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 RC2 (DR. D. KOUTSOYIANNIS' REPORT)

(Note: In the text below, Referees' comments were copied verbatim in **black**. Replies are in **blue**)

I would like to thank the Reviewer for his remarks on the manuscript. Below, I provide a preliminary response to the main concerns. 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.

6. A question not discussed in the paper is if trend identification has some usefulness or not. Let us assume that we have correctly (whatever this might mean) identified statistically significant trends based on past data. If our focus is on the past, what is the usefulness of a trend, once we have a better description by the data themselves? And if the focus is on the future, what is the value of such trend identification? Will a trend identified in the past continue in the future, or will it bend? Does the trend have any predictive value? By the way, this issue has been investigated by Iliopoulou and Koutsoyiannis (2020; see also the motto in that paper), but I would be interested to know the author's opinion on this. The current trend in detecting pointless trends continues to dominate in literature, and it would be useful and beneficial to the community if the paper discussed this issue.

Response

I understand Reviewer's concerns, but I think that the usefulness of detecting trends and prediction are secondary problems, as I believe that "*the trend in detecting pointless trends*" is related to more basic but fundamental problems that the paper tries to highlight.

Let me elaborate a bit more on this point before discussing my point of view on trends.

I'm used to distinguish statistical analysis and decision making, which can rely on statistical tools but is different. Now, if the information feeding decision making is based on a flawed preliminary statistical analysis, any discussion about usefulness, predictability, etc., is just pointless.

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 "*the trend in detecting pointless trends*" 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 with its fallacious 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). The Reviewer described this situation in various papers of him, and his book.
- 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 more compelling than establishing the usefulness of some analysis' output, as they denote that statistical analysis is routinely performed (in some applied sciences) neglecting or ignoring the rationale of statistical inference and 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.

Going back to trends, Reviewer asked**: *'Can the author define what a trend is?*"**

Let me try. In the context of the analyses reported in the paper and most of the literature on the topic, a trend should be intended as a well-defined (deterministic, if one wants) function or rule

linking a variable of interest (or one or more of its properties) to another variable describing a parametric support. For example, the Poisson regression model (non-homogeneous Poisson) used in the paper (and in Farris et al.) assumes that the rate of occurrence is a log-linear function of time

$$\log(\lambda(t)) = \lambda_0 + \phi \cdot t.$$

On the other hand, when we apply Mann-Kendall test, the tested trend is defined as follows (Mann 1945):

"*The hypothesis of a downward trend may be defined in the following way: The sample is still a random sample but X_i has the continuous cumulative distribution function f_i and $f_i(X) < f_{i+k}(X)$ for every i, every X, and every k >0. An upward trend is similarly defined with $f_i(X) > f_{i+k}(X)$.*"

In this case, the property of concern is the so-called *stochastic dominance*. Thus, trends correspond to a well-defined rule implying that the distribution of *X* at time *i stochastically dominates* the distribution of *X* at any subsequent time *i+k*, or vice versa. In other words, it assumes that stochastic dominance is on play, and it depends on time.
Such a definition does not specify a precise formula *y* = g(*t*), but a precise rule, that is, systematic stochastic dominance. Since stochastic dominance often results in monotonic patterns of central tendency measures, Mann-Kendall test is often (but incorrectly) referred to as a test for "trends in the median/mean values".

The case of MK tests allows some considerations:

- In contrast with many "tests" proposed nowadays, tests developed by professional statisticians in the past (century) were compliant with the rationale of statistical inference and were always explicit and rather precise about assumptions and preliminary definitions.
- Based on the definition of trend given by Mann and the fact that the test statistic relies on ranks rather than absolute values, it is expected that such a test is less powerful than other parametric tests based on more restrictive assumptions of (conditional) Gaussianity and precise (non)linear function of the mean (or median or whatever else). On the other hand, parametric tests have higher Type I error under misspecification (which is always the case for non-repeatable processes), because there is no free lunch either in life or statistics.
- Based on the foregoing remarks, it follows how many studies reporting power analysis of MK tests just show what is expected from the theory.

Therefore, different statistical methods and tests rely on their own definition of "trend". When tests are well devised, such definitions are clearly stated as they are necessary to set up a suitable and coherent test statistic and derive its properties.
However, (too) many end-users neglect the different definitions of trend implied by different tests/methods, which become just algorithms to be run over large data sets.

The consequence is not only merging methods that are devised to answer different questions (without recognizing such a difference) but also logical paradoxes.

For example, focusing on the analysis in the paper, some of the fitted non-homogeneous Poisson distributions (fitted under the assumption of non-stationarity) show negative slope, meaning decreasing rate of occurrence. Some of these models yield rate equal to zero in few decades. This has very practical and inconvenient consequences:

- The decreasing rate implies decreasing variance, which depends on time as well. This means that a single population variance does not exist, and autocovariance and autocorrelation are ill-defined as well. In other words, sample ACF has the same lack of meaning as the sample mean under the assumption that data come from a Cauchy distribution. Therefore, all ACFs or lag-1 ACF terms reported in the paper and in Farris et al. (… and in any other paper) under the assumption of non-stationarity do not correspond to any population counterpart. They are just numbers that do not allow any inference because the supposed inferred *population* counterpart does not even exist.
- Such models hinder the calculation of whatever index or summary statistics requiring integration over time, such as return periods and levels. Indeed, if the process no longer exists after e.g. 80 years (because the rate λ become equal to zero after 80 years), calculating return levels over longer return period is just meaningless.

Moving from formal definitions to applications, the keyword in trend analysis of hydroclimatic data is "repeatability". A key aspect of science is experimentation: a result is scientifically sound if it can be replicated. Hydroclimatic processes cannot be repeated; however, some of them repeat themselves in time over time scales that allow multiple observations of the same phenomenon. In these cases, we implicitly or explicitly replace multiple independent experiments with multiple observations of the same process in time. In this respect seasonality is a trend (the distribution of a hydroclimatic variable is assumed to be a function of calendar days, or months, or similar).
What is the difference between seasonal trends and 'monotonic trends' over e.g. 100 years? Repeatability: the formers have been observed many times (let's say, Earth and Sun are so kind to repeat the experiment for us every 365 days), whereas 'monotonic trends' are not. The latter are unique. Making inference on them is like assessing if a die is loaded by throwing it just once.

Finally, there is a third aspect to consider. Referring to trends, I noted that Iliopoulou and Koutsoyiannis (2020) mentioned some econometric literature. This suggested a few additional considerations to me.
The world "trend" is used with different meanings not only in different disciplines but also in the same discipline. For example, when we deal with econometric models (ARCH, GARCH, or whatever else), trend and non-stationarity have necessarily the meaning resulting from the formal definition of stationarity (otherwise these models would not be compliant to mathematical derivations, asymptotic, etc.).
However, trends have a different meaning in technical (visual) analysis used to support investment and trading strategies. The figure below shows the last three months of daily data (Japanese candlesticks) of S&P500 stock index. At a first glance one can think that there is a "monotonic trend", and one can also attempt some prediction, as extrapolation does not seem so difficult.

However, if we expand to the view back to the last 7 months, we have a different picture. Thus, perhaps extrapolation could be not so easy.

Things are even more interesting if we go 13 and 24 months back (see figures below).
In such diagrams, "trends" are not mathematical functions, but only upward and downward fluctuations existing at various scales… I will not digress on the tons of papers and books dealing with scaling properties of these data after Mandelbrot's seminal works). Extrapolation based on the first diagram (last 3 months) can be very harmful.
Is this stuff predictable? Well, generally, the 95% of traders lose money, and "*Normally, forecasts should be as far from one another as they are from the predicted number*" (Taleb, 2010).

Here the massage is that there is a fundamental difference between stock market and hydroclimatic time series: the former result from a man-made process, while the latter do not.
People applying econometric models or statistical tests for trend detection (such as KPSS, Dickey-Fuller, etc.) often miss that such tests involve assumptions like (almost) infinite variance, which are unrealistic for hydroclimatic processes. This is, once again, the effect of using methods without checking their underlying theory, the context where they were developed, and the problems they attempted to solve.

Therefore, when dealing with ‘trends’, we should consider three aspects at least:

- Formal definition, which is required to develop any sound test or model. If such a definition is missing, the resulting methodology might be formally meaningless and useless. Results are just nonsense numbers. A few previous papers of mine (and co-authors) discuss some of these flawed algorithms.
- Type of data: experimental or not. Repeatability.
- Context: type of process we are dealing with (and we want to model), e.g., hydrological, biological, financial, etc.

Are these points exhaustive? No, of course. However, they indicate that these issues have multiple aspects, which should be considered.

7. Since the paper contains a sound epistemological part, it would be suitable to give a definition of a trend on a scientific (not colloquial) basis. I believe this notion, is more unclear than is popular and lacks a proper definition. I am interested to see what the author thinks about this issue.

Response

Please, see the foregoing discussion. I'll add a definition, but, as mentioned above, these issues have several facets, and would require another paper, at least.
In this respect, let me mention an anecdote that further highlights the actual problems affecting the literature on these topics. Recently, I read about a 'functional' definition of stationarity corresponding to a sort of lack of 'visual' trends in a sample.
Thus, even if one can try to be as rigorous as possible, this effort is rather useless if the potential readers do not even recognize the difference between sample and population properties described in the first chapter of every introductory handbook of applied statistics.
Thus, in my experience, the worrying problem is that there are many people working on data analysis (and publishing) with insufficient training (if any)… myself included, perhaps. Of course, the availability of powerful ready-to-use software contributes to this situation. In my opinion, easier access to computational resources and software should correspond to a more and more solid theoretical and epistemological background… however, the 'trend' seems to go in the opposite direction.

8. On the negative side, the paper is difficult to read—and review. While the epistemological parts of the paper and the related questions are well discussed and clear, the technical parts are difficult to assimilate. Near the end of the paper, the author states that he offered an "alternative point of view". However, this is not clarified and is spread in many different sections without coherence. At least a summary of the approach proposed would be helpful and would improve the paper.

Response
I'll try to clarify the text in the revised version (if any).
As mentioned above, the paper attempts to compare two approaches to statistical analysis, the usual one (which is the only one that is theoretically sound) and the fallacious one.
The 'alternative point of view' is just the standard way to make inference, which however looks like an 'alternative' nowadays, as the fallacious approach seems to be the rule in some literature.
The whole paper is a side-by-side comparison of the consequences of the two approaches when they are applied at each stage of the analysis of precipitation data.
In this respect, the structure and rationale of the paper is rather simple.
Moreover, understanding the logic (or lack of logic) of the various steps of analysis is much more important to me than focusing on technicalities. This is why all models and methods are reported in the supplementary material.
The aim is not to compare specific models but two different approaches to statistical inference, highlighting the logical inconsistencies of the second one, which is widespread but contrasts with the principles of scientific inquiry.
I'm aware that this type of content is not common and far from that of the typical technical paper showing "problem-(new) model-results", but I think that the above-mentioned issues are even more important, as they deal with how people use statistics, independently of the specific problem at hand.

9. The structure of the paper is not optimal. Its style is not didactic, and, at places, it is too technical and unclear. The paper needs a substantial overhaul to make an attractive narrative—otherwise I think it would not be read.

Response
I'll try to clarify the text. Please consider that, in the past, when I tried to be more didactic, I have been criticized because I was 'didactic', and this was deemed not appropriate in communications

among peers. When I try to be more technical (to be more consistent with the rules of communication among peers), I was asked to be more didactic.
I'll try to improve the presentation, but I'm sure that there will always be someone who does not like the paper for one reason and the opposite one at the same time.
By the way, the effort to be clear is often useless. For example, the abstract of a paper of mine states that a certain method "*is affected by sample size, distribution shape, and serial correlation*", but that paper is systematically cited as "*such a method is independent of sample size, distribution shape, and serial correlation (…, Serinaldi et al.,…)*".
This is not an isolated case. In my experience, often papers are systematically mis-cited just because readers seem not to recognize the difference between 'it is' and 'it is *not*'. Therefore, attempting to be clear is a good habit, but often pointless.

10. A major negative issue is that the paper is aligned with another paper by Farris et al. (2021). It even attempts to present the methodology of that previous paper (section 3.1). In this respect, it does not look as an independent paper but a discussion on another paper. Yet it is not a formal discussion, as the paper by Farris et al. (2021) was published in another journal. I think the present paper reflects a sound work that could justify publication, but it needs reworking to be presented as a stand-alone paper.

Response
Independent validation of results is another 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 by Yevjevich (1968) and Klemes (1986), for instance, 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. I'll try to better clarify these issues.

11. It appears that the Hurst-Kolmogorov (HK) behaviour (long-term persistence; long-range dependence) is not present in the main paper but only in the Supplement. It has been shown (Iliopoulou and Koutsoyiannis, 2019) that subsetting of a time series (using thresholds or block maxima) distorts the dependence and may hide that behaviour, but this does not mean its influence disappears. The main statistic used in the paper appears to be the lag one autocorrelation (ρ_1), but this does not capture the HK behaviour, so I doubt if it is appropriate. I think this would be useful to discuss in the paper.

Response
HK is reported in the Supplementary along with all methods and technicalities (INAR, NHP, Beta-binomial, IAAFT, etc.).
This was purposedly done to keep the discussion focused on the conceptual problem, that is, the consequences of switching from scientifically sound statistical inference to a logically fallacious approach that mixes up assumptions, models, and results.

The specific models used in the paper are secondary. We could use other models; what matters is how they are used, i.e., do we follow the rationale of statistical inference (which was obvious to any analysist until to the past century) or do we mix up everything confusing models, assumptions, etc, as routinely done in (too) many papers nowadays?

That said, I agree that ρ1 is not representative of HK (of course); moreover, (population) ρ1 does not even exist for NHP, as sample estimates (via standard estimators) do not correspond to any theoretical counterpart (inference is not possible). However, I had to use ρ1 estimates to make a direct comparison with previous results reported in the literature, thus showing precisely the logical and practical inconsistencies resulting from neglecting these theoretical issues.
I'll try to clarify these concepts in the revised version (if any).

12. Overall, it would be a pity if this work and the important points it makes were not published. On the other hand, its current form needs substantial improvement, before the paper can be publishable. I am sorry that I am not more specific in my comments, but, as I said, I had difficulties to read the paper.

Response
I'll do my best to improve the presentation.
Reviewers' comments made me realize that the actual message of the paper did not emerge clearly enough.