Title. Critical questions about CSEP, in the spirit of Dave, Yan, and Ilya.

Authors.

Frederic Schoenberg, corresponding author, Department of Statistics and Data Science, 8125 Math-Science Building, UCLA Los Angeles, CA 90095-1554, USA. frederic@stat.ucla.edu.

Danijel Schorlemmer, Helmholtz Centre Potsdam, GFZ German Research Center for Geosciences, Telegrafenberg, D-14473 Potsdam, Germany.

Both authors contributed equally to this manuscript.

Abstract. In honor of our dear departed friends Yan Kagan, Dave Jackson and Ilya Zaliapin, we propose a selection of broad questions regarding earthquake forecasting and especially the Collaboratory for the Study of Earthquake Predictability (CSEP) in particular, and give our thoughts on their answers. This article reflects our opinions, not necessarily those of Yan Kagan, Dave Jackson, and Ilya Zaliapin, and not necessarily those of the seismological community at large. Rather than to provide definitive answers, we hope to provoke the reader to think further about these important topics. We feel that Dave Jackson in particular might have liked this approach and may have seen this as an appropriate goal.

Conflict of interest statement. The authors acknowledge there are no conflicts of interest recorded.

Keywords. CSEP, earthquake forecasting, earthquake prediction, ETAS, goodness-of-fit assessment, statistical seismology.

1. Is the Collaboratory for the Study of Earthquake Predictability (CSEP) worthwhile? How can it be improved?

While our response to the first question is unequivocally in the affirmative, we must first admit that CSEP has not yet come close to achieving all that was initially hoped of it. When CSEP was formed, many anticipated that the experiments would lead to clear and decisive improvements in earthquake forecasting, would indicate which models are superior and which are inferior, would highlight ways to improve models, and ultimately would lead to marked improvements in our ability to forecast large earthquakes. Perhaps some of these goals will be fulfilled in the future, but the steps so far in these directions have been very small.

It can, at times, seem unclear if anyone is actually making any use of CSEP results in practice, and how much seismic hazard models have been improved using CSEP results. While the Uniform California Earthquake Rupture Forecast 3 (UCERF3) model (Field et al. 2017) has included the model of Helmstetter et al. (2007), which performed best in the 5-year RELM experiment (Schorlemmer et al. 2010, Zechar et al. 2013, Bayona et al. 2022), as one branch for the background seismicity part, UCERF3 remains driven by the fault-based forecasts of the large earthquakes (Field et al. 2017). CSEP has never, to our knowledge, influenced the fault-based parts of hazard models that provide the bulk of the hazard.

On the other hand, from a statistical standpoint, CSEP is a most remarkable scientific achievement. It is a kind of gold standard that other areas of statistical application can only dream of achieving. In wildfire forecasting, for instance, many of the proposed models are not even well-defined and would result in 0 likelihood given data. In epidemiology, the models most often used to forecast the spread of diseases like Ebola or Covid-19 are at least 50 years old and scant attention is given to their goodness-of-fit (Kresin et al. 2021). Further, one cannot reasonably expect government officials and industrial practitioners instantly to make use of scientific advances. Scientific progress has almost always been painfully slow and methodical. There are so many possible examples here, but a recent one is Dave Jackson and Yan Kagan debunking the characteristic earthquake hypothesis (Kagan et al. 2012). It has typically taken decades at least for most scientists and professionals adequately to accept and employ the results of careful scientific work. It might be a bit naive to expect practitioners and other researchers to adjust quickly to reports that a given model does not fit well to data. Such testing is definitely progress nonetheless.

Analyses of CSEP results have pointed out that Epidemic-Type Aftershock Sequence (ETAS) models tend to fit well, at least on relatively short-term scales. ETAS models were initially proposed by Ogata (1988) to describe the times and magnitudes of earthquakes, and were subsequently extended to model spatial-temporal-magnitude catalogs in Ogata (1998). Subsequently, a host of slight modifications have been proposed (e.g. Sornette and Sornette 1999, Helmstetter and Sornette 2002, Console et al. 2003, Ogata et al. 2003, Ogata 2004, Ogata and Zhuang 2006, Marzocchi and Lombardi 2008, Ogata 2011, Zhuang 2012, Nandan et al. 2017, Grimm et al. 2022, Iacoletti et al. 2022, Li and Pu 2022, Aso and Terai 2023). One major problem, however, is that models such as ETAS typically do not help much in forecasting the earthquakes we are most interested in. As their name indicates, ETAS models are mostly useful for describing the frequency and spatial-temporal distribution of aftershocks one expects to see following large earthquakes, or perhaps as a null model to which alternative models might be compared. However, for purposes of planning, public safety, building codes, and most other purposes, what is really sought is the accurate forecasting of the very largest events, or at least the estimation of their long-term frequency, and when it comes to these tasks, most versions of ETAS seem to be scarcely better than a simple homogeneous Poisson model. Indeed, most formulations of the ETAS model assume a Gutenberg-Richter distribution of earthquake magnitudes with the magnitude of each earthquake drawn independently of what occurred previously, and thus essentially the model makes no effort to pinpoint where or when the rate of the largest earthquakes may be higher relative to the rate of smaller events.

This may be an area where more statistical work can be of assistance. Current methods for assessing the fit of earthquake forecast models emphasize overall measures of fit such as the log-likelihood, or total number of events, or other summaries that do not adequately take into account what aspects of the model we care most about. If, for instance, we care exclusively about the model's ability to forecast the largest events, then it may be appropriate to choose a goodness-of-fit measure that properly emphasizes this feature. For instance, suppose one is given data on the times, locations, and magnitudes of n events, (t_i, x_i, y_i, m_i) for $i = 1,\ldots, n$, and let $\lambda(t,x,y,m)$ denote the modeled conditional intensity at spatial-temporal-magnitude (t, x, y, m), with λ_i representing the conditional intensity at point i. The log-likelihood,

$$L = \sum_{i=1}^{n} log\ \lambda_i - \int \lambda(t,x,y,m)\ dtdxdydm,$$

has a first term that properly rewards the model for accurately forecasting earthquakes (where by "accurately forecasting", we mean positing a high value of λ where an earthquake ends up occurring), and a second term that punishes the model for having high values of λ elsewhere. However, the log-likelihood essentially rewards the model equivalently for accurately forecasting a magnitude 3 event or a magnitude 7 event, despite the fact that forecasting the latter event is so much more of interest.

Since forecasting the largest 5% of magnitudes are of most interest, one could instead use, as a measure of fit, a summary such as the quotient

$$Q = \frac{mean\{\lambda_i : m_i > m^{[.95]}\}}{mean\{\lambda(t,x,y,m)\}},$$

for example, where $m^{[.95]}$ is the 95th percentile of the magnitude distribution, perhaps estimated based on prior seismicity, and the denominator mean{λ(t,x,y,m)} may be estimated e.g. using several thousand locations selected at random from the space-time-magnitude observation region. The higher the value of Q, the better the model appears to be forecasting the spatial-temporal locations of the largest 5% of events. For a homogeneous Poisson model with uniform magnitude density, Q will be close to 1. Any model that adequately accounts for the spatial inhomogeneity of seismicity will have $Q > 1$, as it should since such a model will tend to vastly outperform a homogeneous Poisson model at forecasting the largest events. Among competing models, the model that forecasts the larger events more accurately will tend to be the model with higher Q, especially if all the models are similarly calibrated overall [i.e. mean{λ(t,x,y,m)} is close to the overall rate of seismicity] which can readily be checked via other methods, such as the N-test.

Forecasting large earthquakes as point sources as CSEP defines them for the purpose of testing bears its problems. Modelers can employ the knowledge of faults and distribute the hypocenter probabilities along the fault but this is making a model fit to the test design and not the (preferred) other way around. This gets us back to the problem of whether CSEP can influence the fault-based forecasts of hazard models. Clearly, adequate testing procedures for such models are needed and were discussed many times within the CSEP community.

However, the unambiguous identification of fault segments ruptured in earthquakes is already a problem, not to mention the incompleteness of fault models. This is an unfortunate disconnect between common practice in hazard modeling and testing possibilities.

Another problem with some CSEP results is that the observation that model A fits better than model B does not necessarily directly tell us how to improve either of the models. Further, if model A offers superior fit to model B over a 5-year period, it is unclear whether this means model A is likely to outperform model B in the future. If not, then what does testing tell us? We have often observed that models use any seemingly suitable statistical distributions in fitting and then use such fitted models for forecasting. In one paper about testing intensity-prediction equations, it was shown that the models fitted to functions reproducing basic physical principles of wave propagation have higher forecasting capabilities [Mak et al. 2015]. But is CSEP able to discriminate between overfitting and physics-based fitting without very long forecast experiments in which the former are likely to fail compared to the latter? Furthermore, there is a general tendency to increase the complexity of models with more and more parameters. Given that in CSEP all models' forecasts are fully specified with zero degrees of freedom, the number of input parameters does not count into the result. But should it maybe? Again, an overfit model temporarily might fit extremely well or a model might simply have been lucky to fit well in the short term (Sornette et al. 2019), but in the long run the fit will likely deteriorate.

While this criticism may be valid for many analyses of goodness-of-fit, residual methods could perhaps help here, if they can provide useful graphics highlighting where exactly model A seems to outperform model B. Voronoi deviance residuals and super-thinned residuals (Clements et al. 2012, Bray and Schoenberg 2013, Bray et al. 2014, Gordon et al. 2015) seem potentially useful in this regard. Learning how to improve a complex model is no easy task, so even a suggestion of a rather minor improvement should perhaps be seen as a major achievement from a statistical procedure. As for the possibility of overfitting, or for model A to outperform model B in CSEP over several years but not in the long term for whatever reason, while such possibilities may be inevitable, CSEP seems ideally suited to handle these types of problems. Overfitting and lack of reproducibility are enormous problems in conventional research involving retrospective analyses of data. With CSEP and its strict insistence on truly prospective testing, these problems may still exist but they are minimized as much as possible.

2. Are ETAS models valuable?

ETAS models seem frequently to offer the best fit among the proposed models for earthquake occurrences. While so many other models have been suggested based on retrospective analyses, the fact that ETAS performs well in the sense of offering satisfactory fit to data even when used prospectively is extremely impressive, and is solid evidence of its value.

However, as mentioned previously, most versions of ETAS have very little value for forecasting the largest events. (There are exceptions, however, that treat large earthquakes fundamentally differently from smaller ones, such as Nandan et al. 2019, 2022). Furthermore, what exactly is ETAS telling us that goes beyond the heuristic notion that large earthquakes are followed by aftershocks and possibly even larger events? Reasenberg & Jones (1989) have decades ago quantified the chance of a larger earthquake following an already large shock. Does ETAS tell us significantly more despite the much heavier computational load and its complexity? ETAS provides higher resolution, but in what sense is this really useful? Does it change decisions in risk mitigation? Does it allow us to call an area "safe" earlier?

An important topic that has been insufficiently explored is how to quantify a model's *value*, for forecasting. Much attention has been paid to the quantification of a model's goodness-of-fit to data, and this is certainly a component of a model's value, since the better a model fits, the more confidence one has in its forecasting ability. However, goodness-of-fit does not tell the whole story. While ETAS models, for instance, may fit very well to catalogs of earthquakes including aftershocks, and while forecasts can be obtained using ETAS via simulations (Omi et al. 2014, Shcherbakov et al. 2019, Petrillo and Zhuang 2024) or multi-element probability formulas and other techniques (Ogata 2017a, Ogata 2022, Ogata2024), most versions of the ETAS model typically have little value for forecasting the largest earthquakes, which happen to be the ones we care most about (in synthetic tests, Helmstetter and Sornette 2003 quantify that ETAS allows one to forecast about 20% of the largest events). We should explore alternative measures, such as the measure Q proposed above, or variants of the Brier score or information gain restricted to the subset of earthquakes of most concern, perhaps weighting earthquakes differentially based on their energy release, damage potential, or other measures. In particular, it would be of great interest to agree on a measure of a model's forecasting value as a function of time horizon. It may be, for instance, that ETAS has excellent forecasting value for forecasting seismicity several hours or days into the future, but practically no value at forecasting several months or years into the future. Other models possessing the opposite qualities might exhibit worse fit to data overall, yet have more forecasting value in many situations. We believe that including better quantifiers of a model's value into CSEP would be a great step toward discovering and raising awareness about alternative models that are potentially more useful.

The ETAS model should perhaps be seen as the null model, to which alternatives could be proposed and compared. Something similar was proposed by Stark (1997). Indeed, in the original paper proposing ETAS, Ogata (1988) actually used ETAS as a kind of null model, rather than a model to be directly used for forecasting. He identified times of quiescence, which were essentially times when the model appeared to fit poorly, as potential indicators of an impending future large event (even though the quiescence hypothesis has been debunked by van Stiphout et al. 2011 employing a declustering algorithm using Southern California data).

There still seems to be some work to do at constructing a suitable null model, however. Numerous models have been proposed, starting from the silly spatially uniform model (easy to beat), via simple Poisson smoothed seismicity models, all the way to time-varying ETAS models. Each model needs to be somehow calibrated and this creates a plethora of different model flavors for each basic concept. There is presently not a single ETAS model, but rather a host of different varieties, parameterizations, and implementations. Therefore, more thought should go into how to create a suitable version of a null model, and what the requirements should be for such a model.

3. Is ETAS the end?

On the one hand, looking at the collection of recent publications, the answer seems to be yes. Few genuinely new model classes have recently been introduced to capture the time dependence of earthquakes. Instead, more and more flavors of ETAS models have been developed, mainly by attempting to fit the ETAS concept to local, regional, or sequence-specific datasets, perhaps because of the 'because we can' effect: people doing what they know they can achieve even if it is unlikely to lead anywhere substantial. Out of these many ETAS flavors, no consensus model has been selected by the wider community. ETAS seems to reproduce itself constantly without adding significant improvements to a solution of the earthquake forecasting problem. Furthermore, it may be that earthquakes are a natural phenomena too complex to be modeled or forecast accurately, in CSEP or any type of forecasting experiment. No predictability in the pattern of earthquake magnitudes has been discovered, or at least none that has been consistently reproduced. It is possible, as Yan, Dave and their collaborators posited, that earthquakes may be inherently unpredictable, and simply cannot be predicted (Geller et al. 1997, Kagan 1997a).

On the other hand, there may be room for optimism. Some promising recently proposed versions of ETAS take into account focal depth and rupture geometry, for instance (Guo et al. 2015, 2018, 2021, 2024), as well as other physical components such as the magnitude-dependent Omori law (Sornette and Ouillon 2005, Ouillon and Sornette 2005, Ouillon 2009, Tsai et al. 2012) or a two-branched Gutenberg-Richter distribution (Saichev and Sornette 2005, Nandan et al. 2019), but see also Petrillo and Zhuang (2023) for the opposite opinion. Also, while ETAS may fit best among existing models, it has not been shown that ETAS fits better than any *possible* alternative model. Certainly any simple model forecasting the precise occurrence of future seismicity is highly elusive, but this does not mean improvements will not be forthcoming. Scientific progress has often been slow and laborious, and seismology is no exception. Perhaps increased emphasis on *value*-based goodness-of-fit statistics, as described above, may reasonably be anticipated to yield improved models that optimize these more sensible and practical criteria, and CSEP has the potential to be an important leader in this aim. If we increase our focus on models that improve forecasting in useful ways, and reward models for improving forecasting in useful ways, such models are more likely to be discovered. As they say in *Field of Dreams*, "If you build it, he will come."

4. What are the best ways forward for earthquake forecasting?

This may be divided into practical, technical approaches as well as the overall big picture, and we start with the former. While there may be advantages to looking at local results in some situations, testing generally should be done on a global scale to increase the power of the tests so we can have some confidence in the results. Only globally can we test a sufficiently large number of significantly large earthquakes that matter. Yan Kagan consistently advocated this (Kagan 2003, Kagan and Jackson 1991, Kagan and Knopoff 1981, 1987, Kagan et al. 2012). CSEP has, after its inception, expanded from testing earthquake forecasts in California to further testing regions in Japan, Italy, New Zealand, the western Pacific and finally to global tests. However, the majority of tested models were developed for the regional testing regions and only a few are being tested globally.

All testing regions have been defined on 0.1 degree longitude/latitude grid cells (Schorlemmer and Gerstenberger 2007). While this resolution has been chosen to match the location uncertainty of local events, it backfired in the global experiment in which the testing area consists of 6.4+M cells and 100+M bins. However, the grid size should be scaled according to how much data one has. With millions of cells but just a handful of earthquakes, the test has little to no power, so the grid cells should be chosen adaptively in order to maximize the power of the tests. One pathway to solve this problem is to use a multi-resolution grid, e.g. the Quadtree approach taken by Ogata et al. (1996) and Asim et al. (2023), or using Delaunay tessellations (Ogata et al. 2003, Ogata et al. 2019). The forecasts can have high spatial resolution (small cells) in areas with many earthquakes and low resolution in areas with few earthquakes (large cells). This way, meaningful global forecasts can be provided with a few ten thousand of cells, matching the number of observations in the Global CMT catalog.

As far as the big picture and where and whether scientific progress is likely in the future, when we first got interested in earthquakes, we imagined that the historical and modern, high-resolution earthquake record might contain hidden information about when the next big one would occur, and that if we just looked hard enough, we could find some pattern or signal and successfully predict the next major event. Now we realize that was naive, and the Earth tends not to resemble a cartoon villain leaving obvious clues about future calamities. Many natural phenomena are exceedingly complex, and earthquakes seem to fit this mold, so there seems little hope at finding some simple pattern that predicts when the next big event will occur. On the other hand, there is something rather mysterious about large earthquakes. They are, after all, major ruptures of the Earth, and sudden releases of enormous amounts of tectonic energy. And they must be triggered by *something* and must experience a preparation phase that should leave some hints to be observed. Does it not stand to reason that, as our knowledge of the Earth's structure deepens, we ought to be able to figure out and possibly anticipate, or at least see some warning signs, of what is triggering these gigantic outbursts? Naive as it may be, we still believe that a more precise understanding of the Earth's structure should yield an improved ability to forecast major events.

This highlights another possibly important way that CSEP could be improved going forward: by incorporating other types of signals apart from earthquake occurrences. If the information content of earthquake catalogs does not allow for powerful tests and/or useful models, CSEP should reach out to model developments that include further signals that can be observed; potentially important efforts have been made in incorporating such signals recently (Zhuang et al. 2005, Han et al. 2016, Kumazawa et al. 2016, Freund et al. 2021). CSEP has already included models that use strain data as input, however these data have not been included as an authoritative data source to ensure all models use the same strain data. Of course, modelers can technically include any type of data in their models, but only if these data are authoritative and provided by CSEP can comparative testing become really meaningful and ensure full reproducibility of all forecasts generated by the model.

5. Can we use CSEP test results to improve models? Does CSEP provide a useful feedback loop?

The L-test and similar results typically assess the performance of the entire model. However, models are typically compiled of different ingredients and their interplay can be complex. We usually do not know if all components work well and contribute correctly; maybe one component is not calibrated well and lowering the overall performance of the model, but test results rarely indicate this. Instead, each metric typically tests a different feature of a forecast, not a component of the underlying model. But do we know what part of the model causes the feature to perform well or not? Do we know how to translate results of a specific metric into model improvement? The answer is often no!

On the other hand, it definitely seems that, if any substantial progress is ever made at forecasting seismicity, it is going to be largely the result of very careful and intricate model evaluation. Put another way, it seems very unlikely we will ever have substantial improvement without excellent model evaluation experiments like CSEP. Without this kind of rigorous look at the models, seismology would be doomed to keep repeating the mistakes of the 20th century, where model after model was proposed, based on retrospective analysis, only to find out later that the models did not do well prospectively (Kagan and Jackson 1991, Geller et al. 1997, Kagan 1997a, Jackson and Kagan 2006).

The situation is somewhat analogous to examples from statistics in medicine, where before the proper emphasis was placed on randomized controlled experiments, it was almost impossible to tell whether a drug or procedure was effective, and the literature was full of reports promoting procedures like the portacaval shunt, when later experiments clearly showed them to be ineffective (Freedman et al. 1998 very nicely summarizes such examples).

In seismology, with CSEP already firmly in place, the field is poised to move in a positive direction. Even though progress might be slow, we at least have a system in place that could identify improved models once they are proposed, and we have a mechanism for more efficiently sifting through and debunking poorer models. However, the power in medical tests are often much higher, and the success criteria clearer, compared to earthquake forecasting tests.
Will CSEP be as successful for earthquake forecasting as double-blind tests have been for medical research? That remains to be seen.

As mentioned previously, statistical methods for model evaluation still need to be improved so they can more readily lead to model improvements. Graphical methods, such as the smoothed residual field (discussed by Baddeley et al. 2005 for the purely spatial case), Voronoi residuals (Bray et al. 2014) and superthinned residuals (Clements et al. 2012), rather than numerical summary statistics, seem to be the most promising for suggesting model improvements. With superthinned residuals, points are added or removed at random according to the model, so that in the end the residual points should be uniformly scattered if the model fits well, and departures from uniformity indicate places where the model fit poorly. With Voronoi residuals, the domain is divided into cells, one cell for each earthquake, such that each cell consists of all locations closer to the corresponding earthquake than to any other earthquake. The fact that residuals on such an adaptive grid allow one to pinpoint around which earthquakes a model fits well or poorly can sometimes lead directly to ideas for model improvement (Clements et al. 2011, Gordon et al. 2015). However, these methods have their problems as well. For one thing, given a host of competing models, it is cumbersome and difficult to examine a collection of plots, and so for comparing many models, sometimes the simplicity of a numerical summary is desirable. Second, Voronoi residuals might be good for 2-dimensional data, but it remains unclear how exactly to use them for the 3-d case, or even just for 2 spatial dimensions and time. Currently, one typically just ignores time and depth and carves out Voronoi cells using just the epicenters of earthquakes, but this should be improved. Superthinned residuals are easier to implement and do not have the dimensional problem of Voronoi residuals, but because of the randomness introduced in the generation of these residuals, often it is difficult to discern clear patterns or ways to improve models from the superthinned residuals alone. Perhaps we still need more improvements in the realm of visual summaries of goodness-of-fit for point process models.

Some of the current tests in CSEP really are not useful for comparing competing models, and should probably be removed from CSEP. If you are evaluating just one model, then maybe the N-test, L-test, S-test, etc. might be useful (Schorlemmer et al. 2007, Zechar et al. 2010), but when comparing 2 or more models, these tests can be very misleading, since a poorly-fitting but highly variable model will often have a higher p-value than a competing model that actually fits better. So, since CSEP is mainly for model comparison, perhaps going forward we should focus on methods that are better for this purpose. If an overall numerical measure is desired, we think the log-likelihood score is probably best (Ogata 2017b, 2024), at least among previously proposed measures, although the statistic Q mentioned in Section 1 seems potentially more informative. We should probably be content with a summary of the overall fit and not be too concerned with p-values, for model comparison purposes.

6. How much information is in the system and how well is CSEP prepared to harvest it?

This is a tough question to answer. Dave Jackson believed that if strike angle estimates could be improved in the future using more accurate seismometers, we could hope for some substantial information to be gained in that area, since geophysical theory seems to suggest such moment tensors should be extremely useful, though presently their use seems to be somewhat limited. The main information we have seems to be the seismic record, which of course also has errors and missing data, but this has quite thoroughly been studied especially by Kagan (2003, 2004), and obviously is a wealth of information. Certainly the paleoseismic record provides some useful information, and potentially geodetic information should be useful to improve our models as well. While we might be somewhat pessimistic, we must admire the positivity of Ilya Zaliapin, who never seemed to have any doubt that accurate earthquake forecasting should be possible, and whose excellent paper with Ben-Zion highlights some promising avenues to explore, such as localization of seismicity before large events (Ben-Zion and Zaliapin 2020).

The community has used the seismic record since decades to forecasts earthquakes but we have not moved beyond rough statistical methods that are the more successful the smaller the magnitude. Having said this, it seems that the overall number of events per magnitude is the key criterion for successful forecasting. A consequence of this statement would be that we have to wait very very long until we better forecast large events and that the forecast horizons for these events will also be very long.

CSEP seems to be ideally prepared to harvest this type of information. CSEP offers researchers a unique way to use their models to generate actual prospective forecasts in real time, and to evaluate those forecasts as accurately as possible. One thing CSEP really needs going forward is easy public access to these forecasts; currently they seem difficult to obtain for some reason. In addition, we still face the issue of the meaning of test results and what they tell us abut how to improve a model or which part of the model is right or wrong.

7. Can CSEP save us from bogus Articifical Intelligence (AI) forecasts?

It is to be expected that some AI researchers will attempt to forecast earthquakes by feeding whatever they find into their AI algorithms. In addition, we pessimistically anticipate many papers where research simply average over multiple models, possibly in a Bayesian way, forming complex ensemble models that perhaps fit well retrospectively but rarely prospectively, and we do not foresee this being a useful or productive enterprise because the increasingly complex ensemble models become increasingly difficult to improve using residual analyses. In fact, in some sense the idea behind Bayesian ensemble models runs counter to the central idea behind CSEP, which is to evaluate models prospectively and rigorously in order to improve them and to distinguish the ones that seem to fit well from those that do not. Rather than attempting to modify or improve individual aspects of a model, Bayesian model averaging seems to make problems with the models more obscure rather than clearer, and models that should be discarded become instead incorporated into the ensembles, like rotten fruit in a milkshake. Even worse, these kind of models do not contribute to our understanding of the physical processes and how they can be modeled in a useful way. They are merely statistical stunts that are likely doomed to fail sooner rather than later in prospective tests. And even if not, they do not provide any insight that let us improve models that we understand and that mimic the physics we have discovered. CSEP protocols might be the only guardian preventing us from false declarations of success, as many researchers do not fully appreciate the necessity for truly prospective tests.

But what if the CSEP tests cannot provide any conclusive results? How can the wider CSEP community argue against the plethora of AI forecasts that are expected? Serious thinking about the power and value of CSEP tests is needed and needs to be codified somehow in a community process.

On the other hand, CSEP forecasts should be made easily publicly available to welcome further analysis of the results, even if this may spur faulty analyses of the results using AI, deep learning, Bayesian ensemble modeling, etc. Perhaps if CSEP eventually has so many models that it becomes overwhelming, this could be a problem. But otherwise, the fact that CSEP is fully prospective seems to guard against overfitting which is a main problem with AI and deep learning models, and machine learning algorithms could be very useful to incorporate information from various sources, as mentioned above. CSEP forecasts need to be readily available online to enable easy statistical assessment because the statistical analysis of the results will likely be very useful.

8. Are CSEP tests powerless?

On the one hand, we have seen that several of the tests, including the S-test, are not reliably able to reject uninformative models on the global high-resolution grid with 6.5M cells and many more bins because even the longest catalogs contain too few events (Asim et al. 2013, Khawaja et al. 2013). In addition, it is unclear how to measure the power of a test if the true distribution is unknown. It seems that CSEP is caught in the dilemma of either producing powerful tests with low resolution that provide rather meaningless results or have powerless tests at high resolution, again resulting in meaningless results. How can CSEP find the sweet spot and what does this sweet spot tell us about the overall information content of earthquake occurrences and their testing? Much could potentially be gained by finding a formulation that describes the overall information content and the resulting best-possible power of CSEP tests and translate this into uncertainties of the forecasts or the test results. Currently, it is unclear how to use CSEP results and to what extent to trust them given the open question of the power of these tests. Quantifying the power/information content could also help planning for more powerful tests by estimating what amount of input data is needed or how to otherwise compensate with other types of data.

On the other hand, if we include data on lower magnitude events, we actually have plenty of data already and in this context power is not a serious problem. This gets at one of the key philosophical issues in seismology: do the largest earthquakes obey the same fundamental properties as the moderate and small events?

If not, then it might be centuries before we can forecast large earthquakes in any meaningful way. And there are some reasons to be skeptical, especially since so few large events have been observed, their behavior seems difficult to characterize simply, and while small events can reasonably be approximated as point sources, large events can rupture hundreds of kilometers, so reducing them to point sources as done in CSEP may be unrealistic. Incidentally, CSEP's attempts to introduce fault-based testing have so far badly failed as no earthquake-to-fault segment association has ever been developed that could reasonably run without case-by-case human intervention and decision.

But if so, then if we can find models to forecast the moderate earthquakes accurately using the small to moderate events, these same models should be able to forecast the largest earthquakes as well using moderate events. This seems likely and promising, especially since we already have so much data on the moderate and smaller events. Yan Kagan and Dave Jackson were critical of those who kept constantly claiming that we had insufficient data to reject the characteristic earthquake hypothesis, at least for the largest events (Kagan 1997b, Jackson and Kagan 2006). We should not make this same mistake going forward. We have plenty of data. We just need to be scientific about how we use it.

Acknowledgements. We are grateful for our many wonderful conversations and collaborations with Yan Kagan, Dave Jackson, and Ilya Zaliapin.

This material is based upon work supported by the National Science Foundation under grant number DMS-2124433.

Declaration of Interests: The authors declare no competing interests.

Data and resources. No data or special resources were used in this article.

References.

Asim KM, Schorlemmer D, Hainzl S, Iturrieta P, Savran WH, Bayona JA, Werner MJ (2013). Multi-Resolution Grids in Earthquake Forecasting: The Quadtree Approach. *Bull. Seismol. Soc.Am.*, 113(1), 333-347, 10.1785/0120220028.

Aso N, Terai N (2023). Modifications of epidemic-type-aftershock-sequence models for characterizing diffusive shear slips of deep long-period earthquakes. *Geophysical Journal International* 234(2):1254-67.

Baddeley A, Turner R, Møller J, Hazelton M (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67(5):617-66.

Bayona JA, Savran WH, Rhoades DA, Werner MJ (2022). Prospective evaluation of multiplicative hybrid earthquake forecasting models in California. *Geophysical Journal International* 229(3):1736-53.

Ben-Zion Y, Zaliapin I (2020). Localization and coalescence of seismicity before large earthquakes. *Geophysical Journal International* 223(1):561-83.

Bray A, Schoenberg F (2013). Assessment of point process models for earthquake forecasting. *Statistical Science* 28(4), 510-520.

Bray A, Wong K, Barr CD, Schoenberg F (2014). Voronoi cell based residual analysis of spatial point process models with applications to Southern California earthquake forecasts. *Annals of Applied Statistics* 8(4), 2247-2267.

Clements, R.A., Schoenberg, F.P., and Schorlemmer, D. (2011). Residual analysis for space-time point processes with applications to earthquake forecast models in California. *Annals of Applied Statistics* 5(4), 2549--2571.

Clements RA, Schoenberg F, Veen A (2012). Evaluation of space-time point process models using super-thinning. *Environmetrics* 23(7), 606-616.

Console R, Murru M, Lombardi AM (2003). Refining earthquake clustering models. *JGR* 108 B10 2468, doi: 10.1029/2002JB002130.

Freedman D, Pisani R, Purves R (1998). Statistics. W.W. Norton and Co., New York.

Field EH, Milner KR, Hardebeck JL, Page MT, van der Elst N, Jordan TH, Michael AJ, Shaw BE, Werner MJ (2017). A spatiotemporal clustering model for the third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): Toward an operational earthquake forecast. *Bulletin of the Seismological Society of America* 107(3):1049-81.

Freund F, Ouillon G, Scoville J and Sornette D (2021). Earthquake precursors in the light of peroxy defects theory: critical review of systematic observations, Global Earthquake Forecasting System (GEFS) Special Issue: Towards Using Non-seismic Precursors for the Prediction of Large Earthquakes, Eur. Phys. J. Special Topics 230, 7-46. (https://doi.org/10.1140/epjst/e2020-000243-x) (https://arxiv.org/abs/1711.01780)

Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997). Earthquakes cannot be predicted. *Science* 275(5306):1616-1616.

Gordon JS, Clements RA, Schoenberg F, Schorlemmer D. (2015). Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. *Spatial Statistics* 14b, 133-150.

Grimm C, Käser M, Hainzl S, Pagani M, Küchenhoff H (2022). Improving earthquake doublet frequency predictions by modified spatial trigger kernels in the epidemic-type aftershock sequence (ETAS) model. *Bulletin of the Seismological Society of America* 112(1):474-93.

Guo Y., Zhuang J., Zhang H. (2024) Statistical modeling of 3D seismicity and its correlation with fault slips along major faults in California. Earth and Planetary Science Letters, 638: 118747. doi:10.1016/j.epsl.2024.118747

Guo Y., Zhuang J.and Zhang H. (2021). Heterogeneity of aftershock productivity along the mainshock ruptures and its advantage in improving short-term aftershock forecast. Journal of Geophysical Research: Solid Earth, 126:e202JB020494. doi:10.102/202JB020494.

Guo Y., Zhuang J. and N. Hirata (2018). Modeling and forecasting 3D-hypocenter seismicity in the Kanto region. Geophysical Journal International, 214: 520-530. doi:10.1093/gji/ggy154.

Guo Y., Zhuang J. and Zhou S. (2015). A hypocentral version of the space-time ETAS model. Geophysical Journal International, 203: 366-372. doi: 10.1093/gji/ggv319.

Han P, Hattori K, Zhuang J, Chen CH, Liu JY, Yoshida S (2016). Evaluation of ULF seismo-magnetic phenomena in Kakioka, Japan by using Molchan's error diagram, Geophys. J. Int. 208 (1), https://doi.org/10.1093/gji/ggw404 .

Helmstetter A and Sornette D (2002). Sub-critical and supercritical regimes in epidemic models of earthquake aftershocks. *J. Geophys. Res.* 107(B10), 2237.

Helmstetter A and Sornette D (2003). Predictability in the ETAS Model of Interacting Triggered Seismicity, J. Geophys. Res., 108, 2482, 10.1029/2003JB002485.

Iacoletti S, Cremen G, Galasso C. (2022). Validation of the epidemic-type aftershock sequence (ETAS) models for simulation-based seismic hazard assessments. *Bulletin of the Seismological Society of America* 93(3):1601-18.

Jackson DD, Kagan YY (2006). The 2004 Parkfield earthquake, the 1985 prediction, and characteristic earthquakes: Lessons for the future. *Bulletin of the Seismological Society of America* 96(4B):S397-409.

Kagan YY (1997a). Are earthquakes predictable? *Geophysical Journal International* 131(3):505-25.

Kagan YY (1997b). Statistical aspects of Parkfield earthquake sequence and Parkfield prediction experiment. *Tectonophysics* 270(3-4):207-219.

Kagan YY (2003). Accuracy of modern global earthquake catalogs. *Physics of the Earth and Planetary Interiors* 135(2-3):173-209.

Kagan YY (2004). Short-term properties of earthquake catalogs and models of earthquake source. *Bulletin of the Seismological Society of America* 94(4):1207-28.

Kagan YY, Jackson DD (1991). Seismic gap hypothesis: Ten years after. *Journal of Geophysical Research: Solid Earth* 96(B13):21419-21431.

Kagan YY, Jackson DD, Geller RJ (2012). Characteristic Earthquake Model, 1884-2011, RIP. *Seis. Res. Lett.* 83(6), 951-953.

Kagan YY, Knopoff L (1981). Stochastic synthesis of earthquake catalogs. *J. Geophys. Res. Solid Earth* 86(B4), 2853-2862.

Kagan YY, Knopoff L (1987). Statistical short-term earthquake prediction. *Science* 236(4808), 1563-1567.

Khawaja AM, Hainzl S, Schorlemmer D, Iturrieta P, Bayona JA, Savran WH, Werner M, Marzocchi W (2013). Statistical power of spatial earthquake forecast tests, *Geophys. J. Int.*, 233(3), 2053-2066, 10.1093/gji/ggad030.

Kresin C, Schoenberg F, Mohler G (2021). Comparison of Hawkes and SEIR models for the spread of Covid-19. *Advances and Applications in Statistics,* 74, 83-106.

Kumazawa T, Ogata Y, Kimura K, Maeda K, Kobayashi A (2016). Background rates of swarm earthquakes that are synchronized with volumetric strain changes. Earth and Planetary Science Letters 442:51-60. doi:10.1016/j.epsl.2016.02.049.

Li Y, Pu W (2022). Analyzing the 2020 Mw 6.4 Puerto Rico Earthquake Sequence Based on the Epidemic-Type Aftershock Sequence Model. *Seismological Research Letters*. 93(2A):609-19.

Mak S, Clements RA, Schorlemmer D. (2015). Validating Intensity Prediction Equations for Italy by Observations, *Bull. Seismol. Soc. Am.*, 105(6), 2942-2954, 10.1785/0120150070.

Marzocchi W, Lombardi AM (2008). A double branching model for earthquake occurrence. *JGR* 113(B08317).

Molyneux J (2018). Estimation of Spatial-Temporal Hawkes Models for Earthquake Occurrences. PhD thesis, University of California, Los Angeles, pp1-75.

Nandan S, Ouillon G, and Sornette, D (2019). Magnitude of earthquakes controls the size distribution of their triggered events, Journal of Geophysical Research - Solid Earth 124 (3), 2762-2780.

Nandan S, Ouillon G, Wiemer S, Sornette D. (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to California. *Journal of Geophysical Research: Solid Ea*rth 122(7):5118-43.

Nandan S, Ouillon G, and Sornette D (2022). Are large earthquakes preferentially triggered by other large events? Journal of Geophysical Research - Solid Earth 127, e2022JB024380.

Ogata Y (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* 83, 9-27.

Ogata Y (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics* 50, 379-402.

Ogata Y (2004). Space-time model for regional seismicity and detection of crustal stress changes. *Journal of Geophysical Research* 109(B3), B03308, doi: 10.1029/2003JB002621.

Ogata Y (2011). Significant improvements of the space-time ETAS model for forecasting of accurate baseline seismicity. *Earth, Planets and Space* 63(3), 217-229.

Ogata Y (2017a) Forecasting of a large earthquake: an outlook of the research, SRL 88(4) 1117-1126, 88 (4), 1117-1126, doi:10.1785/0220170006.

Ogata Y (2017b) Prediction and validation of short-to-long-term earthquake probabilities in inland Japan using the hierarchical space-time ETAS and space-time Poisson process models, Earth, Planets and Space, 74(1), https://doi.org/10.1186/s40623-022-01669-4 .

Ogata Y (2022) Prediction and validation of short-to-long-term earthquake probabilities in inland Japan using the hierarchical space-time ETAS and space-time Poisson process models, Earth, Planets and Space, 110 https://doi.org/10.1186/s40623-022-01669-4 .

Ogata Y (2024) Prediction and validation of short- medium- and long-term earthquake probabilities using a hierarchical space-time ETAS (HIST-ETAS) models, etc. Report of the Coordinating Committee for Earthquake Prediction, 107(12-8) 547-555, https://cais.gsi.go.jp/YOCHIREN/report/kaihou111/10-06.pdf .

Ogata Y, Katsura K, Tanemura M (2003). Modelling heterogeneous space-time occurrences of earthquakes and its residual analysis. *Applied Statistics* (JRSS-c) 52(4), 499-509.

Ogata, Y, Katsura K, Tsuruoka H, Hirata N (2019). High-resolution 3D earthquake forecasting beneath the greater Tokyo area, *Earth Planets Space* 71(113).

Ogata Y, Utsu T, Katsura K (1996). Statistical discrimination of foreshocks from other earthquake clusters, *Geophys. J. Int.* 127, 17-30.

Ogata Y, Zhuang J (2006). Space-time ETAS models and an improved extension. *Tectonophysics* 413, 13-23.

Omi T, Ogata Y, Hirata Y, Aihara K (2014) Estimating the ETAS model from an early aftershock sequence, Geophy. Res. Lett., 41, 850-857, doi:10.1002/2013GL058958.

Petrillo G., Zhuang J. (2023) Verifying the magnitude dependence in earthquake occurrence. Physical Review Letters. 131:154101.

Petrillo G, Zhuang J (2024) Bayesian earthquake forecasting approach based on the epidemic type aftershock sequence model. Earth Planets Space 76, 78 (2024). https://doi.org/10.1186/s40623-024-02021-8.

Reasenberg PA, Jones LM (1989). Earthquake hazard after a mainshock in California. *Science* 243(4895):1173-1176.

Schorlemmer D, Gerstenberger MC (2007). RELM testing center. *Seismological Research Letters* 78 (1): 30–36.

Schorlemmer D, Gerstenberger MC, Wiemer S, Jackson DD, Rhoades DA (2007). Earthquake likelihood model testing. *Seismological Research Letters* 78 (1): 17–29.

Schorlemmer D, Zechar JD, Werner MJ, Field EH, Jordan TH, and RELM Working Group (2010). First results of the Regional Earthquake Likelihood Models experiment. In: Savage, M.K., Rhoades, D.A., Smith, E.G.C., Gerstenberger, M.C., Vere-Jones, D. (eds)

Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II. Pageoph Topical Volumes. Springer, Basel.

Shcherbakov R, Zhuang J, Zoeller G, Ogata Y (2019) Forecasting the magnitude of the largest expected earthquake, Nature Communications, Vol.10, ArticleNo.4051, pp.1-11, doi: https://doi.org/10.1038/s41467-019-11958-4.

Sornette A, Sornette D (1999). Renormalization of earthquake aftershocks. *Geophys. Res. Lett.* 26(N13), 1981-1984.

Stark PB (1997). Earthquake prediction: the null hypothesis. *Geophysical Journal International*, 131(3), 495-499. Van Stiphout T, Schorlemmer D, Wiemer S (2011) The Effect of Uncertainties on Estimates of Background Seismicity Rate, Bull. Seismol. Soc. Am., 101(2), 482-494, 10.1785/0120090143.Zechar JD, Gerstenberger MC, Rhoades DA (2010). Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts. *Bulletin of the Seismological Society of America* 100 (3): 1184–1195.

Zechar JD, Schorlemmer D, Werner MJ, Gerstenberger MC, Rhoades DA, Jordan TH (2013). Regional Earthquake Likelihood Models I: First-Order Results. *Bulletin of the Seismological Society of America* 103(2A): 787-798.

Zhuang J (2012). Long-term earthquake forecasts based on the epidemic-type aftershock sequence (ETAS) model for short-term clustering. *Research in Geophysics* 2(1):e8-e8.

Zhuang J, Vere-Jones D, Guan H, Ogata Y, Ma L (2005). Preliminary analysis of observations on the ultra-low frequency electric field in a region around Beijing. Pure and Applied Geophysics, 162: 1367-1396, doi:10.1007/s00024-004-2674-3.

Frederic Schoenberg, corresponding author, Department of Statistics and Data Science, 8125 Math-Science Building, UCLA Los Angeles, CA 90095-1554, USA. frederic@stat.ucla.edu.

Danijel Schorlemmer, Helmholtz Centre Potsdam, GFZ German Research Center for Geosciences, Telegrafenberg, D-14473 Potsdam, Germany.