Numerical investigation on the power of parametric and non-parametric tests for trend detection in annual maximum series

Vincenzo Totaro, Andrea Gioia, Vito Iacobellis

Dipartimento di Ingegneria Civile, Ambientale, del Territorio, Edile e di Chimica (DICATECh), Politecnico di Bari, Bari, 70125, Italy

Correspondence to: Vincenzo Totaro (vincenzo.totaro@poliba.it)

Abstract. The need of fitting time series characterized by the presence of trend or change points has generated in latest years an increased interest in the investigation of non-stationary probability distributions. Considering that the available hydrological time series can be recognized as the observable part of a stochastic process with a definite probability distribution, two main topics can be tackled in this context: the first one is related to the definition of an objective criterion for choosing whether the stationary hypothesis can be adopted, while the second one regards the effects of non-stationarity on the estimation of distribution parameters and quantiles for assigned return period and flood risk evaluation. Although the time series trend or change points are usually detected by non-parametric tests available in literature (e.g. Mann-Kendal or CUSUM test), the correct selection of the stationary or non-stationary probability distribution is still required for design purposes. By this light, the focus is shifted toward model selection criteria which implies the use of parametric methods with all related issues in parameters estimation. The aim of this study is to compare the performance of parametric and non-parametric methods for trend detection analysing their power and focusing on the use of traditional model selection tools (e.g. Akaike Information Criterion and Likelihood Ratio test) within this context. Power and efficiency of parameter estimation, including the trend coefficient, were investigated through Monte Carlo simulations using Generalized Extreme Value distribution as parent with selected parameter sets.

1 Introduction

Long and medium-term prediction of extreme hydrological events under non-stationary conditions, is one of the major challenges of our times. Streamflow, as well as temporal rainfall and many other hydrological phenomena, can be considered stochastic processes (Chow, 1964), i.e. families of random variables with an assigned probability distribution, while time series are the observable part of this process. One of the main goals of the extreme events frequency analysis is the estimation of distribution quantiles related to a certain non-exceedance probability. They are usually obtained after fitting a probabilistic model to observed data. Kolmogorov in 1931 introduced the concept of stationarity of a probability distribution, formally defined by Khintchine in 1934, as depicted in the historical review provided by Koutsoyiannis and Montanari (2015).

Quae cum ita sint, detecting the existence of time-dependence in a stochastic process should be considered a necessary task in

30   the statistical analysis of recorded time series. Starting from this statement several considerations can be done in updating some important hydrological concepts while assuming that non-exceedance probability varies with times or other covariates. For example, return period may be reformulated in two different ways, Expected Waiting Time (EWT, Olsen et al., 1998) or Expected Number of Events (ENE, Parey et al., 2007, 2010) which lead to a different evaluation of quantiles within a non-stationary approach. As proved by Cooley (2013), EWT and ENE are differently affected by non-stationarity, possibly

35   producing ambiguity in engineering design practice (Du et al., 2015; Read and Vogel, 2015). Salas and Obeysekera (2014) provided a detailed report about relationships between stationary and non-stationary EWT values within a parametric approach for the assessment of non-stationary conditions. In such a framework, a strong relevance is given to statistical tools for detecting changes in non-normally distributed time series (Kundewicz and Robson, 2004).

So far, the vast majority of research about climate change and detection of non-stationary conditions has been developed

40   through non-parametric approaches. One of the most used non-parametric measures of trend is Sen's slope (Gocic and Trajkovic, 2013). Also, a wide gamma of non-parametric tests for detecting non-stationarity is available (e.g. Kundewicz and Robson, 2004). Statistical tests include Mann-Kendall (*MK;* Mann, 1945; Kendall, 1975) and Spearman (Lehmann, 1975) for detecting trends, Pettitt (Pettitt, 1979) and CUSUM (Smadi and Zghoul, 2006) for change point detection. All of these tests are based on a specific null hypothesis and have to be performed for an assigned significance level. Non-parametric tests are

45   usually preferred to parametric ones because they are distribution-free and do not require knowledge of the parent distribution. In the frequency analysis of extreme events they are also suggested being less sensitive to the presence of outliers with respect to parametric tests (Wang et al., 2005).

On the other hand, the use of null hypothesis significance tests for trend detection has raised concerns and severe criticisms in a wide range of scientific fields since many years (e.g. Cohen, 1994), as outlined by Vogel et al. (2013). Serinaldi et al. (2018)

50   provided an extensive critical review focusing on logical flaws and misinterpretations often related to their misuse.

In general, the use of statistical tests involves different errors, such as type I (rejecting the null hypothesis when it is true) and type II (accepting the null hypothesis when it is false). The latter is related to the test power, i.e. the probability of rejecting the null hypothesis when it is false but, as recognized by a few authors (e.g. Milly et al. 2015; Beven, 2016), in years the importance of power has been largely overlooked in earth system science fields. A strong attention has always been given to

55   the level of significance (i.e. type I error) though, as pointed out by Vogel et al. (2013), "*a type II error in the context of an infrastructure decision implies under-preparedness, which is often an error much more costly to society than the type I error (over-preparedness)*".

Moreover, as already proven by Yue et al. (2002a), the power of the Mann-Kendall test, despite its non-parametric structure, actually shows a strong dependence on the type and parametrization of the parent distribution.

60   In a parametric approach, the estimation of quantiles of the extreme events distribution requires the search for the underlying distribution and for time-dependant hydrological variables, providing the identification of a model, which can be stationary or not (Montanari and Koutsoyiannis, 2014). In other words, it is necessary to define if variables are iid (independent identically

distributed) or i/nid (independent/non identically distributed) and, accordingly, to select between a stationary or not-stationary distribution model (Serinaldi and Kilsby, 2015).

65   In this perspective, the detection of non-stationarity may exploit, besides traditional statistical tests, well known properties of model selection tools. Even in this case several measures and criteria are available for selecting a best-fit model, among these we find Akaike Information Criterion (AIC, Akaike, 1974), Bayesian Information Criterion (BIC, Schwarz, 1978) and Likelihood Ratio test (LR, Coles, 2001), the latter is suitable when dealing with nested models.

The purpose of this paper is to provide further insights on the use of parametric and non-parametric approaches in the
70   framework of extreme events frequency analysis under non-stationary conditions. The comparison between those different approaches is not straightforward. Non-parametric tests do not require the knowledge of the parent distribution and their properties strongly rely on the choice of a null hypothesis. Parametric methods for model selection, on the other hand, require the selection of the parent distribution and the estimation of its parameters, but are not necessarily associated with a specific null hypothesis. Nevertheless, in both cases the evaluation of the rejection threshold is usually based on a statistical measure
75   of trend that, under the null hypothesis of stationarity, follows a specific distribution (e.g. gaussianity of the Kendall statistic for the MK non-parametric test; χ^2 distribution of deviance statistic for the LR parametric test).

Considering *pros* and *cons* of different approaches, we believe that specific remarks should be made about the use of parametric or non-parametric methods for the analysis of extreme event series. For this purpose, we set up a numerical experiment to compare performances of: 1 the MK as a non-parametric test for trend detection, 2 the LR parametric test for model selection,
80   3 the AIC_R parametric test as defined in section 2.3. In particular, the AIC_R is a measure for model selection, based on the AIC, whose distribution was numerically evaluated, under the null hypothesis of a stationary process, for comparison purposes with other tests.

We aim to provide (i) a comparison of test power between MK, LR and AIC_R, (ii) a sensitivity analysis of test power to parameters of a known parent distribution used to generate sample data, (iii) an analysis of the influence of sample size on test
85   power and significance level.

We conducted the analysis using Monte Carlo techniques, by generating samples from parent populations assuming one of the most popular extreme event distributions, the Generalized Extreme Value (Jenkinson, 1955), with linear (and without any) trend in the position parameter. From generated samples we numerically evaluated the power and significance level of tests for trend detection, using MK, LR and AIC_R. For the latter we also checked the option of using the modified version AIC_c,
90   suggested by Sugiura (1978) for smaller samples.

Considering that parametric methods involve the estimation of the parent distribution parameters, we also analysed the efficiency of the Maximum Likelihood (ML) estimator used in trend assessment by comparing the sample variability of the ML estimate of trend with the non-parametric Sen's slope. We also scoped the sample variability of GEV parameters in the stationary and non-stationary cases.

2 Methodological framework

This section is divided into five parts. Subsections 2.1, 2.2 and 2.3 report main characteristics of respectively, *MK*, *LR* and AIC_R based test. In the fourth subsection the probabilistic model used for generations, based on the use of the GEV distribution, is described in the stationary and non-stationary cases. Subsection 2.5 outlines the procedure for numerical evaluation of tests' power and significance level.

2.1 The Mann-Kendall test

Hydrological time series are often composed by non-normally independent realizations of phenomena, and this characteristic makes the use of non-parametric trend tests very attractive (Kundzewicz and Robson, 2004). Mann-Kendall test is a widely used rank-based tool for detecting monotonic, and not necessarily linear, trends. Given a random variable z, and assigned a sample of L independent data $\mathbf{z} = (z_1, ..., z_L)$, the Kendall S statistic (Kendall, 1975) can be defined as:

$$S = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} sgn(z_j - z_i), \tag{1}$$

with *sgn* sign function.

The null hypothesis of this test is the absence of any statistically significant trend in the sample, while it is contemplated by the alternative hypothesis. Yilmaz and Perera (2014) reported that serial dependence can lead to a more frequent rejection of null hypothesis. For $L \geq 8$, Mann (1945) reported how Eq. (1) is approximatively a normal variable with zero mean and variance that, in the presence of t_m *m*-length ties, can be expressed as:

$$V = \frac{L(L-1)(2L+5) - \sum_{m=1}^{n} t_m m(m-1)(2m+5)}{18}.$$

In practice, Mann-Kendall test is performed using the *Z* statistic

$$Z = \begin{cases} \frac{S-1}{\sqrt{V(S)}} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sqrt{V(S)}} & S < 0 \end{cases},$$

which follows a standard normal distribution. With this approach, it is simple to evaluate the p-value and compare it with an assigned level of significance or, equivalently, to calculate the Z_α threshold value to be compared with Z, where Z_α is the $(1 - \alpha)$ quantile of a standard normal distribution.

Yue et al. (2002b) observed that autocorrelation in time series can influence the ability of *MK* test to detect trends. To avoid this problem, a correct approach in trend analysis should contemplate a preliminary check for autocorrelation and, if necessary, the application of pre-whitening procedures.

A non-parametric tool for a reliable estimation of a trend in a time series of length *L* and with *N* pairs of data is the Sen's slope estimator (Sen, 1968):

$$\delta_j = \frac{z_i - z_k}{i - k}, \qquad j = 1, \dots, N \tag{2}$$

being $i > k$. Then, Sen's slope estimator can be defined as the median of δ_j.

2.2 Likelihood Ratio Test

125 The Likelihood Ratio statistical test allows to compare two candidate models. Like its name suggests, it is based on the evaluation of the likelihood function of different models.

The LR test has been used multiple times (Tramblay et al., 2013; Cheng et al., 2014; Yilmaz et al., 2014) to select between stationary and non-stationary models in the context of nested models. Given a stationary model characterized by a parameter set $\boldsymbol{\theta}_{st}$ and a non-stationary model, with parameter set $\boldsymbol{\theta}_{ns}$, if $\ell(\widehat{\boldsymbol{\theta}}_{st})$ and $\ell(\widehat{\boldsymbol{\theta}}_{ns})$ are their respective maximized log-

130 likelihoods, the Likelihood Ratio test can be defined through the deviance statistic

$$D = 2\big[\ell(\widehat{\boldsymbol{\theta}}_{ns}) - \ell(\widehat{\boldsymbol{\theta}}_{st})\big], \tag{3}$$

D is approximately, for large L, χ_m^2 distributed, with $m = dim(\boldsymbol{\theta}_{ns}) - dim(\boldsymbol{\theta}_{st})$ degrees of freedom. The null hypothesis of stationarity is rejected if $D > C_\alpha$, where C_α is the $(1 - \alpha)$ quantile of χ_m^2 distribution (Coles, 2001).

Besides the analysis of power, we also checked (in subsection 3.3) the approximation $D \sim \chi_m^2$ as a function of the sample size

135 L for the evaluation of the level of significance.

2.3 Akaike Information Criterion Ratio test

Information criteria are useful tools for model selection. It is reasonable to retain that Akaike Information Criterion (*AIC*; Akaike, 1974) is the most famous among them. Based on the Kullbach-Leibler discrepancy measure, if $\boldsymbol{\theta}$ is the parameter set of a k-dimensional model ($k = dim(\boldsymbol{\theta})$), AIC is defined as:

140 $$AIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2k. \tag{4}$$

The model that best fits data has the lowest value of *AIC* between candidates. It is useful to observe that the term proportional to the number of model parameters allows to account for the increase of the estimator variance as the number of model parameters increases.

Sugiura (1978) observed that *AIC* can lead to misleading results for small samples; he proposed a new measure for AIC:

145 $$AIC_c = -2\ell(\hat{\theta}) + \frac{2k(k+1)}{L-k-1} \tag{5}$$

where a second-order bias correction is introduced. Burnham e Anderson (2004) suggested to use this version only when $L/k_{max} < 40$, being k_{max} the maximum number of parameters between the compared models. However, for larger L, AIC_c converges to AIC. For a quantitative comparison between *AIC* and AIC_c in the extreme value stationary model selection framework see also Laio et al. (2009).

In order to select between stationary and nonstationary candidate models, we use the ratio

$$AIC_R = \frac{AIC_{ns}}{AIC_{st}}. \tag{6}$$

where the subscripts indicate the AIC value obtained for a stationary (st) and a non-stationary (ns) model, both fitted with maximum likelihood to the same data series.

Considering that the better fitting model has lower AIC, if the time series arises from a non-stationary process, the AIC_R should be less than 1. Vice versa if the process is stationary.

In order to provide a rigorous comparison between the use of MK, LR and AIC_R, we evaluated the $AIC_{R,\alpha}$ threshold value corresponding to significance level α by numerical experiments.

More in detail, we adopted the following procedure:

1. N = 10000 samples are generated from a stationary GEV parent distribution, with known parameters;
2. for each of these samples the AIC_R is evaluated, by fitting the stationary and non-stationary GEV models described in section 2.4, thus providing its empirical distribution (see pdf in fig. 1);
3. exploiting the empirical distribution of AIC_R the threshold associated with a significance level of $\alpha = 0.05$ is numerically evaluated: this value, $AIC_{R,\alpha}$, represents the threshold for rejecting the null hypothesis of stationarity (which in these generations is true) in 5% of the synthetic samples.

This procedure was applied both for AIC and AIC_c. The experiment was repeated for a few selected sets of the GEV parameters, including different trend values, and different sample lengths, as detailed in section 3.

2.4 The GEV parent distribution

The cumulative Generalized Extreme Value (GEV) distribution (Jenkinson, 1955) can be expressed as:

$$F(z, \boldsymbol{\theta}_{st}) = \begin{cases} exp\left\{-\left[1 + \varepsilon\left(\frac{z-\zeta}{\sigma}\right)\right]^{-1/\varepsilon}\right\} & \varepsilon \neq 0 \\ exp\left\{-exp\left[-\left(\frac{z-\zeta}{\sigma}\right)\right]\right\} & \varepsilon = 0 \end{cases} \quad \sigma > 0 \tag{7}$$

where $\zeta, \sigma, \varepsilon$ are respectively known as the position, scale and shape parameter, $\boldsymbol{\theta}_{st} = [\zeta, \sigma, \varepsilon]$, is a general and comprehensive way to express the parameter set in the stationary case. The flexibility of GEV in contemplating Gumbel, Fréchet and Weibull distributions (for $\varepsilon = 0$, $\varepsilon > 0$ and $\varepsilon < 0$ respectively) makes it eligible for a more general discussion about non-stationarity implications.

Traditional extreme value distributions can be used into a nonstationary framework, modelling their parameters as function of time or other covariates (Coles, 2001) producing $\boldsymbol{\theta}_{st} \rightarrow \boldsymbol{\theta}_{ns} = [\zeta_t, \sigma_t, \varepsilon_t]$.

In this study, only a deterministic linear dependence on time t of the position parameter ζ has been introduced, leading Eq. (7) to be expressed as:

$$F(z, \boldsymbol{\theta}_{ns}) = \begin{cases} exp\left\{-\left[1 + \varepsilon\left(\frac{z-\zeta_t}{\sigma}\right)\right]^{-1/\varepsilon}\right\} & \varepsilon \neq 0 \\ exp\left\{-exp\left[-\left(\frac{z-\zeta_t}{\sigma}\right)\right]\right\} & \varepsilon = 0 \end{cases} \qquad \sigma > 0 \tag{8}$$

with

$$\zeta_t = \zeta_0 + \zeta_1 t, \tag{9}$$

and $\boldsymbol{\theta}_{ns} = [\zeta_0, \zeta_1, \sigma, \varepsilon]$.

It is important to note that Eq. (8) is a more general way to define the GEV and has the property of degenerating into Eq. (7) for $\zeta_1 = 0$: in other words Eq. (7) represents a nested model of Eq. (8) which confirms the suitability of the Likelihood Ratio test for model selection.

According to Muraleedharan et al. (2010), the first three moments of GEV distribution are:

$$\text{Mean} = \zeta + \frac{\sigma}{\varepsilon}(g_1 - 1) \qquad \varepsilon \neq 0, \varepsilon < 1 \tag{10}$$

$$\text{Variance} = \frac{\sigma^2}{\varepsilon^2}(g_2 - g_1^2) \qquad \varepsilon \neq 0, \varepsilon < \frac{1}{2} \tag{11}$$

$$\text{Skewness} = sgn(\varepsilon) \cdot \frac{g_3 - 3g_2 g_1 + 2g_1^3}{(g_2 - g_1^2)^{3/2}} \qquad \varepsilon \neq 0, \varepsilon < \frac{1}{3} \tag{12}$$

where $g_k = \Gamma(1 - k\varepsilon)$, with $k \in \mathbb{Z}^+$ and $\Gamma(\cdot)$ is the Gamma function. It is worth noting that, following Eqs. (10), (11) and (12), the trend in the position parameter only affect the Mean, while Variance and Skewness remain constant.

In this work we used the maximum likelihood method (ML) to estimate GEV parameters from sample data. ML allows to treat ζ_1 as an independent parameter, as well as ζ_0, σ and ε. To this purpose we exploited the R package extRemes (Gilleland and Katz, 2016).

2.5 Numerical evaluation of test power and significance level

The power of a test is related to the second type error and is the probability of correctly rejecting the null hypothesis when it is false. In particular, defining α (level of significance) the probability of a Type I error and β the probability of a Type II error, we have *power* = 1 − β. The maximum value of power is 1, which correspond to β = 0, i.e. no probability of Type II error. In

most applications conventional values are α = 0.05 and β = 0.2, meaning that a 1-to-4 trade-off between α and β is accepted. Then, in our experiment we assumed always significance level 0.05, and, for the following description of results and discussion we considered a power level less than 0.8 as too low and hence unacceptable. In the conclusions section we report further considerations about this choice. For each of the tests described in subsections 2.1, 2.2 and 2.3, the power was numerically evaluated according to the following procedure:

1) N = 2000 Monte Carlo synthetic series are generated using the non-stationary GEV in Eqs. (8-9) as parent distribution with fixed parameter set $\boldsymbol{\theta}_{ns} = [\zeta_0, \zeta_1, \sigma, \varepsilon]$ and length L, being $\zeta_1 \neq 0$.

2) The threshold $AIC_{R,\alpha}$ associated with a significance level $\alpha = 0.05$ is numerically evaluated, as described in section 2.3 using the corresponding parameter set $\boldsymbol{\theta}_{st} = [\zeta_0, \sigma, \varepsilon]$ of GEV parent distribution.

3) From these synthetic series the power of the test is estimated as:

$$rejection\ rate = \frac{N_{rej}}{N}$$

where N_{rej} is the number of series for whom the null hypothesis is rejected, as in Yue et al. (2002a).

The same procedure, with N = 10000, was used in order to check the actual significance level of test, which is the probability of first type error i.e. the probability of rejecting the null hypothesis of stationarity when it is true. The task was performed by following the above steps from 1 to 3 while replacing $\boldsymbol{\theta}_{ns}$ with $\boldsymbol{\theta}_{st}$ at step 1), in such a case the rejection rate N_{rej}/N represents the actual level of significance α.

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).

3 Sensitivity analysis, results and discussion

A comparative evaluation of the tests' performance was carried out for all the GEV parameter sets $\boldsymbol{\theta}_{ns}$ obtained considering three values of ε (-0.4, 0, 0.4) and three values of σ (10, 15, 20). The position parameter was kept always constant and equal to $\zeta_0 = 40$. Then, for any possible couple of σ and ε, we considered ζ_1 ranging from -1 to 1 with step 0.1. Such a range of parameters represents a wide domain in the hydrologically feasible parameters space of annual maximum daily rainfall. Upper bounded ($\varepsilon = -0.4$), EV1 ($\varepsilon = 0$), and heavy tailed ($\varepsilon = +0.4$) cases are included. Moreover, for each of these parameter sets $\boldsymbol{\theta}_{ns}$, N samples of different size (30, 50 and 70) were generated.

For a clear exposition of results, this section is divided into four subsections. In the first one we focus on the opportunity of using AIC or AIC_c for the evaluation of AIC_R, in the second one the comparison of test power and its sensitivity analysis to parent distribution parameters and sample size is shown. In the third one, the evaluation of the level of significance for all tests and in particular the validity of the χ^2 approximation for the D statistic is discussed. In the fourth subsection the numerical investigation on the sample variability of parameters is reported.

3.1 Evaluation of AIC_R, with AIC or AIC_c

Considering the non-stationary GEV four-parameter model, in order to satisfy the relation $L/k_{max} < 40$ suggested by Burnham e Anderson (2004), a time series with a record length not less than 160 should be available. Following this simple reasoning the *AIC* should be considered *de facto* not-applicable to any annual maximum series showing a changing point in the '70-80s (e.g. Kiely, 1999). In our numerical experiment, the second-order bias correction of Sugiura (1978) should be always used because for the maximum sample length, L = 70, we have $L/k_{max} = 70/4 = 17.5$ for the non-stationary GEV. Nevertheless, we checked using *AIC* or AIC_c may affect results. To this purpose we evaluated from synthetic series the percentage differences between the power of AIC_R obtained by means of *AIC* and AIC_c. In Fig. 2 the empirical probability density functions of such percentage differences, grouped according to sample length, are plotted for generations with ε = 0.4 and different values of σ. It is interesting to note that only for L = 70 the error distribution shows a regular and unbiased bell-shaped distribution. Then we observe for L = 50 a small negative bias (about -0.02%), while for L = 30 a bias of -0.08 with a multi-peak and negatively skewed pdf. The latter pdf also has a higher variance than the others. The purpose of this figure is to show that the difference between the power obtained with AIC and the power obtained with AIC_c is negligible. Different peaks in one curve (L = 30) can be explained by the merge of sample errors obtained for different values of σ. Similar results were obtained for all values of ε, providing always very low differences and allowing to conclude that the use of *AIC* or AIC_c does not significantly affect the power of AIC_R for the cases examined. This follows the combined effect of the sample size (whose minimum value considered here is 30) and the limited difference in the number of parameters in the selected models. In the following we will refer and show only the plots obtained for the AIC_R in Eq. (6) with *AIC* evaluated as in Eq. (4).

3.2 Dependence of power on parent distribution parameters and sample size

The effect of parent distribution parameters and sample size on the numerical evaluation of power and significance level of *MK*, *LR* and AIC_R for different values of ε, σ and ζ_1 is shown in Fig. 3. The curves represent both significance level which is shown for $\zeta_1 = 0$ (true parent is the stationary GEV) and power for all other values $\zeta_1 \neq 0$ (true parent is the non-stationary GEV). Each subplot in Fig. 3 shows the dependence on the trend coefficient of power and significance level of *MK*, *LR* and AIC_R for one set of parameter values and different sample sizes. In all subplots the test power strongly depends on trend coefficient and sample size. This dependence is also affected by parent parameter values. In all cases the power reaches 1 for strong trend and approaches 0.05 (the chosen level of significance) for weak trend (ζ_1 close to 0). In all combinations of the shape and scale parameters, and especially for short samples, for a wide range of trend values the power exhibits values well below the conventional value 0.8. The curves' slope between 0.05 and 1 is sharp for long samples and slow for short samples. It also depends on the parameter set, being such a slope generally slower for higher values of the scale (σ) and shape (ε) parameters of the parent distribution. Significant difference of power between *MK*, *LR* and AIC_R is observable when the sample size is smaller and still more when the parent is heavy tailed ($\varepsilon = +0.4$).

In particular, for $\varepsilon = 0$, -0.4 and $L = 50, 70$ it is possible to report a slightly larger power of LR with respect to AIC_R and MK, but values are very close to each other. Interesting is the reciprocal position of MK and AIC_R power curves: in fact, the AIC_R power is always larger than the MK one, except when $\varepsilon = -0.4$, without sensible influence of the scale parameter. Higher difference is found for heavy tailed parent distribution ($\varepsilon = +0.4$). While LR keeps having the largest power, the difference with respect to AIC_R remains small while the MK's power almost collapses to values always smaller than 0.5. Practical consequences of such patterns are very important and are discussed in the concluding section.

3.3 Sensitivity and evaluation of the actual significance level

We evaluated the threshold values (corresponding to a significance level of 0.05) for accepting/rejecting the null hypothesis of stationarity according to the methodologies recalled in subsections 2.1 and 2.2 for MK and LR tests and introduced in section 2.3 for AIC_R. Based on such thresholds we exploited the generation of series from a stationary model ($\zeta_1 = 0$) in order to numerically evaluate the rate of rejection of the null-hypothesis, i.e. the actual significance level of the tests considered in the numerical experiment, following the procedure described in subsection 2.5.

Table 1 shows the numerical values of the actual level of significance, obtained numerically, to be compared to the theoretical value 0.05 for all the considered sets of parameters and sample size. Among the three measures for trend detection the LR shows the worst performance. Results in Table 1 show that the rejection rate of the (true) null hypothesis is systematically higher than it should be, and it is also dependent on parent parameter values. Such effect is exalted when the parent distribution is upper bounded ($\varepsilon = -0.4$) and for shorter series ($L = 30$). In practice this implies that when using the LR test, as described in subsection 2.2, there is a quite higher probability of rejecting the null hypothesis of stationarity (if it is true) than expected or designed.

On the other hand, the performances of MK with respect to the designed level of significance are less biased and independent from the parameter set. Similar good performances are trivially obtained for the AIC_R, whose rejection threshold is numerically evaluated.

The plot in Fig. 4 is displayed in order to focus on the actual value of the level of significance and in particular on the LR approximation $D \sim \chi_m^2$ as a function of the sample length L. The difference between theoretical and numerical values of the significance level is represented by the distance between the bottom value of the curve (obtained for $\zeta_1 = 0$, i.e. the stationary GEV model) and the chosen level of significance 0.05 which is represented by the horizontal dotted line. In particular in Fig. 4 results for the parameter set ($\sigma = 15, \varepsilon = -0.4$) show that the actual rate of rejection is always higher than the theoretical one and changes significantly with the sample size, which means that the χ_m^2 approximation leads to significant underestimation of the rejection threshold of the D statistic. Moreover, it seems that the entire curves of the LR power (in red) are upward translated as a consequence of the significance level overestimation, meaning that the LR test power is also overestimated because of the approximation $D \sim \chi_m^2$. These results suggest, for the LR test, the use of a numerical procedure (as the one introduced for AIC_R in subsection 2.3) for evaluating the D distribution and the rejection threshold.

Other considerations can be made on the use of AIC_R. As explained in subsection 2.3 we empirically evaluated by numerical generations the $AIC_{R,\alpha}$ threshold value with significance level 0.05 for each of the parameter sets and sample sizes considered.

295 Similar results were obtained using the AIC_c which are not shown for brevity. We found a significative dependence of $AIC_{R,\alpha}$ on the sample size. Fig. 5 shows curves of $AIC_{R,\alpha}$ obtained for each of the parameter sets vs sample size. It is also worth noting that all curves asymptotically trend to 1 as L increases. This property is due to the structure of AIC and peculiarity of the nested models used in this paper: while using a sample generated with weak non-stationarity (i.e. when $\zeta_1 \to 0$ in Eq. (9)) the maximum likelihood of model (7), $\ell(\widehat{\boldsymbol{\theta}}_{st})$, tends to $\ell(\widehat{\boldsymbol{\theta}}_{ns})$ of model (8) leaving only the bias correction in AIC to be

300 discriminant for model selection. As a consequence, $AIC_{R,\alpha}$ should be always lower than 1, but, increasing sample size, also both the likelihood terms $-2\ell(\widehat{\boldsymbol{\theta}}_{st})$ and $-2\ell(\widehat{\boldsymbol{\theta}}_{ns})$ in Eq. (4) will increase, pushing AIC_R toward the limit 1. On the other hand, Fig. 5 shows that the threshold value $AIC_{R,\alpha}$ is significatively smaller than 1 up to L values well beyond the length usually available in this kind of analysis. Hence the numerical evaluation of the threshold has to be considered as a required task in order to provide an assigned significance level to model selection. On the other hand, the simple adoption of the

305 selection criteria $AIC_R < 1$ (i.e. $AIC_{R,\alpha} = 1$), would correspond to an unknown significance level dependent on the parent distribution and sample size. In order to highlight this point, we evaluated the significance level α corresponding to $AIC_{R,\alpha} = 1$ following the procedure described in subsection 2.5 by generating N = 10000 synthetic series, from a stationary model, for any parameter set and sample length. Results, provided in Tab. 2, show that, in the explored GEV parameter domain, α ranges between 0.16 and 0.26 mainly depending on the sample length and the shape parameter of the parent distribution.

310 **3.4 Sample variability of parent distribution parameters**

Results shown above, with regard to performances of parametric and non-parametric tests, are in our opinion quite surprising and important. It is proved that the preference widely accorded to non-parametric tests, being their statistics allegedly independent from the parent distribution, is not well founded. On the other hand, the use of parametric procedures raises the problem of correctly estimating the parent distribution and, for the purpose of this paper, its parameters. Moreover, as being

315 the trend coefficient ζ_1 a parameter of the parent distribution in non-stationary condition, the proposed parametric approach provides a maximum likelihood-based estimation of the same trend coefficient which is hereafter called ML-ζ_1. For a comparison with non-parametric approaches we also evaluated the sample variability of the Sen's slope measure (δ) of the imposed linear trend. In order to provide insights into these issues, from the same sets of generations exploited above, we also analysed the sample variability of the maximum likelihood estimates ML-ε ML-σ, for different parameter sets and sample

320 length.

We evaluated sample variability $s[\cdot]$, as the standard deviation of the ML estimates of parameter values obtained from synthetic series. In the upper panels of Fig 6 we show s[ML-ζ_1] and the lower panels the Sen's slope median $s[\delta]$. The sample variability of the linear trend is in both cases strongly dependent on sample size and independent from the true ζ_1 value in the range

examined $[-1,1]$. It reaches high values for short samples and in such cases also its dependence on the scale and shape parent parameters is relevant. The ML estimation of the trend coefficient is always more efficient than Sen's slope and this is observed in particular for heavy tailed distributions.

In Fig. 7 we show the empirical distributions of the Sen's slope δ and ML-ζ_1 estimates obtained from samples of size $L = 30$ with parent distribution characterized by $\sigma = 15$ and $\varepsilon = [-0.4, 0, 0.4]$, providing visual information about the range of trend values that may result from a local evaluation. Similar results, characterized by smaller sample variability shown in Fig. 6, are obtained for $L = 50$ and $L = 70$ and are not shown for brevity.

Fig. 8 shows the sample variability of ML-ε and ML-σ, which is still independent from the true ζ_1 for values of $\varepsilon = 0$ and 0.4 while for upper bounded GEV distributions ($\varepsilon = -0.4$) it shows a significant increase for higher values of σ and high trend coefficients ($|\zeta_1| > 0.5$). 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.

In order to better analyse such patterns, for the scale and shape parent parameters we report also the distribution of their empirical ML estimates for different parameter sets vs the true ζ_1 value used in generation. The sample distribution of ML-ε for $\sigma = 15$ is shown in Fig. 9 for $L = 30$ and $L = 70$. The sample distribution of ML-σ for $\sigma = 15$ is shown in Fig. 10 for $L = 30$ and $L = 70$. Subplots show that the presence of a strong trend coefficient may produce significant loss in the estimator efficiency probably due to deviation from normal distribution of the sample estimates also for long samples. This suggests the need of more robust estimation procedures which provides higher efficiency for estimates of ε and σ in case of strong observed trend. It should be highlighted that efficiency in parameter estimation increases with sample size for $\varepsilon = [0, 0.4]$, while it decreases for both ε and σ, in the case $\varepsilon = -0.4$, where the trend of the location parameter implies a shift in time of the distribution upper bound.

4 Conclusions

The results shown have important practical implications. The dependence of test power on the parent distribution parameters may significantly affect results of both parametric and non-parametric tests including the widely used Mann-Kendall.

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 (1992) 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 (1992) 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.

For all the generation sets and tests conducted, under the null hypothesis of stationarity, the power has values ranging between the chosen significance level (0.05) and 1 for large (and larger) ranges of the trend coefficient. The test power always collapses to very low values for weak (but climatically important) trend values (in the case of annual maximum daily rainfall, ζ_1 equal to 0.2 or 0.3 mm per year, for example). In presence of trend, the power is also affected by the scale and shape parameters of the GEV parent distribution. This observation can be made with reference to samples of all the lengths considered in this paper (from 30 to 70 years of observation) but the use of smaller samples significantly reduces the test power and dramatically extends the range of ζ_1 values for which the power is below the conventional value 0.8. The use of this sample size is not rare considering that significative trends due to anthropic effects are typically investigated in periods following a changing point often observed in the '80s.

These results also imply that in spatial fields where the alternative hypothesis of non-stationarity is true, but the parent's parameters (including the trend coefficient) and the sample length are variable in space, the rate of rejection of the false null-hypothesis may be highly variable from site to site and practically out of control. In other words, in such a case, the probability of recognizing the alternative hypothesis of non-stationarity as true from a single observed sample may unknowingly change (between 0.05 and 1) from place to place. For small samples (as $L = 30$ in our analysis) and heavy tailed distributions, the power is always very low for all the investigated range of the trend coefficient.

Hence, considering the high spatial variability of the parent distribution parameters and the relatively short period of reliable and continuous historical observations usually available, a regional assessment of trend non-stationarity may suffer from the different probability of rejection of the null hypothesis of stationarity (when it is false).

These problems affect, in slightly different measures, both parametric and non-parametric tests. While these considerations are generally applicable to all the tests considered, differences also emerge between them. For heavy tailed parent distributions and smaller samples, the *MK* test power decreases more rapidly than for the other tests considered. Low values of power are already observable for $L = 50$. The *LR* test slightly outperforms the AIC_R for small sample size and higher absolute values of the shape parameter. Nevertheless, the higher value of the *LR* power seems to be overestimated as a consequence of the χ_m^2 approximation for the D statistic distribution (see section 3.3).

Results also suggest that theoretical distribution of the *LR* test-statistic based on the null hypothesis of stationarity may lead to a significative increase of the rejection rate compared to the chosen level of significance i.e. an abnormal rate of rejection of the null hypothesis when it is true. In this case the use of numerical techniques, based on the use of synthetic generations performed by exploiting a known parent distribution, should be preferred.

By the light of these results we conclude that in trend detection on annual maximum series the assessment of the parent

390   distribution and the choice of the null hypothesis should be considered as fundamental preliminary tasks. According to this
remark, it is advisable to make use of parametric tests by numerically evaluating both the rejection threshold for the assigned
significance level and the power corresponding to alternative hypotheses. This also requires developing robust techniques for
selecting the parent distribution and estimating its parameters. To this perspective, the use of a parametric measure such as the
AIC_R, may take into account different choices for the parent distribution and, even more importantly, allows one to set the null

395   hypothesis differently from the stationary case, based on a priori information.

The need of robust procedures for assessing the parent distribution and its parameters is also proven by the numerical
simulations we conducted. Sample variability of parameters (including the trend coefficient) may increase rapidly for series
with L as low as 30 years of annual maxima. Moreover, we observed that, in case of high trends, numerical instability and
non-convergence of algorithms may affect the estimation procedure for upper bounded and heavy tailed distributions.

400   Nevertheless, the sample variability of the ML trend estimator was found always smaller than the Sen's slope sample
variability. Finally, it is worth noting that also the non-parametric Sen's slope method, applied to synthetic series, showed
dependence on the parent distribution parameters with sample variability higher for heavy tailed distributions.

This analysis shed lights onto important eventual flaws in the at-site analysis of climate change provided by non-parametric
approaches. Both test power and trend evaluation are affected by the parent distribution as it is the case also in parametric

405   methods. It is not by chance, in our opinion, that many technical studies conducted in years around the world, provide
inhomogeneous maps of positive/negative trends and large areas of stationarity characterized by weak trends that are
considered not statistically significany.

As already said, an advantage of using parametric tests and numerical evaluation of the test-statistic distribution is given by
the possibility of assuming a null hypothesis based on a preliminary assessment of the parent distribution including trend

410   detection by evaluation of non-stationary parameters. This could lead to a regionally homogeneous and controlled assessment
of both significance level and power in a fair mutual relationship. With respect to the estimation of parameters of the parent
distribution, results suggest that at site analysis may provide highly biased results. More robust procedures are necessary like
hierarchic estimation procedures (Fiorentino et al., 1987) providing estimates of ε and σ from detrended series (Strupczewski
et al., 2016; Kochanek et al., 2013).

415   As a final remark, concerning real data analysis, in our numerical experiment we showed that, in some cases, a weak linear
trend in the mean suffices to reduce power to unacceptable values. Yet we explored the simplest nonstationary working
hypothesis by introducing a deterministic linear dependence on time of the location parameter of the parent distribution.
Obviously, when making inference from real observed data other sources of uncertainty may affect statistical inference (trend,
heteroscedasticity, persistence, nonlinearity, etc), and moreover, if considering a nonstationary process with underlying

420   deterministic dynamics, the process turns out to be non-ergodic, implying that statistic inference from sampled series is not
representative of the process's ensemble properties (Koutsoyiannis and Montanari, 2015).

As a consequence, while considering a nonstationary stochastic process as produced by a combination of a deterministic function and a stationary stochastic process, other sources of information and deductive arguments should be exploited in order to identify the physical mechanism underlying such relationships. Even in such a case observed time series have a crucial role in order to calibrate and validate deterministic modeling or, in other words, for confirming or disproving the model hypotheses. In the field of frequency analysis of extreme hydrological events, considering the high spatial variability of sample length, trend coefficient, scale and shape parameters, etc, physically based probability distributions could be further developed and exploited for selection and assessment of the parent distribution in the context of non-stationarity and change detection. Physically based probability distributions we refer to are: (i) those arising from stochastic compound processes introduced by Todorovic and Zelenhasic (1970), which include also the GEV (see Madsen et al., 1997) and the TCEV (Rossi et al., 1984), and (ii) the theoretically derived distributions following Eagleson (1972) whose parameters are provided by clear physical meaning and are usually estimated with support of exogenous information in regional methods (e.g. Gioia et al., 2008; Iacobellis et al., 2011; see also for a more extensive overview Rosbjerg et al., 2013).

Hence, we believe that "learning from data" (Sivapalan, 2003), will remain in future years a key task for hydrologists facing the challenge of consistently identifying both deterministic and stochastic components of change (Montanari et al., 2013). This involves crucial and interdisciplinary research to develop suitable methodological frameworks for enhancing physical knowledge and data exploitation, in order to reduce the overall uncertainty of prediction in a changing environment.

References

Akaike, H.: A new look at the statistical model identification, IEEE Trans. Automat. Control, 19, 716–723, http://dx.doi.org/10.1109/TAC.1974.1100705, 1974.

Beven, K.: Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrol. Sci. J., doi:10.1080/02626667.2015.1031761, 2016.

Burnham, K. P., and Anderson, D. R.: Model selection and multimodel inference, Springer, New York, 2004.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W.: Non-stationary extreme value analysis in a changing climate, Climatic Change, 127, 353–369, https://doi.org/10.1007/s10584-014-1254-5, 2014.

Cohen, J., A power primer, Psychological Bulletin, Vol 112(1), Jul 1992, 155-159, 1992.

Cohen, J.: The earth is round (p < .05), American Psychol., 49, 997–1003, 1994.

Chow, V. T.: Statistical and probability analysis of hydrologic data, in: Handbook of applied hydrology, edited by Chow, V. T., McGraw-Hill, New York, 8.1-8.97, 1964.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, London, 2001

Cooley, D.: Return Periods and Return Levels Under Climate Change, in: Extremes in a Changing Climate, edited by: AghaKouchak A., Easterling D., Hsu K., Schubert S., Sorooshian S., Springer Dordrecht, 97-113, https://doi.org/10.1007/978-94-007-4479-0_4, 2013.

Du, T., Xiong, L., Xu, C. Y., Gippel, C. J., Guo, S., and Liu, P.: Return period and risk analysis of nonstationary low-flow series under climate change, J. Hydrol., 527, 234-250, https://doi.org/10.1016/j.jhydrol.2015.04.041, 2015.

Eagleson, P. S.: Dynamics of flood frequency, Water Resour. Res., doi:10.1029/WR008i004p00878, 1972.

Fiorentino, M., Gabriele, S., Rossi, F., and Versace, P.: Hierarchical approach for regional flood frequency analysis, in: Regional Flood Frequency Analysis, edited by: Singh, V. P., D. Reidel, Norwell, Mass., 35-49, 1987.

Gilleland, E., and Katz, R.W.: extRemes 2.0: An Extreme Value Analysis Package in R, J. Stat. Soft., 72, 1-39. https://www.jstatsoft.org/article/view/v072i08.

Gioia, A., Iacobellis, V., Manfreda, S. and Fiorentino, M.: Runoff thresholds in derived flood frequency distributions, Hydrol. Earth Syst. Sci., doi:10.5194/hess-12-1295-2008, 2008.

Gocic, M. and Trajkovic, S.: Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia, Global Planet. Change, 100, 172–182, https://doi.org/10.1016/j.gloplacha.2012.10.014, 2013.

Iacobellis, V., Gioia, A., Manfreda, S. and Fiorentino, M.: Flood quantiles estimation based on theoretically derived distributions: Regional analysis in Southern Italy, Nat. Hazards Earth Syst. Sci., doi:10.5194/nhess-11-673-2011, 2011.

Jenkinson, A. F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements, Q. J. R. Meteorol. Soc., 81, 158–171, 1955.

Kendall, M.G.: Rank Correlation Methods, 4th ed., Charles Griffin: London, UK, 1975.

Khintchine, A.: Korrelationstheorie der stationären stochastischen Prozesse, Math. Ann., doi:10.1007/BF01449156, 1934.

Kiely, G.: Climate change in Ireland from precipitation and streamflow observations, Adv. Water Resour., 23, 141–151, https://doi.org/10.1016/S0309-1708(99)00018-4, 1999.

Kochanek, K., Strupczewski, W.G., Bogdanowicz, E., Feluch, W. and Markiewicz, I.: Application of a hybrid approach in nonstationary flood frequency analysis – a Polish perspective, *Nat. Hazards Earth Syst. Sci. Discuss*. **1**, 6001-6024, https://doi.org/10.5194/nhessd-1-6001-2013, 2013.

Koutsoyiannis, D., and Montanari A.: Negligent killing of scientific concepts: the stationarity case, Hydrol. Sci. J., 60, 1174-1183, https://doi.org/10.1080/02626667.2014.959959, 2015.

Kundzewicz, Z. W. and Robson, A. J.: Change detection in hydrological records – a review of the methodology, Hydrolog. Sci. J., 49, 7–19, https://doi.org/10.1623/hysj.49.1.7.53993, 2004.

Laio, F., Baldassarre, G. D., and Montanari, A.: Model selection techniques for the frequency analysis of hydrological extremes, Water Resour. Res., 45, https://doi.org/10.1029/2007wr006666, 2009.

Lehmann, E. L., and D'Abrera, H. J. (eds.): Nonparametrics: Statistical methods based on ranks, Holden-Day, Oxford, England, 1975.

Lieber, R. L.: Statistical significance and statistical power in hypothesis testing, J. Orthop. Res., doi:10.1002/jor.1100080221, 1990.

Madsen, H., Rasmussen, P. and Rosbjerg, D.: Comparison of annual maximum series and partial duration series for modelling exteme hydrological events: 1. At-site modeling, Water Res. Res., 1997.

490 Mann, H. B.: Nonparametric tests against trend, Econom., 13, 245-259, http://doi.org/10.2307/1907187, 1945.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Dettinger, M. D. and Krysanova, V.: On Critiques of "stationarity is Dead: Whither Water Management?," Water Resour. Res., doi:10.1002/2015WR017408, 2015.

Montanari, A., Young, G., Savenije, H.H.G., Hughes, D., Wagener, T., Ren, L.L., Koutsoyiannis, D., Cudennec, C., Toth, E.,

495 Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S.J., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z., and Belyaev, V., "Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022. Hydrological Sciences Journal. 58 (6) 1256–1275, 2013.

Montanari, A. and Koutsoyiannis, D.: Modeling and mitigating natural hazards: Stationarity is immortal! Water Resour. Res.,

500 50, 9748–9756, https://doi.org/10.1002/2014wr016092, 2014.

Muraleedharan, G., Guedes Soares, C. and Lucas, C.: Characteristic and moment generating functions of generalised extreme value distribution (GEV), in Sea Level Rise, Coastal Engineering, Shorelines and Tides., 2010.

Olsen, J. R., Lambert, J. H., and Haimes, Y. Y.: Risk of extreme events under nonstationary conditions, Risk Anal., 18, 497-510, https://doi.org/10.1111/j.1539-6924.1998.tb00364.x, 1998

505 Parey, S., Malek, F., Laurent, C., and Dacunha-Castelle, D.: Trends and climate evolution: statistical approach for very high temperatures in France, Clim. Chang., 81, 331-352, https://doi.org/10.1007/s10584-006-9116-4, 2007.

Parey, S., Hoang, T. T. H., and Dacunha-Castelle, D.: Different ways to compute temperature return levels in the climate change context, Environmetrics, 21, 698-718, https://doi.org/10.1002/env.1060, 2010.

Pettitt, A. N.: A non-parametric approach to the change-point problem, Appl. Stat., 28, 126–135,

510 http://dx.doi.org/10.2307/2346729, 1979.

Read, L. K., and Vogel, R. M.: Reliability, return periods, and risk under nonstationarity, Water Resour. Res., 51, 6381-6398, https://doi.org/10.1002/2015WR017089, 2015.

Rosbjerg, D., Blöschl, G., Burn, D., Castellarin, A., Croke, B., Di Baldassarre, G., V. Iacobellis, T. R. Kjeldsen, G. Kuczera, R. Merz, A. Montanari, D. Morris, T. B. M. J. Ouarda, L. Ren, M. Rogger, J. L. Salinas, E. Toth, and Viglione, A.: Prediction

515 of floods in ungauged basins. In G. Blöschl, M. Sivapalan, T. Wagener, A. Viglione, & H. Savenije (Eds.), Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales (pp. 189-226). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139235761.012, 2013

Rossi, F., Fiorentino, M. and Versace, P.: Two-Component Extreme Value Distribution for Flood Frequency Analysis, Water Resour. Res., doi:10.1029/WR020i007p00847, 1984.

520 Salas, J. D., and Obeysekera, J.: Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events, J. Hydrol. Eng., 19, 554-568, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000820, 2014.

Schwarz, G.: Estimating the dimension of a model, Ann. Stat., 6, 461-464, (1978).

Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, J. Am. Stat. Assoc., 63, 1379–1389, http://dx.doi.org/10.2307/2285891, 1968.

525 Serinaldi, F., and Kilsby, C. G.: Stationarity is undead: Uncertainty dominates the distribution of extremes, Adv. Water Resour., 77, 17–36, https://doi.org/10.1016/j.advwatres.2014.12.013, 2015.

Serinaldi, F., Kilsby, C. G., and Lombardo, F.: Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology, Adv. Water Resour., 111, 132–155, https://doi.org/10.1016/J.ADVWATRES.2017.10.015, 2018.

Sivapalan, M., Prediction in Ungauged Basins: A Grand Challenge for Theoretical Hydrology, Hydrol. Process. 17, 3163–
530 3170, 2003.

Smadi, M. M., and Zghoul A.: A sudden change in rainfall characteristics in Amman, Jordan during the mid 1950s, Am. J. Environ. Sci., 2, 84–91, https://doi.org/10.3844/ajessp.2006.84.91, 2006.

Strupczewski, W.G., Kochanek, K., Bogdanowicz, E., Markiewicz, I. and Feluch, W.: Comparison of two nonstationary flood frequency analysis methods within the context of the variable regime in the representative polish rivers, *Acta Geophys*. 64,
535 206–236, https://doi.org/10.1515/acgeo-2015-0070, 2016.

Sugiura, N.: Further analysis of the data by Akaike's information criterion and the finite corrections, Commun. Stat. Theory Methods, A7, 13– 26, https://doi.org/10.1080/03610927808827599, 1978.

Todorovic, P. and Zelenhasic, E.: A Stochastic Model for Flood Analysis, Water Resour. Res., doi:10.1029/WR006i006p01641, 1970.

540 Tramblay, Y., Neppel, L., Carreau, J., and Najib, K.: Non-stationary frequency analysis of heavy rainfall events in southern France, Hydrolog. Sci. J., 58, 280–194, https://doi.org/10.1080/02626667.2012.754988, 2013.

Vogel, R. M., Rosner, A. and Kirshen, P. H.: Brief communication: Likelihood of societal preparedness for global change: Trend detection, Nat. Hazards Earth Syst. Sci., doi:10.5194/nhess-13-1773-2013, 2013.

Wang, W., Van Gelder, P.H., and Vrijling, J.K.: Trend and stationarity analysis for streamflow processes of rivers in Western
545 Europe in the 20th Century, IWA International Conference on Water Economics, Statistics, and Finance, 8–10 July, Rethymno, Greece, 2005.

Yilmaz, A. G., and Perera, B. J. C.: Extreme rainfall nonstationarity investigation and intensity–frequency–duration relationship, J. Hydrol. Eng., 19, 1160–1172, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000878, 2014.

Yilmaz, A. G., Hossain, I., and Perera, B. J. C.: Effect of climate change and variability on extreme rainfall intensity–
550 frequency–duration relationships: a case study of Melbourne, Hydrol. Earth Syst. Sci., 18, 4065–4076, https://doi.org/10.5194/hess-18-4065-2014, 2014.

Yue, S., Pilon, P., and Cavadias, G.: Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series, J. Hydrol., 259, 254-271, https://doi.org/10.1016/S0022-1694(01)00594-7, 2002a.

Yue, S., Pilon, P., Phinney, B., and Cavadias, G.: The influence of autocorrelation on the ability to detect trend in hydrological
555 series, Hydrol. Proc., 16, 1807–1829, https://doi.org/10.1002/hyp.1095, 2002b.

560

565

570

Table 1: Actual level of significance of tests for different sample size, scale and shape parent parameters

	L = 30								
	$\varepsilon = -0.4$			$\varepsilon = 0$			$\varepsilon = +0.4$		
	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
MK	0.048	0.047	0.047	0.047	0.050	0.050	0.046	0.049	0.048
AIC_R	0.050	0.046	0.052	0.051	0.052	0.045	0.052	0.054	0.051
LR	0.104	0.103	0.115	0.061	0.064	0.060	0.084	0.081	0.083

	L = 50								
	$\varepsilon = -0.4$			$\varepsilon = 0$			$\varepsilon = +0.4$		
	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
MK	0.050	0.047	0.046	0.044	0.047	0.050	0.049	0.044	0.048
AIC_R	0.053	0.053	0.046	0.051	0.051	0.057	0.050	0.050	0.053
LR	0.079	0.078	0.074	0.060	0.063	0.063	0.070	0.069	0.070

	L = 70								
	$\varepsilon = -0.4$			$\varepsilon = 0$			$\varepsilon = +0.4$		
	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
MK	0.050	0.052	0.054	0.052	0.051	0.047	0.049	0.048	0.046
AIC_R	0.047	0.051	0.051	0.058	0.058	0.052	0.050	0.054	0.051
LR	0.069	0.069	0.073	0.063	0.065	0.058	0.062	0.062	0.063

575

Table 2: Actual level of significance of AIC_R test for $AIC_{R,\alpha} = 1$

	$\varepsilon = -0.4$			$\varepsilon = 0$			$\varepsilon = +0.4$		
	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
L = 30	0.246	0.254	0.261	0.188	0.191	0.181	0.220	0.221	0.215
L = 50	0.213	0.209	0.206	0.171	0.175	0.170	0.188	0.207	0.195
L = 70	0.192	0.192	0.201	0.168	0.169	0.173	0.184	0.204	0.184

580

585

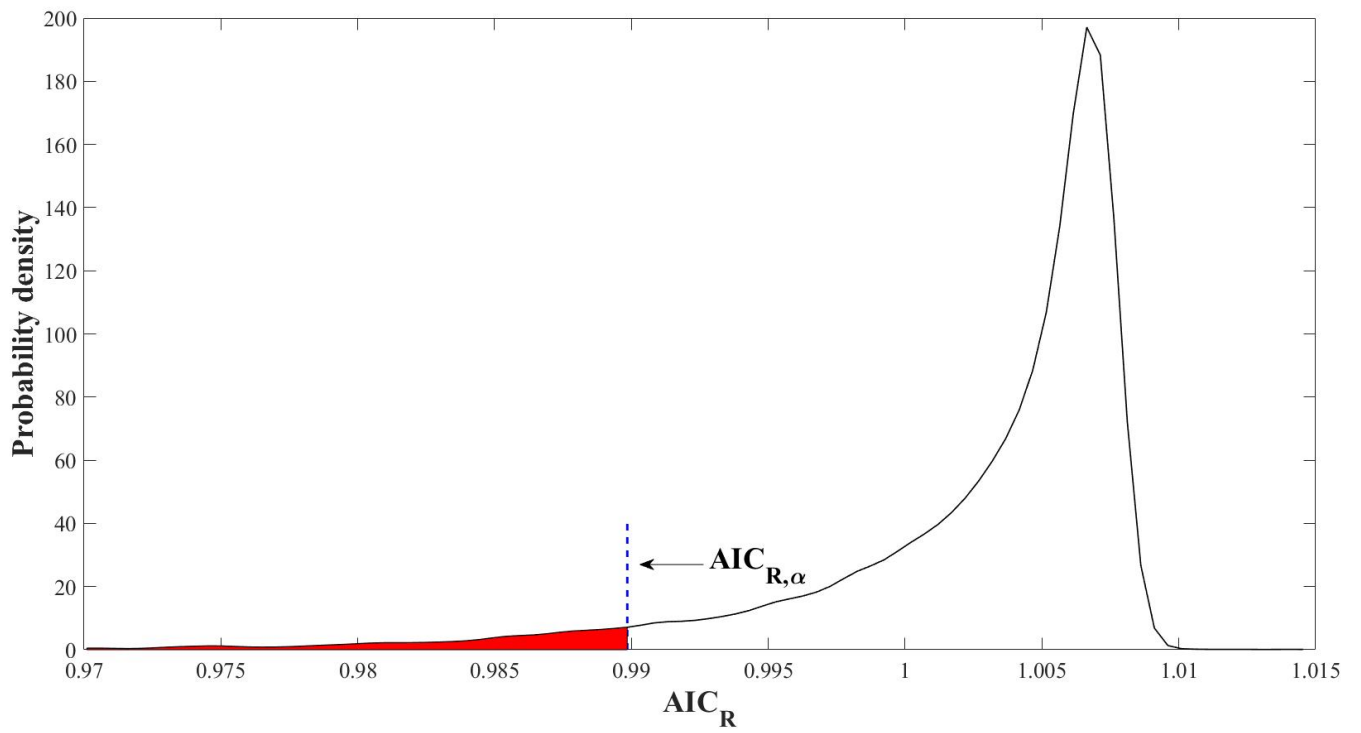
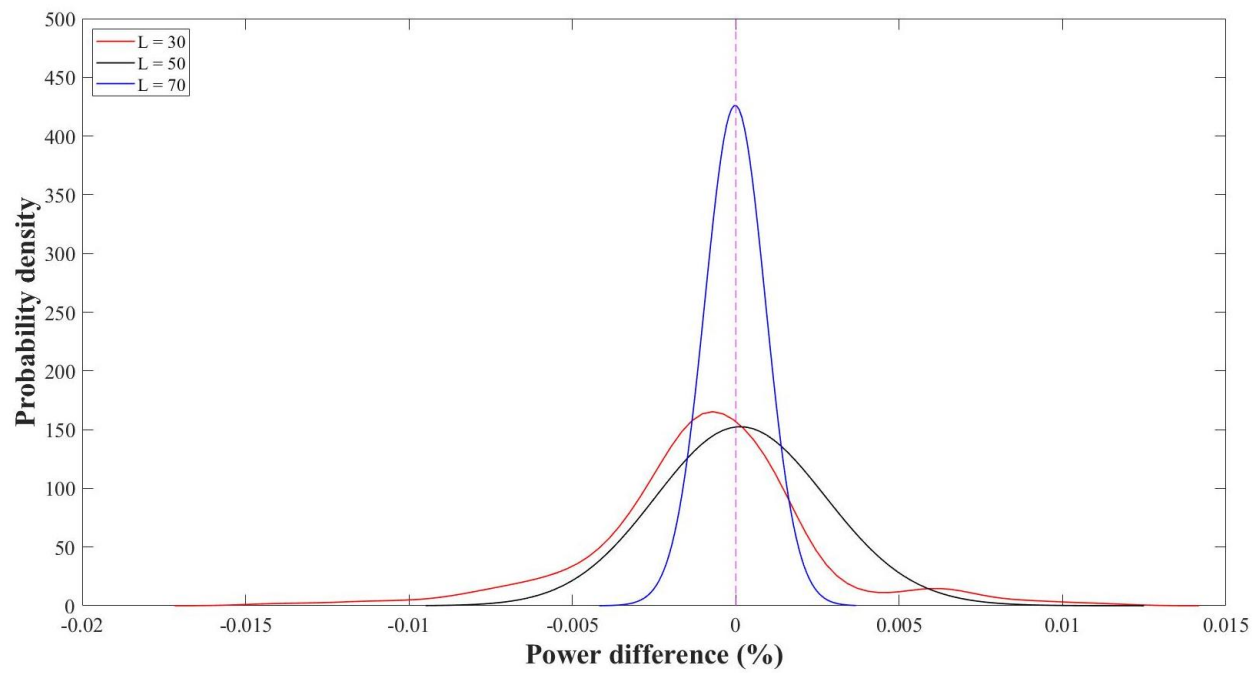



Figure 1: Empirical distribution of AIC$_R$ and rejection threshold $AIC_{R,\alpha}$ of the null hypothesis (stationary GEV parent)

590 **Figure 2: Distributions of the differences between power of AIC_R evaluated with AIC and AIC_c for $\varepsilon = 0.4$**

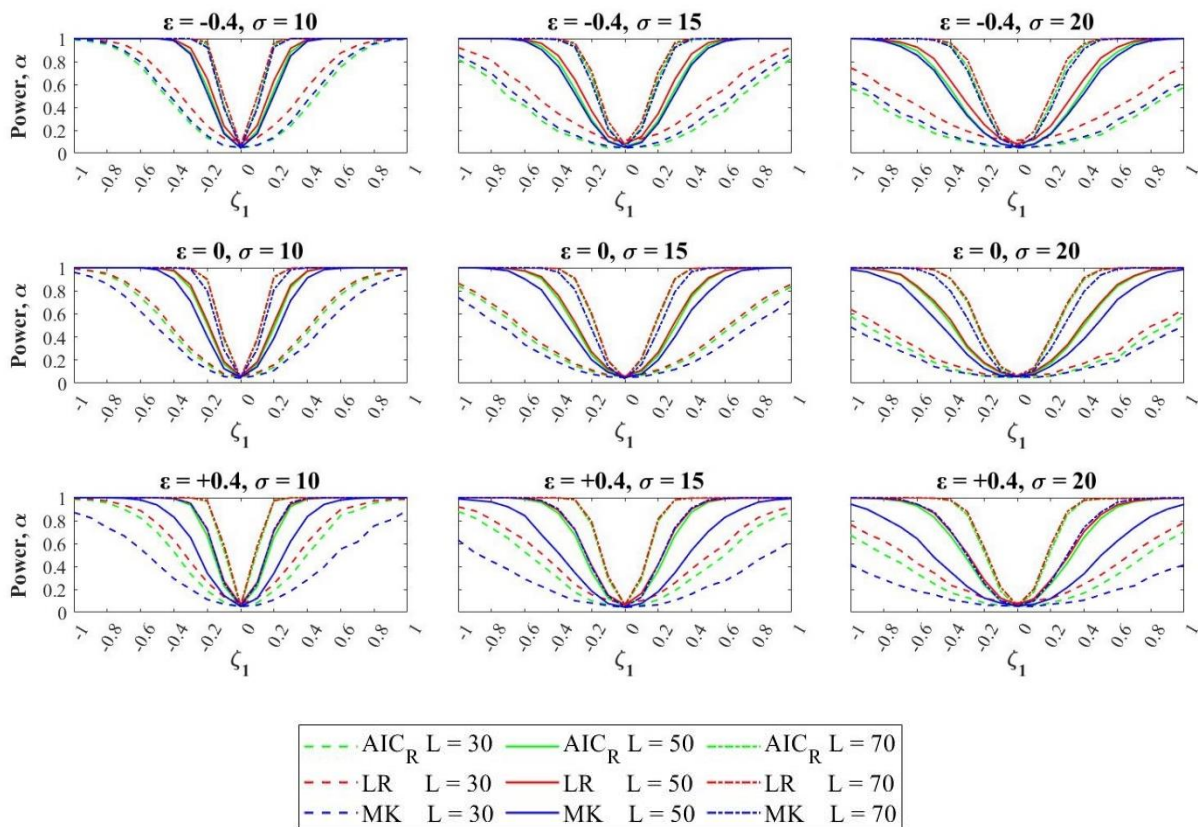
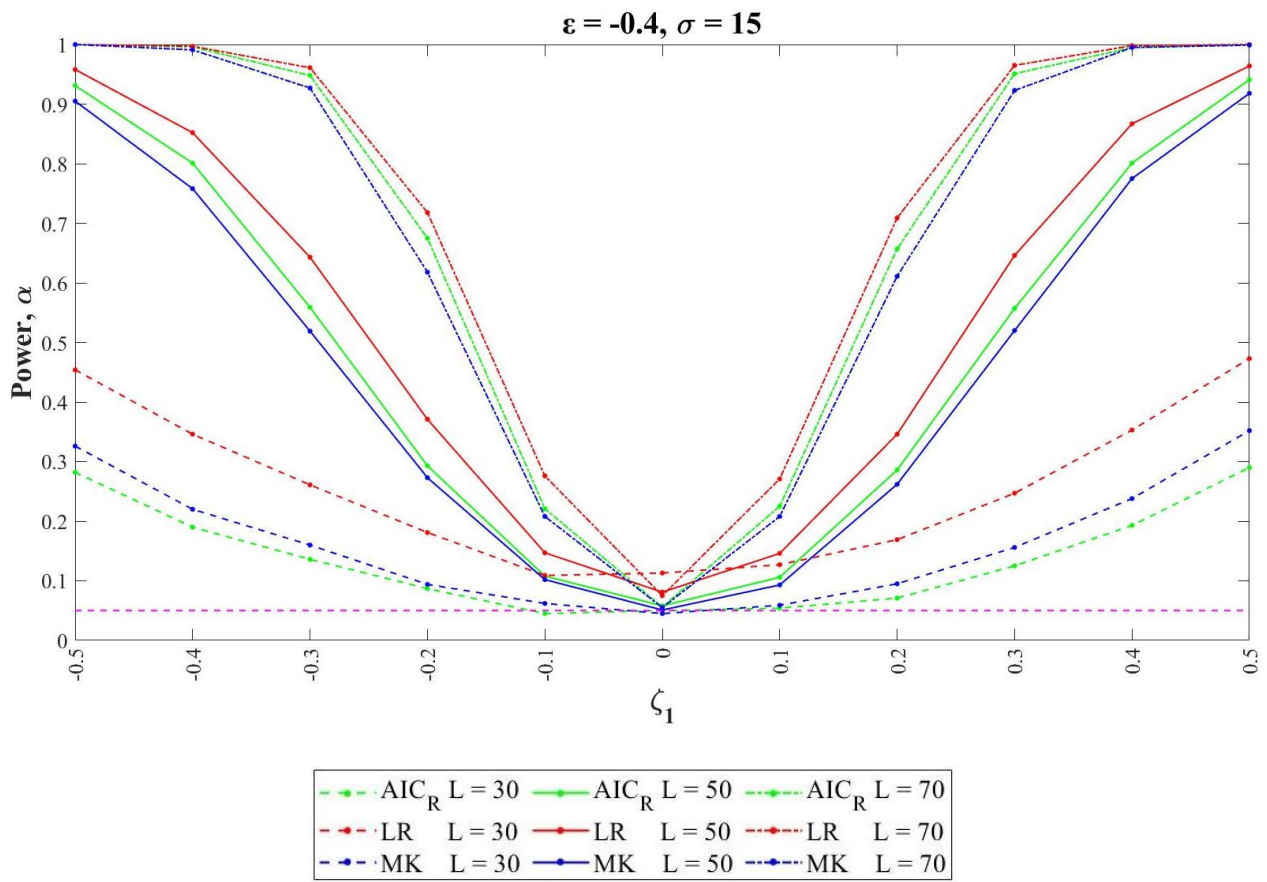
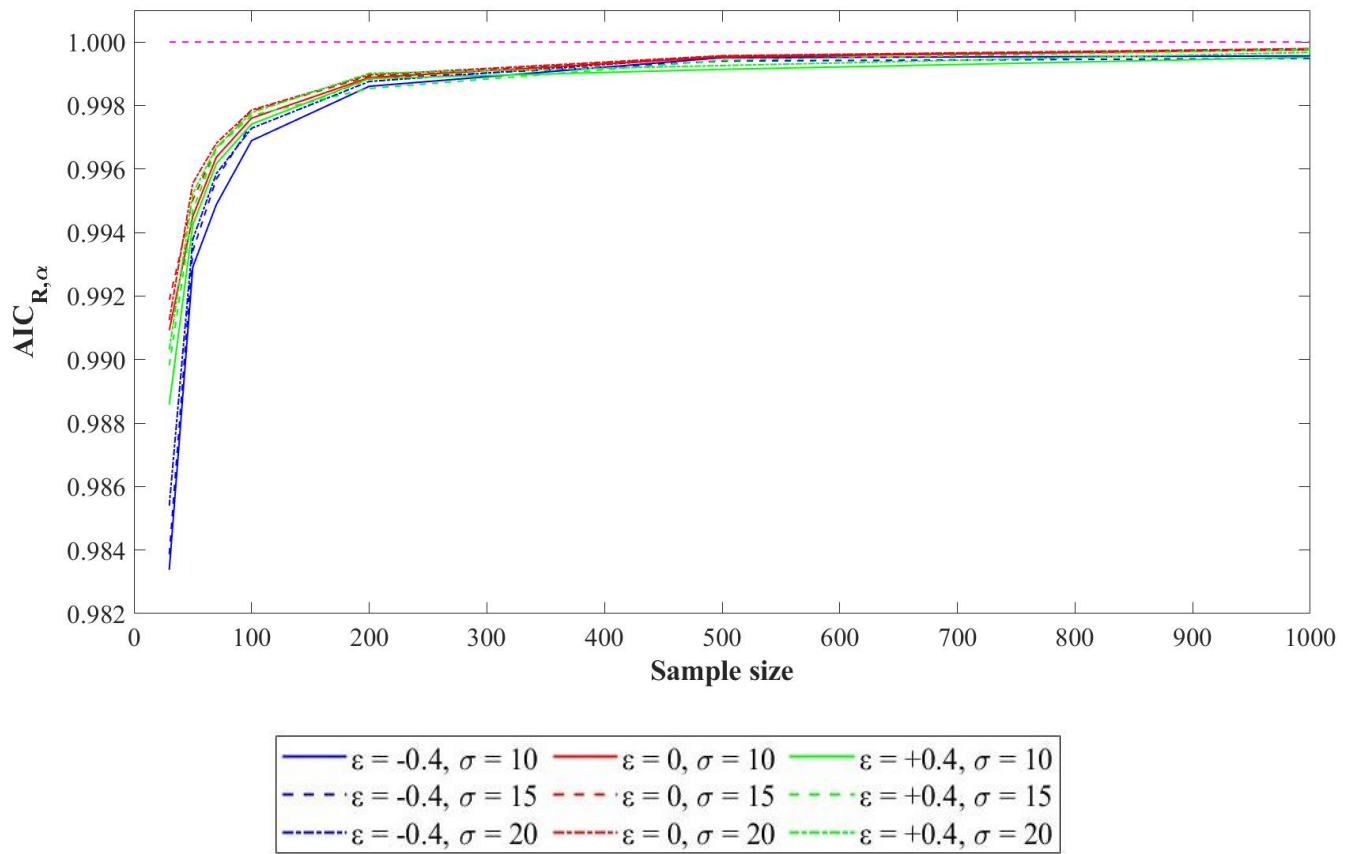
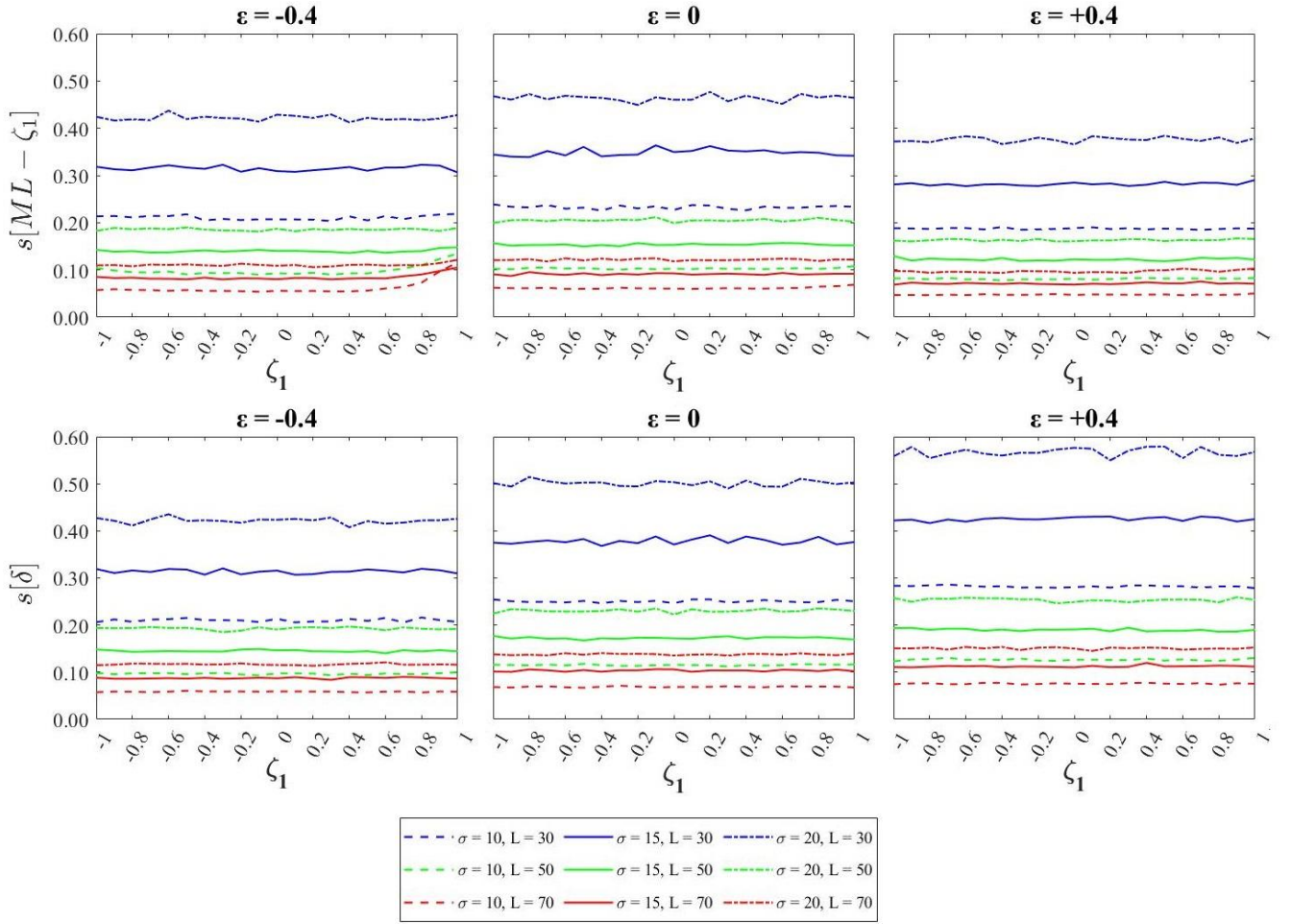



Figure 3: Dependence of test power on trend coefficient, sample size, scale and shape parent parameters.

600   **Figure 4: Focus on the actual level of significance reported for $\zeta_1 = 0$.**

605   **Figure 5: $AIC_{R,\alpha}$ thresholds for different parameter sets vs sample size**

610   **Figure 6: Sample variability of ML-ζ_1 and δ vs trend coefficient ζ_1**

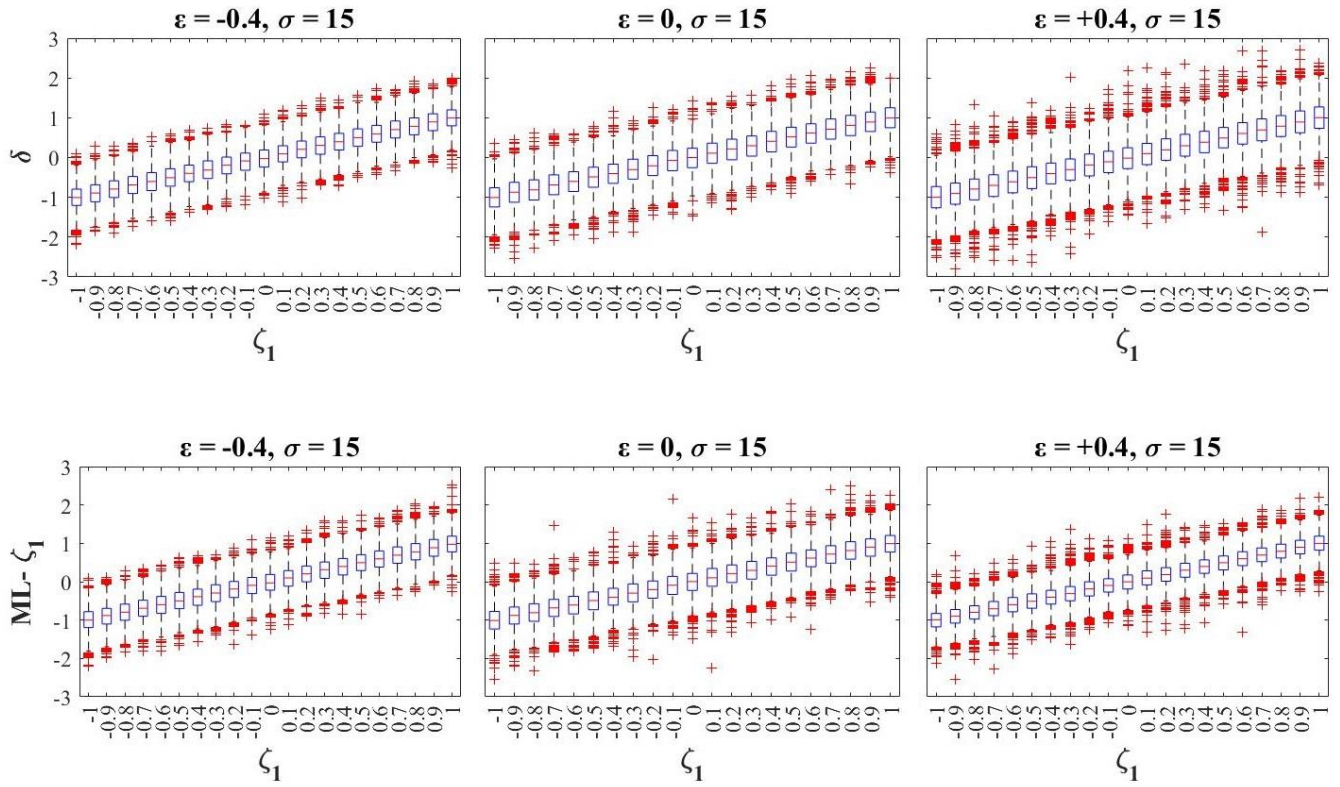



Figure 7: Empirical distributions of δ and ML-ζ_1 evaluated from samples with $L = 30$ and $\sigma = 15$ vs trend coefficient ζ_1

615

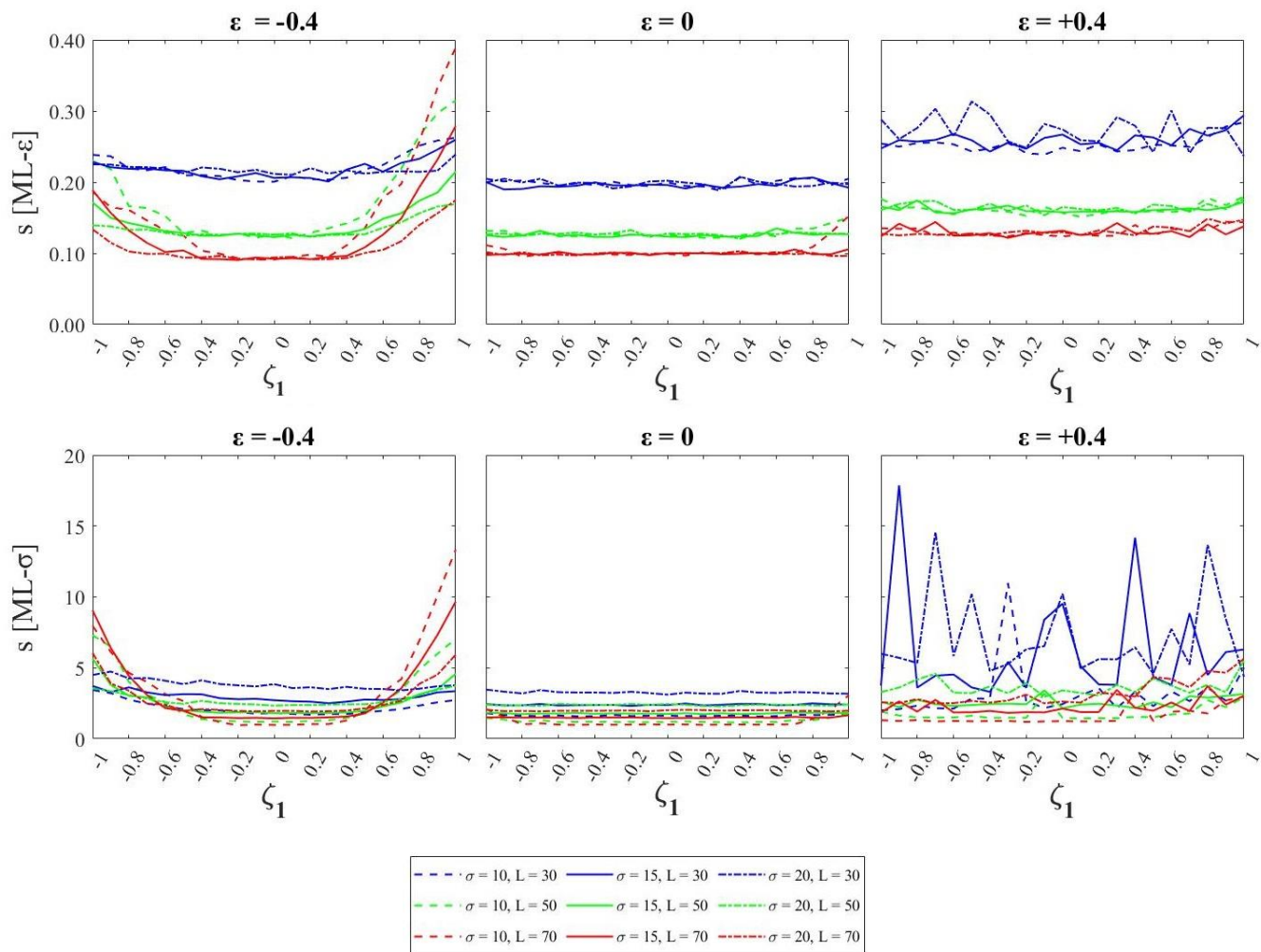



Figure 8: Sample variability of ML-ε and ML-σ vs trend coefficient ζ_1

620

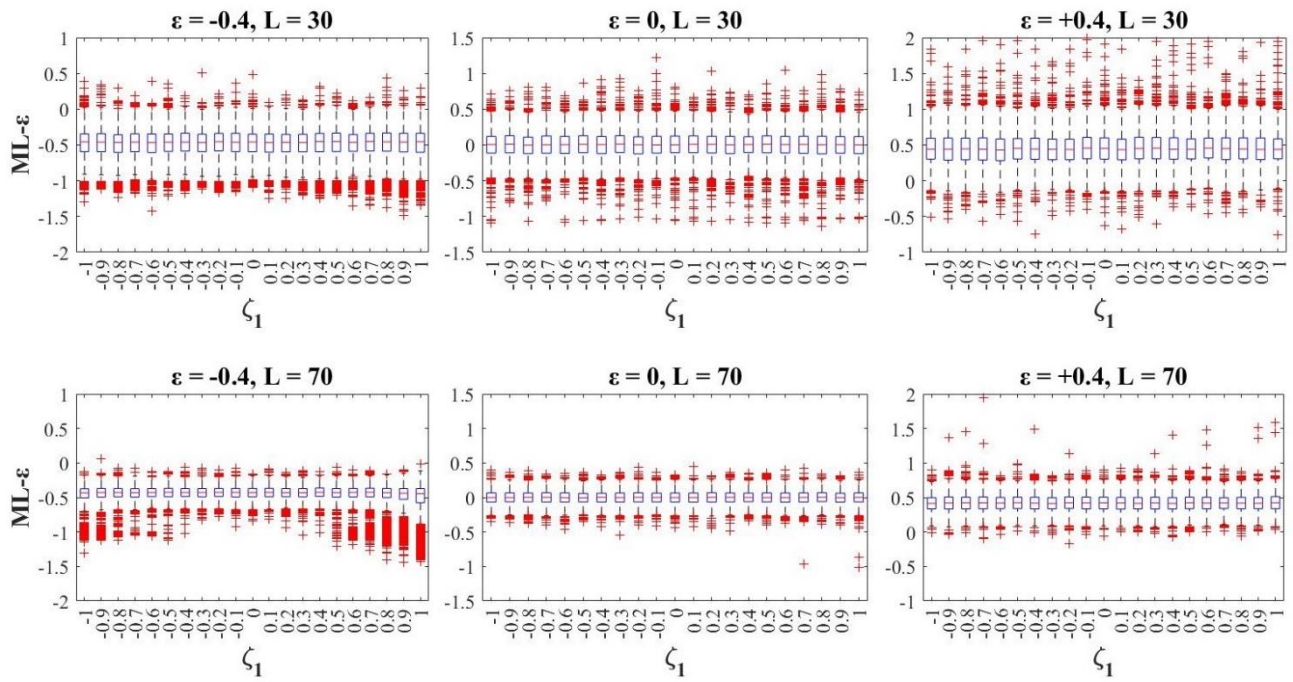



Figure 9: Empirical distributions of ML-ε evaluated for $\sigma = 15$ from samples with $L = 30$ and $L = 70$ vs trend coefficient ζ_1

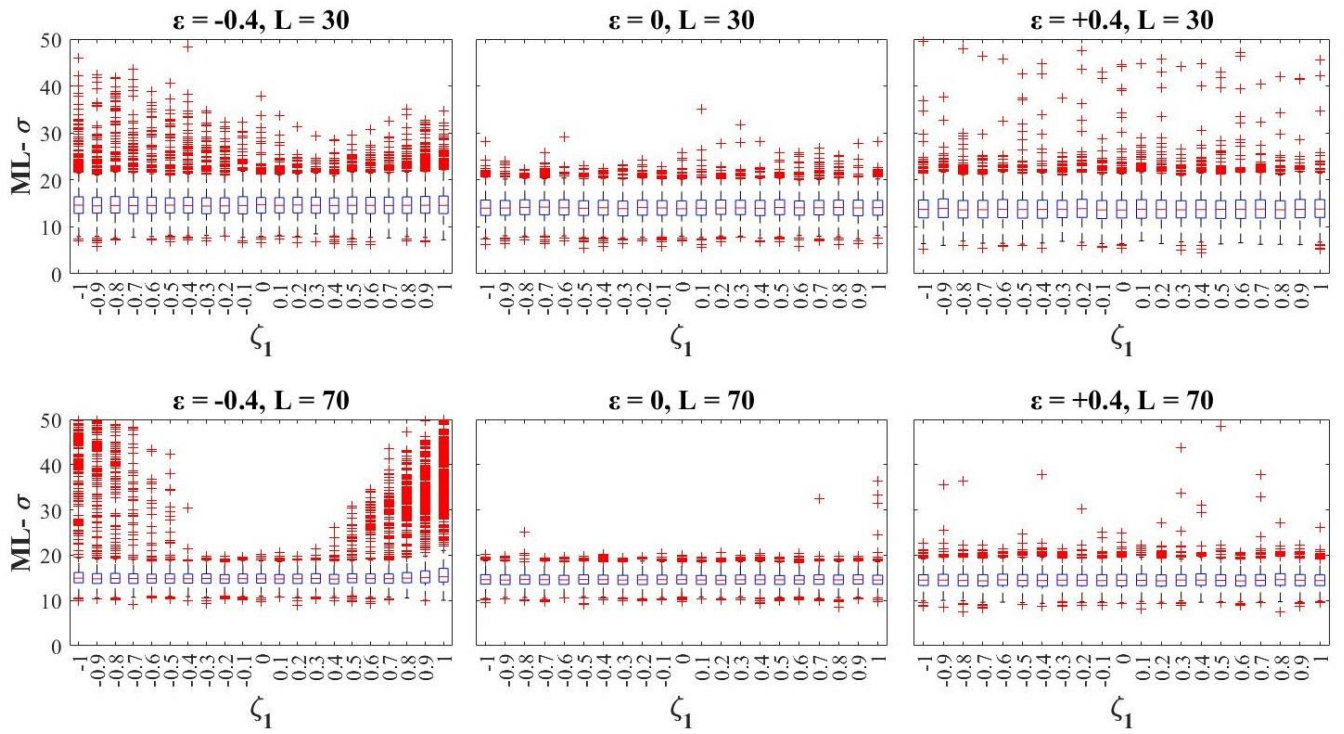



Figure 10: Empirical distributions of ML-σ evaluated for $\sigma = 15$ from samples with $L = 30$ and $L = 70$ vs trend coefficient ζ_1

630