

## Reply to the comments from Francesco Serinaldi

Dear Prof. Serinaldi,

The time and attention that you dedicated to our study, and your constructive and passionate suggestions are extremely appreciated.

To adequately address and discuss the points you raised, we planned, designed, performed and are performing several additional analyses and enriched our literature review. In this reply, we would like to mention our action plan for addressing your comments and some findings we got so far, while an improved, modified and corrected manuscript will be produced in the revision phase.

You can find our reply to the points you raised below.

**F. S.:** *I would like to share some thoughts about this paper with the Authors, hoping that they can contribute to the discussion.*

*A fundamental assumption that is common to almost all methods for regional frequency analysis is spatio-temporal independence. However, the proposed procedure seems to neglect it and introduces a spurious dependence as well, I think. In fact, the sliding window procedure used to compute the sequences of L-moments acts as a moving average (... it is the same procedure used to compute e.g. drought indices such as SPI or SPEI). The resulting time series have a spurious autocorrelation with linear decay of  $1/w$  per time lag. It is known that the autocorrelation affects the estimation of cross-correlation (and vice versa). It can yield spurious cross-correlation, and variance inflation. This means that the sequences of L-moment values computed over sliding (overlapping) time windows and the “rolling mean of the considered teleconnection” might show a spurious cross-correlation. Also note that WeMOI is a low frequency climatic mode characterized by its own “natural” autocorrelation. Therefore, the autocorrelation of the rolling means of WeMOI are characterized by the superposition of two autocorrelation structures.*

*In this context, any statistical test used to check the statistical significance of cross-correlation values should account for the variance inflation affecting the distribution of the test statistics (here, the Spearman correlation; see e.g. works by Khaled H. Hamed in this respect). I may have missed something, but the text seems not to specify whether the Authors accounted for these issues. If not, I think they should be considered, as they often completely change the conclusions of these types of analysis, revealing that dependence might be a much more general and simple way to model the observed behaviour (... and “entia non sunt multiplicanda praeter necessitatem”).*

**Reply of the Authors:** Thank you for this important and relevant comment. The potential impact of spurious autocorrelation was not considered in the analysis. We have repeated our analyses by adopting the Spearman correlation coefficient with the appropriate significance test, as provided by the R library *corrTESTsrd* (Lun et al., 2023). Our preliminary results show that the number of significant correlations detected for WeMOI is no longer overly higher relative to other indexes. However, these results agree with our previous analyses, suggesting that (1) several significant correlations are still present, and (2) some spatial correlation patterns between extreme rainfall statistics and climate indexes exist. Regarding the structure of WeMOI, a mention will be inserted in the revised manuscript.

*F. S.: The Authors denote the GEV models with parameters depending on WeMOI as nonstationary. Nowadays, the term “nonstationary” is used quite arbitrarily and loosely in almost every paper; however, a model is nonstationary if and only if its form (parameters) depends on a parametric support such as time or space. WeMOI is not a parametric support; it is a process with stochastic behaviour (and periodic oscillations at 12 months, and about 20 and 50 years). Models with parameters depending on other “random” processes are not nonstationary but doubly stochastic because the parameters are themselves randomly fluctuating.*

*Please note that this is not just a semantic issue. Double stochastic models can be stationary, thus meaning that we can apply all standard results of mathematical statistics. Conversely, nonstationary models might be problematic and lead to paradoxes and misleading conclusions as they are not consistent with the ergodicity assumption, which is fundamental to establish a correspondence between sample properties and population properties, thus making inference technically possible.*

**Reply of the Authors:** Thank you for this comment. Our original manuscript uses the term “non-stationary” in a way that is similar to many publications (see examples and literature in “The legacy of STAHY” by Volpi et al., 2024). We definitely agree with you on the fact that “doubly stochastic” might be a better characterization of the models we are testing in our analyses (as also suggested in Serinaldi and Kilsby, 2018). The revised version of our manuscript will refer to the suggested terminology.

*F. S.: An example of the consequences of neglecting the importance of underlying assumptions is the interpretation of the performance metrics used to compare stationary and nonstationary models. The RLM index in eq. 6 is related to the test statistic of the likelihood ratio test*

$$\text{Chi}^2 = -2 \cdot \ln(\text{LH}_0/\text{LH}_1) = -2 \cdot \ln(\text{LH}_{st}/\text{LH}_{nst}) = 2 \cdot \ln(\text{LH}_{nst}/\text{LH}_{st}) = 2 \cdot \text{RLM},$$

which is asymptotically distributed as a  $\text{Chi}^2$  variable with degrees of freedom equal to the difference of the number of parameters of the two models. Since the Authors use second-order polynomials for  $\mu$  and/or  $CV$ , the degrees of freedom are 2 (=5-3) for  $\text{GEV}_{1/2}$  vs  $\text{GEV}_0$  or 4(=7-3) for  $\text{GEV}_3$  vs  $\text{GEV}_0$ . Therefore, the 95th quantiles of the two  $\text{Chi}^2$  distributions are about 6 and 9.5, corresponding to critical levels of RLM equal to 3 and 4.25. If we compare these values with those reported in Fig. 6a, we see that most of the RML values are lower than those upper limits. Leaving aside the validity of the asymptotic  $\text{Chi}^2$  for finite samples, the message is that relative measures/metrics should not be interpreted according to their absolute values but should be assessed accounting for their own sampling variability. Indeed, this interpretation reconciles results of Fig. 6a with those in Fig. 6c.

**Reply of the Authors:** We believe that the discussion you raised is very important, and we are sincerely thankful for his comment. We totally agree with the fact that the test statistics should not be considered alone, but with their associated significance value. However, we are not totally persuaded that the reference value of the RML metric you suggested should be used in our analysis. This is mainly due to two reasons: (1) the size of our samples is limited and (2) the nature of the parameters of the polynomial function is different from the other GEV parameters. To share some light on this aspect, we are performing Monte Carlo resampling experiments that are specifically designed for assessing the suitability of the suggested reference value for RML. Our preliminary results seem to indicate that a lower reference value should be used for identifying statistically significant improvements over a stationary model (similarly to what is presented in Ashkar and Aucoin, 2012). Please, see also our replies to the following points on these aspects.

**F. S.:** *In this respect, please note that TN.SW is the only theoretically consistent goodness-of-fit metrics used here, whereas AD is not.*

*Let me explain. When we deal with (supposed) nonstationary models we can only apply diagnostics to standardized versions of the data, conditional on the fitted parameters (see e.g., Coles 2001; An Introduction to Statistical Modeling of Extreme Values). The expression of AD statistic holds true if and only if the data are identically distributed (i.e. the model is stationary) because such a goodness-of-fit test is based on the distance between the parametric model and the empirical cumulative distribution function (ECDF). Now, the values of the ECDF,  $F_n(x)$ , corresponding to each observation are representative of theoretical probabilities if and only if these observations come from the same distribution*

$F(x)$ , thus meaning that larger observations (order statistics) correspond to larger non-exceedance probabilities.

However, suppose that our model is nonstationary; for example, GEV location parameter increases with time because the observations seem to assume greater values through the years. In this case, each observation comes from a different distribution (which is only valid at a specific time... or WeMOI value), and largest values are likely associated to GEV distributions with large location parameters (as the time-varying GEV model attempts to follow the fluctuations of the observations). Therefore, given a sample of size  $n$ , the largest observation, for instance, has non-exceedance probability  $1/(n+1)$  according to the ECDF (... leaving aside plotting position issues), whereas it might have the same probability of a smaller observation under its own local GEV. In other words, the correspondence between empirical and theoretical probabilities has been lost. Therefore,  $\Delta_{AD}$  in Fig. 6a is positive for two reasons:

- $AD_{nst}$  is likely greater than  $AD_{st}$  because the ECDF appearing in the AD formulas refer to (overlooked) stationary assumption.
- The AD statistic is a positive distance that indicates better performance when it is small (but not too small, as it would mean that the  $F(x_i)$  sample is too regular to be a random uniform sample); if  $\Delta_{AD} = AD_{nst} - AD_{st}$  is positive, it means that  $AD_{nst}$  is larger than  $AD_{st}$ , and therefore GEV\_0 is better. In fact, this is the rule used by Ashkar and Ba (2017): “The decision rule for the sample is to choose GP if  $a_{GP} - a_{KAP} < 0$ ”.

That said, as for RLM, the actual question is whether values of  $\Delta_{AD}$  equal to 0.004-0.006 are just within the expected fluctuations for nested models, once we account for factors like dependence and the variance inflation of AD statistics related to the fact that the model parameters are estimated on the same data used to compute the AD statistic itself. Since estimated parameters make AD statistic no longer distribution-free, comparing AD values of different models is also questionable (because identical AD values for two models can correspond to different probabilities in the sampling distributions of AD under these alternative models).

**Reply of the Authors:** Many thanks for this useful and relevant comment. Based on these considerations we decided not to consider the AD metrics in the revised version of our study.

**F. S.:** TN.SW values confirm the message of the other two metrics: the data are not enough to discriminate among GEV\_0, GEV\_1, etc. However, while RLM and AD and their Delta's are expected to be positive but possibly small, SW is centered around zero by construction

(if a model is good enough). Therefore,  $\Delta_{TN.SW}$  is expected to be centered around zero when discriminating between models is not possible. In this case, the comparison is technically sound because TN.SW provides the kind of conditional standardization mentioned above. Thus, results in Fig. 6c are consistent with those in Figs. 6a and 6b, and the interpretation of Fig. 6 provided in section 5.2 should be revised accordingly, I think.

**Reply of the Authors:** We understand the reasons why the TN.SW metric may be considered theoretically consistent with the framework adopted in our study. However, our additional analyses, inspired by your comments, led us to question its true validity and informativeness for our specific study.

In particular, the modified Shapiro-Wilk test consists of two steps (see e.g., Ashkar and Aucoin, 2012): (1) transformation of the AMS into normally distributed samples  $Z_i = FDC_n^{-1}(FDC_{GEV}(x_i))$ ; (2) application of the Shapiro-Wilk test to the obtained sample. We are concerned about the  $FDC_{GEV}$  used in step (1), which is unique for the stationary case, while should be regarded as  $n$  FDCs (i.e., where  $n$  is the length of the time series) in the doubly stochastic case. In order to better understand the nature of these inconsistencies, we are performing a series of Monte Carlo experiments which are designed for assessing the power of these two metrics.

**F. S.:** *A note about the use of return period: return period is the expected value of return intervals, which implies integration over time (by definition). Under nonstationarity, the integral does not yield  $1/(1-F(x))$  because a unique  $F(x)$  does not exist! And replacing it with a set of  $T_i = 1/(1-F_i(x))$  formulas makes no sense. Roughly speaking, under nonstationarity, integrating over time implies averaging over a set of  $F_i(x)$  distributions, and the resulting expectation is a formula reflecting (and function of) all  $F_i$ 's. While I understand the (fallacious but widespread) rationale of drawing a set of return level curves (as those in Fig. 5c-e), the formula  $1/(1-F_i(x))$  does not correspond to any expected value (over time). Under nonstationarity, the return level curve is as unique as in the case of stationarity because the expected value of the (inter)arrival times of exceedances over a specified threshold is always a single value resulting from integration. However, what changes is the expression linking  $T$  with the set of  $F_i$ 's. Even though the diagrams of  $T_i = 1/(1-F_i(x))$  vs  $x$  are reported in almost every paper dealing with nonstationary distributions, this does not make them and the relationships  $T_i = 1/(1-F_i(x))$  meaningful. So, please consider avoiding further spreading such a misconception.*

**Reply of the Authors:** Thank you for raising this point. Indeed, we agree with you about the theoretical nonexistence and incorrectness of a single  $T_i = 1/(1-F_i(x))$ , and yet we disagree with you on the argued meaninglessness of the set of curves. Quite the opposite, we believe that this figure is very informative as it clearly shows

the variability in terms frequency regime of extreme rainfall events that is associated with different climatic conditions, described in our case by the value of WeMOI averaged in the last 30 years. This is very useful from a practical viewpoint in engineering practice for defining possible meaningful future climate scenarios. Nevertheless, we duly noted the theoretical limitations of such a representation, which will be clearly mentioned in the revised version of our manuscript.

**F. S.:** *Finally, the Authors state that “the spatial aggregation into tiles allows to obtain more reliable values of the rainfall statistics”. This is strictly true if the time series within a tile are independent; otherwise, spatial dependence implies information redundancy, meaning that the apparent smoothness comes with uncertainty much larger than one can think, and such averaged/aggregated statistics might be not so reliable. Please note that I do not refer to the spatial dependence of AMAX: these can look approximately uncorrelated (in space and time) even when the underlying processes are strongly dependent. In these cases, clustering in space and time might be an indicator of the underlying (concealed) spatio-temporal dependence. These remarks apply to any type of analysis, including for instance record-breaking observations. In fact, “significant deviations in the number of record-breaking events in an observed time series relative to what is expected under the iid hypothesis indicate non-stationary time series” (Castellarin et al. 2024; <https://doi.org/10.3390/atmos15070865>)... or dependence, I would say! If “I.I.D.” still means independent and identically distributed, discrepancies can come from lack of fulfilment of one of these two assumptions or both, and there is no reason to exclude the former. Based on my experience, people often tend to neglect dependence because adding a few covariates to GLM-like models with an arbitrary polynomial/spline link is a bit easier than deriving the theoretical relationships accounting for dependence.*

*Overall, in my opinion, any method that implies spatio-temporal aggregation, smoothing, and averaging of hydro-climatic data should carefully account for spatio-temporal dependence, as this assumption allows one to keep models simple and parsimonious, it is generally sufficient to describe the behaviour of most of the observed processes, and often reveals that we are overconfident about the reliability of results and the amount of information (effective sample size) actually available.*

**Reply of the Authors:** We are aware that spatial correlation among the timeseries of annual maxima increases the uncertainty of regional estimations, however, it does not introduce bias (see e.g., Hosking and Wallis, 1988). Thus, we believe this aspect has a very limited impact on our analyses, since we adopt spatial discretization only to obtain estimations of the higher order L-moments (and consequently, of the quantiles) that are locally more robust. On the fact that

regional frequency analysis should be preferred to local frequency analysis, the classical literature is clear (see refs. above). Nevertheless, we agree with you that this point is very important and deserves to be discussed in the revised version of the paper. Finally, we designed and are planning on performing an additional set of Monte Carlo simulation experiments for assessing the field significance of the spatial patterns resulted from our study.

*F. S.: However, if we decide to use nonstationary models, we must bear in mind (i) what nonstationarity technically means and implies, (ii) that most of the methods available under stationarity are no longer valid, and (iii) the inference procedure itself along with the interpretation of results are problematic because of lack of ergodicity. The foregoing discussion just mentions some concepts that cannot be transferred when we move from stationarity to nonstationarity. Deeper discussion of these and other issues can easily be found in the literature... some of such a literature (concerning the impact of spatio-temporal dependence on frequency analysis) is from one of the Authors.*

**Reply of the Authors:** We are particularly glad that the topics discussed in our manuscript inspired such a rich, useful and interesting comment. However, we believe that some of these points are not particularly relevant to our analyses. We agree that the utilization of nonstationary (or doubly stochastic) approaches implies the redefinition (or the careful selection) of some methods often adopted with stationary analysis (e.g., the goodness-of-fit metrics, as suggested here or trend tests, as showed in Serinaldi, 2024). However, recent literature urges for approaches and methods that can better capture the effects of climate variability on the frequency of hydrological extremes (e.g., Volpi et al., 2024, and all the references contained there). Hence, the identification of reliable and informative nonstationary frequency models does not seem to be a matter of choice anymore, but rather an open research avenue.

That said, our study does not aim at proposing specific nonstationary models, but rather at investigating the regional characteristics of the link between teleconnection and extreme rainfall regime. What is the best way to implement doubly stochastic RFA models and how to correctly use them for statistical inference are definitely key topics, but they need to be addressed by future studies.

Overall, we totally agree with you that some important points should be discussed, and some elements need to be improved or corrected. As mentioned above, we have been intensively working to refine and complete our analyses. More specifically, these additional investigations aim at (1) addressing autocorrelation issues when computing teleconnections-rainfall correlation, (2) understanding the validity of our metrics, and (3)

assessing the effect of spatial dependence. Our preliminary results seem to confirm most of our previous findings. Thus, we believe that our research questions have universal interest, our methodology is valid (and now improved thanks to your useful comments), and our results are coherent with our current knowledge about the frequency regime of extreme precipitation over the study area.

Again, we express our gratitude for your suggestions. And we look forward to submitting a new version of our manuscript with refined analyses and discussion.

Best regards,

The Authors

## References

Ashkar, F., Aucoin, F., 2012. Choice between competitive pairs of frequency models for use in hydrology: a review and some new results. *Hydrological Sciences Journal* 57, 1092–1106. <https://doi.org/10.1080/02626667.2012.701746>

Hosking, J.R.M., Wallis, J.R., 1988. The effect of intersite dependence on regional flood frequency analysis. *Water Resources Research* 24, 588–600. <https://doi.org/10.1029/WR024i004p00588>

Lun, D., Fischer, S., Viglione, A., Blöschl, G., 2023. Significance testing of rank cross-correlations between autocorrelated time series with short-range dependence. *Journal of Applied Statistics* 50, 2934–2950. <https://doi.org/10.1080/02664763.2022.2137115>

Serinaldi, F., Kilsby, C.G., Lombardo, F., 2018. Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. *Advances in Water Resources* 111, 132–155. <https://doi.org/10.1016/j.advwatres.2017.10.015>

Serinaldi, F., 2024. Scientific logic and spatio-temporal dependence in analyzing extreme-precipitation frequency: negligible or neglected? *Hydrol. Earth Syst. Sci.* 28, 3191–3218. <https://doi.org/10.5194/hess-28-3191-2024>

Volpi, E., Grimaldi, S., Aghakouchak, A., Castellarin, A., Chebana, F., Papalexiou, S.M., Aksoy, H., Bárdossy, A., Cancelliere, A., Chen, Y., Deidda, R., Haberlandt, U., Eris, E., Fischer, S., Francés, F., Kavetski, D., Rodding Kjeldsen, T., Kochanek, K., Langousis, A., Mediero Orduña, L., Montanari, A., Nerantzaki, S.D., Ouarda, T.B.M.J.,

Prosdocimi, I., Ragno, E., Rajulapati, C.R., Requena, A.I., Ridolfi, E., Sadegh, M., Schumann, A., Sharma, A., 2024. The legacy of STAHY: milestones, achievements, challenges, and open problems in statistical hydrology. *Hydrological Sciences Journal* 69, 1913–1949. <https://doi.org/10.1080/02626667.2024.2385686>