Authors' response to comments from Referee #3

The authors would like to thank also the third reviewer for the kind attention and the significative and constructive observations and suggestions.
As a first general remark he/she suggests a careful proofreading for English language which will be certainly done on the revised version of the manuscript from a professional native English speaker.
Specific comments will be addressed in the following lines according to the same numbering provided by the reviewer:
In what follows in *Italic* are the comments provided by the Referee, and in bold **fonts** the authors' response. Changes to the manuscript are quoted and reported in ***"bold italic"*** font. Please consider that text here reported is still undergoing for the check from a professional native English speaker.

1) Lines 113-116. Description of Sen slope estimation and equation (2) should be revised: If N is the number of univocal (non-repeated) couples and j is an index for the j-th couple (x_i, x_k), why should be j > k? Maybe the authors mean i > k? Please check and better specify the role of j index. Remove also "Sorting in ascending order ....", declaring that the median values is the final estimate is enough and the reader understand.
Suggestion accepted.

2) Lines 179-188 + Appendix. These lines + Appendix should be removed. All the analyses in the manuscript are based on ML estimates, thus there is no reason to keep a description of PWM and L-moments.
Suggestion accepted, the appendix and the expressions of L-moments were removed and replaced by theoretical expressions of moments as from first comment from Referee #1.

3) *Section 2.5. It should be written that stationarity is assumed as null hypothesis (e.g. in line 207 and 210)*
Suggestion accepted, we modified lines 205 and 211 by adding, respectively, that "the null hypothesis of stationarity is false" and "the null hypothesis of stationarity (which) is true".

4) Section 2.5, lines 201 and 209. Just a curiosity: why experiments are conducted with a different number of samples (2000 vs 10000)?
We used N = 10000 generations to be sure that the effect of sample variability is negligible when estimating the real level of significance and the corresponding threshold of acceptance. For the evaluation of power (i.e. in all generations with values of $\zeta_1 \neq 0$) we found N = 2000 is a good compromise between quality of results and reasonable computational time. To this purpose we performed sample checks with N = 10000 obtaining negligible differences. We did not report such details for the sake of simplicity and also because power values obtained with generations with $\zeta_1 \neq 0$ are only shown in graphical form and as such they are not even distinguishable from those shown. Moreover, N = 2000 is the same number of generations

used by Yue et al. (2002a). In the revised manuscript we added this short explanation about this point in section 2.5:

"*We used a reduced number of generations (N = 2000) for the evaluation of power as good compromise between quality of results and computational time and, also in analogy with Yue et al. (2002a).*"

Regarding the number of generations N, see also answers to points 7), 8) and 14).

5) Section 2.5, line 198. Authors can provide a reference for the sentence "1-to-4 trade-off between α and β is accepted"?

Such question is everything but trivial, we can provide a reference about this (Cohen, 1994) and we introduced it in the revised manuscript, nevertheless such paper does not come from hydrological literature but from Psychology and is by far the most cited paper in Scopus about "statistical power". Reflecting about this point has stimulated a discussion about the apparent lack of references of this type in the earth system sciences that we have added in the conclusion section. The following lines were added in the conclusion section:

**"*Considering the feasibility of numerical evaluation of power, allowed by the parametric approach, we observe that, while the awareness of the crucial role of type II error is growing in latest years in the hydrological literature, a common debate would deserve more development about which power values should be considered acceptable. Such an issue is much more enhanced in other scientific fields where the experimental design is traditionally required to estimate the appropriate sample size to adequately support results and conclusions. In psychological research, Cohen (1994) proposed 0.8 as a conventional value of power, to be used with level of significance 0.05 thus leading to a ratio 4 to 1 between the risk of type II and type 1 error. The conventional value proposed by Cohen (1994) has been taken as reference by thousands of papers in social and behavioural sciences. In pharmacological and medical research, depending on real implications and nature of the type II error, conventional values of power may be as high as 0.999. This is the value suggested by Liebher (1990), when testing a treatment for patients' blood pressure. He stated, while "guarding against cookbook application of statistical methods", that "it should also be noted that, at times, type II error may be more important to an investigator then type I error".*

We believe that, selecting between stationarity and non-stationarity models for extreme hydrological event prediction, a fair comparison between the null and the alternative hypotheses as $\alpha = \beta = 0.05$ should be taken, which provides power = 0.95. In our discussion we considered 0.8 as minimum threshold for acceptable power values."

6) Lines 214-219. I understand the choice of GEV parameters and it is reasonable to my knowledge of rainfall maxima in Mediterranean climate. I was wandering whether it can be more informative to present results and figures in a more general way, e.g. as a function of the relative trend ζ_1/ζ_0 (which has dimension 1/time)?

This comment is particularly important and stimulating. Nevertheless, such generalization is out of the purpose of this paper and would require more extensive investigation. In facts, presenting results in terms of the ratio ζ_1/ζ_0 would make sense if results of the analyses were the same for different couples of ζ_1 and ζ_0 values producing the same ratio. Actually, this is not the case when σ is fixed. We believe that invariant properties of the frequency distribution of rainfall or floods annual maximum values could be exploited for a generalization of these

analyses, but this would involve consideration of scaling features of different order moments. This could be a quite interesting future development of this study, involving also time dependence of the scale parameter. Not any change has been done to the manuscript.

7) Line 236. "a multi-peak . . .": are you sure that it is not a sampling effect? Increasing the number of simulations is the result the same?

In this figure we show some distributions of sampled errors in terms of difference between the power obtained with AIC and the power obtained with AIC_c. The curves show that the entire range of sampled errors provides negligible values compared to the expected power values. The different peaks in one curve (L = 30) are observed because we merged sample errors obtained for different values of σ characterized by a small different (random) bias. We added such considerations in the revised paper:

"*Aim of this figure is to show that the difference between the power obtained with AIC and the power obtained with AICc is negligible. Anyway, different peaks in one curve (L = 30) can be explained by merging sample errors obtained for different values of σ.*"

8) Lines 267-273 and Table 1. Similar to previous comment: are you sure that variability observed for different σ (but keeping constant the other constraints) are not a sampling effect? Increasing the number of simulations is the result the same?

Results shown in Table 1 are, to some extent, affected by a sampling effect. We had numerically checked the sample variability of the actual level of significance, which is quite smaller than the difference between the designed (0.05) and the actual value of the significance level for the LR test. We didn't report such analysis for the sake of simplicity nevertheless we used a very high number of generations (N = 10000) to produce these values (see also response to comment #4). On the other hand, we agree that there is not such evidence with specific reference to different σ values, while ε and L mostly affect results and accordingly we revised the manuscript rephrasing the sentence as:

"*Such effect is exalted when the parent distribution is upper bounded (ε=-0.4) and for shorter series (L = 30).*"

9) Line 301. I would specify here that series are stationary.

Suggestion accepted

10) *Section 3.4 (and maybe other parts of the manuscript, figures and tables). GEV parameters are always estimated by ML, thus I suggest to avoid the use of the prefix "ML-" before the symbol of the parameter (e.g. in Line 310 and 313). This would made more clear text, figure and tables.*

We would rather maintain the use of the prefix "ML-" in order to distinguish between values (ML-epsilon, ML-sigma, ML-z1) estimated from the series and the theoretical values (EPS, sigma, z1) used in the parent distribution. Not any change was done.

11) Line 316. "Figs. 6 and 9": usually figures should be ordered as they are cited.
Suggestion accepted by revising the text by removing such reference to fig. 9 which is introduced later.

12) Figures. Please consider the opportunity to use larger fonts for labels, they are not readable here, and most probably Figure will be reduced in the final formatting.
Suggestion accepted

13) Figure 6. Use the same range of scales (e.g. 0-0.6) in the y-axis for a fair comparison.
Suggestion accepted

14) Figure 9. Again: are you sure that fluctuations are not due to sampling effects? See e.g. the subplot in the right part of the Figure 9.
We checked such results with different sets of random generations also increasing N up to 10000. Qualitatively results do not change. "*Randomness of results for* $L = 30$ *and* $\sigma = 15, 20$ *is probably due to a reduced efficiency of the algorithm that maximizes the log-Likelihood function, for heavy tailed distributions.*" Such consideration was added in the revised manuscript (Fig. 9 is now Fig. 8, because of the following comment).

15) Figures 7, 8, 10, 11, 12, 13 are not much informative. Please show only a selection of the most representative case. An additional option is to move these figures as supplementary material.
Suggestion accepted. Results shown in figs. 7-8-10-11-12-13 are now shown for representative selected cases in figures 7-9-10.