A comparison of the power of the *t* test, Mann-Kendall and bootstrap tests for trend detection

SHENG YUE
US EPA Mid-Continent Ecology Division, 6201 Congdon Blvd, Duluth, Minnesota 55804, USA
yue.sheng@epa.gov

PAUL PILON
Monitoring Services Division (MSD), Meteorological Services of Canada – Ontario Region, 867 Lakeshore Road, PO Box 5050, Burlington, Ontario L7R 4A6, Canada

Abstract Monte Carlo simulation is applied to compare the power of the statistical tests: the parametric *t* test, the non-parametric Mann-Kendall (MK), bootstrap-based slope (BS-slope), and bootstrap-based MK (BS-MK) tests to assess the significance of monotonic (linear and nonlinear) trends. Simulation results indicate that (a) the *t* test and the BS-slope test, which are slope-based tests, have the same power; (b) the MK and BS-based MK tests, which are rank-based tests, have the same power; (c) for normally-distributed data, the power of the slope-based tests is slightly higher than that of the rank-based tests; and (d) for non-normally distributed series such as time series with the Pearson type III (P3), Gumbel, extreme value type II (EV2), or Weibull distributions, the power of the rank-based tests is higher than that of the slope-based tests. The power of the tests is slightly sensitive to the shape of trend. Practical assessment of the significance of trends in the annual maximum daily flows of 30 Canadian pristine river basins demonstrates a similar tendency to that obtained in the simulation studies.

Key words trend detection; Student's *t* test; Mann-Kendall test; bootstrap test; power of a test; *P* value; trend shape; statistical analysis

Une comparaison de la puissance des tests *t* de Student, de Mann-Kendall et du bootstrap pour la détection de tendance

Résumé Des simulations de Monte Carlo ont été réalisées pour comparer la puissance des tests statistiques suivants pour estimer le niveau de signification de tendances monotones (linéaires et non-linéaires): le test paramétrique *t* de Student, le test non-paramétrique de Mann-Kendall (MK), le test de pente par bootstrap (BS-pente) et le test MK par bootstrap (BS-MK). Les résultats de simulation indiquent que (a) les tests *t* de Student et BS-pente, basés sur la pente, ont la même puissance; (b) les tests MK et BS-MK, basés sur le rang, ont la même puissance; (c) pour des données présentant une distribution normale, la puissance des tests basés sur la pente est légèrement supérieure à celle des tests basés sur le rang; et (d) pour des séries présentant une distribution non-normale, comme une distribution de Pearson III, de Gumbel, de valeur extrême type II, ou de Weibull, la puissance des tests basés sur le rang est supérieure à celle des tests basés sur la pente. La puissance des tests est légèrement sensible à la forme de la tendance. L'estimation pratique de la signification des tendances est similaire pour les études par simulation et pour l'analyse des données de maxima annuels de débits journaliers de 30 bassins vierges canadiens.

Mots clefs détection de tendance; test *t* de Student; test de Mann-Kendall; test bootstrap; puissance d'un test; valeur de *P*; forme de tendance; analyse statistique

INTRODUCTION

The rank-based nonparametric Mann-Kendall (MK) test (Mann, 1945; Kendall, 1975) has been commonly used to assess the significance of monotonic trends in hydro-

Open for discussion until 1 August 2004

meteorological time series (e.g. ven Belle & Hughes, 1984; Cailas *et al.*, 1986; Hipel *et al.*, 1988; Hipel & McLeod, 1994; Taylor & Loftis, 1989; Demarée & Nicolis, 1990; Zetterqvist, 1991; Chiew & McMahon, 1993; Yu *et al.*, 1993; Hirsch *et al.*, 1993; Lettenmaier *et al.*, 1994; Burn, 1994; Yulianti & Burn, 1998; Gan, 1998; Lins & Slack, 1999; Douglas *et al.*, 2000; Pilon & Yue, 2002; Yue *et al.*, 2003; and others). Another rank-based nonparametric test, the Spearman's rho (SR) test (Lehmann, 1975; Sneyers, 1990), has sometimes been applied to detect trends in hydrological data (e.g. Lettenmaier, 1976; El-Shaarawi *et al.*, 1983; Pilon *et al.*, 1985; McLeod *et al.*, 1991; Hipel & McLeod, 1994). The study of Yue *et al.* (2002) documented that these two tests have almost the same power to identify trends in time series. In comparison to the parametric *t* test, the common use of the nonparametric tests is due mainly to the consideration that they are more suitable for the situations of non-normal data, censored data, and missing data problems, which frequently occur in hydro-meteorological studies.

Recently, the attention given to the bootstrap technique has been attributed to the advances made in PC computational capability (see Efron & Tibshirani, 1993; Hjorth, 1994; Davison & Hinkley, 1997). The bootstrap is a computationally intensive approach for assigning measures of accuracy to statistical estimates. The accuracy of statistical inference by the approach depends on the number of bootstrapped samples from original data. That is, as the number of "bootstrapped" samples increases, the accuracy of the statistical inference improves. Its merit is that it is free of the restrictive assumption regarding normality of sample data, and that the method is easy to understand and implement (Simon & Bruce, 1991). The bootstrap techniques have been applied to resolve various problems in the water resources field as demonstrated by Zucchini & Adamson (1989), Vogel & Shallcross (1996), Lall & Sharma (1996), Tasker & Dunne (1997), Stefano *et al.* (2000) and Yue & Wang (2002). The rank-based bootstrap MK test has also been used to detect trends in hydrological time series (e.g. Douglas *et al.*, 2000; Burn & Hag Elnur, 2002; Yue *et al.*, 2003). In these trend-detection studies, sample data are re-sampled by randomly selecting samples from the original data, and then the MK test statistic of the re-sampled data is computed. By re-sampling the original data *N* times and computing the *N* MK statistics, the bootstrap empirical distribution function of the MK statistic can be obtained. This test is referred to as the bootstrap-based MK (BS-MK) test, to distinguish it from the original MK test.

Both the MK and BS-MK tests are used to assess the significance of trend via the MK statistic rather than to directly judge the significance of trend by its magnitude. The assessment of the significance of trend and the computation of the magnitude (slope) of trend are carried out separately. The magnitude of trend may be computed by ordinary least squares (Hirsch *et al.*, 1993) or the nonparametric approach (Sen, 1968). The classical Student's *t* test evaluates the significance of trend via its magnitude, i.e. the t-test statistic is the ratio of the estimate of the magnitude of trend or its slope to its standard deviation.

Given that it is possible to compute the slope of each bootstrapped sample, these values can, in turn, be used to establish the empirical distribution of trend. This can be applied to assess the significance of a specific trend from a target. This study will also propose this approach for trend detection, which is termed the bootstrap-based slope (BS-slope) test. Both the BS-slope test and the BS-MK test are presented in the following sections.

When one wants to perform trend detection, it is natural to ask which of these four tests should be applied to detect a trend in a time series. In other words, which test has the highest power to detect a certain amount of trend? Lettenmaier (1976) compared the power of the *t* test and the Spearman's rho (SR) test for detecting a linear trend in normally distributed series and indicated that the *t* test has slightly higher power than the SR test. Hipel & McLeod (1994) investigated the powers of the MK test and the lag-one serial correlation test for detecting trends in normally-distributed data, and demonstrated that the MK test is more powerful than a lag-one serial correlation test for identifying deterministic trends. Yue *et al*. (2002) documented that the MK and SR tests have the same power and that their power is sensitive to the probability distribution type as well as the statistical properties of sample data.

The objective of this paper is to compare the power of the *t* test, MK test, BS-slope test and BS-MK test for detecting both linear and nonlinear monotonic trends in normal and non-normal series by Monte Carlo simulation. The four tests are also applied to assess the significance of trends in annual maximum flows of 30 near-pristine river basins in Canada.

METHODOLOGY

For a description of the statistics of the parametric *t* test and the conventional MK test, readers may refer to Hirsch *et al*. (1993), or generally available texts on statistics. Only the bootstrap-related tests are introduced here.

Bootstrap-based slope (BS-slope) test

Suppose that an observed sample data set, X $(= x_1, x_2, ..., x_n)$ is available, from which the magnitude of trend, b_o, of interest can be computed using the approach by Theil (1950) and Sen (1968), hereafter referred to as the Theil-Sen Approach (TSA).

$$b_o = \text{Median}\left(\frac{x_j - x_l}{j - l}\right) \ \forall l < j \tag{1}$$

where x_l is the lth value of the sample data X.

The significance of b_o is assessed based on the null distribution of slope, which can be derived by randomly bootstrapping the sample data X. A bootstrapped sample, denoted by $X^*(= x_1^*, x_2^*, \ldots, x_n^*)$, is obtained by randomly sampling n times with replacement and with an equal probability $1/n$ from the observed sample $x_1, x_2, ..., x_n$. By bootstrapping X M times, M independent bootstrap samples $X^{*1}, X^{*2}, X^{*3}, ..., X^{*M}$, each with sample size n can be obtained. The slope ($\hat{b}^*$) for each of the bootstrapped samples is then estimated using equation (1). This results in M estimates of the slope $\hat{b}^*$: $\hat{b}^{*1}$, $\hat{b}^{*2}$, $\hat{b}^{*3}$, ..., $\hat{b}^{*M}$. By arranging them in ascending order, the bootstrap empirical cumulative distribution (BECD$\sim\hat{b}^*$) of the slope will be obtained, as illustrated in Fig. 1. The P value (p_b) of the slope, b_o, of the observed sample data can be estimated using the BECD$\sim\hat{b}^*$ curve:

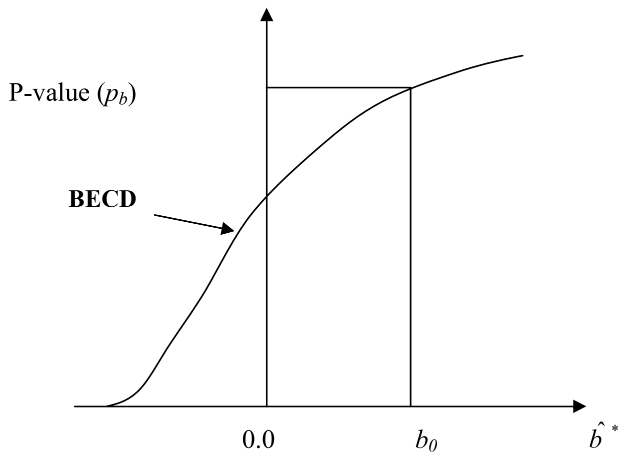

Fig. 1 Schematic illustration for computing the BECD~$\hat{b}^*$.

$$p_b = \Pr[\hat{b}^* \le b_o] = \frac{m_b}{M} \tag{2}$$

where m_b is the rank corresponding to the largest value $\hat{b}^* \le b_o$. For sample data having no trend, the P value should be close to 0.5. A plus or minus b_o value indicates an upward or downward trend, respectively. At the significance level (α) of 0.05 for a one-tailed test, a negative trend is significant when its P value (p_b) $\le$ 0.05, and a positive trend is significant when $p_b \ge 0.95$.

Bootstrap-based MK (BS-MK) test

This test is similar in design to that of the BS-slope test. Rather than being based on the slope, the MK statistic (S_o) of the sample data, X, is computed and used. The significance of S_o can be assessed based on the null distribution of the bootstrap MK statistic, BECD~$\hat{S}^*$, which is derived from the bootstrapped sample data. The P value (p_s) of the S_o of observed sample data is estimated using the BECD~$\hat{S}^*$ curve as:

$$p_S = \Pr[\hat{S}^* \le S_o] = \frac{m_S}{M} \tag{3}$$

where m_S is the rank corresponding to the largest value $\hat{S}^* \le S_o$.

Similar to the BS-slope test, for the sample data without any trend, the P value should be close to 0.5. A plus or minus S_o corresponds to an upward or a downward trend, respectively. At α = 0.05 for a one-tailed test, for a significant negative trend, $p_s \le 0.05$; for a significant positive trend, $p_s \ge 0.95$.

Confidence interval of the bootstrap tests

The percentile method is adopted to construct the bootstrap confidence interval (Efron & Tibshirani, 1993). For a two-tailed test, the percentile method is just the interval between the 100·α/2 and 100·(1 – α/2) percentiles of the bootstrap distribution of C^* ($C^*= S^*$ or b^*); α is pre-assigned significance level. The 100·α/2 percentile of the bootstrap distribution of C^* is estimated by first arranging the C^* in ascending order. Then the percentile is estimated by interpolating between the ($\alpha \cdot M/2$) and the ($\alpha \cdot M/2 + 1$) members of the ordered C^*. If the number of the bootstrap samples, M, is large enough, an accurate confidence interval can be obtained by the percentile method. For 90–95% confidence intervals, Efron & Tibshirani (1993) and Davison & Hinkley (1997) suggest that M should be between 1000 and 2000.

Power computation

The significance level or type I error, α, is the probability of rejecting the null hypothesis when it is true. A type II error (β) is the probability of accepting a null hypothesis when it is false. The power of a test is the probability of correctly rejecting the null hypothesis when it is false, which is equal to 1 – β. When sampling from a population that represents the case where the null hypothesis is false, i.e. the alternative hypothesis is correct, the power can be estimated by (Yue *et al*., 2002):

$$\text{Power} = \frac{N_{rej}}{N} \tag{4}$$

where N is the total number of simulation experiments and N_{rej} is the number of experiments that fall in the critical region, which is either ≤ α/2 or ≥ 1 – α/2.

COMPARISON OF THE POWER OF THE FOUR TESTS TO DETECT LINEAR TRENDS

A linear trend is a special type of monotonic trend having a constant change rate, and it has been widely used to approximate the magnitude of trends in time series analysis. First, the power of these tests for the case of linear trend is investigated. Monte Carlo simulation is used to generate time series of sample size n for a given distribution type having pre-selected characteristics (i.e. coefficient of variation, C_v, and skewness). The effect of sample properties such as sample size, sample variation and sample skewness on the power of statistical tests have been observed by Yue *et al*. (2002). Only positive trends will be inspected here, as for negative trends the power of the tests is identical. In order to assess the ability of the tests to correctly reject the null hypothesis, a linear trend having a specific slope is superimposed onto the generated time series.

Power of the tests for normally-distributed data

Simulation was performed to generate 3000 iid (independent, identically distributed) normal time series having a sample size $n = 50$ with mean $\mu = 1.0$ and coefficient of variation $C_v = 0.5$. Some selected linear trend scenarios ($T_t = b \cdot t$, b = 0.00 (0.004) 0.02, i.e. with b ranging from 0.00 to 0.02 with an increment of 0.004) are superimposed onto each of the generated series. For example, for a time series with $n = 50$, $\mu = 1.0$, and $b = 0.01$, its mean value would increase by 50% over a period of 50 years. For the *t* test and the MK test, their statistics were computed from the simulated samples and the confidence intervals at $\alpha = 0.05$ were established. The power of the tests was then computed using equation (4).

The power of the BS-slope test and BS-MK test was computed as follows. Each of the generated sample series, as described above, was resampled M (=3000) times, resulting in M bootstrap samples. For the BS-slope test, the P value (p_b) for each of the generated 3000 sample series, with a given b, was estimated using equation (2). The percentile interval of p_b at a significance level (α) of 0.05 was constructed using the percentile method on the basis of p_b when $b = 0$. The power of the test for a given $b \neq 0$ was then computed using equation (4). For the BS-MK test, similar to the BS-slope test, the P value (p_S) of S of each generated sample series with a given b was estimated using equation (3). The percentile interval of p_S at $\alpha = 0.05$ was constructed using the percentile method when $b = 0$. The power of the test for a given $b \neq 0$ was then computed using equation (4). Figure 2 shows the powers of these tests. Results indicate that for normally-distributed time series: (a) the slope-based tests, namely the *t* test and the BS-slope test, have almost the same power to detect trends; (b) the rank-based tests, namely the MK and the BS-MK tests, have almost the same power; (c) the power of the slope-based tests is slightly greater than that of the rank-based tests; and (d) when no trend is present, all of the tests have virtually the same power. The above simulation procedures were also replicated for sample sizes $n = 30$ and 80, and the results are the same as in the case of $n = 50$ (not shown here for the sake of brevity).

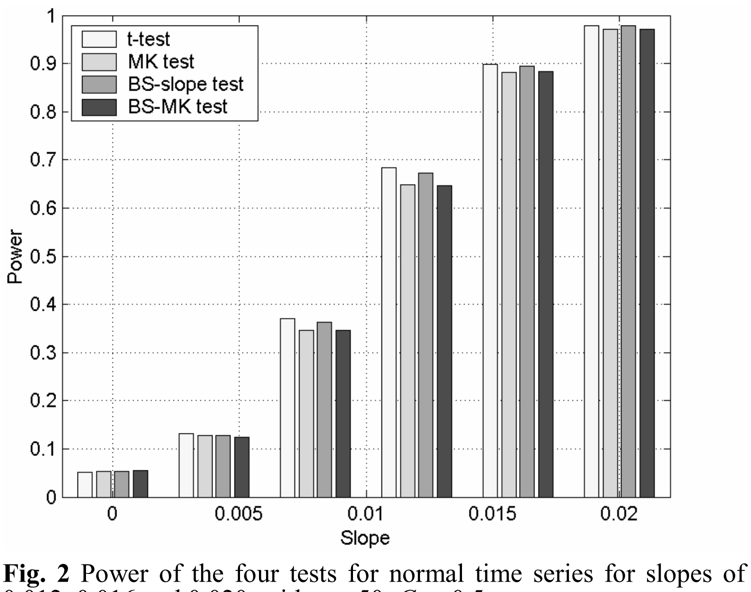

Fig. 2 Power of the four tests for normal time series for slopes of 0, 0.004, 0.008, 0.012, 0.016 and 0.020, with $n = 50$, $C_v = 0.5$.

Power of the tests for the non-normal data

In practice, most hydrometeorological time series may not follow the normal distribution. Distribution types that are frequently encountered in hydrometeorological time series are the Pearson type III (P3), extreme value (Gumbel, EV2 and Weibull) distributions.

Given mean (μ) = 1.0 and $C_v = 0.5$, random variates with Gumbel distributions can be generated using the formulae in Stedinger *et al.* (1993). For the EV2 distribution, $\kappa = -0.3$; for the Weibull distribution, ω (omega) = 0.6; for the P3 distribution, the coefficient of skewness, $\gamma = 1.5$. For each selected distribution type, 3000 samples are generated having sample size $n = 50$. A linear trend scenario, $T_t = b \cdot t$ with b = 0.0 (0.004) 0.02 (t = 0, 1, 2, …, $n - 1$), was then superimposed onto each of the generated series. Figures 3–6 depict the power of the tests for the P3, Gumbel, EV2 and Weibull distributions, respectively. These diagrams indicate that for non-normally distributed series, the two slope-based tests have almost identical power with each other, and this is also the case for the rank-based tests. However, the power of the rank-based tests is consistently higher than that of the slope-based tests when linear trend is present in time series.

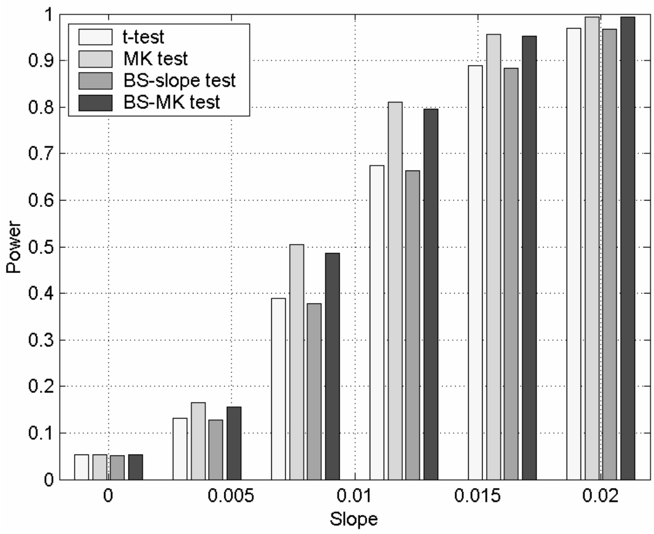

Fig. 3 Power of the four tests for P3-distributed series for slopes of 0, 0.004, 0.008, 0.012, 0.016 and 0.020, with $n = 50$, $C_v = 0.5$ and $\gamma = 1.5$.

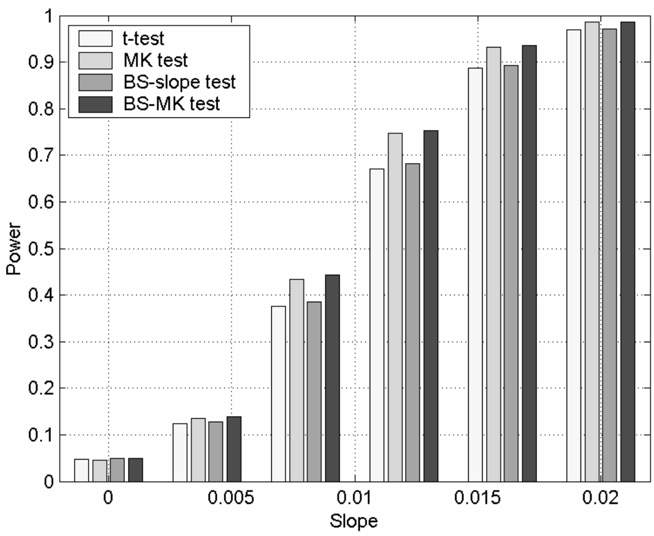

Fig. 4 Power of the four tests for Gumbel-distributed series for slopes of 0, 0.004, 0.008, 0.012, 0.016 and 0.020, with $n = 50$ and $C_v = 0.5$.

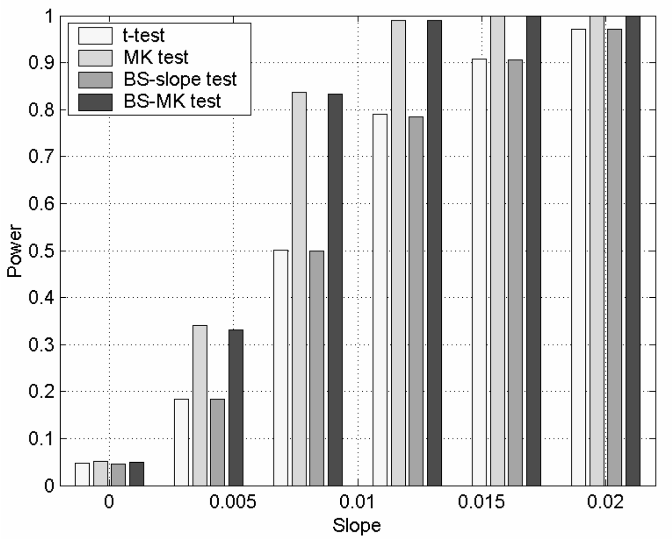

Fig. 5 Power of the four tests for EV2-distributed series for slopes of 0, 0.004, 0.008, 0.012, 0.016 and 0.020, with n = 50, C_v = 0.5 and κ = –0.3.

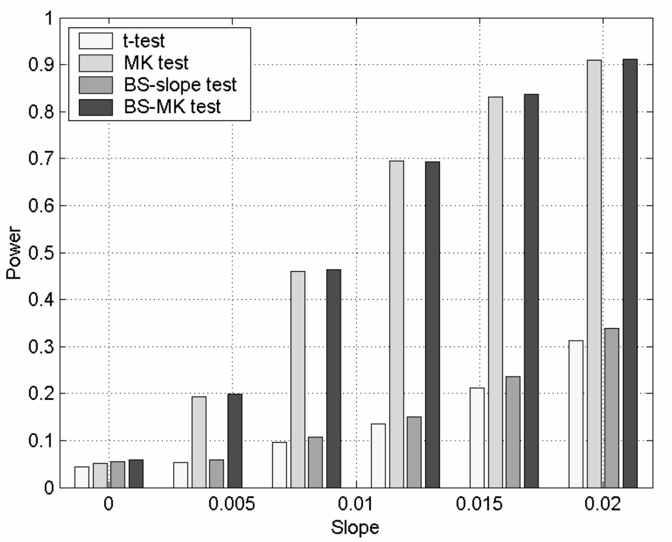

Fig. 6 Power of the four tests for Weibull-distributed series for slopes of 0, 0.004, 0.008, 0.012, 0.016, and 0.020 with n = 50, C_v = 0.5 and ω = 0.6.

COMPARISON OF THE POWER OF THE FOUR TESTS TO DETECT NONLINEAR MONOTONIC TRENDS

In reality, a trend in nature might not be linear. To the authors' knowledge, little attention has previously been paid to ascertaining the influence of the shape of trend on the power of a particular test. It would be useful to know the ability or power of these tests to reject the null hypothesis should a nonlinear monotonic trend exist in a time series. In this study, two types of typical nonlinear monotonic increasing trends (Ratkowsky, 1989) are selected to ascertain the power of the tests:

$$T_1 = B_1\, f_1(t) = \frac{B_1}{(1 + e^{a_1 - c_1 t})^{1/d_1}} \tag{5a}$$

$$T_2 = B_2\, f_2(t) = B_2 e^{a_2 t} \tag{5b}$$

where $f_1(t)$ represents a change rate or slope with time t, which increases at the beginning and then starts to decrease after a certain turning point, i.e. the increasing pace of trend accelerates at the beginning and then decelerates; $f_2(t)$ is a change rate or slope with time t, which increases over the entire period of observation, i.e. the increasing pace of trend accelerates throughout the period; and B_1 and B_2 represent the magnitude of change over the entire period. The two types of trends with given parameters: T_1 (a_1= 0.1, c_1 = 0.15, d_1 = 0.2 and B_1 = 0.2 (0.2) 1.0) and T_2 (a_2 = 0.025 and B_2 = 0.1 (0.1) 0.5) are illustrated in Fig. 7(a) and (b), respectively.

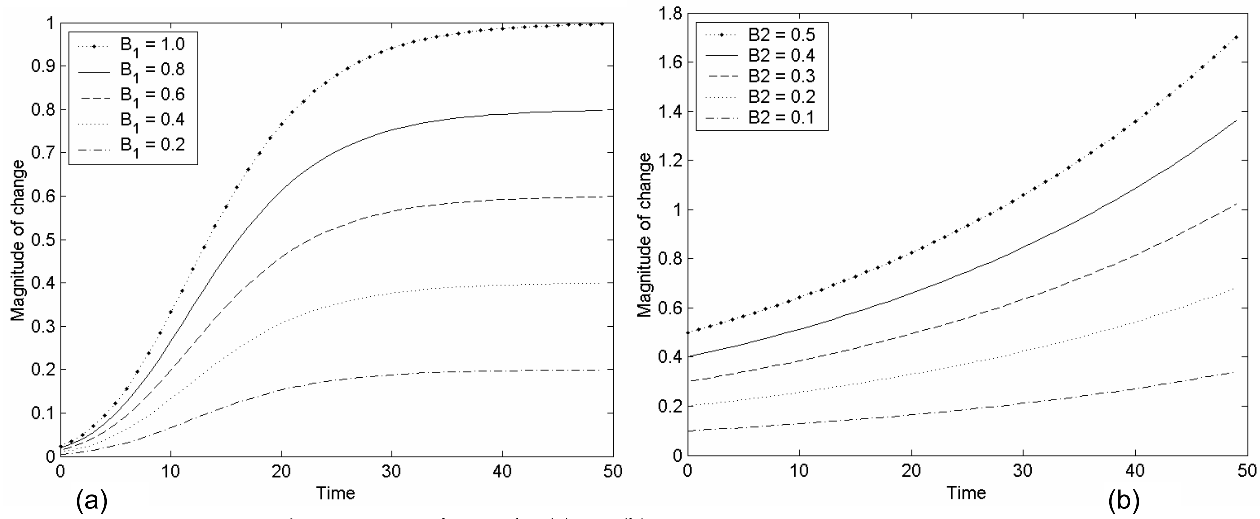

Fig. 7 Monotonic trends: (a) T_1; (b) T_2.

Normally-distributed series

Similar to the case of linear trend, 3000 iid normally-distributed time series were generated having a sample size n = 50, μ = 1.0 and C_v = 0.5. The monotonic trend $T_1 = B_1 f_1(t)$ with B_1 = 0.2 (0.2) 1.0 was superimposed onto each of the generated series. The power of the four tests was then computed and is shown in Fig. 8. The results depicted in Fig. 8 are similar to the previous case of linear trend for normally-distributed data, i.e. the power of the slope-based test is slightly higher than that of the rank-based tests. For the form of the monotonic trend $T_2 = B_2 f_2(t)$ with B_2 = 0.1 (0.1) 0.6, the power of the tests was computed and it was similar to that for the T_1 case. This result is somewhat in contrast to the commonly held view that the parametric t test is only suitable for assessing the significance of a linear trend. The results presented herein indicate that the power of the slope-based tests may be marginally affected by the shape of the monotonic trend (T_1 *vs* T_2) giving the same amount of increase in trend over time, in comparison to the linear trend that is a special case of monotonic trend. Based on the above simulation results, the overall power of a test appears to be more

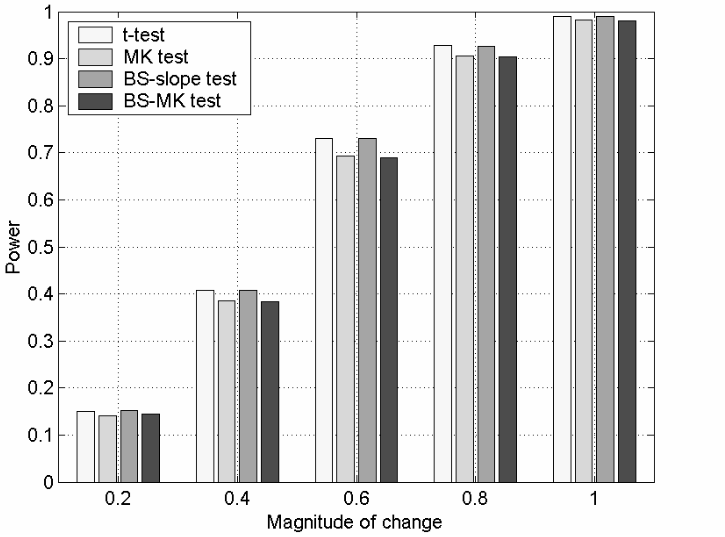

Fig. 8 The same as Fig. 2 but for monotonic trend T_1 with magnitude of changes of 0.2, 0.4, 0.6, 0.8 and 1.0 over time.

influenced by the magnitude of change that occurs over an observational period than by the shape of the monotonic trend.

Non-normally distributed series

Similar to the linear trend case, the monotonic trend $T_1 = B_1 f_1(t)$ or $T_2 = B_2 f_2(t)$ was superimposed onto the generated series. Subsequently, the power for the four tests for the P3, Gumbel, EV2 and Weibull series was computed, which indicates the same tendency as for linear trend (see Figs 3–6). For the sake of conciseness, only the power of the tests with the trend T_1 for the P3-distributed series is illustrated in Fig. 9. The same conclusion as for linear trend can be drawn, i.e. for non-normally distributed data, the power of the rank-based tests is greater than that of the slope-based tests.

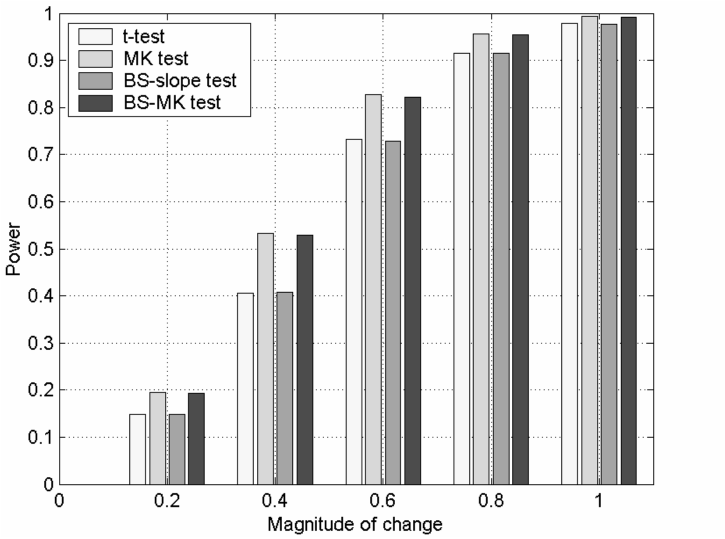

Fig. 9 The same as Fig. 3 but for monotonic trend T_1 with magnitude of changes of 0.2, 0.4, 0.6, 0.8 and 1.0 over time.

From the above simulation experiments, it was found that the slope-based tests, namely the *t* test and the BS-slope test, have the same power to detect the significance of a trend, irrespective of whether a trend is monotonically linear or nonlinear. Similarly, the rank-based tests, namely the MK and the BS-MK tests, have almost identical power. For normally-distributed series, no matter whether a trend is linear or nonlinear, the power of the slope-based tests for detecting the trend is slightly higher than that of the rank-based tests. Finally, for non-normally distributed data, such as the P3, Gumbel, EV2 and Weibull distributions, the rank-based tests have visibly higher power than that of the slope-based tests for detecting the significance of a trend, irrespective of whether a trend is linear or nonlinear. This implies that the existence of trend in non-normally distributed time series can be more effectively identified by the rank-based tests than by the slope-based tests.

The impacts of the shape of trend on the power of the tests

In the previous sections, the power of the tests was investigated for detecting the three types of trend, i.e. linear trend T and nonlinear trends T_1 and T_2. It is useful to know if the shape of the trend affects the power of the tests. To observe this issue, the same magnitude of change is given for the three types of trend, i.e. the mean (1.0) increases by 0.5 over 50 years, as shown in Fig. 10. The same parameters and procedures as used before are applied to generate time series with different distribution types. Only the *t* test and the MK test are inspected here as the BS-slope test and the BS-MK test would have provided similar results. The results for the t-test and the MK test are presented in Figs 11 and 12, respectively. These diagrams demonstrate that the ability to detect trend is somewhat sensitive to the shape of the trend with upward convex shape having the highest power and upward concave shape having the lowest power, except for the Weibull-distributed data. However, the impact of the shape of the trend has relatively little effect on the overall power of the tests for the case studied. This result, along with the observations obtained in the former section, further confirms the

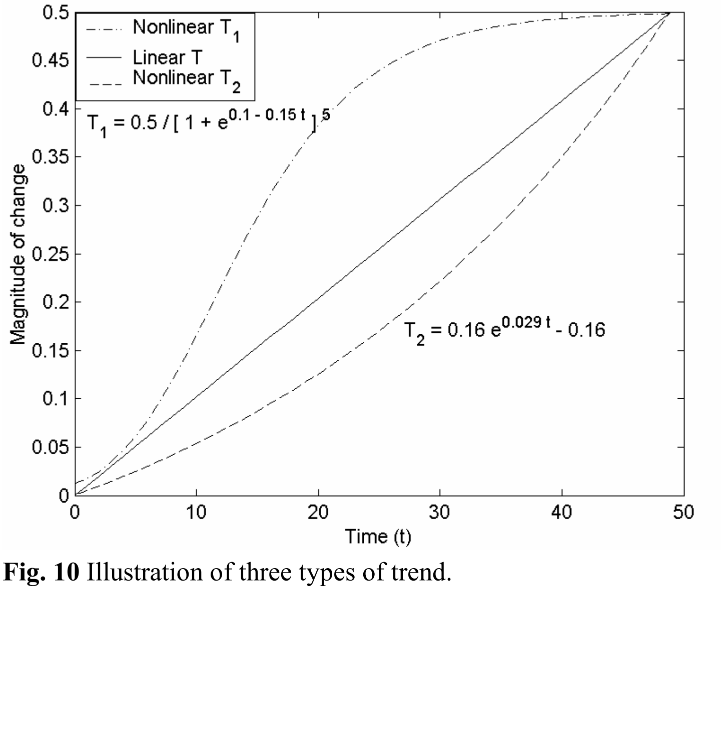

Fig. 10 Illustration of three types of trend.

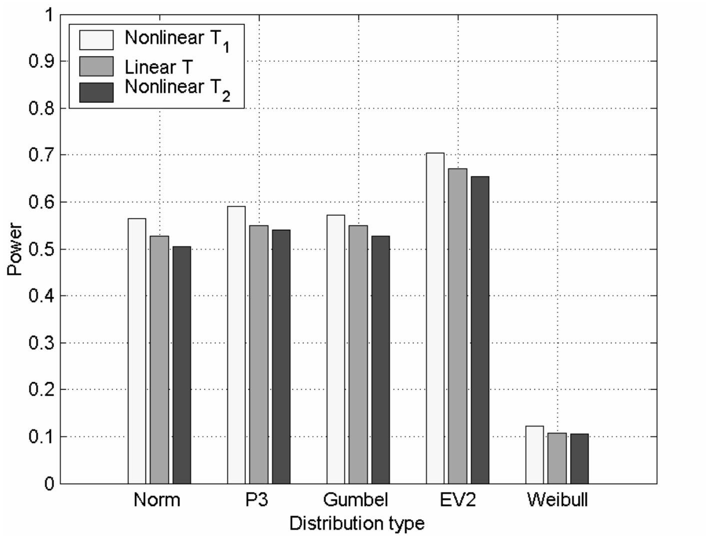

Fig. 11 Comparison of the power of the *t* test for series with different distributions.

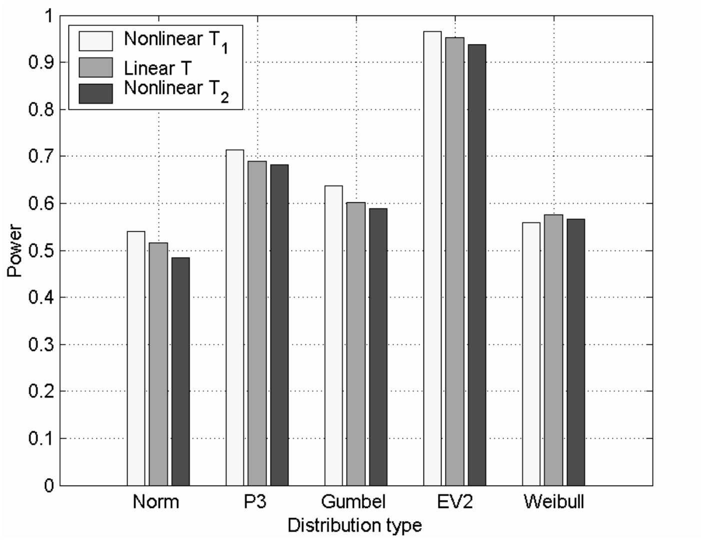

Fig. 12 Comparison of the power of the MK test for series with different distributions.

inference that the power of the tests is only slightly affected by the shape of trend. In addition, in comparison to the shape of trend, the power of a statistical test is much more sensitive to the probability distribution of the sample data. In addition, the MK or rank-based tests prove to be more powerful than the *t* test or slope-based test for non-normal data.

CASE STUDY

Annual maximum daily streamflow of 30 drainage basins representing pristine or stable land-use conditions were selected from the Canadian Reference Hydrometric Basin Network (RHBN) (Environment Canada, 1999). These sites were chosen as their data visually displayed evidence of trend and were useful for demonstrating the

Table 1 Comparison of the power of the *t* test, BS-slope test, MK-test and BS-MK test.

No.	Station ID	Record length	Q_m	C_v	C_s	C_k	Slope ($m^3 s^{-1} year^{-1}$)	*P* value: *t* test	BS-slope test	MK test	BS-MK test
1	08MH091	31	8.9	0.40	0.12	1.78	0.101	***0.081***	***0.075***	0.107	0.109
2	08HA026	19	1.5	0.36	–0.02	2.51	0.037	**0.049**	**0.048**	***0.076***	***0.075***
3	07JC001	22	14.1	0.44	0.13	2.33	0.370	**0.039**	**0.036**	**0.033**	**0.030**
4	06LA001	30	258.2	0.30	0.18	2.89	3.756	**0.010**	**0.010**	**0.002**	**0.002**
5	03QC002	19	480.9	0.31	–0.20	2.62	10.851	**0.039**	**0.034**	***0.062***	***0.055***
6	09AA006	47	226.9	0.18	0.26	2.47	0.736	**0.043**	**0.038**	***0.051***	**0.048**
7	02VC001	37	1566.2	0.30	–0.03	1.96	–15.864	**0.013**	**0.013**	**0.027**	**0.029**
8	03NF001	17	1046.8	0.23	0.17	2.06	–15.806	***0.095***	***0.086***	0.116	0.121
9	03NG001	17	1169.4	0.32	–0.25	1.83	–54.071	**0.000**	**0.001**	**0.002**	**0.001**
10	05DA009	28	264.3	0.15	0.22	3.67	–1.450	***0.057***	***0.051***	***0.080***	***0.083***
11	02JC008	27	170.8	0.25	0.26	2.33	–2.499	**0.008**	**0.009**	**0.012**	**0.011**
12	05AA008	48	33.7	0.51	1.07	4.26	–0.212	0.120	0.114	***0.052***	***0.052***
13	11AA026	63	13.5	1.07	2.01	7.85	–0.130	***0.099***	***0.095***	***0.068***	***0.071***
14	05TD001	35	119.1	0.37	0.48	3.92	–1.935	**0.004**	**0.005**	**0.004**	**0.003**
15	05HA003	34	6.8	0.68	0.52	2.23	–0.089	0.139	0.133	***0.084***	***0.090***
16	06LC001	29	1500.1	0.25	0.39	3.00	–8.586	0.159	0.152	***0.098***	***0.105***
17	10FA002	28	217.7	0.59	1.35	5.25	–2.642	0.193	0.177	***0.093***	***0.092***
18	02YR001	38	29.0	0.29	0.55	2.91	–0.147	0.122	0.120	***0.078***	***0.079***
19	04DA001	29	236.9	0.54	0.91	3.31	–5.533	**0.024**	**0.024**	**0.002**	**0.002**
20	08CD001	33	330.6	0.34	0.52	2.37	–3.448	**0.048**	**0.047**	**0.034**	**0.033**
21	08CE001	42	2357.6	0.24	0.40	2.33	–13.699	**0.030**	**0.029**	**0.026**	**0.026**
22	08NH016	18	4.4	0.31	0.61	3.28	–0.102	**0.049**	**0.045**	**0.038**	**0.039**
23	10AB001	34	697.9	0.27	0.69	3.26	–3.786	0.128	0.125	***0.084***	***0.091***
24	02NE011	29	238.6	0.35	0.83	3.27	–4.739	**0.004**	**0.006**	**0.007**	**0.007**
25	02UC002	28	2278.2	0.28	0.24	2.78	–45.498	**0.000**	**0.001**	**0.000**	**0.000**
26	03MD001	17	2935.9	0.35	0.87	3.23	–104.85	**0.017**	**0.017**	**0.008**	**0.007**
27	05LJ019	41	6.8	1.02	1.58	5.78	0.088	0.170	0.163	***0.093***	***0.095***
28	02ZH001	44	199.1	0.39	0.65	2.90	2.173	**0.007**	**0.009**	**0.009**	**0.010**
29	09CA002	43	280.9	0.21	–0.65	4.10	1.617	**0.012**	**0.012**	**0.005**	**0.004**
30	10SB001	21	2119.6	0.56	0.66	3.06	101.194	**0.007**	**0.009**	**0.007**	**0.005**

practical utility of the results from the above simulation study. Table 1 presents the identifier (ID) of gauging stations in these basins, the record lengths and the statistics (mean, coefficient of variation (C_v), coefficient of skewness (C_s) and coefficient of kurtosis (C_k)) of annual maximum daily flows. The magnitude of trends in these series, estimated using equation (1) are also listed in Table 1. Figure 13 plots flow series, their means, 5-year moving average series and linear trends. These diagrams only intend to visualize the data and to qualitatively assess the possible existence and type of trend. It is evident that monotonic trends, which are either linear or nonlinear, may exist within these series.

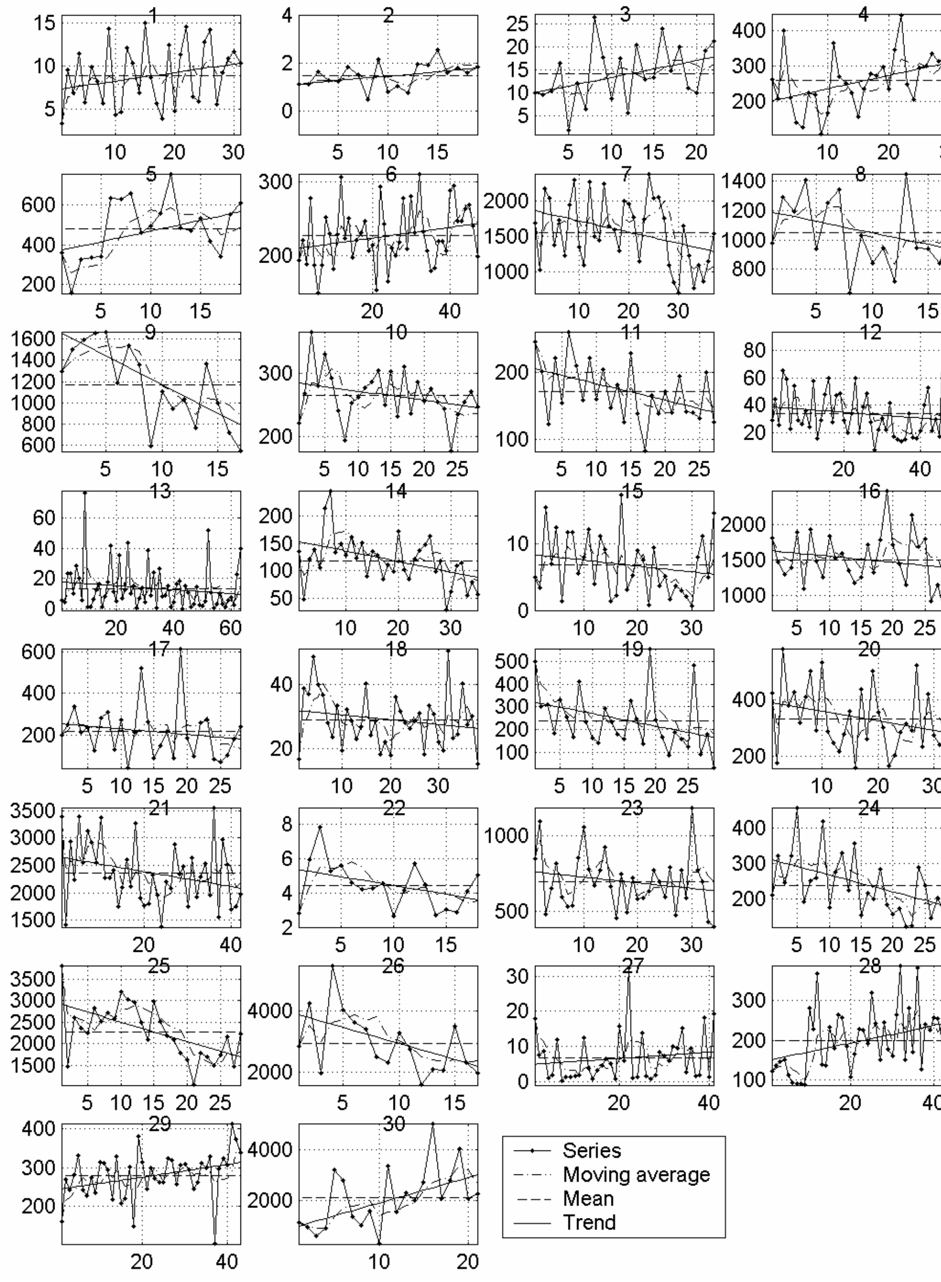

Fig. 13 Visualization of annual maximum daily streamflow series of 30 Canadian pristine river basins.

To assess the statistical significance of the trends in these series, the *P* values for the *t* test, BS-slope test, MK test and BS-MK test were computed. For positive trends, their *P* value (*p*) should be ≥0.50. To be consistent in assessing the significance of positive and negative trends at a given significance level, their *P* value is taken as:

$$p' = \begin{cases} p & \text{for a negative trend} \\ 1-p & \text{for a positive trend} \end{cases} \tag{6}$$

where the probability value *p* is as given by equations (2) and (3) for the BS-slope test and BS-MK tests. The *P* values of these series are presented in the last four columns of Table 1. At a given significance level, the smaller the *P* value, the more significant is the trend. In Table 1, italic bold numbers indicate that the trends are statistically significant at $\alpha = 0.10$ and shaded bold numbers show that the trends are statistically significant at both $\alpha = 0.10$ and 0.05. By comparing the *P* values among these tests, it can be seen that for the data having smaller coefficient of skewness, say $C_s \leq 0.3$, i.e. where the data tend to be nearly symmetrically or normally distributed, the slope-based tests have an increased chance to assess the significance of trends than the rank-based tests, although the difference between them is minor. However, for the series with higher skewness, i.e. when the distribution type is skewed, the rank-based tests are more likely to detect trends. These results are consistent with those obtained from the previous simulation studies.

CONCLUSIONS

In this study, Monte Carlo simulation was applied to assess the power of the parametric *t* test, non-parametric Mann-Kendall (MK), bootstrap-based slope (BS-slope) and bootstrap-based MK (BS-MK) tests to detect monotonic (linear and nonlinear) trends in both normal and non-normal time series. Simulation results indicate that: (a) the *t* test and the BS-slope test, which are slope-based tests, have the same power; (b) the MK and BS-based MK tests, which are rank-based tests, have the same power; (c) for normally-distributed data, the power of the slope-based tests is higher than that of the rank-based tests, but the difference is not great; and (d) for non-normally distributed series, such as time series with the P3, Gumbel, EV2 and Weibull distributions, the power of the rank-based tests is much higher than that of the slope-based tests. The power of the tests is slightly sensitive to the shape of trend, with upward convex shape having the highest power and upward concave shape having the lowest power except for Weibull distributed data. However, in comparison to the impact of the distribution type on the power of the tests, the influence of the shape of trend on the power of the tests is marginal. The assessment of the significance of trends in the annual maximum daily flows of 30 Canadian pristine river basins shows similar results to those obtained in the simulation studies.

The study provides an initial basis for practitioners to select a suitable statistical test based on the sample statistical properties of time series. For approximately normally-distributed series, the slope-based tests should be used to assess the significance of trends, but the rank-based tests can also be applied as the power difference between these two kinds of tests is not great. For non-normal series, the rank-based tests should be employed for trend detection due to their increased ability to detect trends in comparison to the slope-based tests.

Acknowledgements The authors would like to express their thanks to the anonymous reviewers for their comments which improved the quality of the paper.

REFERENCES

Burn, D. H. (1994) Hydrologic effects of climatic change in West Central Canada. *J. Hydrol.* **160**, 53–70.

Burn, D. H. & Hag Elnur, M. A. (2002) Detection of hydrological trends and variability. *J. Hydrol.* **255**(1–4), 107–122.

Cailas, M. D., Cavadias, G. & Gehr, R. (1986) Application of a nonparametric approach for monitoring and detecting trends in water quality data of the St Lawrence River. *Can. Water Poll. Res. J.* **21**(2), 153–167.

Chiew, F. H. S. & McMahon, T. A. (1993) Detection of trend or change in annual flow of Australian rivers. *Int. J. Climatol.* **13**, 643–653.

Davison, A. C. & Hinkley, D. V. (1997) *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, UK.

Demarée, G. R. & Nicolis, C. (1990) Onset of Sahelian drought viewed as a fluctuation-induced transition. *Quart. J. Roy. Met. Soc.* **116**, 221–238.

Douglas, E. M., Vogel, R. M. & Knoll, C. N. (2000) Trends in flood and low flows in the United States: impact of spatial correlation. *J. Hydrol.* **240**, 90–105.

El-Shaarawi, A. H., Esterby, S. R. & Kuntz, K. W. (1983) A statistical evaluation of trends in the water quality of the Niagara River. *J. Great Lakes Res.* **9**, 234–240.

Environment Canada (1999) *Canada's Reference Hydrometric Basin Network.* Atmospheric Monitoring and Water Survey Directorate, Environment Canada, Downsview, Toronto, Canada.

Efron, B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, International Thomson Publication, New York, USA.

Gan, T. Y. (1998) Hydroclimatic trends and possible climatic warming in the Canadian Prairies. *Water Resour. Res.* **34**(11), 3009–3015.

Hipel, K. W. & McLeod, A. I. (1994) Time series modeling of water resources and environmental systems In: *Nonparametric Tests for Trend Detection* (ed. by K. W. Hipel & A. I. McLeod), Ch. 23, 857–931. Elsevier, Amsterdam, The Netherlands.

Hipel, K. W., McLeod, A. I. & Weiler, R. R. (1988) Data analysis of water quality time series in Lake Erie. *Water Resour. Bull.* **24**(3), 533–544.

Hirsch, R. M., Helsel, D. R., Cohn, T. A. & Gilroy, E. J. (1993) Statistical analysis of hydrologic data. In: *Handbook of Hydrology* (ed. by D. R. Maidment), Ch. 17, 17.11–17.37. McGraw-Hill, New York, USA.

Hjorth, J. S. U. (1994) *Computer Intensive Statistical Methods-Validation and Model Selection and Bootstrap*. Chapman & Hall, New York, USA.

Kendall, M. G. (1975) *Rank Correlation Methods*. Griffin, London, UK.

Lall, U. & Sharma, A. (1996) A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.* **32**, 679–693.

Lehmann, E. L. (1975) *Nonparametrics, Statistical Methods Based on Ranks*. Holden-Day, San Francisco, California, USA.

Lettenmaier, D. P. (1976) Detection of trends in water quality data from records with dependent observations. *Water Resour. Res.* **12**(5), 1037–1046.

Lettenmaier, D. P., Wood, E. F. & Wallis, J. R. (1994) Hydro-climatological trends in the continental United States: 1948–88. *J. Climate* **7**, 586–607.

Lins, H. F. & Slack, J. R. (1999) Streamflow trends in the United States. *Geophys. Res. Lett.* **26**(2), 227–230.

Mann, H. B. (1945) Nonparametric tests against trend. *Econometrica* **13**, 245–259.

McLeod, A. I., Hipel, K. W. & Bodo, B. A. (1991) Trend assessment of water quality time series. *Water Resour. Bull.* **19**, 537–547.

Pilon, P. J. & Yue, S. (2002) Detecting climate-related trends in streamflow data. *Water Sci. Technol.* **45**(8), 89–104.

Pilon, P. J., Condie, R. & Harvey, K. D. (1985) *Consolidated Frequency Analysis Package (CFA), User Manual for Version 1.*DEC PRO Series, Water Resources Branch, Inland Water Directorate, Environment Canada, Ottawa, Canada.

Ratkowsky, D. A. (1989) *Handbook of Nonlinear Regression Models*. Marcel Dekker, New York, USA.

Sen, P. K. (1968) Estimates of the regression coefficient based on Kendall's tau. *J. Am. Statist. Assoc.* **63**, 1379–1389.

Simon, J. L. & Bruce, P. (1991) Resampling: a tool for everyday statistical work. Chance. *New Directions for Statistics and Computing* **4**(1), 22–32.

Sneyers, R. (1990) *On the Statistical Analysis of Series of Observations*. Technical Note no. 143, WMO-no. 415, World Meteorological Organization, Geneva, Switzerland.

Stedinger, J. R., Vogel, R. M. & Foufoula-Georgiou, E. (1993) Frequency analysis of extreme events. In: *Handbook of Hydrology* (ed. by D. R. Maidment), Ch. 18, 18.1-18.22. McGraw-Hill, New York, USA.

Stefano, C. D., Ferro, V. & Porto, P. (2000) Applying the bootstrap technique for studying soil redistribution by caesium-137 measurements at basin scale. *Hydrol. Sci. J.* **45**(2), 171–183.

Tasker, G. D. & Dunne, P. (1997) Bootstrap position analysis for forecasting low flow frequency. *J. Water Resour. Plan. Manage.* **123**(6), 359–367.

Taylor, C. H. & Loftis, J. C. (1989) Testing for trend in lake and groundwater quality time series. *Water Resour. Bull.* **25**(4), 715–726.

Theil, H. (1950) A rank-invariant method of linear and polynomial regression analysis, I, II, III, *Nederl. Akad. Wetensch. Proc*. **53**, 386–392; 512–525; 1397–1412.

ven Belle, G. & Hughes, J. P. (1984) Nonparametric tests for trend in water quality. *Water Resour. Res*. **20**(1), 127–136.

Vogel, R. M. & Shallcross, A. L. (1996) The moving blocks bootstrap versus parametric time series. *Water Resour. Res*. **32**(6), 1875–1882.

Yu, Y. S., Zou, S. & Whittemore, D. (1993) Non-parametric trend analysis of water quality data of rivers in Kansas. *J. Hydrol.* **150**, 61–80.

Yue, S. & Wang, C. Y. (2002) Assessment of the significance of serial correlation by the bootstrap test. *Water Resour. Manage.* **16**, 23–35.

Yue, S., Pilon, P. & Cavadias, G. (2002) Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *J. Hydrol.* **259**, 254–271.

Yue, S., Pilon, P. & Phinney, B. (2003) Canadian streamflow trend detection: impacts of serial and cross-correlation. *Hydrol. Sci. J.* **48**(1), 51–63.

Yulianti, J. S. & Burn, D. H. (1998) Investigating links between climatic warming and low streamflow in the Prairies Region of Canada. *Can. Water Resour. J.* **23**(1), 45–60.

Zucchini, W. & Adamson, P. T. (1989) Bootstrap confidence intervals for design storms from exceedence series. *Hydrol. Sci. J.* **34**(1–2), 41–48.

Zetterqvist, L. (1991) Statistical estimation and interpretation of trends in water quality time series. *Water Resour. Res*. **27**(7), 1637–1648.

Received 6 May 2003; accepted 1 August 2003