Bayesian Treatment of Induced Seismicity in Probabilistic Seismic-Hazard Analysis

by Jack W. Baker and Abhineet Gupta

Abstract This article proposes a Bayesian treatment of uncertainty in probabilistic seismic-hazard assessment to account for the possibility of anthropogenic changes in the properties of future earthquake occurrences. Operational factors are used to establish prior distributions for future seismicity, and observations are used to dynamically update the distributions and resulting hazard calculations. The updating process is relevant for regions of the central and eastern United States where large increases in the rate of earthquakes, apparently triggered by anthropogenic activities, are difficult to address using conventional quantitative methods for assessing seismic hazards. The approach is consistent with conventional seismic-hazard analysis initially, allows a number of types of data to be used for updating in an informative and transparent manner, and may incentivize improved monitoring of at-risk regions. Examples are provided to illustrate the approach, including the use of Gibbs sampling to characterize updated distributions in the case for which analytical solutions to the updating equations are not available.

Introduction

Large increases in the rate of earthquakes in a number of areas, at levels inconsistent with historical seismicity, have raised concerns about the impact of fluid injection on earthquake hazards (Committee on Induced Seismicity Potential in Energy Technologies, 2012; Ellsworth, 2013; McGarr *et al.*, 2015). Probabilistic seismic-hazard analysis (PSHA) is a widely used tool for assessing and managing the impacts of natural earthquakes (e.g., McGuire, 2004; Petersen *et al.*, 2014). PSHA generally assumes that past earthquake occurrences and ground motions can be used to infer future seismic hazard but lacks mechanisms to dynamically revise predictions if evidence suggests that induced seismicity has changed the number of earthquakes likely to occur in the near future. Although PSHA has been advocated for management of induced seismicity (Mignan *et al.*, 2015), it is with an admonition that "as more information becomes available (particularly seismic monitoring results), the hazard can be re-calculated and uncertainties reduced" (Majer *et al.* 2012).

This article proposes a PSHA process that allows dynamic updating of results, rather than periodic recalculation of hazard. Uncertainty in future earthquake rates, or other inputs to the hazard assessment, is treated as prior distributions at the start of analysis, and likelihood functions are formulated to quantify the consistency of future observed data with the uncertain parameters. As time passes and new earthquakes are observed, the prior distributions can then be dynamically revised with those likelihoods, using Bayesian updating.

Because of its ability to integrate prior knowledge with observational data, Bayesian analysis of seismic-hazard parameters has been proposed since the earliest days of PSHA (Esteva, 1969), and many earthquake hazard and risk assessments have utilized Bayesian updating and related Bayesian networks (e.g., Bayraktarli *et al.*, 2005, among many). The general updating concept is thus straightforward and widely understood. The contribution of this article is to develop and utilize likelihood functions and updating procedures that are appropriate for dynamic updating, using induced-seismicity data, and that are compatible with modern PSHA procedures. Examples are provided to illustrate the utilization of various types of data and to demonstrate the impact of this updating on ground-motion hazard.

When applying PSHA to induced seismicity, one approach to deal with this challenge is to estimate seismicity rates in narrow windows of time, in order to obtain relatively current estimates of seismicity rates (Convertito *et al.*, 2012; Atkinson *et al.*, 2015; Mignan *et al.*, 2015). Statistical seismicity models can produce short-term dynamic estimates of earthquake activity at well-instrumented sites (e.g., Ogata, 1988; Werner *et al.*, 2011), but they are not clearly relevant in the central United States, where the locations of potential seismicity are not well characterized, where the region is sparsely instrumented, and where it is difficult to locate or even detect small ($M < 3$) earthquakes that might inform occurrence rates of larger earthquakes (Ellsworth, 2013).

In a few cases, relationships have been developed between injection operations and seismicity associated with a specific injection site (e.g., Ake *et al.*, 2005; Brodsky and Lajoie, 2013). There are several models that propose to more

generally relate volumes of injected fluids to the number (Shapiro *et al.*, 2007, 2010) or maximum size (McGarr, 2014) of potentially induced earthquakes. These models are uncertain and difficult to calibrate for new sites, however, and thus at present are not directly adaptable in a PSHA context. Incorporating knowledge about these anthropogenic "operational factors" is a critical need if PSHA is to be used as a tool for risk screening or regulation related to wastewater injection (Walters *et al.*, 2015). The latest U.S. Geological Survey seismic-hazard results for induced seismicity note this limitation but do not yet resolve it (Petersen *et al.*, 2015).

This article proposes to address the above challenges, using a prior distribution for potential future changes in seismicity parameters, based on operational factors. If the prior distribution is diffuse (as it should be in many cases), it need not imply greatly increased seismic hazard. It does, however, facilitate rapid updates in hazard results if data are collected that indicate changed seismicity conditions. With this approach, new data are used to (probabilistically) validate or eliminate the possibility of dangerous conditions, rather than being used solely to estimate future seismicity rates based on recent seismicity. The article begins by describing basic components of traditional PSHA, because many of these components are adopted in the proposed procedure. The traditional approach is then revised to adopt a Bayesian perspective that facilitates the use of additional observational data. Finally, three example calculations are presented to illustrate several ways in which this revised approach can be adopted to utilize various prior distributions and various types of observational data likely to be available in the induced-seismicity context. The first example utilizes an analytical solution to provide a transparent illustration of updating; the second example uses a more realistic formulation of earthquake-rate updating with a numerical solution for the hazard; and the final example utilizes Gibbs sampling to illustrate the feasibility of the approach when analytical or numerical solutions are not available.

Traditional Probabilistic Seismic-Hazard Analysis

The PSHA calculation procedure for natural seismicity is well documented elsewhere (e.g., Cornell, 1968; McGuire, 2004), but is briefly discussed here to facilitate later discussion. The annual rate, λ, that some ground-motion intensity measure [IM] exceeds a given amplitude x at a site is computed as follows:

$$\lambda(\text{IM} > x) = \sum_{i=1}^{n_{\text{sources}}} \lambda(M_i > m_{\min})$$
$$\times \int_{m_{\min}}^{m_{\max}} \int_0^{r_{\max}} P(\text{IM} > x|m,r) f_{M_i,R_i}(m,r) dr dm, \quad (1)$$

in which n_{sources} is the number of earthquake sources considered, and M_i and R_i denote the magnitude and distance distributions for source i. The term $f_{M_i,R_i}(m,r)$ is the probability density function for magnitude and distance of an earthquake on source i, and $\lambda(M_i > m_{\min})$ is the annual rate
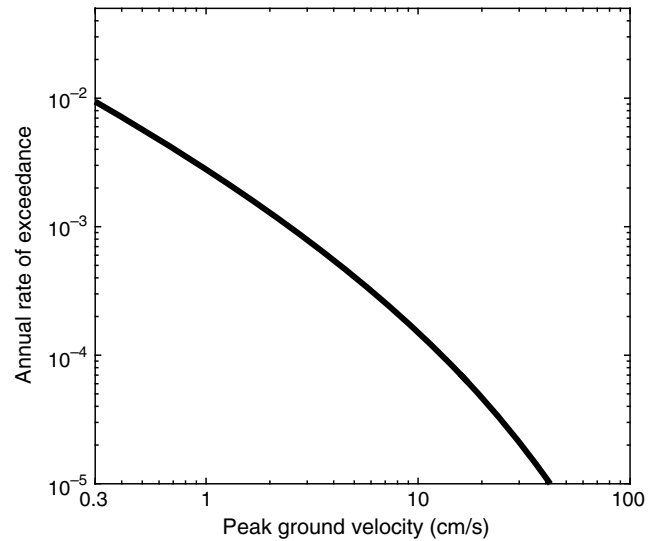

Figure 1. Annual rates of exceeding various levels of peak ground velocity for the example site.

of occurrence of earthquakes on source i with magnitude greater than $m_{\min}$, the minimum magnitude of interest. The $\lambda(M_i > m_{\min})$ and $f_{M_i,R_i}(m,r)$ terms reflect the source characterization portion of the calculation. The probability that an earthquake of magnitude m at distance r could produce a ground motion with intensity greater than x, $P(\text{IM} > x|m,r)$, is the ground-motion characterization portion of the hazard analysis. In general, the source characterization and ground-motion prediction utilize additional variables beyond magnitude and distance (e.g., rupture mechanism and site conditions), but the notation is simplified here for ease of illustration, without affecting the general procedure.

Figure 1 shows results from an example calculation of this type, considering peak ground velocity as the IM of interest. For this calculation, the source characterization is as follows: earthquake locations are uniformly located within a 25 km radius of the site at a depth of 3 km; earthquake magnitudes have a truncated Gutenberg–Richter distribution with $M_{\min} = 3$, $M_{\max} = 6.5$, and $\lambda(M > 3) = 0.1$ per year. Ground-motion predictions are from the model of Atkinson (2015), with no truncation of the ground-motion distribution. A plot like Figure 1, showing the rates of exceeding various levels of ground-motion intensity, is called a ground-motion hazard curve, or simply hazard curve.

The inputs to equation (1) are models and are subject to lack of knowledge (epistemic) uncertainty. In PSHA this epistemic uncertainty is characterized by a logic tree, with each branch of the logic tree corresponding to a plausible set of models and having a resulting ground-motion hazard curve (Kulkarni *et al.*, 1984; Senior Seismic Hazard Analysis Committee [SSHAC], 1997; Mignan *et al.*, 2015). To illustrate, we extend the previous example to treat several input parameters as uncertain. Let us consider the annual rate of earthquakes to be 0.05, 0.1, or 0.2, the maximum possible magnitude to be 6, 6.5, or 7, and three models for predicting
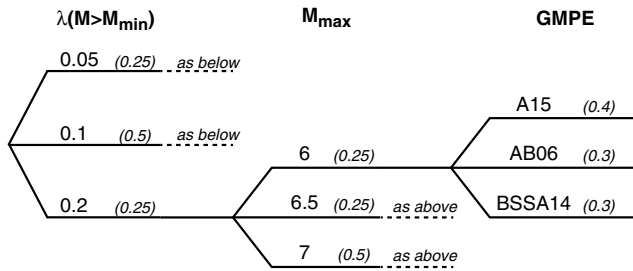
Figure 2. Example probabilistic seismic-hazard analysis logic tree. Parameter values are indicated at the left of each alternative sub-branch, followed by the corresponding weight in italics. Ground-motion prediction equations (GMPEs) are abbreviated as follows: A15, Atkinson (2015); AB06, Atkinson and Boore (2006); BSSA14, Boore et al. (2014).

the ground-motion amplitude for a given earthquake. Figure 2 illustrates a logic tree associated with these alternatives. Each set of alternatives has a corresponding set of weights that sum to 1, and the weight associated with a complete set of inputs is equal to the product of the associated weights (e.g., the weight associated with $\lambda(M_i > m_{\min}) = 0.2$, $M_{\max} = 6.5$, and ground-motion prediction equation [GMPE] = A15 is $0.25 \times 0.25 \times 0.4 = 0.025$). Each complete set of inputs can also be used in equation (1) to compute the associated hazard. These hazard calculations and associated weights, taken collectively over all possible input values, represent a seismic-hazard calculation with epistemic uncertainty considered.

These results are often illustrated by plotting a representative set of hazard curves, fractiles of the exceedance rates, and the mean exceedance rate. Mathematically, we compute the mean hazard as

$$\lambda_{\mathrm{mean}}(\mathrm{IM} > x) = \sum_{j=1}^{n_{\mathrm{branches}}} w_j \lambda_j(\mathrm{IM} > x), \qquad (2)$$

in which j is an index indicating the jth logic-tree branch, w_j is the weight associated with branch j (the w_j's sum to unity), and $\lambda_j(\mathrm{IM} > x)$ is the hazard curve, computed using equation (1), for the source characterization and ground-motion prediction terms from logic-tree branch j. Figure 3 shows the hazard curves and mean hazard for the example described by Figure 2.

The seismic-hazard calculations discussed in this section can be combined with information about exposure (of people, buildings, and other assets) and vulnerability (susceptibility of exposed assets to adverse outcomes under earthquake shaking) to quantify seismic risk (Bommer et al., 2015). Although risk is often the ultimate concern in this application, these later steps are omitted from discussion here in the interest of focusing on the hazard calculation approach.

Bayesian Treatment of Potential Future Seismicity

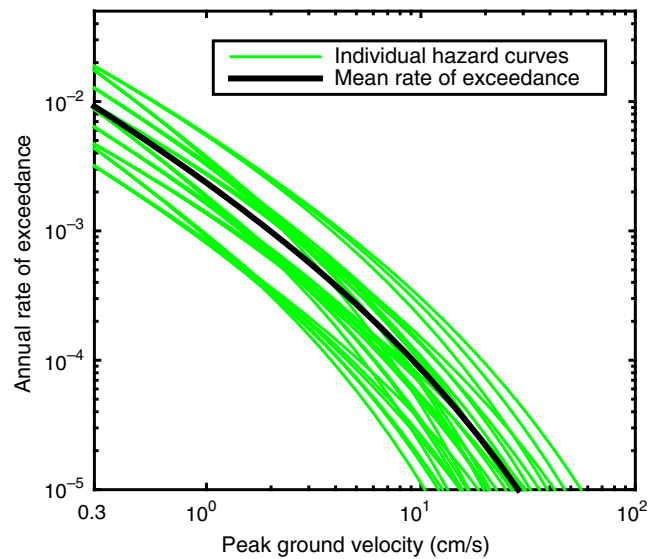To adopt the above PSHA approach to address induced seismicity, we consider how new data would influence the

Figure 3. Annual rates of exceeding various levels of peak ground velocity, considering epistemic uncertainty, for the example site. The color version of this figure is available only in the electronic edition.

weight (w_j in equation 2) associated with our belief that model j represents reality. In the natural seismicity case, new data and models are developed at a relatively steady rate, and typical practice is to repeat hazard calculations on a time-span of years and develop new logic-tree branches and weights (Petersen et al., 2008, 2014). In the case of induced seismicity, where data may provide rapid changes in our estimates of future earthquake activity and shaking at a site, we can treat these models in a Bayesian manner with weights that are dynamically updated. We will treat the initial set of weights as a prior distribution on the models and update the weights using relevant data.

If seismic activity differs from an assumed model, it can be taken as evidence that the model is inconsistent with the current state of nature (Llenos and Michael, 2013) and used to justify reduced consideration of the model in PSHA calculations. That same data, however, will be more consistent with some other model for the state of nature. As long as both models are in the initial logic tree, the data can be used in a likelihood function to update model weights.

If we denote the probability of observing the data, given logic-tree branch j, as L_j, then these values can be used to obtain updated logic-tree branch weights using Bayes rule:

$$w_j' = \frac{L_j w_j}{\sum\limits_{\mathrm{all}_j} L_j w_j}. \qquad (3)$$

The updated weights w_j' describe a posterior distribution for our belief in the candidate models. This standard formula has several intuitive implications. Branches that produce a relatively higher likelihood of observing the data (L_j) receive increased weight ($w_j' > w_j$), whereas branches that are less
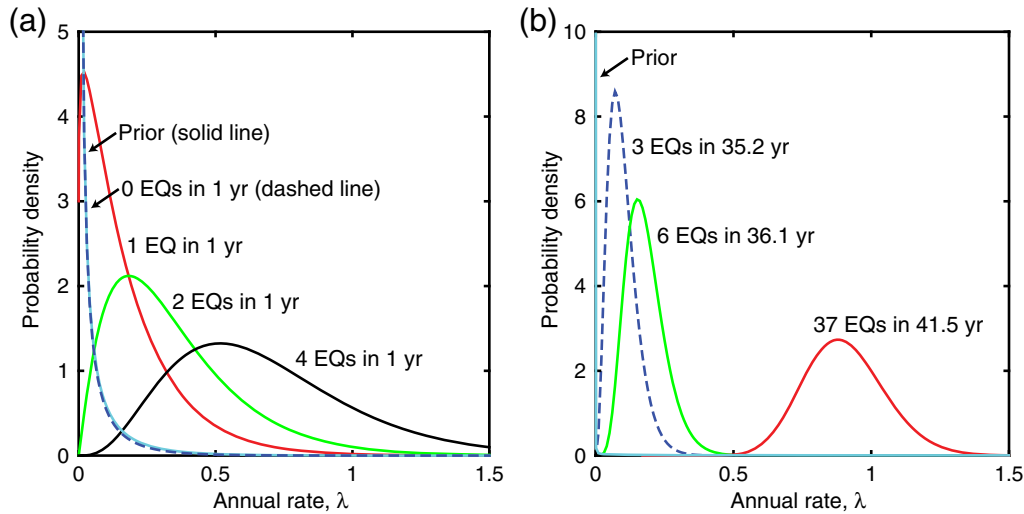
Figure 4. (a) Gamma prior distribution for earthquake rates and posterior distributions, given several possible observation scenarios over a one-year period. (b) Gamma prior and posterior distributions given possible observation scenarios over a multiyear period (i.e., observations corresponding to the data in example 2). The color version of this figure is available only in the electronic edition.

consistent with the data will receive decreased weight. The original logic-tree weights also play a role, however, in quantifying the plausibility of the logic-tree branch prior to the observation of data (as indicated by the presence of w_j in the right-side numerator). The denominator simply normalizes the updated weights to sum to unity. These new weights can then be used to update the seismic-hazard calculation.

Examples

The mathematics of the updating procedure are simple, but complexity can arise when computing the likelihood, L_j, associated with a given data type and logic-tree branch. To illustrate implementation of this procedure, three example calculations are considered here.

Example 1: Potential Change in Seismicity Rate at a Known Point in Time

We first consider a simple case where we have an estimated distribution for the seismicity rate prior to the start of an injection operation and want to update this distribution based on observed seismicity after injection begins. We denote the annual rate of earthquakes from a given source and above some minimum magnitude as λ, simplifying the $\lambda_j(M_i > m_{\min})$ notation from equation (2) for brevity. We will count observed earthquakes satisfying these criteria and use these observations to update our estimate of λ. We assume that earthquakes occur as a Poisson process, that a rate change (if one occurs) will occur immediately after the start of injection, and that the postinjection rate of earthquakes will remain constant. Although these conditions are idealized, they provide analytical results that complement the more general example below.

Under these conditions, the probability of observing n earthquakes in time t after injection begins, given λ, is

$$L(\lambda) = P(n \text{ earthquakes in time } t) \propto \lambda^n e^{-\lambda t}, \quad (4)$$

in which $L(\lambda)$ denotes a likelihood function for λ, conditional on our observed data, and ∞ indicates proportionality (i.e., equality to within a constant that is not needed for later calculations). This serves as the L_j term in equation (3). For this example, we consider a gamma distribution to reflect our prior probabilities for λ:

$$f(\lambda) \propto \lambda^{a-1} e^{-\lambda/b}, \quad (5)$$

in which $f()$ denotes a probability density function, and $a > 0$ and $b > 0$ are parameters controlling the shape of the distribution. The choice of a gamma distribution is convenient here, as it is the conjugate prior of the Poisson-likelihood function, and thus this pair has an analytical solution for the posterior distribution of λ. If λ has a gamma prior distribution with parameters a and b, and we observe n earthquakes in time t, then the posterior distribution for λ also has a gamma distribution but with parameters $a' = a + n$ and $b' = b/(tb + 1)$ (Gelman *et al.*, 2014).

Figure 4 shows a prior gamma distribution having $a = 0.1$ and $b = 0.2$, which gives a mean estimated rate of earthquakes of 0.02 per year but allows for a possibility of a wide variation in rate away from that mean. Figure 4 shows the posterior distributions for several potential cases of postinjection observations. In Figure 4a, we see that when the number of observed earthquakes is larger than the prior estimated mean rate of 0.02 per year, the posterior distribution shifts to the right, as the data suggest a greater rate of earthquakes. For the case of four observed earthquakes in one year, the posterior distribution changes significantly and even indicates a non-negligible probability that the rate of earthquakes is greater than one per year, although the prior distribution indicated that such a rate would be very unlikely.

If zero earthquakes are observed in one year, then the posterior distribution is essentially identical to the prior, as zero earthquakes are expected given the low prior rate, and so the data are not very informative.

In Figure 4b, we consider $a = 0.5$ and $b \to \infty$ and consider some specific observation cases that will be relevant in the following example. This prior distribution gives a very high probability of a rate close to zero (the probability density is on top of the y axis in Fig. 4b) but also a very heavy tail that allows for the possibility of high rates; in fact it has an infinite variance and so does not strongly influence the posterior distribution. Even with a few observations of earthquakes, the posterior distribution differs dramatically from the prior, indicating that the prior is not providing a strong constraint on the posterior distribution.

Example 2: Change in Seismicity Rate at an Unknown Point in Time

The above calculation can be generalized to consider the time of rate change to be unknown, which is useful, as in many cases an increase in seismicity need not occur immediately after injection commences, and there may not be a single causal injection well (Ellsworth, 2013). Assume that earthquakes occur as a Poisson process with rate λ_1 up to time τ and after that occur as a Poisson process with rate λ_2. All three of these parameters are unknown, and we assign λ_1 and λ_2 prior gamma distributions with $a = 0.5$ and $b = \infty$ and τ a prior distribution that is uniform over the duration of the considered catalog. For a given λ_1, λ_2, and τ, a likelihood function can be computed using a generalization of equation (4) that considers the differing λ before and after τ:

$$L(\lambda_1, \lambda_2, \tau) \propto \lambda_1^{n(\tau)} e^{-\lambda_1 \tau} \lambda_2^{n(t_{end}) - n(\tau)} e^{-\lambda_2 (t_{end} - \tau)} \quad (6)$$

(Raftery and Akman, 1986; Gupta and Baker, 2015), $n(t)$ is the number of events between the catalog start and time t, and t_{end} is the time at the end of the considered catalog. The posterior distribution in this case is computed using numerical integration (Gupta and Baker, 2015). The change-point model includes a model selection step in which we evaluate whether or not a constant-rate or change-point model is appropriate. For each set of considered data, we compute a Bayes factor, which is equal to the constant rate likelihood from equation (4) divided by the change point likelihood from equation (6), integrated over the posterior distributions (Raftery and Akman, 1986). When the Bayes factor is less than 0.01, meaning that data are more consistent with the change-point model, we assume a change has occurred and use the posterior distribution for λ_2; otherwise, we conclude that there is no change and use the posterior distribution for λ from example 1.

To illustrate, we consider earthquake observations near Oklahoma City (latitude 35.48°, longitude −97.54°) from 1974 to May 2015. Figure 5a shows the cumulative number of $M \geq 3$ earthquakes observed within a 25 km radius of the reference location at a given point in time. The catalog has been declustered to remove aftershocks using the method of Reasenberg (1985) with the published parameters and a magnitude completeness of 3; this was done so that the occurrences would be Poissonian in the absence of a rate change. With the benefit of viewing the full catalog, one can clearly see a large increase in earthquake rate after 2009. But ideally this change would be detectable shortly after the increase in earthquake activity.

To demonstrate the Bayesian updating approach, we consider three points in time, as noted in Figure 5a. Time A is February 2009, when three earthquakes have been observed over the prior 35 years. Time B is January 2010, when an additional three earthquakes have been observed in the 11 months since time A. Time C is May 2015, at the end of the considered catalog, when a total of 37 earthquakes have been observed.

At each point in time, only the catalog up to that point is considered for estimation. Likelihood values and posterior distributions for the parameters are computed, and here we are interested in the current rate of earthquakes. Figure 5b shows the posterior distributions of earthquake rates at the three considered points in time, with the mean values of the distributions also noted. Figure 5c shows the mean values of the posterior rate distributions considering all possible end dates for the catalog, with the three specific mean values noted by circles. For reference, a frequentist rate is also shown, which is obtained by simply dividing the number of earthquakes to date by the number of years since the start of the catalog. Selected numerical values from both methods are given in Table 1.

A few observations can be made from Figure 5 and Table 1. At time A, when few earthquakes have been observed, the Bayes factor of 0.045 indicates that the no-change model is preferable (consistent with intuition), and the mean posterior rate is 0.10 earthquakes per year. This is similar to the pure frequentist estimate of (3 earthquakes)/(35.2 years) = 0.085 but somewhat higher, reflecting the effect of a prior distribution that allows for some probability of high earthquake rates. At time B, the rapid occurrence of three additional earthquakes reduces the Bayes factor to 0.0016, indicating a clear preference for the change-point model, and that model suggests a high probability that the earthquake rate is dramatically increased after the change. The Figure 5b posterior distribution shows some probability that the earthquake rate is still low (similar to the time A posterior values) but a greater probability that the rate has increased significantly. The mean posterior rate at time B is 4.85 earthquakes per year, versus the 0.17 frequentist rate. (Note that the mean values are shown in Figure 5b and are in visually nonintuitive locations, because the x axis is shown in logarithmic scale.) At time C, the many additional observed earthquakes cause the frequentist estimate to rise a further fivefold, to 0.89, whereas the Bayesian mean posterior rate is 5.58 earthquakes per year (and the Bayes factor has fallen to 10^{-23}). The Bayesian rate is consistent with the 24 earthquakes observed over the previous
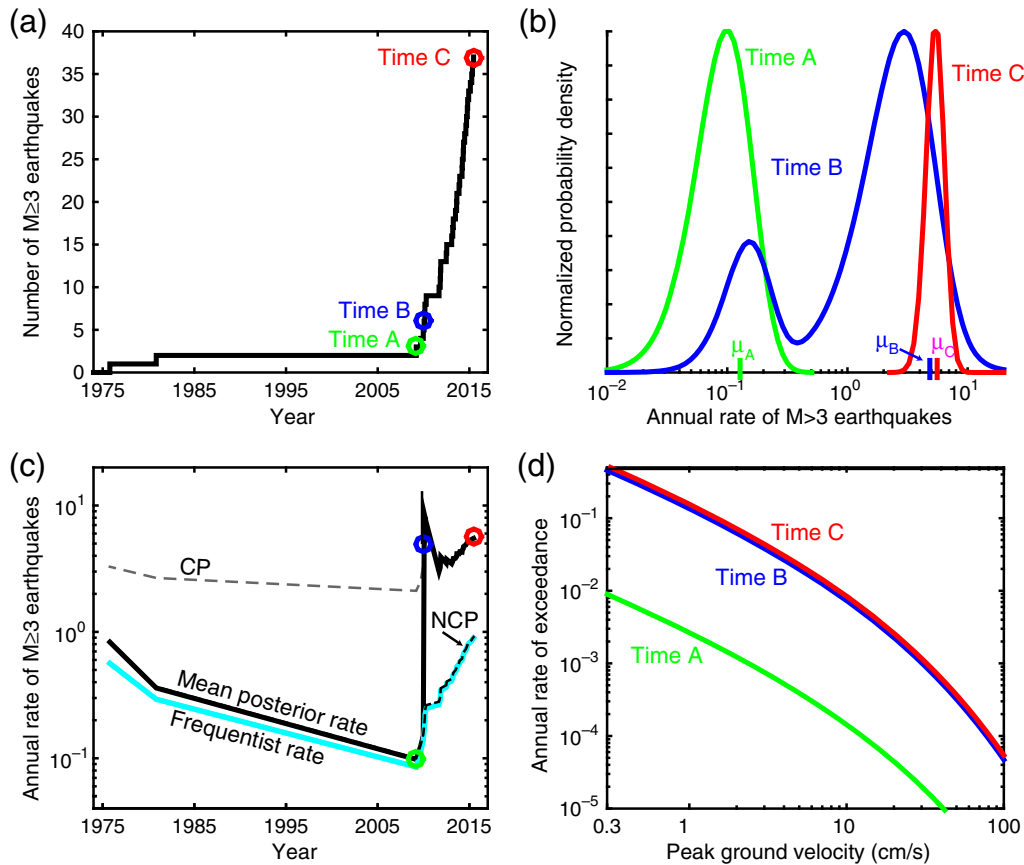
Figure 5. (a) Cumulative number of observed earthquakes within a 25 km radius of Oklahoma City. (b) Posterior earthquake rate distributions at three selected points in time. (c) Mean posterior earthquake rate and frequentist rate versus time. The dashed lines labeled CP and NCP are the mean posterior rates assuming a change point or no change point, respectively. (d) Mean hazard curves given the posterior earthquake rates at the three selected points in time. The color version of this figure is available only in the electronic edition.

four years. Figure 5c shows that the mean Bayesian rate jumps dramatically in 2009 (between time A and time B) and stays somewhat stable after time B, whereas the frequentist estimate fails to reflect the high rate of post-2009 earthquakes. Also shown in Figure 5c are the mean posterior rates assuming a change point (CP) or assuming no change point (NCP). We note that after 2009 the NCP model is effectively equal to the frequentist rate, indicating that the prior distribution assumption has little impact once there are a significant number of observed earthquakes (the mean rates at

times A, B, and C are 0.10, 0.18, and 0.90, respectively, corresponding to the means of the distributions shown in Figure 4b). We also note that the CP model gives unreasonably high estimated rates prior to 2009, indicating the need to perform the Bayes factor computation to evaluate whether a change-point model is appropriate.

Figure 5d shows the mean hazard curves for the site at the three considered times, using the same model assumptions as were used to produce Figure 1 but substituting the posterior earthquake rate distributions considered here. Because nothing

Table 1

Estimated Annual Rates of $M \geq 3$ Earthquakes at Four Points in Time, as Obtained from Four Different Methods

Date	Mean Posterior (baseline)	Mean Posterior (elevated prior)	Frequentist	Windowed Frequentist
February 2009 (time A)	0.10	0.19	0.09	0.04
January 2010 (time B)	4.85	1.19	0.17	1.33
February 2010	5.84	7.16	0.19	2.00
April 2015 (time C)	5.58	5.85	0.89	7.67

Dates corresponding to times A, B, and C from Figure 5 are noted. Mean posterior (baseline) and frequentist are the estimates from Figure 5. Mean posterior (elevated prior) and windowed frequentist are estimates from Figure 6.
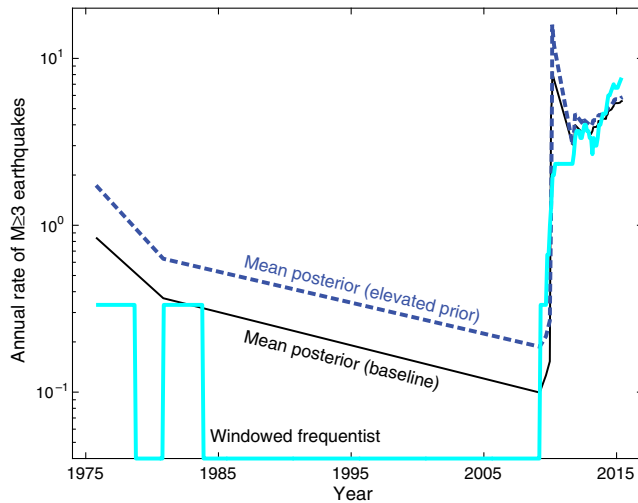
Figure 6. Mean estimated earthquake rates versus time for example 2, using two alternate models. The color version of this figure is available only in the electronic edition.

about the source except for the earthquake rates is being updated, the mean ground-motion exceedance rates are proportional to the mean earthquake rates. This example serves to illustrate how the approach can rapidly update the seismic hazard in a short window of time, without the need to manually change the catalog estimation procedure or to manually specify a date of seismicity rate change.

To examine the stability of the earthquake rate estimates in Figure 5c, Figure 6 shows two sets of alternate results along with the mean posterior rate from Figure 5c for reference. First, we consider a case where operational information (e.g., knowledge that injection volumes in the region are high and that nearby locations have experienced induced seismicity) leads us to believe there is an elevated potential for induced seismicity. We repeat the change-point calculation using a prior gamma distribution with $a = 2$. This factor-of-four increase relative to the value of a above leads to the mean of the distribution being increased by a corresponding factor. Results from this elevated prior case are labeled as mean posterior (elevated prior) in Table 1 and Figure 6. We see that at dates prior to 2009, the elevated prior case estimates rates of approximately twice that of the baseline case. The dramatic increase in rates is fully detected in February rather than January 2010 (see Table 1). After that increase, the elevated prior case continues to estimate higher rates than the baseline case, though the difference is quite small by 2015. This indicates that the prior distribution assumption does have an impact on the results, though the impact is reduced once a significant amount of observed data is available (as would be hoped).

A second alternative calculation shown here is a frequentist rate estimate that considers the average occurrence rate over the previous three years; this is labeled as windowed frequentist in Table 1 and Figure 6. It has the advantage relative to the frequentist estimate of Figure 5 that it can

better quantify the large increase in rates after 2010, but it is less stable than the Bayesian estimates over that time period. Further, it performs poorly prior to 2010, when few earthquakes are observed. When using a window narrower than three years, this windowed estimate becomes more unstable, and when using a wider window it tends toward the baseline frequentist estimate of Figure 5. These results are consistent with the intuition that one should use a windowed earthquake catalog when estimating induced-seismicity hazard using traditional methods (e.g., Petersen *et al.*, 2015), but the choice of window needs to be chosen manually, which is cumbersome for regional assessments and may not be possible until some time after a seismicity increase has occurred.

Example 3: Earthquake Locations Indicate Existence of a Fault Structure

In this example, we consider updating a parameter other than earthquake rate. Injection of wastewater near a preexisting fault (which is well oriented for failure in the ambient state of stress) is a risk factor for inducing earthquakes (e.g., Townend and Zoback, 2000; Walters *et al.*, 2015). Detecting and locating such a fault, if one exists, would thus provide an informative update to seismic hazard. In much of the central and eastern United States, we are limited both in our ability to locate critically stressed faults (other than by observing earthquakes), and limited by sparse instrumentation in detecting small earthquakes or precisely locating detected earthquakes. Here we consider an example where observed earthquake locations are used to update a model for the presence of a fault.

Assume that prior to monitoring we consider the existence of a fault source with unknown parameters. Assume a vertical fault with a strike uniformly distributed between 30° and 90° clockwise from north (an example range of orientations that could be activated under the regional stress state). The length of the fault is uniformly distributed between 3 and 13 km, and it is equally likely to be located anywhere in the 25 km radius region of interest. The vector of parameters associated with the unknown fault are $\theta = \{x, y, l, \alpha\}$, in which $\{x, y\}$ is the coordinate of the lower-left end of the fault, l is the fault length in units of kilometers, and α is the fault's strike angle clockwise relative to north. Under the above assumptions, the prior distribution for θ is

$$f(\theta) \propto 1 \quad \text{if } (x^2 + y^2) < 25, 30° < \alpha < 90°, \quad \text{and} \quad 3 \leq l < 12$$
$$= 0 \quad \text{otherwise} \tag{7}$$

We could take the presence or absence of a fault, and the fraction of on-fault versus background seismicity, to be additional unknown parameters and also update them based on data, but we omit these refinements here for simplicity of illustration.

Figure 7 shows two hypothetical earthquake catalogs that will be used to update potential fault locations. To do
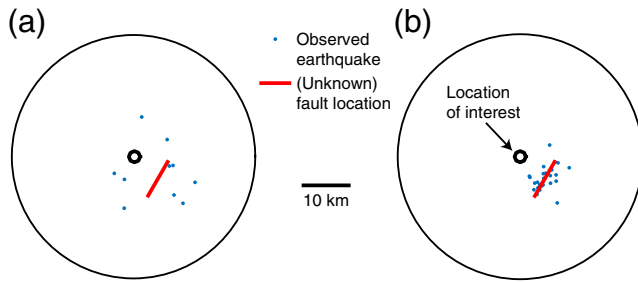
Figure 7. Examples of observed earthquake catalogs. (a) Case 1: 10 observed earthquakes, located with 4 km uncertainty. (b) Case 2: 30 observed earthquakes, located with 2 km uncertainty. The color version of this figure is available only in the electronic edition.
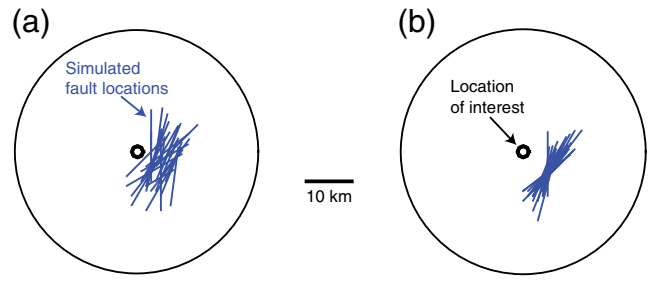

Figure 8. Fifty posterior distribution samples from (a) case 1 and (b) case 2. The color version of this figure is available only in the electronic edition.

this we assume a fault geometry and idealize the earthquakes as points equally likely along the fault length but with estimated locations that are uncertain, due to sparse instrumentation in the area. The location of an assumed 8-km-long fault is shown in Figure 7 for illustration, but it is only used to simulate the example earthquake catalogs (i.e., it is not known for any of the following calculations). In case 1, we assume that 10 earthquakes were observed and that the location uncertainty has a 4 km standard deviation (i.e., the distance from the actual earthquake location to the observed location is assumed to have a Gaussian distribution with this standard deviation and is assumed to be at a uniform random angle from the actual location). In case 2, we assume that better monitoring was utilized, so that 30 earthquakes were observed (i.e., smaller-magnitude earthquakes can be detected), and the location uncertainty has a reduced standard deviation of 2 km.

The earthquake catalogs in Figure 7 can then be used to compute the likelihood of observing these data, conditional upon an assumed fault geometry. For a given set of parameters θ and a set of observed earthquake locations, the likelihood function is as follows (again using a proportionality formula and omitting the normalizing constant):

$$L(\theta) \propto \prod_{i=1}^{n} e^{-0.5(d_i/\sigma)^2}, \qquad (8)$$

in which n is the number of earthquakes, and d_i is the closest distance from earthquake i to the fault defined by θ. The $e^{-0.5(d_i/\sigma)^2}$ term comes from the Gaussian distribution on earthquake locations, with a standard deviation of σ.

In this case there is no simple analytical or numerical solution to the posterior distribution, so we use Gibbs sampling to simulate realizations of the properties of interest from the posterior distribution (Gelman *et al.*, 2014). For each iteration of the process, one parameter of θ is varied across its range of allowable values (leaving all other values constant at the most recently sampled value), the likelihood function of equation (8) is computed for each candidate value of θ, and the product of equations (7) and (8) provides the conditional posterior distribution for that parameter (more

precisely, a discrete approximation using the considered values), conditional on the values of the other parameters of θ. A sample is generated from this conditional posterior distribution using the inverse method, and the process is repeated for the next parameter of θ. After performing several hundred iterations to let the distribution of θ approximately converge, a set of samples from this process is collected as realizations from the posterior distribution.

Sampled realizations from this process are shown in Figure 8 and provide an implicit definition of the posterior distribution. It is clear from these figures that the posterior distribution in case 2 is narrower—that is, the fault parameters have less uncertainty due to the additional monitoring data. We use these sampled results to compute the posterior seismic hazard at the location of interest. To do this, we treat each sampled fault geometry as an equally weighted logic-tree branch representing a potential set of parameter values. This differs slightly from the more common approach of having logic-tree branches corresponding to an exhaustive discretized set of parameter values, but it will produce equivalent results as long as there are sufficient sampled values to capture the posterior distribution (Musson, 2000; Bourne *et al.*, 2015). In addition to the fault geometries from Figure 8, we assume that earthquake magnitudes have a truncated Gutenberg–Richter distribution with $M_{min} = 3$ and $M_{max} = 4.33 + 1.49 \times \log_{10}(l)$ to reflect the potential of longer faults to produce larger earthquakes (Wells and Coppersmith, 1994). Earthquakes are assumed to occur at a rate of $\lambda(M > 3) = 0.1$ per year, uniformly along the length of the fault (which then defines the earthquake distances), and we again use the Atkinson (2015) model for ground-motion prediction. Seismic-hazard results for a location in the center of the study region are shown in Figure 9. Comparing the two subfigures, we see that in case 2 the variability in hazard curves is reduced (due to the reduced uncertainty in fault properties), and the mean hazard is also reduced relative to case 1 (partly due to the reduced uncertainty and partly due to the reduced possibility of the fault coming close to the site). Figure 9b also shows the hazard curve associated with the true fault geometry, indicating the potential for the hazard to be further reduced if the true fault geometry can be more precisely estimated.
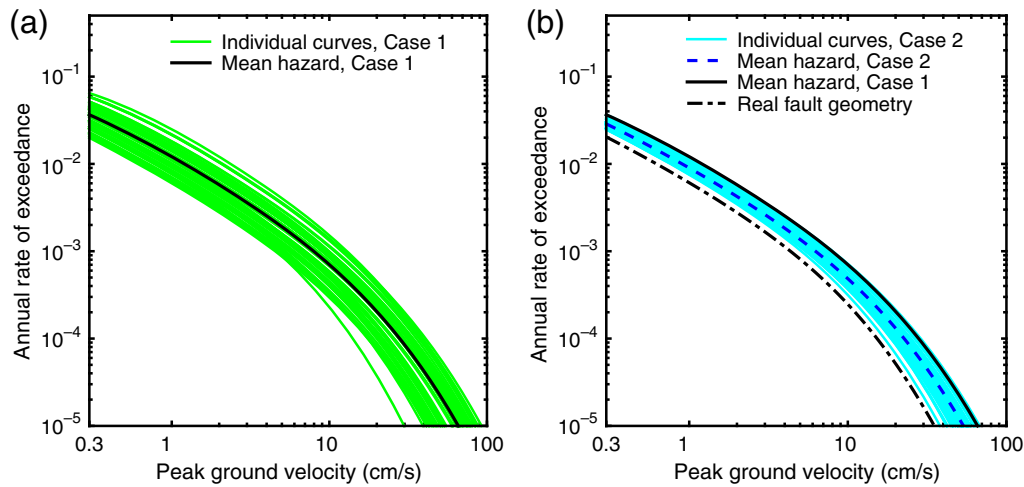
Figure 9. Individual hazard curves and mean hazard associated with the posterior fault distributions of example 3. (a) Results using the case 1 seismic catalog and (b) the case 2 seismic catalog; the mean hazard curve from case 1 and the hazard curve associated with the real fault's geometry are also superimposed for comparison. The color version of this figure is available only in the electronic edition.

This example illustrates that the approach can integrate a number of types of monitoring data into hazard calculations and that it can incentivize monitoring of an injection site to reduce earthquake location uncertainty. If there are several small earthquakes in an area, without precise locations one cannot eliminate the possibility that they are occurring on a fault capable of producing a larger earthquake. If monitoring is performed and confirms that the earthquakes are not occurring on a fault, then the observations need not lead to increased seismic-hazard estimates. Observational data indicating the potential presence of a fault would likely lead to additional data collection for making a more definitive evaluation (e.g., Horton, 2012), so the above calculation would likely not be the final analysis. But the procedure proposed here provides an automated approach to preform initial risk screening and can be used with hypothetical catalogs as the above to evaluate the potential value of information provided by improved monitoring.

Discussion

This Bayesian updating approach is most useful if two conditions hold: the hazard calculation inputs are potentially changing in the future (such that past seismic activity may not be indicative of future hazard), and observational data have the potential to indicate changes if they occur (so that the updating process will be informative). For natural seismicity, there is less value in this approach, as there is no expected change, and so future data on a short timescale are unlikely to provide a substantial update to our belief in a given model (which is presumably based on observations over a longer timescale). For induced seismicity, the opportunities for updating ground-motion-model inputs are likely to also be limited, as the variability associated with ground-motion predictions is large, and so likelihood functions on

these models will not be strongly informative; this is why the above examples considered updating of source characterization rather than ground-motion characterization.

To be useful in a regulatory context, this approach would benefit from standardization of the way in which prior distributions on model parameters are developed (because the prior distributions influence the hazard as well as the degree to which observed data can change the hazard). With a focus on seismicity from wastewater injection in the United States, information on stress orientation, relative stress magnitudes, and mapped faults can be used to identify the locations of potentially active faults (Alt and Zoback, 2014). Locations of prior potentially induced seismicity are documented by, for example, Petersen *et al.* (2015). This information could be used to establish standard prior distribution formulations for earthquake rates—for instance by having prior rate distributions reflecting a given (perhaps high) probability that future rates will be similar to past rates, but also allowing for some probability that rates will dramatically increase. It is the authors' view that understanding of this phenomenon has not yet advanced to the state where quantitative guidance on the choice of prior distributions can be provided here, but the calculation approach described here can be used to test and evaluate prior distribution models as they are developed in the future.

Note that use of a prior distribution suggesting the potential for increased seismicity may lead to mean hazard estimates that are larger than those implied if one only considers past seismicity rates (because there is now some probability of increased rates and thus increased hazard), but this is the natural implication if one believes that there is some possibility of increased future earthquake activity. Future observations of seismicity can then move the hazard higher or lower, depending upon whether increased seismicity is in fact observed.

Conclusions

A Bayesian procedure for seismic-hazard analysis has been described and utilized to update hazard calculations in the presence of induced seismicity, for which some aspect of earthquake occurrences (e.g., rate or location) has the possibility to change in the future, relative to past observations. The proposed approach shares many similarities with traditional PSHA procedures but differs in several ways: (1) epistemic uncertainty is characterized as a prior distribution that will be updated based on data; (2) likelihood functions must be formulated to quantify the consistency of observed data with an assumed model; and (3) the Bayesian updating scheme may produce samples of model parameters from a posterior distribution, rather than an explicit list of weights on model parameter values. In this case, the PSHA logic-tree branches correspond to individual sampled realizations rather than individual parameter values.

The basic principles of Bayesian updating are straightforward and already in use in natural hazard risk assessment. The challenges discussed here are in formulating likelihood functions for updating and in making the updating approach compatible with PSHA principles. By largely adopting the PSHA framework that is widely used for regulations surrounding natural seismicity and also incorporating knowledge from injection operations and real-time monitoring, this approach has the opportunity to provide insights in situations with the potential for induced seismicity. Three examples were provided to illustrate how one can perform hazard updating on a variety of model parameters, using several types of observational data (i.e., earthquake occurrence rates and earthquake locations). Although the examples all have idealizations that would need to be revised in a practical application, they nonetheless illustrate the feasibility of the approach and its compatibility with standard PSHA practice.

All of the examples rely on prior distributions for seismicity parameters, and ideally these prior distributions would be linked to operational factors associated with fluid injection in the region. Prior distribution calibration will thus benefit from continued study of the link between injection operations and future seismicity. The proposed procedure does not eliminate the need for such understanding but rather provides a process for incorporating currently available knowledge when assessing seismic hazard (and resulting risk). By providing a quantitative link between prior assumptions about the potential for induced seismicity, observed data, and resulting hazard calculations, it provides a tool to evaluate the potential benefits provided by enhanced monitoring or new scientific understanding related to induced seismicity.

Data and Resources

The earthquake occurrences surrounding Oklahoma City were obtained from the Oklahoma Geological Survey catalog (http://www.ou.edu/content/ogs/research/earthquakes/catalogs.html, last accessed April 2016). MATLAB source code to perform the example change point and Gibbs sampling calculations is available at https://github.com/bakerjw/BayesianPSHA (last accessed September 2015).

References

Ake, J., K. Mahrer, D. O'Connell, and L. Block (2005). Deep-injection and closely monitored induced seismicity at Paradox Valley, Colorado, *Bull. Seismol. Soc. Am.* **95,** no. 2, 664–683.

Alt, R. C., and M. D. Zoback (2014). Development of a detailed stress map of Oklahoma for avoidance of potentially active faults when siting wastewater injection wells, *AGU Fall Meeting*, Abstract S51A–4434.

Atkinson, G. M. (2015). Ground-motion prediction equation for small-to-moderate events at short hypocentral distances, with application to induced-seismicity hazards, *Bull. Seismol. Soc. Am.* **105,** no. 2A, doi:10.1785/0120140142.

Atkinson, G. M., and D. M. Boore (2006). Earthquake ground-motion prediction equations for eastern North America, *Bull. Seismol. Soc. Am.* **96,** no. 6, 2181.

Atkinson, G. M., H. Ghofrani, and K. Assatourians (2015). Impact of induced seismicity on the evaluation of seismic hazard: Some preliminary considerations, *Seismol. Res. Lett.* **86,** no. 3, 10.1785/0220140204.

Bayraktarli, Y. Y., J. Ulfkjaer, U. Yazgan, and M. H. Faber (2005). On the application of Bayesian probabilistic networks for earthquake risk management, *9th International Conference on Structural Safety and Reliability (ICOSSAR 05)*, 20–23.

Bommer, J. J., H. Crowley, and R. Pinho (2015). A risk-mitigation approach to the management of induced seismicity, *J. Seismol.* **19,** 623–646.

Boore, D. M., J. P. Stewart, E. Seyhan, and G. M. Atkinson (2014). NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes, *Earthq. Spectra* **30,** no. 3, 1057–1085.

Bourne, S. J., S. J. Oates, J. J. Bommer, B. Dost, J. van Elk, and D. Doornhof (2015). A Monte Carlo method for probabilistic hazard assessment of induced seismicity due to conventional natural gas production, *Bull. Seismol. Soc. Am.* **105,** no. 3.

Brodsky, E. E., and L. J. Lajoie (2013). Anthropogenic seismicity rates and operational parameters at the Salton Sea Geothermal Field, *Science* **341,** no. 6145, 543–546.

Committee on Induced Seismicity Potential in Energy Technologies (2012). *Induced Seismicity Potential in Energy Technologies*, The National Academies Press, Washington, D.C.

Convertito, V., N. Maercklin, N. Sharma, and A. Zollo (2012). From induced seismicity to direct time-dependent seismic hazard, *Bull. Seismol. Soc. Am.* **102,** no. 6, 2563–2573.

Cornell, C. A. (1968). Engineering seismic risk analysis, *Bull. Seismol. Soc. Am.* **58,** no. 5, 1583–1606.

Ellsworth, W. L. (2013). Injection-induced earthquakes, *Science* **341,** no. 6142, doi: 10.1126/science.1225942.

Esteva, L. (1969). Seismicity prediction: A Bayesian approach, *Proc. of the Fourth World Conference on Earthquake Engineering*, Santiago, Chile, 13 pp.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, Florida.

Gupta, A., and J. W. Baker (2015). A Bayesian change point model to detect changes in event occurrence rates, with application to induced seismicity, *12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12*, Vancouver, Canada, 8 pp.

Horton, S. (2012). Disposal of hydrofracking waste fluid by injection into subsurface aquifers triggers earthquake swarm in central Arkansas

with potential for damaging earthquake, *Seismol. Res. Lett.* **83,** no. 2, 250–260.

Kulkarni, R. B., R. R. Youngs, and K. J. Coppersmith (1984). Assessment of confidence intervals for results of seismic hazard analysis, *Proc. Eighth World Conference on Earthquake Engineering*, 263–270.

Llenos, A. L., and A. J. Michael (2013). Modeling earthquake rate changes in Oklahoma and Arkansas: Possible signatures of induced seismicity, *Bull. Seismol. Soc. Am.* **103,** no. 5, 2850–2861.

Majer, E., J. Nelson, A. Robertson-Tait, J. Savy, and I. Wong (2012). Protocol for addressing induced seismicity associated with enhanced geothermal systems, *US Department of Energy,* https://www1.eere. energy.gov/geothermal/pdfs/geothermal_seismicity_protocol_012012. pdf (last accessed April 2016).

McGarr, A. (2014). Maximum magnitude earthquakes induced by fluid injection, *J. Geophys. Res.* **119,** no. 2, 1008–1019.

McGarr, A., B. Bekins, N. Burkardt, J. Dewey, P. Earle, W. Ellsworth, S. Ge, S. Hickman, A. Holland, E. Majer, *et al.* (2015). Coping with earthquakes induced by fluid injection, *Science* **347,** no. 6224, 830–831.

McGuire, R. K. (2004). *Seismic Hazard and Risk Analysis*, Earthquake Engineering Research Institute, Berkeley, California.

Mignan, A., D. Landtwing, P. Kästli, B. Mena, and S. Wiemer (2015). Induced seismicity risk analysis of the 2006 Basel, Switzerland, enhanced Geothermal System project: Influence of uncertainties on risk mitigation, *Geothermics* **53,** 133–146.

Musson, R. M. W. (2000). The use of Monte Carlo simulations for seismic hazard assessment in the U.K, *Ann. Geophys.* **43,** no. 1, 1–9.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.* **83,** no. 401, 9–27.

Petersen, M. D., A. D. Frankel, S. C. Harmsen, C. S. Mueller, K. M. Haller, R. L. Wheeler, R. L. Wesson, Y. Zeng, O. S. Boyd, D. M. Perkins, *et al.* (2008). Documentation for the 2008 update of the United States national seismic hazard maps, *U.S. Geol. Surv. Open-File Rept. 2008-1128*, 61 pp.

Petersen, M. D., M. P. Moschetti, P. M. Powers, C. S. Mueller, K. M. Haller, A. D. Frankel, Y. Zeng, S. Rezaeian, S. C. Harmsen, O. S. Boyd, *et al.* (2014). Documentation for the 2014 Update of the United States National Seismic Hazard Maps, *U.S. Geol. Surv. Open-File Rept. 2014-1091*, 243 pp.

Petersen, M. D., C. S. Mueller, M. P. Moschetti, S. M. Hoover, J. L. Rubinstein, A. L. Llenos, A. J. Michael, W. L. Ellsworth, A. F. McGarr, A. A. Holland, *et al.* (2015). Incorporating induced seismicity in the 2014 United States National Seismic Hazard Model: Results of the 2014 workshop and sensitivity studies, *U.S. Geol. Surv. Open-File Rept. 2015-1070*.

Raftery, A. E., and V. E. Akman (1986). Bayesian analysis of a Poisson process with a change-point, *Biometrika* **73,** no. 1, 85–89.

Reasenberg, P. (1985). Second-order moment of central California seismicity, 1969–1982, *J. Geophys. Res.* **90,** no. B7, 5479–5495.

Senior Seismic Hazard Analysis Committee (SSHAC) (1997). Recommendations for probabilistic seismic hazard analysis: Guidance on uncertainty and use of experts, *U.S. Nuclear Regulatory Commission Report, NUREG/CR-6372*, Washington, D.C.

Shapiro, S. A., C. Dinske, and J. Kummerow (2007). Probability of a given-magnitude earthquake induced by a fluid injection, *Geophys. Res. Lett.* **34,** no. 22, doi: 10.1029/2007GL031615.

Shapiro, S. A., C. Dinske, C. Langenbruch, and F. Wenzel (2010). Seismogenic index and magnitude probability of earthquakes induced during reservoir fluid stimulations, *TLE* **29,** no. 3, 304–309.

Townend, J., and M. D. Zoback (2000). How faulting keeps the crust strong, *Geology* **28,** no. 5, 399–402.

Walters, R. J., M. D. Zoback, J. W. Baker, and G. C. Beroza (2015). Characterizing and responding to seismic risk associated with earthquakes potentially triggered by saltwater disposal and hydraulic fracturing, *Seismol. Res. Lett.* **86,** no. 4, 1110–1118.

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,** no. 4, 974–1002.

Werner, M. J., A. Helmstetter, D. D. Jackson, and Y. Y. Kagan (2011). High-resolution long-term and short-term earthquake forecasts for California, *Bull. Seismol. Soc. Am.* **101,** no. 4, 1630–1648.

Stanford University
473 Via Ortega, Room 283
Stanford, California 94305
bakerjw@stanford.edu
    (J.W.B.)

Stanford University
Blume Earthquake Engineering Center
Building 540, Room 206
Stanford, California 94305
abhineet@stanford.edu
    (A.G.)