Towards a Harmonized Operational Earthquake Forecasting Model for Europe: Author's Response

Marta Han[1], Leila Mizrahi[1], and Stefan Wiemer[1]

[1]Swiss Seismological Service (SED), ETH Zurich

Report #1

The revised version is much improved. Below I have some remaining suggestings to improve presentation and clarity. And I do have some queries and comments about the water-level results, specifically the sensitivity of the information gain trends with respect to what seem minor changes in the water-level, which are very surprising and need some technical checks and some interpretation (not a solution). The sensitivity may be reduced by using a more appropriate Poisson baseline (see relevant comment below) or using the water-level only when quakes do appear in zero- forecasts bins (and not in all zero-forecast bins). After that, I can recommend publication.

We thank the reviewer for another round of helpful insights, pointing out the dependency on this parameter in the first place, and further suggestions to help address the water-level sensitivity in our results. We hope that our newly revised manuscript addresses the issue more appropriately.

Figure 1: b-value estimates should include uncertainties estimates.
Fig1: The difference between b and b+ seems large, and here b+ seems to give larger values than usual. Which delta m (ie equivalent mc) did you use for b+? Is it possible that b+ is biased for this very particular dataset? It just seems larger than expected.

We add the uncertainties to the revised manuscript and specifically to the caption of Fig. 1. Both b-value estimates take into account the magnitudes' binning of $\Delta m = 0.2$ and apply the correction introduced in Tinti and Mulargia (1987) to avoid biases. Completeness is spatially and temporally varying, and differences $m - m_c(x, y, t)$ considered in place of pure magnitudes to bring the mixed distribution closer to GR, and completeness of $m - m_c(x, y, t)$ (for b-positive) is 0.2. Since the region considered in our study is very large, the key underlying assumption behind applying the b-positive method, namely that after an event of a certain magnitude, any event larger by a certain increment will be detected, might not apply. However, the results obtained in our experiments suggest that the b-positive estimator yields a more appropriate magnitude distribution than the classical estimator. This suggests that if one estimator is biased, it might be the classical one. However, overall, one of our main conclusions remains that using a single b-value for a large and diverse region such as Europe may be an oversimplification of the problem. The estimation of b-values should be improved by reassessing the completeness estimates and using a spatially (and maybe even temporally) varying b-values and we plan to work on this in the future.

L203: "no substantial evidence" I think requires some references, even if it's your interpretation of some results that do include such updating.

We agree and change the formulation of the statement, now stating that there is no agreed-upon technique to update models to specific sequences. We also add relevant literature, specifically describing the different model updating strategies of Italy, New Zealand, and the United states.

L210: clearly state here that $ETAS_0$ contains one spatially uniform background rate.

We add the clarification to the model specification. However, we would also like to emphasise that while the background rate is uniform during the inversion of parameters, when producing forecasts, background events are simulated at locations where events were observed in the training catalog, weighted by their background probability estimated during the inversion procedure. Hence, in the forecast that is the final output of every model, background events are not located uniformly in space (see Fig. 5).

L276: "due to their under-representation in training data" – There are multiple papers suggesting it's the anisotropy of the aftershocks compared to the isotropic model (Hainzl et al., 2008; Helmstetter et al., 2005; Zhang et al., 2020), as well as the covariance between K and alpha in the likelihood function (Sornette and Werner, 2005 and probably others)

We thank the reviewer for this feasible explanation and add it along with listed literature to the revised manuscript.

L308: pls cite these papers in support of the pyCSEP toolkit efforts:

- *Savran et al., (2022). pyCSEP: A Python Toolkit For Earthquake Forecast Developers. Journal of Open Source Software, 7(69), 3658, https://doi.org/10.21105/joss.03658*

- *Savran, W. H., Bayona, J. A., Iturrieta, P., Asim, K. M., Bao, H., Bayliss, K., ... & Werner, M. J. (2022). pyCSEP: a Python toolkit for earthquake forecast developers. Seismological Society of America, 93(5), 2858-2870.*

Or alternatively/additionally: L496: pls consider adding the JOSS citation here to the SRL citation. The former is the peer-reviewed code base and associated online documentation, while the latter describes the software and motivation.

We agree that these citations are highly relevant to the topic and acknowledge the tremendous efforts done by CSEP community. Thus, we add these citations in the relevant parts of the manuscript.

L383: Actually, the log score is well defined when the forecast is zero and the observed count is zero: the likelihood is exactly 1 and the log likelihood is zero. Secondly, the log score is still well defined, it is negative infinity, when the forecast is zero and there are indeed events. (see comment at the end of the review)

We amend the phrasing in the revised manuscript to avoid confusion.

L393: Could you clarify how the two year period helped set the value? Also, which benchmark model are you referring to (Poisson – as mentioned below)?

Using the 2-year validation set, we inspected plots similar to Fig. S4 to identify water levels for which models perform worse and then verified that it is because either the water level is too low, penalizing its usage heavily, or too high, causing ETAS models to score lower in bins where they forecasted events, but still scoring lower also in bins where water level is used (newly added Fig. S5 in response to the last comment shows this for high water levels). The benchmark model is indeed the time-independent Poisson one (with ESHM seismicity rates per spatial bin). We clarify the statement in the revised manuscript.

L498: true → observed

Corrected here and in a number of other places in the manuscript.

L504: for clarity state why you exclude ETAS$_\alpha$ and ETAS$_{bg,\alpha}$ here (as you do in the caption of Fig 4).

We address this point in the revised manuscript.

Figure 5: Are the white cells in these figures those where no events were simulated? What are the units on the colour scale (expected events per year)? Do these figures suggest a water level based on the background?

The events are counted in total during the entire 25-year training period. In the white cells, indeed no events have been simulated. We extend the caption of said figure to include this information. No water level is used to create these figures. They could, however, be used to create a background-based water level.

Figure 6: I'm quite surprised by these results. Why isn't there a stronger step change at the time of larger quakes, e.g. the M7 in late 2020? It's so evident in late 2018 for a smaller mainshock, but the other large quakes don't seem to generate much information gain for the ETAS models. It's curious – perhaps few aftershocks above the completeness? Please provide a short interpretation.

We add a short analysis in the revised manuscript. Smaller number of aftershocks above m_c seems to be a plausible explanation, as well as the fact that forecasts are issued at midnight. If an event occurs relatively shortly after midnight, the most productive period is not part of the next testing day.

Is the Poisson model uniform or spatially variable? Is the Poisson rate the average rate over the pseudo-prospective period (which would give it more information than the ETAS models got) or over the retrospective training period (which seems fairer, and one I'd recommend)? Please clarify in the text.

In addition, I wonder whether you want to exclude the models you've discounted based on the retrospective tests in Figure 6? You show that some of these excluded models have the highest information gains, but then discount them based on retrospective tests. Is it still useful to show them? In any case - make sure these two sections are very consistent with each other.

The Poisson model is spatially variable, with the rate in each cell being the long-term seismicity rate (normalised from annual to daily) from ESHM20, which includes no information from the pseudo-prospective experiment period, but includes seismicity up to 2015, and additional information such as physical tectonic properties and historic seismicity information. We clarify this better in the revised manuscript in Section 3.2.

Regarding the discounted models, we amend the discussions to be more consistent, but would prefer to keep all models in Figure 6. Better performance of models that fail retrospective tests is still valuable information, allowing comparison between (a) and (b) parts of Figure 6, and our conclusion is that we prefer models used for OEF pass the consistency tests, not that models that fail them should not be disregarded altogether (and it should be investigated further why some are better with shorter time windows or smaller spatial cells).

L628: The water level only needs to be invoked when you have observed quakes but the forecast is zero. If there are zero quakes, the forecast is technically correct and should give probability 1, i.e. log score zero, ie there is no log score penality (a perfect prediction). So the water level should only be applied where you do see events but the forecast is zero. Is this how you've implemented it?

Figure S4 is quite surprising! Slight changes in water level generate substantial changes in overall trends, and even rankings are affected. And the trends (ie overall positive or negative against Poisson) seem to change randomly even for small changes in water level, which is surprising if the baseline stays the same. Did you maintain the simulated forecasts between these plots and only changed the water-level or could there be an effect due to different simulated forecasts here too? To isolate the water-level effect, I think you should keep the simulated forecasts the same, to make it's not differences in stochastic simulations that generate these differences.

More importantly, how do you explain that models perform worse than Poisson in panel top-left when the water level is relatively high (but still less than Poisson?), but then better when the water-level is halved (second panel on the left), then worse again when divided by another 100? It'd be good to label the panels for this discussion.

My recommendation is to check the technical details above (fix simulations between different water-levels; only use it when there are quakes; explain the reversal of trends if it persists). It's interesting to point out this sensitivity, because it helps the community develop better methods (hopefully). You don't need to solve it here.

The simulated dataset is kept fixed already, only the procedure producing the forecasts based on those simulations is adapted to each water level. In the revised manuscript, we specify this information, and add the plot analog to S4 in supplementary materials for when water level is invoked only when needed. The comparison allows for further conclusions in this section answering some of the important questions raised here. In the main plot, we still opt for the version where water level is always distributed over all bins. Namely, our goal is to produce forecasts near real-time and test them truly prospectively; in that case, we cannot know if water level will be needed and distribute it accordingly and therefore, we believe that the current experiment setting better reflects the conditions of truly prospective testing.

Report #2

The authors did a good job in addressing satisfactorily my main comments. I am not still sure to agree with some statements reported in this paper, and with some modeling choices. But, overall, I do think that, in this form, the manuscript can stimulate further research and additional thoughts on this important problem.

We thank the reviewer for the positive feedback and constructive suggestions in the first round of reviews. The efforts described in the paper are an ongoing project and we hope to address some of the worries in our future work.

References

Hainzl, S., Christophersen, A., & Enescu, B. (2008). Impact of Earthquake Rupture Extensions on Parameter Estimations of Point-Process Models. *Bulletin of the Seismological Society of America*, *98*(4), 2066–2072. https://doi.org/10.1785/0120070256

Helmstetter, A., Kagan, Y. Y., & Jackson, D. D. (2005). Importance of small earthquakes for stress transfers and earthquake triggering. *Journal of Geophysical Research: Solid Earth*, *110*(B5). https://doi.org/10.1029/2004JB003286

Sornette, D., & Werner, M. J. (2005). Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences. *Journal of Geophysical Research: Solid Earth*, *110*(B8), 2004JB003535. https://doi.org/10.1029/2004JB003535

Tinti, S., & Mulargia, F. (1987). Confidence intervals of b values for grouped magnitudes. *Bulletin of the Seismological Society of America*, *77*(6), 2125–2134. https://doi.org/10.1785/BSSA0770062125

Zhang, L., Werner, M. J., & Goda, K. (2020). Variability of ETAS Parameters in Global Subduction Zones and Applications to Mainshock–Aftershock Hazard Assessment. *Bulletin of the Seismological Society of America*, *110*(1), 191–212. https://doi.org/10.1785/0120190121