Nat. Hazards Earth Syst. Sci. Discuss., 3, C476–C479, 2015
www.nat-hazards-earth-syst-sci-discuss.net/3/C476/2015/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Brief Communication: Statistical detection and modeling of the over-dispersion of winter storm occurrence" *by* M. Raschke

F. Serinaldi (Referee)

francesco.serinaldi@ncl.ac.uk

Received and published: 22 April 2015

General comments

Dear Dr. Raschke, thank you for your reply. I think this is an interesting discussion.

Specific comments

Section 2: I think you confirmed what I mean. GPD describes a particular case of inhomogeneity corresponding to the random fluctuation of the rate of occurrence λ,

which is modeled by integrating over the d.f. of such a rate (Daley and Vere-Jones 2003, Ch. 2), thus resulting in distributions with constant rate of occurrence. If we want to explicitly model λ fluctuations, we need an explicit model (deterministic or stochastic) for $\lambda(t)$. In this respect, GPD describes a very specific type of inhomogeneity handled in a very specific way.

P1778L19: Thanks for fixing this point. I understand the rationale of your discussion. Unfortunately NHESS is not a statistics journal. In spite of bijective function between return period and the corresponding storm intensities (levels), they are not the same. For design purposes, knowing the degree of rarity of a variable is not enough if we do not specify which variable, and its absolute value and measurement units. This justifies for instance the use of return level plots (return period versus return level) widely used by e.g. Coles (2001), just to mention a statistician who always put attention to the physical meaning of the environmental variables he analyzed... before moving to industry. In a nutshell, in my opinion, real-world data can (sometimes) be treated as random variables but they are not random variables. So, "return periods" describe the degree of rarity, while "return levels" are commonly used to refer to the value of the variable of interest. Of course, we can use different terms, but keeping this distinction can help the NHESS reader, I think.

Section 4: Your reply about performance indices is quite interesting, but I think that we do not disagree very much. In particular, my opinion seems to me in line with "the state-of-the-art of statistics and stochastics". Referring to Fahrmeir et a. (20013) (but I could mention whatever book on GLM/GAM, etc.) log-likelihood ratio (LR), Wald statistic, and score statistics are equivalent and generally distributed (asymptotically) as a χ^2 random variable with suitable degrees of freedom. These quantities describe distances between two alternative models and allow for formal testing based on their distribution under the null. On the other hand, AIC, BIC etc, are not tests (of course) and are not sufficient to state that a model is significantly better than another without applying additional criteria. If the difference between the AIC values of two models

is very small there could be large uncertainty about the best one. In this respect, evidence ratios can help to understand such uncertainty. In your reply, you implicitly applied a testing procedure, but the reported significance levels are just limiting values, as the actual significance of the difference Δ between two AIC values depends on the magnitude of Δ itself. In other words, you showed that ranking the scores is not enough (as is obvious) to make inference (indeed you had to introduce significance levels, which is what I asked in my report). To assess the significance of the difference of AIC or BIC values between models A and B we need a formal test. If we perform this implicitly or explicitly (Δ, LR or something else) does not matter very much, but this information should be provided.

Finally a remark about "The practice of different scientific communities is in contrast to the state-of-the-art of mathematical statistics:". As mentioned above, things are a bit more complex. Statistics provides tools for data analysis that should be chosen according to the aim of the analysis and the nature of the data at hand (financial data, hydrological data, etc, reflect different dynamics and generally require different tools, or ad hoc versions of these tools). Talking about "the state-of-the-art of mathematical statistics" when referring to ML techniques or even more specific performance criteria is a bit reductive. Performance criteria should be chosen in light of the aim of the analysis because "information" is not always the only or most important criterion. In this respect, since you spent some time looking at one of my papers, that is one of the cases in which the focus was not on the information criteria. For design purposes, but also more generally in applied statistics, we usually focus on the quantities of interest for the specific problem and not on the "ML state-of-the-art" per se if the "ML state-of-the-art" does not fit our needs. Unfortunately, people tend to be a purist in their own field of specialization, resulting in no-statisticians badly using statistics, and statistician overlooking practical aspects. For example, in other papers of mine I used the "ML state-of-the-art", AIC, etc. Fortunately "mathematical statistics" goes beyond ML and provides lots of nice tools suitable for different problems. In any case, opinion exchanges, such as this case, are surely useful to break the wall between different

point of views in different fields, share information, and build constructive syntheses.

Sincerely

F

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., 3, 1775, 2015.