On the role of serial correlation and field significance in detecting changes in extreme precipitation frequency

Stefano Farris[1], Roberto Deidda[2], Francesco Viola[3], and Giuseppe Mascaro[4]

[1]Universitá di Cagliari
[2]Universita degli Studi di Cagliari
[3]University of Cagliari
[4]Arizona State University

November 22, 2022

Abstract

Statistical trend analyses of observed precipitation (P) time series are key to validate theoretical arguments and climate projections suggesting that extreme P will increase in a warmer climate. Recent work warned about possible misinterpretation of trend tests if the presence of serial correlation and field significance are not considered. Here, we investigate these two aspects focusing on extreme P frequencies derived from 100-year daily records of 1087 worldwide gauges of the Global Historical Climate Network. For this aim, we perform Monte Carlo experiments based on count time series generated with the Poisson integer autoregressive model and characterized by different sample size, level of autocorrelation, and trend magnitude. The main results are as follows. (1) Empirical autocorrelations are consistent with those of uncorrelated and stationary or nonstationary count time series, while empirical trends cannot be explained as the exclusive effect of autocorrelation; incorporating the impact of serial correlation in trend tests on extreme P frequency has then limited impacts on tests' performance. (2) Accounting for field significance improves interpretation of test results by limiting type-I errors, but it also decreases test power; results of local tests could complement field significance outcomes and help identify weak trend signals where several trends of coherent sign are detected. (3) Based on these findings, evident patterns of statistically significant increasing (decreasing) trends emerge in central and eastern North America, northern Eurasia, and central Australia (southwestern America, southern Europe, and southern Australia). The methodological insights of this work support trend analyses of any hydroclimatic variable

On the role of serial correlation and field significance in detecting changes in extreme precipitation frequency

Stefano Farris[1], Roberto Deidda[1], Francesco Viola[1], and Giuseppe Mascaro[2]

(1) Dipartimento di Ingegneria Civile, Ambientale e Architettura, Università di Cagliari, Italy

(2) School of Sustainable Engineering and the Built Environment, Arizona State University, USA

Corresponding author address: Giuseppe Mascaro, School of Sustainable Engineering and the Built Environment, Arizona State University, 126b, 21 E 6th St, Tempe, AZ 85281. *E-mail*: gmascaro@asu.edu

Abstract

Statistical trend analyses of observed precipitation (P) time series are key to validate theoretical arguments and climate projections suggesting that extreme P will increase in a warmer climate. Recent work warned about possible misinterpretation of trend tests if the presence of serial correlation and field significance are not considered. Here, we investigate these two aspects focusing on extreme P frequencies derived from 100-year daily records of 1087 worldwide gauges of the Global Historical Climate Network. For this aim, we perform Monte Carlo experiments based on count time series generated with the Poisson integer autoregressive model and characterized by different sample size, level of autocorrelation, and trend magnitude. The main results are as follows. (1) Empirical autocorrelations are consistent with those of uncorrelated and stationary or nonstationary count time series, while empirical trends cannot be explained as the exclusive effect of autocorrelation; incorporating the impact of serial correlation in trend tests on extreme P frequency has then limited impacts on tests' performance. (2) Accounting for field significance improves interpretation of test results by limiting type-I errors, but it also decreases test power; results of local tests could complement field significance outcomes and help identify weak trend signals where several trends of coherent sign are detected. (3) Based on these findings, evident patterns of statistically significant increasing (decreasing) trends emerge in central and eastern North America, northern Eurasia, and central Australia (southwestern America, southern Europe, and southern Australia). The methodological insights of this work support trend analyses of any hydroclimatic variable.

1 Introduction

Extreme precipitation (P) is one of the natural hazards with the most significant socioeconomic impacts. Heavy P is the primary input of floods and flash floods, which cause annually large damages to properties and high numbers of fatalities worldwide (Ashley & Ashley, 2008; Peden et al., 2017). For example, the National Oceanic and Atmospheric Administration (NOAA) estimated that, in United States, flooding and severe storms resulted in $437 billion damages and 2379 fatalities from 1980 to 2020 (Smith, 2021). In urban regions, intense P storms lead to pluvial flooding with impacts on traffic (Hooper et al., 2014; Bucar & Hayeri, 2020) and occurrence of power outages (Boggess et al., 2014). Extreme P events have also significant consequences on public health by degrading water quality (Gershunov et al., 2018) and increasing outbreaks of waterborne diseases (Cann et al., 2013). Studies have also shown that extreme P events may reduce crop production (Rosenzweig et al., 2004; Li et al., 2019).

Theoretical arguments suggest that the intensity of P extremes is expected to increase in a future warmer climate (Trenberth et al., 2003; Emori & Brown, 2005; Trenberth, 2011; Nie et al., 2018). According to the Clausius-Clapeyron (CC) equation, as surface temperature rises, the atmospheric water-holding capacity should grow at a rate of 7% K^{-1}. Extreme P is held to increase at a rate close to the CC value or even higher if the strength of moisture convergence will rise (Trenberth et al., 2003). Driven by these theoretical arguments and the evidence of increasing global surface temperature over the last five decades (Hansen et al., 2010; Papalexiou et al., 2020), a number of empirical studies have started to investigate temporal changes of magnitude and frequency in observed records of P extremes based on the application of statistical trend tests. Table 1 summarizes some of these efforts conducted at global and regional scales using mainly daily records of rain gages. Conclusions that emerge across all studies are that (i) trends are mainly increasing but statistically significant only at a limited number of sites; (ii) statistically significant trends are more evident in frequency rather than magnitude of extreme P; (iii) increasing trends are mainly located in eastern and Midwestern U.S. and some regions of Eurasia; and (iv) decreasing trends occur in western U.S. and southern Australia. Despite these common qualitative outcomes, Table 1 emphasizes how these studies vary widely in terms of duration of the investigated time period (ranging from 30 to 112 years); spatial aggregation of the information provided by the rain gages (from point to subcontinental regions); and metrics used to characterize extreme P (targeting magnitude or frequencies above a threshold). As a result, it is difficult to quantitatively compare their results, a task that would be highly needed for practical applications including the update of engineering design standards (Wright et al., 2019).

A key step to improve empirical trend studies of extreme P, facilitate their comparison, and corroborate physical hypotheses on future changes in the driving climate dynamics is to critically assess power and interpret results of statistical trend tests under the possible conditioning of serial correlation, if any, and when applied at multiple sites. We argue that these tasks have received limited attention, likely because these tests are easy to apply numerically via widespread software. These issues have been also recently highlighted by Serinaldi et al. (2018), who discussed potential causes of misuse and misinterpretation of statistical trend tests. One of these causes is the presence of autocorrelation in the analyzed time series, which may occur in hydrologic records as a result of long-term natural climate variability (Koutsoyiannis, 2011; Sun et al., 2018). Several statistical trend tests evaluate the null hypothesis H_0 of random ordering in the time series (note that H_0 is more often defined as “the time series is stationary” or “no trend is present in the time series”). When the time series is autocorrelated while still being stationary, the ordering is not random and the application of trend tests could result in rejecting H_0 more frequently than expected by the significance level (i.e., the type-I error increases). This problem has been investigated for time series of real numbers (e.g., P magnitudes), focusing largely on the Mann-Kendall test (von Storch, 1999; Yue et al., 2002; Hamed, 2009, among others). For this test, the presence of autocorrelation leads to an increase of the test statistic variance, a phenomenon known as variance inflation. To address this issue, two main methods have been proposed including: (i) applying trend tests accounting for a proper estimation of the inflated variance (Hamed & Ramachandra Rao, 1998), and (ii) "prewhitening" the time series, i.e., removing the autocorrelation (Katz, 1988; von Storch, 1999). For both methods, a serial correlation structure of the process has to be adopted based on, e.g., autoregressive or fractional Gaussian models (Hamed, 2009).

As shown in Table 1, most studies that investigated trends in extreme P have not considered the presence of autocorrelation at all or found it to be negligible by simply verifying that the lag-1 autocorrelation, ρ, averaged across all records is close to zero (Groisman et al., 2005; Westra et al., 2013; Papalexiou & Montanari, 2019). Only a small number of efforts have applied techniques to estimate the inflated variance (Tramblay et al., 2013; Kunkel & Frankson, 2015) or prewhitening procedures (Alexander et al., 2006). Unfortunately, several papers have showed that these methods are not easy to apply, because the interaction between possible trends and autocorrelation leads to biases in the estimation of their parameters, which could in turn decrease the trend test power (Yue & Wang, 2002; Bayazit & Önöz, 2007). Moreover, Serinaldi et al. (2018) have demonstrated that the application of different prewhitening techniques to the same dataset could produce markedly diverse outcomes. We have also found that, in the literature that investigated the effect of serial correlation on trend tests, analyses have mainly relied on synthetic experiments in controlled conditions, while observed datasets have been used only in a limited number of cases. In particular, to our knowledge, no study has thoroughly investigated this problem focusing on observed extreme P frequencies.

Another aspect that deserves careful consideration when conducting statistical trend analyses of extreme P is test multiplicity or field significance (Livezey & Chen, 1983; Katz & Brown, 1991; Wilks, 1997; Daniel et al., 2012; Serinaldi et al., 2018). This accounts for the fact that, when a test is applied collectively at M locations (e.g., rain gages or grid points) with a significance level α, the null hypothesis may be rejected, on average, at $\alpha \cdot M$ sites while holding true for the entire set of locations. If the test outcomes are interpreted locally, one can erroneously conclude that a statistically significant trend exists at the $\alpha \cdot M$ sites. This could be even more likely when P records are spatially correlated: in such a case, local tests are not independent and it may be possible to find spatial clusters where H_0 has been erroneously rejected that could mistakenly be considered as physically meaningful spatial features. Results of multiple tests should be instead interpreted globally. To this end, two types of methods have been proposed, including (i) techniques based on counting the number of H_0 rejections and comparing them with thresholds derived from the Binomial distribution (Livezey & Chen, 1983) or from bootstrapping methods (Khaliq et al., 2009; Wilks, 2019), and (ii) methods that minimize the false discovery rate or FDR (Benjamini & Hochberg, 1995; Wilks, 2006, 2016). Modifications of these methods have been proposed to account for spatial dependence. The great majority of previous studies of trend in extreme P have not accounted for field significance, with the exception of Alexander et al., (2006) and Westra et al., (2013), who used bootstrapping methods, and Tramblay et al. (2013), who applied a test based on FDR (Table 1). Additional work is then needed to better investigate the importance of field significance in trend analyses of extreme P records and how its quantification affects power of statistical trend tests.

Driven by these research needs, this study investigates the effect of serial correlation and field significance on power, errors, and interpretation of trend tests applied to observed records of extreme P frequencies at multiple sites. We focus on frequencies (i.e., count time series of exceedances above a threshold) because changes in extreme P have been more effectively detected on counts rather than magnitudes (Papalexiou & Montanari, 2019; Wright et al., 2019). For our analyses, we use 100-year daily P records from 1087 gages the Global Historical Climate Network (GHCN)-Daily dataset (Menne et al., 2012) covering North America, northern and part of southern Europe, northern Asia, and Australia. The core of our methodological framework is based on Monte Carlo simulations, where stationary and nonstationary count time series with different levels of autocorrelation and trend magnitude are generated using the Poisson integer autoregressive (INAR) model of order 1 or Poisson-INAR(1). INAR models were introduced to transfer the structure of autoregressive models for the simulation of integer-valued time series (e.g., McKenzie, 1985; Al-Osh & Alzaid, 1987; Weiß, 2008; Pedeli et al., 2015) and have been rarely applied in hydrology. After showing that the Poisson-INAR(1) model adequately reproduces the autocorrelation structure of most observed count time series, we apply a set of statistical analyses based on Monte Carlo simulations to gain insights on the impact of serial correlation on trend detection in the observed records. We then perform additional Monte Carlo experiments to quantify power and errors of several popular tests (Table 1) conducted locally and at multiple sites, utilizing the FDR test of Wilks (2006) to account for field significance. Finally, we use the knowledge gained with the analyses on serial correlation and field significance to apply trend tests to the observed extreme P frequencies and interpret their results in the studied regions. We repeat the analyses for different sample sizes, ranging from 30 to 100 years, and thresholds used to define the frequencies. While focused on extreme P, this work provides methodological insights supporting trend analyses of any hydroclimatic variable.

2 Data

We use daily P records from the GHCN dataset, which includes more than 100,000 stations in 180 countries with record lengths ranging from a few years to more than 175 years and has been previously used in global (Kunkel & Frankson, 2015; Wilks, 2016; Papalexiou & Montanari, 2019) and regional (Wright et al., 2019; Kunkel et al., 2020) trend analyses. Here, after retaining only records passing all quality controls (Durre et al., 2010), for each station we label as “complete years” those with no more than 10% missing daily data and mark as missing all records collected in those years not satisfying this constraint. Then, we select $M = 1087$ stations with at least 95 complete years in a common 100-year period from 1916 to 2015. Fig. 1 shows the selected gages that are located in three main regions, including North America; northern and part of southern Europe; northern Asia; and Australia. For each record, we derive the count time series of extreme P frequencies $\{o_t\}$ ($t = 1, \ldots, n$, with n being the number of years), defined as the annual occurrences of daily precipitation exceeding the q-th quantiles of its empirical cumulative distribution function (including zeros). These count time series are derived for the nonexceedance probabilities q = 0.9, 0.925, 0.95, 0.975 for n = 100 years and the most recent n = 30 and 50 years.

3 Methodology

The methodology is described in four subsections. In section 3.1, we briefly illustrate the false FDR test that will be applied to evaluate the field significance in selected statistical tests for trend detection. In section 3.2, we investigate the parent distribution of the observed $\{o_t\}$ count time series. In section 3.3 we explain the methods used to generate synthetic count time series simulating statistical properties and potential trends of the observed $\{o_t\}$. In section 3.4, we describe how Monte Carlo simulations based on these synthetic series are used to apply statistical trend tests under different null hypotheses, including possible presence of trend and autocorrelation.

3.1. Evaluation of field significance

As discussed in the Introduction, results of tests conducted at multiple sites are affected by the problem known as test multiplicity or field significance. To account for this, the global null hypothesis H_0 assuming that H_0 is true at all locations should be investigated with a significance level α_{global}. Here, we evaluate the field significance using the FDR test as described in Wilks (2006), since it has been proved more powerful than alternative field significance tests while being computationally efficient (Wilks, 2016). Its application is straightforward; given the p-values from any local test conducted at M sites, the FDR test rejects the local null hypothesis in those sites where the corresponding p-value is lower than a threshold p^*_{FDR} calculated as:

$$p^*_{FDR} = \max_{i=1,\dots,M}\left[p_{(i)}\colon p_{(i)} \le \left(\frac{i}{M}\right)\cdot \alpha_{FDR}\right] \tag{1}$$

where $p_{(i)}$ is the i-th value in the sorted sample of the M p-values, and α_{FDR} is the significance level of the local test (see Wilks, 2016 for details). If the p-value is lower than p^*_{FDR} at one or more sites, then the global H_0 is rejected at a level $\alpha_{global} = \alpha_{FDR}$. In these sites, the local H_0 is also rejected and the potential existence of spatial patterns where H_0 is rejected can be explored. A very attractive property of the FDR test is that it can be easily adapted to the cases of spatial dependence among the gage records. Using numerical simulations, Wilks (2016) suggests that the FDR test is robust to the presence of spatial correlation if a value of $\alpha_{FDR} = 2\alpha_{global}$ is adopted. Unless stated otherwise, for all tests conducted in this study, we assume $\alpha_{global} = 0.05$ and $\alpha_{FDR} = 0.10$.

3.2. Preliminary inference on the parent distribution of exceedance counts

We conduct preliminary analyses to identify a reasonable parent distribution for the observed exceedance counts $\{o_t\}$ at the GHCN gages. Specifically, we apply the Chi-Square and Lilliefors (a generalization of Kolmogorov-Smirnov) goodness-of-fit (GOF) tests to evaluate the null hypothesis H_0 that the Poisson distribution well reproduces the marginal distribution of the observed counts. We do this for the count series with n = 30, 50, and 100 years. Instead of applying the GOF tests in their traditional formulation, we build the null distribution of the GOF test statistics through Monte Carlo simulations (details are provided in Section 3.4), because (i) statistical tables for the Chi-Square null distribution are usually derived and valid when parameters of the fitted distribution are estimated by minimizing the Chi-Square statistic (Fisher, 1922); and (ii) performances of GOF tests can be biased when applied to discrete variables (see e.g. Deidda & Puliga, 2006). We then apply the FDR test for both GOF tests, finding that H_0 cannot be rejected in more than 95% of the gages at $\alpha_{global} = 0.05$ for all values of q and n. Given the very small number of rejections, the Poisson distribution is adopted as the parent distribution of count time series.

3.3 Generation of synthetic count time series

We conduct several Monte Carlo experiments based on the generation of random Poisson-distributed count time series that serve two main goals. The first is to gain insights on the open question raised by several authors (Yue et al., 2002; Hamed, 2009; Serinaldi & Kilsby, 2016) concerning the influence of serial correlation on trend detection and vice versa. In particular, we investigate (i) the degree of autocorrelation that can be detected in time series generated under controlled uncorrelated and nonstationary conditions, and, conversely, (ii) the trend induced by the presence of autocorrelation in time series generated under stationary conditions. The second goal of the Monte Carlo experiments is to generate the null distribution for the statistics of the trend tests (as described in Section 3.4) to account for discretization, sample length, and possible presence of autocorrelation. In such a way, we can also explore the type-I error and power of trend tests applied locally and at multiple sites. The generation of the synthetic count time series is described in the next subsections.

3.3.1 Nonstationary uncorrelated time series

Under the assumption of Poisson distributed counts, we can easily generate synthetic time series with a controlled trend slope ϕ, applying a linear time-varying relation for the Poisson parameter:

$$\lambda_t = \lambda_0 + \phi \cdot t\,, \qquad t = 1, \ldots, n \tag{2}$$

where the intercept λ_0 is derived by constraining the mean value of $\{\lambda_t\}$ to be $\bar{\lambda} = (1 - q) \cdot 365.25$, with q being the selected nonexceedance probability. This results in $\lambda_0 = \bar{\lambda} - \phi \cdot (n + 1)/2$.

3.3.2 Stationary correlated time series

We use the INAR(1) model to generate random autocorrelated stationary count time series. INAR models have been mainly applied in economics and finance (e.g., Blundell et al., 2002; Jung and Tremayne, 2011), epidemiology (e.g., Allard, 1998; Pascual & Akhundjanov, 2019), and insurance (e.g., Gourieroux & Jasiak, 2004; Boucher et al., 2008), but they have received less attention in hydrology and climatology. To define the INAR(1) process, we first introduce the binomial thinning operator, "∘"

(Steutel & van Harn, 1979). If $\rho \in [0, 1]$ and N is a nonnegative integer random variable, this operator is defined as:

$$\rho \circ N = \sum_{i=1}^{N} Y_i \,, \qquad N > 0 \tag{3}$$

where $\{Y_i\}$ are independent and identically distributed (i.i.d.) variates of a Bernoulli distribution B(ρ). While other thinning operators have been proposed (Weiß, 2008), here the binomial thinning operator is used. A process $\{N_t\}$ is defined INAR(1) if:

$$N_t = \rho \circ N_{t-1} + \epsilon_t \tag{4}$$

where $\{\epsilon_t\}$ is an i.i.d. random process of integer values and the binomial thinning operator with parameter ρ is applied to N_{t-1}. Its lag-k autocorrelation is $r(k) = \rho^k$, similar to the AR(1) model for real values.

In light of the results discussed in section 3.2, we adopt a Poisson-INAR(1) model to generate synthetic correlated count series, where $\{\epsilon_t\}$ is an i.i.d. random process according to a Poisson distribution with parameter μ, and the marginal distribution of $\{N_t\}$ is also a Poisson distribution with parameter $\left(\frac{\mu}{1-\rho}\right)$ (Weiß, 2008). Parameters of the Poisson-INAR(1) model reproducing the statistical properties of an observed count time series $\{o_t\}$ can be estimated as: $\rho = \rho_{\text{obs}}$, with ρ_{obs} being the observed lag-1 autocorrelation of $\{o_t\}$; and $\mu = (1 - \rho_{\text{obs}})\overline{\lambda}$, with $\overline{\lambda} = (1 - q) \cdot 365.25$ being the expected number of annual exceedances above the q-th quantile. An example of the capability of the Poisson-INAR(1) model to reproduce the statistical properties of our observed counts is shown in Fig. 2, where the empirical autocorrelation function of two randomly chosen count time series derived from the GHCN P records is compared to the 95% confidence intervals (CIs) built from 10,000 model simulations with the parameters estimated as just described. Fig. 2 shows that the Poisson-INAR(1) model captures very well the empirical autocorrelations at different lags.

3.4 Setup of statistical tests through Monte Carlo simulations

To detect empirical trends in our analyses, we focus on three (two) nonparametric (parametric) statistical tests widely used in trend analyses of P extremes (Table 1). The nonparametric ones include Mann Kendall (Mann, 1945; Kendall, 1975); Kendall's τ (Kendall, 1938; El-Shaarawi & Niculescu, 1992); and Spearman's ρ (Gauthier, 2001). The parametric tests are based on linear and Poisson regression (Wilks, 2019). All these tests have been originally devised to investigate the null hypothesis of trend absence in uncorrelated time series. However, some authors have warned about the possible degraded test performances due to the possible presence of serial correlation in stationary time series (e.g., Serinaldi & Kilsby, 2016). To investigate this issue, we use Monte Carlo simulations to build the distribution of the test statistics under any H_0 that may include uncorrelated and autocorrelated time series. In such a way, we also reduce potential biases introduced by finite sample sizes and discrete records (Deidda & Puliga, 2006), as well as by the presence of ties likely found in count time series.

In the general case of count time series of length n affected by serial correlation, a statistical trend detection test based on Monte Carlo simulations can be applied as follows:

1. The expected number of exceedances above the q-th quantile is estimated as $\bar{\lambda} = (1 - q) \cdot 365.25$.
2. Parameters of the Poisson-INAR(1) model in equation (4) are estimated as: $\rho = \rho_{\text{obs}}$ and $\mu = (1 - \rho_{\text{obs}}) \cdot \bar{\lambda}$.
3. An ensemble of n_{ens} (e.g. n_{ens} = 10,000 in our applications) stationary count time series, each of length n, is generated using the Poisson-INAR(1) model with parameters estimated in step (2).
4. The s test statistic of interest (e.g., $s = \tau$ for Kendall's) is computed for each of the n_{ens} count time series generated in step (3).
5. The empirical cumulative distribution function (ECDF) of the n_{ens} test statistics from step (4) is used to determine the acceptance region of the null hypothesis. For example, for two-sided tests, this is the interval of s-quantiles corresponding to probabilities $\alpha/2$ and $(1 - \alpha/2)$, for any

considered significance level α. The local null hypothesis is therefore accepted or rejected by comparing the test statistic computed on the time series of interest, s_{obs}, with such acceptance region.

6. Similarly, the ECDF of the n_{ens} test statistics from step (4) is used to determine the p-value of s_{obs} (note that, for two-sided tests, as those selected here, the corresponding p-value has to be estimated by doubling the exceedance or nonexceedance probability in the ECDF).
7. If the test is conducted at M sites, the field significance is taken into account through the FDR test applied with the M p-values determined at each site, as described in steps (1)-(6).

This procedure is general and can be implemented for any trend test by using the corresponding test statistic in steps (4)-(6) (see Appendix for details on the tests considered here). Moreover, with this method, different null hypotheses can be tested depending on the properties of the synthetic count series generated in step (3). We will use the following compact notation to describe the null hypothesis tested in this study, including: H_0: "$\rho_0 = 0$; $\phi_0 = 0$" for uncorrelated and stationary signals generated at step (3) from a Poisson distribution with parameter $\bar{\lambda}$ (in this case, step (2) is skipped); and H_0: "$\rho_0 = \rho^*$; $\phi_0 = 0$" for serially correlated and stationary signals generated from the Poisson-INAR(1) model with parameter $\rho = \rho^*$ (e.g., $\rho^* = \rho_{\text{obs}}$ in step (2)).

An analogous procedure can also be implemented to test whether a certain degree of autocorrelation detected in a count time series can be reasonably due to the presence of a given trend. In this case: the null hypothesis is H_0: "$\rho_0 = 0$; $\phi_0 = \phi^*$"; step (2) is skipped; an ensemble of n_{ens} nonstationary uncorrelated count time series of length n is generated in step (3) as described in section 3.3.1, using parameter $\bar{\lambda}$ from step (1) and a given trend slope ϕ^*; finally, the lag-1 autocorrelation is used as test statistic in step (4) and the n_{ens} estimated lag-1 autocorrelations are utilized to compute the p-value associated with the observed autocorrelation.

4 Results and discussion

4.1. Investigation of autocorrelation and its relationship with linear trends

Deciding whether the possible influence of serial correlation in trend detection should be taken into account is not an easy question to answer, because, in principle, there can be a reciprocal feedback between autocorrelation and trend. To investigate this nontrivial issue in our count time series, we use two simple metrics to characterize autocorrelation and trend, namely the lag-1 autocorrelation, ρ, and the linear trend slope, ϕ, respectively. We first compare the empirical distributions of ρ and ϕ of the M = 1087 observed count time series with the corresponding 95% CI of ρ and ϕ, respectively, derived from n_{ens} = 10,000 Monte Carlo simulations under H_0: "$\rho_0 = 0$; $\phi_0 = 0$". In other words, we evaluate whether the observed ρ's and ϕ's can be considered statistically different from those of uncorrelated time series with no trend. Results are shown in Fig. 3 for q = 0.95 and different n (similar patterns are obtained for the other q's; see Figs. S1, S2 and S3 in Supplementary Material). As expected, for both metrics the dispersion of the empirical distributions increases for smaller n. The simple visual comparison of distributions and 95% CIs suggests that H_0 should be locally rejected for ρ in a relatively small number of sites (Figs. 3a-c), while the number of rejections appears to be much higher for ϕ (Figs. 3d-f). These visual speculations are confirmed by results for the local test reported in Table 2 and, more importantly, by the application of the FDR test, which reveals that for n = 100 only 3% (or 0% for n = 50 and 30) of the observed ρ's can be considered statistically significant at α_{global} = 0.05 significance level, while the percentage of statistically significant observed ϕ*'s* is much larger (41%). Results for all considered n and local and FDR tests are reported in Table 2 and consistently show that, while a large number of sites seem to be affected by significant trend, the same conclusion does not hold for empirical serial correlation.

To further explore whether the presence of autocorrelation may introduce bias in the estimation of the linear trend slope, we analyze the joint distribution of ρ and ϕ estimated on the M observed 100-year time series. The scatterplot between these values is plotted in Fig. 4a (grey circles) along with estimates derived from M random stationary and uncorrelated time series (black circles). The visual inspection clearly suggests that the observations do not appear consistent with a hypothesis of both no autocorrelation and no trend. In particular, the observed counts exhibit more cases with higher slope (both positive and negative) that are associated with higher autocorrelation. To gain insights on the potential cause-effect relationship of this outcome (i.e., is the autocorrelation causing an artificial trend or is the opposite true? Or are these effects independent?), we first evaluate whether the presence of trend can artificially induce autocorrelation. For this aim, we generate time series under H_0: "$\rho_0 = 0$; $\phi_0 = \phi$", with ϕ varying from -0.2 to 0.2 events/yr to cover the whole range of observed trend slopes for n =100. For each value of ϕ, we produce n_{ens} = 10,000 samples, estimate ρ on each time series, and derive the 95% CI of ρ (solid lines in Fig. 4b). We find that 95% of the observed (ρ,ϕ) pairs lie within the CI, indicating that the observed ρ's, even if different from zero, are compatible with those of uncorrelated series with trend.

Following a similar framework, we then investigate whether the presence of autocorrelation could artificially induce significant trends. We do so by computing the 95% CI of ϕ from time series randomly generated under H_0: "$\rho_0 = \rho$; $\phi_0 = 0$", with ρ varying from 0 to 0.8 (solid line in Fig. 4c). In this case, a large fraction (40%) of observed (ρ, ϕ) pairs lies outside of this CI, implying that several high values of ϕ cannot be explained solely by the presence of autocorrelation. The same conclusion can be drawn by comparing this CI with that obtained under H_0: "$\rho_0 = 0$; $\phi_0 = 0$" (dotted line in Fig. 4c): the two CIs are very close to each other, meaning that accounting or not for the possible presence of serial correlation has a very limited impact on the assessment of trend significance. The only region where the trend significance could be potentially ascribed to the presence of autocorrelation is the area between the two CIs, which includes only a very limited number of observed cases. It is also worth noticing that such a few cases would be certainly less if one rightly considers only the component of autocorrelation that is not ascribed to the presence of trend, which results in a positive overestimation of ρ, as also clearly reflected in the CIs shown in Fig. 4b (see also Yue & Wang, 2002).

Results presented in Fig. 4 suggest that autocorrelation in observed count time series of extreme P is likely caused by the presence of trends. To complement this conclusion relying on statistical simulations, we provide further evidence based on the physical argument that temporal persistency (if any) in extreme P should significantly decrease after a few years. From each observed time series, we sample the record every four years, thus extracting four sub-series of size n = 25; in such a way, we eliminate the effect of potential autocorrelations at lags from 1 to 3 years. For each sub-series, we estimate ϕ and plot it against the slope estimated on the full series. Results are presented in Fig. 5a, which shows that, despite some expected sampling variability, all values are distributed along the 1:1 line. In addition, we randomly generate M uncorrelated series of duration n = 100 with the same M slopes estimated on the observed series, and, for each synthetic sample, we repeat the same calculation on four sub-series of size n = 25 sampled every four years. The corresponding outcome, reported in Fig. 5b, is consistent with results for the observed series, thus providing further evidence that statistically significant trends exist in our observed count time series, independently of the possible presence of autocorrelation.

4.2. Performance of local trend tests

After analyzing the relations between trend and possible presence of autocorrelation, we now use Monte Carlo simulations to investigate if accounting or not for autocorrelation can affect the power of local trend tests. To this end, we generate 10,000 nonstationary uncorrelated time series for different values of ϕ, n and q using equation (2) as described in Section 3.3.1. For each combination of ϕ, n and q, we estimate the test power as the fraction of rejections of the null hypotheses H_0: “$\rho_0 = 0$; $\phi_0 = 0$” and

H_0: "$\rho_0 = \rho_{obs}$; $\phi_0 = 0$", applying the trend tests as described in Section 3.4. Results are presented in Fig. 6, where dotted and solid lines are used for the two H_0 settings and colors refer to different tests. For n = 100 years, Fig. 6a shows that the power of all tests increases in quasi-linear fashion from 0.05 (the test significance level) at $\phi = 0$ to ~0.9 at $\phi = 0.05$ events/yr, reaching 1 for $\phi > 0.07$ events/yr. As expected, for a given ϕ, the test power decreases with n (Fig. 6b). For $\phi \leq 0.05$ events/yr, the power is less than 0.5 for $n \leq 70$ years, indicating that the statistical tests analyzed here have low ability to detect trends even when n is relatively large. The use of H_0: "$\rho_0 = \rho_{obs}$; $\phi_0 = 0$" leads to a slight power reduction compared to H_0: "$\rho_0 = 0$; $\phi_0 = 0$", a further indication that taking or not taking into account autocorrelation does not significantly impact results. Finally, as better shown in Fig. 6b, parametric (linear and Poisson regression) and nonparametric (Mann Kendall, Kendall's τ and Spearman ρ) tests cluster in two separate groups, with the parametric tests exhibiting slightly higher power than the nonparametric ones. Based on these findings, we will discuss trends in observed count time series in section 4.4 presenting results only for the Poisson regression (PR) and Mann Kendall (MK) tests, which are representative of parametric and nonparametric tests, respectively. The difference in power between these two tests as a function of ϕ for n = 100 years is reported in Fig. 7.

4.3. Performance of trend tests at multiple sites

We gain insights on tests' performance at multiple locations by conducting synthetic experiments on a 50 × 100 grid totaling M = 5,000 sites, where we hypothesize the existence of trend only in an inner rectangular domain containing 30% of the grid points. In each site of this region, we generate count time series with a given linear trend slope ϕ, while, in the remaining grid points placed in the outer region, we generate stationary time series. We do this for q = 0.95 and for n = 50 and 100 years. We discuss here results for PR trend tests (results are similar for other tests) applied locally under H_0: "$\rho_0 = 0$; $\phi_0 = 0$", and globally by accounting for field significance with the FDR test at $\alpha_{FDR} = \alpha_{global}$ (there is no spatial correlation in this experiment). Fig. 8a (Fig. 8b) presents the fraction of H_0 rejections in the inner region with trends (outer region without trends) as a function of ϕ, quantifying test power (type-I error) in that part of the domain. For small trend slopes, local tests lead to higher power (differences of up to 0.5 compared to global results), but such discrepancies approach zero as ϕ increases (Fig. 8a). As found for the local analyses, for a given ϕ, the power is heavily affected by the sample size. For example, for ϕ = 0.05 events/yr, the power of the global test drops from 0.8 for n = 100 years to zero for n = 50 years. On the other hand, applying tests locally without considering field significance leads to much larger type-I errors in the outer region for any ϕ (Fig. 8b). In other words, the use of local tests leads to several false rejections of H_0 that the FDR test is able to prevent. In this case, the effect of the sample size is negligible.

To visually illustrate performance of tests conducted at multiple sites, we refer to the same 50 × 100 grid with time series in the inner region generated with ϕ = 0.05 events/yr, for n = 100 and 50 years. Figs. 8c-f present maps of test results applied locally and globally, with red (green) colors indicating H_0 rejections for the PR when the trend slope estimated on the generated time series is positive (negative). We first focus on the maps for n = 100 years (Figs. 8c,d). When tests are performed locally (Fig. 8c), H_0 is rejected, as expected, at ~5% of the locations in the outer region. This would erroneously indicate statistically significant trends at sites where trend is not present, inducing wrong physical interpretations if these sites coincidentally cluster. Accounting for field significance with the FDR test (Fig. 8d) leads instead to the rejection of H_0 at just a few spurious locations (~1% of the points in the outer region). In this condition, it is more meaningful to interpret these rejections as a result of randomness rather than physical processes. When considering the inner region with trends, the application of the more conservative (i.e., H_0 is rejected less) FDR test returns a higher number of false nonrejections of H_0 compared to the local test (22.8% vs 8.9% of the cases). However, despite the lower power (also highlighted in Fig. 8a), H_0 is rejected at most locations that are spatially clustered, so that the region with trend could be readily identified.

When n = 50 years, results for the local tests (Fig. 8e) do not change in the outer region, with random occurrence of H_0 rejections at ~5% of the points with positive and negative slopes as found for n = 100 years. The reduction of test power due to the smaller sample size leads instead to less H_0 rejections in the inner region. Changes are even more drastic when applying the more conservative FDR test, which results in H_0 nonrejections at all sites (Fig. 8f). This outcome suggests that, when the trend signal is low, the use of methods accounting for field significance will likely indicate the absence of statistically significant trends. In this circumstance, a careful interpretation of results of the more powerful local tests could still allow identifying large areas characterized by statistically significant trends if the sites exhibit coherent positive or negative trend. This is depicted in the example of Fig. 8e, where positive trends are correctly detected at a number of nearby locations that is sufficiently large to identify the inner region. In the outer region, the mixture of both positive and negative trends in sites close to each other should suggest that no trend signal is detectable in such area. This issue will be further discussed in the next section.

4.4. Trend analyses of observed count series

In light of the insights gained in the previous sections, we now analyze the presence of trends in observed count series on the M = 1087 selected stations from the GHCN gage network. Trends are investigated applying the PR and MK tests and, then, the FDR test at α_{global} = 0.05 to account for field significance. We preliminarily considered two null hypotheses: stationary and uncorrelated signals, and stationary and autocorrelated series. Regarding the second null hypothesis, our previous analyses have shown that a large portion (or perhaps all) of the lag-1 autocorrelation estimated on the observed sample, ρ_{obs}, is likely induced by the presence of trend (see Fig. 4b). As a result, when testing the null hypothesis of autocorrelated signals, we should consider only the residual component of ρ_{obs} that cannot be ascribed to the presence of trend (see discussion in Section 4.1). Considering that implementing such an approach is not straightforward, the trend tests were preliminary applied under H_0: "$\rho_0 = 0$; $\phi_0 = 0$" and H_0: "$\rho_0 =$

ρ_{obs}; $\phi_0 = 0$", which represent two extreme conditions. Since we found very similar results and patterns in the two cases (not shown), we hereon discuss only results for H_0: "$\rho_0 = 0$; $\phi_0 = 0$".

Fig. 9 presents maps of statistically significant trends for $q = 0.95$ and $n = 100$ years. Colored circles and triangles locate significant trends for (i) only PR and (ii) both PR and MK tests, respectively. We first note that, as suggested by the synthetic experiments, PR detects a larger number of statistically significant trends than MK, while the opposite never occurs. This is better visualized in the scatterplot between ϕ and ρ of Fig. 10, where H_0 rejections by only PR or both PR and MK tests are plotted with different markers. The occurrence of the different cases is controlled by ϕ, while ρ is not influential, thus providing additional evidence on the limited effect of autocorrelation on trend detection. In particular, H_0 is rejected by both tests for $|\phi| > \sim 0.05$ events/yr, which is a region where the power of all tests is high for $n = 100$ years (Fig. 6a). H_0 is rejected only by PR at several sites where $|\phi|$ is included between roughly 0.02 and 0.05 events/yr, where the power of both tests decreases (Fig. 6a) but is larger for PR than MK (Fig. 7). This behavior can, at least partially, explain why the parametric PR rejects H_0 in more cases than the nonparametric MK test.

Despite PR leads to rejection of H_0 at several sites where our synthetic experiments suggest low test power, Fig. 9 clearly shows that locations where trends are statistically significant are well clusterized in space, with distinct regions where the trend is either increasing (red symbols) or decreasing (green symbols). As shown in the synthetic experiments at multiple sites of Fig. 8, the presence of spatial clusters provides further evidence of trend existence. This empirical result is also supported by the physical argument that extreme P is often controlled by synoptic processes (Barlow et al., 2019), and that their occurrence is changing in time (Zhang & Villarini, 2019). As a result, when trends exist, they should manifest over relatively large regions and, if multiple gages are present, statistical tests should detect statistically significant trends with the same sign at several of these sites (e.g., Kunkel et al., 2020). In particular, consistent with previous work with global and regional datasets (Table 1), our analyses reveal that significant trends are mainly increasing in central and eastern North America (Janssen et al., 2014), northern Europe (Madsen et al., 2014), northern Asia (Zolotokrylin and Cherenkova, 2017), and central regions of Australia (Gallant et al., 2007). Extreme P exhibit instead negative trends in southwestern North America (Hoerling et al., 2016), part of southern Europe (Papalexiou & Montanari, 2019), and southwestern and southeastern regions of Australia close to the coast (Hughes, 2003).

The synthetic experiments indicate that the tests' power could be severely reduced when the sample size decreases. We analyze this issue on the observed count time series by plotting in Fig. 11a the maps of H_0 rejections by the FDR test applied on PR and MK for $q = 0.95$ and $n = 50$ years (results for $n = 30$ years are presented in Fig. S6). When compared to Fig. 9, the number of H_0 rejections dramatically declines. The only regions with a relatively large number of spatially clustered gages that exhibit statistically significant trends are northern Europe (increasing trend) and southern Australia (decreasing trend). In North America, there are some gages where H_0 is rejected, but their location is quite sparse, although there is a relatively clear geographical distinction between increasing and decreasing trends. In this circumstance where the trend signal might be weak, local test results could be used to complement results of the more conservative FDR test. As shown in Fig. 11b, local H_0 rejections have a well-defined spatial pattern with two large regions where the trend sign is the same: central and northeastern (southwestern) North America, with increasing (decreasing) trend, which are the same regions identified in Fig. 9 for $n = 100$ years. To complete our analysis, we investigate the role of the nonexceedance probability q, which controls the threshold used to build the count series of extreme P. Fig. 12 displays maps of global tests results for $n = 100$ years for $q = 0.90$ and 0.975 (results for $n = 30$ years are presented in Figs. S4-S7). As q increases and focus is placed on rarer events, less statistically significant trends are detected, but the spatial patterns of increasing and decreasing trends in the different regions of the worlds are always clearly visible.

5 Summary and conclusions

Increasing evidence and theoretical arguments indicate that global warming is causing and will cause changes in extreme P. Accurate statistical trend analyses of observed and modeled P time series are key to validate hypotheses on the underlying physical mechanisms and improve our ability to predict the magnitude of these changes. In this study, we clarified how autocorrelation and field significance affect application, power, and interpretation of several popular tests for trend detection in count time series. We focused on count time series because stronger trends have been detected in extreme P frequencies rather than magnitudes. We used observed records of extreme P frequency in the 100-year period from 1916 to 2015 collected by 1087 high-quality rain gages of the GHNC network, covering North America, part of Europe and Asia, and Australia. To investigate the role of autocorrelation and field significance and interpret trends in observed records, we designed several Monte Carlo experiments based on the random generation of stationary and nonstationary count time series with different levels of autocorrelation and sample size. The experiments involved the use of the Poisson-INAR(1) model that has been rarely adopted in hydroclimatic applications. Our results can be summarized as follows:

1. Although some observed count time series may exhibit some degree of autocorrelation (quantified through the lag-1 autocorrelation, ρ), we proved that such correlations are mainly consistent with those of uncorrelated and either stationary or nonstationary count time series with the same sample size. We observed that records exhibiting stronger trends (quantified through the linear slope, ϕ) are also characterized by high ρ values; in these cases, using statistical arguments, we proved that the empirical high ρ values are compatible with uncorrelated time series with trends of the same observed magnitude. Conversely, we also proved that high trend slopes cannot be interpreted as a spurious outcome of a stationary autocorrelated signals. As a result, autocorrelation in observed count time series of extreme P appears to be caused by the

presence of trends, indicating that taking or not taking into account its presence when applying statistical trend tests does not significantly impact results.

2. As expected, the power of trend tests is importantly affected by sample size, n, of the analyzed series and trend magnitude, ϕ. For example, considering the occurrences of daily precipitation with nonexceedance probability $q = 0.95$ and a trend slope $\phi = 0.05$ events/yr, the power is lower than 0.5 when $n \leq 70$ years, which is a relatively long record. The power of parametric tests (linear and Poisson regression) is slightly larger than that of nonparametric tests (Mann Kendall, Kendall's τ and Spearman ρ).
3. Trend tests are in most cases applied at multiple locations. Here, we confirmed that, if test multiplicity or field significance is not taken into account, type-I errors could be large and statistically significant trends could be mistakenly detected at several sites, inducing wrong physical interpretations when these locations tend to coincidentally cluster. Accounting for field significance severely reduces this problem. On the other hand, we also showed that the inclusion of field significance leads to a power reduction compared to local tests. While this issue is practically irrelevant when the trend signal is moderate and high, it may result in several incorrect nonrejections of H_0, especially when the sample size is small. To limit this, the careful interpretation of results of local tests could help correctly identify trends in large regions where: (i) several gages are present; (ii) local tests reject H_0 at most locations; and (iii) the trend detected in close gages has the same sign. These recommendations are supported by the empirical analyses of observed records presented here, as well as by the physical evidence that extreme P is mainly driven by large-scale processes whose occurrence has been changing in time. In such a way, the power of regional trend analyses is expected to increase, a task highly desirable to support engineering design against natural hazards (Vogel et al., 2013).

4. The application of several trend tests on the selected 1087 rain gages of the GHNC network reveals statistically significant increasing trends in several parts of the world, including central and eastern North America, northern Europe, part of northern Asia, and central regions of Australia. Decreasing trends are instead found in southwestern North America, part of southern Europe, and southwestern and southeastern regions of Australia. These results are largely consistent with previous studies.

Our work provides useful guidance for a more informed application of statistical trend tests in regional and global trend analyses of hydroclimatic extremes, and for a more realistic interpretation of test results.

Acknowledgment

This work has been supported by Fondazione Banco di Sardegna, funding call 2017, project: Impacts of climate change on water resources and floods, CUP: F71I17000270002. The data used in this study are publicly available via the Global Historical Climatology Network website: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00861.

Appendix A

Given the count time series $\{o_k\}$ with $k = 1, \ldots, n$, we investigate the existence of trend through three nonparametric tests (Mann Kendall, Kendall's τ, and Spearman's ρ) and two parametric tests (test on linear regression slope and Poisson regression). In the following, we report the statistics of each test, which are used to build the null distribution via Monte Carlo simulations as described in section 3.4.

The Mann-Kendall test is based on the test statistic S calculated as:

$$S = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} sign\left(o_j - o_k\right) = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} sign\left(R_j - R_k\right) \tag{A1}$$

where o_j and o_k represent j-th and k-th values of the count time series, R_j and R_k the corresponding ranks, n is the length of the series and:

$$sign\left(R_j - R_k\right) \begin{cases} 1 & \text{for } \left(R_j - R_k\right) > 0 \\ 0 & \text{for } \left(R_j - R_k\right) = 0 \\ \text{-}1 & \text{for } \left(R_j - R_k\right) < 0 \end{cases} \tag{A2}$$

In the Spearman's rank correlation test, the following r_s test statistics is used:

$$r_s \; = \; 1 - \frac{6 \sum_{i=1}^{n} (R_i - i)^2}{n(n^2 - 1)} \tag{A3}$$

The Kendall's τ test is based on a measure of the rank correlation evaluated as follows:

$$\tau \; = \; \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign\left(R_i - R_j\right) sign(i - j) \tag{A4}$$

The trend test on linear regression slope is based on the regression between $\{o_k\}$ and $k = 1, \ldots, n$ as follows:

$$\mu_k = b_0 + b_1 k, \tag{A5}$$

where μ_k is the predicted value, and b_0 and b_1 are parameters estimated through the least squares approach. Here, we apply the trend test using b_1 as test statistic.

The Poisson regression is a generalized linear model that links a Poisson-distributed variable with a set of predictors. Here, we consider only one predictor. The model relates the logarithm of the μ parameter of the parent Poisson distribution of the predictand with the predictor as:

$$\ln(\mu_k) = b_0 + b_1 k \tag{A6}$$

The statistics used to apply the test under the proposed modification with Monte Carlo simulations is b_1.

References

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A. M. G., et al. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, *111*(D5), D05109. https://doi.org/10.1029/2005JD006290

Allard, R. (1998). Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization*, *76*(4), 327–333.

Al-Osh, M. A., & Alzaid, A. A. (1987). First-Order Integer-Valued Autoregressive (INAR (1)) Process. *Journal of Time Series Analysis*, *8*(3), 261–275. https://doi.org/10.1111/j.1467-9892.1987.tb00438.x

Alpert, P., Ben-Gai, T., Baharad, A., Benjamini, Y., Yekutieli, D., Colacino, M., et al. (2002). The paradoxical increase of Mediterranean extreme daily rainfall in spite of decrease in total values. *Geophysical Research Letters*, *29*(11), 31–1. https://doi.org/10.1029/2001GL013554

Asadieh, B., & Krakauer, N. Y. (2015). Global trends in extreme precipitation: climate models versus observations. *Hydrol. Earth Syst. Sci.*, *19*(2), 877–891. https://doi.org/10.5194/hess-19-877-2015

Ashley, S. T., & Ashley, W. S. (2008). Flood Fatalities in the United States. *Journal of Applied Meteorology and Climatology*, *47*(3), 805–818. https://doi.org/10.1175/2007JAMC1611.1

Barlow, M., Gutowski, W. J., Gyakum, J. R., Katz, R. W., Lim, Y.-K., Schumacher, R. S., et al. (2019). North American extreme precipitation events and related large-scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, *53*(11), 6835–6875. https://doi.org/10.1007/s00382-019-04958-z

Bayazit, M., & Önöz, B. (2007). To prewhiten or not to prewhiten in trend analysis? *Hydrological Sciences Journal*, *52*(4), 611–624. https://doi.org/10.1623/hysj.52.4.611

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful

Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Blundell, R., Griffith, R., & Windmeijer, F. (2002). Individual effects and dynamics in count data models. *Journal of Econometrics*, *108*(1), 113–131. https://doi.org/10.1016/S0304-4076(01)00108-7

Boggess, J. M., Becker, G. W., & Mitchell, M. K. (2014). Storm & flood hardening of electrical substations. In *2014 IEEE PES T&D Conference and Exposition* (pp. 1–5). https://doi.org/10.1109/TDC.2014.6863387

Boucher, J.-P., Denuit, M., & Guillen, M. (2008). Models of Insurance Claim Counts with Time Dependence Based on Generalisation of Poisson and Negative Binomial Distributions. *Variance*, *2*(1), 135–162.

Bucar, R. C. B., & Hayeri, Y. M. (2020). Quantitative assessment of the impacts of disruptive precipitation on surface transportation. *Reliability Engineering & System Safety*, *203*, 107105. https://doi.org/10.1016/j.ress.2020.107105

Cann, K. F., Thomas, D. R., Salmon, R. L., Wyn-Jones, A. P., & Kay, D. (2013). Extreme water-related weather events and waterborne disease. *Epidemiology and Infection*, *141*(4), 671–686. https://doi.org/10.1017/S0950268812001653

Daniel, J. S., Portmann, R. W., Solomon, S., & Murphy, D. M. (2012). Identifying weekly cycles in meteorological variables: The importance of an appropriate statistical analysis. *Journal of Geophysical Research: Atmospheres*, *117*(D13). https://doi.org/10.1029/2012JD017574

Deidda, R., & Puliga, M. (2006). Sensitivity of goodness-of-fit statistics to rainfall data rounding off. *Physics and Chemistry of the Earth, Parts A/B/C*, *31*, 1240–1251. https://doi.org/10.1016/j.pce.2006.04.041

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., & Vose, R. S. (2010). Comprehensive

Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology*, *49*(8), 1615–1633. https://doi.org/10.1175/2010JAMC2375.1

El-Shaarawi, A. H., & Niculescu, S. P. (1992). On kendall's tau as a test of trend in time series data. *Environmetrics*, *3*(4), 385–411. https://doi.org/10.1002/env.3170030403

Emori, S., & Brown, S. J. (2005). Dynamic and thermodynamic changes in mean and extreme precipitation under changed climate. *Geophysical Research Letters*, *32*(17). https://doi.org/10.1029/2005GL023272

Fisher, R. A. (1922). On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94. https://doi.org/10.2307/2340521

Gallant, A. E., Hennessy, K., & Risbey, J. S. (2007). Trends in rainfall indices for six Australian regions: 1910-2005. *Australian Meteorological Magazine*, *56*, 223–239.

Gauthier, T. D. (2001). Detecting Trends Using Spearman's Rank Correlation Coefficient. *Environmental Forensics*, *2*(4), 359–362. https://doi.org/10.1006/enfo.2001.0061

Gershunov, A., Benmarhnia, T., & Aguilera, R. (2018). Human health implications of extreme precipitation events and water quality in California, USA: a canonical correlation analysis. *The Lancet Planetary Health*, *2*, S9. https://doi.org/10.1016/S2542-5196(18)30094-9

Gourieroux, C., & Jasiak, J. (2004). Heterogeneous INAR(1) model with application to car insurance. *Insurance: Mathematics and Economics*, *34*(2), 177–192. https://doi.org/10.1016/j.insmatheco.2003.11.005

Groisman, P. Y., Knight, R. W., Easterling, D. R., Karl, T. R., Hegerl, G. C., & Razuvaev, V. N. (2005). Trends in Intense Precipitation in the Climate Record. *Journal of Climate*, *18*(9), 1326–1350. https://doi.org/10.1175/JCLI3339.1

Hamed, K. H. (2009). Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. *Journal of Hydrology*, *368*(1), 143–155. https://doi.org/10.1016/j.jhydrol.2009.01.040

Hamed, K. H., & Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1), 182–196. https://doi.org/10.1016/S0022-1694(97)00125-X

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*(4). https://doi.org/10.1029/2010RG000345

Hennessy, K. J., Suppiah, R., & Page, C. M. (1999). Australian rainfall changes, 1910–1995. *Aust. Met. Mag.*, *48*, 1–13.

Hoerling, M., Eischeid, J., Perlwitz, J., Quan, X.-W., Wolter, K., & Cheng, L. (2016). Characterizing Recent Trends in U.S. Heavy Precipitation. *Journal of Climate*, *29*(7), 2313–2332. https://doi.org/10.1175/JCLI-D-15-0441.1

Hooper, E., Chapman, L., & Quinn, A. (2014). The impact of precipitation on speed–flow relationships along a UK motorway corridor. *Theoretical and Applied Climatology*, *117*(1), 303–316. https://doi.org/10.1007/s00704-013-0999-5

Hughes, L. (2003). Climate change and Australia: Trends, projections and impacts. *Austral Ecology*, *28*(4), 423–443. https://doi.org/10.1046/j.1442-9993.2003.01300.x

Janssen, E., Wuebbles, D. J., Kunkel, K. E., Olsen, S. C., & Goodman, A. (2014). Observational- and model-based trends and projections of extreme precipitation over the contiguous United States. *Earth's Future*, *2*(2), 99–113. https://doi.org/10.1002/2013EF000185

Jung, R. C., & Tremayne, A. R. (2011). Convolution-closed models for count time series with applications. *Journal of Time Series Analysis*, *32*(3), 268–280. https://doi.org/10.1111/j.1467-9892.2010.00697.x

Katz, R. W. (1988). Statistical Procedures for Making Inferences about Climate Variability. *Journal of Climate*, *1*(11), 1057–1064. https://doi.org/10.1175/1520-0442(1988)001<1057:SPFMIA>2.0.CO;2

Katz, R. W., & Brown, B. G. (1991). The problem of multiplicity in research on teleconnections.

International Journal of Climatology, *11*(5), 505–513. https://doi.org/10.1002/joc.3370110504

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1/2), 81–93. https://doi.org/10.2307/2332226

Kendall, M. G. (1975). *Rank correlation methods*. London: Griffin.

Khaliq, M. N., Ouarda, T. B. M. J., Gachon, P., Sushama, L., & St-Hilaire, A. (2009). Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of Canadian rivers. *Journal of Hydrology*, *368*(1), 117–130. https://doi.org/10.1016/j.jhydrol.2009.01.035

Koutsoyiannis, D. (2011). Hurst-Kolmogorov Dynamics and Uncertainty1. *Journal of the American Water Resources Association*, *47*(3), 481–495. https://doi.org/10.1111/j.1752-1688.2011.00543.x

Kruger, A., & Nxumalo, M. (2017). Historical rainfall trends in South Africa: 1921–2015. *Water SA*, *43*, 285. https://doi.org/10.4314/wsa.v43i2.12

Kunkel, K. E., & Frankson, R. M. (2015). Global Land Surface Extremes of Precipitation: Data Limitations and Trends. *Journal of Extreme Events*, *02*(02), 1550004. https://doi.org/10.1142/S2345737615500049

Kunkel, K. E., Karl, T. R., Squires, M. F., Yin, X., Stegall, S. T., & Easterling, D. R. (2020). Precipitation Extremes: Trends and Relationships with Average Precipitation and Precipitable Water in the Contiguous United States. *Journal of Applied Meteorology and Climatology*, *59*(1), 125–142. https://doi.org/10.1175/JAMC-D-19-0185.1

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., & Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. *Global Change Biology*, *25*(7), 2325–2337. https://doi.org/10.1111/gcb.14628

Livezey, R. E., & Chen, W. Y. (1983). Statistical Field Significance and its Determination by Monte

Carlo Techniques. *Monthly Weather Review*, *111*(1), 46–59. https://doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2

Madsen, H., Lawrence, D., Lang, M., Martinkova, M., & Kjeldsen, T. R. (2014). Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. *Journal of Hydrology*, *519*, 3634–3650. https://doi.org/10.1016/j.jhydrol.2014.11.003

Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, *13*(3), 245–259. https://doi.org/10.2307/1907187

McKenzie, Ed. (1985). Some Simple Models for Discrete Variate Time Series. *Journal of the American Water Resources Association*, *21*(4), 645–650. https://doi.org/10.1111/j.1752-1688.1985.tb05379.x

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

New, M., Hewitson, B., Stephenson, D. B., Tsiga, A., Kruger, A., Manhique, A., et al. (2006). Evidence of trends in daily climate extremes over southern and west Africa. *Journal of Geophysical Research: Atmospheres*, *111*(D14). https://doi.org/10.1029/2005JD006289

Nie, J., Sobel, A. H., Shaevitz, D. A., & Wang, S. (2018). Dynamic amplification of extreme precipitation sensitivity. *Proceedings of the National Academy of Sciences*, *115*(38), 9467. https://doi.org/10.1073/pnas.1800357115

Papalexiou, S. M., & Montanari, A. (2019). Global and Regional Increase of Precipitation Extremes Under Global Warming. *Water Resources Research*, *55*(6), 4901–4914. https://doi.org/10.1029/2018WR024067

Papalexiou, S. M., Rajulapati, C. R., Clark, M. P., & Lehner, F. (2020). Robustness of CMIP6 Historical Global Mean Temperature Simulations: Trends, Long-Term Persistence, Autocorrelation, and Distributional Shape. *Earth's Future*, *8*(10), e2020EF001667. https://doi.org/10.1029/2020EF001667

Pascual, F. G., & Akhundjanov, S. B. (2019). Monitoring a bivariate INAR(1) process with application to Hepatitis A. *Communications in Statistics - Theory and Methods*, 1–23. https://doi.org/10.1080/03610926.2019.1645856

Pedeli, X., Davison, A. C., & Fokianos, K. (2015). Likelihood Estimation for the INAR(p) Model by Saddlepoint Approximation. *Journal of the American Statistical Association*, *110*(511), 1229–1238. https://doi.org/10.1080/01621459.2014.983230

Peden, A., Franklin, R., Leggat, P., & Aitken, P. (2017). Causal Pathways of Flood Related River Drowning Deaths in Australia. *PLoS Currents*, *1*. https://doi.org/10.1371/currents.dis.001072490b201118f0f689c0fbe7d437

Rosenzweig, C., Tubiello, F., Goldberg, R., Mills, E., & Bloomfield, J. (2004). Increased crop damage in the US from excess precipitation under climate change. *Global Environmental Change*, *12*, 197–202. https://doi.org/10.1016/S0959-3780(02)00008-0

Serinaldi, F., & Kilsby, C. G. (2016). The importance of prewhitening in change point analysis under persistence. *Stochastic Environmental Research and Risk Assessment*, *30*(2), 763–777. https://doi.org/10.1007/s00477-015-1041-5

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

Smith, A. (2021). *2020 U.S. Billion-Dollar Weather and Climate Disasters—In Historical Context*. https://doi.org/10.13140/RG.2.2.25871.00166/1

Steutel, F. W., & van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability. *The Annals of Probability*, *7*(5), 893–899.

von Storch, H. (1999). Misuses of Statistical Analysis in Climate Research. In H. von Storch & A. Navarra (Eds.), *Analysis of Climate Variability* (pp. 11–26). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sun, F., Roderick, M. L., & Farquhar, G. D. (2018). Rainfall statistics, stationarity, and climate change. *Proceedings of the National Academy of Sciences*, *115*(10), 2305. https://doi.org/10.1073/pnas.1705349115

Tramblay, Y., El Adlouni, S., & Servat, E. (2013). Trends and variability in extreme precipitation indices over Maghreb countries. *Natural Hazards and Earth System Sciences Discussions*, *1*. https://doi.org/10.5194/nhessd-1-3625-2013

Trenberth, K. E. (2011). Attribution of climate variations and trends to human influences and natural variability. *WIREs Climate Change*, *2*(6), 925–930. https://doi.org/10.1002/wcc.142

Trenberth, K. E., Dai, A., Rasmussen, R. M., & Parsons, D. B. (2003). The Changing Character of Precipitation. *Bulletin of the American Meteorological Society*, *84*(9), 1205–1218. https://doi.org/10.1175/BAMS-84-9-1205

Vogel, R. M., Rosner, A., & Kirshen, P. H. (2013). Brief Communication: Likelihood of societal preparedness for global change: trend detection. *Natural Hazards and Earth System Sciences*, *13*(7), 1773–1778. https://doi.org/10.5194/nhess-13-1773-2013

Weiß, C. H. (2008). Serial dependence and regression of Poisson INARMA models. *Journal of Statistical Planning and Inference*, *138*(10), 2975–2990. https://doi.org/10.1016/j.jspi.2007.11.009

Westra, S., Alexander, L., & Zwiers, F. (2013). Global Increasing Trends in Annual Maximum Daily Precipitation. *Journal of Climate*, *26*, 7834. https://doi.org/10.1175/JCLI-D-12-00502.1

Wilks, D. S. (1997). Resampling Hypothesis Tests for Autocorrelated Fields. *Journal of Climate*, *10*(1), 65–82. https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2

Wilks, D. S. (2006). On “Field Significance” and the False Discovery Rate. *Journal of Applied Meteorology and Climatology*, *45*(9), 1181–1189. https://doi.org/10.1175/JAM2404.1

Wilks, D. S. (2016). “The Stippling Shows Statistically Significant Grid Points”: How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It. *Bulletin of the American Meteorological Society*, *97*, 160309141232001. https://doi.org/10.1175/BAMS-D-15-00267.1

Wilks, D. S. (2019). Chapter 7 - Statistical Forecasting. In *Statistical Methods in the Atmospheric Sciences (Fourth Edition)* (pp. 235–312). Elsevier. https://doi.org/10.1016/B978-0-12-815823-4.00007-9

Wright, D. B., Bosma, C. D., & Lopez-Cantu, T. (2019). U.S. Hydrologic Design Standards Insufficient Due to Large Increases in Frequency of Rainfall Extremes. *Geophysical Research Letters*, *46*(14), 8144–8153. https://doi.org/10.1029/2019GL083235

Yue, S., & Wang, C. Y. (2002). Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test. *Water Resources Research*, *38*(6), 4–1. https://doi.org/10.1029/2001WR000861

Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002). The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes*, *16*(9), 1807–1829. https://doi.org/10.1002/hyp.1095

Zhang, W., & Villarini, G. (2019). On the weather types that shape the precipitation patterns across the U.S. Midwest. *Climate Dynamics*, *53*(7), 4217–4232. https://doi.org/10.1007/s00382-019-04783-4

Zolotokrylin, A., & Cherenkova, E. (2017). Seasonal changes in precipitation extremes in russia for the last several decades and their impact on vital activities of the human population. *Geography, Environment, Sustainability*, *10*, 69–82. https://doi.org/10.24057/2071-9388-2017-10-4-69-82

Figures

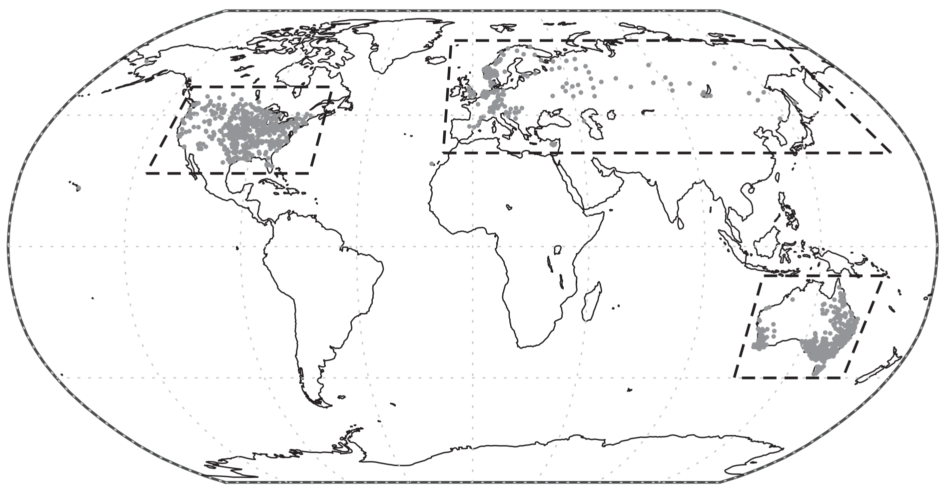

Figure 1. GHCN rain gauges selected for this study with indication of the three regions of (i) North America, (ii) Europe and Asia, and (iii) Australia displayed in subsequent figures.

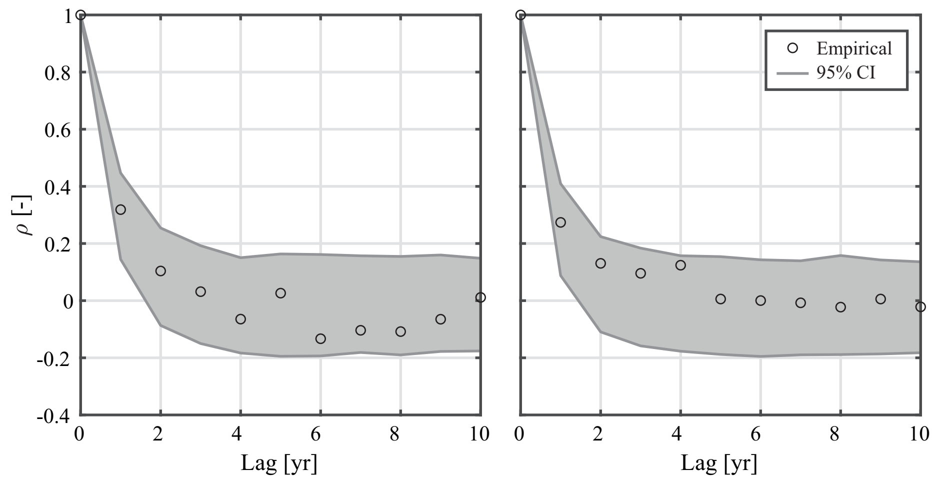

Figure 2. Examples of empirical autocorrelation function of two randomly chosen observed count time series (q = 0.95, n = 100 years) of the GHCN dataset along with 95% confidence interval (CI) derived from 10,000 synthetic time series generated with the Poisson-INAR(1) model. For both series, the slope of the linear trend is smaller than 0.02 events/yr.

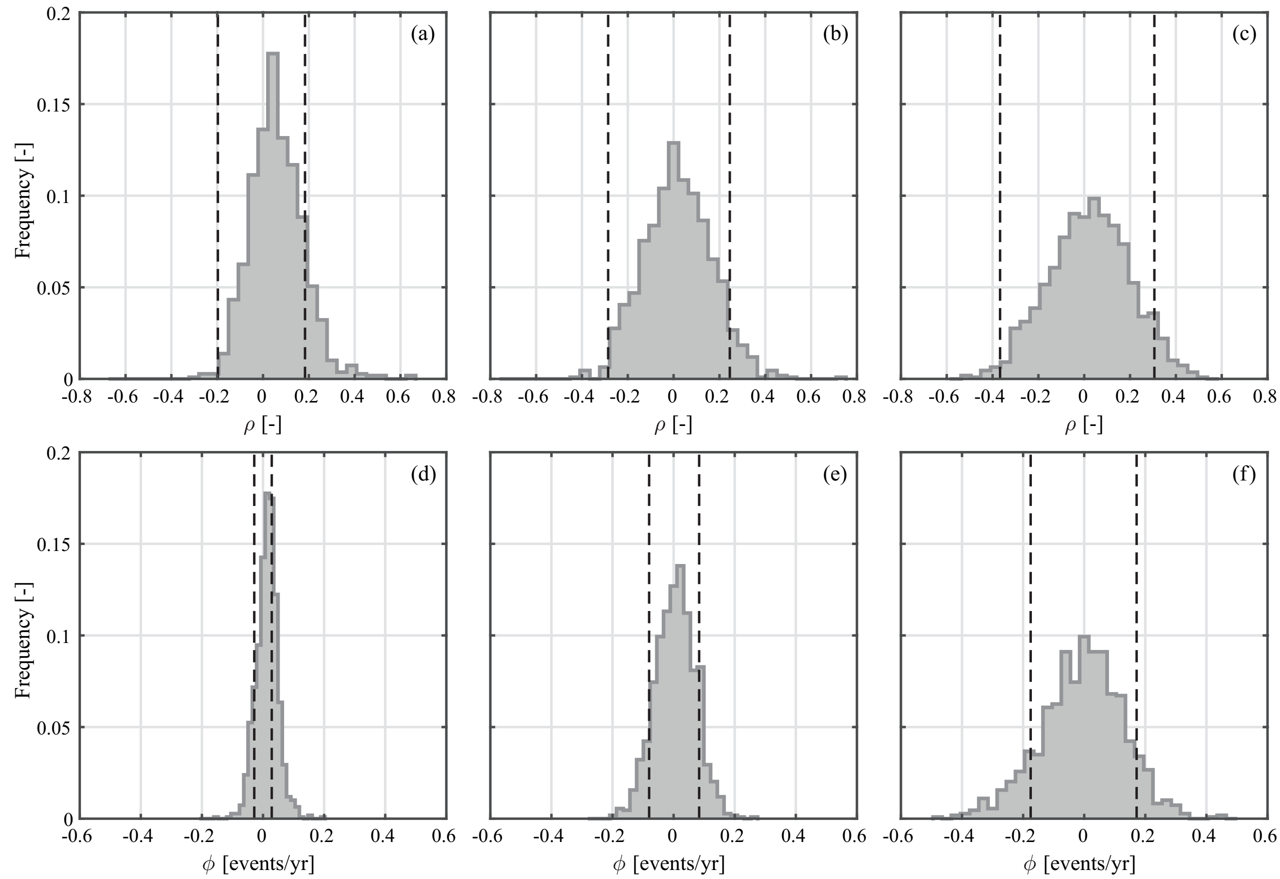

Figure 3. Histograms of (a)-(c) lag-1 autocorrelation ρ and (d)-(f) linear trend slope ϕ estimated on the M = 1087 observed count time series for q = 0.95 and sample size n = 100, 50, and 30 years (from left to right). Vertical lines depict the 95% confidence intervals obtained from 10,000 synthetic uncorrelated and stationary time series (H_0: "$\rho_0 = 0$; $\phi_0 = 0$").

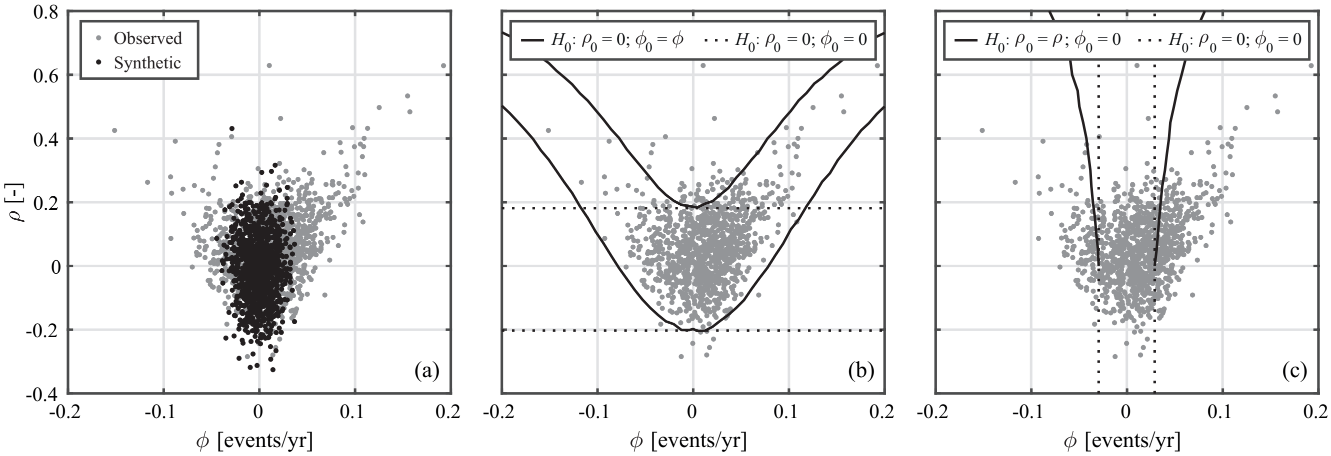

Figure 4. Scatterplot between ϕ and ρ computed on M = 1087 observed count time series for q = 0.95 and n = 100 years (grey circles) along with: (a) scatterplot between ϕ and ρ calculated on synthetic counts with H_0: "$\rho_0 = 0$; $\phi_0 = 0$" (black circles); (b) 95% CIs of ρ computed under H_0: "$\rho_0 = 0$; $\phi_0 = \phi$" (solid line) with ϕ being the value in the x-axis, and H_0: "$\rho_0 = 0$; $\phi_0 = 0$" (dashed line); (c) 95% CIs of ϕ computed under H_0: "$\rho_0 = \rho$; $\phi_0 = 0$" (solid line) with ρ being the value in the y-axis, and H_0: "$\rho_0 = 0$; $\phi_0 = 0$" (dashed line).

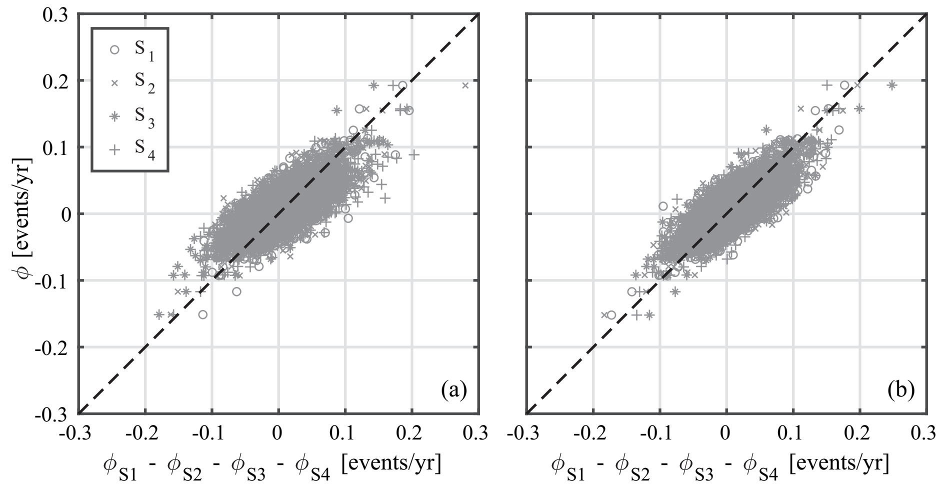

Figure 5. (a) Scatterplot of linear slopes ϕ estimated on M = 1087 observed count time series for q = 0.95 and n = 100 years versus linear slopes ϕ_{S*} (* = 1, 2, 3, 4) estimated on the corresponding 4 x M sub-series of n = 25 years extracted from each full series by sampling one record every four years and denoted with S1-S4. (b) Same as (a) but for synthetic time series generated under H_0: "$\rho_0 = 0$; $\phi_0 = \phi_{obs}$", with ϕ_{obs} being the observed slope.

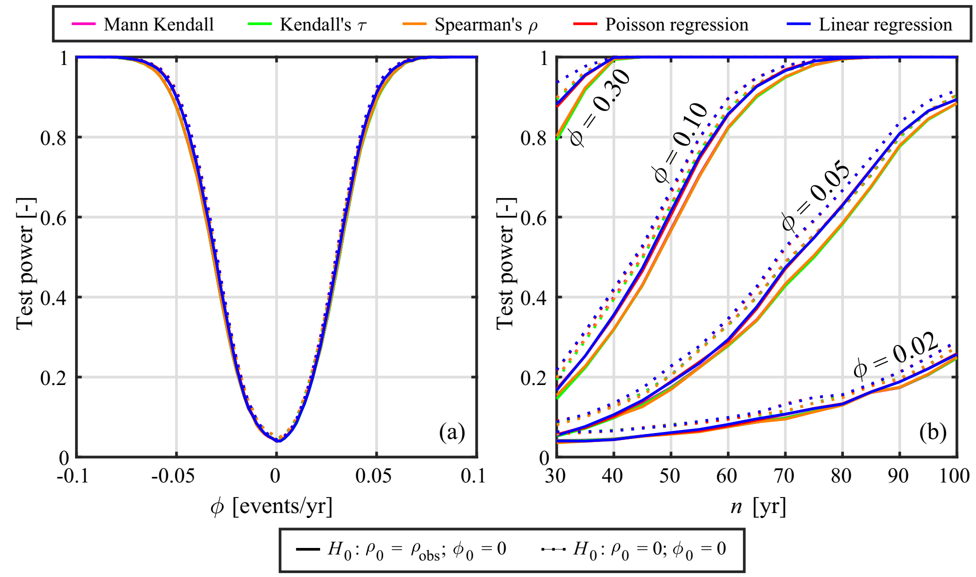

Figure 6. Performances of several trend tests with the null distribution of the test statistics built under H_0: "$\rho_0 = 0$; $\phi_0 = 0$" (dashed line) and H_0: "$\rho_0 = \rho_{\mathrm{obs}}$; $\phi_0 = 0$" (solid line), evaluated on synthetic count time series relative to $q = 0.95$. (a) Power of tests as a function of ϕ for uncorrelated nonstationary time series for length $n = 100$ years. (b) Power of tests as a function of n for uncorrelated nonstationary time series for $\phi = 0.02$, 0.05, 0.10, and 0.30 events/yr.

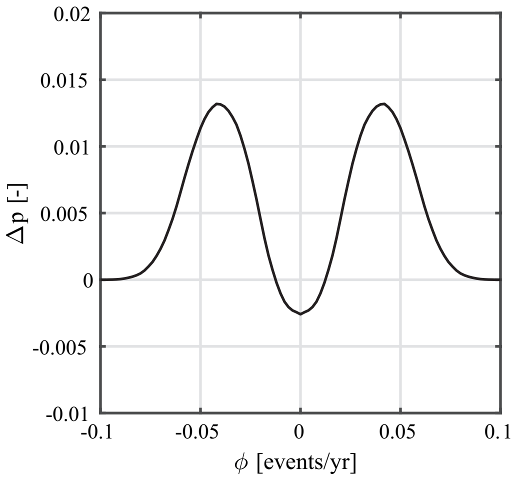

Figure 7. Gaussian-weighted moving average of the differences between power of PR and MK tests (indicated with Δp) reported in Fig. 6a for q = 0.95 and n = 100 years.

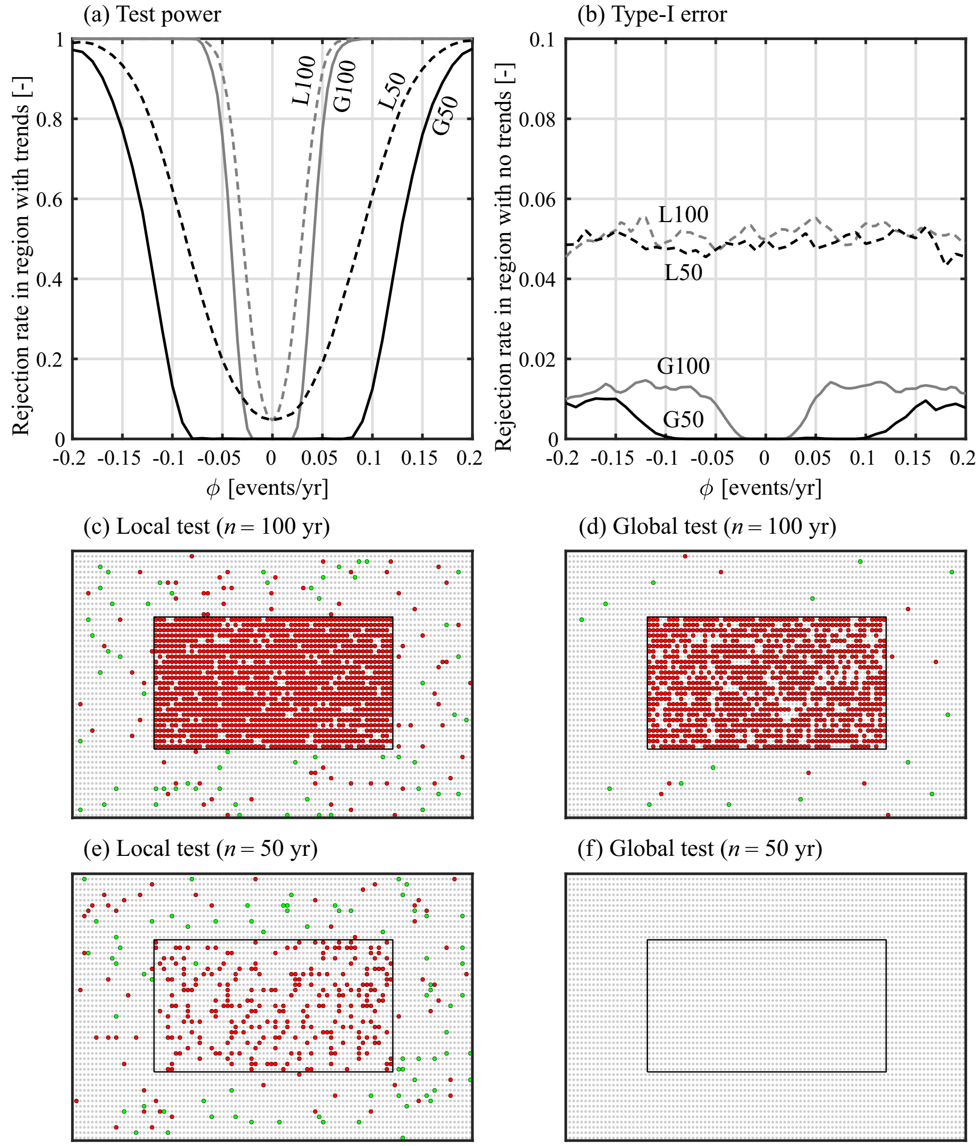

Figure 8. Performance of trend test at multiple sites quantified through a synthetic experiment in a 50 x 100 grid points (see text for details). (a) Fraction of local (L) and global (G) rejections of H_0 as a function of ϕ in the inner region with trend (test power) for n = 100 and 50 years. (b) Same as (a) but for the outer region with no trend (type-I error). (c) Map of local rejections of H_0 for the case where an increasing trend with slope ϕ = 0.05 events/yr is assumed in the inner region and n = 100 years. (d) Same as (c), but for global rejections of H_0 after applying the FDR test. (e)-(f) same as (c)-(d), but for n = 50 years. In (c)-(f), red (green) dots represent rejections of H_0 with increasing (decreasing) trend, while grey dots are used when H_0 is not rejected.

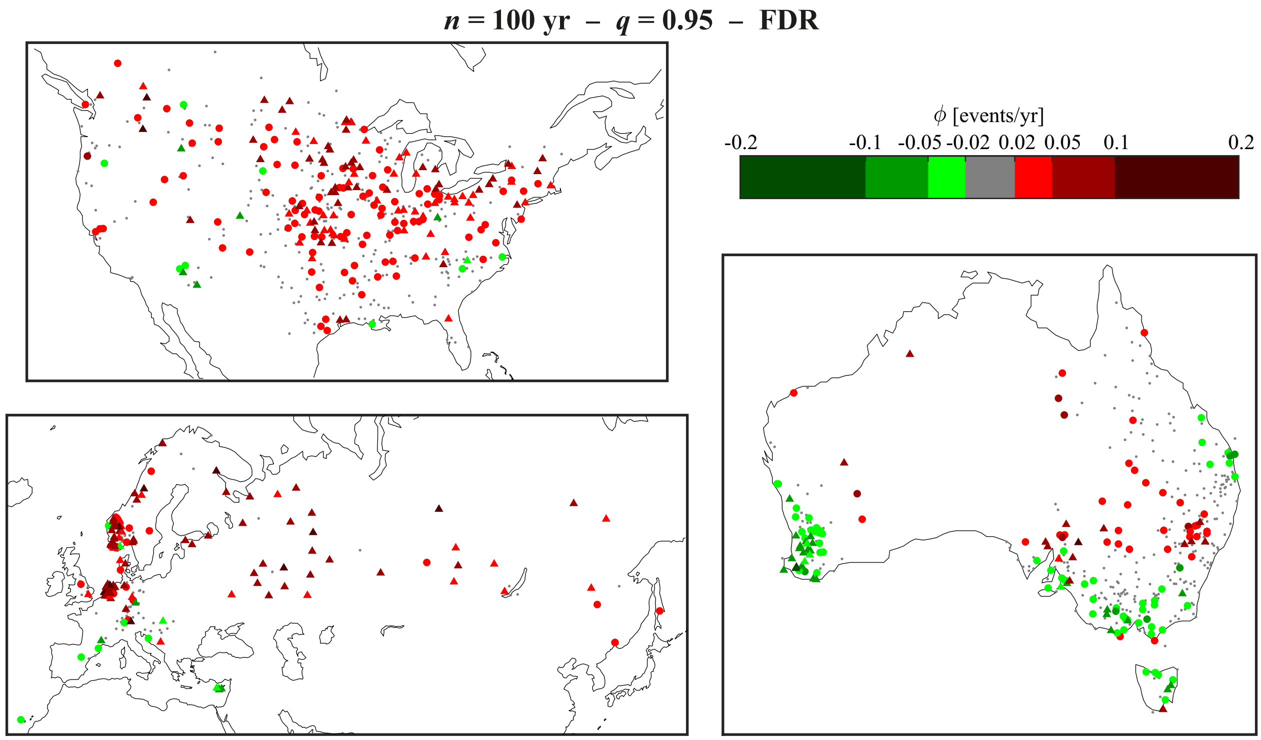

Figure 9. Statistically significant trends at the GHCN gages after applying the FDR tests at $\alpha_{\text{global}} = 0.05$ for $n = 100$ and $q = 0.95$. Larger circles (triangles) are used when H_0 is rejected by PR only (PR and MK), with colors based on the trend slope value and sign. Smaller grey dots are used when H_0 is not rejected by both tests.

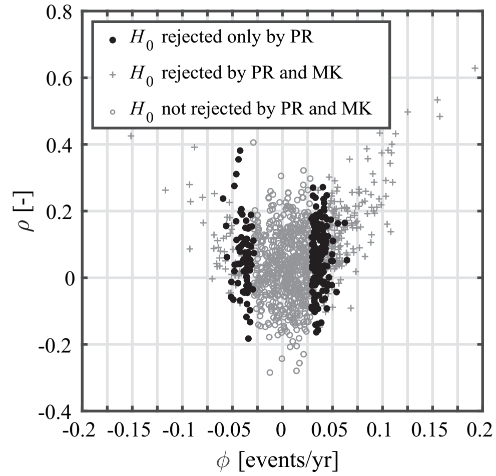

Figure 10. Scatterplot between ϕ and ρ computed on M = 1087 observed count time series for q = 0.95 and n = 100 years, with different markers visualizing possible combined outcomes of PR and MK tests.

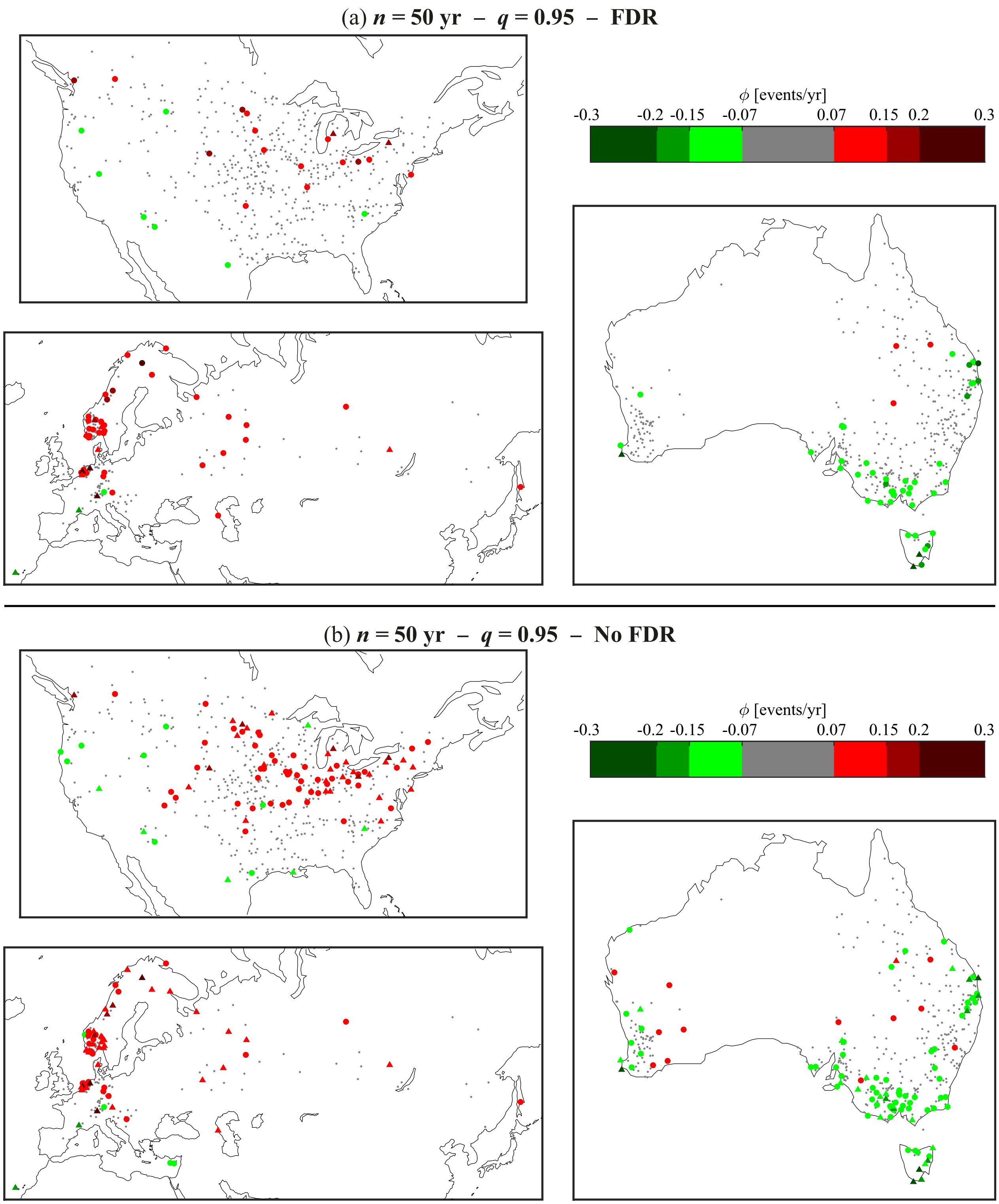

Figure 11. (a) As in Fig. 9 but for n = 50 yr. (b) As in (a) but for local results without the application of the FDR test.

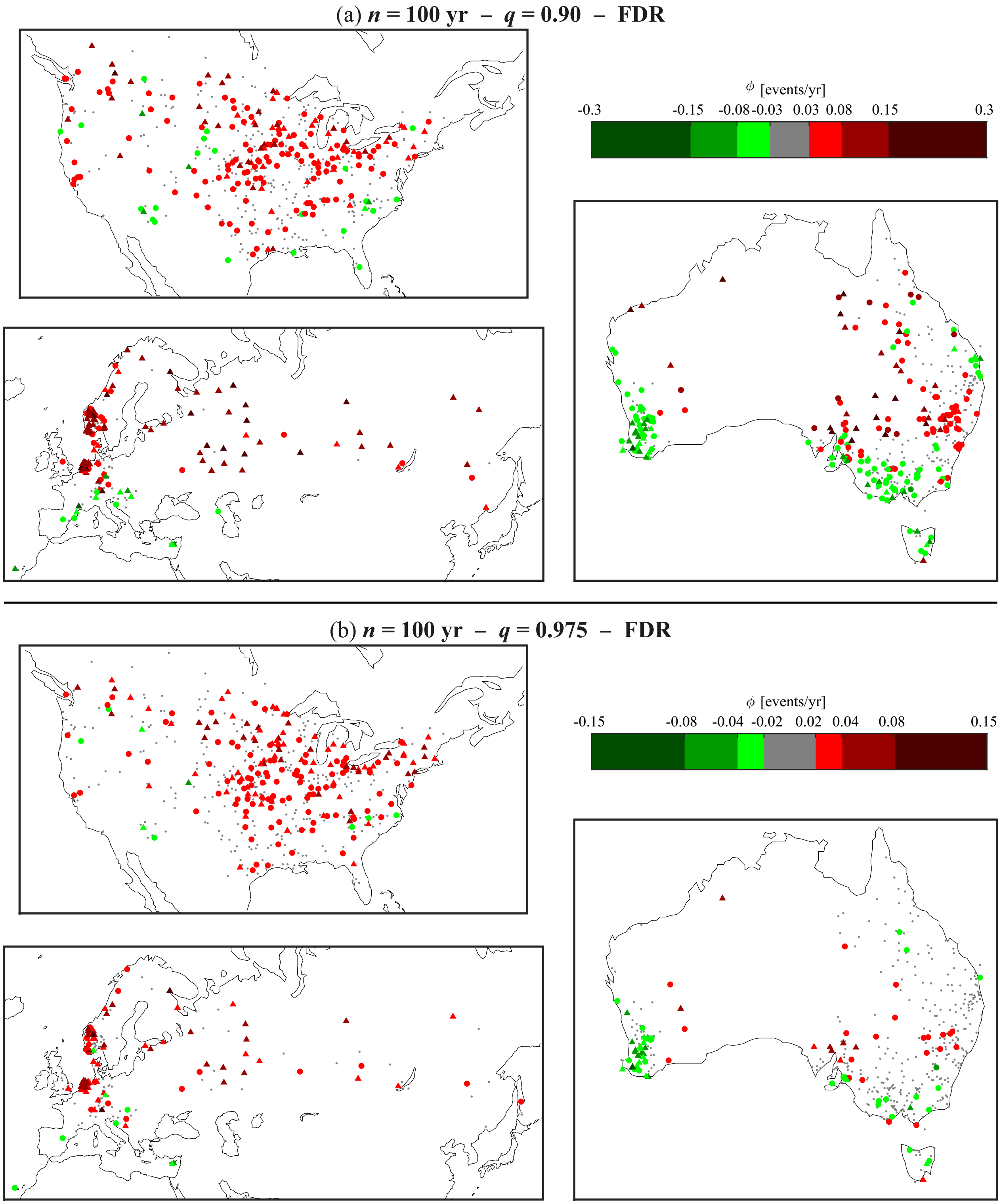

Figure 12. As Fig. 9 but for (a) q = 0.90 and (b) q = 0.975.

Reference	Spatial coverage	Time Period	Number of years	Data	Spatial aggregation	Metrics	Statistical techniques	Account for field significance	Account for serial correlation	Main results
Groisman et al., (2005)	Global	1893-2002	32 to 110	NCDC rain gages	Sub-continental regions	AF_{POT} and AT_{POT}	SP	No	No significant observed ρ's	Statistically significant increasing (decreasing) trends in Europe, America, South-Africa (southwestern Australia).
Alexander et al., (2006)	Global	1901-2003	52 to 102	GHCN, ECA and GSN rain gages (5948)	Point and 5° x 5° grids	11 PCCIs	MK	Bootstrapping	Prewhitening	Trends on extreme events mainly increasing, but statistically significant on 13%–37% of the stations depending on the metric.
Westra et al., (2013)	Global	1900-2009	30 to 110	HadEX2 rain gages (8326)	Point	AM	MK	Bootstrapping	Negligible mean ρ	- Increasing trends in 2/3 of rain gages, but only 8.6% statistically significant; - No evident spatial patterns of significant trends.
Kunkel & Frankson, (2015)	Global	1951-2014	64	GHCN rain gages (6619)	10° x 10° grids	AF_{POT} and AT_{POT}	KT	No	Trend test estimating variance inflation	- Most trends not statistically significant. - Increasing trends in most of the world except western North America, southern Europe, northern Eurasia and western and eastern coasts of Australia.
Asadieh & Krakauer, (2015)	Global	1901-2010	30 to 110	HadEX2 rain gages (~11,600); CMIP5 outputs	2.54° x 3.75° grids	AM	MK	No	No	- Increasing (decreasing) trends in 66.2% (33.8%) of grid cells, but only 18% (4%) statistically significant; - Consistent results with CMIP5 outputs but with trend underestimation.
Papalexiou & Montanari, (2019)	Global	1964-2013	50	GHCN rain gages (8730)	Point and 5° x 5° grids	AF_{POT} and AM_{POT}	MC	No	AF: negligible mean ρ AMM: use of AR(1) in Monte Carlo simulations	- Coherent spatial patterns of trends more evident in frequency (AF_{POT}) than magnitude (AM_{POT}); - For AF_{POT}: increasing trends in central and eastern USA, Europe, eastern Russia and most of China; - For AM_{POT}: increasing trends in western and northern Europe, and eastern and central USA; - Ratio of increasing/decreasing statistically significant trends: 2.4 (1.3) for AF_{POT} (AM_{POT}).
Janssen et al., (2014)	USA	1901-2012	112	HadEX2 rain gages (726) CMIP5 outputs	Sub-continental regions and 1° x 1° grids	AF_{POT}	PD	No	No	- Statistically significant increasing (decreasing) trends in central and eastern (western) USA; - Consistent results from CMIP5 but with a trend underestimation.
Hoerling et al., (2016)	USA	1901-2013; 1979-2013.	35 to 114	GHCN rain gages (~10,000)	Sub-continental regions	AT_{POT}, AF_{POT} and AM_{POT}	ND	No	No	- Statistically significant increasing (decreasing) trends in 1901-2013 in the northeastern (southwestern) USA; - Similar patterns for the 1979-2013 period.
Wright et al., (2019)	USA	1950-2017	68	GHCN rain gages (911)	Sub-continental regions	AF_{POT}	NB	No	No	Statistically significant increasing trends in eastern USA, smaller and less significant changes in western parts

Kunkel et al., (2020)	USA	- 1949-2016; - 1979-2016.	37 to 68	GHCN rain gages (3098)	Sub-continental regions	AF_{POT} and AT	KT	No	No	- In 1949-2016, statistically significant increasing trends in several areas of USA; statistically significant decreasing trends in western USA, but in lower number; - Similar patterns in 1979-2016 but lower number of significant trends.
Kruger & Nxumalo, (2017)	South Africa	1921-2015	95	Rain gages (60)	Point and rainfall districts	11 PI	*t*-test	No	No	Statistically significant increasing trends in indices related to extreme events in southern and middle regions.
New et al., (2006)	South Africa	1961-2000	30 to 40	Rain gages (63)	Point and regions	10 PI	KT	No	No	- Increasing trends in indices related to extreme events at regional scale; - Few statistically significant trends and no evident spatial patterns at local scale.
Tramblay et al., (2013)	North-Africa	1950-2008	33 to 59	Rain gages (22)	Point	11 PI	MK	FDR test	Trend test accounting for variance inflation	- Few decreasing trends on indices related to extreme events; - Statistically significant trends with local tests, but in much lower number after applying the FDR test.
Hennessy et al., (1999)	Australia	1910-1995	86	Rain gages (379)	Countries	Several PI	KT	No	No	Statistically significant increasing (decreasing) trends on indices focused on extreme events in South Australia and New South Wales (Western Australia) regions.
Hughes, (2003)	Australia	-	-	-	-	Review	-	-	-	Increasing trends in most areas, decreasing trends in southwestern and southeastern regions.
Gallant et al., (2007)	Australia	- 1910-2005; - 1950-2005.	56 to 96	Rain gages (92)	Six regions	Several PI	KT	No	No	Statistically significant increasing (decreasing) trends on indices related to extreme events in central (southwestern and southeastern) regions.
Alpert et al., (2002)	Mediterra-nean Basin	1951-1995	45	Rain gages (265)	Countries	AT_{POT}	SP	FDR test	No	Statistically significant increasing trends in two of the four considered countries.
Madsen et al., (2014)	Europe	-	-	-	-	Review	-		-	Overall increase both in frequency and magnitude of extreme precipitation, especially in northern parts.
Zolotokrylin & Cherenkova, (2017)	Russia	1961-2013	53	Rain gages (527)	Point	SF_{POT} and ST_{POT}	Not defined	No	No	Statistically significant increasing trends in 2/3 of rain gages for all seasons.

Table 1. Summary of several empirical trend analyses of precipitation extremes performed at global and regional scale with daily records. Acronyms for datasets, metrics and statistical tests are defined as follows. (1) Datasets: NCDC = National Climatic Data Center; GHCN = Global Historical Climate Network; ECA = European Climate Assessment; GCN = GCOS (Global Observing System for Climate) Surface Network; HadEX2 = Hadley Center Global Climate Extremes Index 2; CMIP5 =Coupled Model Intercomparison Project 5. (2) Metrics: AM = Annual maxima; AF_{POT} (SF_{POT}) = Annual (seasonal) frequencies in peak-over-threshold (POT) series; AT_{POT} (ST_{POT}) = Annual totals of exceedances in POT series; AM_{POT} = Annual average magnitude of exceedances in POT series; AT (ST) = Annual (seasonal) totals; PI = Precipitation-based indices. (3) Statistical tests: MK = Mann-Kendall test; KT = Kendall's τ test; SP = Spearman's ρ test; PD = Poisson's distribution-based test; NB = Negative binomial regression; MC = Monte Carlo simulations.

	n = 100	n = 50	n = 30
Significant ρ's for H_0: "$\rho_0 = 0$; $\phi_0 = 0$"			
Local test	156 (14%)	77 (7%)	80 (7%)
FDR test	29 (3%)	3 (0%)	0 (0%)
Significant ϕ's for H_0: "$\rho_0 = 0$; $\phi_0 = 0$"			
Local test	467 (43%)	244 (22%)	193 (18%)
FDR test	451 (41%)	114 (10%)	94 (9%)

Table 2. Number and percentage of count series derived for $q = 0.95$ with significant ρ and/or ϕ for local and FDR tests assuming H_0: "$\rho_0 = 0$; $\phi_0 = 0$".