where only four stations can possibly see the event, any miss cause the event to not be recorded and the misses at the any of the stations are not on the record. Thus, for this

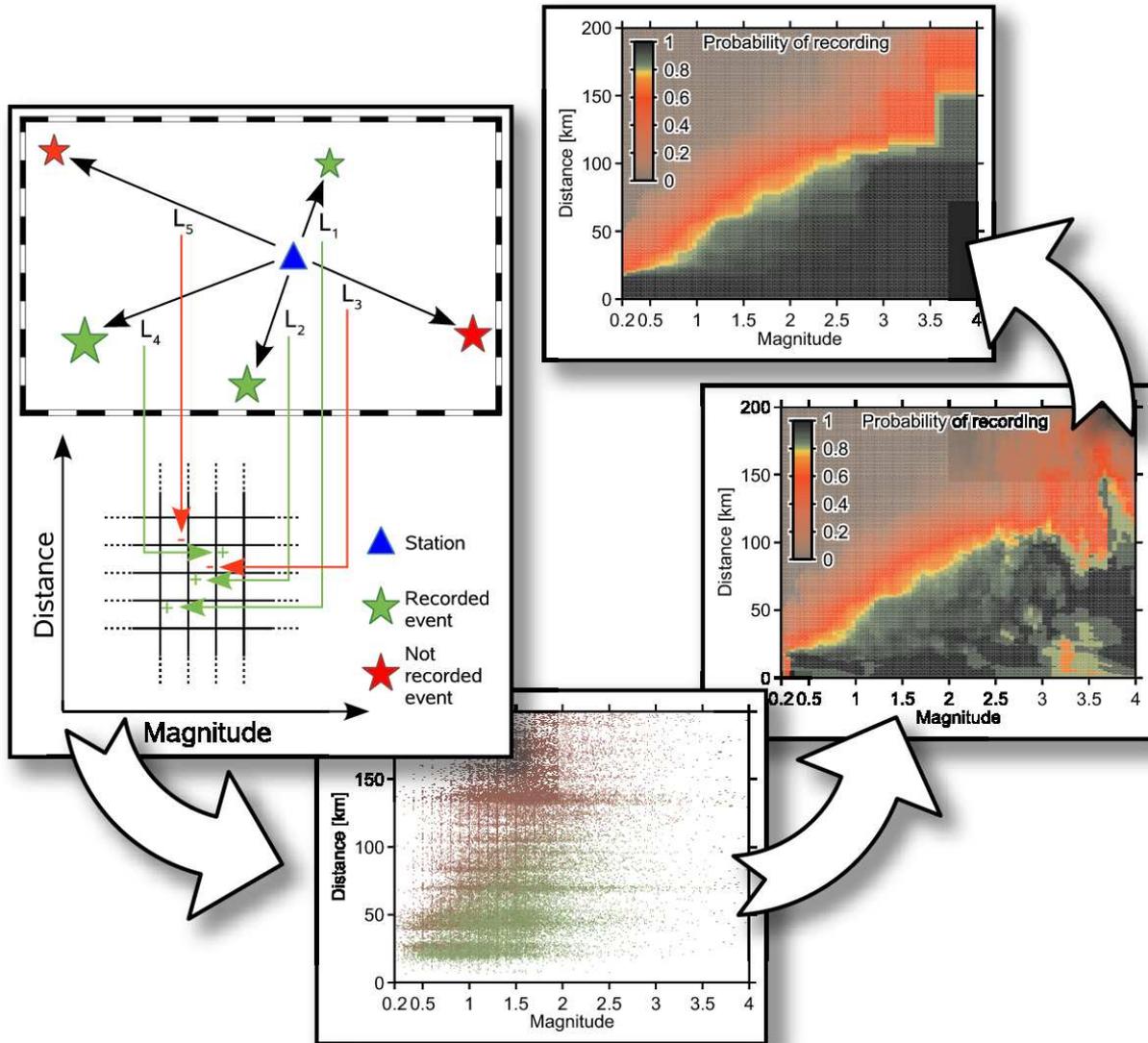


Figure 1: Detection probabilities for each station are derived from earthquakes with and without phase picks at the station that occurred during on-times of the station. (Left) Raw data (green: picked events, red: not-picked events) are generated for each event which occurred while a station (blue triangle) was in operation. (Bottom) Plot of all raw data points of station POB (CI network) for the period 1 January 2001–1 July 2007. (Right) Distribution of detection probabilities over magnitude and distance of station POB derived from raw data points. (Top) Smoothed distribution of detection probabilities.

class of events, only successful picking is recorded (green dots in left frame of Figure 1) and the probabilities of detection are high, because they are directly derived from this raw data. Due to smoothing, the high probabilities propagate into higher magnitudes because for larger magnitudes (events that can be detected at five or more stations) the detection probabilities are lower because of realistic raw data. In the smoothed version, these probabilities become equally high, leading to overly optimistic detection capabilities. This observation leads to the conclusion that this method is limited to detection-probabilities for events large enough to be detected at five or more stations. Although it seems to be a strong limitation, this does not affect completeness estimates at all as events of such low magnitudes are not completely detected and the completeness magnitudes are at much higher values.

The aforementioned observation is supported by the comparison of detection-probability distributions with attenuation curves. The attenuation curves are directly derived from the magnitude definition used at the network because this definition contains the attenuation. Figure 2 shows two detection-probability distributions of two stations of the Southern California Seismic Network. Overlaid are two attenuation curves for arbitrarily chosen amplitudes. These curves fit well lines of constant detection probability in the range of magnitudes 1–3. For smaller magnitudes the curves underestimate the probabilities compared to the distribution. We documented this effect in the previous paragraph. For larger magnitudes, the curves overestimate the detection probabilities. This effect can be explained by the fact that for such events ($M \geq 3$) at larger distances ($L \geq 100$ km) not all stations that receive signals are used for locating the events. This results in unpicked, but visible phases at these stations. As mentioned before, this underestimation does not result in wrong completeness estimates.

The attenuation curves do show clearly a systematic behavior of the detection-probability distributions but they do not readily allow for replacing them with parameterized attenuation. The problems here are manifold:

- Only frequently-used stations exhibit such well defined distributions that can be fitted with attenuation curves.
- The attenuation curve used in this study is derived from the magnitude equation; hence it is based on a general attenuation law and does not account for site-specific attenuation. The advantage of the PMC method is the fact that it takes into account the site-specific attenuation when computing the detection-probability distributions.
- Deviations of the probabilities from the the attenuation curve will also depend on the location of a station and on the number of nearby stations. As mentioned before, a deviation for small magnitudes occurs when only four stations are capable of detecting an event. If the station distribution around a particular station is very dense, this effect will take place for very small magnitudes, with more sparse distributions, this effect will be observed for larger magnitudes.

We plan to correlate detection-probability distributions with local geology settings in order to find common attenuation curves for sets of stations with similar site conditions.

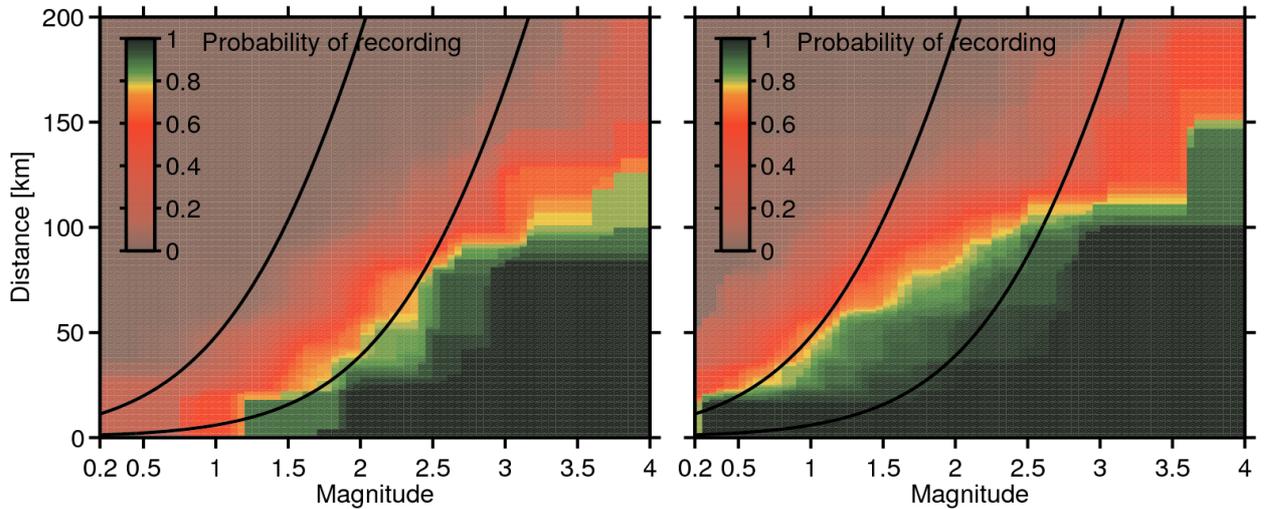


Figure 2: Recording probability distributions with overlaid attenuation curves computed for two different amplitudes. (Left) Probability distribution for station PSP. The attenuation curves very well match the recording probabilities of this particular station. (Right) Probability distribution for station POB. At this station the attenuation curves more strongly deviate from the recording probabilities, indicating a site-specific attenuation.

Cluster Analysis (Aftershocks)

Large clusters of seismicity often cause network operators to change their policies in order to be able to manage locating as many events as possible. This usually means that cluster events will only be located using nearby stations. More distant stations will not be used and therefore their detection probabilities can change. Because a distant station does not detect the cluster events, this fact influences the detection probability for the particular distance range of the cluster events. The question to be answered is whether such clusters affect the computation of overall detection probabilities and probability-based completeness magnitudes.

We analyzed the effect of clustered seismicity on the values of the probability-based magnitude of completeness. We identified two clusters in northern California and computed all detection probabilities with an unchanged catalog and with a catalog in which we removed the clusters. Figure 3 shows one of the investigated clusters, the Geysers cluster. The other cluster is the aftershock sequence of the 2003 San Simeon earthquake. It can be seen on the map that the Geysers cluster consists of a multitude of earthquakes.

Station KIP of the Northern California Seismic Network nicely shows the effect of clusters in the detection-probability distributions. If clusters are present in the catalog, distant station (e.g., KIP for the Geysers cluster) are not consistently used for detecting events in the cluster. This leads to many unpicked raw data points for the distance range of the cluster. The left frame in Figure 4 shows the detection-probability distribution of station KIP derived from a catalog containing cluster

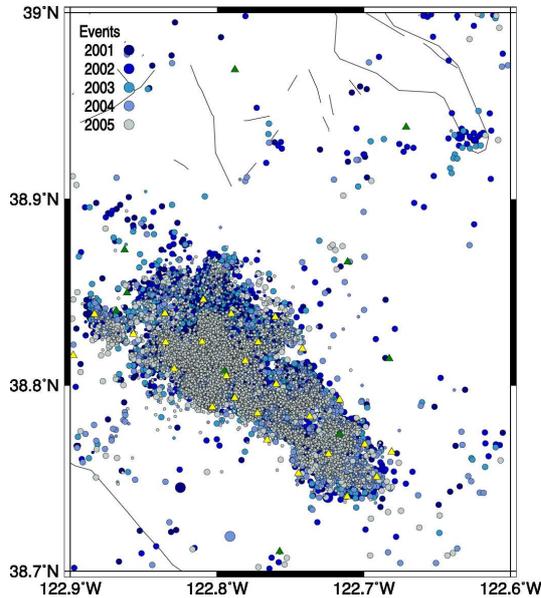


Figure 3: Map showing the seismicity of the Geysers cluster. The events are color coded according to their year of occurrence.

events. At a distance of approximately 120 km, one can see a strong drop in detection probability that does not match the overall trend in this distribution. Figure 5 shows the detections and missed events by station KIP for a distance range, that captures the Geysers cluster. It can be seen that this station has not been used regularly for locating events of this cluster. After removal of the Geysers cluster, the detection-probability distribution shows a more homogeneous picture for this distance range, see right frame in Figure 4.

We came to the conclusion, that removing clusters that cause visible drops in detection-probabilities is improving the results. Therefore, we limited our initial completeness computations for California to the period 1 January 2001–1 July 2007. Further computations will be performed after removal of clusters.

This work on clusters was part of an initial analysis of the completeness of the Northern California Seismic Network and resulted in the publication:

- Bachmann, C. E., D. Schorlemmer, and E. Kissling, Probabilistic Magnitude of Completeness of Northern California, in preparation.

Uncertainties of Completeness Estimates

We developed a method to estimate uncertainties of detection probabilities. Because no analytical solution exists on how to propagate the uncertainties of observations into uncertainties of the target

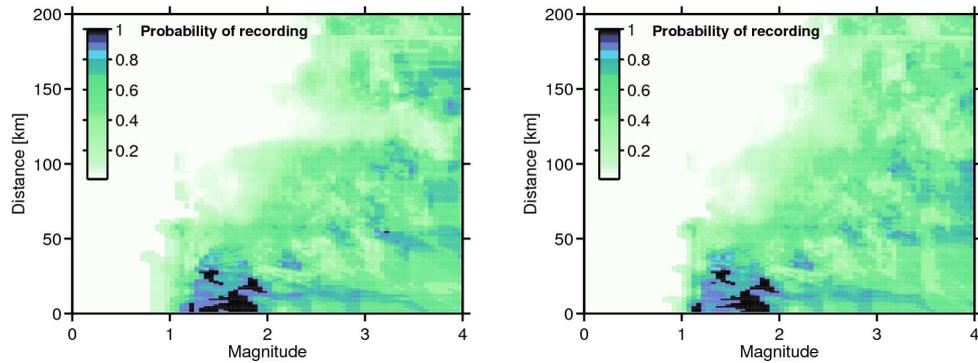


Figure 4: Effect of a cluster on the detection probabilities of a station. (Left) Detection probabilities at station KIP determined from the unchanged catalog. (Right) Detection probabilities at station KIP determined from a catalog in which the Geysers cluster was removed. At the distance of approximately 120 km a remarkable drop in detection probability can be seen.

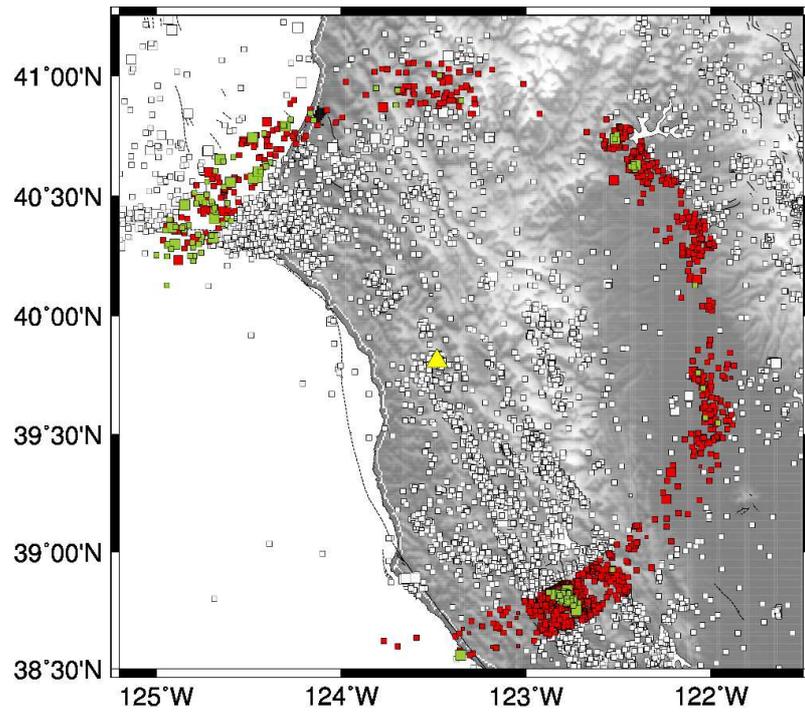


Figure 5: Map showing the picked (green squares) and not picked (red squares) events by station KIP (yellow triangle) for a distance range that matches the Geysers cluster. White squares indicate events outside the given distance range. Although the Geysers cluster is visible as a group of green squares, the number of red squares in the cluster is by far larger.

values, P_D (probability of a station to detect an earthquake at given magnitude and distance), P_E (probability of the network to detect an earthquake of given magnitude at given location), and M_P (probability-based magnitude of completeness of the network at given location), we use a bootstrap approach. We bootstrap the raw observations (information whether earthquakes have been picked or not picked at each station) and compute all values as if we were using the original samples. We repeat this computation 100 times and estimate the uncertainty as a standard deviation σ_D from the distribution of computed values.

This study was carried out on the Swiss Seismological Network, operated by the Swiss Seismological Service. Figure 6 shows the uncertainty computations for the station SENIN. The top frame shows the raw data distribution of this station; the magnitudes are binned in 0.1 magnitude units. The center frame shows the smoothed detection-probability distribution of this station using the original raw data. After bootstrapping the raw data and computing the smoothed distribution 100 times, we obtain the distribution of uncertainties as shown in the bottom frame. The uncertainties are largest for magnitudes and distances that exhibit intermediate detection probabilities ($P \approx 0.5$). In case of very low probabilities, the uncertainty is also very low as bootstrapping leads to basically identical samples. A similar observation can be made for high probabilities, where again the bootstrapped samples are identical and the uncertainty low.

We translated this uncertainty into maps of completeness magnitude using a Monte Carlo approach. For each computation of M_P at any location, we do not directly read the probabilities from the detection-probability distribution of each station but perform a Monte Carlo simulation given the precomputed uncertainties. Repeating the simulation 100 times, results in an uncertainty map as shown in Figure 7.

This work on uncertainty estimates was part of an analysis of the completeness of the Swiss Seismological Service and resulted in the publication:

- Nanjo, K. Z., J. Woessner, S. Wiemer, D. Schorlemmer, and D. Giardini, Earthquake detection capability and its uncertainties of seismic networks in Switzerland, in preparation.

Result Dissemination

We computed completeness maps for each time point of the period 1 January 2001–1 July 2007 for the Southern California Seismic Network. Currently, computations for the same period for the Northern California Seismic Network are underway and eventually these results will be combined into a California Integrated Seismic Network result. All results will be available directly for download on our newly created webservice at completeness.usc.edu/service. This webservice is embedded into a newly created website explaining the PMC method, distributing the codes used for computations, documenting all computations for reproducibility reasons, and making available presentations about this topic, see completeness.usc.edu.

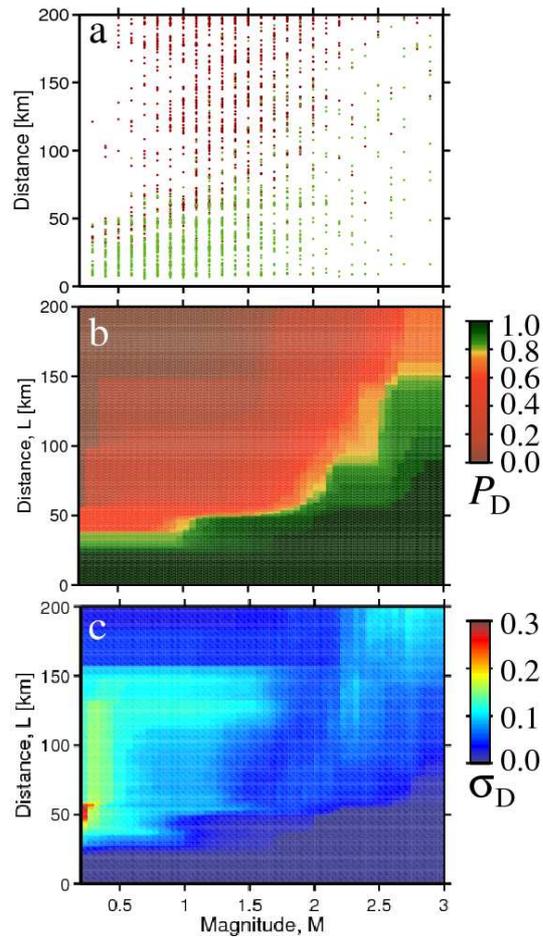


Figure 6: Uncertainty computation for the detection-probability distribution of station SENIN. (a) Raw data distribution. Green dots indicate picked events at this station, red events not picked events, respectively. (b) Smoothed detection-probability distribution. (c) Uncertainties after 100 bootstraps for the distribution in frame (b).

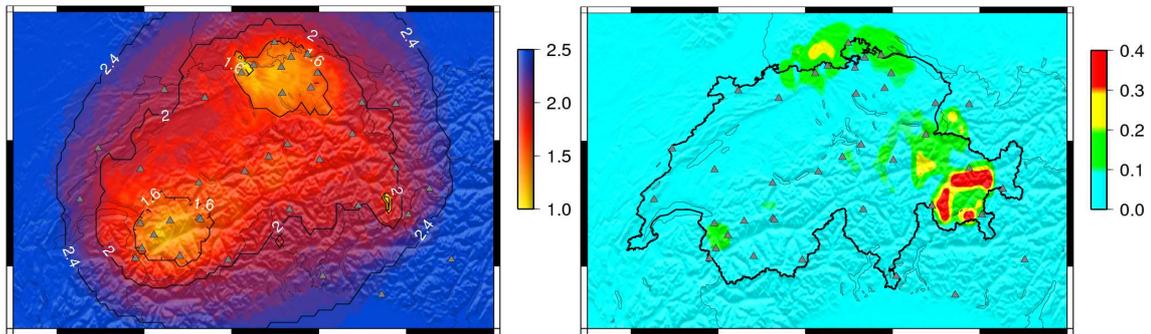


Figure 7: Analysis of the Swiss Seismological Network. (Left) Map of probability-based completeness magnitude, M_P . (Right) Map of uncertainties of M_P .

Software Development

We created an entirely new software package for completeness computations using the PMC Method. The results of the [Schorlemmer and Woessner, in print] paper were computed using our legacy MatLab codes. We have completely rewritten the codes in the programming language Python for multiple reasons:

- Python is open-source and freely available on almost any platform. Unlike MatLab, no licensing fees need to be paid by the user.
- The new codes are embedded in the QuakePy project (www.quakepy.org). QuakePy provides a toolkit for seismic catalog analysis and is also entirely written in Python. QuakePy provides not only pre-processing tools for catalogs that are used in the PMC codes but, more importantly, it also includes the complete QuakeML data model (www.quakeml.org). This allows for using catalogs provided in QuakeML-format which the US networks plan to support within the next year's time.
- Python allows for a clean object-oriented design. Therefore, we designed the codes in such a way that the core computational codes are separated from the network-specific codes. This makes adapting the codes to other networks very easy as only a few functions need to be changed to account for the procedures of a particular network.
- We changed all file formats to XML, for compatibility with QuakeML but also for easy processing of complex data structures. Python comes with many tools to facilitate such processing.

The codes are available via direct subversion checkout

```
svn co https://quake.ethz.ch/svn/quakepy/trunk quakepy
```

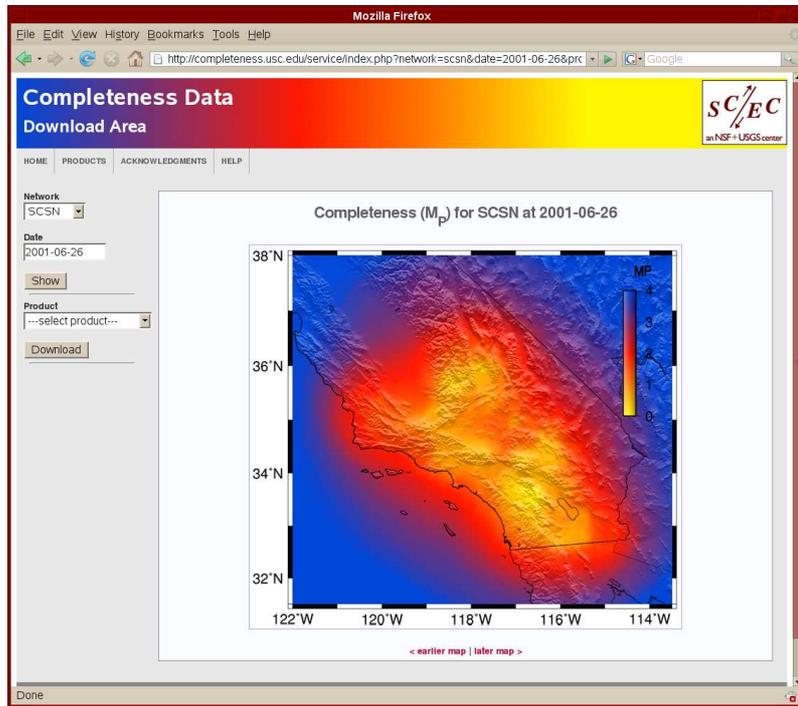


Figure 8: Screenshot of the webservice. On the left menu, the user can choose the network, the date, and a product. Available products are XML result-files, files prepared for GMT, a selection of different image formats, and files containing information about the station configuration of the chosen date.

For data deployment, we developed an openly accessible webservice: `completeness.usc.edu/service`. This service provides interactively the computed values as XML files and GMT [Wessel and Smith, 1991] compatible tab-separated ASCII files, see Figure 8. Additionally, the user can retrieve customized maps of the different computed values. Any data or image retrieval can be done interactively or through a `wget` command for scripting. The codes of the webservice are also available freely under the open-source General Public License.

The computations of detection probabilities and completeness for the Northern and Southern California Seismic Network and work on the webservice for dissemination of these results resulted in the publication:

- Schorlemmer, D. F. Euchner, A Completeness Webservice for California, in preparation.

References

- Schorlemmer, D., and J. Woessner, Probability of detecting an earthquake, *Bull. Seismol. Soc. Am.*, in print.
- Wessel, P., and W. H. F. Smith, Free software helps map and display data, *Eos Trans. AGU*, 72(441), 445–446, 1991.