Impact of spatio-temporal dependence on the frequency of precipitation extremes: Negligible or neglected?

By F. Serinaldi

Submitted to *HESSD*

MS-NR: hess-2023-293

PRELIMINARY REPLY TO RC3

As for the other Reviewers, I would like to thank the Reviewer for the time dedicated to the manuscript. Below, I provide a quick preliminary response. If the manuscript passes the first stage of review and revision is allowed, I will provide more detailed responses and an updated version of the manuscript.

If I correctly interpreted Reviewer's remarks, they raised two main points:

1) Proper acknowledgement of the literature, and
2) Making the paper more stand-alone.

Concerning the first point, some of the papers mentioned by the Reviewer are already cited, while others surely fit and will be added to the bibliography.

I have to say that I did not emphasize any specific method and the related literature just because the actual purpose of the paper is slightly different from what seems to emerge from the interpretation of the three Reviewers that commented the paper so far.

Obviously, if different people made similar interpretation, it means that the text is not clear and need to be revised accordingly.

Technical details concerning all models used in the paper are reported in the 'Supplementary' because the aim of the paper is not the comparison of specific models (or a specific types of analysis), but a discussion about the consequences of neglecting the rationale of statistical inference, using precipitation analysis as an example.

In this respect, let me refer to part of my responses to RC1 and RC2, as they apply to RC3 as well.

The aim of the paper is not to show that a model/method is better than another one, but to show that most of the trend analysis reported in the literature results from systematically neglecting (or

just lack of knowledge of) the epistemological rationale of statistical inference, which is is as follows (e.g., Aitken, 1947; Cramér, 1946; Box, 1976; Papoulis, 1991, von Storch and Zwiers, 2003… among others):

1) Make assumptions.
2) Build models and make inference accounting for the effect and consequences of those assumptions.
3) Interpret results according to the nature of the adopted models and their assumptions.

Most of the literature on trend analysis of (unrepeatable) hydroclimatic processes seems to neglect such a rationale and switches stage 1 with stage 2, resulting in the following fallacious procedure:

1) Select several models and methods based on different and often incompatible assumptions.
2) Make inference neglecting the constraints imposed by the different assumptions.
3) Interpret the results attempting to prove/disprove assumptions.

This approach, which corresponds to a widespread mechanistic use of statistical methods/software, suffers from logical fallacies. It neglects that models cannot be used to prove/disprove their own assumptions in the same way a mathematical theory cannot prove/disprove its own axioms and definitions. Indeed, those models and theories are valid only under those assumptions, axioms, and definitions. Of course, models cannot even be used to prove/disprove alternative assumptions as they might not even exists under those alternative hypotheses.

The paper compares the proper approach to statistical inference (which is the only one conceivable by every analyst until the past century) with its fallacious ('modern') counterpart and shows practical consequences, using a typical trend analysis of precipitation data as an example. I used the analysis reported by Farris et al. because it deals with a large data set and their analysis is rather detailed, thus looking rigorous and solid, but neglecting the effects of the original epistemological fallacies. However, I could have used many other examples: almost every paper reporting "trend analysis" suffers from the same problems.

For the sake of clarity, let me summarize how the twisting of inference logic impacts on the first step of the analysis reported in the paper, i.e. the selection of the marginal distribution, which is a typical exercise familiar to any hydrologist:

In a proper application of statistical inference (as it should be), we can consider for instance the following cases:

Case 1

- Assumption: precipitation occurrences are *assumed* to be consistent with a *stationary and independent* process.
- Under these conditions:
    - Poisson distribution is a suitable candidate for the marginal distribution of count process (of over-threshold occurrences).

    - If one wants, standard Goodness-of-Fit (GoF) tests (such as Kolmogorov-Smirnov, Cramer-von Mises, etc.) can be applied (… even if their application in this context is still problematic).
- Graphical and numerical results can say if Poisson is a defensible model *under stationarity and independence*. If Poisson does not fit satisfactorily, this does not prove/disprove its assumptions; in fact, there can be another distribution that works well under the same assumptions.

Case 2

- Assumption: precipitation occurrences are *assumed* to be consistent with a *stationary and dependent* process;
- Under these conditions:
    - we expect overdispersion because of dependence. The Poisson distribution is known in advance not to be e suitable model from theoretical perspective. Therefore, we should consider some distribution allowing for overdispersion (e.g. negative Binomial, beta Binomial, etc.).
    - In this case, GoF tests should account for variance-inflation of the test statistic due to dependence.
- Also in this case, empirical results do not prove/disprove any assumption; they just say if the adopted models provide a satisfactory description of data within the desired tolerance.

Case 3

- Assumption: precipitation occurrences are *assumed* to be consistent with a *nonstationary and independent* process;
- Under these conditions:
    - we expect overdispersion because of non-stationarity. Indeed, in this case, every observation is *assumed* to come from different distributions. For example, if we *assume* that the rate of occurrence linearly increases in time, data might be *assumed* to come from a set of Poisson distributions with different rate parameter. Therefore, the resulting overall distribution is mixed/compound Poisson, which is over-dispersed. Such a distribution is not even unique, as it depends on the time window where it is computed.
    - In this case, GoF tests cannot be applied to raw data because a unique (population) mixed Poisson does not exist, and such test can be applied at most to filtered (detrended) data (e.g., Coles, 2001) to check the behaviour of the conditional distribution, which is considered unique under the *assumption* that the filtered data are *conditionally stationary* (and a unique conditional distribution does exist).
- Also in this case, empirical results do not prove/disprove any assumption, as the results are valid only under the assumptions used to make inference.

To summarize, following the rationale of statistical inference, both model selection and inference depend on the assumptions we make. Of course, we can use different assumptions (cases 1-3

above), develop the *complete* inference for each one (accounting for the corresponding constraints), and eventually choose the framework based on parsimony, accuracy, and generality of results (or predictability). What we cannot do is mix up models and tools that are valid under some assumptions in a different context and for different assumptions.

Such a misuse (corresponding to the fallacious approach mentioned above) is precisely what is routinely used in (too) many papers and generates logical contradictions and paradoxes. And this is the focus of my paper.

For example, Farris et al. use the Chi-square and Kolmogorov-Smirnov (Lilliefors) GoF tests concluding that the Poisson distribution cannot be rejected for more than 95% of cases at the 5% global significance in all cases (thresholds and samples sizes). Therefore "*the Poisson distribution is adopted as the parent distribution of count time series*". However, their subsequent analysis is based on two models (INAR and NHP) that correspond to two assumptions (dependence and non-stationarity) that are incompatible with both Poisson distribution and the application of GoF tests as done in their Section 3.2.

Indeed, under such assumptions (dependence and non-stationarity), the distribution is expected to be over-dispersed. More precisely, under dependence, we have variance inflation, while under non-stationarity the distribution is not unique and it can be over-dispersed mixed-Poisson, at most. In other words, if the parent distribution of raw count data is assumed to be Poisson (based on GoF tests), NHP cannot be an option and vice versa: the same data cannot be simultaneously Poisson and mixed-Poisson, equi-dispersed and over-dispersed. This would violate the principle of non-contradiction.

Such a contradiction raises from neglecting the fact that (i) candidate models are different under different assumptions, (ii) such assumptions also impact on the form and interpretation of GoF tests, and (iii) outcomes of GoF tests under a specific assumption (e.g., dependence) are not valid under alternative assumptions (such as independence or non-stationarity).

The focus of the paper is on these issues extended to all the steps of the analysis of precipitation data.

In my opinion, these problems are very compelling, as they denote that statistical analysis is routinely performed (in some applied sciences) neglecting or ignoring the rationale of statistical inference and therefore the practical negative consequences of such an approach to data analysis, i.e. a systematic and unavoidable misinterpretation of any result, independently of the models/frameworks used.

These remarks lead to the second concern raised by the Reviewer (and RC2): a stand-alone paper.

Independent validation of results is a key aspect of science. Comparisons and replications of the same methods, analysis and experiments are the core of scientific enquiry, and are routinely applied in medicine, physics, etc.

However, in my experience, this is not the case of hydrological data analysis where we have a sort of binary approach: either a paper proposes something (supposedly) new, or we write a comment, which, however, must always be short because of journals' policies.
In my opinion, scientific debate cannot be relegated to short comments or verbal discussion 'in front of a beer' at some conference, and for sure most of the published papers (included mine) do not propose anything significantly new (despite the clichés emphasizing 'novelty' in abstracts and conclusions).
Therefore, the paper under review is part of a series of mine (with co-authors) falling in the class of so-called "neutral" validation studies, aiming at re-analysing methods, procedures, conjectures, etc., in detail.

This type of work attempts to double check methods and general concepts, using specific examples taken from previous works to provide a side-by-side comparison, to highlight the striking practical differences or possible inconsistencies resulting from different ways of reasoning (or lack of reasoning).

In my opinion, keeping the discussion general, as done in the past for instance by Yevjevich (1968) and Klemes (1986) without referring to specific examples, has historically been proven to be not very effective.
Scientific double-check needs to be as precise as possible, reproducing the results previously reported in the literature (to be sure why they are what they are) and then contrasting them (if required) with alternative methods.
To do that, we need to introduce the original procedures. Let me use an example to explain what I mean:

- Years ago, someone stated that they were able to make cold nuclear fusion (or something like that).
- These results were checked in detail.
- I was found that the claimed procedure did not work.

To do that double-check, it could not be sufficient to say, "We made some experiments, and we were not able to make cold fusion". Discussants had to explicitly refer to the original method and check every detail.
Was not independent check worth communication? Was this negative for science and technological advances?
Did the authors of the original findings feel happy? I do not think so. Probably they were not happy in the same way the supporters of Ptolemaic system were not about Copernican theory.
Should have people avoided to compare Ptolemaic with Copernican theory just not to hurt (the ego of) supporters of the former?

As mentioned above, I used the analysis reported by Farris et al. as quite a sophisticated example of how far can lead ignoring the rationale of statistical inference. Indeed, the message of that work is not the usual conclusion "precipitation is nonstationary" (which is already nonsense), but a bolder statement ("*Accounting for serial correlation in observed extreme precipitation frequency has limited impact on statistical trend analyses*"), which relies on the unscientific concepts that we can use models (generally the most trivial versions) to check their own assumptions or assumptions of other models. The criticisms reported in the paper are valid for most of the literature on the topic, and the concluding discussion is indeed fully general in this respect.

We need to refer to the original approaches if we want to highlight the paradoxes resulting from methodological misconceptions. Otherwise, the discussion would always be vague, and different approaches would look like different (but equally sound) points of view, which is not the case if one of them does not follow the logic of science.
Of course, one can call principles of science into question and introduce new ones. However, in any case, we must agree about which ones we want to use. Otherwise, it would be like attempting to communicate using different languages, or worse, using the same language with words referring to different meanings, that is, missing the link between signifiers and referents.

The paper focuses on these problems and was structured accordingly, and it follows the same structure of previous papers of mine of the same kind… I have to say that, in my experience, years ago this type of papers and discussions were more welcome, while lately they seem to be less 'tolerated'. By the way, "*the times they are a-changin'*" and perhaps a bit 'harsh' but detailed scientific debate is no longer suitable for the "*brave new world*" we are building. Not sure this is positive, but the time will say.

That said, I'll try to better clarify these issues in the revised manuscript if it passes the first stage of reviews.