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 RC1 (DR. G. MASCARO'S 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.

As well summarized in the title, the main goal of this paper is to demonstrate the importance of accounting for spatio-temporal dependence on the frequency of precipitation extremes when investigating the possible presence of non-stationarity. The paper is motivated, in part, by a recent study by Farris et al. (2021) who performed analyses of long-term daily precipitation records covering several regions of the world to investigate the importance of serial correlation and field significance in trend analysis. In the manuscript under review here, most of the analyses of Farris et al. (2021) are repeated using alternative models and approaches that the author believes are more correct for the purpose. The main critiques raised by the author against the work of Farris et al. (2021) are (1) the adoption of the INAR(1)-Poisson and Non-Homogeneous Poisson (NHP) models to simulate time series that reproduce realistic stationary autocorrelated series and uncorrelated series with trends, respectively; and (2) the use of statistical tests to detect trends (and in general the 'standard' approaches used in the literature on trend analyses). To pursue his main goal, the author presents figures and analyses aimed at showing (and erroneous proving) that:

- The INAR(1) and NHP used in Farris et al. (2021) are not proper models of count time series of over-threshold (OT) daily precipitation (P) series since they do not capture the marginal distribution of the dataset.
- The hypothesis of no-trend is verified in all cases using the Iterative Amplitude Adjusted Fourier Transform (IAAFT) model after the power spectrum is bias corrected to account for the sample size and the field significance is considered.
- The Beta-Binomial (BB) distribution parameterized through the empirical spatial, temporal, and spatiotemporal linear correlation structure of the binary process of daily OT P occurrences captures the distribution of annual OT counts. Since the binary process used to parameterize the BB is assumed stationary by the author, it is concluded that there is no need to apply any trend test.

Response

I think that such a summary does not describe the actual content of the manuscript very well.

- I criticize the misuse of INAR(1) and NHP to draw conclusions about their own assumptions and/or alternative assumptions, neglecting the effect of such assumptions on the models themselves, their inference, and the interpretation of results.
- The paper does not verify any hypothesis; conversely, it shows the logical fallacy of attempting to use specific models and tests to check their own assumptions. In the present context, statistical tests are used only to show their inconsistency.
- There is no logical link between using a Beta Binomial distribution (and/or the stationary assumption) and the need to perform whatever statistical test. Such a link or conclusion is not stated anywhere in the paper. What I state throughout the paper (and discuss in detail in section 6.2) is that the outcome of any statistical test is redundant as it just confirms what is expected from its own assumptions (and underlying models). Using trend tests in the context of unrepeatable processes is uninformative even if we assume non-stationarity, non-homogenous Poisson (NHP) models, or whatever else.

Let me summarize once again the content of the paper: it attempts to clarify the practical consequences of neglecting the epistemological rationale of statistical inference, which is as follows (e.g., Aitken, 1947; Cramér, 1946; Papoulis, 1991, von Storch and Zwiers, 2003):

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, including the work by Farris et al. and Reviewer's report, 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 models' 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. This is because those models and theories are valid only under those assumptions, axioms, and definitions. Of course, such 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.

Even if the paper focuses on the data and analysis presented by Farris et al., the discussion is valid for many similar works approaching data analysis in the same (questionable) way.

For the sake of further 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.
- 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. 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.

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, 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 and responses below is on these issues extended to all the steps of the analysis of precipitation data.

Value of the BB distribution for the time series of annual counts {*Zi*}

While I understand the reasonings of adopting the BB as a reasonable distribution to represent the correct marginal distribution of stationary time series of counts that exhibit serial correlation, I disagree with the author attempts to show that this is also true from the practical point of view by comparing it with the Poisson distribution in Figures 2 and 3. I have separate concerns related to these figures.

Scatterplot between mean and variance of Z

In Figure 2, the author presents scatterplots between the mean and variance of *Z* for the observed OT samples along with those of synthetic samples obtained from Poisson, NHP, and BB models. I tried to reproduce Figure 2 for OT above the 95% empirical quantiles and reported results in Figure R1. I first point out that the mean of the observed samples should not vary, because it is prescribed by the quantile adopted for *x**. For the 95% empirical quantiles, it should be *x** = (1 – 0.95)*365.25 = 18.2625. This is not the case in Figure 2 of the paper, where the mean exhibits a negatively skewed spread around the expected value of 18.2625. Based on my interpretation, which I used to generate Figure R1, this is an artifact caused by the fact that the author did not account for the presence of repeated values in the {*Xj*} series when applying the condition (*Xj* > *x**) to ultimately compute {*Zi*}. Note that a positively skewed spread for the mean (essentially, a mirrored version of the spread in Figure 2) is instead obtained by applying the condition (*Xj* ≥ *x**).

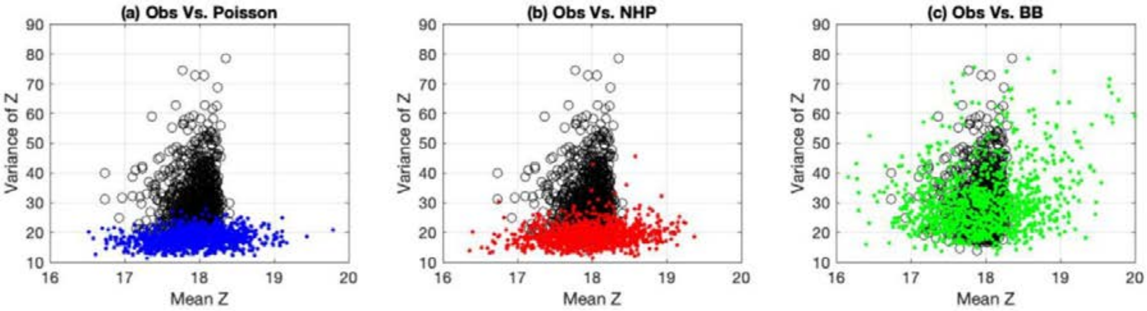


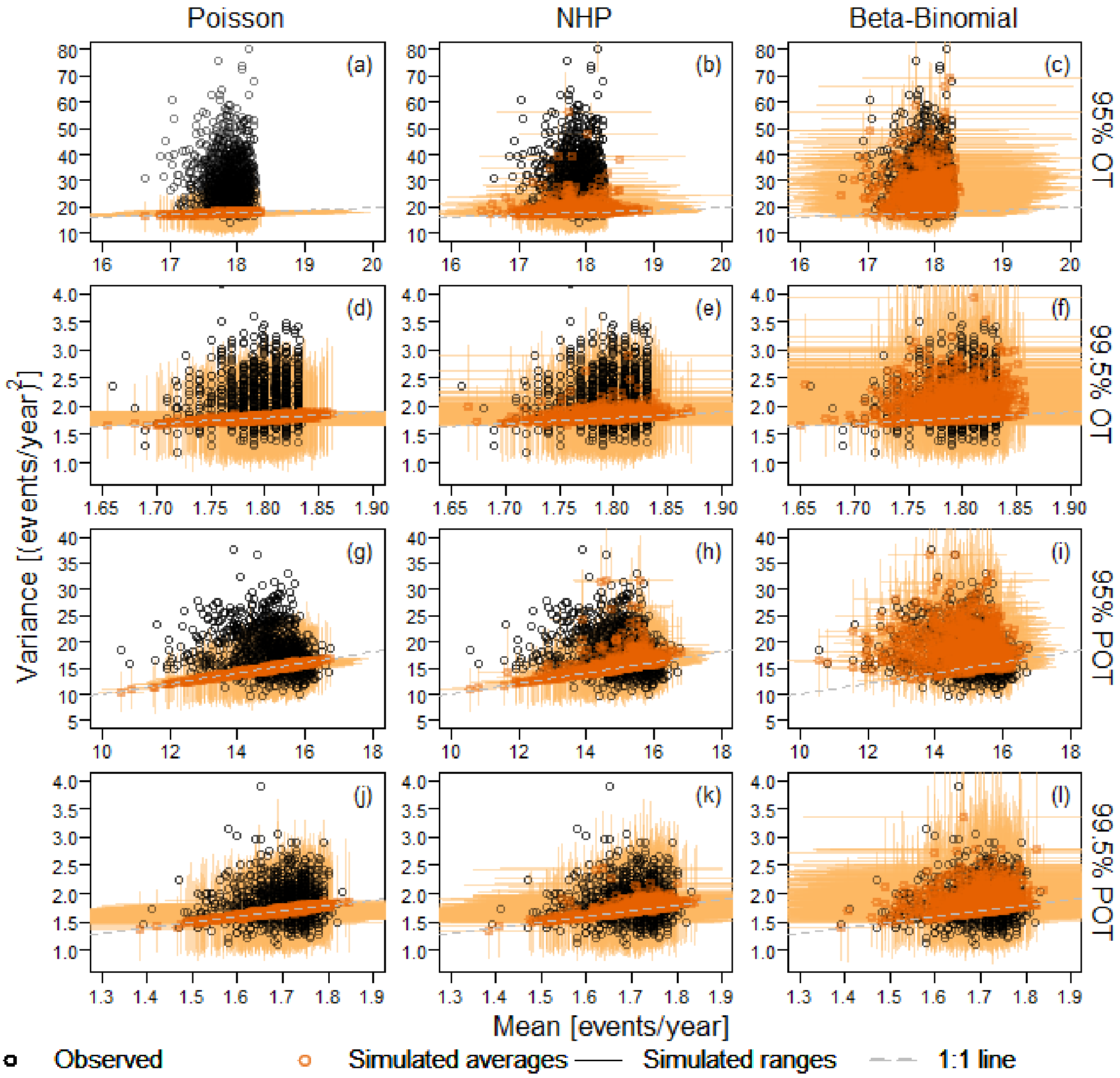
Figure R1. Relationship between mean and variance of observed OT samples and those corresponding to simulated samples from (a) Poisson, (b) NHP, and (c) BB distributions.

That said, I produced Figure R1 by (i) estimating the parameter/s of each model separately on each observed sample, and (ii) generating a sample with the same size as the observation with the estimated parameter/s. This approach mimics what the author indicated in the caption: "*Observed values are compared with those corresponding to simulated samples from Poisson, NHP, and Beta-Binomial (BB) distributions*". If one follows this approach, one random generation of the synthetic samples of the three models should result in a sampling variability for the mean larger than that of observations and symmetric around the expected value defined by the 95% threshold (by the way: note that, in Figure R1, the variability of mean and variance is the same for the Poisson variates, as expected). In Figure 2 of the paper, the means of the randomly generated samples have instead exactly the same range as the observations. This is incorrect.

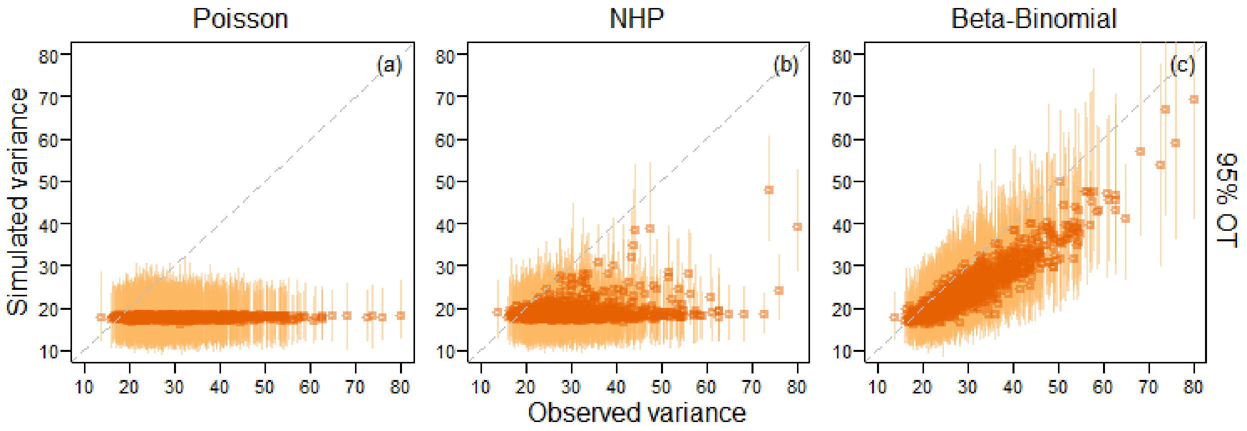
Response

- *x** is the threshold of the precipitation intensity process whereas 18.2625 is the theoretical expectation of the number of over-threshold occurrences over 365.25 time steps (Bernoulli trials).
- I recognize the text is not clear. Obviously, the orange points in Fig.2 are not the coloured points that the Reviewer reports in his Fig. R1 (indeed, the diagrams are obviously different). Let me explain. If we simulate 100 values from a Poisson distribution, the sample average is obviously characterized by large variability. To properly link the sample variances of the simulated samples with the parent distributions for effective visualization, we can choose different strategies. For example: (1) plotting the variance of the simulated sample versus the mean of the observed sample (i.e., the parameter of the generating Poisson/NHP/BB distribution), or (2) simulating *B* time series for each location (e.g., *B*=100), then plotting the averages of the *B* values of mean and variance (along with their ranges) or (3) plotting directly observed variances vs simulated ones.

  The first approach is what is reported in the paper, as it provides very simple and clear representation, while the second method is used to create the figure below, where the orange dots denote the averages over 100 simulations for each location, and the vertical and horizontal lines the range around those averages.



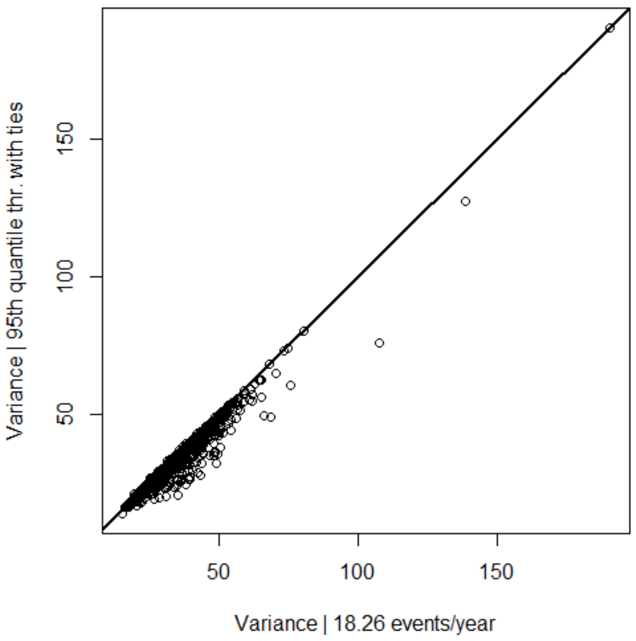
The third method yields the figure below.



One can see that the three approaches provide the same message: Poisson and NHP do not describe the observed variability. Is BB perfect? No, of course: it is just a model. However, BB works much better than Poisson and NHP.

So, why do Reviewer's single "*random generations*" look different from the observed "cloud" of points? Because the 1106 observed means are not random samples from a set of Poisson/NHP/BB distributions (obviously), but the parameters (expectations) of the Poisson/NHP/BB models used to simulate and resulting from a thresholding procedure.

Do things change if we extract exactly 18.2625 events per year? No, really. Both CDFs (not shown) and variances of *Z* (see figure below) do not change very much.



Therefore, independently of OT selection and diagnostic diagrams:

1) the number of OT occurrence is over-dispersed, as expected from theoretical considerations.
2) Poisson and NHP are not able to describe overdispersion, and they are not suitable models for these data.

3) If the marginal distribution is assumed to be Poisson (in contrast with empirical evidence), data cannot be modelled with NHP because the marginal distribution cannot be simultaneously Poisson and mixed-Poisson (this is a logical contradiction). For NHP, one could check conditional Poisson behaviour, at most.
4) Whatever additional analysis based on Poisson or NHP models (as those reported for instance by Farris et al.) is uninformative, as it relies on models that do not describe the observations.

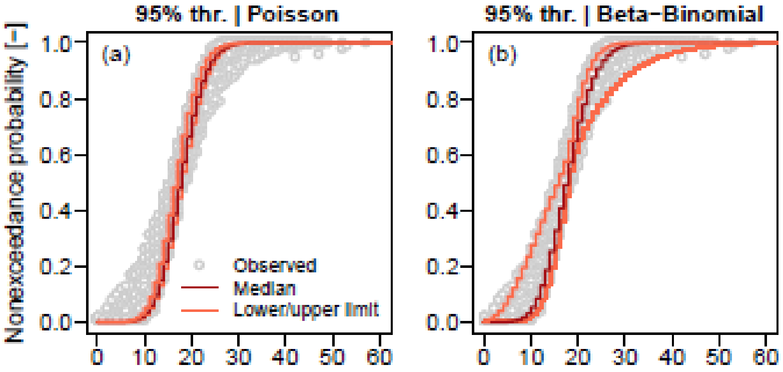
CDFs of observed samples and fitted distributions

Apart from the issue indicated above, Figure 2 (and Figure R1 above) shows that, for several cases, the variance of the observed samples is larger than the mean. The one-parameter Poisson distribution is unable to fully capture this spread, while, as well known, a two-parameter distribution like the BB can be fitted to reproduce both the mean and the variance. To further prove the point that the BB is a better distribution, the author presents in Figure 3 cumulative distribution functions (CDFs) and differences in probability of observed and fitted distribution (I believe). However, the author did not explain how that figure was created. What are the lower and upper limits and how were they calculated?

Response

L316-317: ***"****We compare ECDFs Fn(z) with the CDFs FP(z) and FβB(z) of Poisson and βB models, respectively. Probability plots are complemented with diagrams of the differences ($F_n(z)-F_{model}(z)$) versus z.****"***

Interval calculation is standard: take the 1106 CDFs, calculate the 1106 probability values corresponding to a set of quantiles, and calculate minimum, mean (or median), and maximum probability (or pointwise CIs) for each quantile. Alternatively, one can draw the 1106 CDFs (as further discussed below). The resulting diagram (see Fig.3b, reported below for convenience) shows that the ensemble of BB distributions can describe the global over-dispersion characterizing the observed samples reasonably well. On the other hand, the Poisson distribution cannot. Of course, these results are consistent with the foregoing mean/variance diagrams.



More importantly: can we clearly say that the two-parameter BB distribution provides a large improvement compared to the one-parameter Poisson distribution when looking at panels (e) and (f) of Figure 3 in the paper? Since I could not understand it well, I compared, for some representative gages (selected based on an equiprobability criteria to fairly explore all possible behaviors), the empirical CDF of *Zi* and the CDFs of the fitted Poisson and BB distributions (see comment 3.1 regarding how parameters of the BB were estimated). I chose the gages based on the variance of *Zi*, whose empirical CDF across all gages is shown in Figure R2. I picked the gages with variance associated with a cumulative frequency, *F*, close to 0.1, 0.2, 0.3, …, 0.9. Results are shown in Figure R3: these graphical diagnostics (which the author recommends using in general) indicate that the two distributions are not markedly different, even for the largest values of the variance. This is also quantified by the values of the Cramer-von Mises goodness-of-fit metric (without any penalty) provided in the legend, which are very similar despite the BB distribution having an additional parameter compared to the Poisson, ranging from 8 to 22 for BB and from 9 to 24 for Poisson fitting. In addition, the variability of such metric does not seem related to the variance of *Zi*, suggesting that the gain of applying the BB against the Poisson when the variance of *Zi* becomes increasingly larger than the mean is negligible. <u>Therefore, for most cases, a parsimonious one-parameter distribution like the Poisson does not seem to do a bad job when characterizing the frequency of the empirical counts as compared to the BB (as proposed by the author), which depends on two parameters, and thus should be properly penalized in any comparison.</u>

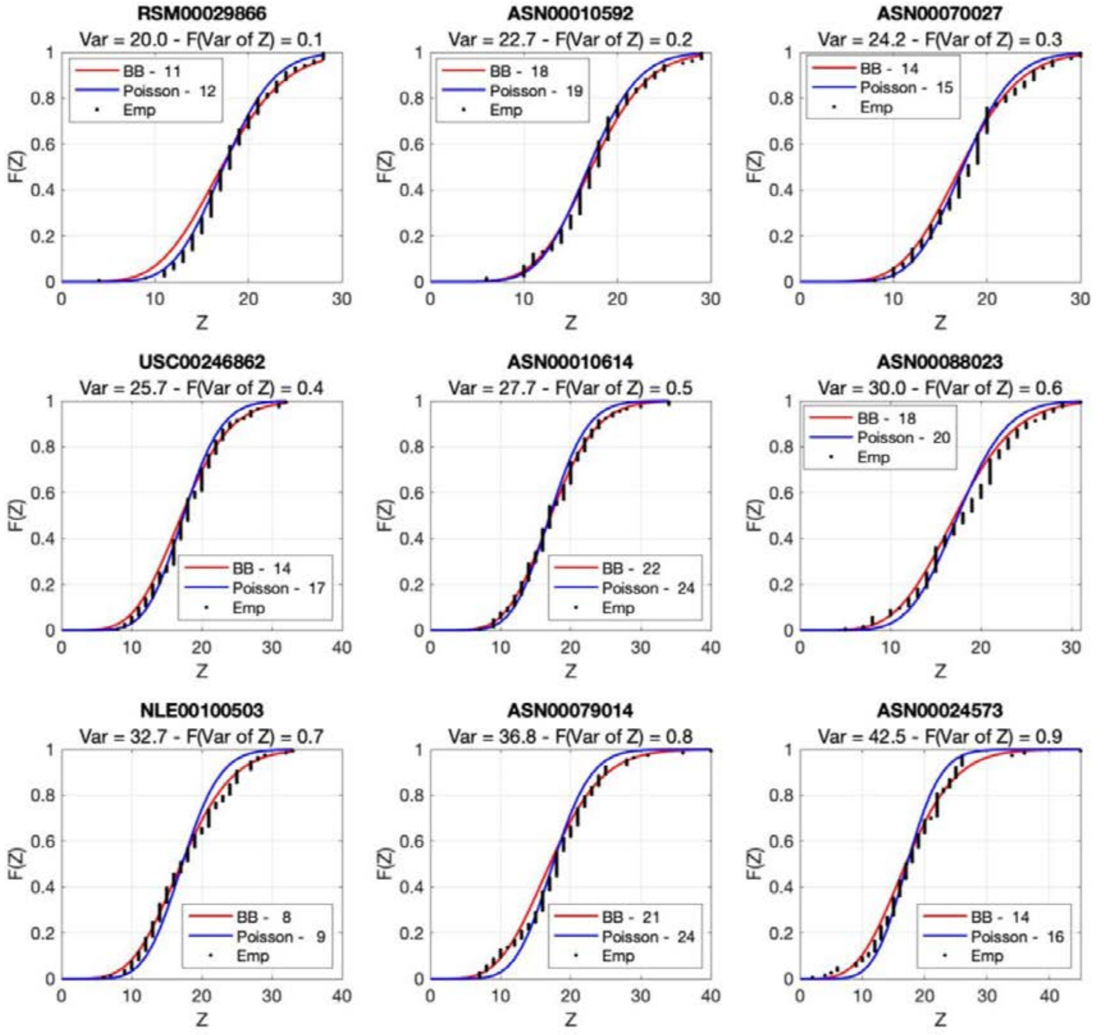
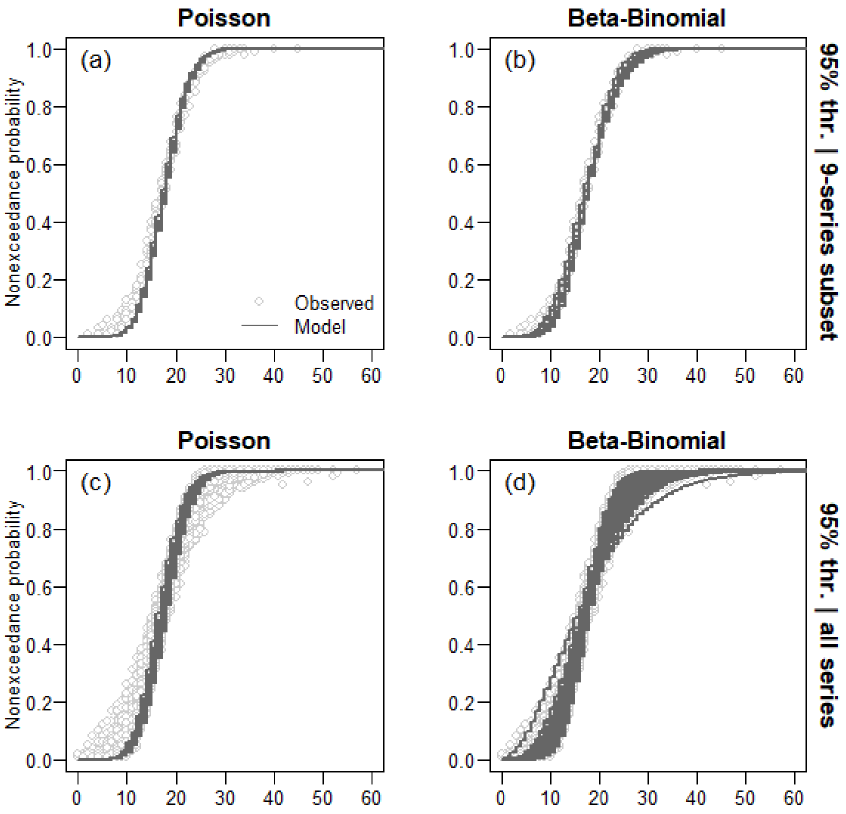


Figure R3. CDF of the observed Z samples, along with fitted Poisson and BB distributions. In each panel, the gage ID is shown along with the values of the variance of Z_i and the cumulative frequency of the variance from Figure R2. The values reported in the legends next to "Poisson" and "BB" are the values of the Cramer-von Mises goodness-of-fit statistics.

Response

- Diagrams are powerful tools either to emphasize evidence or to conceal it. Using nine separate figures with a variable range of the x-axis tends to conceal the fact that the Poisson distribution shows (almost) no variability compared with the BB.

- A fairer visualization is possible by drawing the nine samples in the same panel. Panels (a) and (b) in figure below provide such a fairer comparison, showing that the BB model allows for capturing the observed variability of the lower/upper tails of the empirical distributions (over-dispersion), while the nine Poisson distributions are almost identical (they are identical if we use the rule of extracting exactly 18.2625 events/year).
- More importantly, the comparison of panels (a) and (b) with panels (c) and (d) in figure below show that the nine time series selected by the Reviewer are not representative of the over-dispersion of the 1106 time series. When we consider the whole sample, the difference between BB and Poisson is more evident (obviously).
- Drawing conclusions from the 0.8% of data is rather questionable, especially if we can easily look at the whole sample (by proper diagrams).



- Reporting values of test statistics is uninformative as they are distance metrics affected by their own sampling variability. They can be compared only when they are associated to their p-values or the statistical test is non-parametric (i.e., the sampling distribution of the test statistics does not depend on the tested distribution), which is not the case here.
- Of course, roughly speaking, reporting p-values would mean performing a GoF test. In this respect, the Reviewer does not even need to report such information because Farris et al. already concluded that the Poisson model is a good distribution for the whole sample. However, such conclusions contradict the evidence reported in the foregoing figures as well as the p-values reported in Fig. 4 of the paper (for three different GoF tests), which indicate that the Poisson distribution should be rejected up to 53% (as expected) when using a more powerful test for Poisson hypothesis.
- Parsimony: BB is not a distribution selected among a bunch of models by playing with GoF tests. BB is one of the models theoretically justified under some specific assumptions, which are deemed to be reasonable for the process at hand. This means that, if it works, it is expected to do so over a range of spatio-temporal scales in the same way any sound

mathematical theory or physical model is expected to work in a range of situations. In this respect, the foregoing figures along with Figs. 3, 11, and 13 clearly show that such models work at least from daily to annual scale and from at-site to global scale. Such figures are reported below for the sake of convenience, as some reader can find some evidence in them.

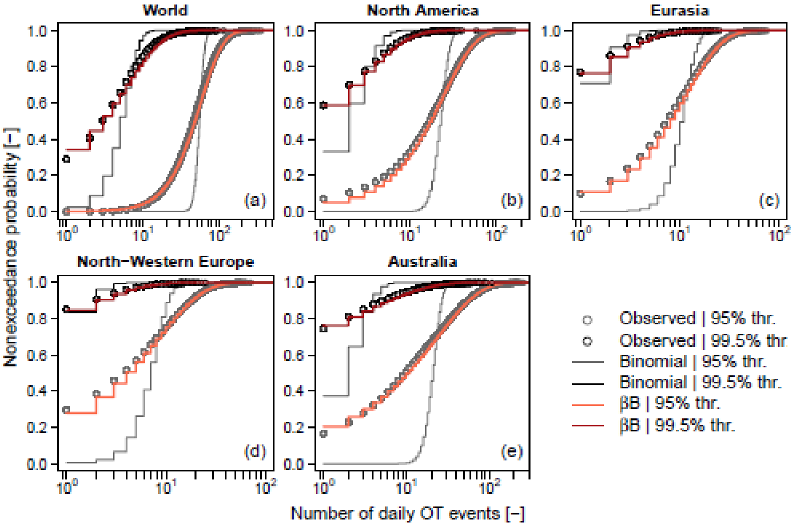


Figure 11. ECDFs of number of OT events (for the 95% and 99.5% thresholds) occurring at daily time scale over different regions along with Binomial and $\beta\mathcal{B}$ CDFs.

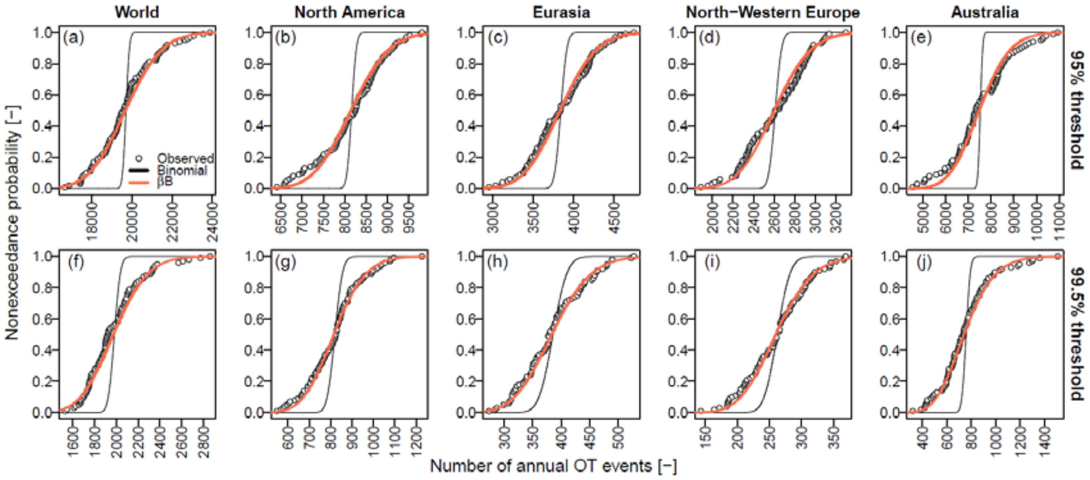


Figure 13. ECDFs of number of OT events (for the 95% and 99.5% thresholds) occurring at annual time scale over different regions along with Binomial and $\beta\mathcal{B}$ distributions.

Why is bias correcting the power spectrum needed?

The author mentions the use of bias correction of the power spectrum, assuming the presence of fractional Gaussian noise, as described in the Supporting Information. However, the author did not properly explain nor demonstrate why this is needed for the series at hand and the level of subjectivity associated with the choice of the bias correction method. In Figures R4-R9, panels (a) and (b) show examples of time series for a few randomly picked synthetic samples along with the upper and lower limits and the median serial correlations of 10,000 synthetic samples generated without any bias correction. If the lag-1 serial correlation ρ1 is used to measure the strength of the serial correlation structure, it is important to mention that ρ1 < 0.2 (0.3) for the *Zi* of ~90% (97%) of the gages. Gages with values of ⍰1 up to 0.3 are reported in Figures R4-R7. For these cases, the autocorrelation function of the synthetic samples does not seem to be affected by any bias. Some bias starts appearing for stronger serial correlation structures, as shown in Figures R8 and R9. However, the time series of Figures R8 and R9 clearly show an increasing trend that induces those strong correlations… While one of the main concerns raised by the author is the subjectivity that is often adopted to choose trend forms, the reasons why the bias correction was applied is not motivated in the paper, nor it is shown the effectiveness of the bias correction across several strengths of the autocorrelation. I also wonder how, in Figure 9, results look like for the case of IAAFT without bias correction plus false discovery rate (FDR) test.

Response

Following the rationale of statistical inference, bias correction is not subjective and is not even an option, but a necessity resulting from the assumption of dependence.

If we *assume* "dependence", we know *a priori* that the estimators of ACF and spectrum are biased when the estimation rely on short samples. This is well known and widely discussed in the cited literature (e.g., Marriott and Pope, 1954; White, 1961; Wallis and O'Connell, 1972; Lenton and Schaake, 1973; Mudelsee, 2001; Koutsoyiannis, 2003; Koutsoyiannis and Montanari, 2007; Papalexiou et al., 2010; Dimitriadis and Koutsoyiannis, 2015; Serinaldi and Kilsby, 2016a).

Let me use an example familiar to the Reviewer (Mascaro 2018, JH). The problem of ACF/spectrum bias is analogous to that of the estimation of the shape parameter of GEV/GP models for short samples. In these cases, estimates might point to apparent exponential tails; however, such a behaviour might be consistent with heavy tailed *population* models. In other words, short samples from heavy tailed models can look "exponential". Therefore, taking the rough estimates as the "truth" might be a mistake.

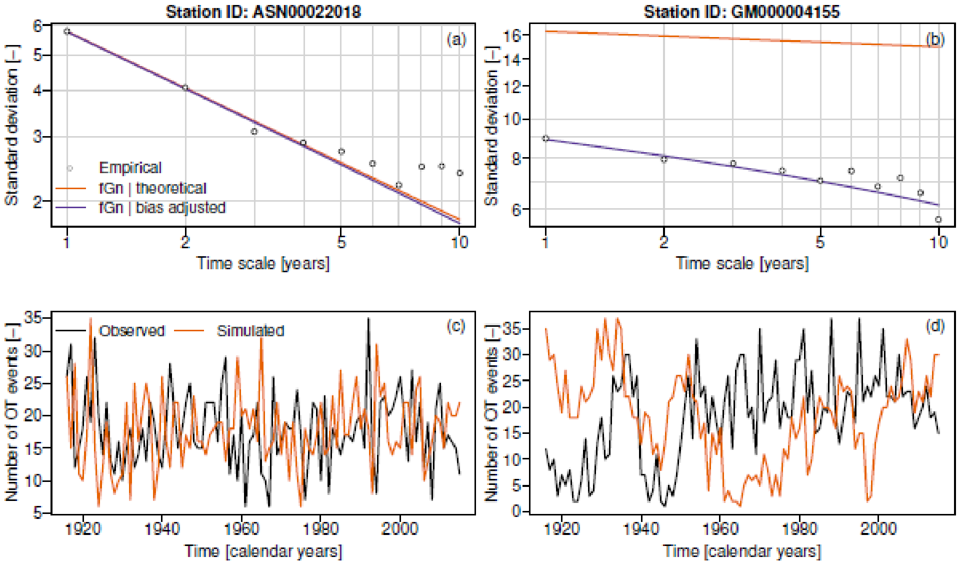
According to the rationale of statistical inference this does not allow to conclude that the analysed process is "heavy tailed" or "exponential" or something else: it just means that under the *assumption* of heavy tails, we can obtain short-sample behaviour coherent with the observations; therefore, if we ***assume*** heavy tailed models, we ***must*** correct for the estimation bias.

The take-home message (that the paper attempted to deliver) is that the choice and the use of estimators ***depend on*** the assumptions we make; we cannot use the same estimator/method under different assumptions, because it might be not valid.

Going back to precipitation data, ***under dependence***, empirical ACF and power spectrum are known to be rather unreliable estimators for short samples (in the same way estimators based on product moments might be inferior in terms of bias and variance to L-moments in hydrological frequency analysis). Therefore, we used the climacogram as a benchmark estimator, as it is known to perform better in these circumstances, allowing for bias correction, etc. In other words, the ACFs shown in fig. R4-R9 are like the estimate of the shape parameter of a GEV distribution made by product moments estimators and neglecting bias correction.

The key point, which is systematically missed in the literature, is that making inference on indices/measures of dependence implies that we ***assume*** (implicitly or explicitly) that dependence does exist, and therefore inference should be made accordingly (accounting for its effects). The same holds for non-stationarity: if we make inference for non-stationary models, we already ***assume*** that non-stationarity is in play; this means for instance that single population moments (mean, variance and covariance) might not exist. As shown below, estimating the ACF for an observed sample under the assumption of non-stationarity (e.g. under NHP) makes no sense, because ***population*** ACF depends on time, and the estimate over a sample is not representative of any theoretical counterpart (like the sample average estimated for instance from a sample drawn from a Cauchy distribution).

That said, contrarily to what stated by the Reviewer, Fig. S1 in the supporting material (reported below for convenience) shows the climacogram for two extreme situations that cover the whole spectrum of cases: Figure S1a,c show a case where bias correction is negligible, while Fig. S1b,d shows an opposite case where the time series looks strongly correlated.



Finally, the Reviewer states "*However, the time series of Figures R8 and R9 clearly show an increasing trend that induces those strong correlations…*"

Saying that the trend induces autocorrelation means ***assuming*** (implicitly, at least), that the process is non-stationary and autocorrelation is an artifact. This is legitimate. However, what about the other way around? Could not be that "trend" an effect of dependence?

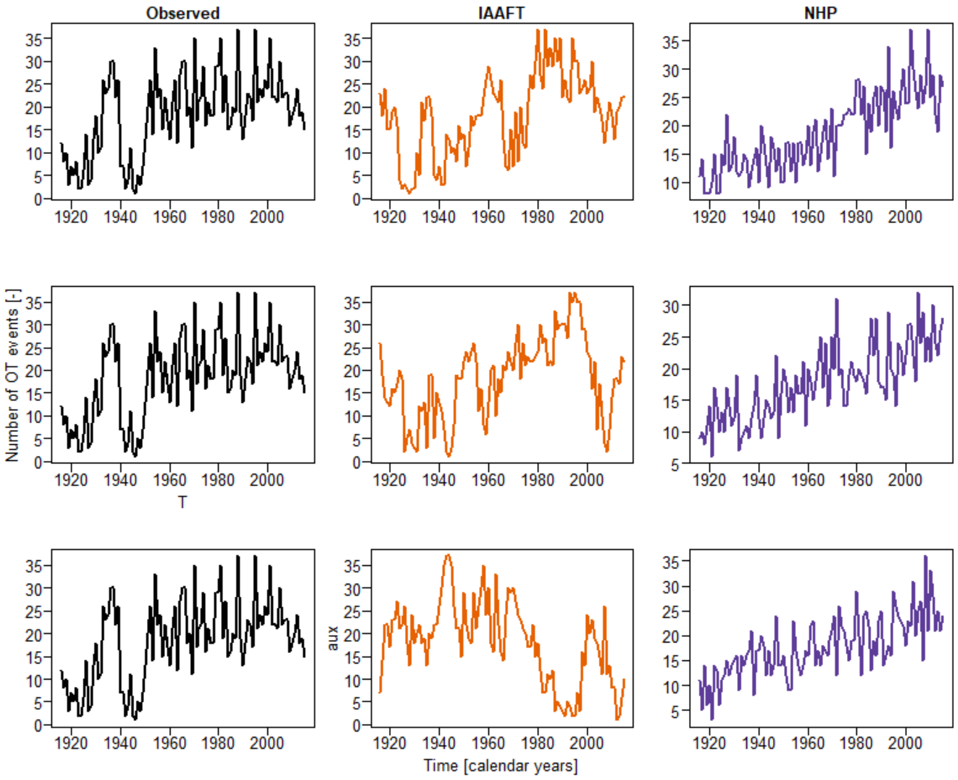
Referring to the series GM00004115 of Fig. S1, the figure below shows some simulations from IAAFT and NHP (with linearly increasing rate of occurrence): Are we sure that the observed data linearly increase? Which approach describes the observed low-frequency fluctuations better?

Based on the analyses reported in Reviewer's report, we should conclude that the second case (stationarity) is not realistic because it does not agree with INAR(1) results. However, the fundamental questions are:

- Can we rely on a model (INAR(1)) that does not even describe the observed marginal distributions?
- is INAR(1) (or whatever specific model) representative of the assumption of stationarity-dependence (and the corresponding infinite models)?
- Are INAR(1) structure and related inference compliant with the consequences of the assumption of dependence?

Please note that I am rather neutral about the use of stationary or non-stationary assumptions, as they are just modelling frameworks commonly used to describe inherently unknown/unrepeatable physical processes.

My criticism is about the lack of coherent and proper use according to the rationale of statistical inference, thus resulting in unscientific conclusions. In other words, using non-stationary models is fully legitimate (of course), but the whole inference procedure should be done accordingly accounting for all the effects of such an assumption, and results cannot be used to discard alternative assumptions, as these generally require other models and alternative inference methods.



Can the IAAFT model be used to generate the null distribution of a trend test?

Another concern that I have is whether the IAAFT model is appropriate for generating the null distribution of a trend test. I estimated the slopes of the linear regression φ (which can be considered a proxy of a trend test metric, being a first order expansion of any trend behavior) for 10,000 synthetic samples generated by the (stationary, but correlated) IAAFT, as suggested by the author, and plotted the empirical density function. Then, I did the same with the (stationary uncorrelated) Poisson distribution, as a reference. Panels (c) of Figures R6-R9 show that the null distribution of φ is bimodal for the IAAFT. Under the hypothesis of stationarity, we should expect a symmetric distribution with the mode at φ = 0, like in the trivial case of the Poisson distribution. Thus, distributions like those shown in Figures R8 and R9 raise serious concerns on the power of any trend test based on the assumption of IAAFT distribution. I want also to stress that this aspect is completely neglected by the author, despite the large effort dedicated to criticizing some assumptions and conclusions of the paper of Farris et al. (2021). Consequently, why not dedicating a very small additional effort to evaluate the power of tests using the distribution/model that the author proposes as an alternative to test the null hypothesis of no trend?

Response

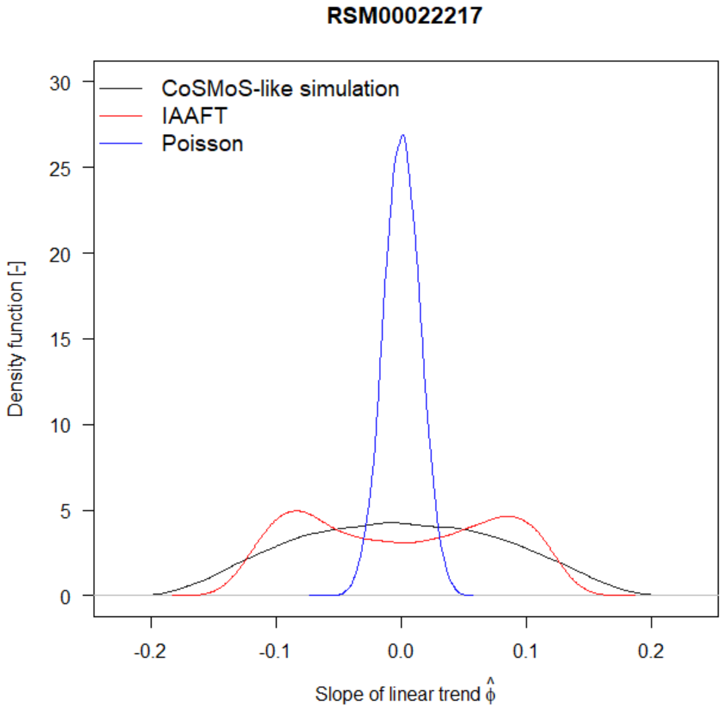
The null distribution depends on the null assumption and the corresponding models. The bimodality reported in Figs R6-R9 depends on the fact that IAAFT yields “constrained” simulations instead of “typical” simulations. This means that all simulated time series share approximately the same power spectrum and therefore the observed “trendy” patters (either increasing or decreasing), while zero slopes are less likely for that observed spectrum.

If we relax that assumption and use "typical" simulation (CoSMoS-like, so to speak), we recover unimodality (see figure below). Do things change? No much, because the aim of the paper is not to find inexistent "perfect models", but to show that there is a "world of options" beyond INAR(1) and NHP, and discarding one of them cannot imply proving/disproving stationarity or non-stationarity.

Thus, "*why not dedicating a very small additional effort to evaluate the power of tests?*"
Because

- tests' outcomes are expected and redundant once we know the behaviour of the theoretical processes behind H_0 (here, the matter of fact is not bimodality or unimodality, but variance inflation), and
- the aim of the paper is exactly to show that the same observed behaviour might be described by different frameworks, and discrimination is not possible in this context.

Indeed, in the context of unrepeatable hydroclimatic processes, power analysis has little usefulness, because the observed data might always be described by a virtually infinite number of frameworks and assumptions, which will never be covered by the usually simple/trivial models used for H_0 and H_1. Moreover, such assumptions refer to models not to the physical processes.
Things are different when we refer to designed experiments, where we can control influencing factors and therefore, we can measure power (because we can control H_0 and H_1), effect size, sample size, etc.
This is the meaning of the discussion in Section 6.2, which summarizes the message reported in previous papers of mine and co-authors.



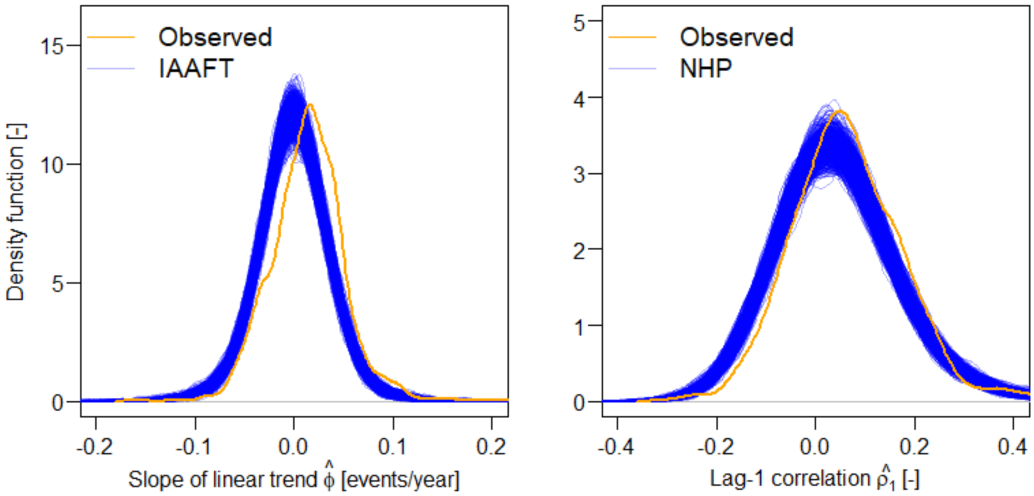
Does the IAAFT model capture the variability of the observed linear slopes?
Finally, since the author presents the IAAFT model as a proper stationary model that can explain the presence of trend in virtue of correlation, then one would expect that it should be able to capture the empirical distribution of the observed slopes of {*Zi*}. I applied the IAAFT model for each gage (without bias correction), generated one synthetic sample, and estimated the slope. I then plotted the empirical PDF of slopes of the observed and synthetic samples. I repeated this nine times to

explore the sampling variability. As shown in Figure R10, the IAAFT model is not able to reproduce the observed sampling variability and, in particular, the larger number of positive slopes.

Response

As shown in the figure below, IAAFT *does not capture* the variability of the observed linear slopes as well as NHP *does not capture* the variability of the 'observed' ρ_1.

Therefore, if we think that IAAFT (stationarity) is not satisfactory, we should also discard NHP (non-stationarity) for its poor reproduction of ρ_1.



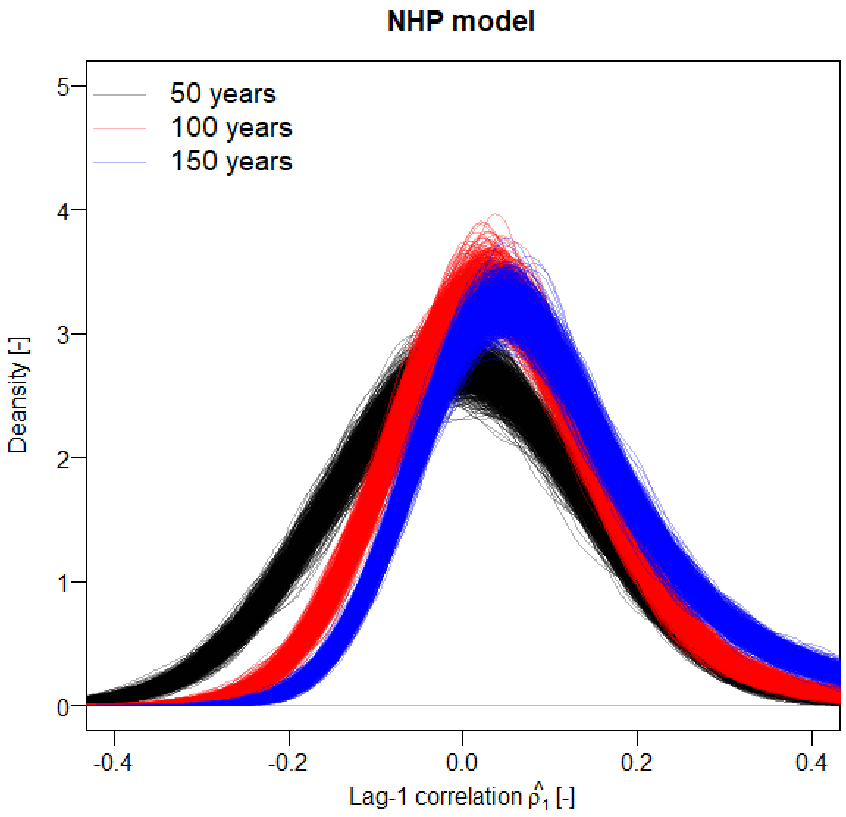
Of course, the terms of the problem are a bit different, and correct interpretation requires (… once again) to account for assumptions:

1) Results in Fig. R10 assume that the raw estimates of ρ_1 are presentative of the population values, neglecting the problems affecting ρ_1 estimators both ***under dependence*** and ***under non-stationarity***.
2) Results in the foregoing figure and Fig. R10 refer to independent at-site simulations, i.e. they do not account for the effect of spatial dependence (which is a reasonable assumption in rainfall fields), which is identical to the effect of temporal dependence, resulting in spatio-temporal 'trends' and 'clusters'. This means that possible local temporal trends are shared by several stations because of spatial correlation. The overall effect of spatial clustering combined with short samples sizes is that we can have asymmetry over some areas (or the whole domain) and/or some time windows.

   More generally, the foregoing simulations reasonably underestimate the actual variability of sampling statistics (not only ρ_1 and φ).

   Furthermore, we should consider the effect of unknow factors affecting precipitation records, including very basic issues related to data collection, handling, and storage. Such global data should be taken with a pinch of salt. And finally, observed data do not come from models, let alone trivial statistical models.
3) ***Under non-stationarity***, the estimates of ρ_1 over 100 time steps do not represent any population counterpart, because the ***population*** moments of NHP are non-stationary and vary with time. For NHP processes, the figure below shows how ρ_1 estimates evolve when we increase the sample size from 50 to 150 years. Why just up to 150 years? Because NHP models with negative slopes φ yield smaller and smaller rate of occurrence as time increases, until the rate of occurrence becomes zero, and the distributions degenerate (i.e., they do not exist). This is a typical problem of non-stationary methods when non-stationarity

(temporal evolution) does not result from robust physical justifications, but just from curve fitting (… usually, untenable straight lines, as for the NHP).



The fallacious approach to statistical inference (i.e. inverting the role of assumptions and inferential tools) leads to think (incorrectly) that the estimates of ρ_1 (via standard estimators) can be used to compare models for which such estimates might be not valid.

On the other hand, proper statistical inference would imply to select an assumption and perform the whole inference coherently, accounting for its effect. As for the GoF tests discussed above, this means to consider dependence and non-stationarity such that:

- ***Under dependence***, we have some estimates of ρ_1 obtained from proper estimators, accounting for possible bias, etc.
- ***Under non-stationarity***, we must be aware that the moments depend on time and any estimate of ρ_1 is not representative of any invariant ACF function, which does not exist, being time dependent.

In other words, under dependence and non-stationarity, ***population*** statistics/properties, such as ACF, have a different meaning and interpretation (and might not even exist). We cannot use the same estimators (usually, valid under i.i.d.), as the resulting estimates refer to different and often incompatible *population* objects, thus making direct comparison technically meaningless.

Lack of details regarding the estimation of the BB parameters

The explanation given in the supporting information of how the intra-cluster correlation parameter of the BB is estimated from a time series {*Yj*} is confusing. The meaning of “cluster” is not explicitly provided, while it should be. Based on Ahn and Chen (1995), a cluster should correspond to {*Yj*} in a given year (i.e., *j* = 1, 2, …, 365): is this the case? If so, each cluster has size *n* = 365 (apart from leap years). The author talks instead about “experiments” on specific days (of the year) *j* and *l*, but they do not introduce a symbol for the number of available clusters. This should be the size of the vector used to compute the correlation coefficients ⍰*jl*. Put simply, if we have *m* = 100 clusters (or years of records), ρ*jl* is the correlation coefficient between the vectors of the *m* Y's at day *j* for all years and the *m* Y's at day *l* for all years. To the my best knowledge, this is the proper approach, and since the authors do not provide any detail, I followed it for the calculations made in this review.

Response

For *m* locations and clusters of size *n*=365, the BB overdispersion parameter is just the mean of the lagged cross-correlation values up to lag 365. That's it. This will be clarified in the revised version (if revision will be allowed).

What serial correlations and slopes of {Zi} are generated by stationary and nonstationary {Xj}?

Response

I think that Reviewer's simulations and their interpretation suffer from the effect of the epistemological problems discussed throughout the paper and in the responses above:

i) An AR(20) model with GG marginals is not even representative of the whole daily rainfall time series recorded in Athens; it was used as a proof of concept to show the reproduction of ACF and marginals for October rainfall.

   Realistic rainfall simulation should account for seasonality, and more importantly for high/low frequency variations at various spatio-temporal scales… it is a bit trickier task than running an AR(20).

   Thinking that such a model, which is not even representative of a single precipitation time series, can be used to discard the assumption of stationarity and the infinite set of corresponding models means iterating the same logical misconception discussed above.

ii) As mentioned above, ρ_1 values have a different meaning and interpretation under the four different assumptions used in the MC experiments. Here the mistake is to think that the same estimator, which is strictly valid for the first set of assumptions, can be used in the other three cases, and that it corresponds to the same population counterpart in all cases. As discussed above, it is not so.

iii) The left and middle panels of Fig. R11 should be compared with panels (b) and (d) of Fig.5 (reported below for convenience), whose interpretation is straightforward: if we account for the effects of dependence, and we make inference accordingly, we obtain a coherent picture showing that:

   a. Poisson marginals are inconsistent with the observed overdispersion, which in turn is instead consistent with dependence.

   b. INAR(1) does not provide a suitable description of *Z*.

   c. If we use a simulation approach coherent with the assumption of dependence, we are able to reasonably reproduce the observed variability (over 1000+ real world series, not just unrealistic AR(20) samples) shown in Fig.R11 (middle panel). Fig. 5b is rather clear in this respect, while Fig.5d shows the difference between the CI obtained by INAR(1) and IAAFT. These results should be read in conjunction with the performance of BB (Fig. 3), which is another model coherent with the assumption of dependence.

To summarize, we can build a coherent modelling framework under stationarity-dependence that can describe observations over several spatio-temporal scales reasonably well.

Does this mean that IAAFT or BB are a panacea? No, of course.

These are just example models used to show that INAR(1) is inappropriate (it does not even describe the marginals) and cannot be used to discard a whole class of stationary-dependent models. This class includes models that can reproduce a variety of patterns much richer than one can think.

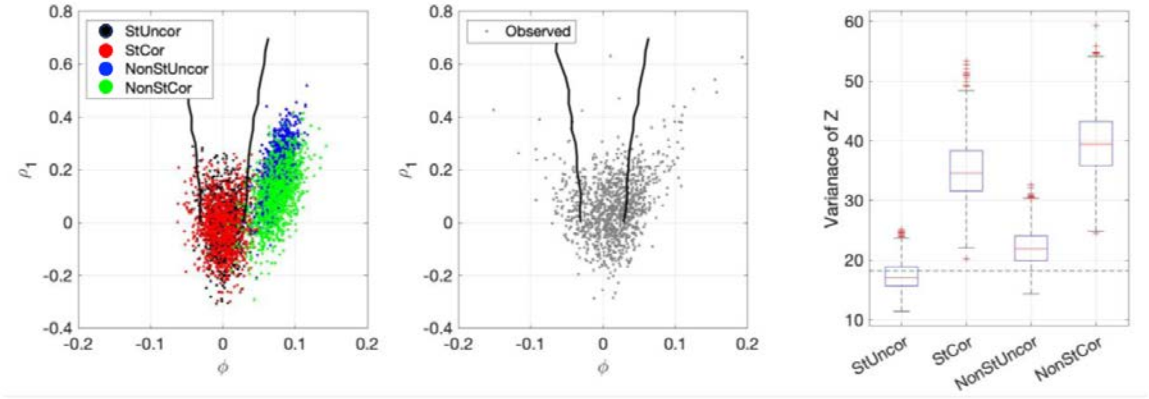
Does this mean that precipitation is "stationary"? No, of course.

Such unrepeatable hydroclimatic processes are neither stationary nor non-stationary! Such concepts only apply to the models we use to describe natural processes.

Everyone can use the modelling framework they like more.

My criticism is about a distorted approach to modelling and inference, which leads to believe that the (poor) performance of a single model (such as INAR) can be used to conclude that a general assumption can be discarded, taking implicitly for granted that (i) such a single model is representative of an infinite class of models, and (ii) it can be used to prove its own assumptions or different assumptions (under which INAR might not even exist).

The scientific method, summarized at the beginning of this reply, implies an opposite approach, whereby inference follows assumptions. Different frameworks should be compared at the end of their own inference, in terms of parsimony, generality, and fit for purposes.



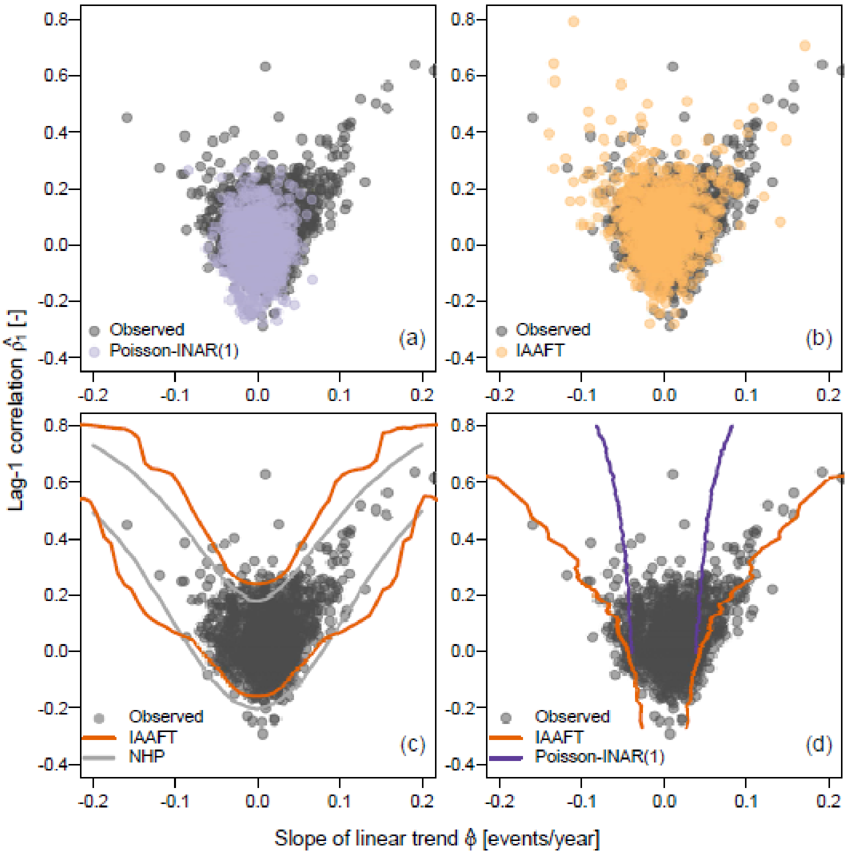


Figure 5. Scatter plots of the pairs $(\hat{\rho}_1, \hat{\phi})$ for the 1,106 observed Z time series over the 95% threshold and 100 years (1916-2015) along with pairs corresponding to Poisson-INAR(1) samples (a), pairs corresponding to IAAFT samples (b), 95% CIs of $(\phi|\mathrm{P}_1 = \rho_1)$ for IAAFT and NHP (c), and 95% CIs of $(\rho_1|\Phi = \phi)$ for IAAFT and Poisson-INAR(1) (d).

Based on these findings, the assumption of stationarity for the application of the BB to model the distribution of *Z* is not supported at all gages. Moreover, the INAR(1) is still useful to assess the nonstationary of the *Z* time series derived from daily P records.

Response

I repeat, this statement results from switching assumptions with models, distorting the rationale of statistical inference:

1- The AR(20) model is not representative of a single time series, let alone worldwide spatio-temporal variability of precipitation "*at all gages*".

2- INAR(1) does not even describe the marginal distribution of *Z*. It does not represent the whole class of stationary models whatsoever, and it cannot say anything about different assumption for which it does not even exist.

3- Figure 5 shows that there might be stationary modelling strategies (different from INAR(1)) that tell a different story.

4- Note that Figures 11 and 13 (shown below once again) report BB derived under the assumption of dependence and stationarity. I would be glad to see an equally simple, general, and parsimonious non-stationary model showing the same goodness-of-fit over the same range of spatio-temporal scales (from daily to annual, and from at-site to worldwide).

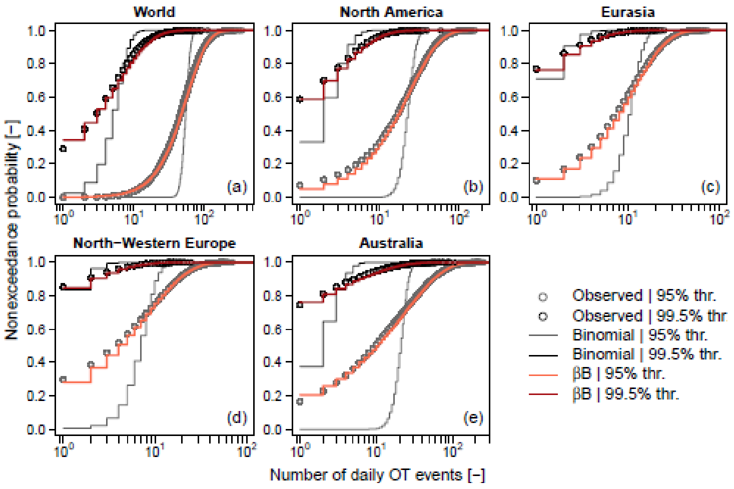


Figure 11. ECDFs of number of OT events (for the 95% and 99.5% thresholds) occurring at daily time scale over different regions along with Binomial and $\beta\mathcal{B}$ CDFs.

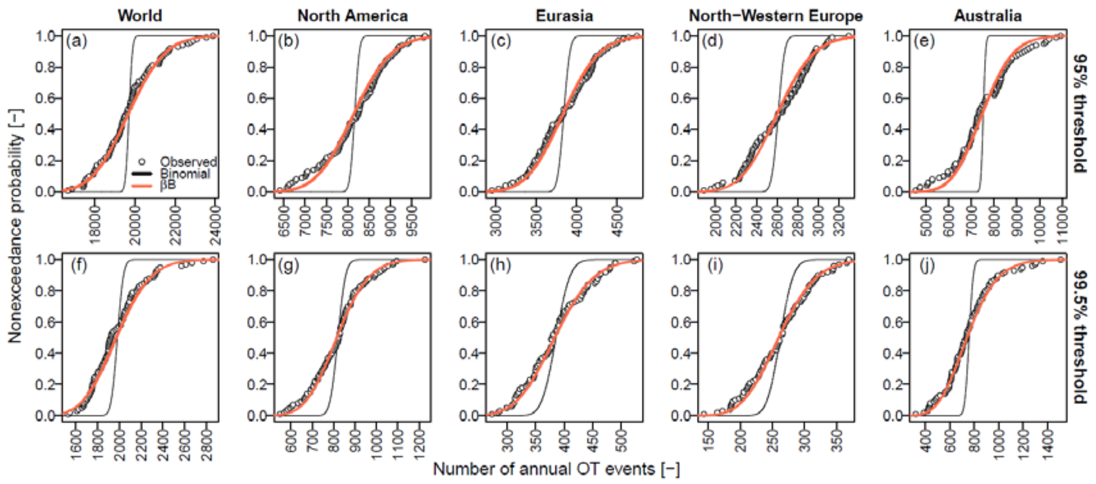


Figure 13. ECDFs of number of OT events (for the 95% and 99.5% thresholds) occurring at annual time scale over different regions along with Binomial and $\beta\mathcal{B}$ distributions.

Is the BB distribution “good” for {Zi} generated by nonstationary uncorrelated {Xj}?

The author indicated that the BB is the theoretically correct distribution for serially correlated count series. The experiments conducted above show that trends in {*Xj*} series introduce artificial serial correlations in the corresponding {*Zi*} even for uncorrelated {*Xj*} (NonStUncor). In these situations, I found that the intra-cluster parameter of the BB can be “successfully” estimated based on the artificial correlations of {*Yj*}. Therefore, since the BB “works”, the theoretical considerations mentioned above would erroneously provide confidence in the presence of serial correlation and nonstationarity of the series. The results presented in section 3.2 indicate that this is most likely the case for several gages.

Response

“*intra-cluster parameter of the BB can be “successfully” estimated based on the artificial correlations of* {*Yj*}.” I guess that the Reviewer has simulated realistic non-stationary binary processes for all 1106 stations, and therefore he has estimated the BB parameters for various spatio-temporal scales, obtaining an equally good or better fit than that shown in Figs. 11 and 13.

If so, it would be interesting to see such results.

However, the matter of fact is different: even if such 1106 non-stationary binary models existed and worked well, what would be the benefit of introducing additional complexity, which is also difficult to theoretically justify?

I recall once again the rationale of statistical inference, which is:

Correct approach

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.

and ***not***:

Fallacious approach

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 the assumptions.

Therefore, I do not use BB to prove "stationarity" or disprove "nonstationarity": this would be scientifically meaningless.

What I do in the paper is to follow the foregoing "correct approach":

1- ***Let's assume*** (<u>not "let's prove!"</u>) stationarity and dependence. Under such assumptions, can we build a simple, parsimonious, and general modelling framework that describes the observations over a range of scales reasonably well?
2- We select models and make inference fulfilling the assumptions.
3- We compare model/inference output with observations. If the models perform satisfactorily for our purposes, they are usable and valid within the limitations of the original assumptions.
4- If the models do not work, we can try other models under the same assumptions or different assumptions bearing in mind parsimony, generality, and reasonable simplicity.

That's it. We do not try to prove model assumptions or alternative model assumptions (… "model assumptions", not "physical process assumptions"!)

Instead, following the "fallacious approach", this is what most of the literature on these topics does:

1- Take several models under different assumptions.
2- Make inference (e.g., GoF tests and ACF estimation) neglecting the effect of those assumptions on tests, estimators, etc.
3- Discard an assumption based on the performance of a single model, which is obviously far from being representative of the whole class of models corresponding to that assumption.
4- Ascribe the retained assumptions to physical processes, whereas they only apply to models.

This fallacious approach is like trying to use the Euclidean geometry to prove or disprove the definition of "point" and "line".

It is well-known that deterministic trends yield spurious 'correlation' and dependence yields spurious 'trends' (based on inappropriate estimators). However, this is irrelevant because precipitation is neither "stationary" nor "non-stationary" and does not come from any model.

One can only use the assumptions that they prefer and check which approach yields the most convenient, parsimonious, and general description for the purposes of interest.

On the contrary, mixing incompatible methods and models attempting to prove assumptions that are incorrectly attributed to physical processes results in confusion, misinterpretation, and misleading conclusions that are inconsistent with scientific reasoning.

The results presented in section 3.2 of Reviewer's report do not indicate anything because they refer to models that do not even represent a single complete precipitation series, and all estimated values suffer from the theoretical inconsistencies discussed above.

Assumption of stationarity for the BB model applied to spatiotemporal precipitation time series
The author applies the BB distribution for the counts at multiple sites with spatially and temporally correlated records under the assumption of stationarity of the correlations. Such an assumption has not been tested in any way by the author, and it might end up being as good or bad (as shown in my comments above) as the assumptions made to test trends that the author criticizes.
In this regard, the CIs in Figure 14 are obtained using the entire record of 100 years, but it could have been instead generated by applying the BB distribution using the first 50 years and results tested with the subsequent 50 years. I assume that other ways to test the assumption of stationarity of the correlations could be designed or, perhaps, found in the literature.

Response
My previous remarks apply here as well:

- Observed correlations are not (and cannot be) either stationary or non-stationary!
- We can only observe sampling fluctuations, which should not be confused with population properties.
- The estimation of the correlation itself depends on the assumption of stationarity, dependence, non-stationarity, etc.
- Inference depends on the assumptions behind diagrams, estimators, models, etc. **not vice versa**.
- I do not criticize "*the assumptions made to test trends*" anywhere, because models' assumptions are what they are.
  I criticize the idea that statistical tests can provide information about whatever assumption for unrepeatable processes! Throughout the paper, I actually stress several times that the output of the tests is just what is expected from their underlying ***assumptions***.
- If we assume stationarity and dependence, we obtain the kind of fit shown in Figs. 11 and 13 (reported below once again). The merit of such assumptions is that they correspond to simple models that describe the observed behaviour over a range of scales.
- Thinking that the model assumptions can be tested on observations of unrepeatable processes (that do not correspond to any model, for sure), means overlooking the rationale of scientific enquiry:

"*The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work - that is, correctly to describe phenomena from a reasonably wide area. Furthermore, it must satisfy certain aesthetic criteria - that is, in relation to how much it describes, it must be rather simple.*" (von Neumann 1955).

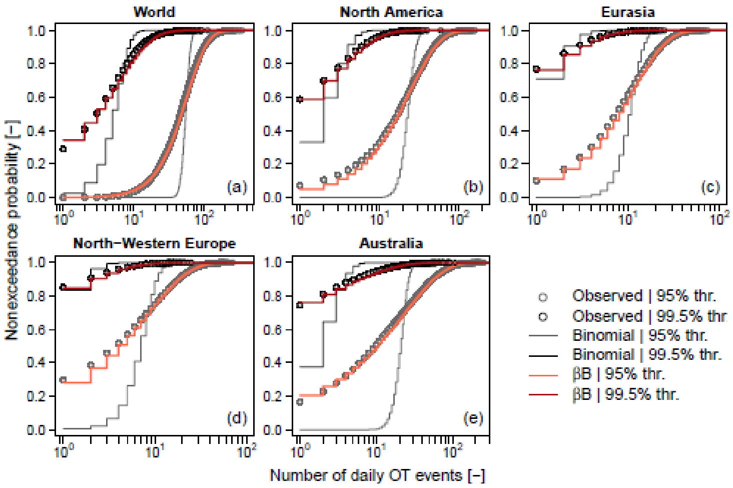


Figure 11. ECDFs of number of OT events (for the 95% and 99.5% thresholds) occurring at daily time scale over different regions along with Binomial and $\beta\mathcal{B}$ CDFs.

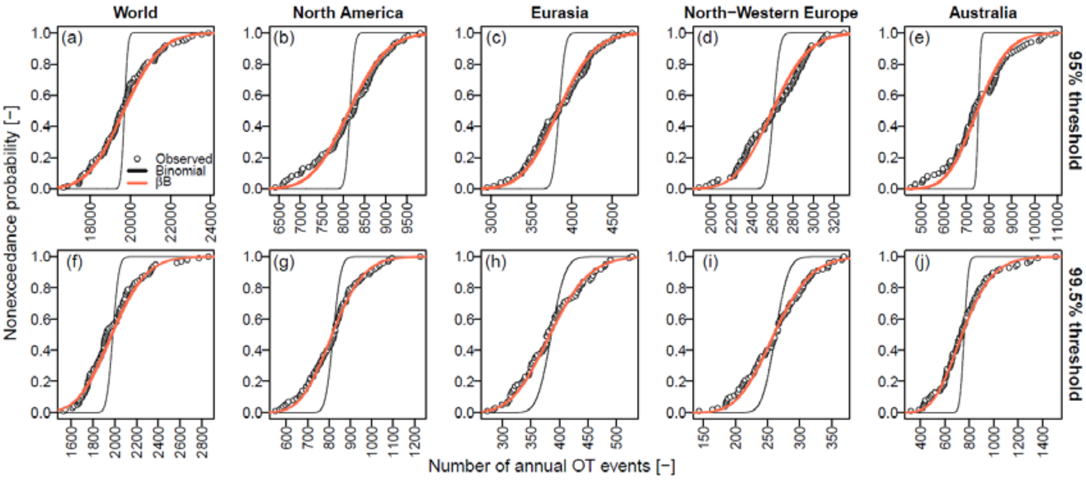


Figure 13. ECDFs of number of OT events (for the 95% and 99.5% thresholds) occurring at annual time scale over different regions along with Binomial and $\beta\mathcal{B}$ distributions.