OPINION

"All Models Are Wrong, but Some Are Useful"

I have found myself repeating this famous Box (1980) quote a lot lately, so much so that one colleague said he was going to keep a tally (perhaps implying that it is getting annoying). Maybe explaining the three main reasons here will help shut me up.

WE SHOULDN'T BE TOO SANCTIMONIOUS ABOUT OUR NEW MODEL

Building a new model, especially one used for policy purposes, takes considerable time, effort, and resources. In justifying such expenditures, one inevitably spends a lot of time denigrating previous models. For example, in pitching the third Uniform California Earthquake Rupture Forecast (UCERF3) (http://www.WGCEP.org/UCERF3), criticisms of the previous model included fault-segmentation assumptions and the lack of multifault ruptures. In the context of including spatiotemporal clustering for operational earthquake forecasting (e.g., Jordan *et al.*, 2011), another criticism has been that previous candidate models not only ignore elastic rebound but also produce results that are antithetical to that theory. For instance, the short-term earthquake probabilities model (Gerstenberger *et al.*, 2005), which provided California aftershock hazard maps at the U.S. Geological Survey web site between 2005 and 2010, implies that the time of highest likelihood for any rupture will be the moment after it occurs, even for a big one on the San Andreas fault. Furthermore, Monte Carlo simulations imply that excluding elastic rebound in such models also produces unrealistic triggering statistics (Field, 2012).

Given all models are wrong, what we really hope is that any new model is more useful than its predecessors and that the value added exceeds the total development costs.

While UCERF3 includes solutions to these and other issues, it also embodies questionable assumptions and approximations of its own. For example, simplified rules are used to quantify which multifault ruptures are possible, including a 5 km distance threshold that allows faults separated by 4.999 km to rupture together, but not those separated by 5.001 km. Although undoubtedly incorrect, and likely to make the wrong call at some points in the fault system, this approximation is presumably better than excluding multifault ruptures altogether. Likewise, an epidemic-type aftershocks sequence (ETAS; Ogata, 1988) component has been added to represent spatiotemporal clustering in the elastic-rebound, finite-fault-based framework of UCERF3 (e.g., Field, 2012). However, ETAS is not how the Earth actually triggers events; at best it represents a useful statistical proxy for whatever physics governs real earthquake sequences.

Uttering the Box quote in this context represents both full disclosure and hopefully a degree of humility, especially in the aftermath of casting aspersions on previous models.

WE SHOULDN'T LET PERFECTION BE THE ENEMY OF A MORE USEFUL MODEL

As scientists, we naturally strive to maximize model elegance and range of predictability, and we typically delay publication until these are to our liking. Although generally admirable, such perfectionism can deprive stakeholders and interest groups of useful information. As any seismic hazard or risk analyst knows, you go with the model you have, not the one you might want or wish to have at a later time (paraphrasing Donald Rumsfeld's famous words regarding war).

Pointing out the technical problems with any model is easy, as exemplified above, and it is scientifically fun and stimulating to come up with possible solutions. Both occur in spades in large collaborations. Box's quote can, therefore, provide huge relief with respect to that-has-to-be-fixed sentiments, especially when a project deadline is looming. Given all models are wrong, what we really hope is that any new model is more useful than its predecessors and that the value added exceeds the total development costs.

WE HAVE A LONG WAY TO GO WITH RESPECT TO TESTING MODEL USEFULNESS

Building on previous efforts (e.g., Schorlemmer and Gerstenberger, 2007), the collaboratory for the study of earthquake predictability (CSEP) has made great strides in formally testing earthquake-forecast algorithms worldwide (e.g., Jordan, 2006; Zechar *et al.*, 2010). With respect to all models being wrong, Marzocchi and Jordan (2014) have already defended the meaningfulness of such tests, as well as the inclusion of subjective epistemic uncertainties, as long as whatever metric we are trying to predict adheres to various exchangeability criteria.

CSEP tests evaluate both reliability, meaning consistency with observations, and skill, meaning performance relative to

doi: 10.1785/02201401213

other models (nomenclature from Jordan *et al.*, 2011); this evaluation is currently done with respect to the location and frequency of earthquake nucleation. Are such tests informative with respect to evaluating risk-mitigation usefulness?

The smoothed seismicity-based forecast of Helmstetter *et al.* (2007) outperformed several others in a 5 year prospective test for California (Zechar *et al.*, 2013). Is this model, therefore, more useful than the others? It is based solely on observed seismicity, with the consequent implication that the Coachella Valley is relatively safe in terms of large earthquakes (due to low-seismicity rates there). Geology and paleoseismology, on the other hand, imply the San Andreas fault is locked and loaded in this valley, and UCERF3, consequently, gives this area the highest likelihood of hosting a large earthquake. Which model would you choose if you were designing a hotel for Palm Springs?

The point is that exhibiting reliability and skill in CSEP is not currently a sufficient condition for model usefulness, as the forecast could be totally misleading at the larger magnitudes that dominate risk. In fact, reliability and skill does not seem to be a necessary condition either, as one can imagine a fault-based model that doesn't even bother to forecast smaller earthquakes (thus failing current CSEP tests).

It still remains to be seen how long we will need to wait for definitive results on usefulness under various conditions.

Not only are all models wrong, but their relative usefulness varies depending on location and the specific loss metric of interest.

The basic problem is a lack of prospective test data at the magnitudes that dominate risk. This is not to say that CSEP tests are currently worthless. For one, quantifying model performance at small magnitudes is still scientifically useful in terms of understanding how the Earth works at those magnitudes. The tests will also be meaningful for any models that enlist small earthquakes to forecast larger ones, as reliability at the low end will be a necessary condition for reliability at the high end (but not sufficient because the assumed extrapolation may be misleading). Time will also bring more test data, as will expanding the spatial domain of forecast models, but it still remains to be seen how long we will need to wait for definitive results on usefulness under various conditions (a question I find myself asking at every CSEP meeting).

Another limitation of CSEP is the current focus on earthquake nucleation. State-of-the-art hazard and loss assessments don't even consider hypocenters, depending instead on proximity to the nearest rupture surface. Suppose you have two different models that are equally skilled and reliable in terms of forecasting the hypocenter of a large earthquake. However, one model says we will have unilateral rupture to the north, whereas the other says it will be unilateral to the west (e.g., because geologists argue over the orientation of the fault). The practical implications could be huge depending on where your site is located. Likewise, one can imagine a model that is dead wrong in terms of the hypocenter but actually more useful in terms of getting the closest point of rupture correct (e.g., because rupture nucleates on an unknown extension of a fault).

To complicate matters even further, model usefulness depends heavily on where you are and what you are concerned about (the risk metric). In other words, reliability and skill will differ depending on whether you are considering a single-family dwelling, a skyscraper, a nuclear power plant, a long-span bridge, lifelines crossing a fault, light rail, or a statewide portfolio of insurance policies. For example, imagine a densely populated basin bounded by strike-slip faults on two opposite sides (think San Bernardino). Geodesy gives a good constraint on the total deformation across the basin, but it cannot resolve which fault is more active, so you construct alternative models to acknowledge this epistemic uncertainty (all other things being equal). If we have a high-value asset sitting next to one of the faults, then the risk implications of the different models may vary tremendously, meaning one will have superior reliability and skill with respect to potential losses. However, if we have a portfolio of insurance policies distributed evenly across the valley, or an asset located exactly in the middle, then all the models will have the exact same reliability and skill with respect to potential losses.

Thus, not only are all models wrong, but their relative usefulness varies depending on location and the specific loss metric of interest. Ideally, CSEP would, therefore, move beyond testing earthquake nucleation and provide on-demand evaluations for any and all metrics of interest. Progress has already been made with respect to testing hazard (e.g., Stirling and Petersen, 2006; Stein *et al.*, 2015, and references therein), but we need to go beyond that to specific loss metrics. Expanding the scope of CSEP would take considerable resources, so any such expenditure should be weighed against the question of how long we will need to wait for definitive results. The only alternative, which may be more fruitful in the short term, will be to test the assumptions that go into our models via traditional science (outside the collaboratory). Until then, we will be forced to rely on expert judgment in situations where definitive tests are lacking, something even more physics-based models can't rescue us from.

The news is not all bad, as it implies we can sometimes, and without shame, get away with simplified models. For example, I'm certain that a model that assumes segmentation and excludes multifault ruptures will do just fine in some circumstances. In fact, there may be cases in which ignoring faults altogether, as in the global earthquake activity rate model (GEAR1, Bird *et al.*, submitted), may prove just as useful as a model like UCERF3. Likewise, perhaps elastic rebound can be left out of spatiotemporal clustering models in some situations. The point here is that we have only just begun thinking about quantifying such relative model usefulness.

ACKNOWLEDGMENTS

This paper was improved by thoughtful reviews by Ross Stein and Warner Marzocchi.

REFERENCES

Bird, P. D., D. Jackson, Y. Y. Kagan, C. Kreemer, and R. S. Stein. GEAR1: A global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity, *Bull. Seismol. Soc. Am.* (submitted).

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness, *J. Roy Stat. Soc. Ser. A* **143,** 383–430.

Field, E. H. (2012). Aftershock statistics constitute the strongest evidence for elastic relaxation in large earthquakes—Take 2, *2012 Meeting of the Seismological Society of America*, San Diego, California, 17–19 April 2012.

Gerstenberger, M., S. Wiemer, L. Jones, and P. Reasenberg (2005). Real-time forecasts of tomorrow's earthquakes in California, *Nature* **435,** 328–331.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007). High-resolution time-independent grid-based forecast for $M \geq 5$ earthquakes in California, *Seismol. Res. Lett.* **78,** 78–86.

Jordan, T. H. (2006). Earthquake predictability, brick by brick, *Seismol. Res. Lett.* **77,** 3–6.

Jordan, T. H., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2011). Operational earthquake forecasting: State of knowledge and guidelines for implementation, final report of the International Commission on Earthquake Forecasting for Civil Protection, *Ann. Geophys.* **54,** no. 4, 315–391, doi: 10.4401/ag-5350.

Ogata, Y. (1988). Statistical models of point occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.* **83,** 9–27.

Marzocchi, W., and T. H. Jordan (2014). Testing for ontological errors in probabilistic forecasting models of natural systems, *Proc. Natl. Acad. Sci. Unit. States Am.* **111,** 11973–11978, doi: 10.1073/pnas.1410183111.

Schorlemmer, D., and M. C. Gerstenberger (2007). RELM testing center, *Seismol. Res. Lett.* **78,** 30–36.

Stein, S., B. Spencer, and E. Brooks (2015). Metrics for assessing earthquake hazard map performance, *Bull. Seismol. Soc. Am.* (submitted).

Stirling, M., and M. Petersen (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,** 1978–1994.

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling, and T. H. Jordan (2010). The collaboratory for the study of earthquake predictability perspectives on computational earth science, *Concurrency Comput. Pract. Ex.* **22,** 1836–1847, doi: 10.1002/cpe.1519.

Zechar, J. D., D. Schorlemmer, M. J. Werner, M. C. Gerstenberger, D. A. Rhoades, and T. H. Jordan (2013). Regional earthquake likelihood models I: First-order results, *Bull. Seismol. Soc. Am.* **103,** 787–798.

Edward H. Field
U.S. Geological Survey
Denver Federal Center
P.O. Box 25046, MS 966
Denver, Colorado 80225-0046 U.S.A.
field@usgs.gov