Testing for ontological errors in probabilistic forecasting models of natural systems

Warner Marzocchi[a,1] and Thomas H. Jordan[b,1]

[a]Centro per la Pericolosità Sismica, Istituto Nazionale di Geofisica e Vulcanologia, 00143 Rome, Italy; and [b]Southern California Earthquake Center, Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089

Contributed by Thomas H. Jordan, June 9, 2014 (sent for review March 7, 2014; reviewed by James O. Berger and Frank Scherbaum)

Probabilistic forecasting models describe the aleatory variability of natural systems as well as our epistemic uncertainty about how the systems work. Testing a model against observations exposes ontological errors in the representation of a system and its uncertainties. We clarify several conceptual issues regarding the testing of probabilistic forecasting models for ontological errors: the ambiguity of the aleatory/epistemic dichotomy, the quantification of uncertainties as degrees of belief, the interplay between Bayesian and frequentist methods, and the scientific pathway for capturing predictability. We show that testability of the ontological null hypothesis derives from an experimental concept, external to the model, that identifies collections of data, observed and not yet observed, that are judged to be exchangeable when conditioned on a set of explanatory variables. These conditional exchangeability judgments specify observations with well-defined frequencies. Any model predicting these behaviors can thus be tested for ontological error by frequentist methods; e.g., using *P* values. In the forecasting problem, prior predictive model checking, rather than posterior predictive checking, is desirable because it provides more severe tests. We illustrate experimental concepts using examples from probabilistic seismic hazard analysis. Severe testing of a model under an appropriate set of experimental concepts is the key to model validation, in which we seek to know whether a model replicates the data-generating process well enough to be sufficiently reliable for some useful purpose, such as long-term seismic forecasting. Pessimistic views of system predictability fail to recognize the power of this methodology in separating predictable behaviors from those that are not.

system science | Bayesian statistics | significance testing | subjective probability | expert opinion

Science is rooted in the concept that a model can be tested against observations and rejected when necessary (1). However, the problem of model testing becomes formidable when we consider natural systems. Owing to their scale, complexity, and openness to interactions within a larger environment, most natural systems cannot be replicated in the laboratory, and direct observations of their inner workings are always inadequate. These difficulties raise serious questions about the meaning and feasibility of "model verification" and "model validation" (2), and have led to the pessimistic view that "the outcome of natural processes in general cannot be accurately predicted by mathematical models" (3).

Uncertainties in the formal representation of natural systems imply that the forecasting of emergent phenomena such as natural hazards must be based on probabilistic rather than deterministic modeling. The ontological framework for most probabilistic forecasting models comprises two types of uncertainty: an aleatory variability that describes the randomness of the system, and an epistemic uncertainty that characterizes our lack of knowledge about the system. According to this distinction, which stems from the classical dichotomy of objective/subjective probability (4), epistemic uncertainty can be reduced by increasing relevant knowledge, whereas the aleatory variability is intrinsic to the system representation and is therefore irreducible within that representation (5, 6).

The testing of a forecasting model is itself a statistical enterprise that evaluates how well a model agrees with some collection of observations (e.g., 7, 8). One can compare competing forecasts within a Bayesian framework and use new data to reduce the epistemic uncertainty. However, the scientific method requires the possibility of rejecting a model without recourse to specific alternatives (9, 10). The statistical gauntlet of model evaluation should therefore include pure significance testing (11). Model rejection exposes "unknown unknowns"; i.e., ontological errors in the representation of the system and its uncertainties. Here we use "ontological" to label errors in a model's quantification of aleatory variability and epistemic uncertainty (see *SI Text*, *Glossary*). [Other authors have phrased the problem in different terms; e.g., Musson's (12) "unmanaged uncertainties." In the social sciences, "ontological" is sometimes used interchangeably with "aleatory" (13).]

The purpose of this paper is to clarify the conceptual issues associated with the testing of probabilistic forecasting models for ontological errors in the presence of aleatory variability and epistemic uncertainty. Some relate to long-standing debates in statistical philosophy, in which Bayesians (14) spar with frequentists (15), and others propose methodological accommodations that draw from the strengths of both schools (10, 16). Statistical "unificationists" of the latter stripe advocate the importance of model checking using Bayesian (calibrated) *P* values (16, 17, 18, 19) as well as graphical summaries and other tools of exploratory data analysis (20). Bayesian modeling checking has been criticized by purists on both sides (21, 22), but one version, prior predictive checking, provides us with an appropriate framework for the testing of forecasting models for ontological errors.

Significance

Science is rooted in the concept that a model can be tested against observations and rejected when necessary. However, the problem of model testing becomes formidable when we consider the probabilistic forecasting of natural systems. We show that testability is facilitated by the definition of an experimental concept, external to the model, that identifies collections of data, observed and not yet observed, that are judged to be exchangeable and can thus be associated with well-defined frequencies. We clarify several conceptual issues regarding the testing of probabilistic forecasting models, including the ambiguity of the aleatory/epistemic dichotomy, the quantification of uncertainties as degrees of belief, the interplay between Bayesian and frequentist methods, and the scientific pathway for capturing predictability.

Author contributions: W.M. and T.H.J. designed research, performed research, contributed new reagents/analytic tools, and wrote the paper.

Reviewers: J.O.B., Duke University; and F.S., Institute of Earth and Environmental Sciences, University of Potsdam.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: warner.marzocchi@ingv.it or tjordan@usc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1410183111/-/DCSupplemental.

Among the concerns to be addressed is the use of expert opinion to characterize epistemic uncertainty, a common practice when dealing with extreme events, such as large earthquakes, volcanic eruptions, and climate change. Frequentists discount the quantification of uncertainties in terms of degrees of belief as "fatally subjective—unscientific" (23), and they oppose "letting scientists' subjective beliefs overshadow the information provided by data" (21). Bayesians, on the other hand, argue that probability is intrinsically subjective: all probabilities are degrees of belief that cannot be measured. Jaynes's (24) view is representative: "any probability assignment is necessarily 'subjective' in the sense it describes only a state of knowledge, and not anything that could be measured in a physical experiment." Immeasurability suggests untestability: How can models of probabilities that are not measureable be rejected?

Our plan is to expose the conceptual issues associated with the uncertainty hierarchy in the mathematical framework of a particular forecasting problem—probabilistic seismic hazard analysis (PSHA)—and resolve them in a way that generalizes the testing for ontological errors to other types of probabilistic forecasting models. A glossary of uncommon terms used in this paper is given in *SI Text, Glossary*.

Probabilistic Seismic Hazard Analysis

Earthquakes proceed as cascades in which the primary effects of faulting and ground shaking may induce secondary effects, such as landslides, liquefactions, and tsunamis. Seismic hazard is a probabilistic forecast of how intense these natural effects will be at a specified site on earth's surface during a future interval of time. Because earthquake damage is primarily caused by shaking, quantifying the hazard due to ground motions is the main goal of PSHA. Various intensity measures can be used to describe the shaking experienced during an earthquake; common choices are peak ground acceleration and peak ground velocity. PSHA estimates the exceedance probability of an intensity measure X; i.e., the probability that the shaking will be larger than some intensity value x at a particular geographic site over the time interval of interest, usually beginning now and stretching over several decades or more (5, 25, 26). It is often assumed that earthquake occurrence is a Poisson process with rates constant in time, in which case the hazard model is said to be time independent.

A plot of the exceedance probability $F(x) = P(X > x)$ for a particular site is called the hazard curve. Using hazard curves, engineers can estimate the likelihood that buildings and other structures will be damaged by earthquakes during their expected lifetimes, and they can apply performance-based design and seismic retrofitting to reduce structural fragility to levels appropriate for life safety and operational requirements. A seismic hazard map is a plot of the exceedance probability F at a fixed intensity x as a function of site position (27, 28) or, somewhat more commonly, x at fixed F (29–31). Official seismic hazard maps are now produced by many countries, and are used to specify seismic performance criteria in the design and retrofitting of buildings, lifelines, and other infrastructure, as well as to guide disaster preparedness measures and set earthquake insurance rates. These applications often fold PSHA into probabilistic risk analysis based on further specifications of loss models and utility measures (32, 33). Here we focus on assessing the reliability of PSHA as a forecasting tool rather than its role in risk analysis and decision theory.

The reliability of PSHA has been repeatedly questioned. In the last few years, disastrous earthquakes in Sumatra, Italy, Haiti, Japan, and New Zealand have reinvigorated this debate (34–39). Many practical deficiencies have been noted, not the least of which is the paucity of data for retrospective calibration and prospective testing of PSHA models, owing to the short span of observations relative to the forecasting time scale (40, 41). However, some authors have raised the more fundamental question of whether PSHA is misguided because it cannot capture the aleatory variability of large-magnitude earthquakes produced by complex fault systems (35, 38, 42). Moreover, the pervasive role of subjective probabilities in specifying the epistemic uncertainty in PSHA has made this methodology a target for criticism by scientists who adhere to the frequentist view of probabilities. A particular objection is that degrees of belief cannot be empirically tested and, therefore, that PSHA models are not scientific (43–45).

We explore these issues in the PSHA framework developed by earthquake engineers in the 1970s and 1980s and refined in the 1990s by the Senior Seismic Hazard Analysis Committee (SSHAC) of the US Nuclear Regulatory Commission (5). For a particular PSHA model H_m, the intrinsic or aleatory variability of the ground motion X is described by the hazard curve $F_m(x) = P(X > x \mid H_m)$. The epistemic uncertainty is characterized by an ensemble $\{H_m\colon m \in M\}$ of alternative hazard models consistent with our present knowledge. A probability π_m is assigned to $F_m(x)$ that measures its plausibility, based on present knowledge, relative to other hazard curves drawn from the ensemble.

Ideally, the model ensemble $\{H_m\}$ could be constructed and the model probabilities assigned according to how well the candidates explain prior data. Various objective methods have been developed for this purpose: resampling, adaptive boosting, Bayesian information criteria, etc. (46, 47). However, the high-intensity, low-probability region of the hazard space $F \otimes x$ most relevant to many risk decisions is dominated by large, infrequent earthquakes. The data for these extreme events are usually too limited to discriminate among alternative assumptions and fix key parameters. Therefore, in common practice, the model ensemble $\{H_m\}$ is organized as branches of a logic tree, and the branch weights $\{\pi_m\}$, which may depend on the hazard level x, are assigned according to expert opinion (5, 48, 49).

If the logic tree spans a hypothetically mutually exclusive and completely exhaustive (MECE) set of possibilities (50), π_m can be interpreted as the probability that $F_m(x)$ is the "true" hazard value $\hat{F}(x)$, and operations involving the plausibility measure $\{\pi_m\}$ must obey Kolmogorov's axioms of probability; e.g.,

$$\sum_{m \in M} \pi_m = 1. \qquad [1]$$

In PSHA practice, logic trees are usually constructed to sample the possibilities, rather than exhaust them, in which case the MECE assumption is inappropriate. One can then reinterpret π_m as the probability that $F_m(x)$ is the "best" among a set of available models (12, 47, 51) or "the one that should be used" (52). This utilitarian approach increases the subjective content of $\{F_m(x), \pi_m\}$, as well as the possibilities for ontological error. We will call $\{F_m(x), \pi_m\}$ the experts' distribution to recognize this subjectivity, regardless of whether $\{H_m\}$ comes from a logic tree or is constructed in a different way.

For conceptual simplicity, we will assume the discrete experts' distribution $\{F_m(x), \pi_m\}$ samples a continuous probability distribution with density function $p(\phi)$, where $\phi = F(x_0)$ at a fixed hazard value x_0. We denote this relationship by $\{\phi_m\} \sim p(\phi)$ and call p the extended experts' distribution. Various data-analysis methods have been established to move from a discrete sample to a continuous distribution, although the process also compounds the potential for ontological error. Given the extended experts' distribution, the expected hazard at fixed x_0 is

$$\overline{\phi} = \int_0^1 \phi \; p(\phi) d\phi. \qquad [2]$$

This central value measures the aleatory variability of the hazard, conditional on the model, and the dispersion of p about $\overline{\phi}$ describes the epistemic uncertainty in its estimation.

Epistemic uncertainty is thus described by imposing a subjective probability on the target behavior of H_m, which is an

objective exceedance probability, ϕ_m. This procedure is well established in PSHA practice (5, 53), but it encounters frequentist discomfort with degrees of belief and Bayesian resistance to attaching measures of precision to probabilities. [In Bayesian semantics, the frequency parameters of aleatory variability are usually labeled as "chances" rather than as probabilities. According to Lindley (14), "the distinction becomes useful when you wish to consider your probability of a chance, whereas probability on a probability is, in the philosophy, unsound." Jaynes (24) also rejects the identification of frequencies with probabilities.] Although these concerns do not matter much in the routine application of risk analysis, where only the mean hazard is considered (32, 54, 55), they must be addressed within the conceptual framework of model testing.

In our framework, testability derives from a null hypothesis, here called the "ontological hypothesis," which states that data-generating hazard curve $\hat{\phi}$ (the "true" hazard) is a sample from the extended experts' distribution, $\hat{\phi} \sim p(\phi)$. If an observational test rejects this null hypothesis, then we can claim to have found an ontological error.

Testing PSHA Models

One straightforward test of a PSHA model is to collect data on the exceedance frequency of a specified shaking intensity, x_0, during N equivalent trials, each lasting 1 y. If $\phi_m = F_m(x_0)$ is the 1-y exceedance probability for a particular hazard model H_m, and the data are judged to be unbiased and exchangeable under the experimental conditions—i.e., to have a joint probability distribution invariant with respect to permutations in the data ordering (56–58)—then the likelihood of observing k or more exceedances for each member of the experts' ensemble is given by the tail of a binomial distribution:

$$P(k|\phi_m) = \sum_{n=k}^{N} \binom{N}{n} \phi_m^n (1-\phi_m)^{N-n}. \quad [3]$$

Under the extended experts' distribution, the unconditional probability is the expectation,

$$P(k) = \int_0^1 P(k|\phi) p(\phi) d\phi. \quad [4]$$

For notational simplicity, we suppress the dependence of P on N.

There have been only a few published attempts to test PSHA models against ground motion observations (59–62). To our knowledge, all have assumed a test distribution $P(k|\overline{\phi})$ that has been computed from [**3**] using the mean exceedance probability rather than from the unconditional distribution [**4**]. This reflects the view shared by many hazard practitioners that the mean hazard is the only hazard needed for decision making (54, 55, 63, 64).

However, a test based on $P(k|\overline{\phi})$ is often overly stringent, as can be seen from a simple example. We consider a PSHA model that comprises an experts' ensemble of 20 equally weighted hazard curves, each an exponential function, $F_m(x) = \exp(-\lambda_m x)$ (Fig. 1). The values sampled at x_0 (arbitrarily chosen to be 0.29) give a mean exceedance probability of $\overline{\phi} = 0.085$ and can be represented by a beta distribution Be(α, β) with parameters $\alpha = 1.0$ and $\beta = 10.7$ (Fig. 2*A*). Suppose this ground motion threshold is exceeded $k = 10$ times in $n = 50$ y; then the P value conditioned on the mean hazard is $P(k|\overline{\phi}) = 0.008$, whereas the unconditional value is $P(k) = 0.123$ (Fig. 2*B*). Thus, although this observational test rejects $\hat{\phi} = \overline{\phi}$ at a fairly high (99%) confidence level, it cannot reject the ontological hypothesis $\hat{\phi} \sim p(\phi)$, even at a low (90%) level.

Finding an ontological error may not directly indicate what is wrong with the model. The ontological error might indicate that the parameters of the model were badly estimated, or that the basic structure of the model is far from reality. In our example test, a small P value could imply that either the beta distribution of ϕ, which characterizes the epistemic uncertainty, or the exponential distribution of $F_m(x)$, which characterizes the aleatory variability, is wrong (or that both are). A small P value might also indicate that the data-generating process is evolving with time, so the data used for testing do not have the same distribution as the data used to calibrate the model.

Bayesian updating under the ontological hypothesis can improve the parameter estimates and thereby sharpen the extended experts' distribution, but it cannot discover ontological errors, such as the inadequacy of the exponential distribution or a time dependence of the parameter λ not included in the time-independent model. To do that we must subject our "model of the world," given here by $p(\phi)$, to a testing regime guided by an experimental concept that appropriately conditions nature's aleatory variability.

Primacy of the Experimental Concept in Ontological Testing

The experimental concept in our PSHA example is very simple. We collect a set of yearly data $\{x_n: n = 1, 2, \ldots, N\}$ and construct a binary sequence $\{e_n: n = 1, 2, \ldots, N\}$ by assigning $e_n = 1$ if x_n exceeds a threshold value x_0 and $e_n = 0$ if it does not. We observe that the sequence sums to k. We judge that the joint probability distribution of $\{e_n\}$ is unchanged by any permutation of the indices; the yearly data are thus exchangeable and the sequence is Bernoulli. In the earthquake forecasting problem, we further assert, through a leap of faith, that future years are exchangeable with past years; i.e., the data sequence is exchangeable in Draper et al.'s (57) second sense. The predictive power of the time-independent PSHA model hangs on this assertion.

In general terms, the experimental concept specifies collections of data, observed and not yet observed, that are judged to be exchangeable when conditioned on a set of explanatory variables. Exchangeable events can be modeled as identical and independently distributed random variables with a well-defined frequency of occurrence (65, 66). This event frequency or chance—the limiting value of k/N in our example—represents the aleatory variability of the data-generating process (5, 6). Theoretical considerations and a finite amount of data can only constrain this probability within some epistemic uncertainty, quantified by the extended expert's distribution.

Exchangeability thus distinguishes the aleatory variability, given by $P(k \mid \phi)$, from the epistemic uncertainty, given by $p(\phi)$. In our PSHA example, the exchangeability judgment links Eqs. **3** and **4** to de Finetti's (65) representation of a Bernoulli process conditioned on $p(\phi)$ and thus connects the testing experiment to the frequentist concept of repeatability.

The primacy of the experimental concept can be illustrated by extending our PSHA example. The test in Fig. 2 (test 1) derives from an experimental concept based on a single exchangeability judgment. Now consider a second experimental concept (test 2)

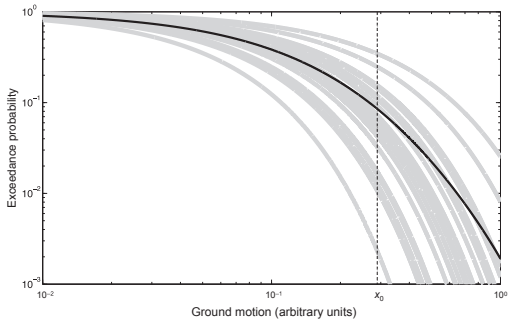

Fig. 1. A simple PSHA model used in our testing examples. Gray lines show the experts' ensemble of 20 exponential hazard curves $\{F_m(x): m = 1, \ldots, 20\}$. Black line is the corresponding mean hazard curve $\overline{F}(x)$. The distribution of values sampled at x_0 (dashed line) is given in Fig. 2.

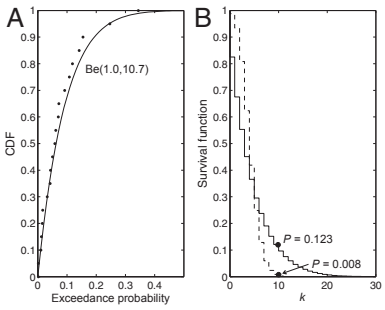

Fig. 2. Results from test 1. (*A*) Discrete points are the cumulative distribution function (CDF) of the experts' ensemble $\{F_m(x_0)\colon x_0 = 0.29,\ m = 1, \ldots, 20\}$ from Fig. 1. When assigned equal weights ($\pi_m = 1/20$), this ensemble can be represented by an extended experts' distribution Be(1.0, 10.7), shown by the solid line. (*B*) Dashed line is the survival function (1 – CDF) of k or more exceedances in $n = 50$ trials conditioned on the mean hazard; solid line is the survival function of the unconditional exceedances, computed from Eq. **4**. For the observation $k = 10$, the P value of the former is 0.008 and that of the latter is 0.123.

that distinguishes exceedance events in years when some observable index A is zero from those in years when A is unity. The data-generating process provides two sequences, $\{e_n^{(0)}\colon n = 1, 2, \ldots, N_0\}$ when $A = 0$ and $\{e_n^{(1)}\colon n = 1, 2, \ldots, N_1\}$ when $A = 1$. Both are judged to be Bernoulli, and they are observed to sum to k_0 and k_1 respectively. If A correlates with the frequency and/or magnitude variations in the earthquake rupture process, e.g., if the occurrence of large earthquakes and consequent ground shaking stronger than x_0 are more likely when $A = 1$, then the expected frequency of k_1/N_1 might be greater than that of k_0/N_0.

As this example makes clear, it is not the aleatory variability intrinsic to the model that matters in testing, but rather the aleatory variability defined by the exchangeability judgments of the experimental concept. In other words, aleatory variability is an observable behavior of the data-generating process—nature itself—conditioned by the experimental concept to have well-defined frequencies. A model predicting this behavior can thus be tested for ontological error by frequentist (error statistical) methods.

Suppose that, in a 50-y PSHA experiment, there are $N_0 = 35$ y when $A = 0$ and $N_1 = 15$ y with $A = 1$. Further suppose that no exceedances of the ground motion threshold are observed in the former set ($k_0 = 0$), and 10 are recorded in the latter ($k_1 = 10$). Test 1 applied to these datasets returns an identical result ($k = 10$), so the model passes with a prior predictive P value of 0.123, as shown in Fig. 2. However, in test 2, the $A = 1$ observation is much less likely under the ontological hypothesis ($P_1 = 0.0006$) than the $A = 0$ observation ($P_0 = 0.231$), and the P value for the combined result is quite small, 0.0012 (Fig. 3). Therefore, the model can be rejected by test 2 with high confidence.

From this experiment, we infer that the data-generating process $\hat{F}$ is probably A dependent and therefore time dependent (all years are not exchangeable). Hence, we might seek an alternative model that captures this type of time dependence in its aleatory variability, and we might reelicit expert opinion to characterize the epistemic uncertainty in its (two or more) aleatory frequencies.

In ontological testing, Box's famous generalization that "all models are wrong, but some are useful" (67) can be restated as "all models are wrong, but some are acceptable under particular experimental concepts." Qualifying a model under an appropriate set of experimental concepts is the key to model validation, in which we decide if a model replicates the data-generating process well enough to be sufficiently reliable for some useful purpose, such as long-term seismic forecasting. In our example, a model that passes test 1 may be adequate for time-independent forecasting, but, by failing test 2, it should be rejected as a viable time-dependent forecast.

Discussion

For frequentists, a probability is the limiting frequency of a random event or the long-term propensity to produce such a limiting frequency (9, 68); for Bayesians, it is a subjective degree of belief that a random event will occur (14). Advocates on both sides have argued that degrees of belief cannot be measured and, by implication, cannot be rejected (9, 24, 69). In the words of one author (70), "the degree of belief probability is not a property of the event (experiment) but rather of the observer. There exists no uniquely true probability, only one's true belief." Within the subjectivist Bayesian framework, one's true belief can be informed, but not rejected, by experiment.

For us, the use of subjective probability such as expert opinion poses no problems for ontological testability as long as the experimental concept defines sets of exchangeable data with long-run frequencies determined by the data-generating process. These frequencies, which characterize the aleatory variability, have epistemic uncertainty described by the experts' distribution. Expert opinion is thus regarded as a measurement system that produces a model that can be tested.

To illustrate this point with an example far from PSHA, we recall an experiment reported by Sir Francis Galton in 1907. During a tour of the English countryside, he recorded 787 guesses of the weight of an ox made by a farming crowd and found that the average was correct to a single pound (71). The experts' distribution he tabulated passes the appropriate Student's t test (retrospectively, of course, because this test was not invented until 1908).

We note one difference between farmers and PSHA experts. The experts' distribution measures the epistemic uncertainty at a particular epoch, which may be larger or smaller depending on how the experts are able to sample the appropriate information space. In Galton's experiment, a farmer looks individually at the ox, reaching his estimate more or less independently of his colleagues. As more farmers participate, they add new data, and the epistemic uncertainty is reduced by averaging their guesses. At a particular epoch, adding more PSHA experts will better determine, but usually not reduce, the epistemic uncertainty, because they rarely make independent observations but work instead from a common body of knowledge, which may be very limited. This and other issues related to the elicitation of expert opinion, such as how individuals should be calibrated and how a consensus should be drawn from groups, have been extensively studied (5, 72, 73).

The ontological tests considered here are conditional on the experimental concept, which may be weak or even wrong. An experimental concept will be incorrect when one or more of its exchangeability judgments violate reality. The constituent assumptions can often be tested independently of the model; e.g., through correlation analysis and other types of data checks (57, 74, 75). In our example, if the frequency estimator k/N increases in the long run (e.g., if the proportion N_1/N_0 increases), then the event set is not likely to be exchangeable,

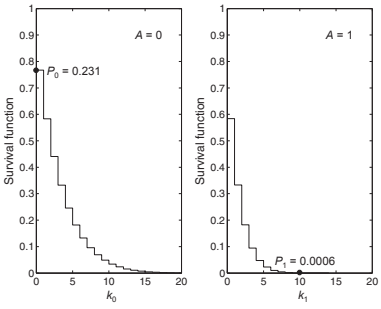

Fig. 3. Results from test 2. In a 50-y PSHA experiment, no ground motion exceedances are observed in the 35 y when $A = 0$ (*Left*), which gives $P_0 = 0.231$, and 10 exceedances are observed in the 15 y with $A = 1$ (*Right*), which gives $P_1 = 0.0006$. The P value for the combined result is 0.0012.

and the experimental concept of test 1 needs to be rethought to allow for this secular time dependence.

Through exchangeability judgments, the experimental concept ensures that aleatory variables have well-defined frequencies, which is just what we need to set up a regime for testing the ontological hypothesis. Bayesian model checking provides us with several options for pure significance testing. The ontological hypothesis can be evaluated using

i) the prior predictive *P* value, computed directly from Eq. **4** (17);
ii) the posterior predictive *P* value, computed after the experts' distribution has been updated according to Bayes; e.g., $\pi'_m \propto \phi_m^k (1-\phi_m)^{N-k} \pi_m$ (18, 76); or
iii) a partial posterior predictive *P* value, computed for a data subset after the experts' distribution has been updated using a complementary (calibration) data subset (77).

An ontological error is discovered when the new data are shown to be inconsistent with $p(\phi)$ or its updated version $p'(\phi)$; e.g., when a small *P* value is obtained from the experiment.

Among these options, the posterior predictive checks (*ii*) and (*iii*) are most often used by Bayesian objectivists because their priors are often improper and uninformative (20, 77). Moreover, in many types of statistical modeling, the goal is to assess the model's ability to fit the data a posteriori, not its ability to predict the data a priori. According to Gelman (78), for example, "All models are wrong, and the purpose of model checking (as we see it) is not to reject a model but rather to understand the ways in which it does not fit the data. From a Bayesian point of view, the posterior distribution is what is being used to summarize inferences, so this is what we want to check." In particular, the power of the test, or more generally its severity (9), is not very important. "If a certain posterior predictive check has zero or low power, this is not a problem for us: it simply represents a dimension of the data that is automatically or virtually automatically fit by the model" (75).

In forecasting, however, we are most interested in a model's predictive capability; therefore, severe tests based on the prior predictive check (*i*) are always desirable. This testing regime is entirely prospective; the models are independent of the data used in the test (the testing is blind), and there are no nuisance parameters. The extended experts' distribution $p(\phi)$ is subjective, informative, and always proper. "Using the data twice" is not an issue, as it is with the posterior predictive check (*ii*) (16). If based on the same data, prior predictive tests are always more severe than posterior predictive tests. Continual prospective testing now guides the validation of forecasting models in civil protection and other operational applications (79–82).

When the experimental concept is weak, lots of models, even poor predictors, can pass the test. An experimental concept provides a severe test, in Mayo's (9) sense, if it has a high probability of detecting an ontological error of the type that matters to the model's forecasting application. Severe tests require informative ensembles of exchangeable observations (although we must admit that these are often lacking in long-term PSHA). In practice, batteries of experimental concepts must be used to specify the aleatory variability of the data-generating process and organize severe testing gauntlets relevant to the problem at hand.

When the problem is forecasting, the most crucial features of an experimental concept are the assertions that past and future events are exchangeable. Scientists make these leaps of faith not blindly, but through careful consideration of the physical principles that govern the system behaviors they seek to predict. Therefore, the experimental concepts used to test models, as much as the models themselves, are the mechanism for encoding our physical understanding into the iterative process of system modeling. Validating our predictions through ontological testing is the primary means by which we establish our understanding of how the world works, and thus an essential aspect of the scientific method (1). It seems to us that the more pessimistic views of system predictability (2, 3) fail to recognize the power of this methodology in separating behaviors that are predictable from those that are not.

This power can be appreciated by considering cases where the exchangeability of past and future events is dubious. A notorious example is the prediction of financial markets. Exchangeability judgments are problematic in these experiments, because the markets learn so rapidly from past experience (83, 84). Without exchangeability, no experimental concept is available to discipline the system variability. Processes governed by physical laws are less contingent and more predictable than these agent-based systems; for example, exchangeability judgments can be guided by the characteristic scales of physical processes, leading to well-configured experimental concepts.

The points made in this paper are basic and without mathematical novelty. In terms of the interplay between Bayesian and frequentist methods, our view aligns well with the statistical unificationists (10, 85). However, it differs in the importance we place on the experimental concept in structuring the uncertainty hierarchy—aleatory, epistemic, ontological—and our emphasis on testing for ontological errors as a key step in the iterative process of forecast validation.

As with many conceptual discussions of statistical methodology, the proof is in the pudding. Does the particular methodology advocated here help to clarify any persistent misunderstandings that have hampered practitioners? We think so, particularly in regard to the widespread confusion about how to separate aleatory variability from epistemic uncertainty. Consider two examples:

- In a review of the SSHAC methodology requested by the US Nuclear Regulatory Commission, a panel of the National Research Council asserted that "the value of an epistemic/aleatory separation to the ultimate user of a PSHA is doubtful... The panel concludes that, unless one accepts that all uncertainty is fundamentally epistemic, the classification of PSHA uncertainty as aleatory or epistemic is ambiguous." (54). Here the confusion stems from SSHAC's (5) epistemic/aleatory classification scheme, which is entirely model-based and thus ambiguous.
- In his discussion of "how to cheat at coin and die tossing," Jaynes (24) describes the ambiguity of randomness in terms of how the tossing is done; e.g., how high a coin is tossed. "The writer has never thought of a biased coin 'as if it had a physical probability' because, being a professional physicist, I know that it does *not* have a physical probability." His confusion arises because he associates the aleatory frequency with the physical process, which is ambiguous unless we fix the experimental concept.

Both the model-based and physics-based ambiguity in setting up the aleatory/epistemic dichotomy can be removed by specifying an experimental concept. By testing the ontological hypothesis under appropriate experimental concepts, we can answer the important question of whether a model's predictions conform to our conditional view of nature's true variability.

ACKNOWLEDGMENTS. We thank Donald Gillies for discussions on the philosophy of probability. W.M. was supported by Centro per la Pericolosità Sismica dell' Istituto Nazionale di Geofisica e Vulcanologia and the Real Time Earthquake Risk Reduction (REAKT) Project (Grant 282862) of the European Union's Seventh Programme for Research, Technological Development and Demonstration. T.H.J. was supported by the Southern California Earthquake Center (SCEC) under National Science Foundation Cooperative Agreement EAR-1033462 and US Geological Survey Cooperative Agreement G108C20038. The SCEC contribution number for this paper is 1953.

1. American Association for the Advancement of Science (1989) *Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics and Technology* (American Association for the Advancement of Science, Washington, DC).
2. Oreskes N, Shrader-Frechette K, Belitz K (1994) Verification, validation, and confirmation of numerical models in the Earth sciences. *Science* 263(5147):641–646.
3. Pilkey OH, Pilkey-Jarvis L (2007) *Useless Arithmetic: Why Environmental Scientists Can't Predict the Future* (Columbia Univ Press, New York).

4. Hacking I (1975) *The Emergence of Probability* (Cambridge Univ Press, Cambridge, UK).
5. Senior Seismic Hazard Analysis Committee (1997) *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts* (U.S. Nuclear Regulatory Commission, U.S. Dept. of Energy, Electric Power Research Institute), NUREG/CR-6372, UCRL-ID-122160.
6. Goldstein M (2013) Observables and models: Exchangeability and the inductive argument. *Bayesian Theory and Its Applications*, eds Damien P, Dellaportas P, Olson NG, Stephens DA (Oxford Univ Press, Oxford, UK), pp 3–18.
7. Murphy AH, Winkler RL (1987) A general framework for forecast verification. *Mon Weather Rev* 115:1330–1338.
8. Schorlemmer D, Gerstenberger MC, Wiemer S, Jackson DD, Rhoades DA (2007) Earthquake likelihood model testing. *Seismol Res Lett* 78:17–29.
9. Mayo DG (1996) *Error and the Growth of Experimental Knowledge* (Univ of Chicago Press, Chicago).
10. Gelman A, Shalizi CR (2013) Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol* 66(1):8–38.
11. Cox DR, Hinkley DV (1974) *Theoretical Statistics* (Chapman & Hall, London).
12. Musson R (2012) On the nature of logic trees in probabilistic seismic hazard assessment. *Earthquake Spectra* 28(3):1291–1296.
13. Dequech D (2004) Uncertainty: Individuals, institutions and technology. *Camb J Econ* 28(3):365–378.
14. Lindley DV (2000) The philosophy of statistics. *Statistician* 49(3):293–337.
15. Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *Br J Philos Sci* 57:323–357.
16. Bayarri MJ, Berger J (2000) P-values for composite null models. *J Am Stat Assoc* 95(452):1127–1142.
17. Box GEP (1980) Sampling and Bayes inference in scientific modelling and robustness. *Roy Statist Soc Ser A* 143(4):383–430.
18. Gelman A, Ming XL, Stern HS (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Stat Sin* 6:733–807.
19. Sellke T, Bayarri MJ, Berger JO (2001) Calibration of *p* values for testing precise null hypotheses. *Am Stat* 55(1):62–71.
20. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) *Bayesian Data Analysis* (Chapman & Hall/CRC, London), 2nd Ed.
21. Mayo D (2011) Statistical science and philosophy of science: Where do/should they meet in 2011 (and beyond)? *Rationality. Mark Morals* 2:79–102.
22. Morey RD, Romeijn J-W, Rouder JN (2013) The humble Bayesian: Model checking from a fully Bayesian perspective. *Br J Math Stat Psychol* 66(1):68–75.
23. Jaffe AH (2003) A polemic on probability. *Science* 301(5638):1329–1330.
24. Jaynes ET (2003) *Probability Theory: The Logic of Science* (Cambridge Univ Press, New York).
25. Cornell CA (1968) Engineering seismic risk analysis. *Bull Seismol Soc Am* 58(5): 1583–1606.
26. McGuire RK (2004) *Seismic Hazard and Risk Analysis* (Earthquake Engineering Research Institute, Oakland, CA).
27. Working Group on California Earthquake Probabilities (1995) Seismic hazards in southern California: Probable earthquakes, 1994-2024. *Bull Seismol Soc Am* 85(2):379–439.
28. Headquarters for Earthquake Research Promotion (2005) *National Seismic Hazard Maps for Japan* (Report issued on 23 March 2005, available in English translation at http://www.jishin.go.jp/main/index-e.html).
29. Frankel AD, et al. (1996) *National Seismic Hazard Maps; Documentation June 1996*. (US Geological Survey Open-file Report1996-532).
30. MPS Working Group (2004) Redazione della mappa di pericolosita sismica prevista dall'Ordinanza PCM del 20 marzo 2003, rapporto conclusivo per il Dipartimento della Protezione Civile (Istituto Nazionale di Geofisica e Vulcanologia, Milano-Roma, http://zonesismiche.mi.ingv.it).
31. Petersen MD, et al. (2008) *Documentation for the 2008 Update of the United States National Seismic Hazard Maps*. (US Geological Survey Open-file Report 2008-1128).
32. Bedford T, Cooke R (2001) *Probabilistic Risk Analysis: Foundations and Methods* (Cambridge Univ Press, Cambridge).
33. Baker JW, Cornell CA (2008) Uncertainty propagation in probabilistic seismic loss estimation. *Struct Saf* 30:236–252.
34. Stein S, Geller R, Liu M (2011) Bad assumptions or bad luck: Why earthquake hazard maps need objective testing. *Seismol Res Lett* 82(5):623–626.
35. Stein S, Geller R, Liu M (2012) Why earthquake hazard maps often fail and what to do about it. *Tectonophysics* 562-563:1–25.
36. Stirling M (2012) Earthquake hazard maps and objective testing: The hazard mapper's point of view. *Seismol Res Lett* 83(2):231–232.
37. Hanks TC, Beroza GC, Toda S (2012) Have recent earthquakes exposed flaws in or misunderstandings of probabilistic seismic hazard analysis? *Seismol Res Lett* 83(5): 759–764.
38. Wyss MA, Nekrasova A, Kossobokov V (2012) Errors in expected human losses due to incorrect seismic hazard estimates. *Nat Hazards* 62(3):927–935.
39. Frankel A (2013) Comment on "Why earthquake hazard maps often fail and what to do about it" by S. Stein, R. Geller, and M. Liu. *Tectonophys* 592:200–206.
40. Savage JC (1991) Criticism of some forecasts of the National Earthquake Prediction Evaluation Council. *Bull Seismol Soc Am* 81(3):862–881.
41. Field EH, et al. (2009) Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2). *Bull Seismol Soc Am* 99(4):2053–2107.
42. Geller RJ (2011) Shake-up time for Japanese seismology. *Nature* 472(7344):407–409.
43. Krinitzsky EL (1995) Problems with logic trees in earthquake hazard evaluation. *Eng Geol* 39:1–3.
44. Castaños H, Lomnitz C (2002) PSHA: Is it science? *Eng Geol* 66(3-4):315–317.
45. Klügel J-U (2009) Probabilistic seismic hazard analysis for nuclear power plants – Current practice from a European perspective. *Nucl Eng Technol* 41(10):1243–1254.
46. Cox LA, Jr (2012) Confronting deep uncertainties in risk analysis. *Risk Anal* 32(10): 1607–1629.
47. Marzocchi W, Zechar JD, Jordan TH (2012) Bayesian forecast evaluation and ensemble earthquake forecasting. *Bull Seismol Soc Am* 102(6):2574–2584.
48. Kulkarni RB, Youngs RR, Coppersmith KJ (1984) Assessment of confidence intervals for results of seismic hazard analysis. *Proceedings of the Eighth World Conference on Earthquake Engineering*, vol 1, 263-270, International Association for Earthquake Engineering, San Francisco.
49. Runge AK, Scherbaum F, Curtis A, Riggelsen C (2013) An interactive tool for the elicitation of subjective probabilities in probabilistic seismic-hazard analysis. *Bull Seismol Soc Am* 103(5):2862–2874.
50. Bommer JJ, Scherbaum F (2008) The use and misuse of logic trees in probabilistic seismic hazard analysis. *Earthquake Spectra* 24(4):997–1009.
51. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. *Stat Sci* 14(4):382–417.
52. Scherbaum F, Kuhen NM (2011) Logic tree branch weights and probabilities: Summing up to one is not enough. *Earthquake Spectra* 27(4):1237–1251.
53. Working Group on California Earthquake Probabilities (2003) *Earthquake Probabilities in the San Francisco Bay Region: 2002–2031* (USGS Open-File Report 2003-214).
54. National Research Council Panel on Seismic Hazard Evaluation (1997) *Review of Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts* (National Academy of Sciences, Washington, DC).
55. Musson RMW (2005) Against fractiles. *Earthquake Spectra* 21(3):887–891.
56. Lindley DV, Novick MR (1981) The role of exchangeability in inference. *Ann Stat* 9(1): 45–58.
57. Draper D, Hodges JS, Mallows CL, Pregibon D (1993) Exchangeability and data analysis. *J R Stat Soc Ser A Stat Soc* 156(1):9–37.
58. Bernardo JM (1996) The concept of exchangeability and its applications. *Far East J Math Sci* 4:111–122.
59. McGuire RK, Barnhard TP (1981) Effects of temporal variations in seismicity on seismic hazard. *Bull Seismol Soc Am* 71(1):321–334.
60. Stirling M, Petersen M (2006) Comparison of the historical record of earthquake hazard with seismic hazard models for New Zealand and the continental United States. *Bull Seismol Soc Am* 96(6):1978–1994.
61. Albarello D, D'Amico V (2008) Testing probabilistic seismic hazard estimates by comparison with observations: An example in Italy. *Geophys J Int* 175:1088–1094.
62. Stirling M, Gerstenberger M (2010) Ground motion-based testing of seismic hazard models in New Zealand. *Bull Seismol Soc Am* 100(4):1407–1414.
63. McGuire RK, Cornell CA, Toro GR (2005) The case for using mean seismic hazard. *Earthquake Spectra* 21(3):879–886.
64. Cox LA, Brown GC, Pollock SM (2008) When is uncertainty about uncertainty worth characterizing? *Interfaces* 38(6):465–468.
65. de Finetti B (1974) *Theory of Probability: A Critical Introductory Treatment* (John Wiley and Sons, London).
66. Feller W (1966) *An Introduction to Probability Theory and its Applications* (John Wiley and Sons, New York), 3rd Ed, Vol II.
67. Box GEP (1976) Science and statistics. *J Am Stat Assoc* 71(356):791–799.
68. Popper KR (1983) *Realism and the Aim of Science* (Hutchinson, London).
69. de Finetti B (1989) Probabilism. *Erkenntnis* 31:169–223.
70. Vick SG (2002) *Degrees of Belief: Subjective Probability and Engineering Judgment* (ASCE Press, Reston, VA).
71. Galton F (1907) Vox populi. *Nature* 75(1949):450–451.
72. Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford Univ Press, New York).
73. U.S. Environmental Protection Agency (2011) *Expert Elicitation Task Force White Paper* (Science and Technology Policy Council, U.S. Environmental Protection Agency, Washington, DC).
74. O'Neill B (2009) Exchangeability, correlation, and Bayes' effect. *Int Stat Rev* 77(2): 241–250.
75. Gelman A (2011) Induction and deduction in Bayesian data analysis. *Rationality. Mark Morals* 2:67–78.
76. Rubin DB (1984) Bayesian justifiable and relevant frequency calculations for the applied statistician. *Ann Stat* 12(4):1151–1172.
77. Bayarri MJ, Castellanos ME (2007) Bayesian checking of the second levels of hierarchical models. *Stat Sci* 22(3):322–343.
78. Gelman A (2007) Comment: Bayesian checking of the second levels of hierarchical models. *Stat Sci* 22(3):349–352.
79. Jordan TH, et al. (2011) Operational earthquake forecasting: State of knowledge and guidelines for implementation. *Ann Geophys* 54(4):315–391.
80. Inness P, Dorling S (2013) *Operational Weather Forecasting* (John Wiley and Sons, New York).
81. Marzocchi W, Lombardi AM, Casarotti E (2014) The establishment of an operational earthquake forecasting system in Italy. *Seismol Res Lett*, in press.
82. Zechar JD, et al. (2010) The Collaboratory for the Study of Earthquake Predictability perspectives on computational earthquake science. *Concurr Comput* 22:1836–1847.
83. Davidson P (1996) Reality and economic theory. *J Post Keynes Econ* 18(4):479–508.
84. Silver N (2012) *The Signal and the Noise: Why So Many Predictions Fail But Some Don't* (Penguin Press, New York).
85. Bayarri MJ Berger JO (2013) Hypothesis testing and model uncertainty. *Bayesian Theory and Its Applications*, ed Damien P, Dellaportas P, Olson NG, and Stephens DA (Oxford Univ Press, Oxford, UK), pp. 361–194.