Review of: “Impact of spatio-temporal dependence on the frequency of precipitation extremes: Negligible or neglected?” by Dr. Francesco Serinaldi

Reviewer: Giuseppe Mascaro

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.

I carefully read the paper by Dr. Serinaldi to properly understand the critiques of some of the results of Farris et al. (2021), of which I am a co-author, and whether the conclusions reached in this paper are supported by the data. As I demonstrate below, while some of the conclusions are reasonable, there are a number of assumptions made by the author that are (1) not empirically supported, and (2) similar in nature to the subjectivity that the author criticizes in the ‘standard’ approaches used in the literature on trend analyses. Based on these issues which are better explained in the points below, I recommend the paper rejection.

I first introduce some symbols following for the most part those used by the author:

- X is the random variable of daily precipitation depth, P. $\{X_j\}$ is the random process of X for discrete times j = 1, 2, 3, … (here, days).

- Y is the binary random variable based on the condition $X > x^*$, with x^* being a given threshold. $\{Y_j\}$ is the binary random process of Y for discrete times j = 1, 2, 3, … (here, days).
- Z is the nonnegative integer counts of P observations exceeding x^* in a specified time window; here, within each year. $\{Z_i\}$ is the random process of Z for discrete times i = 1, 2, 3, … (here, years). $\{Z_i\}$ is related to $\{Y_j\}$ as $Z_i = \sum_{j=1}^{n} Y_j$, with n = 365 (or 366). In words, Z_i is the sum of all the ones in $\{Y_j\}$ that occur in year i. These are the OT samples mentioned above.

Now, I explain my concerns.

1. Value of the BB distribution for the time series of annual counts $\{Z_i\}$

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.

1.1. 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 $\{X_j\}$ series when applying the condition ($X_j > x^*$) to ultimately compute $\{Z_i\}$. 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 ($X_j \geq x^*$).

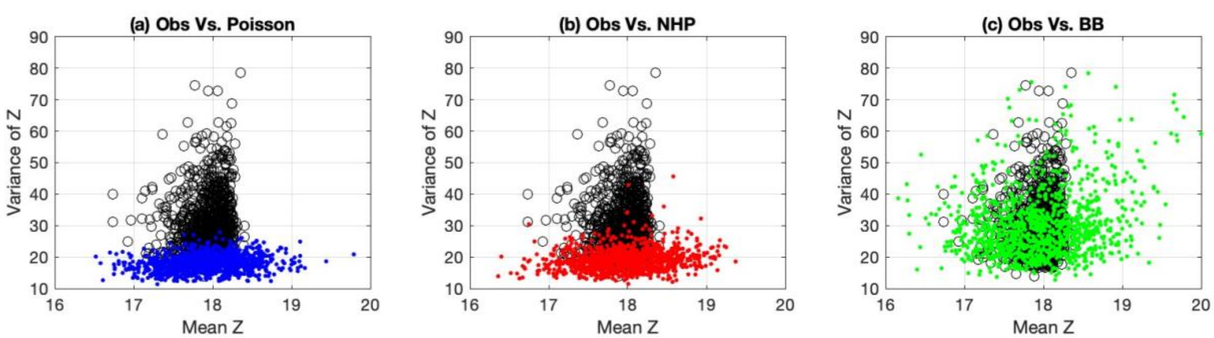

Figure R1. Relationship between mean and variance of observed OT samples and those corresponding to simulated samples from (a) Poisson, (b) NHP, and (c) BB distributions.

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. <u>This is incorrect.</u>

1.2. 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?

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 Z_i 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 Z_i, 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 Z_i, suggesting that the gain of applying the BB against the Poisson when the variance of Z_i 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>

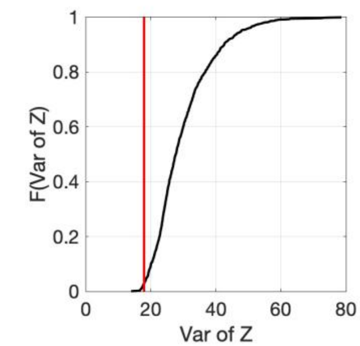

Figure R2. Empirical CDF of the variance of the OT observed samples, Z. The vertical red line is the average of the sample means.

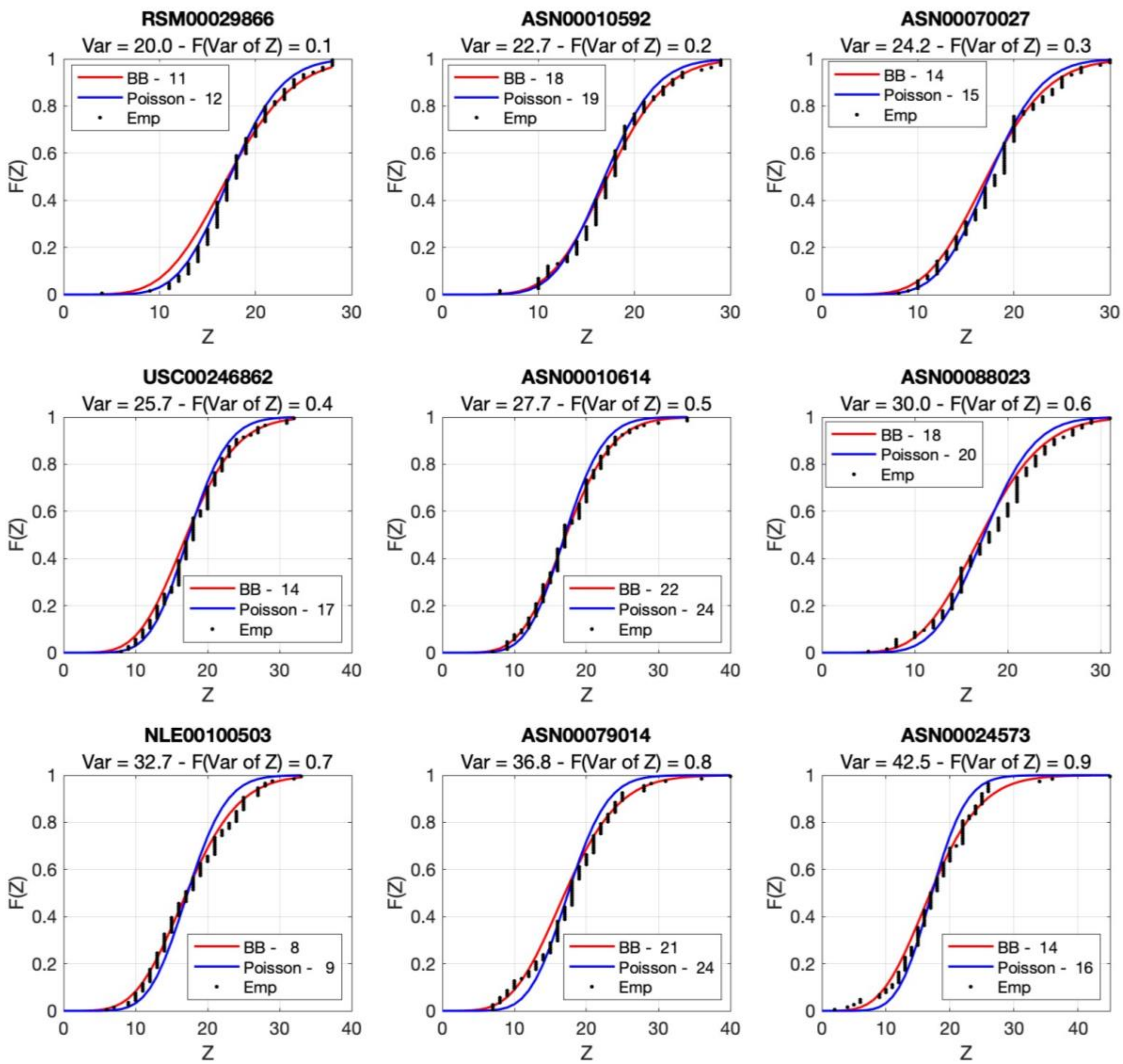

Figure R3. CDF of the observed Z samples, along with fitted Poisson and BB distributions. In each panel, the gage ID is shown along with the values of the variance of Z_i and the cumulative frequency of the variance from Figure R2. The values reported in the legends next to "Poisson" and "BB" are the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

2. Use of the IAAFT model

The author uses the Iterative Amplitude Adjusted Fourier Transform (IAAFT) model to make the case that there are alternative stationary models that reproduce the same linear trend slope and serial correlation of the observed samples. As mentioned by the author, "*IAAFT allows the simulation of synthetic time series that preserve the empirical marginal distribution and, to some error level, the empirical power spectrum of the original data*". The empirical marginal distribution is preserved because the data are reshuffled. As such, it seems to work like a bootstrapping of the data that preserves some lags of the serial correlation (see Section 5.3.5 in Wilks, 2011).

I have three concerns regarding this model. To show this, I picked the $\{Z_i\}$ for a few gages with different strengths of the serial correlation structure and used the IAAFT model to generate synthetic samples.

2.1. 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 ρ_1 is used to measure the strength of the serial correlation structure, it is important to mention that $\rho_1 < 0.2$ (0.3) for the Z_i of ~90% (97%) of the gages. Gages with values of ρ_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, <u>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</u>. I also wonder how, in Figure 9, results look like for the case of IAAFT without bias correction plus false discovery rate (FDR) test.

2.2. 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 $\phi = 0$, 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?

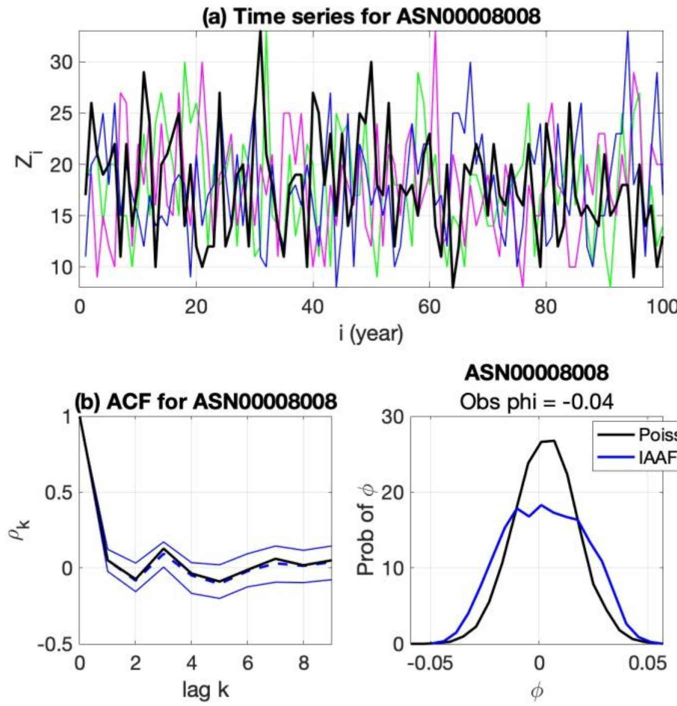

Figure R4. (a) Time series of an observed {Z_i} with ρ_1 close to 0 along with three synthetic time series generated with the IAAFT model. (b) Autocorrelation function of the observed {Z_i} along upper and lower limits (solid blue lines) and the median (dashed blue line) computed from 10,000 samples generated with the IAAFT model. (c) Empirical pdf of the slopes of the linear trend, ϕ, of the 10,000 time series generated with the IAAFT model and the Poisson distribution fitted to the observed {Z_i}. The value of the linear slope of the observed time series is also reported in the title.

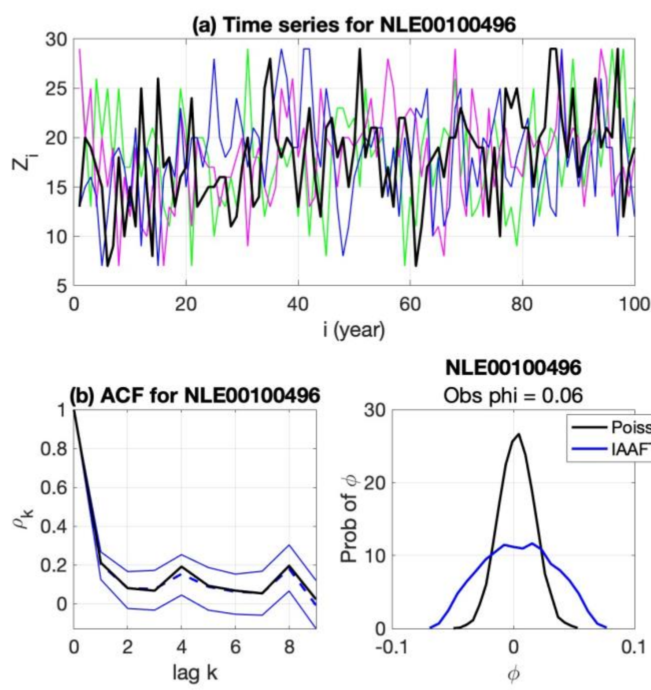

Figure R5. Same as Figure R4 but for another gage where ρ_1 of {Z_i} with close to 0.2.

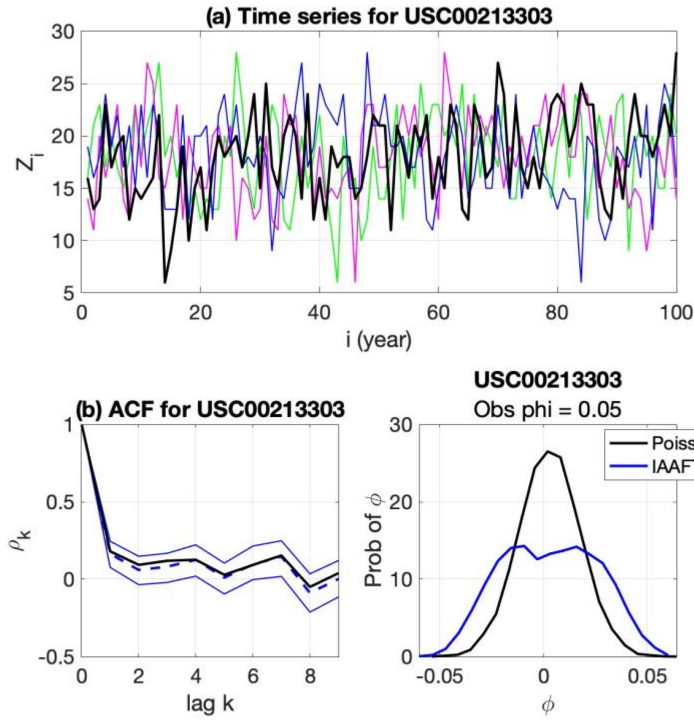

Figure R6. Same as Figure R4 but for another gage where ρ_1 of $\{Z_i\}$ with close to 0.2.

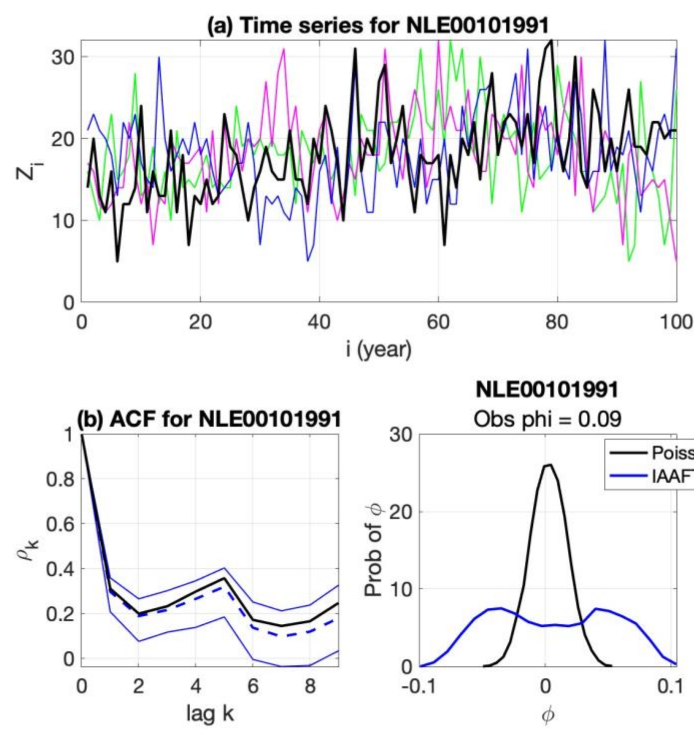

Figure R7. Same as Figure R4 but for another gage where ρ_1 of $\{Z_i\}$ with close to 0.3.

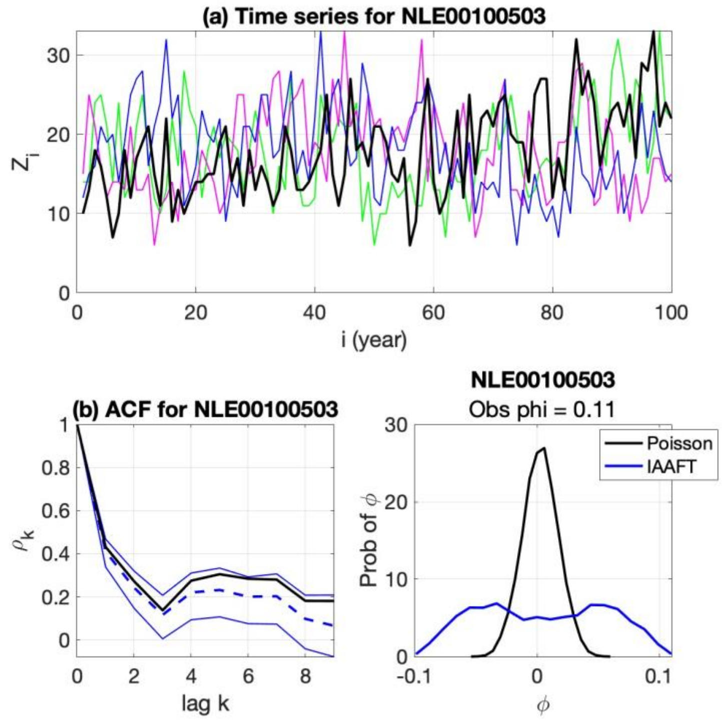

Figure R8. Same as Figure R4 but for another gage where ρ_1 of $\{Z_i\}$ with close to 0.4.

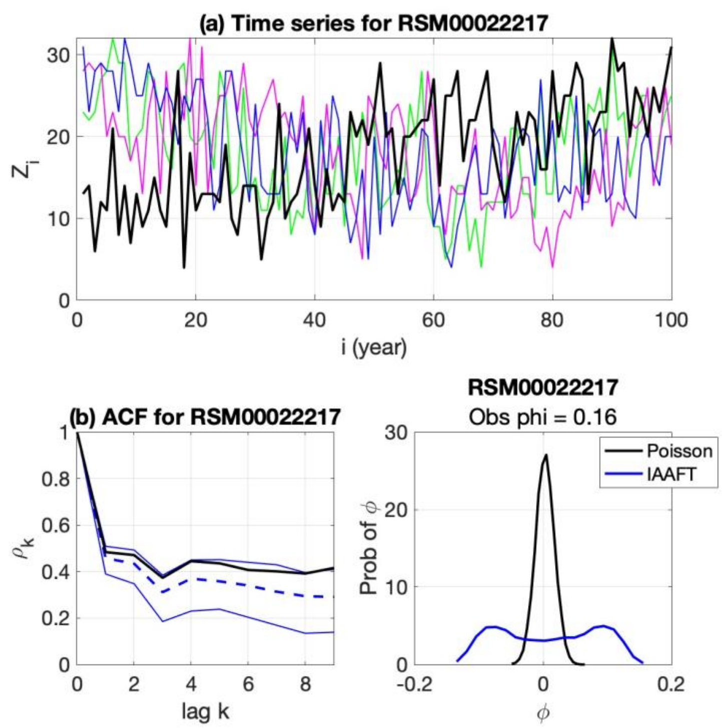

Figure R9. Same as Figure R4 but for another gage where ρ_1 of $\{Z_i\}$ with close to 0.5.

2.3. *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 $\{Z_i\}$. 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, <u>the IAAFT model is not able to reproduce the observed sampling variability and, in particular, the larger number of positive slopes</u>.

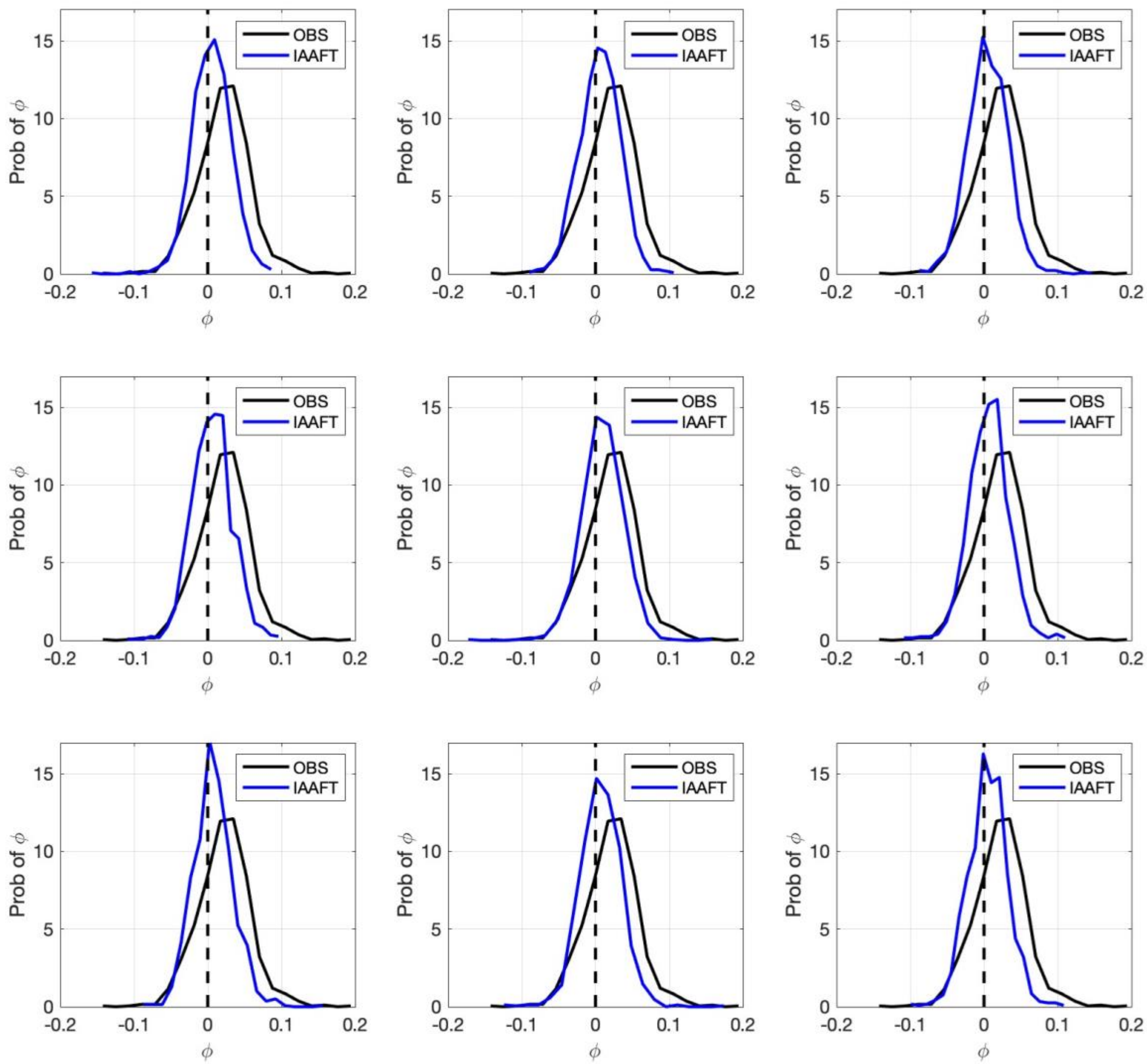

Figure R10. Empirical PDFs of the slopes of linear trend of $\{Z_i\}$, ϕ, for the observed records and the corresponding synthetic sample (one per gage) generated by the IAAFT model. The nine panel refers to a different set of generations with the IAAFT model.

3. Assumption of stationarity for the BB model applied to count time series

One of the points that the author wants to make is that the BB distribution parameterized based on the mean serial correlation of the binary process *Y* assumed to be stationary well captures the distribution of *Z*. I have several concerns related to this point.

3.1. 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 $\{Y_j\}$ 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 $\{Y_j\}$ 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 ρ_{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.

3.2. What serial correlations and slopes of $\{Z_i\}$ are generated by stationary and nonstationary $\{X_j\}$?

As mentioned, the estimation of the BB intra-cluster correlation parameter is based on $\{Y_j\}$, assumed to be stationary. However, the stationarity of $\{Y_j\}$ derives from the stationarity of $\{X_j\}$. This is a very important link that the author has not mentioned nor explored. To do this, I performed Monte Carlo simulations with the Complete Stochastic Modelling Solution (CoSMoS; Papalexiou, 2018) to generate stationary and nonstationary time series of $\{X_j\}$. CoSMoS has been shown able to provide realistic precipitation time series that capture serial correlation and marginal distribution, and Dr. Serinaldi has co-authored a few papers where this model has been further developed and tested (Papalexiou et al., 2021; Papalexiou & Serinaldi, 2020). For simplicity, I used the parameters of CoSMoS presented by Papalexiou (2018) for the daily rain gage of the National Observatory of Athens, Greece (see section 4.3 and Figure 6 of that paper), including those controlling the serial correlation and the marginal distribution of the P depths, *X*. For the latter one, Papalexiou (2018) uses the Generalized Gamma (GG) distribution, which depends on two shape parameters and one scale parameter, β. I performed the following experiments, where, in each case, I generated 1000 time series $\{X_j\}$ of 100 years each (i.e., size of 100 x 365 which is basically the same as the observed series in the paper):

1) **StUncor**: stationary and uncorrelated time series obtained as variates from the GG distribution with observed parameters for Athens.
2) **NonStUncor**: non-stationary and uncorrelated time series obtained as variates from the GG distribution with β that increases linearly each year by an arbitrary quantity starting from the β estimated in Athens as the initial value, while the shape parameters of the GG are constant and equal to those estimated for Athens.
3) **StCor**: stationary and correlated time series using CoSMoS with the parameters of serial correlation structure and marginal distribution estimated for Athens.

4) **NonStCorr**: nonstationary and correlated time series using CoSMoS where: (1) the scale parameter β increases linearly each year by an arbitrary quantity starting from the β estimated in Athens as the initial value, (2) the shape parameters of the GG are constant and equal to those estimated for Athens, and (3) the serial correlation structure of *X* is based on the observed parameters for Athens. Note that changes in β do not affect the other CoSMoS parameters, including those controlling the inflation of the serial correlation structure involved in the application of CoSMoS (details are given by Papalexiou, 2018).

For each time series, I computed $\{Y_j\}$ based on x^* = 95% empirical quantile, and, from these, $\{Z_i\}$. I also calculated the lag-1 serial correlation ρ_1 and the slope of the linear trend ϕ from the $\{Z_i\}$. Results of these experiments are shown in the left panel of Figure R11 via the scatterplot between ρ_1 and ϕ; the middle panel shows, for reference, the same scatterplot for the observed series; and the right panel presents instead the boxplots of the variance of the synthetic $\{Z_i\}$.

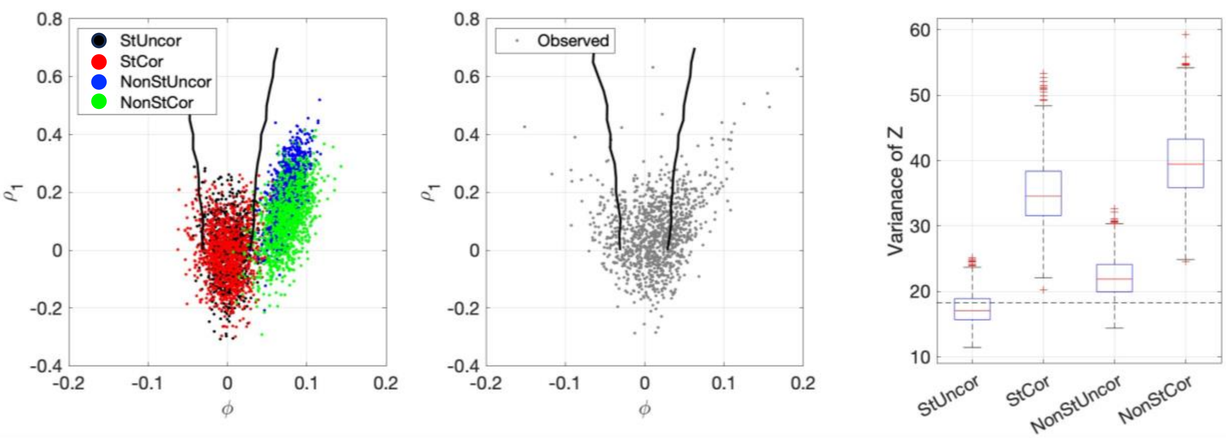

Figure R11. Left: relationship between ρ_1 and ϕ of the 1000 {Zi} time series derived by continuous daily precipitation time series generated by the CoSMoS model for the four experiments (dots with different colors) along with the 95% confidence intervals of the INAR(1) model (black lines). Middle: relationship between ρ_1 and ϕ of the observed {Zi} time series (dots), along with 95% confidence intervals of the INAR(1) model (black lines). Right: boxplots of the variance of the 1000 {Zi} time series for the four experiments.

<u>The take-home messages of Figure R11 are as follows (note that I tested stronger serial correlations for $\{X_j\}$ or different parameters of the GG and obtained the same results</u>):

1. Stationary time series of $\{X_j\}$ that are either correlated (StCor) or uncorrelated (StUncor) lead to practically the same variability of the (ρ_1, ϕ) pairs estimated on $\{Z_i\}$.
2. The variance of $\{Z_i\}$ for StUncor is similar to the mean and, thus, the distribution of *Z* is well modeled by a Poisson distribution. On the other hand, the variance of $\{Z_i\}$ for StCor is larger than the mean, implying that the sample is overdispersed and, thus, better modeled by the BB distribution, as the author pointed out. However, the confidence interval (CI) of the INAR(1) model captures very well the variability of (ρ_1, ϕ) for both StUncor and StCor.

3. The presence of non-stationarity in the time series of $\{X_j\}$ leads to higher (ρ_1, ϕ) of $\{Z_i\}$. These are outside the CI of the INAR(1) model based on the increment that was chosen for the scale parameter of the GG, and lie close to some of the observed pairs.
4. The presence of trends in $\{X_j\}$ inflates the variance of $\{Z_i\}$, although the largest factor that increases it is the presence of serial correlation.

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.

3.3. Is the BB distribution "good" for $\{Z_i\}$ generated by nonstationary uncorrelated $\{X_j\}$?

The author indicated that the BB is the theoretically correct distribution for serially correlated count series. The experiments conducted above show that trends in $\{X_j\}$ series introduce artificial serial correlations in the corresponding $\{Z_i\}$ even for uncorrelated $\{X_j\}$ (NonStUncor). In these situations, I found that the intra-cluster parameter of the BB can be "successfully" estimated based on the artificial correlations of $\{Y_j\}$. 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.

4. 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.

5. Other comments:

5.1: How were the CIs of the IAAFT model obtained in Figure 5?

5.2: Lines 302-303: How were independent peak-over-threshold events identified?

5.3: Lines 377-378: The author mentions that the INAR(1) model does not reproduce the autocorrelation of the observed data, but this was never proved.

5.4: More details are needed to explain how Figure 11 was generated. Also, the values of the estimated parameters of the BB model should be provided.

5.5: Line 539-540: is this a possible explanation of something that could not be proved? In other words, is this as “subjective” as the interpretation based on the presence of non-stationarity?

5.6: Line 594-596: I disagree with the author. Even if not perfect in terms of marginal distribution and based on some level of subjectivity on the trend (i.e., the simplest case of linearity), the calculations presented in this review confirm that the INAR(1) and NHP models, adopted by Farris et al. (2019), could be reasonably used as extreme cases to investigate where the statistics of the observed records are located compared to simple stationary and non-stationary models. Of course, better models are welcome to further improve the analyses.

Reference

Farris, S., Deidda, R., Viola, F., & Mascaro, G. (2021). On the role of serial correlation and field significance in detecting changes in extreme precipitation frequency. *Water Resources Research*, e2021WR030172. https://doi.org/10.1029/2021WR030172

Papalexiou, S. M. (2018). Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*, *115*, 234–252. https://doi.org/10.1016/J.ADVWATRES.2018.02.013

Papalexiou, S. M., & Serinaldi, F. (2020). Random Fields Simplified: Preserving Marginal Distributions, Correlations, and Intermittency, With Applications From Rainfall to Humidity. *Water Resources Research*, *56*(2), e2019WR026331. https://doi.org/10.1029/2019WR026331

Papalexiou, S. M., Serinaldi, F., & Porcu, E. (2021). Advancing Space-Time Simulation of Random Fields: From Storms to Cyclones and Beyond. *Water Resources Research*, e2020WR029466. https://doi.org/10.1029/2020WR029466

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*. International Geophysics Series, Academic Press. https://doi.org/10.1098/rspa.2007.0277