Non-poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California

Authors: J. E. Ebel, Daniel W. Chambers, A. L. Kafka, Jenny A. Baglivo

Published in *Seismological Research Letters*, vol. 78, no. 1, pp. 57-65, 2007

THE SEISMOLOGICAL SOCIETY OF AMERICA
400 Evelyn Ave., Suite 201
Albany, CA 94706-1375
(510) 525-5474; FAX (510) 525-7204
www.seismosoc.org

Non-Poissonian Earthquake Clustering and the Hidden Markov Model as Bases for Earthquake Forecasting in California

John E. Ebel,[1] Daniel W. Chambers,[2] Alan L. Kafka,[1] and Jenny A. Baglivo[2]

INTRODUCTION

The quest to find successful methods to forecast earthquakes has proven to be very challenging. Useful earthquake forecasts require detailed specification of a number of variables, namely the epicenter, depth, time, and magnitude of the coming earthquake. While forecasting the times of strong aftershocks within the rupture zone of a strong earthquake has been developed with some success (*e.g.*, Reasenberg and Jones 1989, 1994), forecasting the times of future strong earthquakes, even when their locations are known to occur within broad geographic areas, has not been very successful. The apparent success of the M8 algorithm in forecasting the 2003 **M** 6.7 San Simeon earthquake (Keilis-Borok *et al.* 2004) followed by the failure of this same algorithm after it mistakenly forecast a strong earthquake in southern California before September 2004, shows the promise and disappointment of the current state of earthquake forecasting.

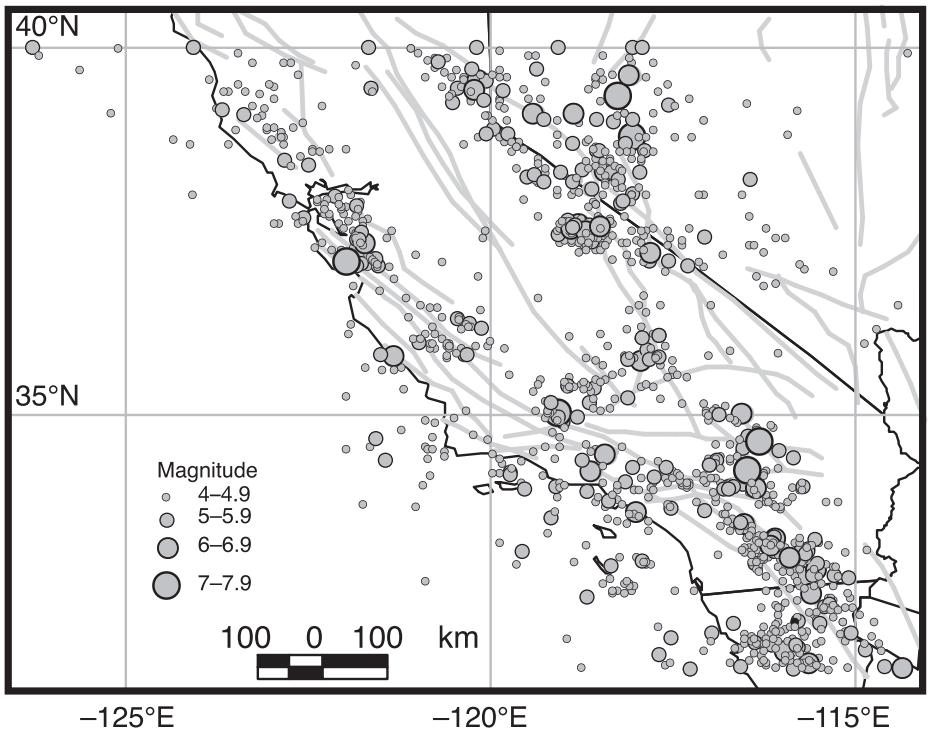
This paper describes a set of long-term (five-year) forecasts of **M** ≥ 5.0 earthquakes and two different methods for short-term (one-day) forecasts of **M** ≥ 4.0 earthquakes for California and some adjacent areas. We are submitting this set of forecasts to the Regional Earthquake Likelihood Models (RELM) project of the Southern California Earthquake Center (SCEC) for testing against other proposed forecasting methods. Two forecast maps showing the expected rate of **M** ≥ 5.0 earthquakes in the study region during the next five years were submitted for RELM testing in December 2005. One map presents a forecast of **M** ≥ 5.0 mainshocks only, while the other map has a forecast of **M** ≥ 5.0 mainshocks and aftershocks. These long-term forecast maps are based on an extrapolation into the future of the average rates from 1932 to 2004 of **M** ≥ 5.0 mainshocks in the forecast area.

The basis underlying both of the short-term earthquake forecasting methods that are described in this paper is the observation that mainshocks in California and western Nevada of **M** ≥ 4.0 are more temporally clustered than expected from a memoryless, Poisson distribution of earthquakes with time. To illustrate this, we obtained the Advanced National Seismic System (ANSS) earthquake catalog of **M** ≥ 4.0 from 1932 to 2004 for the region of figure 1 and removed all foreshocks and aftershocks using the time-space windows of Gardner and Knopoff (1974). Triggered earthquakes, defined as any earthquake that took place within one day of an **M** ≥ 6 mainshock, were also removed. The declustered catalog of epicenters is shown in figure 1, and the distribution of interevent times for this declustered catalog is shown in figure 2. It is evident that even after declustering, there are more instances of two mainshocks taking place with short interevent times than expected for a Poisson distribution with the same mean earthquake rate. The discrepancy between the observations and the Poisson distribution is greatest for event pairs with interevent times of one day or less, and it decreases back to the Poisson distribution for interevent times of five days or more. Declustering an earthquake catalog is not straightforward; rather, it depends on the definition of aftershocks and foreshocks that is used. However, it is our experience that there is an excess of interevent times of five days or less relative to a Poisson distribution no matter what declustering algorithm is used. The two different computer codes for daily forecasts of the expected rates of **M** ≥ 4.0 earthquakes will be submitted by the end of December 2006 to the RELM testing center. One daily forecast method, described in the next section of this paper, is based on an extrapolation into the future of short-term non-Poissonian earthquake clustering that was observed in the earthquake catalog of California and Nevada from 1932 to 2004. The other daily forecast method, described in more detail later, uses a hidden Markov model (HMM) with parameters derived from past seismicity to make daily forecasts of **M** ≥ 4.0 earthquakes for the region of the RELM experiment. Both computer codes will forecast mainshocks as well as aftershocks. Because of its simplicity of implementation, we base all of our mainshock forecasts on an earthquake catalog that was declustered of foreshocks and aftershocks using the Gardner and Knopoff (1974) definition of aftershocks.

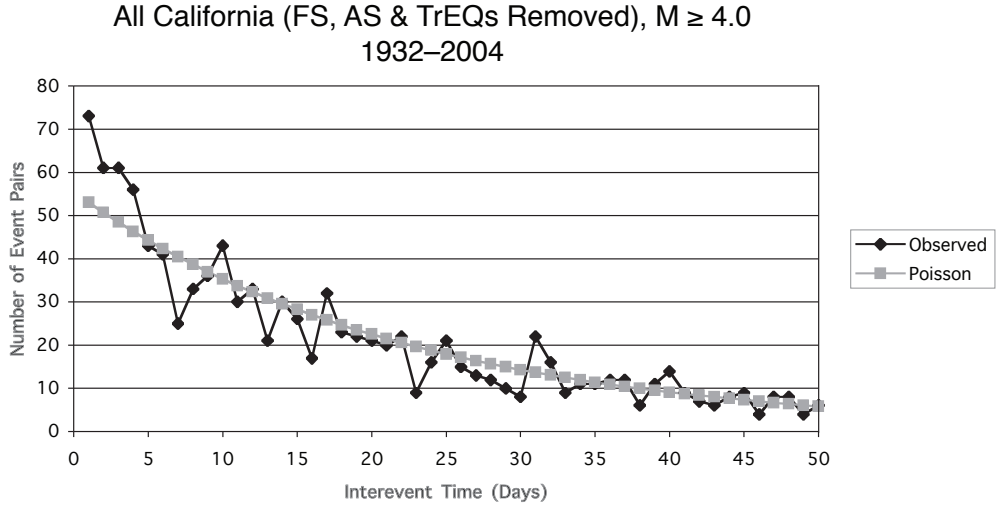
RELM EARTHQUAKE FORECASTS BASED ON SHORT-TERM NON-POISSONIAN EARTHQUAKE CLUSTERING IN CALIFORNIA

The philosophy behind the first method of earthquake forecasting is the assumption that the average statistical properties of the spatial and temporal occurrences of earthquakes with **M** ≥ 4.0 during the future forecast period are the same as the average properties of those variables over the past 70 or so years. This

1. Weston Observatory, Department of Geology and Geophysics, Boston College
2. Department of Mathematics, Boston College



▲ **Figure 1.** Map of the declustered earthquake catalog (foreshocks, aftershocks, and triggered events removed) of **M** ≥ 4.0 events from 1932 to 2004.



▲ **Figure 2.** Distribution of observed earthquake interevent times (diamonds) for the declustered earthquake catalog from 1932 to 2004 (figure 1) plotted along with the theoretical Poisson distribution for a catalog with the observed mean rate of earthquake occurrence.

assumption means that the short-term **M** ≥ 4.0 spatial forecasting generated for RELM with this method will look identical to maps of the past **M** ≥ 4.0 seismicity. Kafka (2002) has shown that for many parts of the world, including California, most new earthquakes tend to occur near locations that have experienced past earthquakes. This is true for earthquake catalogs that contain foreshocks and aftershocks as well as for catalogs from which foreshocks and aftershocks have been removed. Thus, by following our philosophy we expect on average to have a high success rate with our spatial forecasts. As for the temporal part of our earthquake forecasts, the average occurrence rates of aftershocks in and around California appears well-described by a form of Omori's law (Reasenberg and Jones 1989, 1994), and so that will form the basis of the temporal forecasting of earthquake activity near the epicenter of a larger earthquake in the time immediately following that event. For the short-term temporal forecasts of other mainshocks, the Poisson distribution of interevent times modified with an excess of earthquake pairs with short interevent times (*i.e.*, figure 2) is the statistical distribution from which these forecasts will be made.

We first describe here how aftershocks and foreshocks will be handled in this forecast method. Each time an earthquake of **M** ≥ 4.0 takes place, a circle of radius R will be drawn around the epicenter. The radius R is based on the aftershock distance defined by Gardner and Knopoff (1974) and is a function of the mainshock magnitude. Since RELM requires forecasts of seismicity rates in 0.1° by 0.1° cells, all cells that touch or contain a part of the area within R will be considered a part of the aftershock region. The formulation of Omori's law of Reasenberg and Jones (1989) with their generic California parameters will then be used to calculate the expected rate of earthquakes with any magnitude greater than 4.0. Obviously, for those magnitudes that are less than the mainshock magnitude, the forecast rate will be for aftershocks, while for those magnitudes that are greater than the magnitude of the first event, the forecast rate assumes that the first earthquake was a foreshock. That foreshocks and aftershocks can be described by the same version of Omori's law has been argued by Felzer *et al.* (2004). Table 1 shows the aftershock radii *R* and the one-day **M** ≥ 4.0 earthquake activity rates for earthquakes that are expected after the first day, fifth day, tenth day, and fiftieth day after mainshocks of different magnitudes. In our application of this approach for RELM, when the forecast aftershock/foreshock rate drops below the background mainshock rate for a cell, then the background mainshock rate will be used.

For those locations that are outside all aftershock zones, a different method will be used to compute the expected daily rate of M ≥ 4.0 earthquakes. For these areas we will use the average rate λ for the entire study area from the earthquake catalog from 1932 to 2004 after the catalog has been declustered of foreshocks, aftershocks, and one-day triggered events, as described above. We assume that this mean mainshock rate can be distributed throughout the study area proportional to the past local seismicity. To do this, we will divide the region into cells that are 0.3° by 0.3° on a side, and then for each cell i we will compute the total number of mainshocks n_i from the declustered catalog from 1932 to 2004. We then compute the expected mean rate λ_i of earthquakes in cell *i* using the formula

$$\lambda_i = \frac{\lambda n_i}{N} \qquad (1)$$

where *N* is the total number of earthquakes in the declustered catalog; if a cell contains no earthquakes, a small, arbitrary rate (several orders of magnitude smaller than that for cells with at least 1 event) is assigned to that cell. The rates per cell can then be upsampled to the 0.1° by 0.1° cell size specified for the short-term RELM forecasts. We have chosen to carry out the above calculation on 0.3° by 0.3° cells rather than 0.1° by 0.1° cells directly to get a better estimate of the rates. Of course, the downside of this approach is that it smoothes the spatial forecasts.

The mainshock seismicity rate λ_i for each cell that needs to be specified for the daily **M** ≥ 4.0 RELM forecasts depends on the seismicity during the preceding few days before the forecast. There are several cases that must be considered. First, if there was no **M** ≥ 4.0 mainshock anywhere within the entire forecast region during the preceding four days (96 hours) before the time of the forecast, then the mean daily rate λ, designated

TABLE 1
Aftershock Rates

Mm	*R* (km)	Rate (day 1)	Rate (day 5)	Rate (day 10)	Rate (day 50)
4.0	15.0	0.020	0.004	0.002	Background
4.5	17.5	0.058	0.011	0.005	0.001
5.0	20.0	0.165	0.030	0.014	0.003
5.5	23.5	0.470	0.086	0.041	0.007
6.0	27.0	1.340	0.246	0.117	0.021
6.5	30.5	3.820	0.701	0.333	0.059
7.0	35.0	10.892	1.997	0.950	0.168
7.5	40.5	31.054	5.694	2.708	0.478
8.0	47.0	88.535	16.235	7.721	1.363

M: Earthquake mainshock magnitude.
R: Maximum aftershock distance from the mainshock epicenter (Gardner and Knopoff, 1974).
Rate: Expected rate of **M** ≥ 4.0 earthquakes per day on days 1, 5, 10, and 50 after a mainshock of magnitude Mm.

TABLE 2
Rates per day of M ≥ 4.0 mainshocks in the study area

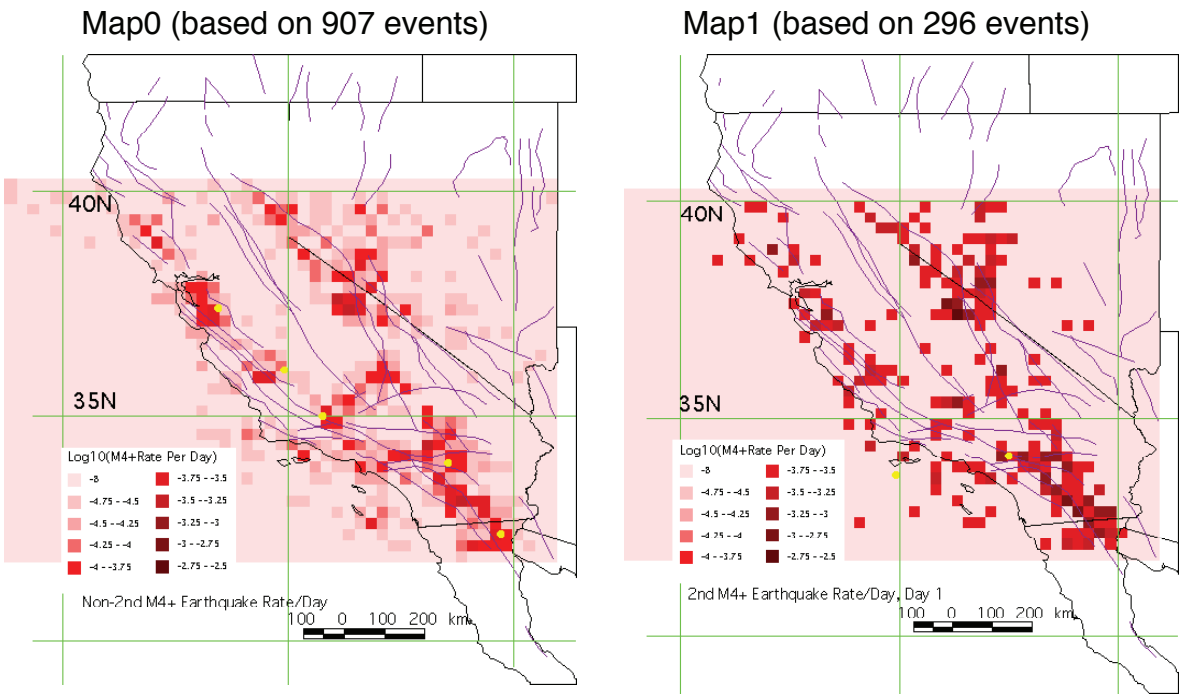
$\lambda^{(0)}$— 0.045 events/day
$\lambda^{(1)}$— 0.063 events/day
$\lambda^{(2)}$— 0.055 events/day
$\lambda^{(3)}$— 0.055 events/day
$\lambda^{(4)}$— 0.055 events/day

$\lambda^{(0)}$, $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, and $\lambda^{(4)}$, are defined in the text.

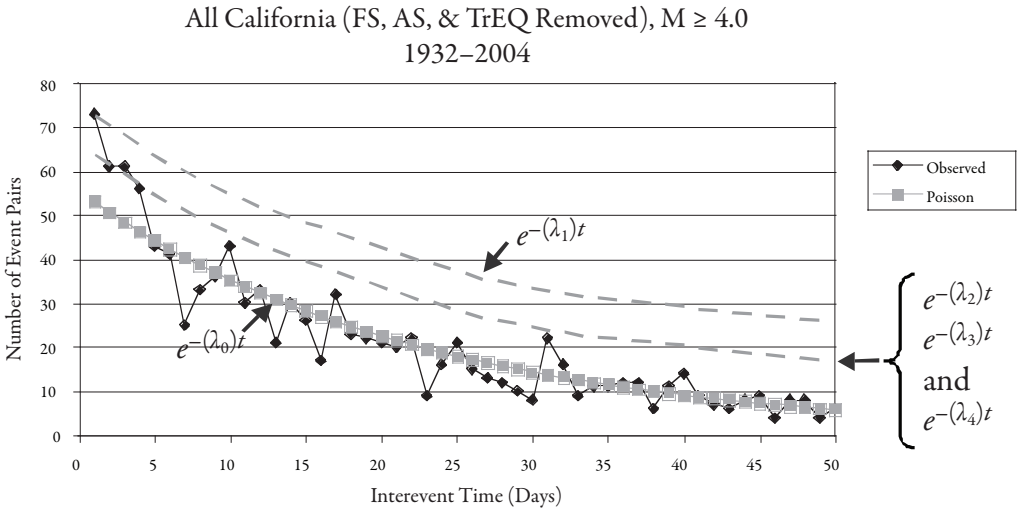
$\lambda^{(0)}$, is found from the mean daily **M** ≥ 4.0 earthquake rate for the 1932–2004 declustered catalog. The spatial seismicity rate in each cell *i* is determined by counting the number of **M** ≥ 4.0 earthquakes $n_i^{(0)}$ in the 1932–2004 declustered catalog that were not preceded within four days by another **M** ≥ 4.0 mainshock. A map of the spatial distribution of these seismicity rates, which we call Map0, is shown in figure 3, while table 2 lists the rate $\lambda^{(0)}$.

In the second case, there was an **M** ≥ 4.0 mainshock somewhere in the study region during the 24 hours before the forecast is made. In this case, the mean **M** ≥ 4.0 mainshock rate $\lambda^{(1)}$. is found as shown in figure 4, where the Poisson curve is moved upward until it intersects the data point for day one. The spatial seismicity rate in each cell *i* is determined by counting the number of **M** ≥ 4.0 earthquakes $n_i^{(0)}$ in the 1932–2004 declustered catalog that were preceded within one day by another **M** ≥ 4.0 mainshock. The designation Map1 is given to the map of the spatial distribution of the seismicity parameters for this forecast case (figure 3).

A third case is one where there was an **M** ≥ 4.0 mainshock 24–48 hours preceding the forecast period but there was no such event during the immediately preceding 24 hours. For this case, the mean **M** ≥ 4.0 mainshock rate $\lambda_i^{(2)}$ is found as shown in figure 4, where the Poisson curve is moved upward until it



▲ **Figure 3.** Plots of Map0 and Map1. The construction of Map0 and Map1 is described in the text. The yellow dots show **M** ≥ 4.0 earthquakes from January to September 2005. On Map0 the yellow dots represent the locations of mainshocks that were not preceded by another earthquake within five days, while those on Map1 show the locations of mainshocks that followed another mainshock within five days.



▲ **Figure 4.** Illustration of how the average regional seismicity rates $\lambda^{(0)}$, $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, and $\lambda^{(4)}$ for Map0, Map1, Map2, Map3, and Map4 were calculated.

intersects the data point for day two. The spatial seismicity rate in each cell i is found by counting the number of **M** ≥ 4.0 earthquakes $n_i^{(2)}$ in the 1932–2004 declustered catalog that were preceded within one day by another **M** ≥ 4.0 mainshock. The spatial distribution of these seismicity rates is called Map2. In a similar manner, the spatial seismicity rates for Map3 (**M** ≥ 4.0 mainshock in the preceding 48–72 hours but no subsequent event) and for Map4 (**M** ≥ 4.0 mainshock in the preceding 72–96 hours but no subsequent event) are created.

Table 2 lists the seismicity rates $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, and $\lambda^{(4)}$ for these forecast maps. In practice, because of the relatively small number of earthquakes per cell for finding $n_i^{(1)}$, $n_i^{(2)}$, $n_i^{(3)}$, and $n_i^{(4)}$ we have decided to improve the statistical sample by counting all earthquakes that were preceded any time within the previous five days by another mainshock and to use this number as n_i for Map1, Map2, Map3 and Map4. Thus, for our forecasts with this method, Map1, Map2, Map3, and Map4 have identical spatial patterns and differ only in their absolute seismicity

rates $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, and $\lambda^{(4)}$. In this method, the total one-day **M** ≥ 4.0 earthquake forecast for each day for RELM is a combination of any aftershock forecasts for those places with recent mainshocks combined with the appropriate mainshock forecast as described in the previous paragraphs. Thus, a map of the forecast that would be issued on a given day would be either Map0, Map1, Map2, Map3, or Map4 modified to show local increases in the forecast seismicity rate at locations where recent mainshocks had taken place. An example of some forecast maps that would have been generated using this method are forecast maps for the days before and just after the December 2003 San Simeon earthquake (figure 5). One can see the changes in the forecast maps as the seismicity took place through the time period depicted in figure 5.

The RELM project calls for **M** ≥ 4.0 earthquake forecasts to be issued daily at a predetermined time. For the method described in this section, the daily forecasts are in essence the issuance of one of the five maps described above, modified by aftershock forecasts at those locations where recent mainshocks have occurred. Which of the five maps gets issued depends on the **M** ≥ 4.0 mainshock throughout the region during the previous few days. Table 3 lists the forecast maps that would be generated based on a hypothetical daily report of earthquake activity in the region during the previous 24 hours before each forecast is issued. As called for by RELM, the seismicity rates shown in Map0, Map1, Map2, Map3, and Map4 will not be altered in any way during the duration of the RELM daily earthquake forecast experiment.

The RELM project also calls for the issuance of maps showing the expected rate of **M** ≥ 5.0 earthquakes in the study region during the next five years. One map forecasts only mainshocks, while another map forecasts all **M** ≥ 5.0 events (foreshocks, mainshocks, and aftershocks). Once issued, these maps are unchanged during the course of the RELM forecast experiment. In this case, our mainshock forecast method proposed in this section consists simply of computing the average rate per five years of independent **M** ≥ 5.0 mainshocks in each 0.1° by 0.1° cell in the study region after upsampling from 0.3° cells, as described above. A version of our mainshock forecast map, based on the average rate of **M** ≥ 5.0 earthquakes from 1932 to 2004, is shown in figure 6. In some areas (such as around Long Valley in California), the expected number of **M** ≥ 5.0 earthquakes in five years approaches or exceeds 1. Our **M** ≥ 5.0 forecast map that includes both mainshocks and aftershocks modifies the mainshock forecast map by adding aftershock seismicity to each cell where the rate of aftershocks is computed using the generic California aftershock model of Reasenberg and Jones (1989). Foreshocks are not included in this forecast. In these and all our forecasts, we assume that the largest earthquake than can take place is **M** 8.0, and the largest aftershock that can occur is **M** 7.0.

Daily forecasts of **M** ≥ 4.0 seismicity as well as the single forecast of **M** ≥ 5.0 earthquakes issued for the RELM experiment are required to specify the rate of earthquake activity for each 0.1 magnitude unit starting at the lowest magnitude specified for that forecast. In the method we propose here, we will use a single *b* value from a Gutenberg-Richter magnitude distribution (Gutenberg and Richter 1944) from our declustered catalog to calculate the expected number of earthquakes at each magnitude level. Thus, for each cell shown in Map0, Map1, Map2, Map3, and Map4, a Gutenberg-Richter (GR) distribution of the earthquake magnitudes will be assumed, with the value for *a* determined from the seismicity observed in that cell from 1932 to 2004. For mainshocks, a *b* value of 0.77 is used, while for aftershocks the *b* value is 0.91 from the generic aftershock model of Reasenberg and Jones (1989).

TABLE 3
Hypothetical Set of Earthquake Forecasts

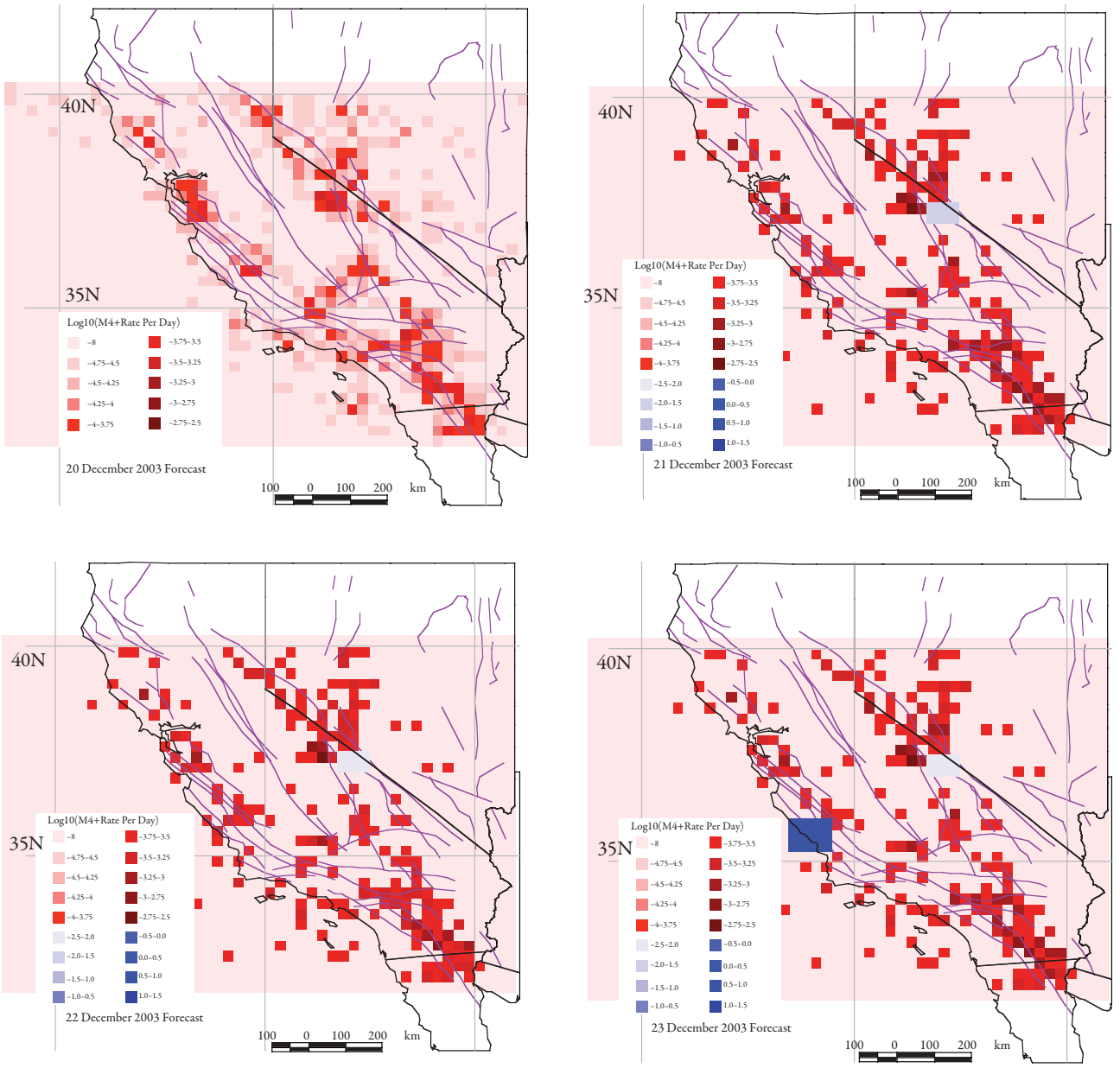
Day	Earthquake	Map Issued for Next Day
Day 1	None	Map0
Day 2	M4.2	Map1
Day 3	None	Map2
Day 4	None	Map3
Day 5	None	Map4
Day 6	None	Map0
Day 7	M5.3	Map1
Day 8	M4.2	Map1
Day 9	None	Map2
Day 10	None	Map3
Day 11	M4.6	Map1

Earthquake refers to a mainshock of **M** ≥ 4.0 anywhere in the study area.

HIDDEN MARKOV MODEL EARTHQUAKE FORECASTS FOR CALIFORNIA

The second method of earthquake forecasting uses the hidden Markov model (HMM) (see, for example, Baum and Petrie 1966). Hidden Markov models are a rich class of statistical models that have been applied in fields as diverse as speech recognition (Rabiner 1989), ion channel analysis (Fredkin and Rice 1992a, 1992b), bioinformatics (Durbin *et al.* 1998), and seismology (Granat and Donnellan 2002). HMMs were shown by Granat and Donnellan (2002) to fit earthquake data in southern California and were used to find classes of similar earthquakes. Here we use the HMM to forecast future earthquakes in a dynamic way, basing each forecast on the data available up to that point in time.

A hidden Markov model consists of a sequence of observations and a sequence of unknown (hidden) states. The distribution of a future observation depends on the state of the system at that time. The system moves from state to state according to a Markov chain. At any given time, the state is unknown, but the probability of being in each state can be computed given the previous observations. The concept of a state here is a statistical construct, not a physical one; however, it corresponds to the idea that physical conditions imply that the next earthquake has an increased probability of occurring in a particular

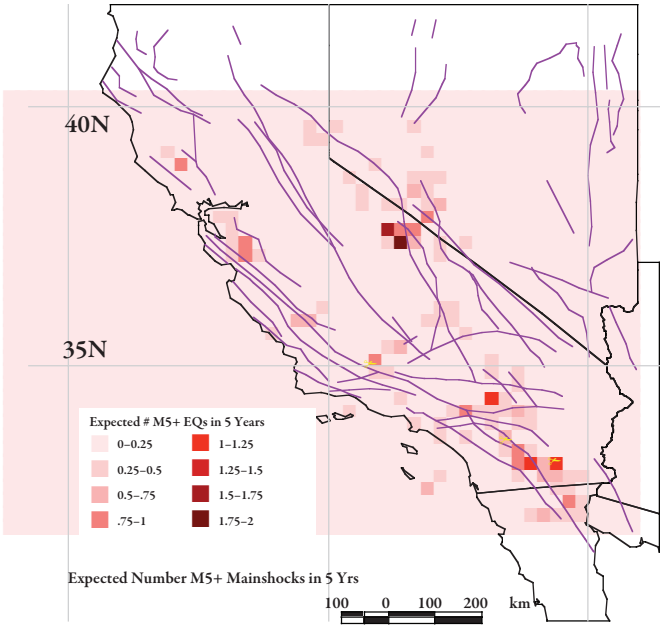


▲ **Figure 5.** Sample daily forecast maps of **M** ≥ 4 earthquake activity for the time period from 20–24 December 2003. On 20 December, an **M** 4.0 earthquake occurred in eastern California, followed by an **M** 6.5 earthquake near San Simeon, California, on 22 December. The red colors show the forecast rates for mainshocks, while the blue squares show the forecast rates for aftershocks.

spatial quadrant and is more likely to occur either sooner (*i.e.*, within the next few days) or later (*i.e.*, a couple weeks hence). The role of the states is crucial in our forecasting method since (a) we can estimate the probability of a future state given current observations and (b) we know the distribution of future observations based on each state. By combining these, we forecast the probabilities of future observations based on current observations, which is the heart of forecasting. The presence of states in our hidden Markov model provides the bridge from the past seismicity to the projected future observations.

In our implementation, the observations associated with an earthquake are its interevent time (the number of days since the previous earthquake) and the spatial quadrant in which it occurs. Figure 7 shows the four spatial quadrants into which we have divided California and the location of the 1,202 **M** ≥ 4 earthquakes in California from 1932 to 2004 in the declustered catalog used to estimate the parameters of the model. The axes that define the four spatial quadrants were derived from a principal components analysis (Rao 1973) of the declustered catalog.

We used eight states in our model, corresponding to an expected shorter or longer interevent time and an increased likelihood of being in one of the four spatial quadrants. Given a particular state, we took the probability distribution of the interevent time to the next earthquake to be an exponential distribution with a mean assigned to that state and the probability

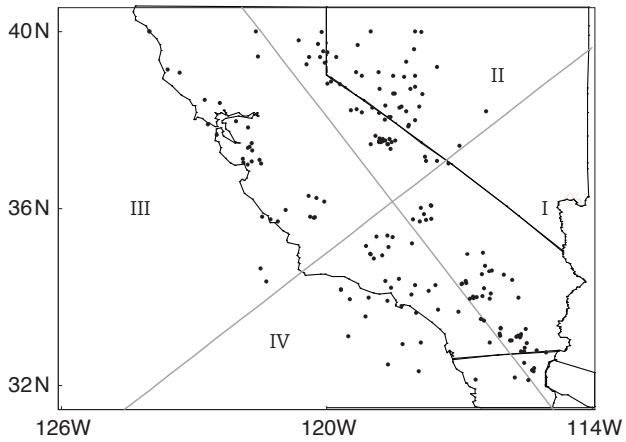


1 January 2005–16 September 2005

▲ **Figure 6.** Map of the expected number of **M** ≥ 5.0 earthquakes in the study area during the next five-year period. The yellow dots show the locations of earthquakes of **M** ≥ 5.0 from January to September 2005.

distribution of the location as a vector of the probabilities that the next earthquake occurs in each of the four quadrants.

The distributions of interevent time and location for each state are represented in table 4. For example, if the system is in State 1, the interevent time for the next earthquake has an exponential distribution with a mean of 4.5 days; the probabilities that the next **M** ≥ 4.0 earthquake will occur in quadrants 1–4 are (0.676, 0, 0.324, 0), respectively. By integrating the exponential density over one day, we find the probability of an earthquake within 24 hours anywhere in the RELM region to be 0.199.



▲ **Figure 7.** The four spatial quadrants, numbered by Roman numerals, used in the HMM.

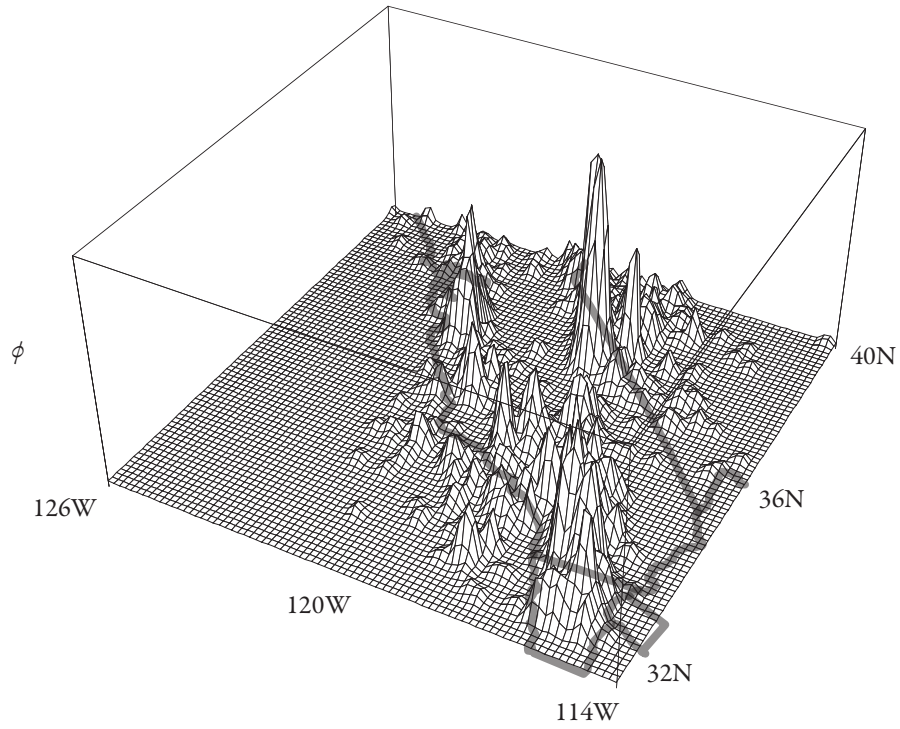
Finally, we multiply this by the spatial quadrant distribution for this state to find the probability of an earthquake within 24 hours in each of the four quadrants. Similar results, as shown in table 4, hold for the other states. These distribution parameters and all other parameters for the HMM were estimated using standard HMM techniques (see, for example, Rabiner 1989; Granat and Donnellan 2002).

Our forecasting procedure is quite simple. Each time a daily forecast is made, our code uses the observations of past seismicity available to it and computes the probability of being in each of the eight states at the time of the next earthquake. It then uses these eight probabilities in a weighted average of the probabilities in the last column of table 4 to compute the probability of an earthquake of unspecified magnitude at or above **M** 4.0 within 24 hours in each of the four spatial quadrants. It is important to note what changes and what stays the same in our HMM procedure. The state-specific distributions in table 4 remain constant; what changes with each forecast is the probability of being in each of the states, conditional on the updated

TABLE 4
HMM state parameters

State	Mean	1-day Probability	Quadrant Distribution	1-day Probability in Quadrant
1	4.5	0.199	(0.676, 0, .0324, 0)	(0.135, 0, 0.065, 0)
2	2.1	0.384	(0.084, 0.916, 0, 0)	(0.032, 0.352, 0, 0)
3	20.5	0.048	(0, 0.564, .0436, 0)	(0, 0.027, 0.021, 0)
4	9.1	0.104	(0, 0.182, 0, 0.818)	(0, 0.019, 0, 0.085)
5	24.4	0.042	(0.828, 0.172, 0, 0)	(0.033, 0.007, 0, 0)
6	29.9	0.033	(0.001, 0.999, 0, 0)	(0, 0.033, 0, 0)
7	29.7	0.033	(0, 0, 1.0, 0)	(0, 0, 0.033, 0)
8	24.8	0.040	(0, 0.005, 0.068, 0.927)	(0, 0, 0.003, 0.037)

For each state:
Mean: mean of the exponential distribution
1 day probability: probability of an earthquake within next 24 hours
Quadrant distribution: probability that the next earthquake occurs in each of the spatial quadrants
1 day probability in quadrant: probability of an earthquake within next 24 hours in each quadrant



▲ **Figure 8.** Graph of the spatial smoothing function ϕ, a mixture of 1,202 bivariate Gaussian densities.

seismicity information, and hence the forecast probability of an earthquake within a day in each of the four quadrants.

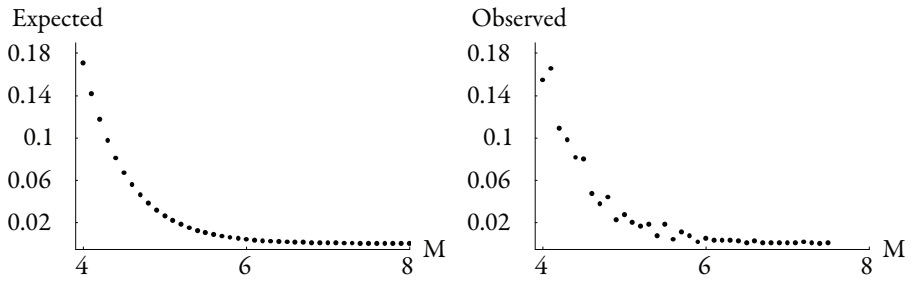
The rest of the code translates these four probabilities into rates for each of the RELM 0.1° by 0.1° and 0.1 magnitude unit interval cells by multiplying the probability for the quadrant in which the cell appears by a cell location factor and a cell magnitude factor, determined as follows.

For each of the 1,202 $\mathbf{M} \geq 4$ earthquakes in the declustered 1932–2004 California catalog, we defined an uncorrelated bivariate Gaussian density with mean at the epicenter and with a common standard deviation of 0.1°. These were averaged to give a bivariate Gaussian mixture density $\phi(x, y)$ of the spatial probabilities of an earthquake at position (x, y). (See figure 8 for a graph of ϕ.) For each RELM 0.1° by 0.1° cell, the integral of ϕ over the cell is divided by the integral of ϕ over the quadrant. The result is used as the location factor for that cell, and it represents the probability that an earthquake occurs in that cell, conditional on its occurring in that spatial quadrant. The magnitude factor for each cell was derived from the GR line that was fit to the declustered 1932–2004 catalog. The a value was found by normalizing this GR line to have unit area between M = 4 and M = 8. For each magnitude (specified to 0.1 magnitude units), the magnitude factor is simply the value of this normalized GR line at that magnitude. See figure 9 for a graph of the magnitude cell factor.

The coded application of the HMM procedure for the RELM earthquake forecasting experiment works as follows. Each time a forecast is to be issued, the updated ANSS catalog is used to forecast the probability of an earthquake of $\mathbf{M} \geq 4$ within 24 hours in each of the four spatial quadrants. The code then takes each RELM bin consisting of a 0.1° by 0.1° location and a 0.1 unit magnitude range and multiplies the probability for that quadrant where the cell is located by the location factor and the magnitude factor for that bin. The result is the reported rate for that bin. Finally, using the same procedure described earlier in this paper for the first forecasting method, the forecast activity rates of cells within the aftershock zones of recent mainshocks are adjusted to the aftershock activity rate based on the time since the mainshock.

DISCUSSION AND CONCLUSIONS

The two short-term earthquake forecast models that are described in this paper are both extrapolations of past seismic activity into the future, but each does the extrapolation in a different way. The short-term non-Poissonian earthquake clustering model is effectively an empirical extrapolation of the average behavior of the past 72 years of earthquake activity into the future. It is very simple, as there is no underlying model other than the statistical properties of the past seismicity. Furthermore, it is nonadaptive in that once the forecast maps Map0, Map1, Map2, Map3, and Map4 have been defined, they will not change during the course of the RELM experiment. The HMM is also an extrapolation to the future of the average behavior of the past earthquake activity, but with the underlying idea that the seismicity at any given time can be in any one of several states, with the probability of each state being calculated as part of the model. In our HMM formulation, the state parameters are determined from the past 72 years of earthquake activity and remain unchanged throughout the course



▲ **Figure 9.** Left: graph of the magnitude cell factor, expected number of earthquakes of given magnitudes under the GR relationship; right: observed distribution of magnitudes in the 1932–2004 declustered California catalog.

of the RELM experiment. However, the model is adaptive in that the forecast probabilities and therefore forecast earthquake rates change each day based on the seismic activity up to that time. Therefore, the HMM allows for the creation of a very wide range of forecast maps compared to the non-Poissonian earthquake clustering model described above.

The handling of aftershocks for both methods proposed here is rather simple, and this is by design because it is the forecasting of mainshocks that is our primary interest. We do not plan to make our aftershock forecasts adaptive (*i.e.*, updating the Omori-law parameters each day as an aftershock sequence plays itself out). Since there are many quantitative models to describe the temporal evolution of an aftershock sequence, a concerted effort is needed just to determine the best model to apply. It was our decision to take a simple, widely used aftershock model and to use it with average aftershock parameters that have been found previously for the forecast region.

The major focus of this study is to see if times of increased probabilities of earthquake mainshocks in California can be identified based on extrapolations from the past seismicity history. If such times can indeed be identified, even if the earthquake probability is only somewhat enhanced over the background Poisson probability, there will certainly be public interest in this capability. There would also be scientific interest in this capability, since it would then be possible to look for other geological and geophysical indicators that correlate with the times of enhanced earthquake probability.

REFERENCES

Baum, L. E., and T. Petrie (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* **37**, 1,554–1,563.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Ebel, J. E. (2003). Seismologists must begin forecasting earthquakes. *Seismological Research Letters* **74**, 3–5.

Felzer, K., R. E. Abercrombie, and G. Ekström (2004). A common origin for aftershocks, foreshocks, and multiplets. *Bulletin of the Seismological Society of America* **94**, 88–98.

Fredkin, D. R., and J. A. Rice (1992a). Bayesian restoration of single-channel patch cloamp recordings. *Biometrics* **48**(2), 427–448.

Fredkin, D. R., and J. A. Rice (1992b). Maximum likelihood estimation and identification directly from single-channel recordings. *Proceedings: Biological Sciences* **249**, 125–132.

Gardner, J. K., and L. Knopoff (1974). Is the sequence of earthquakes in southern California with aftershocks removed Poissonian? *Bulletin of the Seismological Society of America* **64**, 1,363–1,367.

Granat, R., and A. Donnellan (2002). A hidden Markov model-based tool for geophysical data exploration. *Pure and Applied Geophysics* **159**, 2,271–2,283.

Gutenberg, B., and C. F. Richter (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America* **34**, 185–188.

Kafka, A. (2002). Statistical analysis of the hypothesis that seismicity delineates areas where future large earthquakes are likely to occur in the central and eastern United States. *Seismological Research Letters* **73**, 990–1,001.

Keilis-Borok, V. I., P. N. Shebalin, P. N. Mitpan, K. Aki, A. Jin, A. Gabrielov, D. L. Turcotte, Z. Liu, and I. Zaliapin (2004). Documented prediction of the San Simeon earthquake six months in advance; premonitory change of seismicity, tectonic setting, and physical mechanism. *Seismological Research Letters* **75**, 266.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE* **77**, 257–286.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley & Sons.

Reasenberg, P. A., and L. M. Jones (1989). Earthquake hazard after a mainshock in California. *Science* **243**, 1,173–1,176.

Reasenberg, P. A., and L. M. Jones (1994). Earthquake aftershocks: Update. *Science* **265**, 1,251–1,252.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, and D. Jackson (2004). Earthquake likelihood model testing. RELM Web site http://www.RELM.org .

Weston Observatory
Department of Geology and Geophysics
Boston College
Chestnut Hill, Massachusetts 02467 USA
ebel@bc.edu
(J.E.E.)
kafka@bc.edu
(A.L.K.)

Department of Mathematics
Boston College
Chestnut Hill, Massachusetts 02467 USA
chambers@bc.edu
(D.W.C.)
baglivo@bc.edu
(J.A.B.)