Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-024-02709-z

ORIGINAL PAPER

On the power of popular two-sample tests applied to precipitation and discharge series

Giuseppe Mascaro[1,2]

Accepted: 12 March 2024

Abstract
Two-sample tests are widely used in hydrologic and climate studies to investigate whether two samples of a variable of interest could be considered drawn from different populations. Despite this, the information on the power (i.e., the probability of correctly rejecting the null hypothesis) of these tests applied to hydroclimatic variables is limited. Here, this need is addressed considering four popular two-sample tests applied to daily and extreme precipitation, and annual peak flow series. The chosen tests assess differences in location (t-Student and Wilcoxon) and distribution (Kolmogorov–Smirnov and likelihood-ratio). The power was quantified through Monte Carlo simulations relying on pairs of realistic samples of the three variables with equal size, generated with a procedure based on suitable parametric distributions and copulas. After showing that differences in sample skewness are monotonically related to differences in spread, power surfaces were built as a function of the relative changes in location and spread of the samples and utilized to interpret three case studies comparing samples of observed precipitation and discharge series in the U.S. It was found that (1) the t-Student applied to the log-transformed samples has the same power as the Wilcoxon test; (2) location (distribution) tests perform better than distribution (location) tests for small (moderate-to-large) differences in spread and skewness; (3) the power is relatively lower (higher) if the differences in location and spread or skewness have concordant (discordant) sign; and (4) the power increases with the sample size but could be quite low for tests applied to extreme precipitation and discharge records that are commonly short. This work provides useful recommendations for selecting and interpreting two-sample tests in a broad range of hydroclimatic applications.

Keywords Two-sample tests · Test power · Precipitation · Discharge

1 Introduction

Two-sample tests are used to investigate whether there is sufficient statistical evidence to claim that two samples of a variable could be considered random and independent realizations of the same population or if they might have been drawn from different populations. To address this general question, several tests have been designed that quantify differences in either location (e.g., Student 1908; Kruskal 1957), spread (e.g., Bartlett 1937), or the entire distribution (e.g., Massey 1951) of the two analyzed samples. Two-sample tests could be parametric, thus relying on specific assumptions on the underlying populations, or non-parametric, thus being potentially distribution-free (Wilks 2011), although their performance might be affected by the specific form of the underlying population (Totaro et al. 2020).

Since the problem targeted by two-sample tests is quite general, these statistical tools have been utilized for a large variety of research and practical applications across most disciplines. In hydrologic and climate studies involving precipitation, two-sample tests have been used, for instance, to separate storm types and inform flood prediction (Knighton and Walter 2016); evaluate whether the distributions of daily precipitation from different gridded products (e.g., from interpolated ground observations, reanalysis, or climate model outputs at different resolutions) could be considered the same (e.g., Orskaug et al. 2011; Rauscher et al. 2016; Thober and Samaniego 2014); estimate the shortest duration that characterizes the internal climate variability of daily precipitation (Schindler et al. 2015); investigate changes in the distribution of daily and extreme precipitation associated

✉ Giuseppe Mascaro
gmascaro@asu.edu

[1] School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, USA

[2] Center for Hydrologic Innovations, Arizona State University, Tempe, AZ, USA

Published online: 30 March 2024

with different climate regimes or synoptic circulation patterns (e.g., Shen et al. 2016; Chu et al. 2010; Sýkorová and Huth 2020); detect trends in extreme precipitation due to climate change (Shao et al. 2015; Chu et al. 2010; Xu et al. 2003; Park et al. 2011; Roth et al. 2012; Beguería et al. 2011); identify spatial heterogeneities in the distribution of precipitation extremes (Chu et al. 2010); and quality control rain gage records (Sugahara et al. 2015), among other goals. Hydrologic studies have also applied two-sample tests to investigate changes in the distributions of variables related to discharge records (e.g., Angelina et al. 2015; Knoben et al. 2018).

A key piece of information that is needed to select the right test and correctly interpret its results is the statistical power, which is the probability of correctly rejecting the null hypothesis H_0 (e.g., the two samples belong to the same population or have the same location) when it is false. The power of two-sample tests has been investigated since the 1970s by theoretical statisticians (e.g., Feltovich 2003; Freidlin and Gastwirth 2000; Baumgartner and Kolassa 2021; Lee et al. 1975; O'Gorman 1995) and for applications in medical research (e.g., Fagerland and Sandvik 2009; Fagerland 2012), psychology (e.g., van den Brink and van den Brink 1989), biology (e.g., Collings and Hamilton 1988), economics (e.g., Feltovich 2003), social sciences (e.g., Penfield 1994), and computer sciences (e.g., Gretton et al. 2012), among other disciplines. These efforts highlighted that the test performance depends on different factors related to the sample size and the shape (mainly, spread and skewness) of the population and null distributions. A good synthesis of what many of these studies suggested is provided by Fagerland and Sandvik (2009), who concluded that *"simple rules about which test should be used in which situation cannot be accurately stated"*. As a result, to properly quantify the performance of two-sample tests for a given application, it is desirable to carry out analyses that target specific variables and analyze the properties of the data at hand.

In applications to hydrologic and climate variables, a few studies have investigated the power of trend tests (e.g., Prosdocimi et al. 2014; Vogel et al. 2013; Amorim and Villarini 2023), while, somewhat surprisingly, the power of two-sample tests has received less attention. This work addresses such relatively straightforward and yet critical need by quantifying the power of four popular two-sample tests applied to three hydroclimatic variables, including non-zero precipitation (NZP), annual precipitation maxima (APM), and annual peak flows (APF). The analyzed tests are (1) the Student *t* (Student 1908) and (2) Wilcoxon rank-sum (Kruskal 1957) tests, which assess differences in location (hereafter, location tests); and (3) the Kolmogorov–Smirnov (Massey 1951) and (4) likelihood-ratio (Wilks 2011; LR) tests, which evaluate differences in the entire distribution (hereafter, distribution tests). The tests' power was quantified through Monte Carlo simulations where realistic samples of NZP, APM, and APF series of equal size were generated as variates of parametric probability distributions that were shown to adequately model these variables across the world. Since the distribution parameters exhibit correlation, a novel method based on copulas, fitted on observed precipitation and discharge series in the United States (U.S.), was designed to generate sets of parameter values from realistic ranges that preserve such correlation. After showing that the differences in spread and skewness of the samples are monotonically related, the tests' power computed through the Monte Carlo simulations was first summarized through surfaces as a function of the differences in mean and spread for different sample sizes. The insights gained from the power surfaces were then used to interpret the tests' outcomes in three case studies in the U.S. aimed to quantify differences in the observed seasonal distributions of precipitation and discharge series and of precipitation distributions before and after ~ 1980–1990 likely due to climate change. Finally, the analyses of the Monte Carlo simulations and the case studies were utilized to develop recommendations for selecting two-sample tests and interpreting their results in hydrologic and climate applications based on precipitation and discharge series.

2 Methods

2.1 Overview of two-sample statistical tests

We first provide some basic definitions of hypothesis testing that are useful to properly understand the meaning and utility of the test power. Hypothesis testing is the process of statistical inference where limited data samples are analyzed to draw conclusions about the properties of the underlying population/s (Wilks 2011). This process is performed through statistical tests, which are all based on the definition of (1) the test statistic: a metric considered appropriate for the analyzed inferential problem; (2) the null hypothesis, H_0: a logical statement used as a reference for the test statistic; (3) the alternative hypothesis, H_A: another logical statement alternative to H_0; and (4) the null distribution: the sampling distribution for the test statistic if H_0 is true. Once the test statistic is computed from the available samples, its value is used to calculate the *p*-value from the null distribution, i.e., the probability that the test statistic assumes values that are against H_0. To draw a binary conclusion about the test outcome, the significance level, α, is defined (e.g., $\alpha = 0.01$ or 0.05) and compared with the *p*-value: if $p \leq \alpha$ ($p > \alpha$), then there is (there is not) enough statistical evidence to reject H_0.

The significance level also represents the Type I error of the test, i.e., the probability of rejecting H_0 when it is actually true. Another type of error is Type II, denoted with β, which is the probability of not rejecting H_0 when it is false; its complement to 1, (1-β), is defined as test power, i.e., the

probability of correctly rejecting H_0 when it is false. Type II errors depend on the definition of H_A and the value of α. To better explain these concepts and statistics, Fig. 1 shows two illustrative examples of tests with different power. In both panels, the null distribution (labeled H_0) of the test statistic is shown in black, while the distributions of two alternative hypotheses are plotted in red ($H_{A,1}$) and blue ($H_{A,2}$). The gray area, α, denotes the region of the null distribution where H_0 is rejected, which is also the assumed Type I error, while the red and blue areas, β_1 and β_2, are the Type II errors under the two alternative hypotheses. In both panels, $\beta_1 < \beta_2$ since the sampling distribution of $H_{A,1}$ is farther from the corresponding null distribution. The test in Fig. 1a is more powerful (i.e., the values of β are smaller and of (1-β) are larger) than the test in Fig. 1b because the spread of its null distribution is smaller.

Here, we assessed the power of four popular two-sample tests applied to series of NZP, APM, and APF, which are widely analyzed in hydroclimatic studies. Let $\boldsymbol{x}_1 = \{x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}\}$ and $\boldsymbol{x}_2 = \{x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}\}$ be two independent samples with size n_1 and n_2, respectively. Broadly speaking, two-sample tests assess the null hypothesis H_0 that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are random and independent realizations of the same population. There are several parametric and non-parametric statistical tests that can be applied to achieve this goal, some of which are designed to assess differences in the entire distribution and others in specific population statistics, with the most popular targeting location and spread. Here, we considered (1) the parametric Student t (t-S) and (2) non-parametric Wilcoxon rank-sum (Wi) location tests, and (3) the non-parametric Kolmogorov–Smirnov (KS) and (4) parametric likelihood-ratio (LR) distribution tests. The definition of the statistics and null distributions of these tests are summarized in Appendix 1. As well known, the t-S test is based on the assumption of normality of the samples, which is hardly met with precipitation and discharge series that are quite often positively skewed. To investigate the impact of this assumption on the test power, the t-S test was applied to the original and the log-transformed samples, with the latter aimed at reducing the skewness and meeting the normality condition. Results for the t-S test were here shown for the case of equal variance, although it was verified that they did not substantially change when the test was applied for unequal variances. Finally, to apply the LR test, we used the parametric distributions described in Section. 2.3.

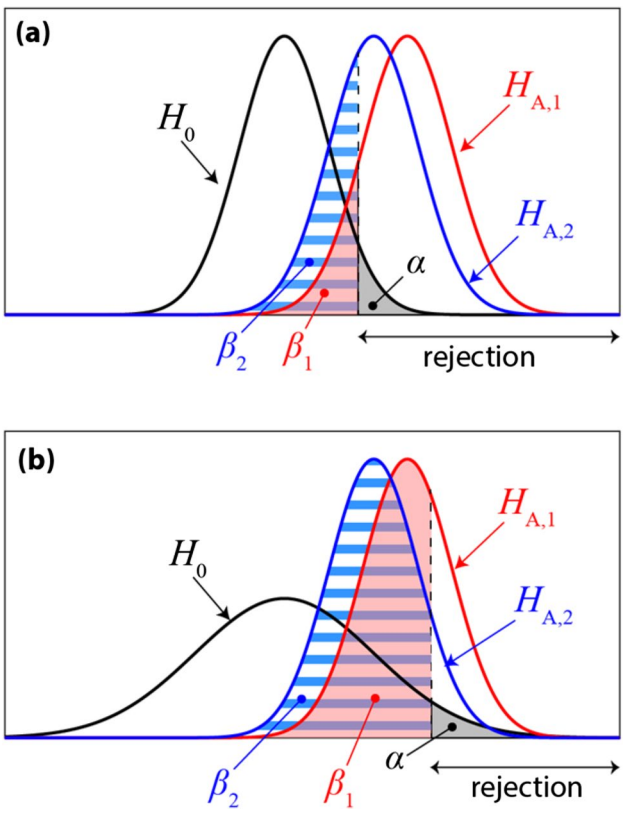

Fig. 1 Illustration of Type I and II errors in hypothesis testing. The panels show the probability distribution functions of the test statistic under the null distribution (H_0) and two alternative hypotheses in red ($H_{A,1}$) and blue ($H_{A,2}$), along with the Type I (α) and II (β_1 and β_2) errors for (**a**) a more and (**b**) a less powerful test. The region of rejection of H_0 is also shown

2.2 Estimation of tests' power with Monte Carlo simulations

The power of the tests applied to NZP, APM, and APF series was quantified through Monte Carlo simulations as a function of metrics accounting for differences in location and spread, and skewness of the samples. Preliminary tasks involved identifying parametric probability distributions that were shown in the literature to well model observed NZP, APM, and APF series; fitting these distributions to observed precipitation and discharge series; and using copulas to model the multivariate distributions of their parameters (see Section. 2.3 for details). For each series type, the Monte Carlo experiments consisted of the following steps:

1. N_{ens} variates with sample size n were generated from the corresponding probability distribution with parameters randomly extracted from the associated copulas. This implies that the power was evaluated for samples with equal size, $n_1 = n_2 = n$.
2. The L-moments (Hosking 1990) of each variate, λ_r, with $r = 1, 2, 3$, were computed and, from these, the L-CV, $\tau = \lambda_2/\lambda_1$, and L-skewness, $\tau_3 = \lambda_3/\lambda_2$. The statistics λ_1 (the sample mean), τ, and τ_3 were used to quantify the location, spread, and skewness of the samples, respectively.
3. For each possible pair of variates (total of $N_{ens} \cdot (N_{ens} - 1)/2$), the p-values of the two-

sample tests were computed, along with the percent difference between the λ_1's and τ's as $\Delta\lambda_1 = (\lambda_{1,1} - \lambda_{1,2})/[(\lambda_{1,1} + \lambda_{1,2})/2] \times 100$ and $\Delta\tau = (\tau_1 - \tau_2)/[(\tau_1 + \tau_2)/2] \times 100$, respectively, and the simple difference between the τ_3's, $\Delta\tau_3 = (\tau_{3,1} - \tau_{3,2})$. In these metrics, $\lambda_{1,1}$ and $\lambda_{1,2}$ (τ_1 and τ_2; $\tau_{3,1}$ and $\tau_{3,2}$) are the values of λ_1 (τ; τ_3) of the two samples.

4. After showing that $\Delta\tau_3$ is monotonically related to $\Delta\tau$ (see Section. 3.1), surfaces of test power for a given α were constructed as a function of $\Delta\lambda_1$ vs. $\Delta\tau$. This involved (a) creating a regular grid of $\Delta\lambda_1$ and $\Delta\tau$ values; (b) finding the number of variate pairs $N_{i,j}$ falling in each grid element (i, j); and (c) computing the fraction of test rejections as $\#(p \le \alpha)/N_{i,j}$, with the constraint that $N_{i,j} > 20$ to have a sufficiently high number of pairs. The resulting value provides an estimate of the power (1-β) in the (i, j) grid point.

The surfaces of test power were also built for the differences in the population *L*-moments with the goals of quantifying the theoretical power of the tests and how this power decreases with the sample size, as well as to partially address the issues related to the post-experiment power computation (e.g., Hoenig and Heisey 2001).

2.3 Datasets and parametric distributions used for the Monte Carlo simulations

To generate realistic variates for the Monte Carlo simulations, we identified suitable parametric probability distributions for observed NZP, APM, and APF series. We used daily precipitation observations from 1499 gages of the Global Historical Climatology Network daily (GHCNd; Menne et al. 2012) in the contiguous U.S. with more than 50 years of data. For each gage, we extracted the NZP and APM series of each season (Winter: DJF; Spring: MAM; Summer: JJA; and Fall: SON). To identify the dominant seasons of NZP and APM required in the first case study (see Section. 3.2), the gages were grouped based on the nine climatic regions defined by the National Centers for Environmental Information (NCEI; Fig. 2a) based on soil moisture anomalies (Karl and Koss 1984) and used in several prior studies on precipitation regimes in the U.S. (Kunkel et al. 2012, 2020). For each region, the two seasons with the largest differences between the gage-averaged λ_1, τ, and τ_3 were found (Fig. 2a). We also used daily discharge records from 672 gages with 20 to 111 years of data representing near-natural streamflow conditions that are part of the U.S. Geological Survey (USGS) Hydro-Climatic Data Network 2009 (HCDN2009; Slack and Landwehr 1992). The gages monitor basins with drainage areas ranging from 2.2 to 25,791 km^2, covering a wide range of climatic conditions in the U.S. To select the dominant seasons needed for the second case study (see Section. 3.3), for each stream gage, we extracted the APF series of the four seasons and, from these, we obtained the two dominant seasons with the largest λ_1's (displayed in Fig. 2b).

Following empirical evidence presented in the literature, we adopted the Burr Type XII ($\mathcal{B}\nabla$XII) distribution to model the NZP series (Papalexiou and Koutsoyiannis 2016; Mascaro et al. 2023; Papalexiou and Koutsoyiannis 2012), the generalized extreme value distribution ($\mathcal{GEV}$) for the APM series (Papalexiou and Koutsoyiannis 2013; Blanchet et al. 2016; Mascaro 2020; Deidda et al. 2021), and Log-Person Type 3 ($\mathcal{LP}3$) to characterize the APF records (Vogel et al. 1993; Griffis and Stedinger 2007; England et al. 2019). The cumulative distribution functions (CDFs) of these three-parameter distributions are

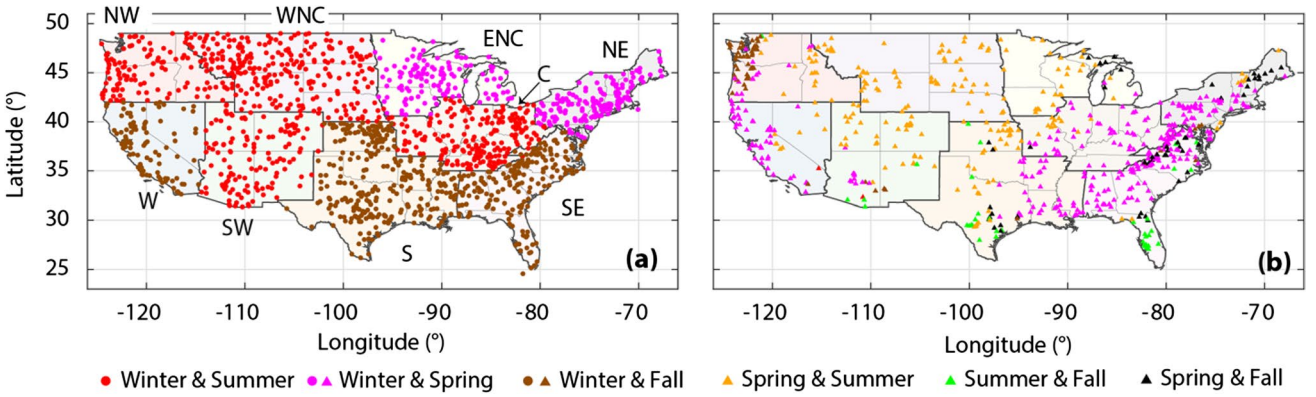

Fig. 2 (**a**) Map of daily rain gages (circles) from the Global Historical Climatology Network daily (GHCNd) in the contiguous United States (U.S.) color-coded based on the two dominant seasons in the NCEI regions defined as NW: Northwest (106 gages); WNC: West North Central (250 gages); ENC: East North Central (124 gages); C: Central (188 gages); NE: Northeast (154 gages); SE: Southeast (181 gages); S: South (272 gages); SW: Southwest (132 gages); and W: West (92 gages). (**b**) Stream gages (triangles) of the U.S. Geological Survey (USGS) Hydro-Climatic Data Network 2009 (HCDN2009) with the indication of the two dominant seasons

provided in Appendix 2. As illustrated in the *L*-moment ratio diagrams of Fig. 3, the empirical *L*-moment ratios (τ and τ_3) of the observed NZP series of the four seasons are largely included within the theoretical surface of the $\mathcal{B}\nabla$XII, while those (τ_3 and *L*-kurtosis, $\tau_4 = \lambda_4/\lambda_2$, with λ_4 being the fourth *L*-moment) of the APM and APF series are scattered around the theoretical lines of the $\mathcal{GEV}$ and $\mathcal{LP}3$, respectively. This suggests that the selected distributions are suitable for modeling the observed precipitation and discharge series. For each seasonal series, we estimated the three parameters of the corresponding parametric model (i.e., $1499 \times 4 = 5996$ sets of $\mathcal{B}\nabla$XII and $\mathcal{GEV}$ parameters, and $672 \times 4 = 2688$ sets of $\mathcal{LP}3$ parameters) and fit copulas to represent their multivariate distribution as described in detail in Appendix 2 and Figs. S1-S3 of the Supplementary Information. The copulas allowed sampling sets of parameters for each distribution whose values are within plausible ranges and preserve their dependence structure; this, in turn, led to realistic samples for the analyzed hydrologic series.

3 Results

3.1 Power of two-sample tests applied to precipitation and discharge series

We first present in Fig. 4 the relationships among $\Delta\lambda_1$, $\Delta\tau$, and $\Delta\tau_3$ for all types of variables for a given samples size, *n*, and the population (results are similar for other *n*'s). Due to the positively skewed nature of the samples, for a fixed $\Delta\lambda_1$, higher values of $\Delta\tau_3$ are associated with larger $\Delta\tau$ and vice versa; in other words, if a sample has larger (smaller)

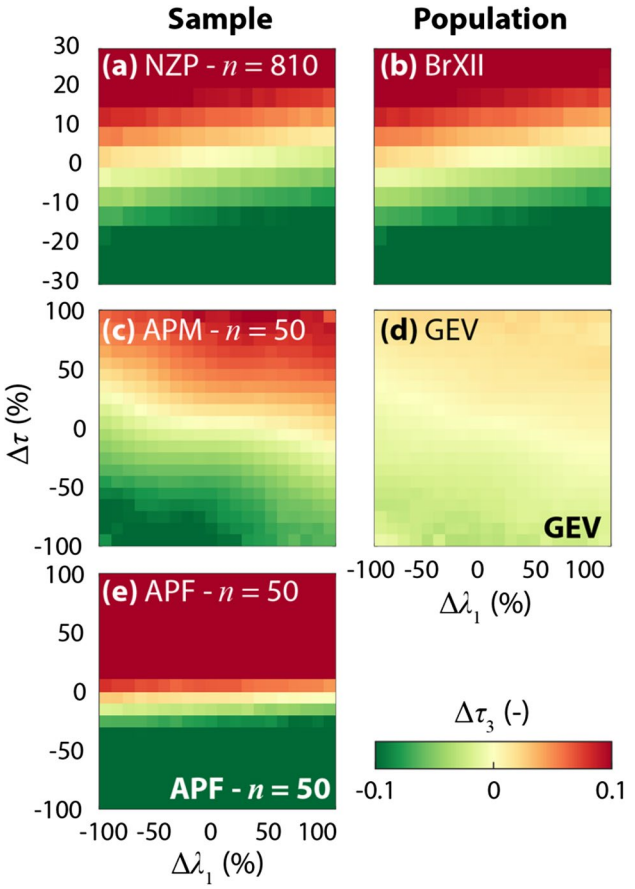

Fig. 4 Relationships among $\Delta\lambda_1$, $\Delta\tau$, and $\Delta\tau_3$ shown as surfaces of $\Delta\tau_3$ as a function of $\Delta\lambda_1$ and $\Delta\tau$ for (**a**) pairs of NZP samples with size $n = 810$ and (**b**) the corresponding $\mathcal{B}\nabla$XII population; (**c**) pairs of APM samples with size $n = 50$ and (**d**) the corresponding $\mathcal{GEV}$ population; and (**e**) pairs of APF samples with $n = 50$. For the latter case, the population is not shown because $\Delta\lambda_1$, $\Delta\tau$, and $\Delta\tau_3$ for the $\mathcal{LP}3$ distribution are referred to the log-transformed samples

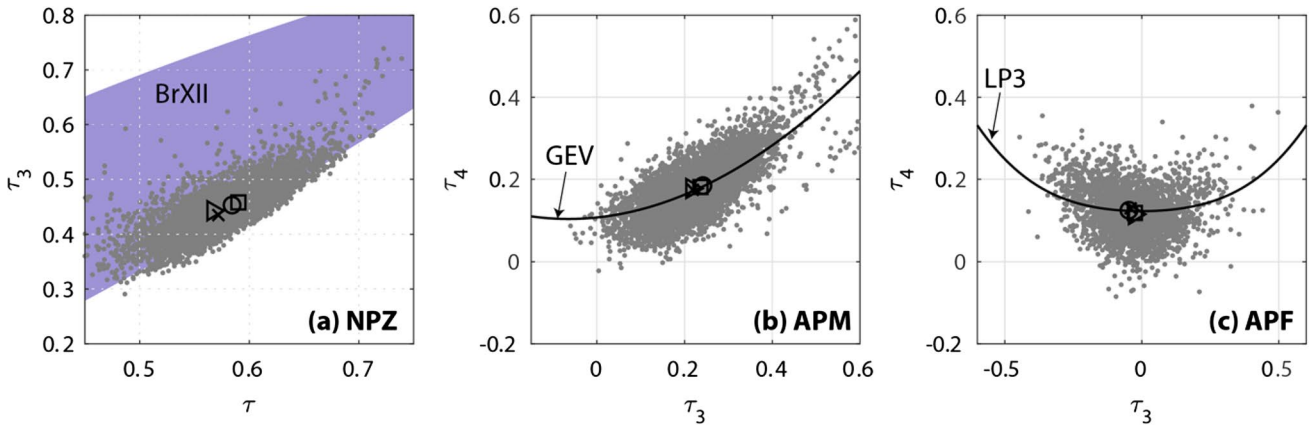

Fig. 3 The *L*-moment ratio diagrams (τ vs. τ_3 and τ_3 vs. τ_4) for (**a**) observed NZP records in the four seasons and theoretical surface of the Burr Type XII ($\mathcal{B}\nabla$XII) distribution, (**b**) observed APM series in the four seasons and theoretical relation for the generalized extreme value ($\mathcal{GEV}$) distribution, and (**c**) the logarithm of the observed APF series in the four seasons and theoretical relations for the Log-Person Type 3 ($\mathcal{LP}3$) distribution. In each panel, the mean of the sampling *L*-moment ratios for each of the four seasons are plotted with different markers to emphasize the suitability of the chosen distributions

skewness than the other one, it has also higher (lower) spread. From the practical standpoint, this implies that the power surfaces could be visualized as a function of either $\Delta\tau$ or $\Delta\tau_3$. Here, we used $\Delta\tau$ in the figures but refer also to the skewness for their interpretation. The surfaces of the power (1-β) of the analyzed tests for the significance level $\alpha = 0.05$ as a function of $\Delta\lambda_1$ and $\Delta\tau$ are shown in Figs. 5, 6, and 7 for the NZP, APM, and APF series, respectively. Different n's were considered ranging from 270 to 1350 rainy days (roughly, 10 to 50 years of seasonal records assuming 30 rainy days in a season) for the NZP series, and from 30 to 100 years for the annual maxima series, APM and APF. The Monte Carlo simulations were based on $N_{ens} = 5000$ variates. Some general considerations can be made:

(1) As expected, the power diminishes as n decreases.
(2) When the samples have very similar skewness and, consequently, close spread ($|\Delta\tau|$ and $|\Delta\tau_3|$ close to 0), the power diminishes as $|\Delta\lambda_1|$ decreases. As differences in skewness and spread become larger (smaller) than 0, the range of $\Delta\lambda_1$ for which the power decreases is shifted towards positive (negative) values.
(3) The power of t-S applied to the original sample is not significantly affected by $\Delta\tau$ (or $\Delta\tau_3$), while the power surfaces of all other tests exhibit an ellipsoidal shape with the major axis extending from southwest to northeast in the $\Delta\lambda_1$-$\Delta\tau$ plane. From the practical standpoint, this means that, if sample 1 has a higher location than sample 2, the test power is relatively smaller if sample 1 has also larger spread (and skewness), whereas the power is relatively larger if sample 1 has lower spread (and skewness) than sample 2 (and vice versa, since the order of the samples is irrelevant). To explain the reasons for this outcome, examples of $\mathcal{GEV}$ probability density functions where $\Delta\tau$ and $\Delta\tau_3$ have the same sign which is either discordant or concordant with the sign of the $\Delta\lambda_1$ (assumed constant) are presented in Fig. S4. When the signs are concordant, the modes of the two populations are similar and most of their density is concentrated within similar ranges; as a result, samples drawn from these populations tend to be similar leading to a lower test power. In contrast, when the sign of $\Delta\tau$ and $\Delta\tau_3$ is discordant with that of $\Delta\lambda_1$, the density of population 2 is distributed around a much larger range of values than the density of population 1; this likely results in samples with different statistics and, thus, in higher test power.
(4) Applying t-S with the log-transformed sample (t-S–Log) leads to power surfaces, plotted as a function of $\Delta\lambda_1$ and $\Delta\tau$ computed for the original samples, which

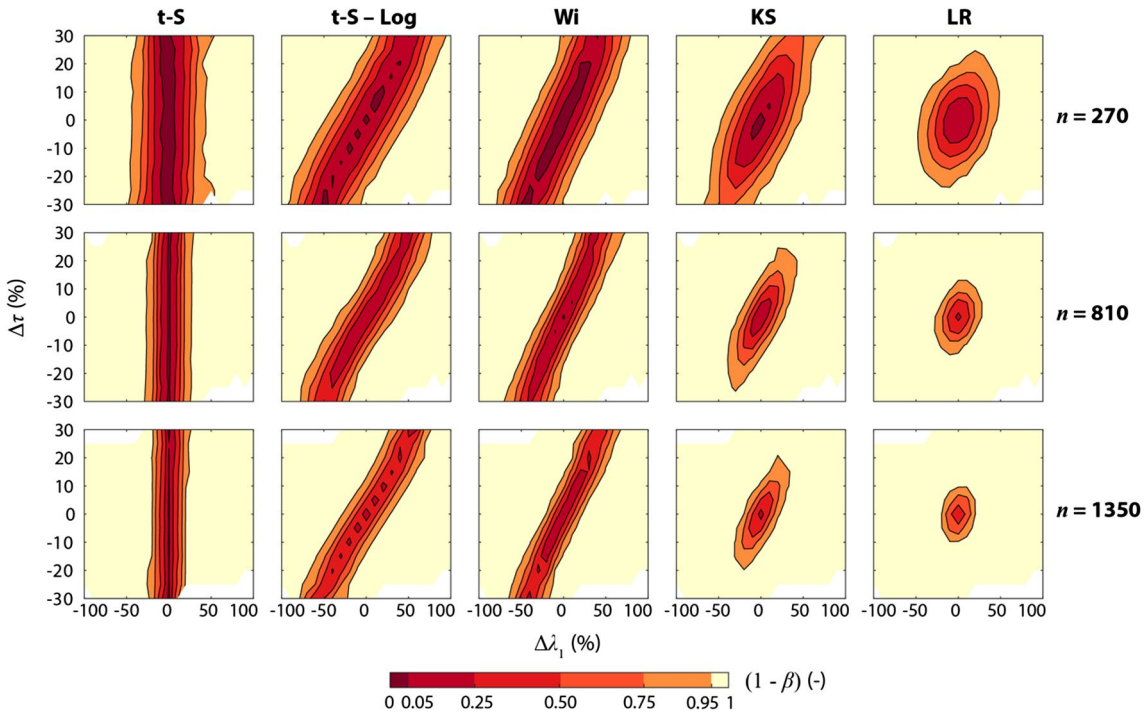

Fig. 5 Surfaces of test power (1-β) for the significance level $\alpha = 0.05$ as a function of $\Delta\lambda_1$ and $\Delta\tau$ of NZP series for the t-Student, Wilcoxon (Wi), Kolmogorov–Smirnov (KS), and likelihood-ratio (LR) tests. The t-Student test was applied to the original (t-S) and log-transformed (t-S–Log) samples. The rows refer to different sample sizes, n. The surfaces were built through the Monte Carlo experiments described in Section. 2.2 with $N_{ens} = 5{,}000$ variates

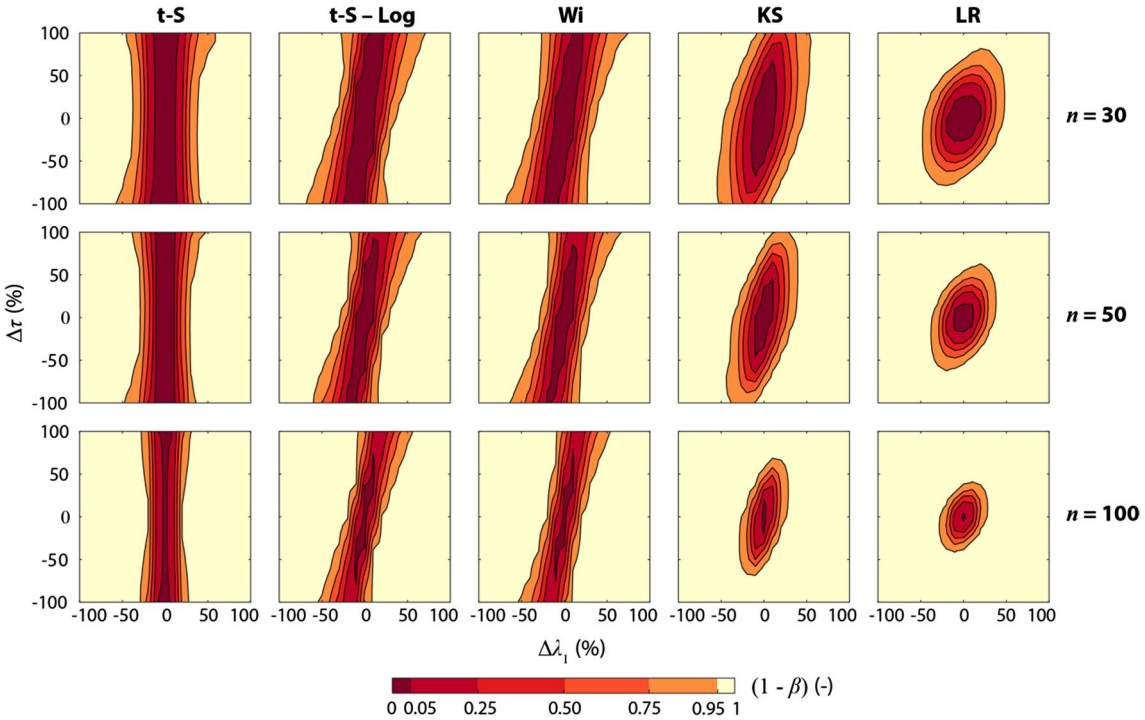

Fig. 6 Same as in Fig. 5, but for the APM series

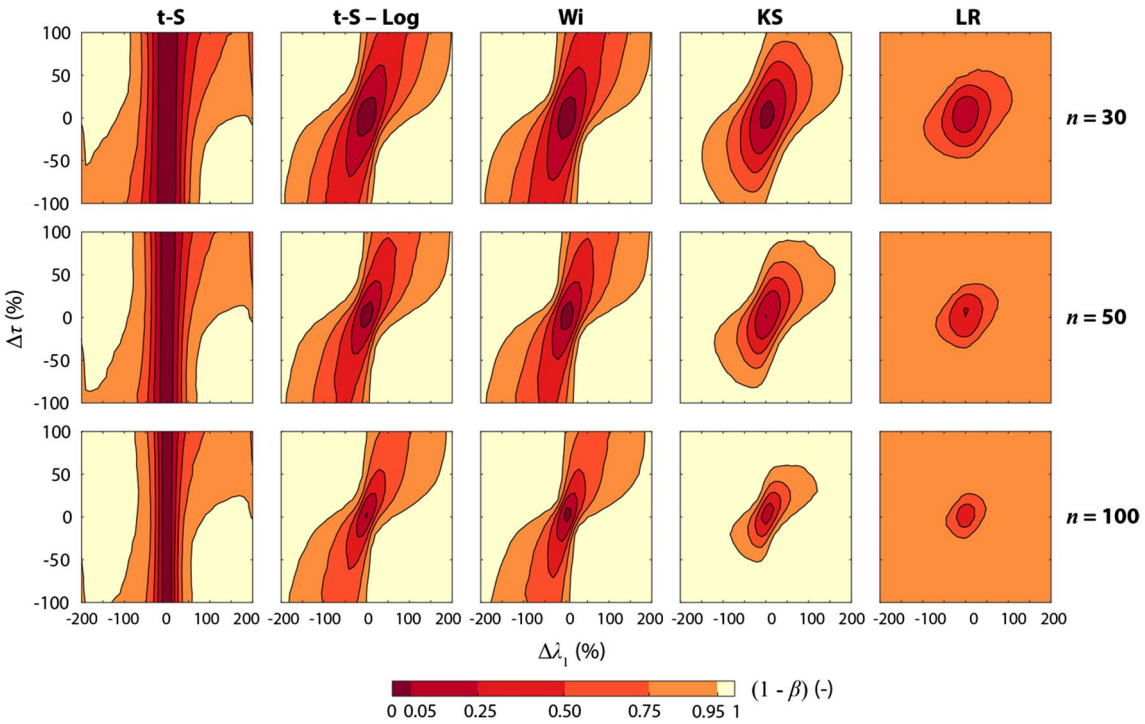

Fig. 7 Same as in Fig. 5, but for the APF series

are remarkably similar to those obtained for the non-parametric Wi test.

(5) The location tests are more powerful than the distribution tests when $|\Delta\tau|$ is relatively low (especially for APM and APF series), while the opposite is true as $|\Delta\tau|$ increases (the considerations made for $|\Delta\tau|$ hold also for $|\Delta\tau_3|$). LR is in most cases the most powerful test, as also depicted in Figs. S5-S7 that present the surfaces of the difference between the power of each test and LR. However, this result is in part biased because LR was applied with the same parametric distributions used to generate the variates in the Monte Carlo simulations, i.e., the test is more "informed" about the underlying populations. However, the case studies based on observed data described in Sections 3.2–3.4 suggest that the choice of the parametric distribution appears appropriate because the outcomes of the LR test are well in line with the test power and supported by physical evidence.

To quantitively complement the visual inspection of Figs. 5–7, Fig. 8 displays the relationships between test power and one of $\Delta\lambda_1$, $\Delta\tau$, and n for fixed values of the other two metrics. The chosen fixed value of n is 30 for the APM and APF series and 810 (~30 years) for the NZP records. If the samples have the same spread ($\Delta\tau = 0\%$, panels in the first column on the left), the test power varies with changes in location, $\Delta\lambda_1$, according to a reversed bell-shaped function with a minimum value close to α (here, 0.05) at $\Delta\lambda_1 = 0\%$. Two-sample tests applied to NZP series reach a high power (> 0.75) for $|\Delta\lambda_1| > 25\%$, while this threshold for $|\Delta\lambda_1|$ increases to 75% and 100% for the annual maxima of precipitation, APM, and discharge, APF, respectively. For these two series of extremes, location tests are more powerful than distribution tests when there is no difference in spread ($\Delta\tau = 0\%$). On the other hand, if the spreads of the two samples are different ($\Delta\tau > 0\%$; panels in the second column), the impacts on the test power depend on the series. For the NZP records and $\Delta\tau = 10\%$, the reversed bell-shaped functions of all tests except for t-S shift

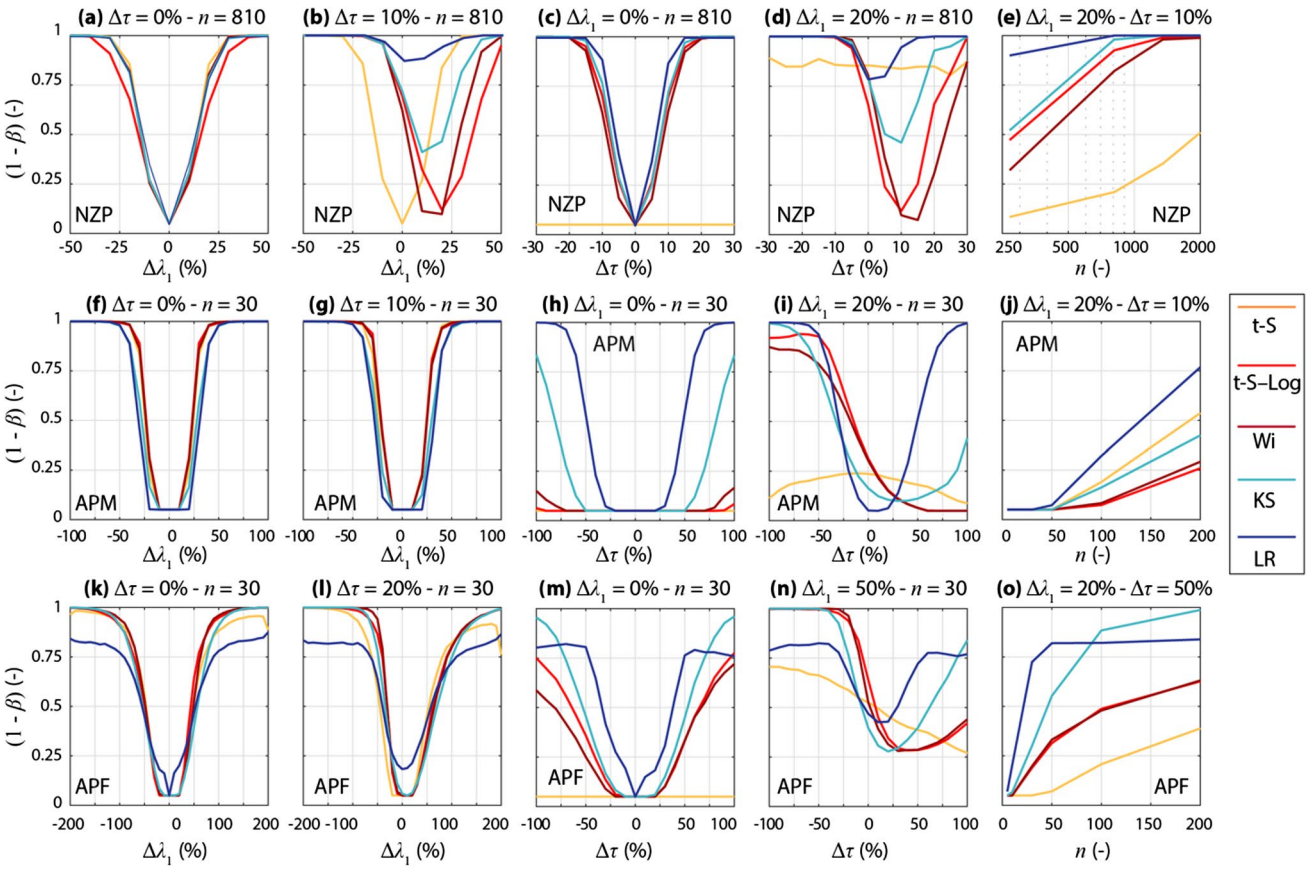

Fig. 8 Relationships between the power $(1 - \beta)$ of the analyzed tests and (1) $\Delta\lambda_1$ for fixed $\Delta\tau$ and n (panels in the first two columns on the left), (2) $\Delta\tau$ for fixed $\Delta\lambda_1$ and n (panels in the third and fourth columns), and (3) n for fixed $\Delta\lambda_1$ and $\Delta\tau$ (panels in the fifth column). Panels (**a**)-(**e**) show results for the NZP series, (**f**)-(**j**) for the APM series, and (**k**)-(**o**) for the APF records. Note that different intervals are used in the x-axes to better visualize the differences across the tests

towards positive $\Delta\lambda_1$ (the region with the same sign as $\Delta\tau$), and their minimum power values increase for the distribution tests, which are more powerful than the location tests. For the APM (APF) series, differences in spread of 10% (20%) do not substantially modify the outcomes reported for $\Delta\tau = 0\%$.

The relations between test power and $\Delta\tau$ for samples with the same location ($\Delta\lambda_1 = 0\%$; panels in the third column) are symmetric. In these conditions, the tests assessing differences in the distribution (notably, LR) have the best ability to correctly reject H_0. The t-S test applied to the original sample does not have any power (i.e., $(1 - \beta) \cong \alpha$), while t-S–Log and Wi exhibit a power that is relatively high for the NZP series, moderate for the APF records, and extremely low for the APM series. The presence of differences in location ($\Delta\lambda_1 > 0\%$; panels in the fourth column) introduces asymmetries and shifts in the relations between test power and $\Delta\tau$ in ways that vary with the type of series and test. When compared to the cases for $\Delta\lambda_1 = 0\%$, the distribution tests become more powerful for most values of $\Delta\tau$, while the location tests gain power in the region where $\Delta\tau$ and $\Delta\tau_1$ have opposite sign (here, $\Delta\tau < 0$) and lose it where their sign is concordant. Finally, the effect of the sample size, n, on the power for fixed $\Delta\lambda_1 > 0$ and $\Delta\tau > 0$ is presented in the panels of the fifth column. For the cases shown in Fig. 8, the power of the distribution tests (particularly, LR) increases faster than the location tests, although it remains relatively low for the APM series even for $n = 200$ years. Interestingly, for this variable, the t-S test applied to the original sample performs similarly to K-S. It is important to emphasize again that the considerations made for $\Delta\tau$ in the interpretation of Fig. 8 are also qualitatively valid for $\Delta\tau_3$ given the monotonic relationship between these two metrics (Fig. 4).

To complete the assessment of the tests' power, the power surfaces were also computed as a function of $\Delta\lambda_1$ and $\Delta\tau$ obtained from the population L-moments, i.e., the L-moments of the $\mathcal{B}\nabla\mathrm{XII}$, $\mathcal{GEV}$, and $\mathcal{LP}3$ distributions. The differences between these power surfaces and those derived as a function of the sample $\Delta\lambda_1$ and $\Delta\tau$ are shown in Figs. S8-S10. Note that, since the $\mathcal{LP}3$ distribution is fitted to the log-transformed samples, the metrics $\Delta\lambda_1$ and $\Delta\tau$ in Fig. S7 are referred to the log-transformed variables. The differences in power are both positive and negative and do not exceed |0.25| (|0.1|) for NZP and APM (APF). The largest differences are found in regions of relatively low $\Delta\lambda_1$, especially where $\Delta\lambda_1$ and $\Delta\tau$ have discordant signs. As expected, the differences in power decrease as the sample size increases.

3.2 Case study 1: Two-sample tests applied to observed seasonal precipitation series

The practical importance of the insights gained through the Monte Carlo experiments is first demonstrated by investigating the null hypothesis H_0 that seasonal observations of NZP and APM series belong to the same population. The rain gages and corresponding dominant seasons are shown in Fig. 2. For each gage and series type, the values of $\Delta\lambda_1$ and $\Delta\tau$ between the two seasonal samples were calculated, along with the p-values of all tests. For each test, the field significance was also accounted for by applying the false discovery ratio test (Wilks 2006, 2016) with a global significance level $\alpha_{\mathrm{global}} = 0.10$ as in Farris et al. (2021) to account for the spatial dependence among the gage records. Results for the NZP and APM records are presented in Figs. 9 and 10, respectively. For each of the climate zone shown in Fig. 2a, the power surfaces of the Wi (Figs. 9a and 10a) and LR (Figs. 9b and 10b) tests are displayed along with the empirical ($\Delta\lambda_1$, $\Delta\tau$) points, plotted in black (green) if H_0 was rejected (not rejected). For simplicity, the power surface was built assuming the same size for both seasonal samples, which was set equal to the mean record length across all gages of each region. This resulted in n ranging from 810 to 1800 for the NZP series, depending on record length and number of rainy days, and in a constant $n = 50$ years for the APM series.

First, we note that, for both series types in the NW, WNC, and ENC regions, H_0 was rejected at practically all gages; this outcome is supported by large statistical evidence because the empirical ($\Delta\lambda_1$, $\Delta\tau$) estimates lie in an area of high power for both tests (i.e., the probability of correctly rejecting H_0 is high). In the other climate regions, H_0 is rejected only at some of the gages but the outcomes differ depending on the type of series and test. For example, in NE, the NZP series exhibit two clusters of gages where H_0 is rejected (not rejected) in a region of high (low) power for both the location and distribution tests (Fig. 9). Note that, in this region, the use of t-S applied to the original series would have failed to reject H_0 at most gages (not shown); however, in this case, the use of t-S applied in this way should be avoided because almost all empirical ($\Delta\lambda_1$, $\Delta\tau$) points are in a region of very low power for this test. The two clusters do not instead emerge when analyzing the APM records, where H_0 cannot be rejected at practically all gages by both tests (Fig. 10). This outcome should be interpreted considering that most empirical ($\Delta\lambda_1$, $\Delta\tau$) estimates lie along the southwest-northeast axis where the test power is lower, thus revealing the limited ability of both tests to investigate H_0 for samples of extreme P in the NE region.

Results for the SE region are instead useful to demonstrate two important issues: (1) depending on the location of the ($\Delta\lambda_1$, $\Delta\tau$) estimates, tests with different power can lead to diverse outcomes; and (2) when a test is applied at multiple sites, the inspection of the geographical patterns of its outcomes provides a straightforward, yet critical, physical support of the statistical analyses, provided that the spatial correlation among the station records has been taken into

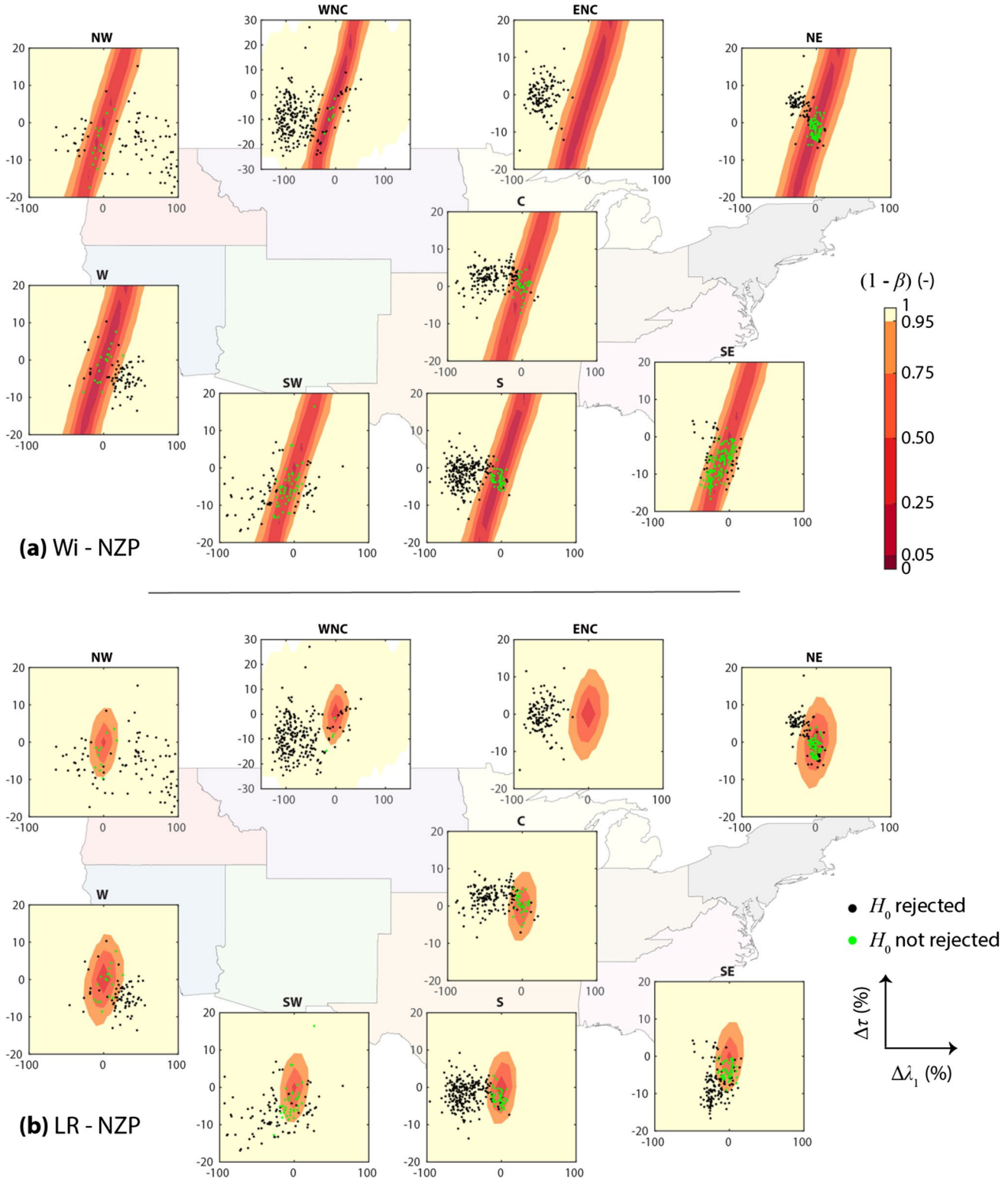

Fig. 9 Results of the (**a**) Wilcoxon (Wi) and (**b**) likelihood-ratio (LR) tests applied to the NZP series of the dominant seasons in the nine climate regions, plotted in the ($\Delta\lambda_1$, $\Delta\tau$) space along with the power surface of each test

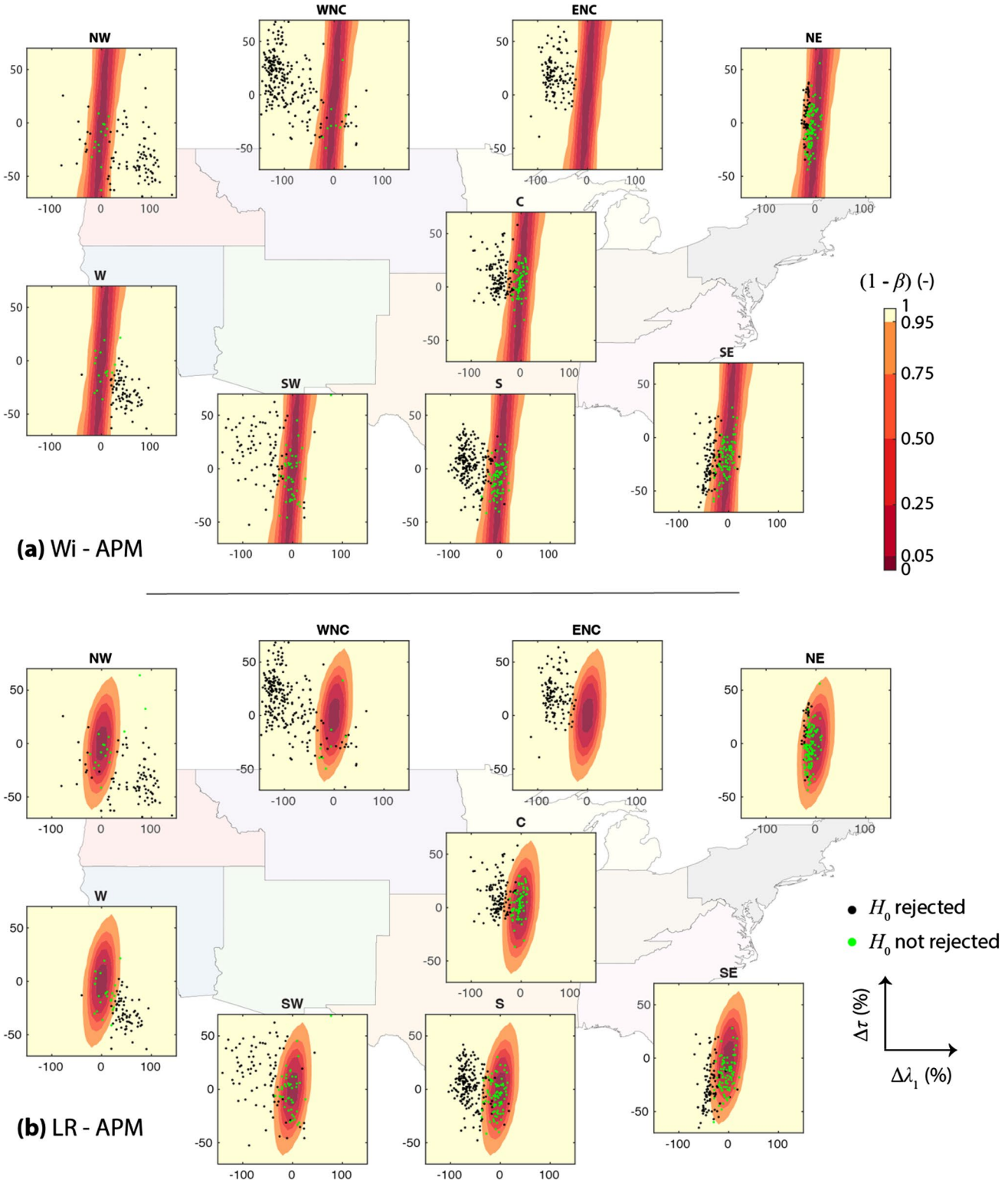

Fig. 10 Same as in Fig. 9, but for the APM series

account (here, via the significance level of the FDR test). To prove these points, Fig. 11 displays the maps of the power and rejections of H_0 at the gages of the SE region for the Wi location test (results are similar for t-S–Log) and for the K-S and LR distribution tests. When applied to the NZP seasonal samples (Fig. 11a), Wi indicates that H_0 cannot be rejected

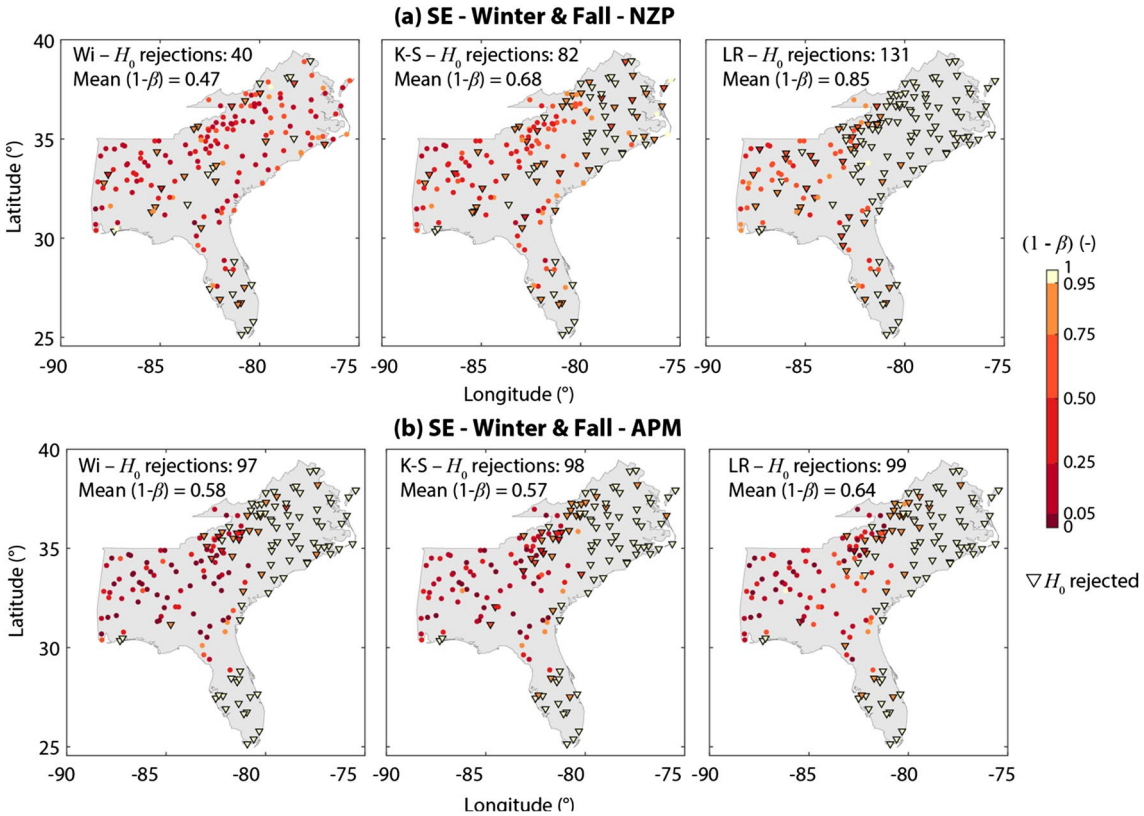

Fig. 11 Maps of power and outcomes of the Wilcoxon (Wi), Kolmogorov–Smirnov (K-S), and likelihood-ratio (LR) tests applied to the (**a**) NZP and (**b**) APM records of the two dominant seasons observed at the rain gages in the SE climate region. The number of H_0 rejections and mean test power ($1-\beta$) are reported in each panel. The total number of gages is 181

at most gages apart from some random locations. If these conclusions are accepted without taking into account the low power of this test, one could end up making wrong physical interpretations, especially when attempting to explain the patterns of the H_0 rejections. On the other hand, the use of distribution tests with higher power suggests that H_0 can instead be rejected at a much larger number of sites; such number of rejections is higher for LR which is also more powerful than K-S. The spatial variability of the H_0 rejections is consistent across the distribution tests and exhibit a clear geographical pattern likely controlled by the distance from the coast, thus providing physical support to the outcomes of the statistical analyses. When the two-sample tests are instead applied to the APM series (Fig. 11b), all tests have similar power and lead to comparable spatial patterns of the H_0 rejections which are quite close to those found for the NZP records.

3.3 Case study 2: Two-sample tests applied to observed seasonal discharge series

As a second case study, we tested H_0 for the two dominant seasons of the APF series observed at the stream gages of Fig. 2b. Results are summarized in Fig. 12 which displays the empirical ($\Delta\lambda_1$, $\Delta\tau$) estimates, color-coded based on the rejection of H_0, on top of the power surfaces of the Wi and LR tests built for the average sample size, $n=50$ years (note that $n<50$ for 35% of the gages). When applied to the observed daily peak flow records, no test is clearly the most powerful across all cases. There is not enough evidence to reject H_0 at a percentage of gages that ranges from 57 to 62% across all tests and dominant seasons. However, since almost all these cases occur in areas of low power ($|\Delta\lambda_1|\leq 50\%$ and $|\Delta\tau|\leq 30\%$), it should be also noted that the tests adopted here have restricted capacity to investigate H_0. The rejections of H_0 occur instead in regions of higher test power where $\Delta\lambda_1$ is relatively large and its sign is discordant with that of $\Delta\tau$ (see, e.g., cases in the fourth quadrant for Winter & Fall and Spring & Summer).

To visualize the locations of the H_0 rejections, Fig. 13 presents the maps of power and outcomes of the two tests for Spring & Summer. For these seasons, results are very similar across location and distribution tests as shown by the close number of H_0 rejections in each climate zone. The gages where H_0 is rejected are clustered in space, which gives physical evidence that

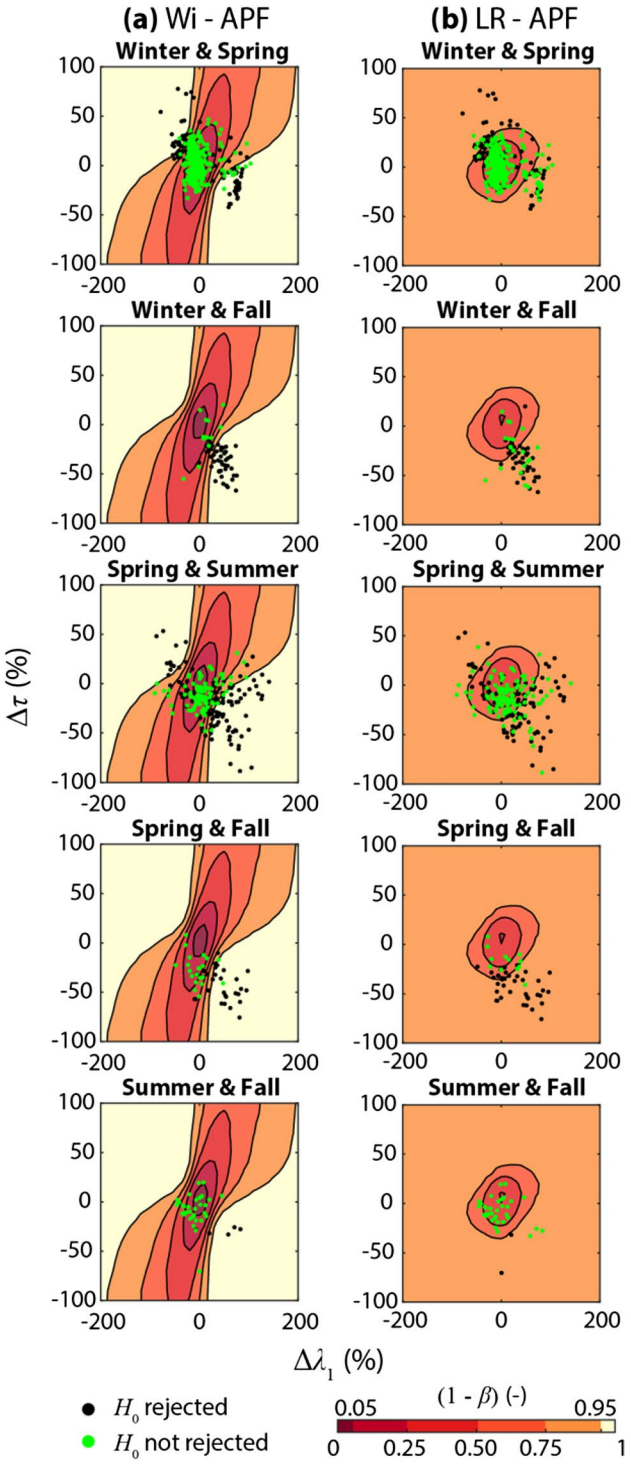

Fig. 12 Results of the (**a**) Wilcoxon (t-S) and (**b**) likelihood-ratio (LR) tests applied to the APF series of the dominant seasons recorded at the stream gages shown in Fig. 2b, plotted in the ($\Delta\lambda_1$, $\Delta\tau$) space along with the power surface of each test

supports the statistical outcomes since differences in the seasonal peak flow regimes at close gages are expected to be determined by local climatic and physiographic features. The findings for Spring & Fall, reported in Fig. 14, permit emphasizing that, when t-S is applied to the original skewed sample, its power might be very low (here, mean of 0.41) if $|\Delta\lambda_1|$ is small even if $|\Delta\tau|$ is not negligible; as a result, t-S fails to reject H_0 in a cluster of stream gages located in the SE and S regions. If t-S is instead applied to the log-transformed sample, its power is much higher (mean of 0.69) because the test is now sensitive to differences in spread, and comparable to that of the distribution tests (mean of 0.73 for both K-S and LR). For these two more powerful tests, H_0 is rejected at essentially the same sites.

3.4 Case study 3: Two-sample tests applied to assess climate change on observed precipitation series

The third case study was inspired by the experiment proposed by Wilks (2011; Example 5.6) where the LR test was applied to evaluate whether the first and second half of a precipitation record are drawn from different populations as a possible consequence of climate change. Here, this experiment was performed separately for the NZP and APM series of the rain gages of Fig. 2. The resulting series were split in the middle of their corresponding records, which largely falls between 1980 and 1990, leading to two samples of about 30 years. Results for the LR test are displayed in Fig. 15 (the outcomes are very similar for the other three tests). For the NZP series, H_0 is rejected at about two-thirds of the sites (Fig. 15a), a result supported by the corresponding empirical ($\Delta\lambda_1$, $\Delta\tau$) estimates being located in the region of high test power in the second and fourth quadrants. Despite the high test power, when such rejections are plotted spatially, no organized pattern emerges (Fig. 15b), thus complicating the physical interpretation of the test outcomes. While further investigations that are out of the scope of this paper would be required to better explain this finding via, e.g., trend or change-point tests, here we can emphasize that failure to reject H_0 at the sites where the power is low does not necessarily imply that the distribution of NZP has not changed over time, but it could also be a reflection of the relatively small difference between the two samples that the LR test is not able to detect because of its low power. Focusing on the APM series, H_0 cannot be rejected at practically all gages (Figs. 15c,d). Again, although there is no significant statistical evidence to assume that the distribution of extreme precipitation has varied in time, the empirical ($\Delta\lambda_1$, $\Delta\tau$) points are in a region of very low test power and very high type-II error β, i.e., the probability of not rejecting H_0 when it is actually false is large. Therefore, it might still be possible that the differences between the samples are too small to be detected by the statistical tests considered here, especially given the small sample size ($n = 30$).

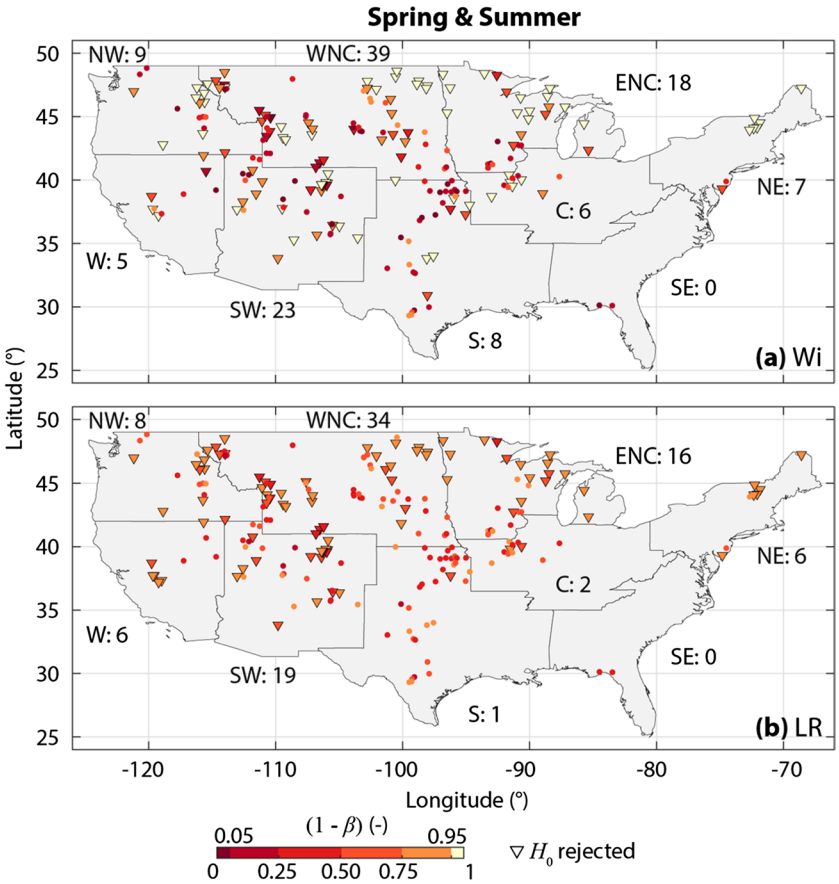

Fig. 13 Maps of power and outcomes of the (**a**) Wilcoxon (Wi; mean power across all gages of 0.57) and (**b**) likelihood-ratio (LR; mean power of 0.61) tests applied to the APF records in Spring & Summer observed at the stream gages of Fig. 2b. The rejection of H_0 after accounting for field significance is indicated with a triangle; their number is also reported for each climate region

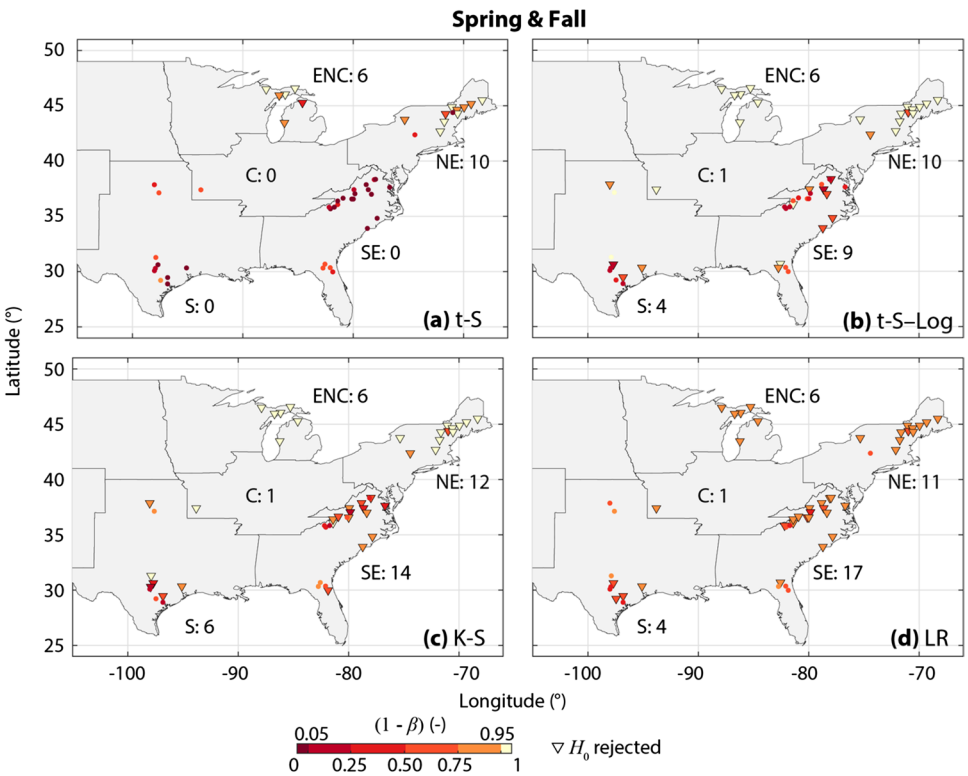

Fig. 14 Same as in Fig. 13 but for the Spring & Fall seasons. The tests shown are the t-Student applied to the (**a**) original (t-S; mean power of 0.41) and (**b**) log-transformed (t-S – Log; mean power of 0.69) samples, (**c**) Kolmogorov–Smirnov (K-S; mean power of 0.73), and (**d**) likelihood-ratio (LR; mean power of 0.73)

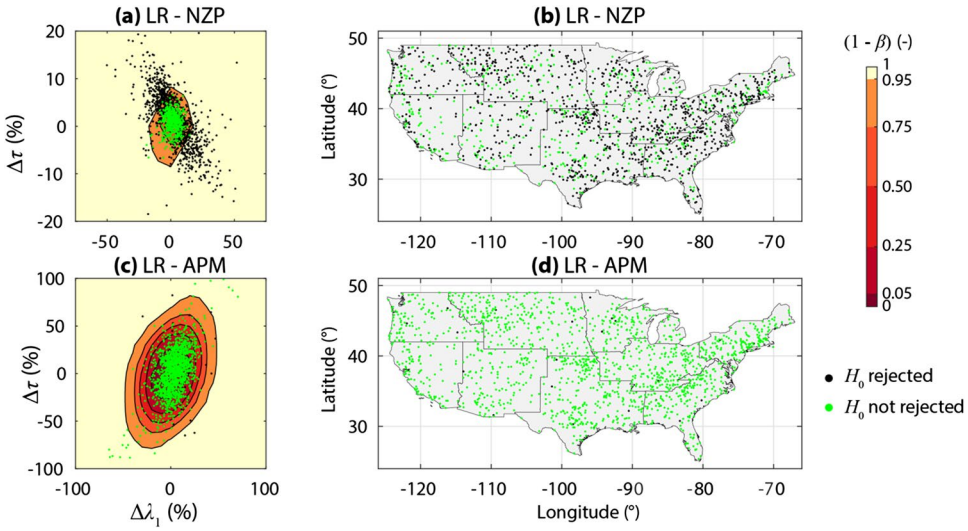

Fig. 15 Results of the likelihood-ratio (LR) test applied to the first and second halves of the (**a**)-(**b**) NZP and (**c**)-(d) APF records observed at the rain gages shown in Fig. 2a. The rejections and non-rejection of H_0 are plotted in the ($\Delta\lambda_1$, $\Delta\tau$) space along with the power surface (left panels) and as maps (right panels)

4 Summary and discussion

In this work, the power of four popular two-sample tests assessing differences in location and distribution was evaluated, to our knowledge for the first time, for daily (NZP) and extreme (APM) precipitation series, and for extreme discharge records (APF). For this aim, Monte Carlo simulations were performed with pairs of synthetic samples of these three variables generated through a novel procedure based on suitable parametric distributions and copulas that leads to realistic variates (see Appendix 2). The power was evaluated as a function of the relative changes in location ($\Delta\lambda_1$), spread ($\Delta\tau$), and skewness ($\Delta\tau_3$) of samples with the same size, which was varied within ranges of commonly available records. The most important results are as follows.

(1) Due to the positively skewed nature of the samples, differences in skewness between the samples lead to differences in spread according to a monotonical relationship. Therefore, the test power is qualitatively related to $\Delta\tau_3$ and $\Delta\tau$ in a similar way.

(2) While based on the assumption of normality and homoscedasticity, the t-S applied to highly skewed precipitation and discharge samples is rather robust since its power is not affected by differences in spread and skewness (Figs. 5 and 6), except for a slight influence found in the APF series (Fig. 7). This is consistent with earlier work, although it was reported that the robustness of t-S might decrease in the case of samples with different sizes which was not analyzed here (Cressie and Whitford 1986; Rasch et al. 2007; Fagerland 2012). Our findings also revealed that the t-S test has a very low ability to correctly detect differences between samples that have similar locations but diverse spreads (and skewness), as shown in the example involving the APF series of Fig. 13.

(3) If the parametric t-S test is applied to the log-transformed samples (t-S–Log), which have reduced skewness, the relationship between its power and $\Delta\lambda_1$ and $\Delta\tau$ of the original samples is very similar to that of the non-parametric Wi test (see Figs. 5–7). Therefore, any of these two well-known tests could be chosen to assess differences in location of precipitation and discharge series, as also proved by the case studies.

(4) For a given value of the difference in location $\Delta\lambda_1$, all tests (except for the original t-S) have lower power if the difference in spread $\Delta\tau$ (and skewness $\Delta\tau_3$) has the same sign as $\Delta\lambda_1$ and higher power if the sign is opposite. This is easily visualized by the southwest-northeast orientation of the major axis of the ellipsoids of the power surfaces. Interestingly, the ($\Delta\lambda_1$, $\Delta\tau$) points of several observed APM seasonal series lie in a region of lower power along such axis (see results for the NE, SE, C, and S regions in Fig. 10). The opposite is true for a number of observed APF seasonal series whose empirical values of $\Delta\lambda_1$ and $\Delta\tau$ have discordant signs (see Fig. 12).

(5) Location tests perform slightly better than distribution tests if the samples have a very close spread and skewness, while distribution tests overperform location tests when the difference in spread and skewness is moderate to large.

(7) The Monte Carlo experiments indicate that, overall, LR is the most powerful test, although this test is more "informed" about the shape of the populations whose parametric form is used for its application. However, the case studies investigating differences in seasonal observed precipitation and discharge samples support this conclusion due to the good correspondence between low p-values (here shown through H_0 rejections) and high test power, as well as to the physical support provided by the presence of geographical patterns of H_0 rejections, after accounting for

field significance and the presence of spatial correlations among the gage records.

(8) As expected, the power increases with the sample size. It becomes high (≥ 0.75) in most conditions for relatively short samples of the NZP series (e.g., $n=810$, which corresponds to about 30 years of seasonal samples and 8 years of annual samples). It is instead low for many combinations of $\Delta\lambda_1$ and $\Delta\tau$ for the APM and APF series even up to $n=50$ or 100 years (see examples of Fig. 8).

(9) The case studies revealed that significant differences between the samples (i.e., H_0 is rejected) are detected with high statistical confidence (i.e., the power or the probability of correctly rejecting H_0 is high) more frequently when considering observed NZP samples, and less frequently with APM and APF samples. For these two variables characterizing extreme events, the tests' power is often very low in part because of the small sample size, thus suggesting that these statistical tools have limited ability to investigate H_0. It was also found that the visual inspection of the geographical patterns of the H_0 rejections could provide valuable information to physically support the outcomes of the statistical analyses.

5 Conclusions

This study provides one of the first quantifications of the power of two-sample tests applied to precipitation and discharge series. Analyses based on Monte Carlo simulations and observed data allowed deriving a set of recommendations that are useful for the selection of two-sample tests and the interpretation of their results in a wide range of hydrologic and climate studies. In particular, it is expected that this work will support the increasing number of studies evaluating changes in precipitation and discharge regimes due to global warming and land cover modifications. The proposed methodology is quite general and could be used to quantify the power of two-sample tests for other applications involving different hydroclimatic variables. Future work should investigate the effects on the test power of samples with different sizes and serial correlation in the analyzed variables, as well as the effect of the significance level, α. We also highlight that the methodology for the generation of realistic precipitation and discharge variates based on copulas could also be useful for other statistical analyses that require multiple samples of these variables.

Appendix 1: Summary of two-sample statistical tests

In the following, we provide a summary of null hypothesis H_0, test statistics, and null distribution used to derive the p-value of the two-sample statistical tests used in the paper. Let $\boldsymbol{x}_1 = \{x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}\}$ and $\boldsymbol{x}_2 = \{x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}\}$ be two independent samples with size n_1 and n_2, respectively. The Student t-test (t-S) evaluates the null hypothesis H_0 of equal mean between the two samples. The t-S test statistic is $z = \frac{\bar{x}_1 - \bar{x}_2}{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^{1/2}}$, where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, and s_1^2 and s_2^2 are the sample variances. The null distribution for z is the t distribution with degrees of freedom $\nu = \min(n_1, n_2) - 1$ for small sample sizes while, for moderately large sizes (as the case of our records), the null distribution of z is well modeled by the standard Gaussian. The t-S test is two-sided. The Wilcoxon (Wi) test is a non-parametric test whose H_0 is that both samples have equal median. The statistic is $U = R_1 - \frac{n_1}{2}(n_1 + 1)$, where R_1 is the sum of the ranks of the first sample and n_1 is its size, while the null distribution of U is Gaussian for sample sizes larger than 10. The Kolmogorov–Smirnov (KS) test is also a non-parametric test that investigates the null hypothesis H_0 that the samples come from the same distribution; its statistic is $D_s = \max_x |F_1(x) - F_2(x)|$, where $F_1(x)$ and $F_2(x)$ are the empirical cumulative distribution functions of the first and second samples, respectively; the p value is computed through the Kolmogorov distribution or other approximations. The likelihood ratio (LR) test has the same H_0 as the Kolmogorov–Smirnov test and requires assuming parametric forms for the distribution of samples 1 and 2 and the two samples combined. Let $G_1\left(x;\hat{\boldsymbol{\theta}}_1\right)$, $G_2\left(x;\hat{\boldsymbol{\theta}}_2\right)$, and $G_0\left(x;\hat{\boldsymbol{\theta}}_0\right)$ be such distributions with parameters $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, and $\hat{\boldsymbol{\theta}}_0$ estimated on the corresponding samples $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_0 = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$. The test statistic is $\Lambda^* = 2[L_1\left(\hat{\boldsymbol{\theta}}_1;\boldsymbol{x}_1\right) + L_2\left(\hat{\boldsymbol{\theta}}_2;\boldsymbol{x}_2\right) - L_0\left(\hat{\boldsymbol{\theta}}_0;\boldsymbol{x}_0\right)]$, where $L_k\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}};\boldsymbol{x}_{\boldsymbol{k}}\right)$ is the log-likelihood of the corresponding distribution $G_k\left(x;\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)$, with $k=0$, 1, and 2. The null distribution is the χ^2 with degrees of freedom $\nu = m_1 + m_2 - m_0$, where m_k is the number of parameters of the k-th distribution.

Appendix 2: Parametric probability distributions and multivariate distributions of their parameters

The cumulative distribution function (CDF) of the Burr Type XII ($\mathcal{B}\nabla\mathrm{XII}$), generalized extreme value ($\mathcal{GEV}$) and Pearson Type 3 ($\mathcal{P}3$) distributions are:

$$F_{\mathcal{B}\nabla\mathrm{XII}}\left(x;\gamma_1,\gamma_2,\theta\right) = 1 - \left[1 + \gamma_2\left(\frac{x}{\theta}\right)^{\gamma_1}\right]^{-\frac{1}{\gamma_1\gamma_2}} \tag{1}$$

$$F_{\mathcal{GEV}}(x;k,\mu,\sigma) = \begin{cases} \exp\left\{-\left(1 + k\frac{x-\mu}{\sigma}\right)^{-\frac{1}{k}}\right\} & k \neq 0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & k = 0 \end{cases} \tag{2}$$

$$F_{P3}(x;\alpha,\beta,\xi)=\begin{cases} G\left(\alpha,\frac{x-\xi}{\beta}\right)/\Gamma(\alpha) & \gamma>0 \\ G\left(\alpha,\frac{\xi-x}{\beta}\right)/\Gamma(\alpha) & \gamma<0 \end{cases} \quad (3)$$

The $\mathcal{B}\nabla$XII distribution has two shape parameters, $\gamma_1>0$ and $\gamma_2>0$, and one scale parameter, $\theta>0$, and is defined for $x\geq 0$. The $\mathcal{GEV}$ distribution has a shape parameter, $k\in(-\infty,+\infty)$, a location parameter, $\mu\in(-\infty,+\infty)$, and a scale parameter, $\boldsymbol{\sigma}>0$; it is defined in the sets $-\infty<x<\infty$ if $k=0$, $\mu-\frac{\sigma}{k}\leq x<\infty$ if $k>0$, and $-\infty<x\leq\mu-\frac{\sigma}{k}$ if $k<0$. The $\mathcal{P}3$ distribution has a shape parameter, $\alpha>0$, a scale parameter, $\beta\in(-\infty,+\infty)$, and a location parameter, $\xi\in(-\infty,+\infty)$; it is defined in the sets $\xi\leq x<\infty$ if the skewness coefficient $\gamma>0$, and $-\infty<x\leq\xi$ if $\gamma<0$. In equation (A3), G(•, •) and Γ(•) are the incomplete and complete gamma functions, respectively. The Log-Pearson Type 3 ($\mathcal{LP}3$) is the $\mathcal{P}3$ distribution where x is the log-transformed value of the original sample (here, the APF series).

Parameters of the $\mathcal{B}\nabla$XII were estimated using the numerical procedure proposed by Zaghloul et al. (2020) based on the method of L-moments. Parameters of the $\mathcal{GEV}$ and $\mathcal{LP}3$ were also estimated with the method of L-moments following Hosking and Wallis (1997). For the $\mathcal{GEV}$, a bias correction of the shape parameter was applied to account for the short sample size using the empirical relations proposed by Papalexiou and Koutsoyiannis (2013) and recently applied by Ansh Srivastava and Mascaro (2023). For each distribution, copulas were used to model the multivariate distribution of the parameters. For the $\mathcal{B}\nabla$XII distribution, we used a three-dimensional Gaussian copula with marginal distributions given by the Generalized Exponential Type 4 ($\mathcal{GE}4$; Papalexiou 2022) for γ_1, the Generalized Gamma ($\mathcal{GG}$) for γ_2, and the empirical CDF for β. For the $\mathcal{GEV}$ distribution, we used a bidimensional Gaussian copula for $\boldsymbol{\sigma}$ and μ with their empirical CDFs adopted as marginal distributions. No significant relationship was found between k and the other two parameters; therefore, after generating a correlated pair of $\boldsymbol{\sigma}$ and μ, the synthetic value of k was randomly drawn from the empirical CDF of the observed estimates. Finally, for the $\mathcal{LP}3$ distribution, a three-dimensional Gaussian copula was used separately for the cases of positive and negative skewness coefficients with marginal distributions given by the $\mathcal{B}\nabla$XII for α and β, and the empirical CDF for ξ. For all three distributions, the selected copulas captured quite well the dependence structure among the parameters, as shown through the scatterplots, histograms, and values of the Spearman rank correlation coefficients reported in Figs. S1. S2, and S3.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00477-024-02709-z.

Acknowledgements The author thanks the Editor and three anonymous reviewers whose comments greatly helped to improve the quality of the manuscript. The MATLAB codes used for the analyses could be shared upon request to the author.

Author contributions Giuseppe Mascaro performed all the work.

Funding This work has been supported by the National Science Foundation (NSF) awards #2212702: "CAS-Climate: A Novel Process-Driven Method for Flood Frequency Analysis Based on Mixed Distributions" and #2221803: "Collaborative Research: CAS—Climate: Improving Nonstationary Intensity-Duration-Frequency Analysis of Extreme Precipitation by Advancing Knowledge on the Generating Mechanisms".

Data availability Precipitation data from the GHCNd rain gage network are available at: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily. Discharge data for the HCDN2009 stream gages are available at: https://waterdata.usgs.gov/nwis/sw.

Declarations

Competing interests The authors declare no competing interests.

References

Amorim R, Villarini G (2023) Assessing the performance of parametric and non-parametric tests for trend detection in partial duration time series. J Flood Risk Manag, e12957. https://doi.org/10.1111/JFR3.12957

Angelina A, Gado Djibo A, Seidou O, Seidou Sanda I, Sittichok K (2015) Changes to flow regime on the Niger River at Koulikoro under a changing climate | Modifications du régime d'écoulement du fleuve Niger à Koulikoro sous changement climatique. Hydrol Sci J 60:1709–1723. https://doi.org/10.1080/02626667.2014.916407

Ansh Srivastava N, Mascaro G (2023) Improving the utility of weather radar for the spatial frequency analysis of extreme precipitation. J Hydrol (amst) 624:129902. https://doi.org/10.1016/J.JHYDROL.2023.129902

Bartlett M (1937) Properties of sufficiency and statistical tests. Proc R Soc Lond A Math Phys Sci 160. https://doi.org/10.1098/rspa.1937.0109

Baumgartner D, Kolassa J (2021) Power considerations for Kolmogorov-Smirnov and Anderson-Darling two-sample tests. Commun Stat Simul Comput. https://doi.org/10.1080/03610918.2021.1928193

Beguería S, Angulo-Martínez M, Vicente-Serrano SM, López-Moreno JI, El-Kenawy A (2011) Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: A case study in northeast Spain from 1930 to 2006. Int J Climatol 31:2102–2114. https://doi.org/10.1002/joc.2218

Blanchet J, Ceresetti D, Molinié G, Creutin JD (2016) A regional GEV scale-invariant framework for Intensity–Duration–Frequency analysis. J Hydrol (amst) 540:82–95. https://doi.org/10.1016/j.jhydrol.2016.06.007

van den Brink WP, van den Brink SGJ (1989) A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. Br J Math Stat Psychol 42:183–189. https://doi.org/10.1111/j.2044-8317.1989.tb00907.x

Chu PS. Chen YR, Schroeder TA (2010) Changes in precipitation extremes in the Hawaiian Islands in a warming climate. J Clim, 23. https://doi.org/10.1175/2010JCLI3484.1

Collings BJ, Hamilton MA (1988) Estimating the power of the two-sample Wilcoxon test for location shift. Biometrics 44:847–860. https://doi.org/10.2307/2531596

Cressie NAC, Whitford HJ (1986) How to Use the Two Sample t-Test. Biometric J, 28. https://doi.org/10.1002/bimj.4710280202

Deidda R, Hellies M, Langousis A (2021) A critical analysis of the shortcomings in spatial frequency analysis of rainfall extremes based on homogeneous regions and a comparison with a hierarchical boundaryless approach. Stochas Environ Res Risk Assess 35(12):2605–2628. https://doi.org/10.1007/S00477-021-02008-X

England JF Jr, Cohn TA, Faber BA, Stedinger JR, Thomas WO Jr, Veilleux AG, Kiang JE, Mason RR Jr (2018) Guidelines for determining flood flow frequency—Bulletin 17C (ver. 1.1, May 2019): U.S. Geological Survey Techniques and Methods, book 4, chap. B5, p 148. https://doi.org/10.3133/tm4B5

Fagerland MW (2012) T-tests, non-parametric tests, and large studiesa paradox of statistical practice? BMC Med Res Methodol, 12. https://doi.org/10.1186/1471-2288-12-78

Fagerland MW, Sandvik L (2009) Performance of five two-sample location tests for skewed distributions with unequal variances. Contemp Clin Trials, 30. https://doi.org/10.1016/j.cct.2009.06.007

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

Feltovich N (2003) Nonparametric tests of differences in medians: Comparison of the Wilcoxon-Mann-Whitney and robust rank-order tests. Exp Econ 6:273–297. https://doi.org/10.1023/A:1026273319211

Freidlin B, Gastwirth JL (2000) Should the Median Test be Retired from General Use? Am Statistic, 54. https://doi.org/10.1080/00031305.2000.10474539

Gretton A, Sejdinovic D, Strathmann H, Balakrishnan S, Pontil M, Fukumizu K, Sriperumbudur BK (2012) Optimal kernel choice for large-scale two-sample tests. Adv Neural Inform Process Syste 25

Griffis VW, Stedinger JR (2007) Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. I: Distribution Characteristics. J Hydrol Eng 12. https://doi.org/10.1061/(asce)1084-0699(2007)12:5(482).

Hoenig JM, Heisey DM (2001) The Abuse of Power. Am Stat 55:19–24. https://doi.org/10.1198/000313001300339897

Hosking JRM (1990) L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. J Roy Stat Soc: Ser B (methodol) 52:105–124. https://doi.org/10.1111/J.2517-6161.1990.TB01775.X

Hosking JRM, Wallis JR (1997) Regional Frequency Analysis. Cambridge University Press. https://doi.org/10.1017/CBO9780511529443

Karl TR, Koss WJ (1984) Regional and national monthly, seasonal, and annual temperature weighted by area, 1895–1983. Historical Climatol Series 3–3

Knighton JO, Walter MT (2016) Critical rainfall statistics for predicting watershed flood responses: rethinking the design storm concept. Hydrol Process 30:3788–3803. https://doi.org/10.1002/hyp.10888

Knoben WJM, Woods RA, Freer JE (2018) A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data. Water Resour Res 54:5088–5109. https://doi.org/10.1029/2018WR022913

Kruskal WH (1957) Historical Notes on the Wilcoxon Unpaired Two-Sample Test. J Am Stat Assoc 52(279):356–360. https://doi.org/10.1080/01621459.1957.10501395

Kunkel KE, Easterling DR, Kristovich DAR, Gleason B, Stoecker L, Smith R (2012) Meteorological Causes of the Secular Variations in Observed Extreme Precipitation Events for the Conterminous United States. J Hydrometeorol 13:1131–1141. https://doi.org/10.1175/JHM-D-11-0108.1

Kunkel TR, Karl MF, Squires X, Yin ST, Stegall, and D. R. Easterling, (2020) Precipitation extremes: Trends and relationships with average precipitation and precipitable water in the contiguous United States. J Appl Meteorol Climatol 59(125–142):2020. https://doi.org/10.1175/JAMC-D-19-0185.1

Lee ET, Desu MM, Gehan EA (1975) A Monte Carlo study of the power of some two-sample tests. Biometrika 62:425–432. https://doi.org/10.1093/biomet/62.2.425

Mascaro G (2020) Comparison of local, regional, and scaling models for rainfall intensity–duration–frequency analysis. J Appl Meteorol Climatol 59:1519–1536. https://doi.org/10.1175/JAMC-D-20-0094.1

Mascaro G, Papalexiou SM, Wright DB (2023) Advancing Characterization and Modeling of Space-Time Correlation Structure and Marginal Distribution of Short-Duration Precipitation. Adv Water Resour 177:104451. https://doi.org/10.1016/J.ADVWATRES.2023.104451

Massey FJ (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. J Am Stat Assoc 46:68. https://doi.org/10.2307/2280095

Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG (2012) An overview of the global historical climatology network-daily database. J Atmos Ocean Technol, 29. https://doi.org/10.1175/JTECH-D-11-00103.1

O'Gorman TW (1995) The effect of unequal variances on the power of several two–sample tests. Commun Stat Simul Comput 24:853–867. https://doi.org/10.1080/03610919508813279

Orskaug E, Scheel I, Frigessi A, Guttorp P, Haugen JE, Tveito OE, Haug O (2011) Evaluation of a dynamic downscaling of precipitation over the Norwegian mainland. Tellus, Ser: Dyn Meteorol Oceanograph 63:746–756. https://doi.org/10.1111/j.1600-0870.2011.00525.x

Papalexiou SM (2022) Rainfall Generation Revisited: Introducing CoSMoS-2s and Advancing Copula-Based Intermittent Time Series Modeling. Water Resour Res, 58, e2021WR031641. https://doi.org/10.1029/2021WR031641

Papalexiou SM, Koutsoyiannis D (2012) Entropy based derivation of probability distributions: A case study to daily rainfall. Adv Water Resour. https://doi.org/10.1016/j.advwatres.2011.11.007

Papalexiou SM, Koutsoyiannis D (2013) Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resour Res 49:187–201. https://doi.org/10.1029/2012WR012557

Papalexiou SM, Koutsoyiannis D (2016) A global survey on the seasonal variation of the marginal distribution of daily precipitation. Adv Water Resour, 94. https://doi.org/10.1016/j.advwatres.2016.05.005

Park J-S, Kang H-S, Lee YS, Kim M-K (2011) Changes in the extreme daily rainfall in South Korea. Int J Climatol 31:2290–2299. https://doi.org/10.1002/joc.2236

Penfield DA (1994) Choosing a two-sample location test. J Exp Educ 62(4):343–360. https://doi.org/10.1080/00220973.1994.9944139

Prosdocimi I, Kjeldsen TR, Svensson C (2014) Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK. Nat Hazard 14:1125–1144. https://doi.org/10.5194/nhess-14-1125-2014

Rasch D, Teuscher F, Guiard V (2007) How robust are tests for two independent samples?. J Stat Plan Inference, 137. https://doi.org/10.1016/j.jspi.2006.04.011

Rauscher SA, O'Brien TA, Piani C, Coppola E, Giorgi F, Collins WD, Lawston PM (2016) A multimodel intercomparison of resolution effects on precipitation: simulations and theory. Clim Dyn 47:2205–2218. https://doi.org/10.1007/s00382-015-2959-5

Roth M, Buishand TA, Jongbloed G, Klein Tank AMG, van Zanten JH (2012) A regional peaks-over-threshold model in a nonstationary climate. Water Resour Res, 48. https://doi.org/10.1029/2012WR012214

Schindler A, Toreti A, Zampieri M, Scoccimarro E, Gualdi S, Fukutome S, Xoplaki E, Luterbacher J (2015) On the internal variability of simulated daily precipitation. J Clim, 28. https://doi.org/10.1175/JCLI-D-14-00745.1

Shao Y, Wu J, Ye J, Liu Y (2015) Frequency analysis and its spatiotemporal characteristics of precipitation extreme events in China during 1951–2010. Theor Appl Climatol 121:775–787. https://doi.org/10.1007/s00704-015-1481-3

Shen SSP, Wied O, Weithmann A, Regele T, Bailey BA, Lawrimore JH (2016) Six temperature and precipitation regimes of the contiguous United States between 1895 and 2010: a statistical inference study. Theor Appl Climatol 125:197–211. https://doi.org/10.1007/s00704-015-1502-2

Slack JR, Landwehr JM (1992) Hydro-Climatic Data Network (HCDN); a U.S. Geological Survey streamflow data set for the United States for the study of climate variations, pp 1874–1988. https://doi.org/10.3133/ofr92129

Student, 1908 The Probable Error of a Mean. Biometrika, 6. https://doi.org/10.2307/2331554

Sugahara S, da Rocha RP, Ynoue RY, da Silveira RB (2015) Statistical detection of spurious variations in daily raingauge data caused by changes in observation practices, as applied to records from various parts of the world. Int J Climatol 35:2922–2933. https://doi.org/10.1002/joc.4183

Sýkorová P, Huth R (2020) The applicability of the Hess–Brezowsky synoptic classification to the description of climate elements in Europe. Theor Appl Climatol, 142. https://doi.org/10.1007/s00704-020-03375-1

Thober S, Samaniego L (2014) Robust ensemble selection by multivariate evaluation of extreme precipitation and temperature characteristics. J Geophys Res 119:594–613. https://doi.org/10.1002/2013JD020505

Totaro V, Gioia A, Iacobellis V (2020) Numerical investigation on the power of parametric and nonparametric tests for trend detection in annual maximum series. Hydrol Earth Syst Sci 24:473–488. https://doi.org/10.5194/HESS-24-473-2020

Vogel RM, Thomas WO, McMahon TA (1993) Flood-Flow Frequency Model Selection in Southwestern United States. J Water Resour Plan Manag, 119. https://doi.org/10.1061/(asce)0733-9496(1993)119:3(353)

Vogel RM, Rosner A, Kirshen PH (2013) Brief Communication: Likelihood of societal preparedness for global change: trend detection. Nat Hazard 13:1773–1778. https://doi.org/10.5194/nhess-13-1773-2013

Wilks DS (2006) On "Field Significance" and the False Discovery Rate. J Appl Meteorol Climatol 45:1181–1189. https://doi.org/10.1175/JAM2404.1

Wilks (2011) Statistical methods in the atmospheric sciences. Academic Press, International Geophysics Series, p 676

Wilks DS (2016) "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It. Bull Am Meteorol Soc 97:2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1

Xu ZX, Takeuchi K, Ishidaira H (2003) Monotonic trend and step changes in Japanese precipitation. J Hydrol (amst) 279:144–150. https://doi.org/10.1016/S0022-1694(03)00178-1

Zaghloul M, Papalexiou SM, Elshorbagy A, Coulibaly P (2020) Revisiting flood peak distributions: A pan-Canadian investigation. Adv Water Resour 145:103720. https://doi.org/10.1016/J.ADVWATRES.2020.103720