STEP: Further Development of Short-Term Earthquake Probabilities Codes and Statistical Tests of Forecasts

Award Number: 06HQGR0141

Final Technical Report

Matt Gerstenberger

GNS Science

1 Fairway Drive

Avalon, Lower Hutt

New Zealand

Phone: ++64-4-570-4554

FAX: ++64-4-570-4600

m.gerstenberger@gns.cri.nz

http://www.gns.cri.nz

Abstract

This project described in this report consisted of four topic areas: 1) maintaining the USGS Short-term Aftershock Probability (STEP) real-time hazard pages; 2) developing an operational STEP code in the JAVA programming language; 3) implementing STEP into the Collaboratory for the Study of Earthquake Predictability (CSEP) testing center; and, 4) developing a relationship between ground-shaking amplitude and Modified Mercalli Intensity (MMI). Periodic maintenance on the STEP web pages has been performed to ensure the pages are continually operating correctly. The JAVA code is implemented within OpenSHA and is undergoing final validation. A STEP model is currently operating and being tested in real-time in the CSEP California testing center. Lastly, a probabilistic relationship between amplitude and MMI has been developed and a manuscript is currently under review with the Bulletin of the Seismological Society of America.

Online Code Maintenance

As with any continually operating online and real-time code, the current online version of STEP (`http://earthquake.usgs.gov/eqcenter/step/`) will never be 100 percent free of computer code problems: these problems may arise from such things as bugs in the STEP algorithms, real-time situations which were not considered ahead of time, errors in the various codes and scripts that keep the real-time operations running, or from errors in the supplied input catalog data and associated programming codes. The STEP code is effectively in continual operation with the code being launched every 30 minutes and a complete calculation run taking between 20 and 40 minutes, depending on the number and size of active aftershock sequences at the time. The operation of the code has been under continual monitoring with routine error reports being generated and continual supervision of the online maps to determine any undetected errors. The time since the beginning of the project has required fairly routine maintenance of the product with minor bug fixes (e.g., appropriate handling of an earthquake swarm), improvement to the code to better handle the timing of the automated map generation and handling of computer and environment related errors.

Implementation of STEP into JAVA

The online version of STEP is a collection of Linux shell scripts, PERL scripts, GMT and predominantly MATLAB code. To aide long-term operation of the STEP code, the MATLAB code has been ported to the Java programming language. This porting is important for three reasons: 1) Operation of the code in Java will make it easier for others, currently within the USGS to perform the routine maintenance of the STEP code; 2) porting the code to another implementation allows for an easy way to validate the functionality of the code; and, 3) by developing in Java, under the OpenSHA project, STEP gains a considerable amount of functionality and efficiency that otherwise would have been impossible to implement for any single product.

Progress of the porting to Java was steady throughout the project. The conversion has required a rethinking of the most basic flow of the programming. This is due to the fact that MATLAB programming operates with a different principle than Java; MATLAB is generally based on linear programming (of which the STEP code is 100% linear) while Java requires object oriented programming. Fortunately, by rethinking the basic flow of the programming, we are able to also optimize the code and reduce the time it takes to perform a complete map calculation.

The first step in porting the code was adapting to the new programming environment being utilized by OpenSHA: the Eclipse programming environment. Initially a design of the code was developed with the major classes being defined along with rules for how these classes would interact with each other. This step in the development took place with heavy involvement from OpenSHA experts Ned Field and Nitin Gupta of the USGS Pasadena. After the design, the classes were implemented resulting in approximately 30 classes being written that are specific to the STEP implementation but that interface with OpenSHA. M. Gerstenberger made two trips to Pasadena to work with N. Gupta and N. Field and the STEP Java code. The STEP code is now written in Java and is undergoing final validation to ensure that it is operating correctly.

Testing Center Implementation and Development

During August and September, 2006, M. Gerstenberger spent several days working with Danijel Schorlemmer at ETH-Zürich, Switzerland beginning the development of the computational code that is necessary to control STEP for its implementation within the RELM testing center, and subsequently the California Testing Center section of the Collaboratory for the Study of Earthquake Predictability (CSEP). CSEP is a Southern California Earthquake Center (SCEC) related collaboratory that has taken over the RELM efforts. In order to implement STEP within the testing center it is necessary to design an interface that will allow the model, and all other models, to communicate with the testing code and also to receive any data that is necessary for creating a forecast. We have designed a system that:

1. Explicitly prescribes how a model forecast is to be stated. This includes the actual textual format the forecast must be provided in to the testing center.

2. Automatically downloads and archives an updated catalog of seismicity data for the RELM testing region within California. This catalog can be updated every X number of minutes, depending on what is required based on a particular test definition (e.g., 24 hours).

3. Automatically declusters the input catalog using a Monte Carlo approach to the Reasenberg (1985) declustering algorithm and archives the catalog.

4. Automatically initiates a forecast calculation for a real-time forecast model and archives the resulting forecast.

5. Performs a likelihood based test of the forecast using multiple time intervals and comparing to the observed earthquakes. All data input into the test is archived within the system.

The original development of the testing center ideas and methodologies are explained in Schorlemmer and Gerstenberger (2007) and Schorlemmer et al. (2007). CSEP staff have since taken over development of all testing center related code developments and are now regularly making official code updates and releases: `http://us.cseptesting.org/`.

Since December of 2007 an version of the STEP model has been implemented in the CSEP California testing center and has been used as a prototype model for development of the testing software for aftershock models. As a result of this, the model has been undergoing daily testing since it was first implemented. Via this testing a bug has been found in the layout of the spatial grid for the STEP model as implemented within the testing center; once this bug is corrected it will be possible to rerun the test back to the initial starting date. This bug does not affect the USGS online version of STEP.

Ground-shaking to MMI Relationship

M. Gerstenberger, Bruce Worden and David Wald from the USGS initially developed a new linear relationship between ground-shaking amplitude and Modified Mercalli Intensity (MMI). This relationship improved upon a past model (Wald et. al., 1999) and derived an optimized fit of a line with a single hinge point relating peak ground velocity to MMI. It was decided that this initial model could be improved upon and that errors in the standard linear regression approach could be minimized by deriving a probabilistic relationship between ground-shaking amplitude and MMI. The argument is that before deriving a linear regression, a functional form of the model must be assumed (e.g., a line with a single hinge point) and that by taking a probabilistic approach no basic assumption about the model type is necessary. Additionally, with a probabilistic approach, one can obtain different descriptions of the model by deriving the model at different probability levels such as the mean (50%) model or the 67% model (one standard deviation).

The CIIM and strong ground motion data sets have provided an extensive data set for developing the methodology and contain approximately 34,000 CIIM-amp data pairs within 2km of each other. The methodology we developed takes advantage of this large data set to develop a MMI probability density function for each amplitude level we are interested in. We have developed a relationship between peak ground velocity (PGV) and MMI as well as other ground motion measures: peak ground acceleration (PGA), and 0.3-second, 1-second and 3-second pseudo-spectral acceleration. Additionally we have developed reverse relationships, Amp(MMI) which allow one to calculate a desired ground motion measure given a MMI.

The basic steps involved are for the MMI(Amp) relationship are:

1. Associate all amplitudes and MMIs from each event

2. find all MMI observations that are within 2km of each amplitude

3. bin the amplitudes in 100 bins evenly spaced on a logarithmic scale
4. find the number of observed MMI in .1 MMI bins, within each amplitude bin
5. calculate the cumulative probability of observing MMI≥1 using the number of observations in each amplitude bin
6. for each amplitude bin, using the cumulative probabilities, we find:
   (a) the median MMI (i.e., 50% probability of exceedance) as shown in Figure 1.
   (b) the 84th and 16th percentile probability of exceedance, which are one standard deviation if a normal distribution is assumed.

This differs from traditional MMI regression techniques in which each amplitude is associated with just the single closest MMI recording. We have chosen to use a 2km radius because we never have an exact spatial mapping of the observed amplitude and observed MMI; using all observations allows us to account for the considerable variation in MMI observations that can occur within a small distance.

Finally, we have fitted a smooth curve to the points shown in Figure 1. We chose to fit a logistic regression to the points, as this function is ideally suited to fitting a response variable which is constrained to lie between two limits. It asymptotes to the maximum and minimum expected values of the response variable (e.g., MMI). For the case of MMI(Amp) it is fairly clear what the asymptotes should be: at the low end, MMI = 1 is the smallest possible value, while at the high end and using current definitions, MMI = 10 is the maximum possible integer value. Because values greater than X are no longer considered valid by the U.S. Geological Survey in the MMI scale (Dewey et al., 1995), we choose an asymptote of 10.5, due to the convention of rounding of MMI. We suggest that values greater than 10 be rounded down to 10. In the future, if MMIs of 10 are reported, we may adjust our fitting to accommodate the new data. The logistic equation for MMI(Amp) is expressed as follows:

$$MMI = \frac{MMI_{min} + MMI_{max} \times 10^{a+b\times AMP}}{1 + 10^{a+b\times AMP}}$$

where MMI_{min} is the minimum MMI possible, MMI_{max} is the maximum MMI possible, AMP is the log amplitude value and a and b are the fitted constants. The function was optimized by the least squares method using a logarithmically transformed MMI, so that the objective function minimized was $\sum\left(\log\frac{MMI_{obs}-1}{10.5-MMI_{obs}} - \log\frac{MMI_{pred}-1}{10.5-MMI_{pred}}\right)^2$.

The parameters for the logistic regression fitted to the median curve are shown in Table 1. Three goodness of fit statistics are shown, σ_{mean} which is the fit to the median curve as explained previously, σ_{tot} which is the RMS error calculated against all felt report and MMI data, and $\sigma_{>3}$ which is the RMS error calculated against all data ≥ 3 cm/s or % g. The statistic $\sigma_{>3}$ is most informative because it shows us how well the model performs against the real data in the range that we are interested in. PSA(.3s) and PSA(3s) are not shown as the data quality and fit were very poor. From this table it is clear that we are best able to fit the raw data in the range of interest using the PGV amplitudes. Interestingly PGA performs slightly better when compared against all data; however, it is important to

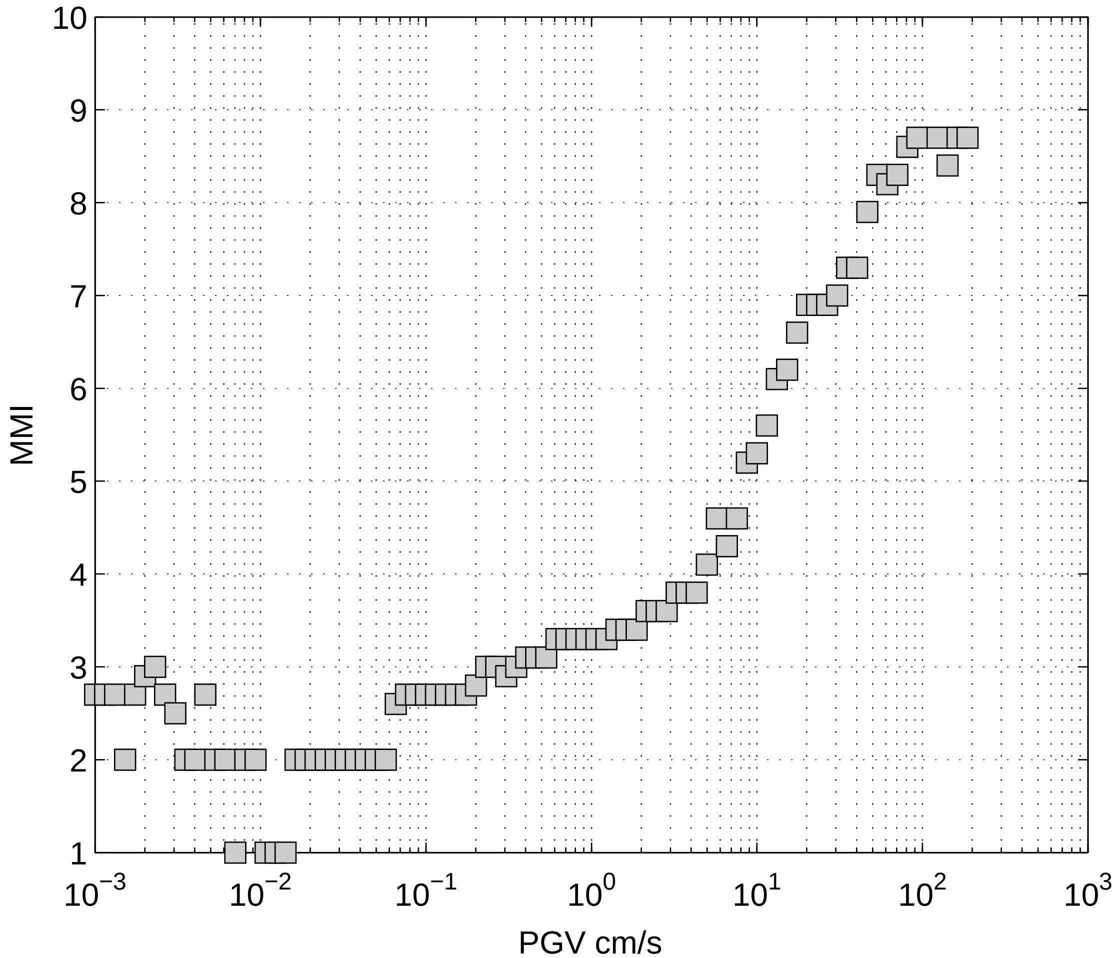

Figure 1: The median probability of MMI, given an observed PGV

	a	b	σ_{mean}	σ_{tot}	$\sigma_{>3}$
PGA	-0.98	0.75	.010	1.24	1.26
PGV	-0.80	0.68	.0014	1.28	1.04
PSA(1s)	-0.81	0.70	.0055	1.37	1.09

Table 1: Estimated parameter values for median MMI(Amp) and misfits. $\sigma_{>3}$ is calculated against all data ≥ 3 cm/s or % g and is the most informative; σ_{tot} and σ_{mean} are the misfits to all data and to the median curve, respectively.

remember that the data is incomplete for lower MMI so this result can be misleading. If all smaller MMI were recorded, it is possible that the PGA fit would be poorer; we cannot know. The fit to the PGV data is shown in Figure 2. The fit is compared to the Wald et al. (1999) relationship. It can be seen that for larger amplitudes the fit is very similar to the Wald relationship which was based on a much more limited data set. The primary differences are slightly lower predicted MMI for amplitudes up to approximately 100 cm/s, but much lower predictions for larger amplitudes.

The reverse fit, PGV(MMI), was done using the same methodology with only the following changes:

1. find all amplitudes within 2km of each MMI
2. apply two corrections to minimize the differences due to the observed MMI and the amplitude not being co-located
   (a) a distance correction
   (b) a site response correction

The correction is a scalar based on the difference in source-to-site distance of the recorded ground motion and the MMI. The scalar is calculated from the ratio of the predicted amplitude at the ground motion location to the predicted amplitude at the MMI location. The predicted amplitudes were calculated using the attenuation models of Joyner and Boore (1988) for PGV and Boore et al. (1997) for PGA and PSA (but because the scaling factor is the ratio of predicted ground motions at relatively similar distances, we do not expect the choice of attenuation model to have a significant effect on the results). Two sets of parameters were used in the Boore et al. (1997) attenuation relation depending on earthquake magnitude: the original set for for M>5.5 and another set determined for earthquakes M<5 by Wald et al. (2005), with the amplitude values for 5<M<5.5 calculated by a linear combination of the two.

After applying the distance correction we scaled the amplitude to the site conditions at the MMI location. To do this we multiply the distance-corrected amplitude, AMP_r, by the ratio of the site amplification factor , SAF_{mmi}, for the site class at the MMI location to the site amplification factor, SAF_{gm},for the site class at the recorded amplitude location:

$$AMP_{sc} = AMP_r(SAF_{mmi}/SAF_{gm}).$$

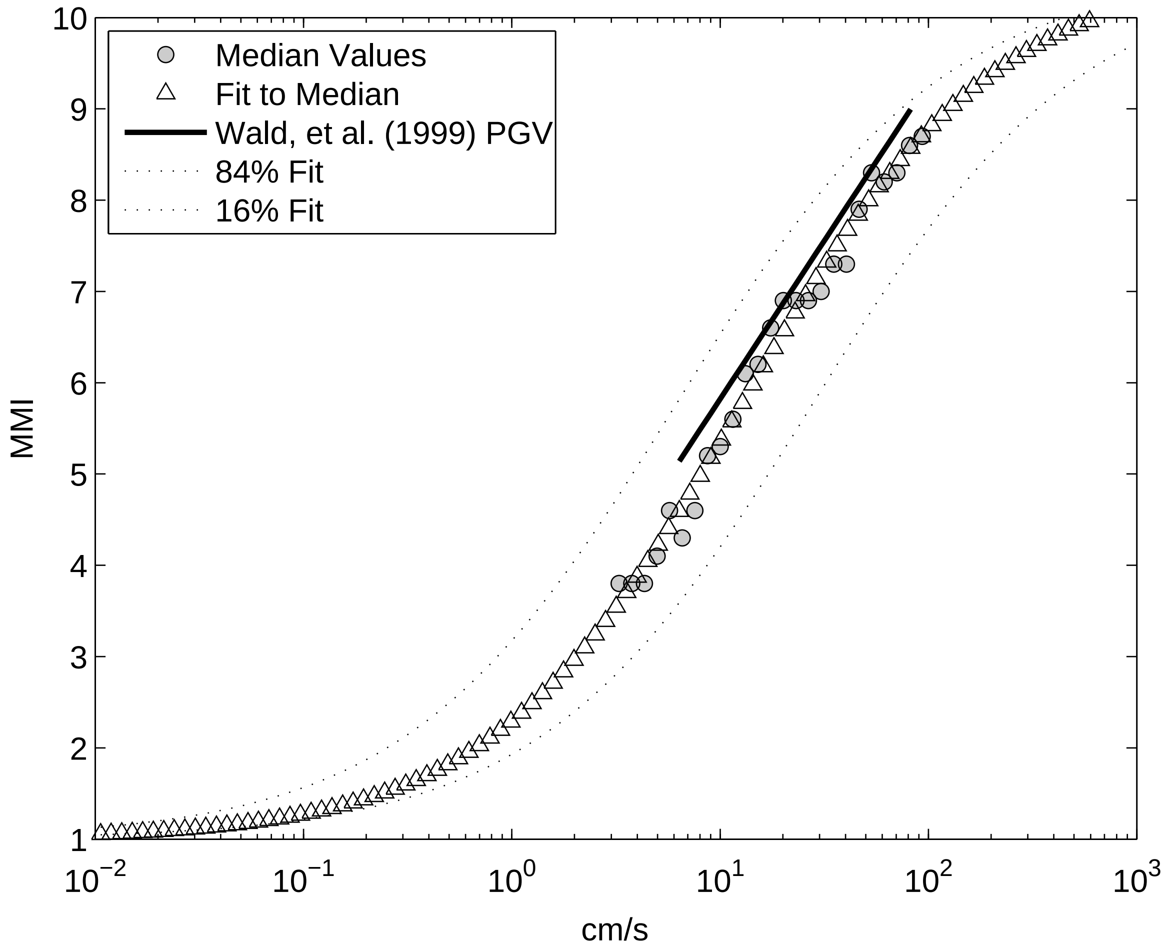

Figure 2: MMI(Amp). The curve for calculating the median MMI given PGV. The circles show the median MMI taken directly from the summed data; the triangles show the logistic regression fitted to the median data. Also shown is the Wald et al. (1999) regression; only the PGV regression of Wald is shown whereas their relationship would use PGA data for the lower portion of the line. We also show the fit to the 16th and 84th percentiles.

	a	b	AMP_{min}	AMP_{max}	σ_{mean}	σ_{tot}
PGA	-1.17	0.26	-1.15	1.90	0.0040	0.57
PGV	-1.38	0.27	-1.25	1.97	0.0032	0.52
PSA(.3s)	-0.67	0.17	-1.67	2.59	0.0036	0.63
PSA(1s)	-1.49	0.27	-1.23	2.2	0.0025	0.57
PSA(3s)	-1.42	0.29	-2.57	1.35	0.0053	0.81

Table 2: Estimated parameter values for median Amp(MMI) and misfits.σ_{tot} and σ_{mean} are the misfits to all data and to the median curve, respectively.

National Earthquake Hazards Reduction Program (NEHRP) site classifications were extracted for both the amplitude locations and the MMI locations from the Statewide Site Conditions Map for California (Wills et al., 2000) and the Borcherdt (1994) site amplification factors were applied. Once the site conditions and distance correction are applied, there remain amplitude differences between nearby stations (likely the result of differences in geometry and wave propagation). To account for these variations, as with the first relationship, we have used a probabilistic approach, assigning a probability density function (PDF) to the amplitude at the MMI location. The median of the amplitude PDF is the corrected amplitude, as described above, and the standard deviation ($\sigma = .31$) comes from the differences in the logarithm of the inter-station amplitudes and is derived from a data set consisting of all observed amplitudes, grouped by earthquake and with inter-station offsets of less than 2km. This requires the assumption that the data are log-normally distributed; however, after a thorough investigation of the data set, this assumption appears to be sound and is in agreement with literature (Boore et al., 1997; Field and Hough, 1997).

The fitting of the logistic regression was done in the reverse direction as for the MMI(PGV) relationship. Additionally, for this regression we allowed the minimum and maximum asymptotes to be fit. The fitted parameters for all ground motion measures are shown in Table 2 and the final fitted curves are shown in Figure 3.

An extensive review of the results has been submitted and is currently in review with the Bulletin of the Seismological Society of America: Probabilistic Relationships Between Peak Ground Motion and Modified Mercalli Intensity, Gerstenberger, M.C., Worden, C.B., Rhoades, D.A., and Wald, D.J.

Bibliography

Gerstenberger, M.C., Worden, C.B., Rhoades, D.A., and Wald, D.J, Probabilistic Relationships Between Peak Ground Motion and Modified Mercalli Intensity, BSSA submitted, 2008.

Availability of Data

All computer code developed for STEP will be open source and available to anyone interested. The JAVA code will be available via the OpenSHA project. All ground motion data and geocoded data used in the MMI study is also freely available. All results of testing of the

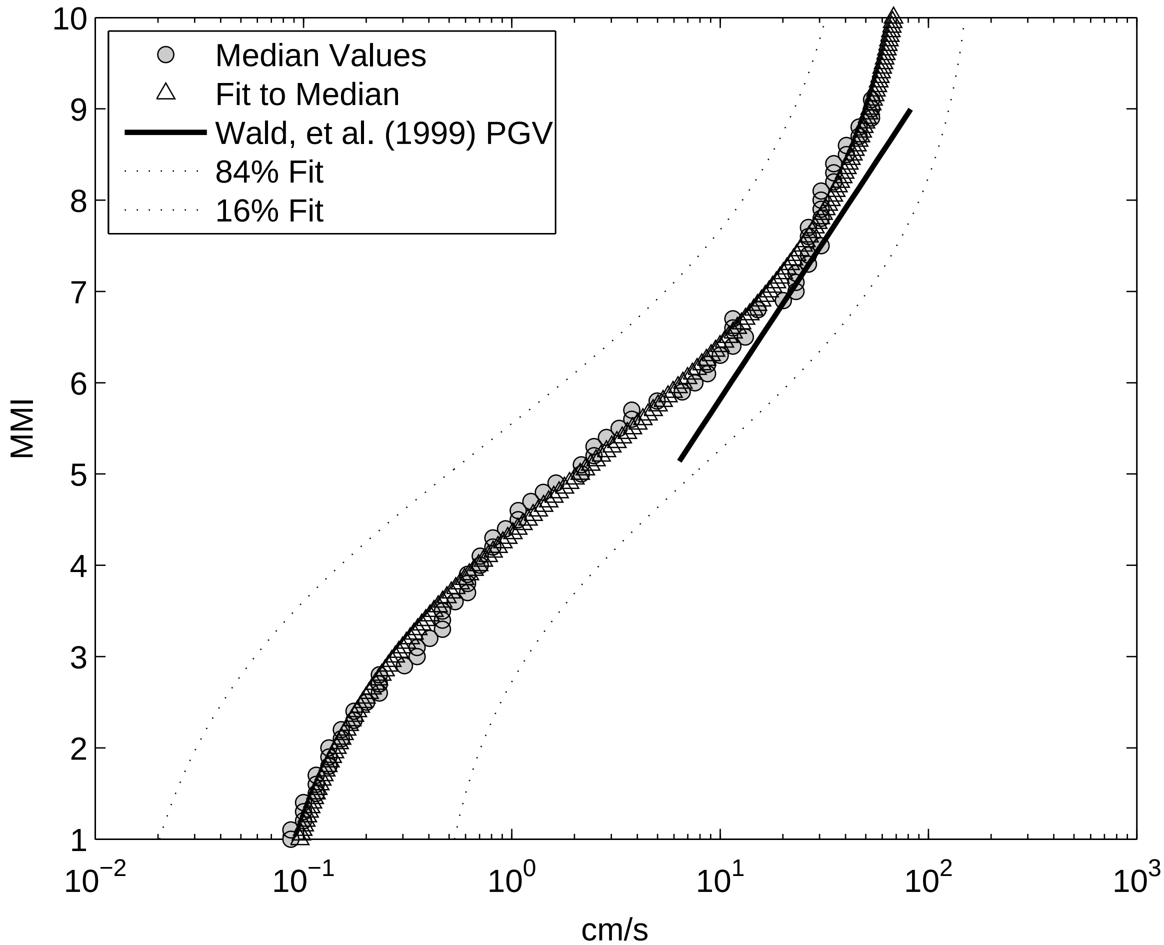

Figure 3: Amp(MMI). The curve for calculating the median PGV given MMI. The circles show the median PGV taken directly from the summed data; the triangles show the logistic regression fitted to the data. Also shown is the Wald et al. (1999) regression; only the PGV regression of Wald is shown whereas their relationship would use PGA data for the lower portion of the line. We also show the fit to the 16th and and 84th percentiles.

STEP model can be obtained by contacting the CSEP project at `http://us.cseptesting.org/`. For details on how to obtain any code or data please contact: Matt Gerstenberger, ++64-4-570-4554, m.gerstenberger@gns.cri.nz.

References

Boore, D. M., Joyner, W. B., and Fumal, T. E. (1997). Equations for estimating horizontal response spectra and peak acceleration from western north america earthquakes: A summary of recent work. *Seismological Research Letters*, 68(1):128–153.

Borcherdt, R. D. (1994). Estimates of site-dependent response spectra for design (methodology and justification). *Earthquake Spectra*, 10(4):617–654.

Dewey, J., Reagor, B., Dengler, L., and Moley, K. (1995). Intensity distribution and isoseismal maps for the northridge, california, earthquake of january 17, 1994. *U.S. Geological Survey Open-File Report 95-92*, page 35.

Field, E. and Hough, S. (1997). The variability of psv response spectra across a dense array deployed during the northridge aftershock sequence. *Earthq. Spectra*, 13(2):243–257.

Joyner, W. B. and Boore, D. M. (1988). Measurement, characterization, and prediction of strong ground motion. *Proceedings of Earthquake Engineering and Soil Dynamics, II, Park City, Utah*.

Reasenberg, P. A. (1985). Second-order moment of central california seismicity. *Journal of Geophysical Research*, 90:5479–5495.

Schorlemmer, D. and Gerstenberger, M. (2007). RELM Testing Center. *Seismol. Res. Letts.*, 78(1):30–36.

Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D. D., and Rhoades, D. A. (2007). Earthquake likelihood model testing. *Seismol. Res. Letts.*, 78(1):17–29.

Wald, D., Worden, B., Quitoriano, V., and Pankow, K. L. (2005). Shakemap manual: Technical manual, user's guide, and software guide:. U.S. Geological Survey Techniques and Methods 12A01. http://pubs.usgs.gov/tm/2005/12A01/.

Wald, D. J., Quitoriano, V., Heaton, T., and Kanamori, H. (1999). Relationships between peak ground acceleration,peak ground velocity, and modified mercalli intensity in california. *Earthq. Spectra*, 15(3):680–697.

Wills, C. J., Petersen, M. D., Bryant, W. A., Reichle, M. S., Saucedo, G. J., Tan, S. S., Taylor, G. C., and Treiman, J. A. (2000). A site-conditions map for california based on geology and shear wave velocity. *Bulletin of the Seismological. Society of. America*, 90:S187–S208.