ABSTRACT: Studies detecting trends in climate elements typically concentrate on their local significance, ignoring the question of whether the significant local trends may or may not have occurred as a result of chance. This paper fills this gap by examining several approaches to detecting statistical significance of trends defined on a grid (i.e., on a regional scale). To this end, we introduce a novel simple procedure of significance testing that is based on counting signs of local trends (sign test), and we compare it with five other approaches to testing collective significance of trends: counting, extended Mann–Kendall, Walker, false detection rate (FDR), and regression tests. Synthetic data are used to construct null distributions of trend statistics, to determine critical values of the tests, and to assess the performance of tests in terms of type-II error. For lower values of spatial and temporal autocorrelations, the sign test and extended Mann–Kendall test perform slightly better than the counting test; these three tests outperform the Walker, FDR, and regression tests by a wide margin. For high autocorrelations, which is a more realistic case, all tests become similar in their performance, with the exception of the regression test, which performs somewhat worse. Some tests cannot be used under specific conditions because of their construction: the Walker and FDR tests for high spatial autocorrelations, and the sign test under high spatial autocorrelations.

KEYWORDS: Statistical techniques; Time series; Trends

1. Introduction

Detection of ongoing and recent climate change is one of the most important tasks in present-day climatology. Climate change detection studies are conducted on various spatial scales of aggregation, from global means (e.g., Jones et al. 2012; Hawkins and Jones 2013; Jiménez-Muñoz et al. 2013; Karl et al. 2015; Rahmstorf et al. 2017) through continental and regional means (e.g., Giorgi 2002; Jones and Moberg 2003; Scherrer et al. 2005; de Barros Soares et al. 2017; Hong et al. 2017) to means for individual countries or their parts (e.g., Brázdil et al. 2009; El Kenawy et al. 2012; Vose et al. 2012; González-Hidalgo et al. 2016; Tao et al. 2017) and local time series. A typical study detecting climate change in local scale is performed at a set of stations (or, alternatively, on a regular grid) over a region, a country, or a group of countries, up to a continent. Linear trends (i.e., slopes of linear regression lines of the variable against time) are calculated and their statistical significance is assessed. Statistical significance usually means a rejection of the null hypothesis that there is no correlation between the climate variable and time. Only sites (stations or grid points) with significant trends are often discussed then, those with insignificant trends are neglected. The lack of significance may not indicate the absence of a trend, which is a frequent interpretation of such a situation; it may rather reflect the fact that the data series is simply not long enough for the trend, which is embedded in shorter-term climate fluctuations, to achieve significance. Arguments against the overuse of significance testing of trends similar to Nicholls (2001) were also raised by Griffiths et al. (2005) and Moberg and Jones (2005). Quite a little seems to have changed in the practice of local climate change detection studies since the time of publication of these papers, however. The dichotomous view, considering trends either significant or absent, leads to difficulties in interpretation, which are not admitted by most authors. For example, how to interpret the situation that just one of two close stations has a significant trend? Is climate change proceeding at one station but absent at the other? Or, more intriguingly, what if the trend is significant in maximum temperature and insignificant in minimum temperature at one station, which are not admitted by most authors. For example, Unkašević and Tosić (2013), Dumitrescu et al. (2015), Fathian et al. (2015), Caloiero (2017), Burger et al. (2018), and Caloiero et al. (2019). This common approach to trend detection and climate change analysis suffers from several weaknesses, which provide a motivation and a starting point for this study.

First, as stressed by Nicholls (2001), who criticized a misuse and overuse of significance testing in climate change studies, a considerable amount of information may be lost when insignificant trends are neglected. The lack of significance may not indicate the absence of a trend, which is a frequent interpretation of such a situation; it may rather reflect the fact that the data series is simply not long enough for the trend, which is embedded in shorter-term climate fluctuations, to achieve significance. Arguments against the overuse of significance testing of trends similar to Nicholls (2001) were also raised by Griffiths et al. (2005) and Moberg and Jones (2005). Quite a little seems to have changed in the practice of local climate change detection studies since the time of publication of these papers, however. The dichotomous view, considering trends either significant or absent, leads to difficulties in interpretation, which are not admitted by most authors. For example, how to interpret the situation that just one of two close stations has a significant trend? Is climate change proceeding at one station but absent at the other? Or, more intriguingly, what if the trend is significant in maximum temperature and insignificant in minimum temperature at one station, while the opposite is the case at a neighbor station? This is a realistic situation that appears, for example, on a map of annual temperature trends for Bangladesh in Rahman and Lateh (2016), without being noticed and commented on by the authors.

Second, the multiplicity of individual (local) tests is ignored. The question on the collective significance of local tests (also referred to as “global” or “field” significance), that is, whether the observed number of significant trends (i.e., of rejected null
This study deals with the detection of trends of one sign over the analyzed domain. Although it may seem to be a limitation, it is in fact a fairly common task since the expected manifestations of ongoing climate change as well as potential policy-relevant adverse effects are frequently related to one specific sign of a trend only. Simply speaking, it is more relevant to the society to know whether temperature is rising rather than whether it is changing; whether hot extremes occur more often rather than whether they change their frequency; whether wind becomes stronger rather than changing its speed; whether precipitation intensity is growing rather than changing; and whether drought periods are becoming longer rather than changing their duration (e.g., Spinoni et al. 2015b; Soulis et al. 2016; Brázdil et al. 2019; Finkel and Katz 2018). The fact that the detection of trends of one sign is a relevant task is supported by the fact that another test, extended Mann–Kendall test (which is included in our study), was developed for exactly the same purpose (e.g., Khaliq et al. 2009). On the other hand, the detection and assessment of significance of trends of both signs at the same time is also a common and policy-relevant task; for example, this relates to trends in pressure, frequency of synoptic events, or temperature variability (e.g., Gillett et al. 2005; Alexander and Perkins 2013; Kučerová et al. 2017).

2. Statistical tests

First of all, six statistical tests that allow for assessing regional-scale significance of trends defined at individual sites are described in this section. While the sign test and extended Mann–Kendall test are directly designed to test for trends of one sign in the domain, the other four tests—the counting test, Walker test, test based on the false detection rate (referred to as the FDR test here), and test of regression patterns—can be used for testing the significance of trends of both signs together. Three of the tests use local tests at individual sites as an input (counting test, Walker test, and FDR test), while the other three handle the data at all sites simultaneously (extended Mann–Kendall test, test of regression patterns, and the sign test, proposed by us). We start with the description of the local tests that we use in the first group of tests of collective significance, and then we list all the tests of collective significance. In the tests utilizing local tests, we suppose trends are evaluated for their significance locally at level \( \alpha_l \) at \( N \) sites.

a. Local tests

The classical \( t \) test, nonparametric Mann–Kendall test, Mann–Kendall test with prewhitening, and Mann–Kendall test with the equivalent sample size are employed as local tests for the counting, Walker, and FDR tests. All of the tests evaluate the relationship between the tested time series and the order of values in time (e.g., the year if the series is composed of annual values). The \( t \) test evaluates the linearity of the relationship and is based on the correlation coefficient \( r \) between the tested time series and the order of the values, \( k = 1, \ldots, K \): the quantity

\[
t = \frac{r}{\sqrt{1 - r^2}}\sqrt{K - 2}
\]

is \( t \)-distributed with \( K - 2 \) degrees of freedom.
The Mann–Kendall test (M-K test) is a nonparametric test for assessing significance of a trend, assuming its monotonicity, not linearity. The local M-K test statistic is the sum of signs of differences for all pairs of values: it is defined at site \( i \) for time series \( x_{ik}, k = 1, \ldots, K, \) as
\[
T_i = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \text{sgn}(x_{ik} - x_{il}).
\]
(2)
The test statistic is normally distributed with zero mean and variance
\[
\text{Var}(T_i) = \frac{K(K - 1)(2K + 5) - \sum_{m=1}^{M} q_m m(m-1)(2m + 5)}{18},
\]
(3)
where the sum over \( m \) accounts for the effect of ties: \( q_m \) is the number of ties of extent \( m \) and \( M \) is the number of tied groups.

The presence of temporal autocorrelation affects results of significance testing for the above local tests. Typically, it inflates test statistics so that the null hypothesis is rejected (and trend is detected) more often than it should correctly be. There are various approaches to cope with autocorrelation. Some of them handle it a posteriori, such as by replacing the real sample size by its effective size in the calculation of critical values (e.g., Wilks 2006a; Deceser et al. 2014; Tabari et al. 2015); some of them handle it prior to testing by modifications of the time series entering the test.

We take both approaches, using procedures that have been widely used in climatological literature. Temporal autocorrelation is removed by the prewhitening procedure that retains trend magnitude (Yue and Wang 2002a; Yue et al. 2002): the linear trend is removed from the time series first, then lag-1 autocorrelation \( r_1 \) is calculated in detrended series \( x'_i \) and the new prewhitened time series \( y_i \), from which the lag-1 autocorrelation is removed, is formed by
\[
y_i = x'_i - r_1 x'_{i-1}.
\]
(4)

Last, the trend is added to the prewhitened series. This new variable then enters the standard testing procedure.

The test with equivalent sample size is based on replacing the actual sample size \( K \) by effective sample size \( K' \), which accounts for the inflation of variance of the test statistic due to temporal autocorrelation. The general formula (e.g., Yue and Wang 2004; Tabari et al. 2015) reduces for the first-order autoregressive process to
\[
K^* = K \frac{1 - r_1}{1 + r_1},
\]
(5)
where \( r_1 \) is the lag-1 autocorrelation coefficient of the detrended series.

b. Counting test

The counting test is a classical test for collective significance, which was introduced into atmospheric sciences by Livezey and Chen (1983). Its essence is that the collective significance is achieved if significance is detected for a sufficiently large number of local tests.

If the null hypothesis of no trend is true the chance of wrongly rejecting the local null hypothesis (i.e., detecting a significant trend) at any site is \( \alpha_L \). For independent local tests, the number of locally significant trends under the null hypothesis follows binomial distribution: the probability that their number \( M \) exceeds preset value \( m \) is
\[
P(M \geq m) = \sum_{i=m}^{N} \frac{N!}{i!(N-i)!} \alpha_L^i (1 - \alpha_L)^{N-i}.
\]
(6)
The mean value of the distribution of \( M \) is \( N\alpha_L \). However, the number of locally significant tests \( m_c \) necessary to achieve collective significance level \( \alpha \) is larger than \( N\alpha_L \). The collective significance level, which in formal terms is the probability that the given number of local null hypotheses are falsely rejected, can be determined from Eq. (6) by putting its right-hand side equal to \( \alpha \). Livezey and Chen (1983) demonstrate that for 50 local tests, collective significance \( \alpha = 5\% \) is achieved if at least 12\% of local tests are significant at level \( \alpha_L = 5\% \). Even for 1000 tests, at least 7\% of local tests must be significant to achieve the collective significance of 5\%.

The above numbers apply to the situation that all local tests are independent, which is rarely the case in atmospheric sciences. For dependent tests, that is, in the presence of spatial autocorrelation between the sites where trends are evaluated, the test becomes too permissive, that is, rejects local null hypotheses too frequently. Consequently, the number of significant local tests needed to achieve the global significance is higher than in the case of independence. Wilks (2006a) demonstrates that the sensitivity of the counting test to the mutual dependence of local tests is high.

c. Walker test

The counting test has the property of not taking into account the magnitude of significance achieved by local tests, which may be considered a disadvantage. The same number of highly significant local tests and of the tests that exceed the local significance level only marginally result in the same collective significance according to the counting test.

The Walker test approaches the issue of collective significance in a different way. It is built on the premise that the local significance can be achieved even if just one of the local tests is sufficiently highly significant. The Walker test detects collective significance at level \( \alpha_C \) if the smallest \( p \) value of all local tests, \( p_{\min} \), satisfies the condition (Katz and Brown 1991; Wilks 2006a)
\[
p_{\min} \leq 1 - (1 - \alpha_C)^{1/N}.
\]
(7)

d. False detection rate

The test based on false detection rate (FDR) can be viewed of as a generalization of Walker test. It was introduced into atmospheric sciences by Ventura et al. (2004) and later advocated by Wilks (2006a, 2016). The test is based on Benjamini and Hochberg (1995) procedure, which, as Wilks (2006a) demonstrates, also provides a field significance test. Whereas the Walker test looks at the most significant local test only, the
FDR test takes into account the strength of significance of all local tests and denies the null hypothesis of no trend if any of the local tests yields a sufficiently small \( p \) value, the threshold depending on its order.

The FDR test detects collective significance if the following inequality holds for the \( p \) value of at least one local test:

\[
p \leq \max_{i=1}^{N} \left[ p_{(i)}; p_{(i)} \leq \frac{i}{N} \right],
\]

where \( p_{(i)} \) denotes sorted \( p \) values of local tests; that is, \( p_{(i)} \) is the \( i \)th-smallest \( p \) value.

A considerable advantage of Walker and FDR tests is that they are much less sensitive to the mutual dependence of local tests than the counting test (Wilks 2006a). Unlike the counting test, the FDR test tends to be conservative if local tests are not independent. Under high spatial autocorrelations, Wilks (2016) recommends to double the statistical significance for the numbers of trends of either sign follow binomial distribution with parameters \( p = 0.5 \) and \( N \) being the number of sites. The probability that the number of trends of the given sign (positive or negative), \( M \), exceeds the preset value \( n \) is

\[
P(M \geq n) = \sum_{i=n}^{N} \frac{N!}{i!(N-i)!} 0.5^N.
\]

The advantage of this test is its simplicity: no significance tests are needed at individual sites; just a sign of a trend must be calculated. Any method of determination of a trend, whether parametric or nonparametric, is applicable here.
3. Synthetic data

Null distributions of the test statistics outlined above are determined and the power of the tests is evaluated on a set of synthetic time series, which are generated for a regular grid consisting of 20 × 20 points. In generating these data, the local variance, temporal autocorrelation, spatial autocorrelation, and trend magnitude are prespecified.

The synthetic data are produced by the Spatial Generator for Trend Analysis stochastic weather generator (SPAGETTA; Dubrovský et al. 2020). Temporal and spatial autocorrelations are represented by a first-order multivariate autoregressive process, in which the modeled vector consists of values related to individual grid points (i.e., having 20 × 20 grid points, the dimension of the vector is 400). Spatial autocorrelations between grid points are assumed to isotropically decrease with their distance increasing:

\[ \text{cor}\{x_i(k), x_j(k)\} = C_0^{d(i,j)} \text{ and} \]
\[ \text{cor}\{x_i(k), x_j(k-1)\} = L_0^{d(i,j)} \]

where \(C_0<1\) and \(0<L_0<1\) are user-defined parameters, \(k\) is the time index, and \(d(i,j)\) is distance of any pair of grid points \(i\) and \(j\). If we set \(d=1\) for the neighboring (closest) grid points, the two above equations apparently imply that \(C_0\) represents spatial autocorrelation between two closest grid points, \(p\) and \(q\), at zero lag:

\[ \text{cor}\{x_p(k), x_q(k)\} = C_0 \]

and \(L_0\) represents lag-1 temporal autocorrelation at any grid point:

\[ \text{cor}\{x_i(k), x_i(k-1)\} = L_0. \]

For the sake of simplicity, we refer to \(C_0\) and \(L_0\) as spatial autocorrelation and temporal autocorrelation, respectively, in further text. Using the algorithm just described, synthetic series of 200 time units (which may represent years in our setting) long are generated. They are produced for a wide range of values of temporal and spatial autocorrelations (listed in Table 1), which cover values occurring in real world for various variables on grids of reasonable horizontal resolutions. Since for trend detection, the signal-to-noise ratio, which is equivalent to the ratio of the trend magnitude to the standard deviation, is relevant, we only list the ratio in Table 1. Standard deviation is identical for all time series.

The stationary series, that is, the data with the above statistical properties without any trend imposed, are used to construct null distributions and determine critical values of the statistical tests. The time series with imposed linear trends of various magnitudes (see Table 1) are then used to evaluate the power of the tests.

For each combination of values of statistical properties, 10,000 random time series that are 200 time units long at 20 × 20 grid points are generated. Additional analyses are conducted on subsets of these data for a smaller grid (10 × 10 points) and for shorter series (30, 40, 50, 70, 100, and 150 time units).

4. Implementation of tests

For each random time series, the test statistics described in the following, pertaining to the individual statistical tests, are calculated for all values of the statistical properties (spatial and temporal autocorrelation and trend magnitude), listed in Table 1. The counting, Walker, and FDR tests are conducted with all four local tests, namely the \(t\) test, M-K test, pre-whitened M-K test, and M-K test with equivalent sample size at the 5% level. The possible ranges for all tests are listed in Table 2, which also indicates the tail where the region of rejection is located.

For the counting test, the number of grid points where the local significance of a trend is detected, \(n_{\text{sig}}\), is counted. Then the test statistic \(C\) is simply the percentage of grid points with locally significant tests:

\[ C = \frac{n_{\text{sig}}}{N} \times 100. \]

The statistic of the Walker test \(W\) is the minimum \(p\) value of all \(N\) local \(t\) tests:

\[ W = p_{(1)} = \min_{i=1,...,N} p_i. \]

The statistic of the FDR test follows from its description in Eq. (8): we seek for the \(p\) value for which the inequality in Eq. (8) is fulfilled most strongly; that is, we seek for such \(i\) that the \(i\)th smallest \(p\) value divided by \(i\) is minimum. The minimum value of this ratio is then multiplied by the number of grid points for convenience; the statistic of the FDR test, \(f\), is therefore

\[ f = \min_{i=1,...,N} \left[ \frac{p_{(i)}}{i} \right] N. \]

For the extended M-K test, we use test statistic \(T\) calculated from Eq. (9) standardized to unit variance by dividing it by its standard deviation from Eq. (10), that is,
Criteria typically yield widely different numbers (e.g., Serrano from noise, that is, how many components to retain; these measure. Numerous criteria exist on how to separate signal (DelSole and Yang 2011). Principal component analysis is one would obtain if the original criterion were used. et al. 1999). Here we choose the simple rule to keep that many nomenclature; Richman 1986) using correlation as similarity columns and time realizations in rows (S-mode in the standard conducted on the data matrix consisting of gridpoint values in 5 components to retain differs from that used originally by DelSole and Yang (2011); hence results might differ from what one would expect if all trends are assumed to be either positive or negative, that is, no trends have exactly zero magnitude (this assumption is fully realistic; no trend with exactly zero magnitude in fact appeared in our very large sample). For the ease of display and for consistency with the counting test, we express this statistic as a percentage:

\[ S = \left( \frac{2^{n^2}}{N} - 1 \right) \times 100. \]  

Since we focus on trends with one sign over the analyzed domain, we opt for one-sided hypotheses for all tests. That is, the null hypothesis of no trend is tested against the alternative of a positive trend.

5. Null distributions and critical values

Figure 1 displays empirical distributions (histograms) of test statistics for the six tests for an illustrative selection of temporal and spatial autocorrelation values. All tests appear to be sensitive to the presence of autocorrelations, although to a different extent and in a different way. The distributions of the sign test are symmetric around zero because the probability of positive and negative trends is the same under the null hypothesis. The sign test, thanks to its construction, is insensitive to temporal autocorrelation as the magnitude of autocorrelation does not affect the number of positive and negative trends. Growing spatial autocorrelations make the distribution of the sign test widen because the probability of a simultaneous detection of trends of one sign at a large number of sites grows with increasing spatial autocorrelation. For high spatial autocorrelations, the distribution becomes U-shaped, with most values concentrated at the lower and upper bound (−100 and +100). That is, there is a strong tendency for local trends to be either all negative or all positive under high spatial autocorrelations, even though no real trend is present.

Histograms for the extended M-K test confirm that its modification succeeds in eliminating the sensitivity to spatial autocorrelation: the histogram of the test statistic does not change with its increasing values. Nonzero temporal autocorrelations result in a larger variance of the empirical distribution, without considerable changes in its shape.

The sensitivity of the regression test to spatial autocorrelations is relatively weak: The distribution of its statistic moves its peak toward zero and extends its variance with increasing spatial autocorrelations. Nonzero temporal autocorrelation results in the null distribution moving to the right, while keeping the ratio of its standard deviation to the mean approximately constant.

For the remaining three tests, we first discuss their behavior for the \( t \) test used as a local test. Nonzero temporal autocorrelations result in the statistic of the counting test being shifted to higher values; that is, the number of trends detected by local tests increases with increasing temporal autocorrelation. High spatial autocorrelations make the distribution of the test statistic widen and shift to zero. For very high spatial autocorrelations, there is a large probability that no local test will detect a trend if the null hypothesis is true (for spatial autocorrelation of 0.97, this probability is as large as 70%), but at the same time there is a nonzero, although small, probability that trend is detected at any number of sites, including the possibility that it is detected everywhere.

Unlike the other four tests, the Walker and FDR tests have their region of rejection in the left tail. The null distribution of the Walker test tends to stick to zero: the probabilities are largest for near-zero values and decrease toward the right tail. In other words, the smallest \( p \) value of all \( N \) local tests is usually very small even if the null hypothesis is true. Relatively large \( p \) values are nevertheless also plausible, despite occurring with much smaller probability. The autocorrelation values control the degree to which the Walker test statistic sticks to zero. For large spatial autocorrelations, the distribution becomes wider; for autocorrelation of 0.97, all values of the test statistic up to 1.0 are possible (i.e., the minimum \( p \) value may take on any value), although the probability of near-zero values remains the largest. The opposite effect appears for large temporal autocorrelations: the probability concentrates in the vicinity of zero.

A similar behavior can be observed for the FDR test. Its distribution is approximately uniform in the absence of autocorrelations. The probability concentrates at the upper bound.
Distributions of test statistics for counting, Walker, and FDR tests are insensitive to whether ‘t’ tests or Mann–Kendall tests are used as local tests (not shown): the distributions are almost identical. The two approaches of accounting for temporal autocorrelations result in different modifications of distributions of test statistics.

The use of prewhitening makes a rather minor difference. For nonzero temporal autocorrelations, the distribution for the counting test shifts to somewhat higher values, while the distributions for the Walker and FDR tests stick more strongly to zero (cf. middle and left columns in Fig. 2). This makes the distributions for the prewhitened series even more different from the distributions for the series with zero temporal autocorrelation than the distribution of the original series are. The likely reason is illustrated in Fig. 3. The histogram of autocorrelations in time series generated with temporal autocorrelation equal to 0.3 is rather wide (middle panel in Fig. 3); the actual autocorrelations range from 0.04 to 0.50 because of the random component in data. The histogram for the series from which the temporal autocorrelation was removed by prewhitening (right panel of Fig. 3) is much narrower than the histogram for series generated under the assumption of zero temporal autocorrelation (left panel). This effect occurs because the whitening procedure mistakes a part of stochastic variability for autocorrelation and removes it, although it should have been retained.

The use of effective sample size in local tests alters the distributions of test statistics substantially, making them close to distributions for series with zero autocorrelation. For the counting test, the distribution is shifted left close to the origin, although it still has somewhat higher mean and variance than that for zero autocorrelation. For the Walker test, the distribution after accounting for temporal autocorrelation has somewhat higher concentration of probability near the origin, while an extra peak appears near the upper bound for the FDR test.

Whether and how temporal autocorrelation is accounted for has no effect on how the distributions depend on spatial autocorrelations: for nonzero spatial autocorrelations, the distributions for prewhitened local tests and tests with effective

---

**Fig. 1.** Null distributions for test statistics, 20 × 20 grid; from left to right: sign test, counting test, M-K test, Walker test, FDR test, and regression test. A local ‘t’ test is used for the counting, Walker, and FDR tests. The third row (highlighted by a rectangle) is for zero temporal and spatial autocorrelations; the upper two rows illustrate the effect of nonzero temporal autocorrelations (0.3 and 0.6); the lower three rows illustrate the effect of nonzero spatial autocorrelations (0.6, 0.9, and 0.97). Note that the range of horizontal and vertical axes varies among plots. Values of the Walker test statistic are multiplied by 1000.
FIG. 2. Effect of prewhitening on the test statistics. Null distributions are shown for the (top) counting, (middle) Walker, and (bottom) FDR tests for the (left) local M-K test, (center) local prewhitened Mann–Kendall test, and (right) local Mann–Kendall test with effective sample size, all for temporal autocorrelations of 0.0, 0.3, and 0.6 (for the first, second, and third rows in each section, respectively).
The empirical distributions constitute null distributions, from which critical values of the tests are to be deduced. The critical value for the significance level of 5% is the 95th percentile of the empirical distributions for the sign, counting, extended M-K, and regression tests, and the 5th percentile for Walker and FDR tests. A strong concentration of probability at the bound near which the critical value lies constitutes a potential problem because the test may become inapplicable. This concerns Walker and FDR tests under high temporal autocorrelations, the sign test under high spatial autocorrelations, and the counting test when both temporal and spatial autocorrelations are high. The use of local tests with effective sample size alleviates the limitations of the counting. Walker, and FDR tests under high temporal autocorrelations: these tests become applicable under high temporal autocorrelations if an effective sample size is used. If the largest (when the critical value is in the upper tail) or smallest (when the critical value is in the lower tail) possible value of the test statistic occurs in more than 5% of cases, the critical value is equal to the maximum (or minimum for the Walker and FDR tests) possible value, that is, when the test is unusable as explained above, are denoted by italics in Table 3. It is worth noting that Fig. 1 and Table 3 suggest that under high spatial autocorrelations (0.97 and more), even the presence of a positive trend at all 400 grid points does not ensure statistical significance of the sign test. Whether the critical values depend on the length of time series is examined by calculating Spearman correlations of the critical values with the series length; they are displayed in Table 4 for selected values of spatial and temporal autocorrelations. (Recall that we examine seven different series lengths—namely, 30, 40, 50, 70, 100, 150, and 200 time units—from which the correlations are calculated.) One would expect that, in general terms, the shorter the time series the wider the null distribution because of a larger uncertainty. First, this is not true for all tests; for example, the variance of the null distribution for the sign test under zero spatial autocorrelations, and hence the critical value, is independent of the series length since the probability of detecting a positive or negative trend at a single site when no trend is imposed is equal to 0.5 regardless of the length of the series. Second, the way that the width of the null distribution projects onto the critical value depends on the shape of the distribution and is different for unbounded distributions (such as those for extended M-K test; see Fig. 1), distributions bounded from the side where the critical value lies (e.g., Walker test), distributions bounded from the other side (e.g., the FDR test for positive spatial autocorrelations), and U-shaped distributions (such as sign test for high spatial autocorrelations). This means that there is no
simple general rule describing what the dependence of the critical value on the series length is; a similar argument applies to the dependence on the grid size, discussed below. Spearman correlations are insignificant for the sign test (this holds for other values of autocorrelations not shown in Table 4 as well), indicating that the variability of critical values relative to the series length is a demonstration of sampling variability. The dependence of critical values on the series length is more systematic for the other tests; it is particularly strong for the regression test and for tests with a local test with prewhitening.

<table>
<thead>
<tr>
<th>Temporal autocorrelation</th>
<th>Spatial autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sign test</td>
</tr>
<tr>
<td></td>
<td>Counting test with local t test</td>
</tr>
<tr>
<td></td>
<td>Counting test with local M-K test</td>
</tr>
<tr>
<td></td>
<td>Counting test with local prewhitened M-K test</td>
</tr>
<tr>
<td></td>
<td>Extended Mann–Kendall test</td>
</tr>
<tr>
<td></td>
<td>Walker test (×1000) with local t test</td>
</tr>
<tr>
<td></td>
<td>Walker test (×1000) with local M-K test</td>
</tr>
<tr>
<td></td>
<td>Walker test (×1000) with local prewhitened M-K test</td>
</tr>
<tr>
<td></td>
<td>FDR test with local t test</td>
</tr>
<tr>
<td></td>
<td>FDR test with local M-K test</td>
</tr>
<tr>
<td></td>
<td>FDR test with local prewhitened M-K test</td>
</tr>
<tr>
<td></td>
<td>Regression test</td>
</tr>
</tbody>
</table>

Table 3. Critical values for individual tests, for time series that are 200 units long, and for a 20 × 20 grid for various values of temporal (in rows) and spatial (in columns) autocorrelation. Cases in which the critical value is equal to the largest (or smallest for Walker and FDR tests) possible value, because of which the test becomes unusable, are denoted in italics.
The effect of the grid size on the critical values can be deduced from Table 5, comparing critical values for the 20 × 20 and 10 × 10 grids. Extended Mann–Kendall and FDR tests appear to be insensitive to the grid size and the regression test is little sensitive to it, especially for small autocorrelations. Critical values of all the other tests seem to be systematically affected by the grid size.

If the test is not sensitive to the series length and/or the grid size it may be sensible to determine the critical values from the longest possible time series and/or for the largest grid because they are determined with the smallest uncertainty. Nevertheless, we prefer to determine the critical values for every series length and grid size separately for all tests in order to handle all the tests in the same way.

6. Evaluation of tests: Type-II errors

The power of a test can be evaluated against a specific alternative hypothesis as a type-II error, that is, the probability that the null hypothesis is not rejected if the alternative hypothesis is true. Various magnitudes of trend (see Table 1) are used as an alternative hypothesis. Type-II errors for the 20 × 20 grid and the time series 200 units long are displayed in Table 6 for a selection of tests; a full table for all tests, including all combinations with local tests, is provided as Table S1 in the online supplemental material. In the absence of autocorrelations, the sign, counting, and extended M-K tests are able to detect even the weakest trend considered in our study, 0.005 per decade, with certainty (type-II error being 0.2% and less). Dimensionless trends of 0.075 per decade and larger are detected with type-II errors below 5% by all tests even for the strongest autocorrelations, with the exception of the regression test and the largest autocorrelations considered.

We set the type-II error of 5% as a limit of detectability of a trend, analogously to the significance level (type-I error), which is also set to 5%. The trend magnitude corresponding to this limit can be determined by interpolation between values in Table 6. For example, the value of trend corresponding to the 5% detectability for the sign test for zero temporal autocorrelation and spatial autocorrelation of 0.97 is between 0.03 (for which type-II error is 8.9%) and 0.04 (type-II error is 1.3%). Using linear interpolation, the detectable trend, that is, the trend corresponding to type-II error of 5%, is approximated to be 0.0351 for this case. The values of trends detectable at the 5% level, estimated by linear interpolation, are displayed in Fig. 4 for a range of lengths of the series, two grid sizes, and four combinations of temporal and spatial autocorrelation.

The ability of all tests to detect a trend decreases as the time series shortens. For the high autocorrelations (0.6/0.97), even an extremely steep trend of one standard deviation per decade cannot be detected for the 30-yr series by any test. It is also clearly visible that trends are easier to detect on the larger grid.

For a given trend magnitude, the ability to detect a trend worsens (i.e., type-II error grows) with spatial and temporal autocorrelation growing. The sign, counting, and extended M-K tests perform better than Walker and FDR tests for all autocorrelations considered, the regression test tending to perform better than Walker and FDR tests. Extended M-K test is comparable to the sign test in its performance for most configurations of autocorrelations, grid size, and series length; it tends to be slightly more powerful for higher spatial autocorrelations. The extended M-K and sign tests perform consistently better than the counting test, although by a relatively narrow margin. The performance of the FDR test is consistently but only slightly better than that of Walker test, which is an expected finding given that the former is a generalization and improvement of the latter. The differences between tests tend to decrease with increasing autocorrelations. Note that for temporal autocorrelation of 0.6 and spatial autocorrelation of 0.97, performance of Walker and FDR tests is shown for the local tests with effective sample size because it is the only variant available (all other variants of local tests are unusable because of the construction of test as discussed above).

The effect of the choice of the local test on the detectability of trends is small. It is illustrated for the counting test for zero spatial autocorrelation. The effect is similar but smaller for Walker and FDR tests and for higher spatial autocorrelations.
<table>
<thead>
<tr>
<th>Relative trend magnitude</th>
<th>Sign test</th>
<th>Counting test with local $t$ test</th>
<th>Extended Mann–Kendall test</th>
<th>Regression test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.010</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.015</td>
<td>0.6</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.020</td>
<td>0.0</td>
<td>0.97</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.030</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.040</td>
<td>0.3</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.050</td>
<td>0.6</td>
<td>0.97</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.075</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.100</td>
<td>0.6</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.125</td>
<td>0.6</td>
<td>0.97</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 6. Type-II errors (%) for various values of spatial and temporal autocorrelation and various trend magnitudes as specific alternative hypotheses for selected tests. Crosses denote cases in which the test is unusable; see the text for more explanation.
The performance of the local $t$ test appears to be slightly better than the Mann–Kendall test. Prewhtening results in a marginal improvement of performance, while the use of effective sample size leads to a slight deterioration. This is illustrated for the counting test in Figs. 4a and 4b, but applies to Walker and FDR tests as well. For higher autocorrelations, the trend detectability is very similar for all the local tests.

In summary, Table 6 and Fig. 4 suggest that extended M-K and sign tests perform almost equally well, whereas the
performance of the other tests is consistently worse. However, the applicability of the sign test is limited because it is unusable under very high spatial autocorrelations.

7. Application to real data

It is important to note that the tests were compared under assumptions that may not be met in reality. We take annual mean temperatures from HadCRUT4 surface temperature data (ensemble median from version HadCRUT.4.5.0.0; Morice et al. 2012) for the period 1961–2010 on a 20 × 20 grid with a 5° spacing, covering southeastern Asia, most of Australia, and adjacent parts of Indian and Pacific Oceans, as an example. The grid extends from 32.5°S to 62.5°N and from 82.5° to 177.5°E. The reason for choosing this domain is that it is the only domain on Earth comprising 20 × 20 grid points where data in this dataset are almost complete for the given 50-yr period. As Fig. 5 indicates, the range of values of autocorrelations is considerable: temporal autocorrelations range from slightly negative values to almost 0.8, while spatial autocorrelations have much more skewed distribution, with values from below 0.1 to over 0.98. The assumption of spatially homogeneous autocorrelations is clearly not met.

The assumption of AR(1) process in both spatial and temporal domain is not realistic either: temperature is known to exhibit long-range dependence (e.g., Lennartz and Bunde 2009; Mann 2011; Gil-Alana et al. 2019) whereas the assumption of isotropy of spatial autocorrelations is broken by the existence of teleconnections, that is, highly correlated but mutually distant regions, of which Southern Oscillation is the most relevant one in southeastern Asia and Australia. The spatial autocorrelations are also likely to differ between the north–south and west–east directions, in particular if a regular latitude–longitude grid is used, in which the distance of grid points in the west–east direction decreases toward the pole.

The study of whether the rating of methods is still valid under the long-range persistence and presence of teleconnections, and in general if other assumptions are not met, may be the topic of further work.

The idealized assumptions of the constancy of autocorrelations and variance and of the lack of long-range dependence in both time and space are necessary for the simulation study and comparison of tests to be feasible and tractable: analogous assumptions form the grounds of similar studies (e.g., Yue et al. 2002; Wilks 2006a, 2016). In practical applications it is advisable that the critical values be determined from empirical null distributions obtained by resampling the real data rather than taken from synthetic data created under idealized assumptions. The resampling must respect the properties of data relevant for the estimation of trends and their significance (e.g., McKinnon and Deser 2018), that is, temporal and spatial autocorrelations must be preserved in resampled data in particular.

Despite the simplifying assumptions, results of this study provide reliable guidelines for the selection of an appropriate test of collective significance of trends.

8. Conclusions

This paper focuses on collective (global, field) significance of tests of local trends. Our objective was to introduce a novel simple procedure of significance testing for the presence of trend on a grid, which is based on counting signs of local trends (sign test), and compare it with other approaches to testing collective significance of trends (counting, extended Mann–Kendall, Walker, FDR, and regression tests).

First, the null distributions of trend statistics are constructed using synthetic data. The critical values of tests are then determined as corresponding quantiles of the null distributions. The performance of tests is quantified in terms of trend magnitudes, for which type-II error (i.e., the probability of failing to
reject the null hypothesis when the alternative hypothesis is true) attains 5%: the lower the trend magnitude, which is detectable with the given type-II error, the better performance of the test. For lower values of spatial and temporal autocorrelations, the sign test and extended M-K test outperform the counting test by a narrow margin; these three tests are considerably more efficient than Walker, FDR, and regression tests for the entire range of autocorrelations. The spread of performance of individual tests becomes narrower for high autocorrelations. Some tests become unusable under high spatial or temporal autocorrelations because their critical values are identical to the highest or lowest possible value of the null distribution. Taken together, the extended (multisite) version of M-K test appears to be the best choice. The sign test proposed by us, which has an advantage of computational simplicity, is an almost equally appropriate choice whenever it is usable, that is, for low to moderate autocorrelations when its critical values remain below the maximum possible value of the test statistic. Another advantage of the sign test is the independence of the critical values on the length of the series.

An example of annual mean temperatures in HadCRUT4 data is used to demonstrate that real data are unlikely to comply with the assumptions on which the synthetic data are constructed and critical values determined. Therefore, real-world applications of the testing procedures will require data resampling (Monte Carlo) procedures to create null distributions and to determine critical values. Our results highlight the necessity of testing for collective significance in trend detection studies and should serve as a guidance for the selection of the appropriate collective significance test.

Acknowledgments. This study was supported by the Czech Science Foundation, project 16-04676S. Tomáš Krauskopf kindly assisted with processing the real-world data. The authors thank Prof. Richard Vogel of Tufts University for insightful discussions about Eq. (10) and the anonymous reviewers whose comments led to substantially improved clarity of the paper.

REFERENCES


