

Weaknesses and *some possible* solutions in current CSEP experiments

Warner Marzocchi,
Istituto Nazionale di Geofisica e Vulcanologia, Italy

□ The Poisson hypothesis does not hold in some (many?) cases, in particular for short-term forecasting models like ETAS.

Forecasts should not only consist of one number, but of a distribution

□ The 'likelihood' used is not actually a likelihood, because there is correlation in space. I have no solution for that, but it would be wise to consider alternative metrics

□ It is not clear the role of all classical tests. In theory, a test should be carried out once. I would prefer that we assign to each model a performance measure like the Bayes factor. We may run classical test for selected hypothesis that are of specific interest. For instance, for 1-day forecasts model, I may be more interested in running a consistency test just once for the whole period instead of having multiple consistency tests run every day.

- ❑ The role of consistency tests should be clarified. I may anticipate that all models are wrong, so it is only matter of how many data I have to collect in order to find a significant discrepancy between forecasts and data. Moreover, it happens that stupid models may not fail, but other much more clever models do. To me, consistency tests are useful to understand where my model is getting wrong.
- ❑ Comparison test are probably more interesting, but classical tests may be misleading (see my previous comments). We need to develop different metrics for comparison, possibly avoiding classical tests
- ❑ Should CSEP incorporate some ensemble model? I would say yes, but it is not clear yet in my mind how to proceed.

- ❑ In general, we have to clarify what we can learn from each specific test/statistics
- ❑ Easy: create a repository containing all papers that show the results of CSEP experiments