GEAR1: a Global Earthquake Activity Rate model constructed from geodetic strain rates and smoothed seismicity

P. Bird[1], D. D. Jackson[1], Y. Y. Kagan[1], C. Kreemer[2], and R. S. Stein[3]

[1]Department of Earth, Planetary, and Space Sciences, University of California-Los Angeles

[2]Nevada Bureau of Mines and Geology, and Seismological Laboratory, University of Nevada-Reno

[3]U.S. Geological Survey, Menlo Park, California

Final manuscript of 3 July 2015, accepted by ***BSSA***

This submission includes Electronic Supplement files, including additional text and references on scoring of forecasts, tables of scoring results, and source code and data files needed to reproduce our forecast.

ABSTRACT

GEAR1 estimates the rate of shallow earthquakes with magnitudes 6 through 9 everywhere on Earth. It was designed to be reproducible and testable. Our preferred hybrid forecast is a log-linear blend of two parent forecasts based on the Global CMT catalog (smoothing 4602 $m \geq 5.767$ shallow earthquakes, 1977-2004) and the Global Strain Rate Map version 2.1 (smoothing 22415 GPS velocities), optimized to best forecast the 2005-2012 GCMT catalog. Strain rate is a proxy for fault stress accumulation, and earthquakes indicate stress release, so a multiplicative blend is desirable, capturing the strengths of both approaches. This preferred hybrid forecast outperforms its seismicity and strain rate parents; the chance that this improvement stems from random seismicity fluctuations is less than 1%. The preferred hybrid is also tested against the independent parts of the ISC-GEM catalog ($m \geq 6.8$ during 1918-1976) with similar success. GEAR1 is an update of this preferred hybrid. Comparing GEAR1 to the Uniform California Earthquake Rupture Forecast version 3 (UCERF3), net earthquake rates agree within 4% at $m \geq 5.8$ and at $m \geq 7.0$. The spatial distribution of UCERF3 epicentroids most resembles GEAR1 after UCERF3 is smoothed with a 30-km kernel. As UCERF3 has been constructed to derive useful information from fault geometry, slip rates, paleoseismic data, and enhanced seismic catalogs (not used in our model), this is encouraging. To build parametric catastrophe bonds from GEAR1, one could calculate the magnitude for which there is a 1% (or any) annual probability of occurrence in local regions.

INTRODUCTION

Forecasting of seismicity is one of the more important practical applications of geophysical research. Given a good forecast, societies have an opportunity to optimize their investments in safer buildings and more resilient infrastructure; they also have a context in which to consider offers and purchases of insurance. Objective scoring of forecasts can advance science by testing hypotheses about earthquake generation and interaction (*e.g.*, Kelleher *et al.*, 1973; McCann *et al.*, 1979; Kagan and Jackson, 1991; Nishenko and Sykes, 1993; Jackson and Kagan, 1993). Our objective in this paper is to build a testable global reference model of the expected long-term rates of shallow earthquakes (those with hypocentroids no more than 70 km below sea level) as a function of space and magnitude.

Only seismic catalogs, global plate boundary models, and Global Positioning System (GPS) geodetic velocities provide uniform global coverage. Despite the obvious importance of databases of active faults in seismic hazard studies, a comprehensive global inventory of active faults does not yet exist. Few faults are well-mapped and fewer still have reliable slip rates, geometries, and rakes needed to transform those faults into earthquake sources. Thus, the only faults represented in this model are the principal plate boundaries such as subduction zones and oceanic transforms, and even these are designated only as belts of straining, not as specific planes. Also, only a global model that forecasts moderate-magnitude earthquakes implies a sufficient rate of shocks to meet the testing requirement. We will demonstrate below that competing global forecasts can be reliably ranked after only 1 to 8 years of testing, provided that those forecasts have magnitude thresholds of approximately 5.8 to 7.0, respectively.

Previous forecasts have been constructed in two fundamentally different ways: by smoothing of past catalog seismicity, or by applying seismic-coupling coefficients to faults with estimated slip rates and other zones whose tectonic deformation rates have been measured. The creation of a smoothed-seismicity forecast from a seismic catalog is straightforward, though it requires careful research into optimization of the smoothing algorithm. A strength of smoothed-seismicity methods is that they can capture hazards far from plate boundaries such as igneous intrusions and gravity tectonics, such as the earthquakes of magnitude up to 7.4 which have occurred in Hawaii. But a weakness is that existing catalogs are too short to include seismicity along all plate boundaries and fault zones. Another issue is that if small earthquakes are used to increase the sample size, induced earthquakes can be included as sources; it is not yet known whether a better forecast of large earthquakes would be obtained by including, or omitting, this induced seismicity. A further complication is that induced seismicity typically has a different time-dependence than natural seismicity.

Tectonic forecasts require a reasonably complete database of deforming zones (*i.e.*, active faults with their slip rates, and/or deforming areas with their strain-rates, and/or adjacent plates and their Euler vectors). They also require a seismic catalog to calibrate the coupling coefficients that will be used to convert fault slip rates and/or distributed strain rates to long-term seismicity. However, if these sources can be grouped into a few tectonic zones of global extent, then earthquakes accumulate rapidly in each zone, and it may be that only a few decades of seismic catalog will suffice for calibration of a model with a modest number of degrees of freedom. One weakness of tectonic forecasts is their potential to overestimate seismicity of regions where faults creep aseismically. This is particularly important in subduction zones, where some seismologists and geodesists believe that there is broad diversity in the extent to which they are seismically coupled, while others (*e.g.*, McCaffrey, 2008) question whether this is measurable with present datasets.

One recent development in tectonic forecasting is the incorporation of relative plate rotations and GPS-derived interseismic velocities. Such velocity-based tectonic models (*e.g.*, Bird, 2009; Field *et al.*, 2013) impose kinematic compatibility on their faults and zones of straining, reducing the risk that incomplete information about one fault, or one benchmark, will result in incorrect seismicity forecasts.

Some forecasts and hazard models use a spatial composite approach, in which the well-known faults are explicitly represented by traces, dips, and slip rates; but other deformation is approximated by distributed sources derived from smoothed seismicity. The recent Uniform California Earthquake Rupture Forecast version 3 (UCERF3) by Field *et al.* (2013) is such a forecast.

Several groups have begun to pursue "hybrid," "mixture," or "ensemble" approaches, in which two or more forecasts are combined to forecast the earthquake rate in every spatial cell (rather than partitioning space as in a spatial-composite forecast). Rhoades and Gerstenberger (2009) proposed a linear combination of two time-dependent models. Bird *et al.* (2010b) gave a preview of global linear and log-linear hybrids of smoothed-seismicity and tectonic components, with encouraging retrospective test results. Marzocchi *et al.* (2012) proposed a Bayesian method for creating a linear-combination ensemble of existing forecasts with optimized weights, and applied it to 6 existing Regional Earthquake Likelihood Models (RELMs) for the southern California region (Field, 2007; Schorlemmer *et al.*, 2010). Rhoades *et al.* (2013, 2014) combined these same RELMs into many multiplicative hybrid models, and found a greater improvement with multiplicative mixing than with linear combinations; prospective testing of these hybrids is planned at the Collaboratory for the Study of Earthquake Predictability (CSEP). Taroni *et al.* (2013) discuss four methods for linearly combining global forecasts, and some reservations concerning available tests.

In this project we combine only two "parent" forecasts into a variety of hybrid forecasts. The smoothed-seismicity parent forecast (or "Seismicity" for brevity) is a global forecast, on a 0.1° × 0.1° grid, of shallow earthquakes with scalar moment $M > 10^{17.7}$ N m, based on GCMT shallow seismicity (Ekstrom *et al.*, 2012, and references therein) during 1977-2004, computed by the methods of Kagan and Jackson (1994, 2000, 2011). Basically, each epicentroid point (whose magnitude is above the threshold) is convolved with a previously-optimized generic smoothing kernel, and the results are summed to produce a map of forecast shallow earthquake rates (above the same threshold). Each smoothing kernel is a product of functions of radius, source earthquake magnitude, and azimuth. As a function of radius, the smoothing kernel is that of equation (3) in Kagan and Jackson (2011), with parameter $r_s = 6$ km. The overall amplitude of each smoothing kernel is a linear function of source earthquake magnitude, so larger events are considered to forecast greater future seismicity. Also, since each GCMT centroid includes two possible fault planes, each smoothing kernel is anisotropic, with greater future seismicity forecast along the inferred strikes of the possible faults. There is no time-dependence in this long-term forecast, and all earthquakes in the source catalog (including possible aftershocks) are equally important, regardless of their sequence. A minimum or background level of intraplate seismicity, integrating to 1% of total shallow seismicity, is uniformly distributed. This Seismicity parent forecast is shown in ***Figure 1A***.

Our second parent forecast is a tectonic forecast (or “Tectonics” for short) based on version 2.1 of the Global Strain Rate Map (GSRM2.1) of Kreemer *et al.* (2014). This strain-rate map was based on plate-tectonic concepts and 22415 interseismic Global Positioning System (GPS) velocities. Thus, it can be the basis for a self-consistent velocity-based forecast. It was converted by Bird and Kreemer (2015) to a long-term tectonic forecast of seismicity using the Seismic Hazard Inferred From Tectonics (SHIFT) hypotheses presented by Bird and Liu (2007). The specific algorithm is very similar to that which Bird *et al.* (2010a) used to create an earlier global tectonic forecast from GSRM (version 1). Basically, each strain-rate tensor is converted to a long-term seismic moment rate by multiplication with the elastic shear modulus, the grid-cell area, a dimensionless geometric factor, and a depth parameter called the coupled seismogenic thickness. This coupled thickness value is taken from the "most comparable class" of plate boundary in previous published compilations. Then, the seismic moment rate is converted to earthquake rates by taking the normalized frequency/magnitude distribution of the same "most comparable class" of plate boundary as a model. The algorithmic innovations of Bird and Kreemer (2015) are that: (1) spatial smoothing was applied to the activity (both strain-rate and seismicity) of offshore plate boundaries, and (2) velocity-dependence of seismic coupling in subduction zones and continental convergent boundaries (Bird *et al.*, 2009) was included. This forecast (***Figure 1B***) was originally created on a global grid of 0.25° × 0.20° cells, but has been resampled on a 0.1° × 0.1° grid for this project. Numerical smoothing due to resampling was minimal because 80% of the values transfer unchanged (when expressed as epicentroid rate densities in m^{-2} s^{-1}), and the other 20% are simple equally-weighted averages of the values in two adjacent cells within one row.

Both our two parent forecasts (discussed above) and our hybrid forecasts (discussed below) share a common feature: They are forecasts of total seismicity, with no use of declustering and no distinction between "mainshocks" and "aftershocks." This feature is motivated by the lack of *in-situ* physical distinctions between these two classes, by the lack of community agreement on an optimal declustering scheme, and by consideration of likely misclassifications that would result from catalog boundaries in space, time, and magnitude. We concede that this departure from the RELM tradition (Schorlemmer & Gerstenberger, 2007) may also have disadvantages, although the only one now apparent is the need for caution in the selection of testing algorithms, as detailed in our electronic supplement.

HYBRID FORECASTS

The forecasts discussed in this paper do not include any explicit time-dependence. (However, forecasts prepared using different calibration time windows will differ slightly as a result.) . All forecasts have only a single depth bin: hypocentroids no more than 70 km below sea level. In the form which we will retrospectively test (below), they have only a single magnitude bin: all earthquakes at or above a magnitude threshold, without distinction. All forecast rate densities are expressed on a common global grid of 0.1° × 0.1° cells, are uniform within each cell, and are discontinuous at cell boundaries. We will refer to the Seismicity parent forecast as an 1800 × 3600 matrix of positive numbers S_{ij} which give the forecast rate density of shallow earthquake epicentroids in the cell at row i and column j of this grid, in units of $m^{-2}\ s^{-1}$. (Because we use seismicity rate density, the values are laterally smooth near the poles, instead of becoming very small as they would if we tabulated expected earthquake numbers.) We will refer to the Tectonics parent forecast as matrix T_{ij}, and to the Hybrid forecast as matrix H_{ij}. The rough or initial version of each hybrid H''_{ij} is produced by parallel numerical operations on all corresponding pairs of Seismicity ($\tilde{S}$) and Tectonics ($\tilde{T}$) cells, without any lateral interactions between neighboring cells.

Another simplification in the first phases of this study was that we scaled the Seismicity forecast to have the same global earthquake rate as the Tectonics forecast, before combining them. That is, both parent forecasts followed the global frequency/magnitude curve of the Tectonics forecast, which in turn was based on the union of different tapered Gutenberg-Richter distributions (Bird & Kagan, 2004) for different plate-boundary analogs. This approach is not necessarily the best for scaling to higher threshold magnitudes, and in a later section below, we will propose a potentially more accurate (but more complex) solution. However, because the algorithms that we will use for scoring forecasts are insensitive to overall forecast earthquake rate, this choice of scaling method has little or no effect on our test results.

To obtain the final form H_{ij} of each hybrid forecast we apply two regularizing transformations, the second of which introduces some weak lateral interaction. First, we require that every cell of every hybrid have a positive value, no less than a minimum epicentroid-rate-density value f, which we have chosen as $f = \inf(\inf(S_{ij}), \inf(T_{ij}))$, where "inf" stands for "infimum" (the lesser, as in the intrinsic MIN function of Fortran):

$$H'_{ij} = \sup(H''_{ij}, f) \quad (1)$$

and "sup" stands for "supremum" (the greater, as in the intrinsic MAX function of Fortran). Second, we normalize the global integral of the forecast to a desired global shallow earthquake rate R, while preserving minimum seismicity density f, by a linear transformation:

$$H_{ij} = f + \left(H'_{ij} - f\right)\left(R - Gf\right) \Big/ \left(\left(\sum\sum H'_{ij} A_j\right) - Gf\right) \quad (2)$$

where A_i is the area of each cell in row i, and $G = 4\pi r^2 = 3600\ \Sigma A_i$ is the area of the Earth based on a spherical approximation with radius r. We will abbreviate this second step by representing it as application of a normalizing forecast operation $H = \mathrm{N}(H')$, and abbreviate the result of both regularizing steps by $H_{ij} = \mathrm{N}(\sup(H''_{ij}, f))$. For all retrospective tests, the global earthquake rate R imposed by the N() operator was chosen to be the same as that of both parent forecasts.

One traditional hybrid is a weighted linear-combination of S and T:

$$H_{ij} = \mathrm{N}(\sup((c\, S_{ij} + (1 - c)\, T_{ij}), f)) \tag{3}$$

where c is to be determined. Linear mixing of forecasts can be justified by either of two arguments: (a) he two parent forecasts can be regarded as expressing alternative measurements of the same underlying process, possibly with different error sources; or (b) seismicity can be regarded as the sum of two independent components, which are described by the Seismicity and Tectonics parents, respectively. Marzocchi *et al.* (2012) presented a complex algorithm that could be used to estimate c. However, since there is only one parameter to optimize, we prefer to create and test alternative hybrids using multiple values of c.

Another possible view might be that the two parent forecasts capture *independent* prerequisites for seismicity: there must be a continuing energy source for lithospheric deformation (some of which is elastic), *and* there must also be triggering by sudden stress changes (either static or dynamic) due to nearby earthquakes to start a new earthquake rupture. Probability theory predicts that the chance of an event requiring two independent preconditions is proportional to the product of their two separate probabilities. Also, when space/time discretization is fine enough so that all probabilities are much less than unity, then rates are proportional to probabilities. Therefore, it is plausible to suppose that earthquake rates might be proportional to the product of two precondition probabilities which might be captured in the Seismicity and Tectonics forecast maps, respectively. In this view, it is more appropriate to *multiply* the S and T estimates:

$$H_{ij} = \mathrm{N}(\sup((S_{ij}^{d}\, T_{ij}^{(1-d)}), f)) \tag{4}$$

or equivalently,

$$H_{ij} = \mathrm{N}(\sup((10^{[\, d \log S + (1-d) \log T\,]}), f\,)) \tag{5}$$

where d is an exponent to be determined. This set of hybrids will be called "log-linear." Note that d = 0.5 gives the geometric mean of Seismicity and Tectonics. We have only considered exponents which sum to unity because the parent forecasts have each already been optimized to contain the proper dynamic range of seismicity densities, and we want all log-linear hybrids to share this property.

Finally, it is possible that both the Seismicity and Tectonics forecasts underestimate the true rates in different localities, and so taking the larger of the two in every cell might more successfully forecast future quakes. For example, the Tectonics forecast might underestimate seismicity in regions of volcanism and landslides which lie in plate interiors, like Hawaii. The Seismicity forecast may seriously underestimate the future seismicity of those plate-boundary segments which happened not to have any large earthquakes during the learning period. So, one additional hybrid selects the greater of Seismicity or Tectonics:

$$H_{ij} = \mathrm{N}(\sup(\, S_{ij}\, ,\, T_{ij})) \tag{6}$$

Note that all hybrid models we produce use weights that are global and independent of magnitude. Both spatially-variable weighting and magnitude-dependent weighting could be considered in the future. The great difficulty lies in testing the value of such additional degrees of freedom; if they only affect forecast rates of very rare earthquakes (*i.e.*, either at high magnitudes, or in plate interiors), they are likely to remain untestable for centuries.

RETROSPECTIVE TESTING AGAINST EARTHQUAKES OF 2005-2012

The ideal way to evaluate success of forecasts is prospective testing by independent authorities, such as the Collaboratory for the Study of Earthquake Predictability (CSEP). However, since we intend to select a preferred model based on subtle differences seen in 8-year retrospective tests, it could take a similar number of years to get definitive confirmation or refutation of our selection. Also, to justify the effort of independent prospective testing, models generally should demonstrate success in retrospective tests as a necessary (but not sufficient) condition. Thus, we begin with retrospective tests. Of course, the largest 2005-2012 earthquakes are known to those who created the model, and this opens the door to subtle biases. The tests we perform here might be called "pseudo-prospective" because we test models that were created without using those years of the seismic catalog that will be used for testing. However, we permit the use of other kinds of data, such as GPS velocities, that were collected during the test years. This is because we use the GPS data to infer the secular or long-term process, and the most recent data tend to be more accurate, permit the use of longer time series, and are more geographically complete.

The primary test catalog we use is the full Global Centroid Moment Tensor (GCMT) catalog (Ekström *et al.*, 2012, and references therein). It gives the location of the centroid (also known as hypocentroid) which is the point source best representing the low-frequency and permanent offsets due to one earthquake. For planar faults, this (hypo)centroid is typically located in the middle of the slip distribution. We refer to the overlying surface point as the epicentroid. GCMT is the catalog that was used for calibration in the Tectonics forecast and as a basis for smoothing in the Seismicity forecast. Prior studies have shown it to be relatively complete above scalar seismic moment $M = 10^{17.7}$ N m (Kagan, 2003). In this paper, we use the moment (M)-to-magnitude (m) conversion of the U.S. Geological Survey:

$$m = \left(2/3\right)\left(\log_{10}(M) - 9.05\right), \qquad (7)$$

while noting that some other authors and authorities have used slightly different formulas, and that some authors have preferred the symbols M_0 for scalar moment and M_w for magnitude. Under conversion (7), $M = 10^{17.7}$ N m corresponds to moment magnitude $m = 5.767$.

The GCMT test period we use is 2005-2012 inclusive, or 8 full years. This yields 1694 shallow (≤ 70 km) test earthquakes ***(Figure 2A)***, but leaves the prior ~78% of the GCMT catalog available for calibration and learning along each forecasting branch. Also, this test window postdates the calibration study of Bird and Kagan (2004), who used years 1977-2003 of the GCMT catalog to determine boundary half-widths, coupled thicknesses, corner magnitudes, and asymptotic spectral slopes (β of Bird & Kagan, 2004) of different kinds of plate-boundary seismicity; all of these determinations are employed in the Tectonics forecast under test.

The forecast-scoring metrics that we have used include the information scores $\boldsymbol{I}_0$ (specificity) and $\boldsymbol{I}_1$ (success) of Kagan (2009), and the space statistic "S" of Zechar *et al.* (2010). The reasons for selecting these metrics, and a general discussion of their algorithms and characteristics, can be found in the electronic supplements to this article (see "Discussion of forecast-scoring metrics").

Because we will give the greatest weight to the $\boldsymbol{I}_1$ (success) scores of the hybrid models when we choose the preferred hybrid, an informal overview of this metric may be appropriate here. $\boldsymbol{I}_1$ is the mean (over all test-earthquakes) of the information gain (expressed as a number of binary bits, including fractional bits) of the forecast under test, using the forecast relative probability of the spatial cell into which the epicentroid of the test earthquake falls. The reference model (for defining information gain) has equal seismicity spread uniformly across the globe. Both the forecast under test and the reference model are expressed as maps of conditional probability of the (longitude, latitude) location of the next test epicentroid, with spatial integrals of unity; therefore, the overall rate of earthquakes forecast is not a factor in any $\boldsymbol{I}_1$ score; only the quality of its map-pattern is important. In our work, we have typically found that the contributions to $\boldsymbol{I}_1$ from individual test earthquakes range from about -5.8 (when the test earthquake is a surprise earthquake in a plate-interior) to about +10.5 (when the test earthquake occurs in one of the most seismic subduction zones). Mean $\boldsymbol{I}_1$ scores (averaged over all test earthquakes during the test time window) range from about 3.4 to 4.3 for the models rated here. Thus, the best results we discuss have mean forecast earthquake probabilities (in the cells containing the actual test earthquakes) that are about $2^{(4.3-3.4)} = 2^{0.9} =$ 1.866 times higher (87% higher) than in the worst results. Yet, even the worst model is better by a mean factor of $2^{3.4} = 10.6$ than pure ignorance. The S statistic that we discuss and present in our electronic supplements uses a totally different algorithm, but we found that it gave similar or identical rankings of our hybrid forecasts. In contrast, the $\boldsymbol{I}_0$ ("specificity") metric of Kagan (2009) is an abstract measure of the potential information gain of a forecast, computed without regard to (*e.g.*, perhaps in advance of) any test earthquakes.

The threshold magnitude for our preferred test is the estimated catalog-completeness threshold of $m \geq 5.767$ (Kagan, 2003; note our equation 7), because this yields the greatest number of test earthquakes (1695) and so the greatest statistical power. We will abbreviate this threshold as *m*5.767+ in supplementary tables (available in the electronic supplement to this article) and in discussions. But, it is also important to learn whether the forecasts perform equally well for larger, more damaging earthquakes. Therefore, we also perform parallel tests on forecasts prepared using moment threshold $M \geq 3.548 \times 10^{19}$ N m (magnitude threshold $m \geq 7.00$, or *m*7+). Unfortunately, at this level there are only 90 test earthquakes available, and the statistical power of these tests is much lower. Testing at higher thresholds (*e.g.*, *m*8+ or *m*9+) will not be meaningful for at least a century, even for the globe as a whole.

Results of all measures of forecast success and specificity for these GCMT tests are shown in ***Table S1***, available in the electronic supplement to this article. The variations of $\boldsymbol{I}_0$ (specificity) and $\boldsymbol{I}_1$ (success) scores for all linear-combination and log-linear models are shown in ***Figure 3***. Four broad conclusions are apparent from Table S1 and from Figure 3:

1. Each method of hybridization we tried resulted in hybrids that score better than the parent forecasts. This is true whether we measure success by $\boldsymbol{I}_1$ or S. We suspect this occurs because our two parent forecasts are very different, and reflect nearly independent approaches to forecasting seismicity, with substantially uncorrelated biases and errors.

2. The log-linear mixing method produced the most successful hybrid, which was the one with $d = 0.6$; that is, exponent of 0.6 on the Seismicity component and exponent of 0.4 on the Tectonics component. This preferred hybrid will be referred to as H* below. We suspect that multiplicative mixing outperformed linear mixing because our two parent forecasts capture independent requirements for seismicity: secular accumulation of elastic strain (Tectonics), and time-specific triggering or advancement of slip instabilities (Seismicity). Future prospective

testing will be the best way to determine whether a log-linear hybrid is always superior to a linear hybrid; we do not claim this as a definitive result.

3. The success of H* at threshold *m*5.767+ seems hold up as the threshold is raised to *m*7+. (Its S-statistic drops from 0.97 to 0.59, but the latter result is still excellent.) Naturally, the higher-magnitude results are less definitive due to the limited number of test earthquakes. Still, this encourages us to propose H* as a viable candidate model for higher magnitudes, even though forecasts for threshold magnitudes *m*8+ and *m*9+ cannot be conclusively tested with current catalogs.

4. Hybridization, by any of the three methods we tested, results in lower specificity ($\boldsymbol{I}_0$) of the hybrid forecast, compared to the parents. This is natural, as each parent forecast predicts moderate seismicity in a few regions which are at the intraplate-background level in the other forecast. The loss of specificity is less with log-linear mixing than with the other two methods we tried. The specificity $\boldsymbol{I}_0$ of H* is 3.801 for *m*5.767+, which is only slightly less than the 3.829 specificity of its Seismicity parent.

In selecting our preferred hybrid model (H*), we gave primary weight to the $\boldsymbol{I}_1$ success scores, which measure the mean (over all test epicentroids) number of binary bits of information gain from using this forecast instead of a spatially-uniform null forecast: this was about 4.2 bits at both thresholds. We note that H* also has the highest S-statistic in each set of tests (0.971 for *m*5.767+, 0.59 for *m*7+). Specificity $\boldsymbol{I}_0$ was not a selection criterion, because specificity exceeding success is not particularly desirable, and suggests some systematic problem with a forecast. However, we see (in Table S1, available in the electronic supplement to this article) that very little specificity was sacrificed in preferring H* relative to the Seismicity parent forecast.

SIGNIFICANCE OF HYBRID IMPROVEMENT

An important question is, whether the improvement we have obtained through hybridization is significant, considering the inherent time-variability of forecast scores? Can we show, in advance of prospective testing, that our identification of the best model is likely to be stable, and therefore that our preferred model H* is truly superior? Actually, it is not very helpful just to estimate the variance of each test metric individually; it is more useful to know their correlations and the statistics of their differences. Here we argue that the difference is significant, based on the small amount of time-history available to us and a simple scaling argument. We focus on $\boldsymbol{I}_1$ success, as it is the simpler measure to interpret. First, we look at the year-to-year behavior of the critical score difference, and estimate its standard deviation based on 8 test windows of one-year length. Then, we consider how standard deviations of scores are expected to scale with the length of the test window; for this we appeal both to theory and to the 36-year GCMT history of retrospective success of the Tectonics parent. This leads to model standard deviations for our identified improvements in 8-year $\boldsymbol{I}_1$ tests, and thus an educated guess as to their significance.

Table S2, available in the electronic supplement to this article, shows the time-history through 2005-2012 of $\boldsymbol{I}_1$ success scores of the preferred hybrid H* and of the previous best parent forecast, which was Seismicity. These annual tests with threshold m5.767+ used an average of 212 earthquakes per test. The time-history of the difference $\boldsymbol{I}_1$(H*) – $\boldsymbol{I}_1$(S) had a mean of 0.294

and a sample standard deviation of 0.087 across these one-year tests. The sample correlation coefficient of the $\boldsymbol{I}_1$ successes of these two forecasts is 0.958. This happens because some years (*e.g.*, 2008) had several (4~5) unexpected intraplate earthquakes which lowered the scores of both models, while some years (*e.g.*, 2011) had only ~1 intraplate earthquake, but had many earthquakes on known plate boundaries which both models correctly forecast. This finding is encouraging, because it suggests that meaningful distinctions between competing models can be made after brief tests. For example, if the long-term average of the difference $\boldsymbol{I}_1$(H*) – $\boldsymbol{I}_1$(S) is actually 0.294, as we currently estimate, and if this difference has a normal distribution with standard deviation 0.087 across one-year tests, then the chance of finding a negative difference (*i.e.*, preferring the other model) in any future one-year test would be less than 0.1%, because such a result would be more than 3 apparent standard deviations from the apparent mean of the difference.

If the threshold is raised to *m*7+ so that there are only ~11 earthquakes per test, then all results are more variable and uncertain. These 8 one-year tests on the right side of Table S2 (available in the electronic supplement to this article) show that standard deviations of the $\boldsymbol{I}_1$ scores of these two competing models rise by factors of 2.3 and 3.1, respectively, and the standard deviation of their difference rises by a factor of 3.6, to 0.32. Still, the correlation of $\boldsymbol{I}_1$(H*) with $\boldsymbol{I}_1$(S) remains high, at 0.947. Consequently, the sign of the score difference $\boldsymbol{I}_1$(H*) – $\boldsymbol{I}_1$(S) only reversed in one of the 8 years. Formally, we can estimate that, if the long-term mean score difference is actually 0.47, and its standard deviation is actually 0.32 across multiple one-year tests, then we would expect to see a preference for the Seismicity model (relative to the hybrid H*) in just 7% of one-year tests at threshold *m*7+.

The $\boldsymbol{I}_1$ scores in Table S1 (available in the electronic supplement to this article) are even more reliable for indicating relative model quality, because they are all from 8-year tests. We can estimate the improvement in certainty by estimating how the standard deviations of score differences scale with the number of years in the test. One might suppose that the standard deviation of any test metric (or difference in metrics) should scale in proportion to $N^{-1/2}$, where N is the number of test earthquakes. Of course, this can only be proven under the assumption that earthquakes are independent. Also, scaling with number of earthquakes can only be translated into scaling with number of years if earthquakes occur at a constant global rate. Therefore, the simple hypothesis that standard deviations of test metrics should scale as $W^{-1/2}$, where W is the length of the test time window, needs to be checked. ***Figure 4*** displays the standard deviation of the $\boldsymbol{I}_1$ success of the Tectonic parent forecast over the whole GCMT period of 1977-2012; to obtain these small-sample standard deviations the 36-year history was subdivided many times, into shorter windows with W = 1 yr, 2 yr, 3, yr, … 9 yr, and these windows were created using every possible start-year. (To obtain these bootstrap estimates, we overlook the slight circularity of testing the Tectonics forecast against some of the same earthquakes that were used to calibrate its 5 zonal seismicity-correction factors, as described by Bird & Kreemer, 2015.) In fact, $W^{-1/2}$ scaling seems consistent with our results. This was expected based on Kagan's (2009; his Fig. 3) result, using simulated catalogs rather than real ones, that the $\boldsymbol{I}_1$ score is a random variable whose distribution is close to a normal distribution.

Based on this scaling, we estimate that the standard deviations of the critical score difference $\boldsymbol{I}_1$(H*) – $\boldsymbol{I}_1$(S) over multiple future 8-year tests should be 0.031 at threshold m5.767+ and 0.11 at threshold m7+. This means that the hybrid improvements in $\boldsymbol{I}_1$ that we found in 8-year tests

(Table S1, available in the electronic supplement to this article) have signal/noise ratios of 9.5 (at threshold *m*5.767+) and 4.3 (at *m*7+).

Finally, we are able to assess the "statistical significance" of hybrid improvement, using first the physical-sciences and then the statistical meaning of that term. Fortunately, both communities share some common concepts and vocabulary: Our null hypothesis is that preferred hybrid H* is no better than the parent Seismicity forecast according to the $\boldsymbol{I}_1$ metric. Our complementary hypothesis is that H* is better than Seismicity according to the $\boldsymbol{I}_1$ metric. The p value is the model chance of obtaining the actual signal/noise ratio (or a higher one) if the null hypothesis were correct; it is obtained from the Gaussian cumulative distribution function (with mean of 0 and standard deviation of 1) when the independent variable is the negative of the signal/noise ratio, so in this case $p =$ 1×10^{-21} (at *m*5.767+) and $p =$ 8.5×10^{-6} (at *m*7+). The complement of the p -value is $(1-p)$.

In physical-sciences usage, "statistical significance" is a positive real number, expressed using any of 3 popular metrics: signal/noise ratio, p -value, or the complement of the p -value (often described as %-confidence). The significance level is considered to be p . Therefore we can say, in physical-science usage, that there is more than 99%-confidence that hybrid improvement is real, at either magnitude threshold.

In statistical usage, "statistical significance" is limited to the logical values True or False. To determine which is appropriate requires a pre-selected significance level based on community standards. For purposes of illustration, let us select $\alpha = 0.01$. Then, the statistical significance of hybrid improvement is True at either threshold, because both p -values are less than α .

RETROSPECTIVE TESTING AGAINST EARTHQUAKES OF 1918-1976

Storchak *et al*. (2012) released the International Seismological Centre-Global Earthquake Model (ISC-GEM) catalog, which is a comprehensive revision of the longstanding ISC catalog. Their work included consultation of original sources; inclusion of more phases; uniform relocation of all earthquakes with a single modern algorithm; and assignment of moment magnitude (m) to every event, either through review of the literature or by use of regression relations. This new catalog is believed to be relatively complete for moment threshold $M \geq 1.778{\times}10^{19}$ N m (m6.8+) from 1918 onward (Michael, 2014; Di Giacomo et al., 2015). In those years which predate the routine production of GCMT solutions (1918-1976), there are 881 shallow earthquakes of m6.8+ in this catalog which we have not previously used, either for model-construction or for testing. We take this opportunity to assess whether the hybrid improvements that we demonstrated in the previous sections are specific to the last decade and to the GCMT catalog, or are more universal.

Parent and hybrid models were prepared for threshold m6.8+, but otherwise in exactly the same ways as for the previous tests. That is, the catalog-calibration window for both parents was GCMT 1977-2004.

Table S3, available in the electronic supplement to this article, gives all of these test results. The patterns we see are almost identical to those from m7+ tests against GCMT 2005-2012 (Table S1), except that these tests have more statistical power due to 10 times as many test earthquakes, and that $\boldsymbol{I}_1$ successes and S-statistics are generally lower. As before, we find that: (1) Both parent forecasts have comparable success; (2) All hybrid forecasts perform better than either parent, with a maximum improvement of +0.4 in $\boldsymbol{I}_1$; (3) Log-linear hybrids perform best; (4) The best log-linear hybrid is a relatively even blend of Tectonics and Seismicity; and (5) The loss of $\boldsymbol{I}_0$ specificity for the H* preferred hybrid, relative to the Seismicity parent, is small.

These results are important because they demonstrate that the value of each parent forecast, and the improvement in hybrid mixtures, is relatively independent of time and technology. The generally lower level of $\boldsymbol{I}_1$ success scores (offset by -0.4) and S-statistics (offset by -0.23) compared to the GCMT 2005-2012 tests in Table S1 (available in the electronic supplement to this article) can probably be attributed to two causes: (1) There are less accurate epicenters, depths, and magnitudes in the ISC-GEM catalog. Even though events have been relocated with modern algorithms, errors in phase arrival times due to analog recording and/or clock drift in the period 1918-1976 are much more difficult to correct. Also, accurate magnitude estimation is difficult with narrow-band seismometers, and (in the early decades) with non-standard seismometers. (2) The Seismicity parent forecast gets less help from long-running aftershock sequences when the test window is longer.

SCALING THE SEISMICITY PARENT FORECAST TO HIGH MAGNITUDES

The previous discussion has focused entirely on testing and optimizing the map-patterns of forecasts at those moderate magnitudes where test earthquakes are abundant. Yet, the high-resolution global forecast template of CSEP requires estimation of earthquake rate maps at thresholds up to *m*8.95+. Also, the GEM Foundation has a goal of building global seismic hazard and risk models which will require similar high-magnitude rate estimates. Computation of a preferred hybrid forecast H* for a high threshold magnitude requires that we have corresponding versions of both parent forecasts. Since the high-magnitude scaling of the Tectonics forecast is already defined (Bird and Kreemer, 2015), it remains to specify how the Seismicity parent forecast will be extrapolated to high magnitudes. To reduce artifacts and problems, it is important to take account of the different corner magnitudes m_c (which locate the roll-offs of frequency/magnitude curves) in different tectonic settings (Bird *et al.*, 2002; Bird & Kagan, 2004; Kagan *et al.*, 2010).

A straightforward way to incorporate this information is to scale the local (per-cell) epicentroid rate densities, from the original Seismicity forecast with threshold m_t = 5.767 to a higher threshold m (*e.g.*, 8), by use of the G factor from the tapered Gutenberg-Richter frequency-moment relation:

$$\frac{S_{ij}(m)}{S_{ij}(m_t)} = G(m, m_t, m_c, \beta) = \left(\frac{M(m)}{M(m_t)}\right)^{-\beta} \exp\left(\frac{M(m_t) - M(m)}{M(m_c)}\right) \tag{8}$$

where the $S_{ij}(m)$ are per-cell shallow earthquake rate densities (in $m^{-2}\ s^{-1}$) above magnitude m; $M(m)$ is the scalar moment associated (7) with magnitude m; m_c is the corner magnitude in the cell, and β is the asymptotic spectral slope of the frequency-moment relation (for $m \ll m_c$) in the same cell (Jackson & Kagan, 1999; Kagan & Jackson, 2000; Bird & Kagan, 2004).

We first implemented scaling (8) using the maximum-likelihood corner magnitudes (6.79 ~ 8.75) and spectral slopes (0.639 ~ 0.767) of the 5 tectonic zones in Table 1 of Kagan *et al.* (2010), together with the tectonic-zone map of the same paper. After a number of experiments, we decided to moderate this simplistic application of tectonic zonation in five ways: [1] We raised the corner magnitude of zone 4 (Trench) to 9.5, based on later research of Kagan & Jackson

(2013). This value is also more consistent with results of Bird & Kagan (2004); yet it still falls within the uncertainty of Kagan *et al.* (2010). [2] We merged tectonic zone 0 (Intraplate) with tectonic zone 1 (Active continent) using weighted-averages m_c = 7.72 and β = 0.645 in their union, in order to eliminate artifacts which had been appearing along the 0/1 zone boundaries during extrapolation. These two mean values are within the uncertainty ranges of the 4 unmerged estimates in Kagan *et al.* (2010). [3] We spatially smoothed the map of zone-based corner magnitude and the map of zone-based spectral slope to eliminate remaining discontinuities; this smoothing is done by convolution with an isotropic Gaussian kernel of scale length 200 km. [4] We applied a constant stretching factor to variations from the mean within each smoothed map, in order to restore their original standard deviations. (Before smoothing, corner magnitudes had an area-weighted mean of 7.806 and standard deviation of 0.463; smoothing reduced this standard deviation to 0.322; amplification of remaining variations by factor 1.439 brought the standard deviation back to 0.463.) [5] During extrapolation of the Seismicity forecast to high magnitudes, we applied the extrapolated epicentroid rate density of the united zone 0/1 (outside the halos of any catalog earthquakes) as a lower limit on the forecast epicentroid rate density of all cells. This is to recognize the possibility of occasional energetic ruptures on new faults, even in the vicinity of old plate boundaries. It also limits the dynamic range of the extrapolated forecast to be more similar to the dynamic range of the forecast for *m*5.767+ which was previously optimized, and which we have tested. Details of this algorithm are contained in the source code provided as an electronic supplement to this article.
Figure 5 shows an example of a Seismicity parent forecast (for years 2005+) extrapolated to *m*8+ by these methods.

The extrapolated Seismicity parent forecast is now in reasonable agreement with the frequency-magnitude statistics of global catalogs (***Table 1***). However, this exercise highlighted the importance of both the corner magnitude we apply in zone 4 (Trench), and the generic frequency/magnitude curve that we assumed for all zones. Great *m*9+ earthquakes are rare, a few per century, and the rate difference, if we accept $m_c = 9.5$, between a straight-line Gutenberg-Richter frequency/magnitude distribution and a tapered Gutenberg-Richter distribution seems small. As calculated previously (Jackson & Kagan, 2012; Kagan & Jackson, 2013), the global rate of *m*10+ events is 0.057 per century or 0.21 per century for the gamma distribution or tapered Gutenberg-Richter distribution, respectively. But, it increases to 0.57 per century for the classical straight-line Gutenberg-Richter law. We recognize the desirability of further research and testing regarding these issues.

GLOBAL EARTHQUAKE ACTIVITY RATE MODEL

Earlier in this paper we established that the best-performing hybrid H* in the most powerful retrospective test (against shallow GCMT earthquakes, *m*5.767+, in 2005-2012) was the log-linear hybrid (equations 4, 5) with exponent of $d = 0.6$ on the Seismicity component. This gives us a basis for proposing a global reference model that presently appears optimal, at least for those moderate magnitudes where testing is currently meaningful. However, the earthquake rates of the two parent forecasts diverge slightly at high threshold magnitudes; thus, we must also specify a choice regarding the combination of these two forecasts of the global shallow earthquake rate (R_S from Seismicity; R_T from Tectonics) for $m > 5.767$. By analogy with the formula that determines the map-pattern of H*, we choose the global rate formula

$$R_{\mathrm{H*}}(m > 5.767) = R_{\mathrm{S}}^{0.6}(m) \times R_{\mathrm{T}}^{0.4}(m)\,. \tag{9}$$

Up until this point, we have illustrated, tested, and discussed models based on GCMT catalog years 1977-2004, which left the years 2005-2012 (and 1918-1976) available for testing. To improve our preferred model in advance of prospective testing, it is also important to make use of all available years in the modern broad-band digital-seismology catalog. Thus, we recomputed both parents, Seismicity and Tectonics, and the preferred hybrid H* based on all available complete GCMT years: 1977-2013. One change was that $R_{\mathrm{GCMT}}(5.767)$ based on 1977-2013 is 6.5% higher than the rate based on 1977-2004 (Figure 2) because of the rate increase of 26.7% that occurred at the end of 2004. Another change was that local maxima in forecast seismicity appear near large earthquakes of 2005-2013 because of the influence of the updated Seismicity parent forecast.

This update of our preferred hybrid model H*, with $R_{\mathrm{H*}}(m>5.767)$ based on (9) above, is named Global Earthquake Activity Rate model 1 (GEAR1). Figure 2B and ***Figure 6*** show maps of this model at thresholds of *m*5.767+ and *m*8+, respectively. As threshold magnitude rises above the calibration level of *m*5.767+, GEAR1 global earthquake rates forecast for the future match past instrumental catalog rates fairly well through thresholds *m*7+, *m*8+, and *m*9+ (Table 1).

There are a number of reasons why GEAR1 will eventually be superseded by revised versions (“GEARn”). Continuing enlargement of the global GPS dataset may eventually prompt an update of the Tectonics parent component. Also, an improved hybrid might use a future Tectonics forecast employing both GPS strain-rates and the GEM Faulted Earth and/or GEM Subduction Sources datasets in a unified kinematic finite-element deformation model. The extrapolation of the Seismicity parent forecast to high magnitudes may be revised or further optimized. Catalog seismicity from before 1977 may eventually be incorporated into the Seismicity parent forecast. Also, strong seismicity in risk-sensitive parts of the globe could prompt an update of the Seismicity component, again leading to a new GEAR. In any case, long-term independent prospective testing of GEAR1, whether superseded or not, should have value in verifying the expected long-term stability of hybrid improvement.

In the CSEP forecast format (XML file), all forecasts must have defined start- and end-dates. The forecast start-date and end-date for GEAR1 must be chosen by the user, in the GEAR1_parameters.dat file which is input before the XML file is created. The start-date should be no earlier than 2014.01.01 to avoid circularity. All forecast earthquake counts in each magnitude bin of each spatial cell will be proportional to the length of the forecast time window. However, conceptually the time-window for this GEAR1 forecast is 2014+, which is indefinite or open. (This is why we prefer to display our results as maps of earthquake rates rather than earthquake counts.)

An important question for future testing and research is: For how long into the future should a forecast of the GEAR type be trusted? Large earthquakes (especially those in unexpected places) modify the forecast map of the Seismicity parent forecast, and thus any GEAR forecast; however, after their aftershocks have died out, it might be a very long time until the next large earthquake in that area. Thus, it is conceivable that very-long-term seismicity (*e.g.*, 100 years into the future) might be overpredicted in some intraplate regions. This is an open question, as many previous seismic-hazard models, created in other ways, have also anticipated elevated hazard for two or more centuries following famous historic earthquakes. By omitting any stated expiration-date for the current GEAR1 forecast, we do not mean to guarantee that there is no such date; we only note that this is a complex question which cannot yet be answered.

USE OF GEAR1 FOR CATASTROPHE BONDS

GEAR1 can easily be used to calculate the earthquake magnitude for which there is a 1% (or any) annual probability of occurrence in circles of 100-km (or any) radius, and so can be used to estimate the risk of triggering a parametric catastrophe bond (Franco, 2010) payment based on this criterion. A global map of this type is shown in ***Figure 7***. If the radius of integration circles were increased, all magnitudes would rise. Importantly, this map refers to epicentroids, rather than ends of ruptures which might, or might not, extend into a given integration circle. On the basis of published USGS procedures, authoritative earthquake centroid locations and magnitude assignments are routinely reported for global earthquakes by the USGS 'ComCat' within minutes to hours, and are fixed and finalized six weeks after the mainshock. Thus, both the estimate of the likelihood of the trigger, and the timely confirmation of its occurrence, can be fully, unambiguously, and transparently specified.

GEAR1 could therefore serve as a basis for catastrophe bonds, in which investors receive a high rate of interest on their principal until and unless the specified earthquake strikes, in which case they would lose their principal. A GEAR-based bond could open the market to quake-threatened developing nations, and creating new and more diversified opportunities for investors. There could be composite global bonds, or many smaller bonds or reinsurance securities customized for the regions of interest to investors (those taking the risk) and cedants (those reducing their risk). Ultimately, GEAR1 could be an efficient and transparent platform for the exchange of financial risk.

COMPARISON TO REGIONAL FORECASTS: California

One purpose of this global seismicity model is to provide first-order estimates of seismicity in regions that lack their own regional seismic-hazard programs. Another purpose is to initiate comparisons with detailed national and regional models created by other methods. Naturally, many seismologists will regard these comparisons as tests of GEAR1. We advocate a more neutral approach: Large differences between GEAR1 and regional forecasts (if not readily explained by differences in format or data scope, or simple explanations based on the temporal limitations of GEAR) should lead to further investigation of both GEAR1 and these other independent forecasts. In any case, future prospective testing of these competing forecasts should be conducted because of its very low marginal cost.

Our GEAR1 forecast does not use any database of active faults. However, many regional models do use fault traces, and sometimes associated slip rates. Thus, one expected difference is that the GEAR1 forecast is likely to be spatially smoother, and lack sharp maxima along traces of active faults. The Tectonics parent of the GEAR1 forecast was based on an approximation (Bird & Kreemer, 2015) that secular strain-rates recorded by GPS (or implied by relative plate rotation) are good proxies for long-term tectonic strain; however, interseismic elastic strain accumulation is known to be spatially smoother than eventual seismic strain release. The Seismicity parent of the GEAR1 forecast is also necessarily smooth because its source catalog (GCMT, *m*5.767+, 1977-2013) only captures a modest number of earthquakes in most regions, and these point sources must be spatially smoothed to provide an optimized forecast of future seismicity. For example, in the California-centric rectangle defined by limits [126°W ≤ longitude ≤ 114°W] and [32°N ≤ latitude ≤ 42°N], only 52 such earthquakes have been recorded by GCMT. Because of this contrast in resolution, it may be most valuable to compare overall seismicity rates and patterns of low spatial frequency (such as those obtained by smoothing the detailed regional forecast).

Another expected difference is that many regional models refer to past earthquakes inferred from analog-instrumental catalogs, from historical catalogs, or from paleoseismic field studies. But GEAR1 uses no data regarding events before 1977. In the U.S.A., a prominent example is that the National Seismic Hazard Maps (*e.g.*, Petersen *et al.*, 2008) show high forecast hazard around the epicenters of the 1811-1812 earthquakes in the area of New Madrid, MO, but GEAR1 does not forecast high seismicity there. In such cases, a higher forecast seismicity in the regional model is easily understood, although it is still subject to prospective testing. However, any difference in which the regional model projects a lower overall seismicity than GEAR1 should be investigated; it may be found to depend critically on a questionable assumption.

Here we present a brief comparison of GEAR1 to the Unified California Earthquake Rupture Forecast version 3 (UCERF3) of Field *et al.* (2013) which is widely considered to be one of the most technically complex regional forecasts. This model used an expanded database of active faults, not limited to faults with measured geologic slip rate. Its logic-tree considered 4 alternative deformation models, with 70% total weight on a set of 3 kinematically self-consistent deformation models that merged geologic, geodetic, and plate-tectonic constraints. Also, it simulated the earthquake-rupture process in detail in order to include multi-fault ruptures, creating thousands of virtual catalog realizations, constrained by seismic catalogs, fault slip rates from the deformation models, and geologic recurrence intervals. Both forecasts of long-term epicentroid rate density are presented for comparison in ***Figure 8***. The GEAR1 forecast has been windowed to display only the area of 7.50×10^{11} m^2 that is also covered by UCERF3.

At magnitude threshold *m*5.8+, these two forecasts anticipate very similar total earthquake rates: 121 epicentroids/century in GEAR1, and 126 epicentroids/century in UCERF3. The UCERF3 forecast has higher spatial variance; if we divide the spatial standard deviation of each forecast by its respective mean rate, these relative standard deviations are 155% for GEAR1 but 181% for UCERF3. Consistent with this, the $\boldsymbol{I}_0$ specificities are 0.896 for GEAR1 (in the California region of Figure 8) but 1.069 for UCERF3. Both statistics confirm the visual impression that the UCERF3 forecast seismicity is more strongly concentrated along traces of modeled faults. The correlation coefficient between these two forecasts is 0.482. However, we also tried smoothing the UCERF3 forecast and then re-computing correlations of these smoothed versions of UCERF3 with (unchanged) GEAR1; we found that the correlation coefficient rises smoothly to a maximum of 0.625 when the smoothing is done by convolution with a 2-D Gaussian bell-curve function of characteristic length 30 km. The specificity of this particular smoothed version of UCERF3 would drop to 0.608, which is actually below the local specificity of GEAR1.

At threshold magnitude *m*7.0+, the results are similar. The spatially-integrated total rates are 7.64 epicentroids/century for GEAR1 and 7.49 epicentroids/century for UCERF3. The relative standard deviation is stable at 159% for GEAR1, but rises to 224% for UCERF3. Specificity $\boldsymbol{I}_0$ is stable at 0.909 for GEAR1 (in California) but rises to 1.755 for UCERF3. Both of these statistics indicate an even stronger concentration of UCERF3 seismicity on modeled faults at threshold *m*7+. The correlation coefficient between the two models is 0.462, but this rises to a peak of 0.600 when the UCERF3 model is smoothed using a characteristic length of 25 km; this same amount of smoothing would also lower the UCERF3 specificity to 1.024 which is not much more than the local specificity of GEAR1.

Thus, these two forecasts have strong similarities, but the UCERF3 forecast provides a sharper focus because it was based on traces of known active faults, while GEAR1 was not. The ideal level of forecast smoothness is currently uncertain, and needs to be tested and optimized in future prospective experiments. A formal prospective test of all recent California forecasts, also including those of Marzocchi *et al.* (2012) and of Hiemer *et al.* (2013) and of Rhoades *et al.* (2013, 2014), would be valuable, even though a lengthy duration (*e.g.,* 50~200 years) will probably be required for conclusive ranking of all these models.

It is worth noting that the plausibility of GEAR1 seen in this California comparison may depend strongly on the very widespread and precise network of GPS observations in the region, which both models incorporate, although in different ways. Unless both of these forecasts are contradicted by future seismicity, this comparison leaves an impression that geodetic observation may partially substitute for full knowledge of active fault locations and rates, at least for applications in which the precise locations of future ruptures are not required.

CONCLUSIONS AND PROSPECTS

This project has succeeded in merging disparate long-term seismicity models into testable global forecasts of long-term shallow seismicity, and has made a start on testing them, retrospectively. We find that multiplicative blends of smoothed-seismicity and tectonic forecasts outperform linear blends. The improvement in information score is large, and quite unlikely to be due to one-time random fluctuations in seismicity. It is encouraging that our preferred model, though chosen for its improved performance in forecasting catalog years 2005-2012, also outperforms previous methods in forecasting catalog years 1918-1976. Furthermore, a local comparison to the recent UCERF3 long-term forecast in California shows that both anticipate the same overall earthquake rates, with the map-pattern of our GEAR1 model closely resembling a smoothed version of the map-pattern of UCERF3.

In the near future, this GEAR1 forecast will be submitted for independent prospective testing at CSEP; preliminary results should be available after only one year of testing because of its global scope. Assuming success similar to that we have seen retrospectively, others may wish to build rupture models and seismic-hazard models based on GEAR1, by supplementing its maps of epicentroid rate density with specific fault sources (where known) or focal mechanisms (elsewhere), with rupture depths and extents, and with attenuation relations. It will be important to add supplemental data (and/or assumptions) about the depths of shallow ruptures; GEAR1 has made no distinctions between earthquakes within its depth range of 0~70 km because of the limitations of available test catalogs; however, a rupture model built from GEAR1 would need to be more precise. It will also be important to make policy decisions regarding whether historical and/or paleoseismic events (like those around New Madrid, MO in the U.S.A.) should result in locally-elevated model hazard, despite the absence of complete and consistent global databases of historical and/or paleoseismic events, and the absence of rigorous prospective testing of related hypotheses. Another possibility for future development is that the availability of transparent estimates of the occurrence of large shallow earthquakes in specific local regions could contribute to greater trade in parametric catastrophe bonds. Looking beyond GEAR1 to potential future versions, there is an opportunity for further improvement by incorporating seismic catalog years before 1977 into the smoothed-seismicity parent forecast, and by incorporating new geodetic data and revised plate models into the tectonic parent forecast.

DATA AND RESOURCES

The source code and data files used to create the Tectonics parent forecast were described by Bird and Kreemer (2015). These same data files are needed to compute the GEAR1 hybrid model, although the application code is different. These file-names are listed in the small parameter file GEAR1_parameters.dat (available in the electronic supplement to this article).

The only dataset used to compute the Seismicity parent forecast was the Global Centroid Moment Tensor catalog. We provide this parent forecast, for years 2014 and after, as the large (439 MB) ASCII table file GL_HAZTBLT_M5_B2_2013.TMP (available in the electronic supplement to this article, in a compressed .zip format occupying 56 MB).

Our GEAR1 forecast is provided in the form of Fortran 90 source code GEAR1_for_CSEP.f90, available in the electronic supplement to this article. This is an extension and expansion of program SHIFT_GSRM2f_for_CSEP.f90 described and published by Bird and Kreemer (2015). A compiled 64-bit executable for Windows is available from the first author. This program will produce a 3.7 GB file containing a global grid of 0.1° x 0.1° cells, with forecast shallow seismicity of each cell divided into 31 magnitude bins ranging from m = 6.00±0.05 in steps of 0.10 up to the final open-ended bin *m*8.95+, in the XML format required by CSEP. Utility program XML_2_GRD, available from the web site of the first author, can be used to extract a spatial grid for any desired threshold magnitude from *m*5.75+ to *m*9.15+. Another utility program, extract_regional_GRD, can be used to extract a rectangular subregion at the same threshold magnitude. GRD file format is documented at http://peterbird.name/guide/grd_format.htm (last visited February 2015). The website of the first author also provides a mapping tool (NeoKineMap) and two forecast-scoring tools (Kagan_2009_GJI_I_scores, and pseudoCSEP) that work with this GRD file format, and which were used in this study to create maps and tables, respectively.

ACKNOWLEDGEMENTS

Graeme Weatherill performed PSHA hazard calculations with a preliminary version of this model, helping greatly to define its strengths and limitations. Mark Stirling organized a Powell Conference on this model and advised on its construction, as did Martin Käser, Harold Magistrale, Morgan Page, Yufang Rong, and Graeme Weatherill. Edward Field and Peter Powers provided a gridded-seismicity forecast for the California region based on UCERF3. Edward Field, Wayne Thatcher, Mark Stirling, and an anonymous reviewer provided thoughtful reviews leading to clarification. We gratefully acknowledge the U.S. Geological Survey (USGS) John Wesley Powell Center for Analysis and Synthesis, where the GEAR project received its first external review, as well as financial support from the USGS, the Global Earthquake Model (GEM) Foundation, the National Science Foundation (NSF), the Southern California Earthquake Center (SCEC), and the University of California Los Angeles (UCLA). DDJ and YYK received support from NSF through grants EAR-0944218 and EAR-1045876. This research was also supported by SCEC, which is funded by NSF Cooperative Agreement EAR-1033462 and USGS Cooperative Agreement G12AC20038. The SCEC contribution number for this paper is 2075. Publication was supported by the USGS. All opinions, expressed or implied, are those of the authors and do not reflect official positions of the USGS, GEM Foundation, NSF, SCEC, or UCLA.

REFERENCES CITED

Bird, P. (2009). Long-term fault slip rates, distributed deformation rates, and forecast of seismicity in the western United States from fitting of community geologic, geodetic, and stress direction datasets, *J. Geophys. Res.*, **114**, B11403, doi: 10.1029/2009JB006317.

Bird, P., and Y. Y. Kagan (2004). Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings, *Bull. Seismol. Soc. Am.*, **94**(6), 2380-2399.

Bird, P., and C. Kreemer (2015). Revised tectonic forecast of global shallow seismicity based on version 2.1 of the Global Strain Rate Map, *Bull. Seismol. Soc. Am.*, 105(1), 152-166, doi: 10.1785/0120140129.

Bird, P., and Z. Liu (2007). Seismic hazard inferred from tectonics: California, in: S. E. Hough and K. B. Olsen (ed.), Special Issue on: Regional Earthquake Likelihood Models, *Seismol. Res. Lett.*, **78**(1), 37-48.

Bird, P., Y. Y. Kagan, and D. D. Jackson (2002). Plate tectonics and earthquake potential of spreading ridges and oceanic transform faults, in: S. Stein and J. T. Freymueller (editors),

Plate Boundary Zones, *Geodynamics Series*, **30**, 203-218, Am. Geophys. U., Washington, DC.

Bird, P., Y. Y. Kagan, D. D. Jackson, F. P. Schoenberg, and M. J. Werner (2009). Linear and nonlinear relations between relative plate velocity and seismicity, *Bull. Seismol. Soc. Am*., **99**(6), 3097-3113, doi: 10.1785/0120090082.

Bird, P., C. Kreemer, and W. E. Holt (2010a). A long-term forecast of shallow seismicity based on the Global Strain Rate Map, *Seismol. Res. Lett*., **81**(2), 184-194, plus electronic appendices.

Bird, P., Y. Y. Kagan, and D. D. Jackson (2010b). Time-dependent global seismicity forecasts with a tectonic component: Retrospective tests (abstract), *Eos Trans. Am. Geophys. U.*, **91**(52), Fall Meeting Supplement, Abstract S44B-03.

Di Giacomo, D., I. Bondár, D. A. Storchak, E. R. Engdahl, P. Bormann, and J. Harris (2015). ISC-GEM: Global instrumental earthquake catalogue (1900-2009), III: Re-computed M_S and m_b , proxy M_W , final magnitude composition and completeness assessment, *Phys. Earth Plan. Int., **239**(2), 33-47, doi: 10.1016/j.pepi.2014.06.005*.

Ekström, G., M. Nettles, and A. M. Dziewonski (2012). The global CMT project 2004-2010: Centroid moment tensors for 13,017 earthquakes, *Phys. Earth Planet. Inter.*, **200-201**, 1-9.

Field, E. H. (2007). Overview of the Working Group for the Development of Regional Earthquake Likelihood Models (RELM), in: S. E. Hough and K. B. Olsen (ed.), Special Issue on: Regional Earthquake Likelihood Models, *Seismol. Res. Lett.*, **78**(1), 7-16.

Field, E. H., G. P. Biasi, P. Bird, T. E. Dawson, K. R. Felzer, D. D. Jackson, K. M. Johnson, T. H. Jordan, C. Madden, A. J. Michael, K. R. Milner, M. T. Page, T. Parsons, P. M. Powers, B. E. Shaw, W. R. Thatcher, R. J. Weldon, II, and Y. Zeng (2013). Unified California Earthquake Rupture Forecast, version 3 (UCERF3)-The time-independent model, *U.S. Geol. Surv. Open-File Rep.*, **2013-1165** (*Cal. Geol. Surv. Spec. Rep.* **228**, and *Southern California Earthquake Center Pub.* **1792**), 97 pages; http://pubs.usgs.gov/of/2013/1165/.

Franco, G. (2010). Minimization of trigger error in cat-in-a-box parametric earthquake catastrophe bonds with an application to Costa Rica, *Earthquake Spectra*, **26**, 983-998, doi: 10.1193/1.3479932.

Hiemer, S., D. D. Jackson, Q. Wang, Y. Y. Kagan, J. Woessner, J. D. Zechar, and S. Wiemer (2013). A stochastic forecast of California earthquakes based on fault slip and smoothed seismicity, *Bull. Seismol. Soc. Am.*, **103**(2A), 799-810, doi: 10.1785/0120120168.

Jackson, D. D., and Y. Y. Kagan (1993). Reply [to Nishenko & Sykes], *J. Geophys. Res.*, **98**(B6), 9917-9920: doi: 10.1029/93JB00699.

Jackson, D. D., and Y. Y. Kagan (1999). Testable earthquake forecasts for 1999, *Seismol. Res. Lett.*, **70**, 393-403.

Jackson, D. D., and Y. Y. Kagan (2012). 10 (abstract), Amer. Geophys. U. Fall Meeting 2012 abstract #S24A-08.

Kagan, Y. Y. (2003). Accuracy of modern global earthquake catalogs, *Phys. Earth Planet. Int.*, **135**(2-3), 173-209, doi:10.1016/S0031-9201(02)00214-5.

Kagan, Y. Y. (2009). Testing long-term earthquake forecasts: Likelihood methods and error diagrams, *Geophys. J. Int*., **177**(2), 532-542.

Kagan, Y. Y., and D. D. Jackson (1991). Seismic gap hypothesis: Ten years after, *J. Geophys. Res.*, **96**(B13), 21,419-21,431.

Kagan, Y. Y., and D. D. Jackson ([1994). Long-term probabilistic forecasting of earthquakes, J. *Geophys. Res.*, **99**, 13,685-13,700.

Kagan, Y. Y., and D. D. Jackson (2000). Probabilistic forecasting of earthquakes, *Geophys. J. Int*., **143**, 438-453.

Kagan, Y. Y. and D. D. Jackson (2011). Global earthquake forecasts, *Geophys. J. Int.*, **184**(2), 759-776, doi: 10.1111/j.1365-246X.2010.04857.x

Kagan, Y. Y. and D. D. Jackson (2013). Tohoku earthquake: A surprise?, *Bull. Seismol. Soc. Am.*, **103**(2B), 1181-1194.

Kagan, Y. Y., P. Bird, and D. D. Jackson (2010). Earthquake patterns in diverse tectonic zones of the globe, *Pure Appl. Geoph.*, **167**(6/7), 721-741.

Kelleher, J., L. R. Sykes, and J. Oliver (1973). Possible criteria for predicting earthquake locations and their application to major plate boundaries of the Pacific and the Caribbean, *J. Geophys. Res.*, **78**, 2547-2585.

Kreemer, C., Blewitt, G., and Klein, E.C. (2014). A geodetic plate motion and Global Strain Rate Model, *Geochem. Geophys. Geosyst*., 15(10), 3849-3889, doi: 10.1002/2014GC005407.

Marzocchi, W., J. D. Zechar, and T. H. Jordan (2012). Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. Seismol. Soc. Am.*, **102**(6), 2574-2584.

McCaffrey, R. (2008). Global frequency of magnitude 9 earthquakes, *Geology*, **36**(3), 263-266, doi: 10.1120/G24402A.1.

McCann, W. R., S. P. Nishenko, L. R. Sykes, and J. Krause (1979). Seismic gaps and plate tectonics: Seismic potential for major boundaries, *Pure Appl. Geoph*., **117**, 1082-1147.

Michael, A. J. (2014). How complete is the ISC-GEM Global Earthquake Catalog?, *Bull. Seismol. Soc. Am.*, **104**(4), 1829-1837.

Nishenko, S. P., and L. R. Sykes (1993). Comment on “Seismic gap hypothesis: Ten years after” by Y. Y. Kagan and D. D. Jackson, *J. Geophys. Res.*, **98**(B6), 9909-9916.

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, N. Luco, E. H. Field, C. J. Wills, and K. S. Rukstales (2008). Documentation for the 2008 update of the United States National Seismic Hazard Maps, *U.S. Geol. Surv. Open-File Rep*., **2008-1128**, 128 pages.

Rhoades, D. A., and M. C. Gerstenberger (2009). Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.*, **99**(2A), 636-646.

Rhoades, D. A., M. C. Gerstenberger, A. Christopherson, J. D. Zechar, D. Schorlemmer, M. J. Werner, and T. H. Jordan (2013). Multiplicative hybrids of models from the five-year RELM experiment (abstract), Southern California Earthquake Center, Annual Meeting 2013 Program, poster 265.

Rhoades, D. A., M. C. Gerstenberger, A. Christophersen, J. D. Zechar, D. Schorlemmer, M. J. Werner, and T. H. Jordan (2014). Regional Earthquake Likelihood Models II: Information Gains of Multiplicative Hybrids, Bull. Seismol. Soc. Amer., 104(1), doi: 10.1785/0120140035.

Schorlemmer, D., J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, and the RELM Working Group (2010). First results of the Regional Earthquake Likelihood Models experiment, *Pure Appl. Geoph.*, **167**, 859-876, doi: 10.1007/s00024-010-0081-5.

Storchak, D. A., D. Di Giacomo, I. Bondar, J. Harris, E. R. Engdahl, W. H. K. Lee, A. Villasenor, P. Bormann, and G. Ferrari (2012). ISC-GEM Global Instrumental Earthquake Catalog, 1900-2009, *GEM Technical Report*, **2012-01**, 128 pages, doi: 10.13117/GEM.GEGD.TR2012.01, Global Earthquake Model Foundation, Pavia, Italy.

Taroni, M., J. D. Zechar, and W. Marzocchi (2013). Assessing global M6+ seismicity forecasts, *Geophys. J. Int.*, **136**, 422-432, doi: 10.1093/gji/ggt369.

Zechar, J. D. (2008). Methods for evaluating earthquake predictions, Ph.D. dissertation, U. Southern California, Los Angeles.

Zechar, J. D., and T. H. Jordan (2008). Testing alarm-based earthquake predictions, *Geophys. J. Int.*, **172**, 715-724, doi: 10.1111/j.1365-246X.2007.03676.x.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010). Likelihood-based tests for evaluation space-rate-magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.*, **100**(3), 1184-1195, doi: 10.1785/0120090192.

AUTHORS

Peter Bird; Department of Earth, Planetary, and Space Sciences; University of California; Los Angeles, CA 90095-1567.

David D. Jackson; Department of Earth, Planetary, and Space Sciences; University of California; Los Angeles, CA 90095-1567.

Yan Y. Kagan; Department of Earth, Planetary, and Space Sciences; University of California; Los Angeles, CA 90095-1567.

Corné Kreemer; Nevada Bureau of Mines and Geology; Seismological Laboratory; University of Nevada; 1664 North Virginia Street; Reno, NV 89557-0178.

Ross S. Stein; U.S. Geological Survey; Office 3A113; 345 Middlefield Road, MS 977; Menlo Park, CA 94025.

TABLE

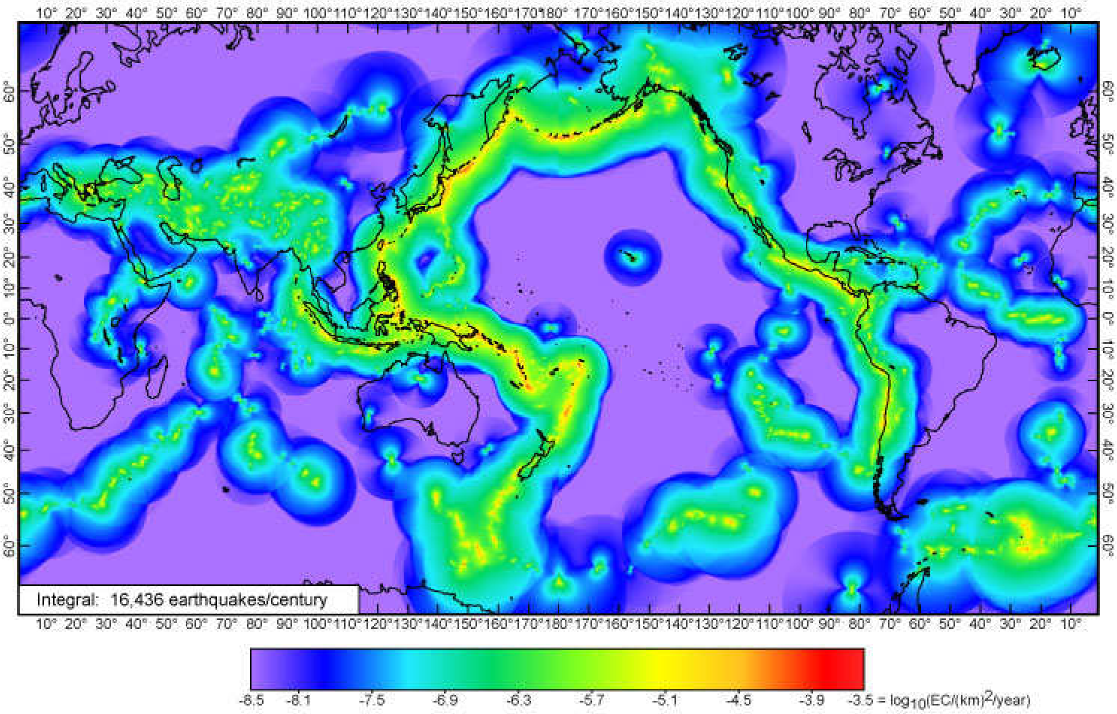
Table 1. Global shallow earthquake rates *(R)*, per century:

Catalog or model	*m*5.767+	*m*7+	*m*8+	*m*9+
ISC-GEM 1918-1976	N/A*	942	80	3
GCMT 1977-2013	17503	951	65	5
merged catalogs 1918-2013	N/A*	946	74	4
GEAR1, for 2014+	17589	1087	92	5
Seismicity, for 2014+	17647	1043	85	4
Tectonics, for 2014+	17503	1155	103	6

*N/A = Not Available (catalog incomplete).

FIGURES & CAPTIONS

(a)



(b)

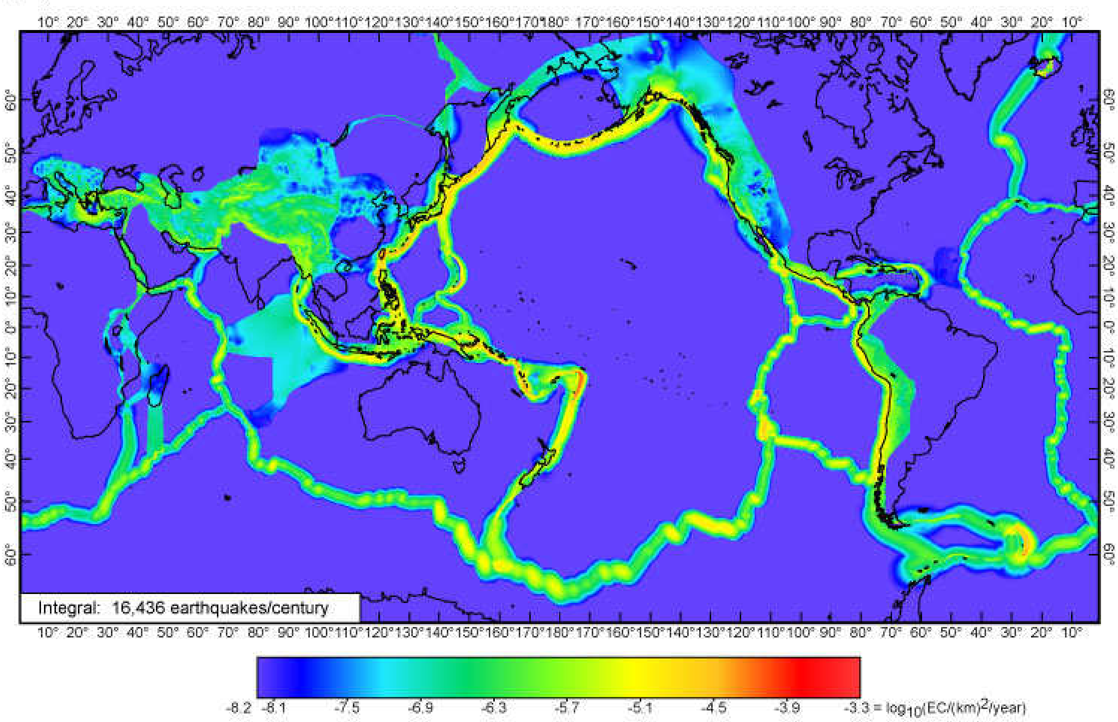
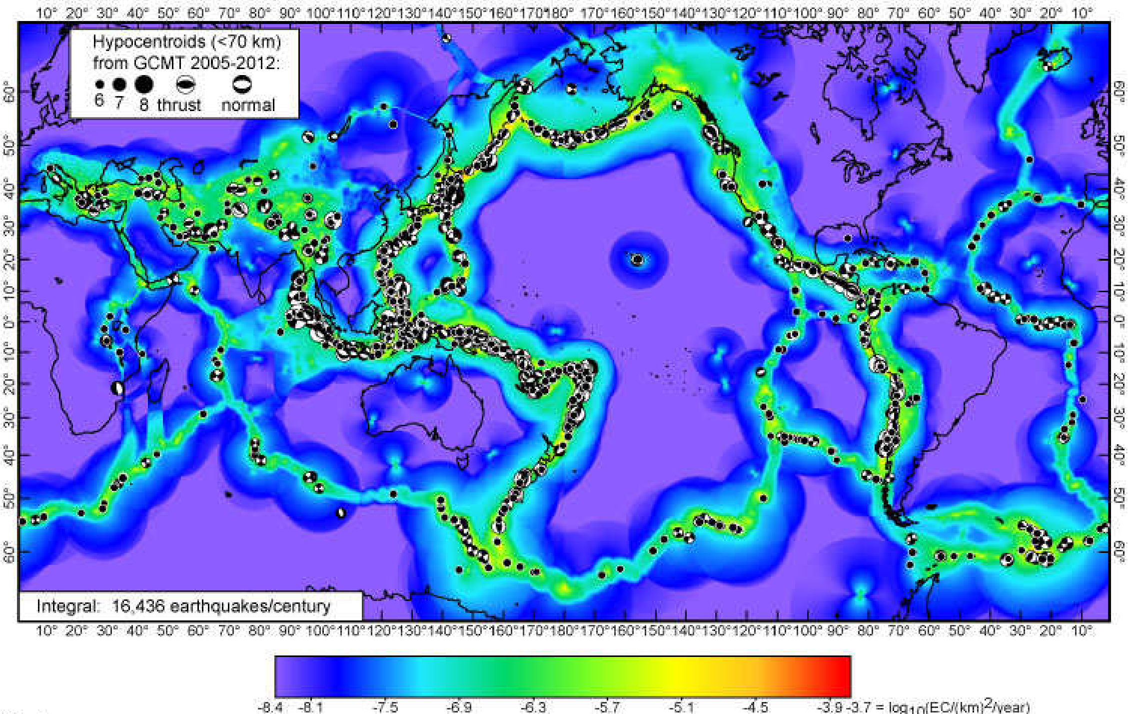


Figure 1. Two parent forecasts with threshold magnitude *m*5.767+: (a) Seismicity parent forecast for years 2005+. Mercator projection. Logarithmic color- (or gray-) scale shows the rate density of epicentroids corresponding to shallow (≤70 km) hypocentroids, in units of $(km)^{-2}$ $year^{-1}$. (b) Tectonics parent forecast for years 2005+. Conventions as in part (A), and identical color- (or gray-) scale. Equal to model SHIFT-GSRM2f of Bird and Kreemer (2015), except that its original spatial grid of 0.25 × 0.20-degree cells is here resampled to a finer 0.1 × 0.1-degree grid. Colored maps appear in the online version.

.

(a)



(b)

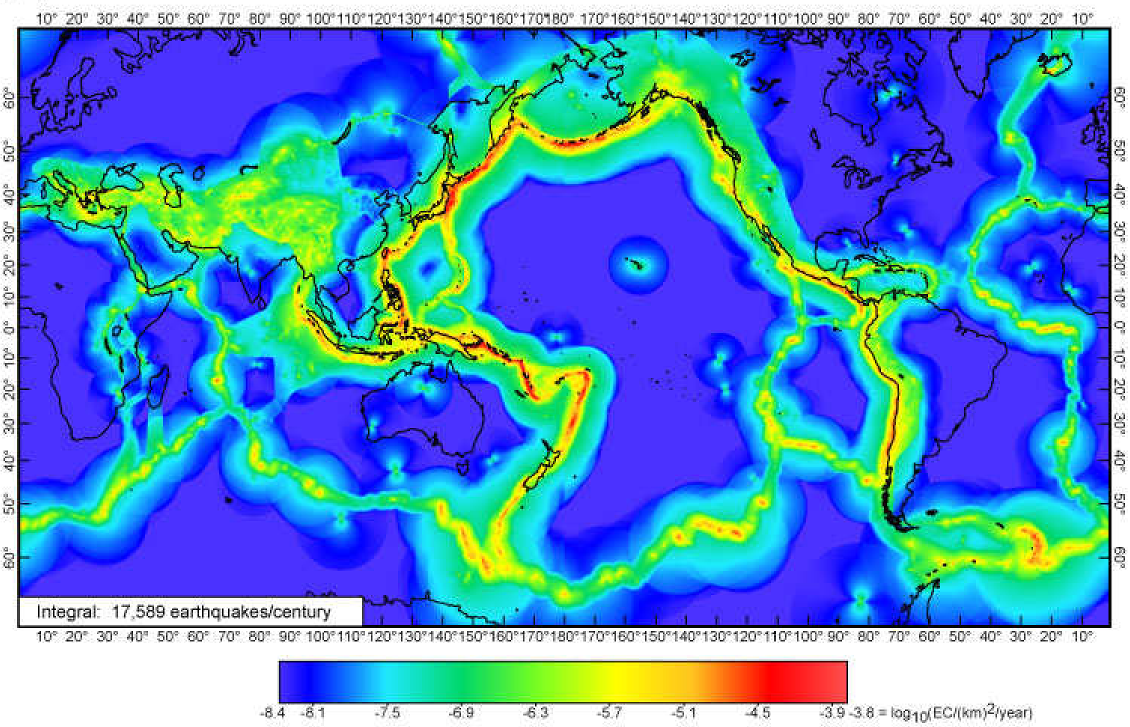


Figure 2. Preferred hybrid forecasts for threshold magnitude *m*5.767+, both with and without overlay of test earthquakes: (a) Preferred hybrid forecast H* (log-linear, with exponent *d* = 0.6 on Seismicity)

for years 2005+ compared to 1694 shallow test earthquakes from GCMT catalog years 2005-2012. For test earthquakes of $m > 6$, focal mechanism is shown on lower focal hemisphere. Scores from this comparison (and many others) are shown in Table S1 (available in the electronic supplement to this article) and Figure 3. (b) GEAR1 forecast (preferred hybrid H*, updated to end-2013) for years 2014 and after. Mercator projection. Logarithmic color- (or gray-) scale shows the rate density of epicentroids corresponding to shallow (≤70 km) hypocentroids, in units of $(km)^{-2}$ $year^{-1}$. Colored maps appear in the online version.

.

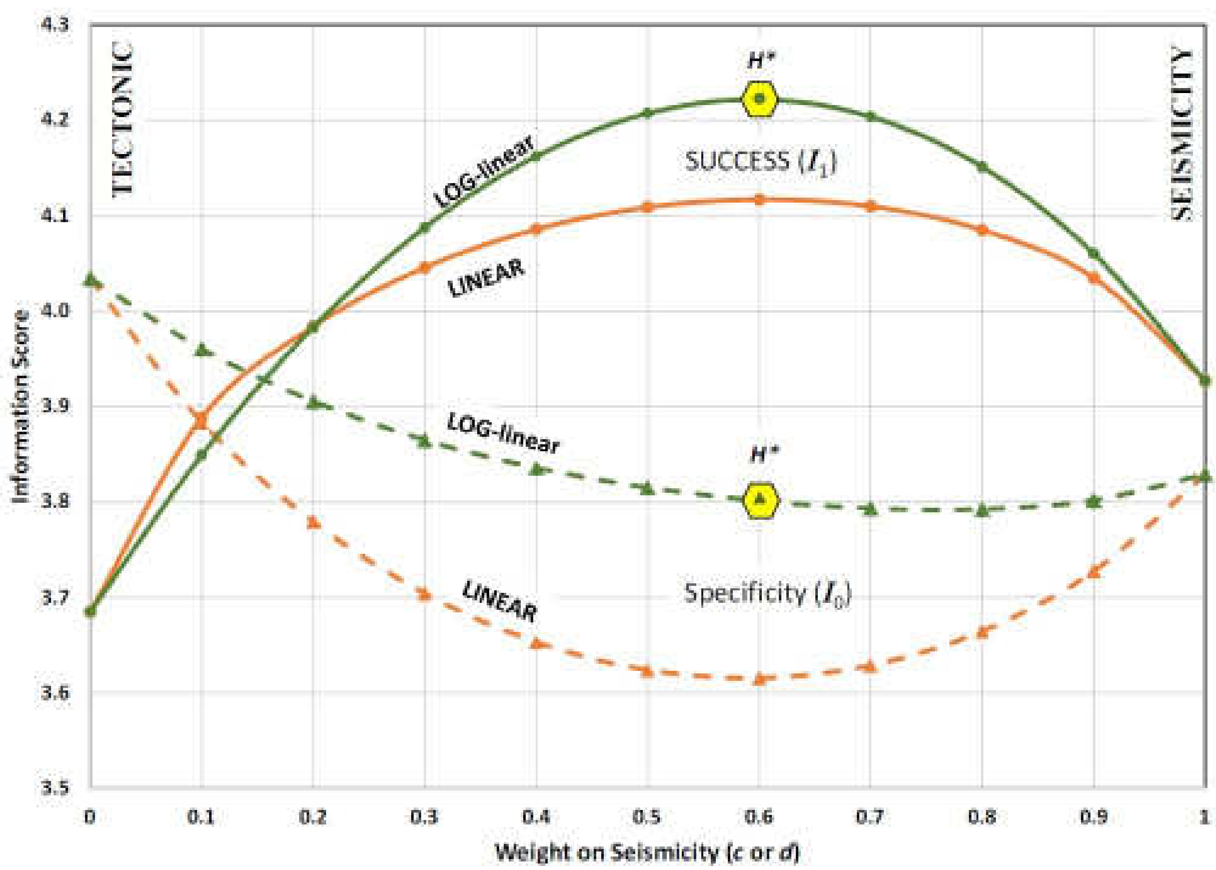


Figure 3. Success I_1 and specificity I_0 of both linear and log-linear hybrid models as a function of mixing parameter c or d, in tests against GCMT catalog years 2005-2012 at threshold m5.767+. Both

of these $\boldsymbol{I}$ information scores were defined by Kagan [2009]. The preferred hybrid model H* is highlighted.

.

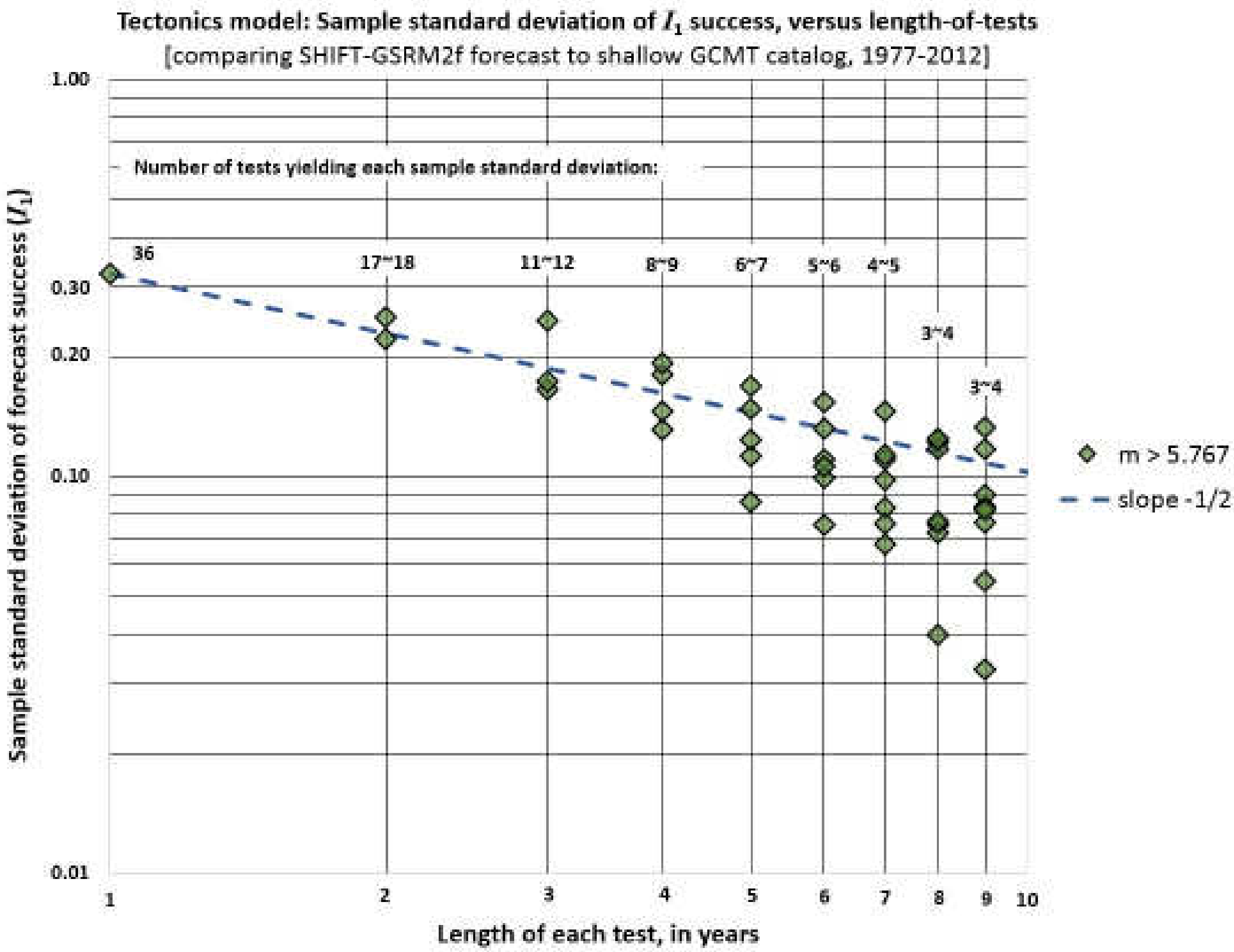


Figure 4. Small-sample standard deviations (diamonds) of the success of the Tectonics forecast, σ($\boldsymbol{I}_1$(T)), as a function of test window length *W,* in the range of 1 to 9 years. Based on multiple subdivisions (with re-use) of GCMT catalog years 1977-2012. Dashed line with slope -1/2 appears consistent with these bootstrap experimental results. There are 2 points at *W* = 2 years, 3 points at *W* = 3 years, *etc.*, because these longer windows can be defined using *W* different start-years. The number of scores compared to compute each sample standard deviation decreases, with increasing *W*, from 36 to 3~4, which explains

the increasing scatter of these sample standard deviations.

881 .

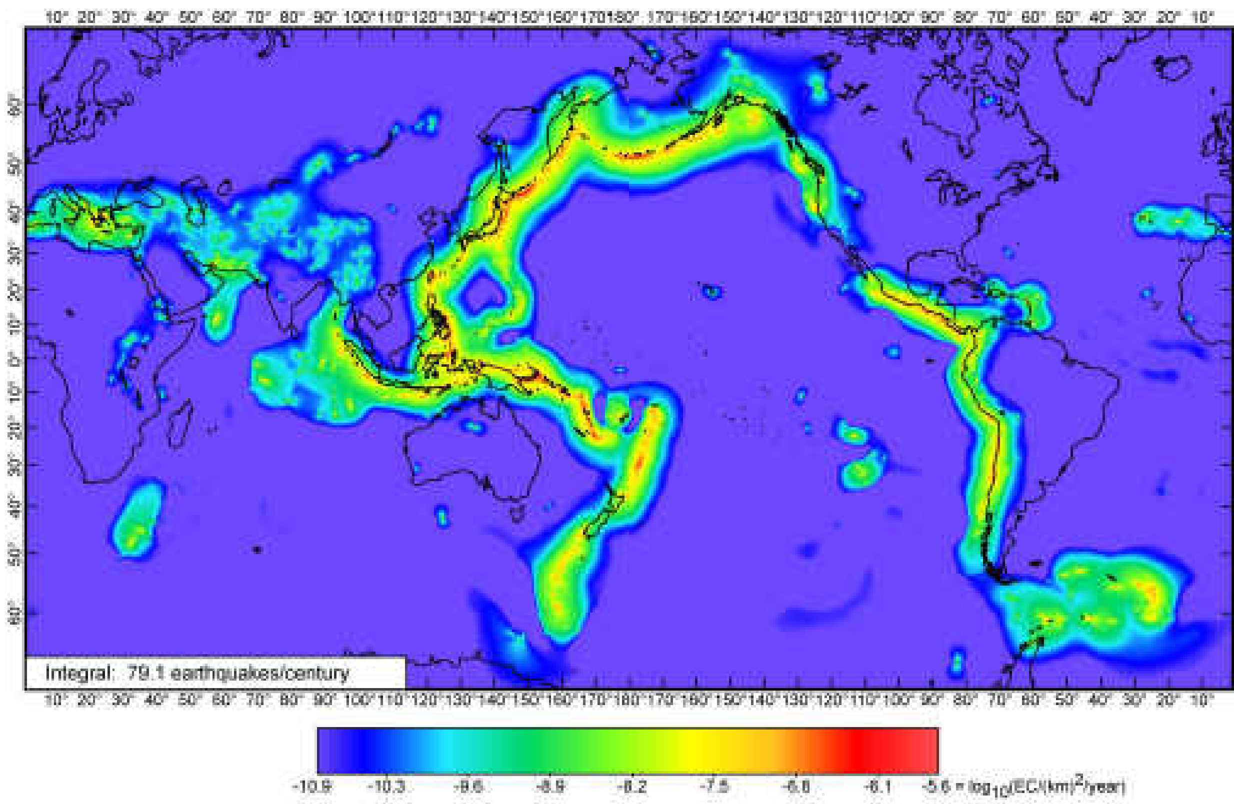


Figure 5. Extrapolation of the Seismicity parent forecast (from Figure 1A, for years 2005+) to threshold *m*8+. Epicentroid rate density of each cell was extrapolated with the tapered Gutenberg-Richter frequency/magnitude distribution (8) using corner magnitudes and spectral slopes based on Table 1 of Kagan *et al.* (2010) and the tectonic zone map of the same paper, but only after edits and smoothing had been applied to the maps of corner magnitude and spectral slope, as described in text. Most spreading ridges have disappeared from this map because their corner magnitudes are less than the *m*8+ threshold. A few active spots remain where oceanic transform slip is transpressive; because slip-partitioning into thrust earthquakes is expected, these regions were assigned to tectonic zone 4 by Kagan *et al.* (2010). Colored map appears in the online version.

882 .

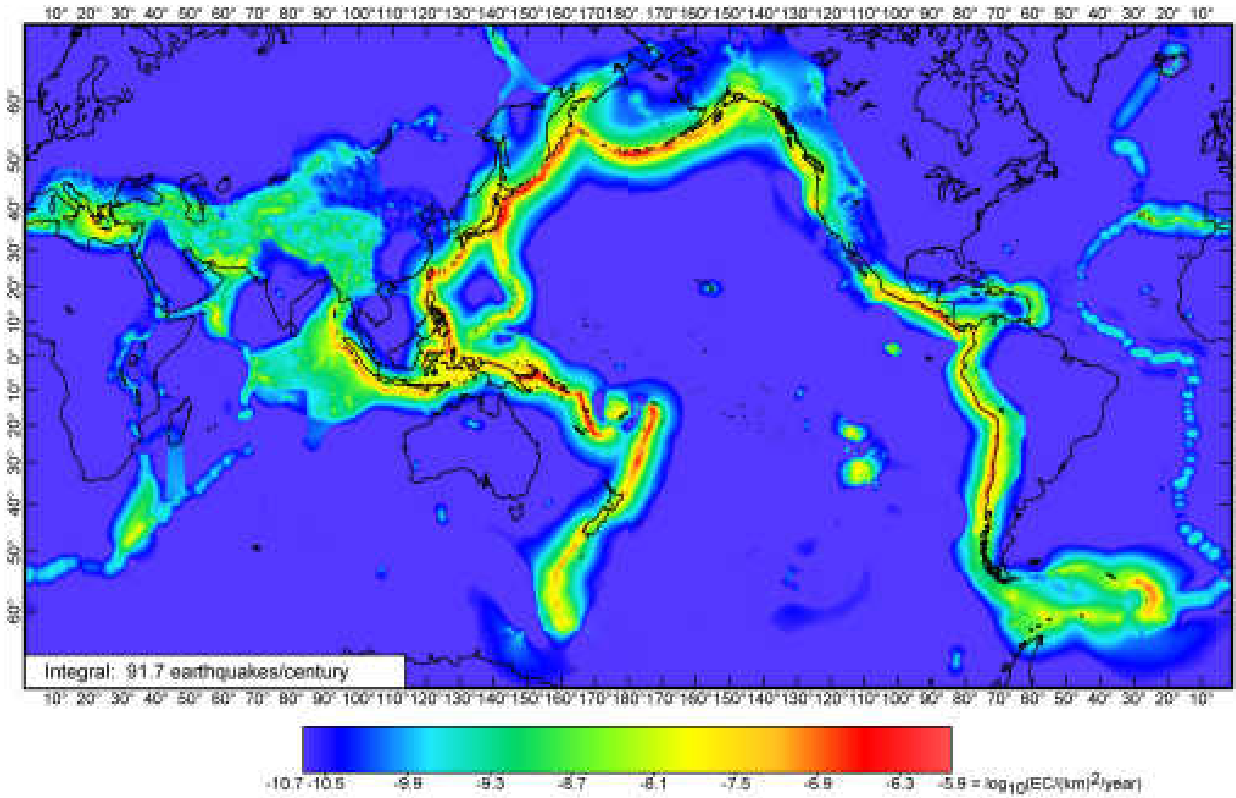


Figure 6. GEAR1 forecast for threshold magnitude *m*8+ and for years 2014 and after. Conventions as in Figure 2B. Global earthquake rate is based on (9). This map has strong similarities to the parent Seismicity forecast of Figure 5, but also reflects the influence of the Tectonics parent forecast in its better depiction and resolution of plate boundary zones, and also the updating of both parent forecasts to the end of 2013. Colored map appears in the online version.

.

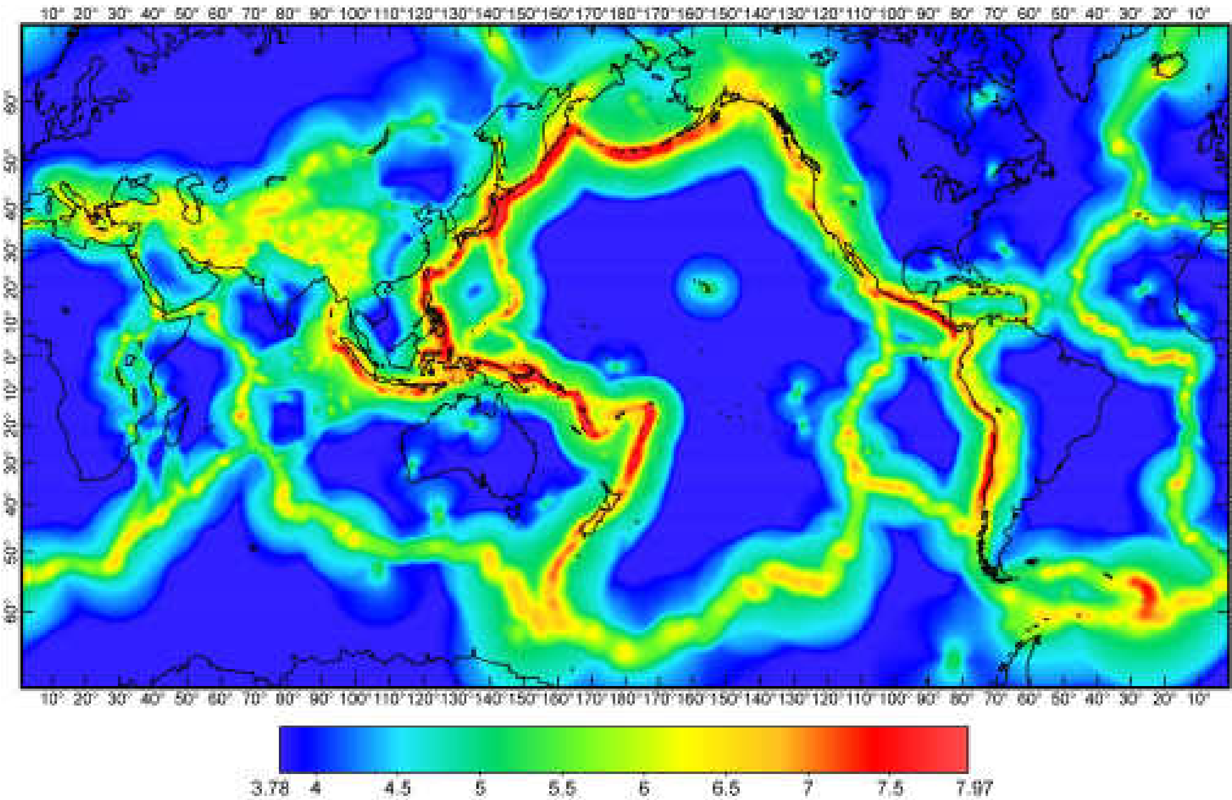


Figure 7. GEAR1 forecast for years 2014 and after, represented as the magnitude that has forecast epicentroid rate of 0.01/year (*i.e.*, probability of approximately 1% per year) within a local circle of radius 100 km about each test point. Colored map appears in the online version.

.

.

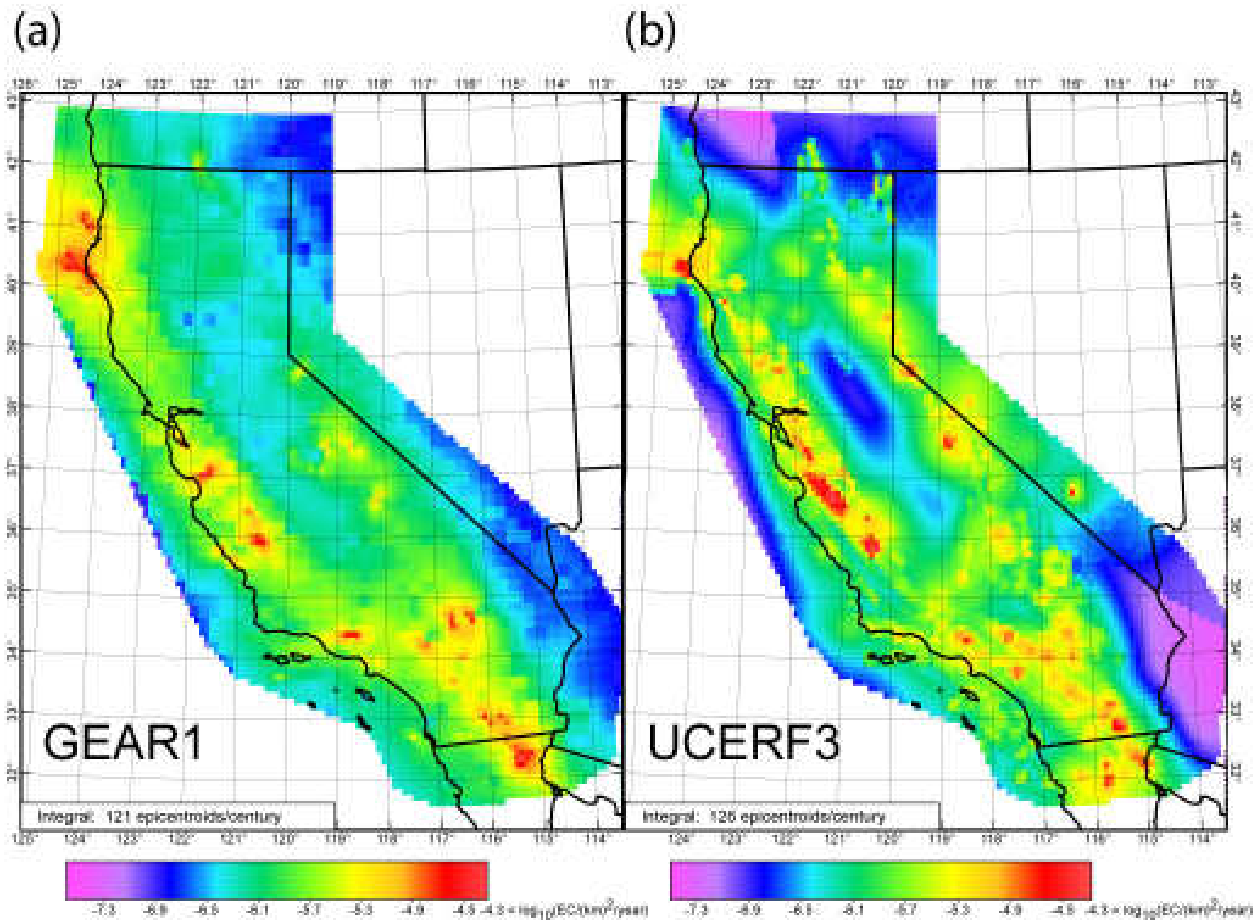


Figure 8. Comparison of: (a) GEAR1 long-term epicentroid rate densities in the California region at threshold magnitude *m*5.8+, with: (b) the branch-weighted mean time-independent seismicity forecast UCERF3 by Field *et al.* (2013) at the same threshold. The GEAR1 forecast has been windowed to match the area covered by UCERF3. Note that the UCERF3 forecast has not been smoothed for this figure, although smoothing is discussed in the text. Statistics of the comparison are presented in the text. Colored maps appear in the online version.

.

<u>Electronic Supplement</u> to:

GEAR1: a Global Earthquake Activity Rate model constructed from geodetic strain rates and smoothed seismicity

P. Bird, D. D. Jackson, Y. Y. Kagan, C. Kreemer, and R. S. Stein

These Electronic Supplement files include additional text and references on scoring of forecasts, tables of scoring results, and source code and data files needed to reproduce our forecast.

Table Captions:
Table S1: Results of 8-year retrospective tests against shallow events from GCMT, 2005-2012;
Table S2: One-year $\boldsymbol{I}_1$ success scores against shallow events in GCMT, 2005-2012;
Table S3: Results of 59-year retrospective tests against shallow ISC-GEM, 1918-1976;

Other:
*Fortran 90 source code in file GEAR1_for_CSEP.f90 (for computing GEAR1);
*GEAR1_parameters.dat (primary input file for computing GEAR1);
*Smoothed-seismicity parent forecast in file GL_HAZTBLT_M5_B2_2013.TMP.zip (56 MB compressed version of 349 MB ASCII file GL_HAZTBLT_M5_B2_2013.TMP).

Discussion of forecast-scoring metrics

Some seismic forecast test metrics consider only the number of earthquakes; some consider only the map-pattern; others consider both. When testing forecast numbers of earthquakes, it is critical to take into account earthquake clustering on all scales, which causes annual earthquake counts to have a distribution much broader than a Poisson distribution, which is probably best described by the negative-binomial distribution (Kagan, 2010). Such a revised number-test would be a valuable tool for testing forecasts of total seismicity. However, to date no such test is a recognized standard (*e.g.*, operational at the Collaboratory for the Study of

Earthquake Predictability, or CSEP). Instead, CSEP employs the N-test (Field, 2007; Schorlemmer and Gerstenberger, 2007; Schorlemmer *et al.*, 2007, 2010; Zechar *et al.*, 2010) which assumes that test earthquakes are mutually independent. This assumption of independence requires both the forecast and the test catalog to be declustered prior to testing. Unfortunately, there is no obvious method available for declustering our two parent forecasts or our hybrid forecasts, so we cannot use such tests. We also decline to emulate the L-test and R-test of CSEP (ibid), which have similar issues requiring declustering. Here, we will consider test metrics that compare only the map-patterns of forecast and test seismicity, independent of earthquake counts.

Our preferred measures of forecast specificity and success are two of the information scores $(\boldsymbol{I})$ defined by Kagan (2009). "Success" measure $\boldsymbol{I}_1$ is the mean, over all test earthquakes, of the base-2 logarithm of the ratio of conditional probability density in the cell in which the test earthquake epicentroid occurred, to the mean conditional probability density in the whole forecast region (which, in this paper, is the shallow part of the Earth). Here, "conditional probability" is the probability of an earthquake appearing with epicentroid at a particular (longitude, latitude) point, conditional on the occurrence of one new earthquake somewhere in the forecast domain with magnitude at or above threshold. Thus, $\boldsymbol{I}_1$ is the mean number of binary bits of information gain per actual test earthquake, over an ignorant model that has only a single global earthquake rate. "Specificity" $\boldsymbol{I}_0$ is the sum over all forecast cells of the normalized forecast rate times the base-2 logarithm of the ratio of normalized forecast rate to normalized cell area. Thus, $\boldsymbol{I}_0$ is the mean number of binary bits of information gain (per virtual, expected earthquake), over an ignorant model that has only a single global earthquake rate. Note that specificity I_0 does not require or use the test catalog, so it is an interesting descriptor of the forecast, but less important than success I_1. These information scores have several advantages: scores are independent of any difference between the total numbers of forecast and test earthquakes; no simulated virtual catalogs based on the forecast are needed; no random perturbations of the test catalog are needed; declustering is not used; the success has a normal distribution across repeated tests when the number of test earthquakes is large (Kagan, 2009); and the results are on an absolute scale. The rate-corrected average-information-gain tests (T- and W-tests) of Rhoades *et al.* (2011) are similar but not identical.

As another measure of forecast map-patterns, we use the space statistic "S" (Zechar *et al.*, 2010) which is a variant of the L-test implemented at CSEP, but with the dependence on earthquake count removed. Specifically, we scale each forecast under test to the actual rate of earthquakes during the test period, to eliminate any direct earthquake-count factors in the likelihoods. As in the implementation of the L-test used at CSEP we simulate a number (1000) of virtual test catalogs from the forecast to experimentally describe the sample-size effect on test precision. No clustering or aftershock sequences are simulated in these virtual catalogs. The S-statistic is the fraction of simulations in which the (single) test-catalog "log-likelihood" exceeds the virtual-catalog "log-likelihood." If this statistic is less than 0.05, some consider the forecast should be rejected. However, note that (unlike CSEP) we have not declustered either the forecast or the test catalog. Consequently, our "log-likelihoods" are only biased estimates of true log-likelihoods, and our S-statistic is not really a probability, so applying any hard cutoff is inappropriate. Still, models that give a very low value of the S statistic are likely to have lower quality. Interestingly, a very high value of the S statistic (> 0.5) can be a sign of potential for improvement; it shows that test earthquakes consistently fell in the areas of maximum forecast probability, but that the intermediate-probability “shoulders” of the forecast were lightly populated with fewer than predicted test earthquakes, and may be broader than necessary.

References cited

Field, E. H. (2007). Overview of the Working Group for the Development of Regional Earthquake Likelihood Models (RELM), in: S. E. Hough and K. B. Olsen (ed.), Special Issue on: Regional Earthquake Likelihood Models, *Seismol. Res. Lett.*, **78**(1), 7-16.

Kagan, Y. Y. (2009). Testing long-term earthquake forecasts: Likelihood methods and error diagrams, *Geophys. J. Int.*, **177**(2), 532-542.

Kagan, Y. Y. (2010). Statistical distributions of earthquake numbers: Consequences of branching process, *Geophys. J. Int*., **80**(3), 1313-1328, doi: 10.1111/j.1365-246X.2009.04487.x.

Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christopherson, J. D. Zechar, and M. Imoto (2011). Efficient testing of earthquake forecasting models, *Acta Geophys*., **59**, 728-747.

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

Schorlemmer, D., M. C. Gerstenberger, W. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). Earthquake likelihood model testing, in: S. E. Hough and K. B. Olsen (ed.), Special Issue on: Regional Earthquake Likelihood Models, *Seismol. Res. Lett.*, **78**(1), 17-29.

Schorlemmer, D., J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, and the RELM Working Group (2010). First results of the Regional Earthquake Likelihood Models experiment, *Pure Appl. Geoph.*, **167**, 859-876, doi: 10.1007/s00024-010-0081-5.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010). Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.*, **100**(3), 1184-1195, doi: 10.1785/0120090192.

Table S1. Results of 8-year retrospective tests against shallow events from GCMT, 2005-2012:

Threshold:	***m*5.767+ (*N* = 1694)**			***m*7+ (*N* = 90)**		
Model Class, *c* or *d*	**Speci-ficity (I_0)**	**Success (I_1)**	**S statis-tic**	**Speci-ficity (I_0)**	**Success (I_1)**	**S statis-tic**
Envelope	3.607	4.00	0.960	3.745	4.0	0.61
Tectonic (*c* = 0)	4.034	3.69	0.023	4.347	3.8	0.35
Linear, *c* = 0.1	3.884	3.89	0.426	4.154	4.0	0.46
Linear, *c* = 0.2	3.779	3.98	0.806	4.012	4.1	0.54
Linear, *c* = 0.3	3.704	4.05	0.932	3.903	4.1	0.56
Linear, *c* = 0.4	3.653	4.09	0.971	3.819	4.1	0.56
Linear, *c* = 0.5	3.624	4.11	0.983	3.758	4.1	0.62
Linear, *c* = 0.6	3.616	4.12	0.986	3.720	4.1	0.62
Linear, *c* = 0.7	3.629	4.11	0.977	3.704	4.1	0.63
Linear, *c* = 0.8	3.665	4.08	0.969	3.713	4.0	0.59
Linear, *c* = 0.9	3.728	4.03	0.884	3.750	3.9	0.57
Seismicity (*c*=1)	3.829	3.93	0.584	3.829	3.8	0.48
Tectonic (*d* = 0)	4.034	3.69	0.023	4.347	3.8	0.35
Log-linear, *d*=0.1	3.960	3.85	0.234	4.232	4.0	0.43
Log-linear,	3.906	3.98	0.572	4.143	4.1	0.52

d=0.2						
Log-linear, *d*=0.3	3.865	4.09	0.827	4.072	4.2	0.55
Log-linear, *d*=0.4	3.836	4.16	0.927	4.013	4.2	0.54
Log-linear, *d*=0.5	3.815	4.21	0.968	3.962	4.2	0.58
Log-linear, *d*=0.6 (H*)	3.801	4.22	0.971	3.917	4.2	0.59
Log-linear, *d*=0.7	3.793	4.20	0.962	3.877	4.2	0.59
Log-linear, *d*=0.8	3.792	4.15	0.931	3.843	4.1	0.58
Log-linear, *d*=0.9	3.802	4.06	0.833	3.823	3.9	0.54
Seismicity (*d*=1)	3.829	3.93	0.584	3.829	3.8	0.48

m: magnitude; *N*: number of test earthquakes; *c, d*: mixing coefficients of hybrid forecasts, defined in main text; $\boldsymbol{I}_0$, $\boldsymbol{I}_1$: forecast scores of Kagan (2009); S-statistic defined by Zechar *et al.* (2010).

Table S2. One-year I_1 success scores against shallow events in GCMT, 2005-2012:

Threshold:	*m*5.767+ ($N \cong 212$/year)			*m*7+ ($N \cong 11$/year)		
Year	I_1(H*)	I_1(S)	I_1(H*)-I_1(S)	I_1(H*)	I_1(S)	I_1(H*)-I_1(S)
2005	3.8843	3.7445	0.1398	3.3317	2.4809	0.8508
2006	3.9901	3.5784	0.4117	3.1494	2.3293	0.8201
2007	4.2486	3.9903	0.2583	4.9593	4.5226	0.4367
2008	3.9290	3.6740	0.2550	3.5888	3.6086	-0.0198
2009	4.3836	4.0526	0.3310	4.7456	4.5973	0.1483
2010	4.2299	3.8480	0.3819	4.2105	3.4980	0.7125
2011	4.8201	4.4914	0.3287	4.6979	4.3939	0.3040
2012	4.1738	3.9273	0.2465	3.9932	3.4583	0.5349
Mean:	4.2074	3.9133	0.2941	4.0846	3.6111	0.4734
Standard Deviation:	0.3021	0.2833	0.0870	0.6855	0.8761	0.3168
Correlation:	0.9578			0.9467		

m: magnitude; *N*: number of test earthquakes; I_1: forecast score of Kagan (2009); H*: preferred hybrid forecast; S: Seismicity parent forecast.

Table S3. Results of 59-year retrospective tests against shallow events from ISC-GEM, 1918-1976:

	m6.8+ (N = 881)		
Model Class, c or d	**Speci-ficity (I_0)**	**Success (I_1)**	**S statis-tic**
Envelope	3.729	3.60	0.305
Tectonic (c = 0)	4.292	3.41	0.000
Linear, c = 0.1	4.108	3.59	0.034
Linear, c = 0.2	3.975	3.66	0.130
Linear, c = 0.3	3.872	3.70	0.289
Linear, c = 0.4	3.794	3.72	0.351
Linear, c = 0.5	3.738	3.72	0.455
Linear, c = 0.6	3.705	3.71	0.458
Linear, c = 0.7	3.694	3.69	0.429
Linear, c = 0.8	3.706	3.64	0.402
Linear, c = 0.9	3.747	3.56	0.237
Seismicity (c = 1)	3.829	3.39	0.073
Tectonic (d = 0)	4.292	3.41	0.000
Log-linear, d = 0.1	4.180	3.56	0.016
Log-linear, d = 0.2	4.092	3.67	0.065
Log-linear, d = 0.3	4.023	3.75	0.178

Log-linear, *d* = 0.4	3.967	3.80	0.253
Log-linear, *d* = 0.5	3.921	3.82	0.327
Log-linear, *d* = 0.6	3.883	3.81	0.359
Log-linear, *d* = 0.7	3.851	3.76	0.344
Log-linear, *d* = 0.8	3.828	3.68	0.274
Log-linear, *d* = 0.9	3.817	3.55	0.166
Seismicity (*d* = 1)	3.829	3.39	0.073

m: magnitude; *N*: number of test earthquakes; *c*, *d*: mixing coefficients of hybrid forecasts, defined in main text; $\boldsymbol{I}_0$, $\boldsymbol{I}_1$: forecast scores of Kagan (2009); S-statistic defined by Zechar *et al.* (2010).