Stochastic Environmental Research and Risk Assessment (2020) 34:733–754
https://doi.org/10.1007/s00477-020-01797-x

ORIGINAL PAPER

Dissecting innovative trend analysis

Francesco Serinaldi[1,2] · Fateh Chebana[3] · Chris G. Kilsby[1,2]

Published online: 4 May 2020

Abstract
Investigating the nature of trends in time series is one of the most common analyses performed in hydro-climate research. However, trend analysis is also widely abused and misused, often overlooking its underlying assumptions, which prevent its application to certain types of data. A mechanistic application of graphical diagnostics and statistical hypothesis tests for deterministic trends available in ready-to-use software can result in misleading conclusions. This problem is exacerbated by the existence of questionable methodologies that lack a sound theoretical basis. As a paradigmatic example, we consider the so-called Şen's 'innovative' trend analysis (ITA) and the corresponding formal trend tests. Reviewing each element of ITA, we show that (1) ITA diagrams are equivalent to well-known two-sample quantile-quantile (q–q) plots; (2) when applied to finite-size samples, ITA diagrams do not enable the type of trend analysis that it is supposed to do; (3) the expression of ITA confidence intervals quantifying the uncertainty of ITA diagrams is mathematically incorrect; and (4) the formulation of the formal tests is also incorrect and their correct version is equivalent to a standard parametric test for the difference between two means. Overall, we show that ITA methodology is affected by sample size, distribution shape, and serial correlation as any parametric technique devised for trend analysis. Therefore, our results call into question the ITA method and the interpretation of the corresponding empirical results reported in the literature.

Keywords 'Innovative' trend analysis (ITA) · Şen 'test' · Quantile-quantile plots · Linear regression · Uncertainty · Temporal dependence · Methodological inconsistencies · Neutral validation

1 Introduction

Testing trend hypothesis on observed time series is one of the most common exercises reported in the hydro-meteorological literature mainly owing to the interest in detecting possible consequences of human activities on the dynamics of climate and hydrological cycle. Referring to Khaliq et al. (2009) and Bayazit (2015) for an overview of methods, trend analysis usually relies on the application of some statistical hypothesis tests for slowly-varying and/or abrupt changes (e.g. Mann–Kendall (MK), Pettitt, or similar) to summary statistics of hydrological time series (e.g. annual averages, maxima and minima).

However, trend analysis is often performed in a mechanistic way with little attention to the underlying assumptions and the limits of Significance Tests (STs; Cox and Hinkley 1974) for trends. Referring to Wasserstein and Lazar (2016), Wasserstein et al. (2019) and references therein for a thorough discussion on misuse and logical flaws of STs, Serinaldi et al. (2018) attempted to warn practicing hydrologists against a mechanistic use of classical trend tests in the analysis of hydro-climatic time series.

Focusing on practical standpoint, it should be noted that some trend STs suggested in the literature are technically incorrect. An example of these methods is the (still) widely used trend-free prewhitening (TFPW) technique (Yue et al. 2002), whose formal flaws are discussed by Serinaldi and Kilsby (2016a). TFPW technical inconsistencies resulted in contrasting empirical results that led to various but incorrect interpretations about the origin/cause of the detected

✉ Francesco Serinaldi
francesco.serinaldi@ncl.ac.uk

[1] School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK

[2] Willis Research Network, 51 Lime St., London EC3M 7DQ, UK

[3] Institut National de la Recherche Scientifique (INRS), Centre Eau Terre Environnement (ETE), 490, Rue de la Couronne, Quebec, QC G1K 9A9, Canada

trends (e.g. Khaliq et al. 2009; Kumar et al. 2009; Sagarika et al. 2014; Basarin et al. 2016; Pathak et al. 2016; Tananaev et al. 2016; Xiao et al. 2017). Therefore, the problem of technical flaws is particularly important in this context, as trend analysis is often used to support conclusions concerning the evidence of anthropogenic activity on hydro-climatic processes, and their extensive application (due to their relative simplicity) resulted in a large body of literature supporting this hypothesis, irrespective of the fact that STs are not devised to analyze non-randomized samples coming from non-repeatable experiments such as the majority of hydro-climatic records (Flueck and Brown 1993; von Storch 1999; Greenland et al. 2016; Serinaldi et al. 2018).

Şen's innovative trend analysis (ITA) (Şen 2012) is one of many techniques proposed to detect deterministic trends in observed time series. This method attracted the attention of analysts as it was introduced with the appealing (but questionable) claim that this technique "*does not require restrictive assumptions because now classical approaches including most frequently used Mann–Kendall trend test and Sepeard's* [Spearman's] *rho test. The new methodology is valid whatever the sample size, serial correlation structure of the time series, and non-normal probability distribution functions (PDFs). Although the classical methods require prewhitening prior to their applications, such a procedure is not necessary in the proposed methodology in this paper. The validity of the methodology is presented first through extensive Monte Carlo simulation methods*" (Şen 2012). A method that is not affected by sample size, serial correlation and type of distribution surely appears a sort of panacea for trend analysis. However, every statistical method (1) deals with sampling uncertainty and must be sensitive to it, otherwise it would be deterministic, and (2) every statistical analysis relies on weak or strong assumptions (see "Appendix 1"). By the way, Şen (2012) does not report any Monte Carlo simulation despite what is stated in the conclusions of that paper (some Monte Carlo simulations were reported two years later by Şen (2014), and they are discussed below).

Since we were attracted by the apparently amazing properties of ITA and its presentation and justification, we reviewed all the elements and principles of this method, thus performing a so-called *neutral* (independent) validation/falsification analysis (see e.g. Boulesteix et al. 2018, and references therein). This study reports the results of such a review, showing that ITA diagrams are equivalent to two-sample quantile-quantile (q–q) plots, while ITA formal tests, once corrected for mathematical inconsistencies, reduce to standard parametric tests for the difference between two means, and therefore ITA diagnostics are affected by sample size, serial correlation and type of distribution.

This work is structured as follows. Section 2 introduces the rationale of ITA diagnostic plots and explains that they are simple two-sample q–q plots. We also recall the correct interpretation of these diagrams and show that they are affected by sample size, serial correlation and type of distribution. Section 3 further investigates the effect of serial correlation on ITA, showing that several contradicting statements reported in the literature (Şen 2012, 2014, 2017b) depend on the model used to combine deterministic trends and serial correlation, thus stressing that ITA relies on strong assumptions. In Sect. 4, we revise the formal ST proposed within the ITA framework, and show that it is a standard parametric test for the difference between two means, once the mathematical inconsistencies of the ITA formulas are corrected. Section 5 explains why the ITA confidence intervals (CIs) are incorrect and recalls how to build correct CIs. In Sect. 6, we discuss how the above mentioned inconsistencies affect also some methods and analyses derived from ITA. Conclusions and recommendations are summarized in Sect. 7.

2 Setting the stage: overview of ITA and two-sample quantile-quantile plots

Şen's ITA comprises a graphical tool and a formal hypothesis test (Şen 2012, 2014, 2017c). The same material with minor changes has then been collected in a book (Şen 2017b), and directly applied by other authors without any independent assessment of its rationale and mathematical formulation (e.g., Cui et al. 2017; Tosunoglu and Kisi 2017; Wu and Qian 2017; Alashan 2018; Caloiero 2018; Caloiero et al. 2018; Güçlü 2018a; Morbidelli et al. 2018; Zhou et al. 2018; Li et al. 2019).

ITA consists of splitting a time series of (even) size n, $\{x_i\}_{i=1}^{n}$, in two halves of size $n' = n'' = n/2$, $\mathbf{x}' = \{x_i\}_{i=1}^{n/2}$ and $\mathbf{x}'' = \{x_i\}_{i=n/2+1}^{n}$, sorting them in ascending order (i.e. computing the order statistics $x'_{(i)}$ and $x''_{(i)}$, $i = 1, \ldots, n/2$), and then plotting the pairs of sorted values, $(x'_{(i)}, x''_{(i)})$, $i = 1, \ldots, n/2$, against each other. The foregoing procedure is the same used to draw two-sample q–q plots widely applied to check whether two samples (with the same size) have the same distribution (Wilk and Gnanadesikan 1968). In our case, the two samples are the first and second halves of a time series. Even though the interpretation of two-sample q–q plots is well-known and reported in introductory handbooks of applied statistics, it is worth recalling basic properties to support the subsequent discussion. Referring to Fig. 1 and assuming that $\mathbf{x}'$ follows a standard Gaussian distribution, (1) a shift with respect to the 1:1 line corresponds to a shift in the first moment, $\Delta\mu = \mu'' - \mu'$ (or location parameter of the underlying theoretical

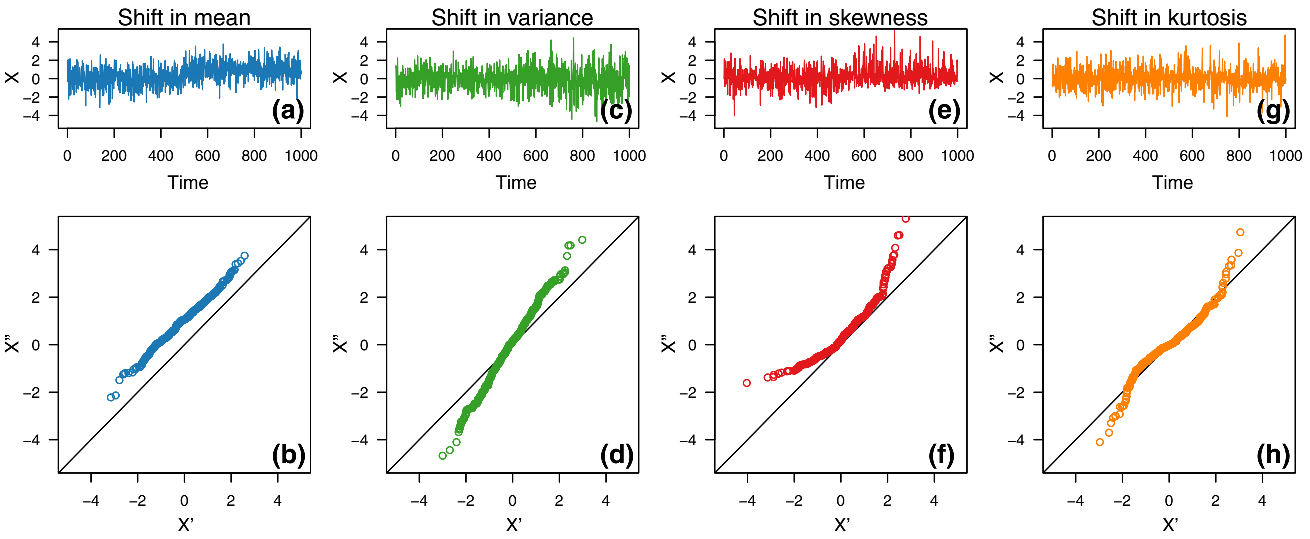

Fig. 1 Examples of time series and corresponding two-sample q–q plots where the distributions of $\mathbf{x}'$ and $\mathbf{x}''$ have different properties. **a**, **b** $\mathcal{N}(0,1)$ versus $\mathcal{N}(1,1)$ (shift in the mean value but same variance). **c**, **d** $\mathcal{N}(0,1)$ versus $\mathcal{N}(0,1.4)$ (same mean but different variance). **e**, **f** $\mathcal{N}(0,1)$ versus $\mathcal{G}(0,0.78)$, where $\mathcal{G}$ is the Gumbel distribution (same mode and variance but different skewness). **g**, **h** $\mathcal{N}(0,1)$ versus $\mathcal{PE}(0,1,0.9)$, where $\mathcal{PE}$ is the power-exponential distribution (same mean, variance, and skewness but different kurtosis)

distribution), where μ' and μ'' are the first moments of $\mathbf{x}'$ and $\mathbf{x}''$, respectively; (2) q–q plots showing approximately linear patterns with slopes different from 1:1 denote discrepancies in the second moment (or scale parameter); (3) J-shaped q–q plots denote differences in the skewness; and (4) S-shape configurations correspond to discrepancies in terms of kurtosis (e.g. Bennett et al. 2013). Of course, there are also other possible patterns depending on the nature of the distribution support (e.g. upper/lower bounded), and the presence of outliers or mixtures of distributions (see e.g., D'Agostino and Stephens 1986, pp. 24–57). Despite this variety of possible cases, in Şen's interpretation, whatever departure from the 1:1 line is considered as an exclusive sign of the presence of a deterministic trend. We stress again that both ITA and two-sample q–q plots take two time series (in this specific case the two halves of a single time series), arrange them in ascending order and plot one of these ordered series versus the other one. Therefore, two-sample q–q plots compare the empirical distributions of two series and do not involve any theoretical distribution. In this respect, it is important to distinguish the general rationale of q–q plots, i.e. comparing two generic distributions, with their standard use for assessing the agreement between the empirical distribution and a theoretical distribution.

Obviously, two-sample q–q plots and therefore ITA diagrams are influenced by the shape of the distribution, serial dependence, and sample size, as these factors influence the uncertainty of the scatter plots of $\mathbf{x}''$ versus $\mathbf{x}'$. A simple Monte Carlo simulation can help visualizing these issues. We consider two distributions, standard Gaussian ($\mathcal{N}(0,1)$) and standard exponential ($\mathcal{E}(1)$), different samples sizes $n = \{50, 100, 500, 1000\}$, and different dependence structures corresponding to first-order autoregressive (AR(1)) processes with parameter $\rho_1 = \{0, 0.5, 0.7, 0.9\}$. For each combination of parameters, 1000 samples are simulated and ITA diagrams (i.e. two-sample q–q plots) are drawn. For $\mathcal{N}(0,1)$, Fig. 2 shows that (1) the scattering of ITA patterns around the 1:1 line decreases as the sample size increases, (2) the scattering and the range of simulated values increase as the serial dependence increases because of variance-inflation effect and reduction of effective size, and (3) the scattering is larger around the tails because of the larger uncertainty of extreme values. Figure 3 shows how the shape of the ensemble of ITA plots changes when the distribution is no longer bell-shaped and symmetric (e.g. Gaussian) but right skewed (e.g. exponential). In particular, as $\mathcal{E}(1)$ is lower bounded, the variability in the lower tail becomes null (Fig. 3). The same remarks hold for the upper tail when the distribution is left skewed and/or upper bounded, *mutatis mutandis*.

These examples highlight that it is rather difficult to obtain ITA plots laying on the 1:1 line even if the two halves of time series, $\mathbf{x}'$ and $\mathbf{x}''$, are drawn from the same distribution without introducing any deterministic trend. Although these properties are well known, according to Şen (2014), ITA diagnostic plot (i.e. the two-sample q–q plot) "*does not require any assumption, and it can be applied in cases of serial dependence, non-normal data distribution, and small sample lengths*". The following arguments support this statement (Şen 2012): "*the basis of the approach rests on the fact that if two time series are*

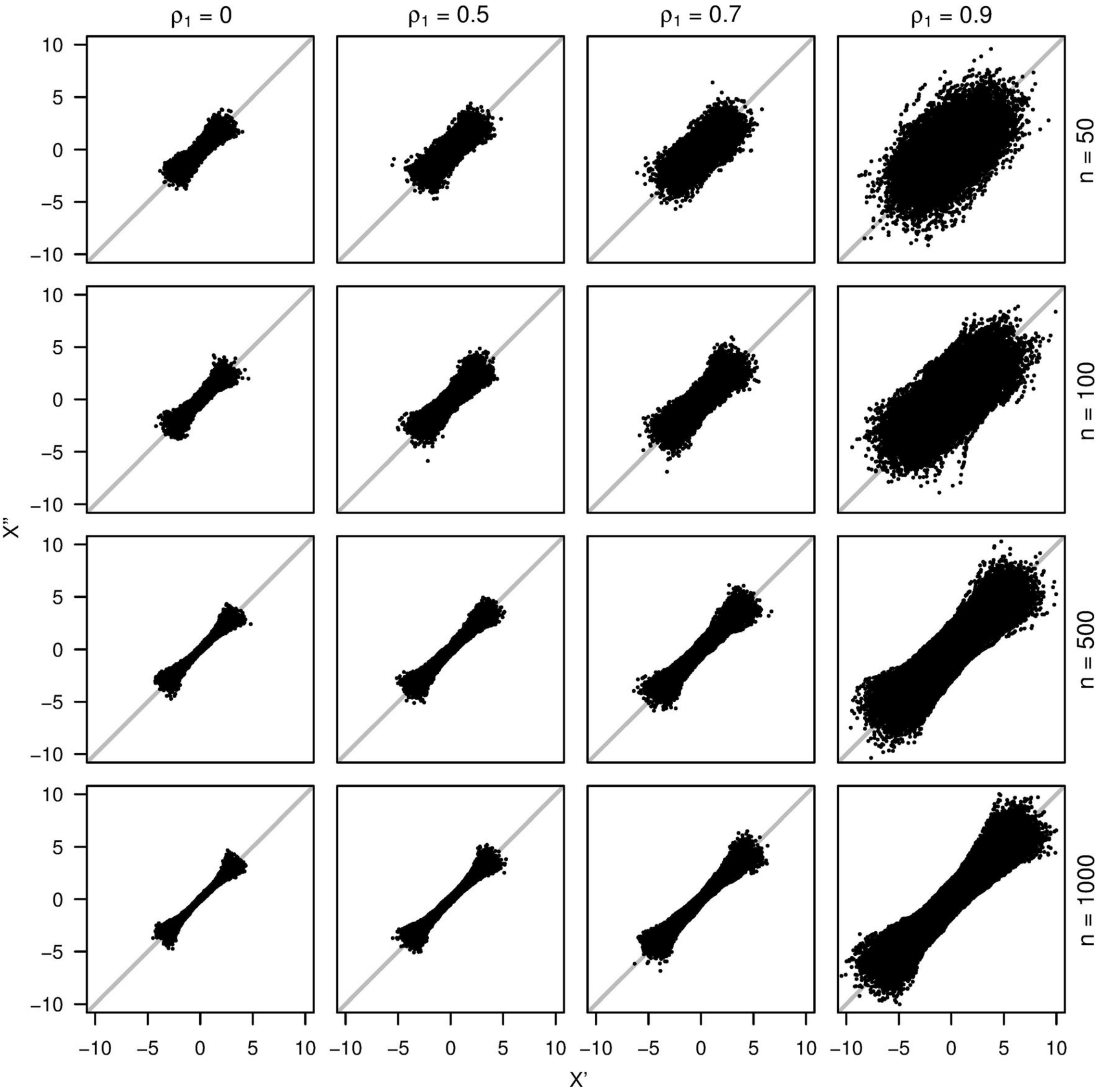

Fig. 2 ITA plots (two-sample q–q plots) for samples drawn from an AR(1) process with parameter $\rho_1 \in \{0, 0.5, 0.7, 0.9\}$, $\mathcal{N}(0,1)$ marginal distribution, and sample size $n \in \{50, 100, 500, 1000\}$. The diagrams show the dependence of sampling uncertainty of ITA plots on ρ_1 and n. The case $\rho_1 = 0$ corresponds to the *i/id* process

identical to each other, their plot against each other shows scatter of points along 1:1 (45°) *line on the Cartesian coordinate system... Whatever the time series are whether trend free or with monotonic trends, all fall on the 1:1 line when plotted. There is no distinction whether the time series are non-normally distributed, having small sample lengths, or possess serial correlations*". This statement, which is true for two *identical* finite-size time series (i.e. perfectly correlated data) or infinite-size sequences (i.e. when dealing with population properties), is then transposed *tout court* to the case of the two halves of the same finite-size time series, overlooking that $\mathbf{x}'$ and $\mathbf{x}''$ are never identical, and their fluctuations and the corresponding ITA plot patterns depend on sample size, serial correlation, and shape of the parent distribution, as shown in Figs. 2 and 3.

The above remarks strongly influence the interpretation of ITA plots. According to Şen (2012), if the patterns fall above (below) the 1:1 line they denote the presence of monotonic increasing (decreasing) deterministic trend, while mixed patterns (i.e., part of the points laying above the 1:1 line and part below) can be related to non-monotonic trends. However, for finite-size samples, the location of the ITA plot with respect to the 1:1 line is neither necessary nor sufficient condition to make conclusions about the presence of deterministic trends. In fact, departures from the 1:1 line can be related to sampling fluctuations, autocorrelation, and shape of distribution without

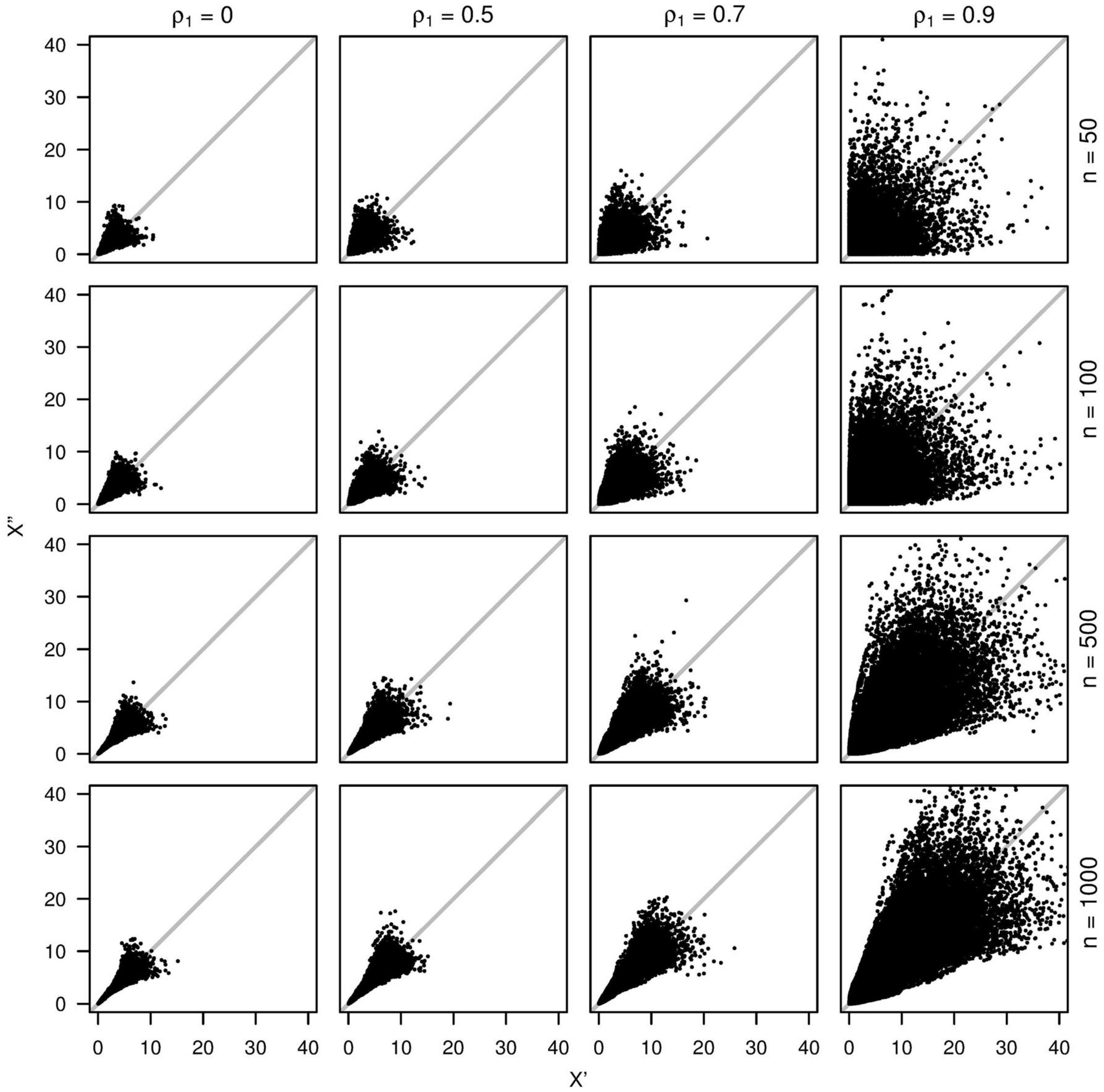

Fig. 3 As for Fig. 2 but with $\mathcal{E}(1)$ marginal distribution

any deterministic trends. On the other hand, the ITA plot can lie on the 1:1 line when there is a deterministic trend resulting in identical distributions of $\mathbf{x}'$ and $\mathbf{x}''$. Figure 4a–d shows that the ITA plots of $\mathbf{x}'$ and $\mathbf{x}''$ can depart from the 1:1 line for samples drawn from a stationary AR(1) process (with $\rho_1 = 0.95$) or a sequence resulting from independent and identically distributed (*i/id*) random variables with standard Gumbel distribution ($\mathcal{G}(0,1)$). Conversely, an increasing linear trend in $\mathbf{x}'$ followed by a decreasing linear trend in $\mathbf{x}''$ can yield indistinguishable sorted samples (Fig. 4e–f). The same can hold true for combinations of linear and nonlinear trends (Fig. 4g–h). The comparison of Fig. 4b and j highlights that almost indistinguishable ITA plot patterns can result from finite-size time series characterized by true deterministic linear trends and sequences from a serially correlated stationary process, thus preventing any discrimination based on this type of diagrams.

The diagrams discussed above suggest another remark. Şen (2012) suggests splitting the ITA plot in three areas corresponding to low, medium, and high values, and therefore studying each subset, interpreting departures from the 1:1 line as possible trends in each class of values. This procedure has three problems, at least: (1) the identification of the three areas in the ITA plot is arbitrary; (2) splitting the samples generally means performing the analysis on very few data points (Figure 4 in Şen (2012) shows examples where the clusters of high quantiles include 3–7 data points); and more importantly (3) Figs. 2, 3, and 4d show that different classes of values (i.e. low, medium and high) exhibit very different departures from the 1:1 line even in cases where there is no ‘trend’, and the

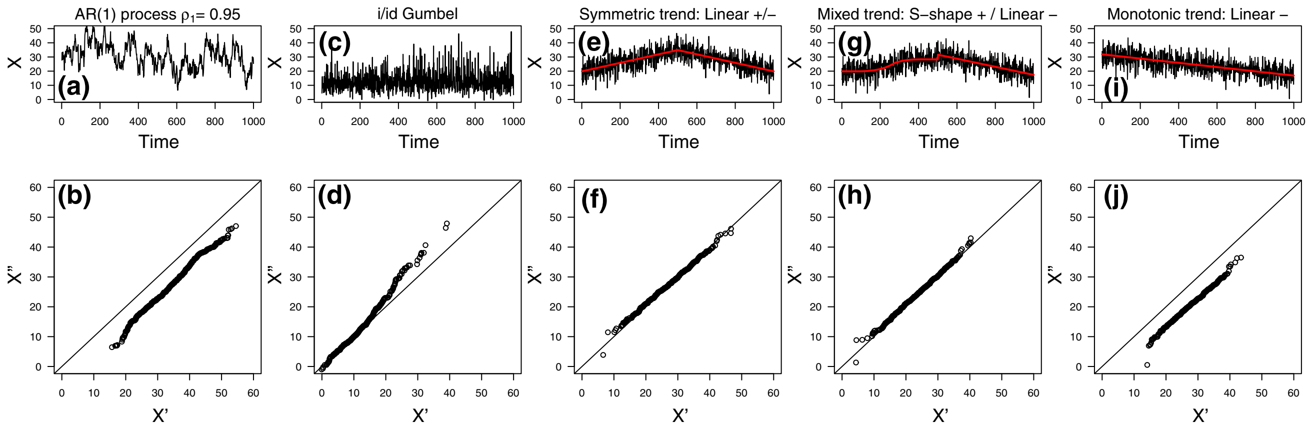

Fig. 4 Counter examples showing how ITA plots (two-sample q–q plots) can be misleading in drawing conclusions about the presence of deterministic trends. **a**, **b** Time series and ITA plots of a time series of size $n = 1000$ simulated from an AR(1) process with $\rho_1 = 0.95$. **c**, **d** Similar to panels (**a**) and (**b**) but for an *i/id* process with $\mathcal{G}(0, 1)$ marginal distribution. **e**, **f** Time series and ITA plots of a sequence resulting from the superposition of an *i/id* process with $\mathcal{N}(0, 1)$ marginal distribution and a non-monotonic deterministic trend linearly increasing (decreasing) in the first (second) half of the time series. **g**, **h** Similar to panels (**e**) and (**f**) but with a S-shaped nonlinear increasing trend in the first half of the time series. **i**, **j** Similar to panels (**e**) and (**f**) but with a linear decreasing trend spanning the entire time series. Panels (**b**) and (**d**) show cases where ITA plots (seem to) indicate departures from the expected 1:1 line even if no deterministic trend is in place. Panels (**f**) and (**h**) show that (almost) perfect alignment along the 1:1 line is possible when non-monotonic trends are in place. Panel (**j**) shows that true deterministic linear trends can yield ITA plots almost indistinguishable from those corresponding to serially correlated time series from a stationary process reported in panel (**b**). See text for further discussion

magnitude of these discrepancies depends to the shape of the generating distribution and further increases as the (possible) serial correlation increases and the sample size decreases. Therefore, interpreting departures of few points from the 1:1 line without considering that such a type of diagrams are affected by serial dependence, shape of the distribution and sample size, is generally misleading. We further discuss the role of the sampling uncertainty and its proper quantification in Sect. 5.

3 Effects of autocorrelation: challenging the principle of non-contradiction

According to Şen (2012), ITA should be unaffected by serial correlation. However, Şen (2017b, pp. 194–196) shows that the shift of the ITA patterns from the 1:1 line increases as the correlation increases for fixed (linear) trend values. Şen (2017b, p. 196) also provides a table showing the values of the shift corresponding to a set of linear trend slopes β and lag-1 autocorrelation ρ_1, claiming that such a "*table can be used to determine the magnitude of monotonic trend in any time series provided that the serial correlation coefficient and the slope on the square area template are determined*". In order to better understand the apparently contradictory statements about dependence or independence of ITA from serial dependence, we repeat the Monte Carlo experiments presented in Şen's original works with the same setting.

In this section, we show that some of the statements about sensitivity of ITA to autocorrelation refer to two different models, one of which is not mentioned in the ITA literature, while conclusions about the ability of ITA to recognize the sign of serial correlation result from incorrect diagrams. We further stress that ITA results are strongly dependent on the model used to merge deterministic trends and serial correlation.

3.1 Are ITA diagrams independent of autocorrelation? Distinguishing population and finite-size sample properties

According to Şen (2017b, pp. 192–193), the effect of serial correlation is studied by superimposing a sequence drawn from a (discrete in time) trend-free stationary first-order autoregressive process with parameter ρ_1 to a linear trend with slope parameter β. This model (hereinafter, M1) is widely used in the literature on trends (Zhang et al. 2000; Wang and Swail 2001; Yue and Wang 2002; Zhang and Zwiers 2004) and reads as follows:

$$\begin{cases} y_t = x_t + \beta t \\ x_t = \rho_1 x_{t-1} + \varepsilon_t, \end{cases} \tag{1}$$

where ε_t is an *i/id* standard Gaussian process. Following Şen (2017b), we simulated single time series of size $n = 10,000$ for various combinations of values of β and ρ_1. Comparing the simulations for ρ_1 equal to zero (independence) and 0.9 [Figures 5.15 and 5.16 in Şen (2017b)], Şen (2017b) concluded that "*Comparison of Figs. 5.15 and*

5.16 indicate that whether the time series is independent or dependent, there is no difference in the square area procedure and as long as the basic time series has a monotonic trend, the appearance of the two-halves sorted magnitude plots will appear along 45° straight-lines without any distinction. This statement alleviates the drawback of the MK trend test, which requires independent data".

Figure 5j and t reproduce the original ITA plot of figures 5.15 and 5.16 in Şen (2017b) [note that the figure 5.15 is identical to figure 3 in Şen (2014)]. Time series corresponding to each ITA plot are reported in panels 5a–i and k–s. Figure 5t shows that the ITA plots corresponding to $\rho_1 = 0.9$ cover a wider range of values (especially for low β values) and are less aligned along the 1:1 line, thus showing some irregular fluctuations when compared with the ITA plots in Fig. 5j (with $\rho_1 = 0$). These differences can appear small and negligible; however, they are the effect of the variance inflation due to the serial correlation and seem to be small only because the analysis refers to relatively long time series. In fact, for $n = 10,000$ and $\beta = 0.003$, the signal-to-noise ratio, here defined as the ratio between the variance of the signal (i.e. the linear trend line) and that of the autoregressive noise, $r_{SN} = \sigma_S^2/\sigma_N^2$, is 75 for $\rho_1 = 0$ and $\cong 14$ for $\rho_1 = 0.9$. Although r_{SN} dramatically decreases as ρ_1 increases, it is still high because the amplitude of the deterministic trend dominates the amplitude of the stochastic component. For $n = 10,000$, even though values of β equal to e.g. 0.003 seem small in terms of absolute value, they correspond to a shift of 30 units between the beginning and the end of the time series, while the range of the superposed noise is one order of magnitude smaller. Therefore, from a theoretical point of view, serial correlation does not change the magnitude of β under M1.

However, in real world analysis, where r_{SN} is usually much smaller, serial correlation affects the ITA plot patterns, and therefore the estimation of the true values of β. Let us consider shorter time series of size $n = 1000$ (Fig. 5u–ad). For $n = 1000$ and $\rho_1 = 0.9$, the fluctuations related to the deterministic trend and stochastic component have the same order of magnitude, thus concealing the linear trend. The lack of alignment of the ITA plots and their mutual overlapping in Fig. 5ad are the affect of the variance inflation due to the serial correlation. In other words, serial correlation increases the variance of the stochastic part, thus concealing the linear pattern of the deterministic component. The latter becomes evident only if the length of the time series is long enough, so that the deterministic shift is much larger than the range of the stochastic fluctuations.

To summarize, from a theoretical point of view (i.e. looking at the population properties), the structure of M1 implies that β does not change with ρ_1; however, from an

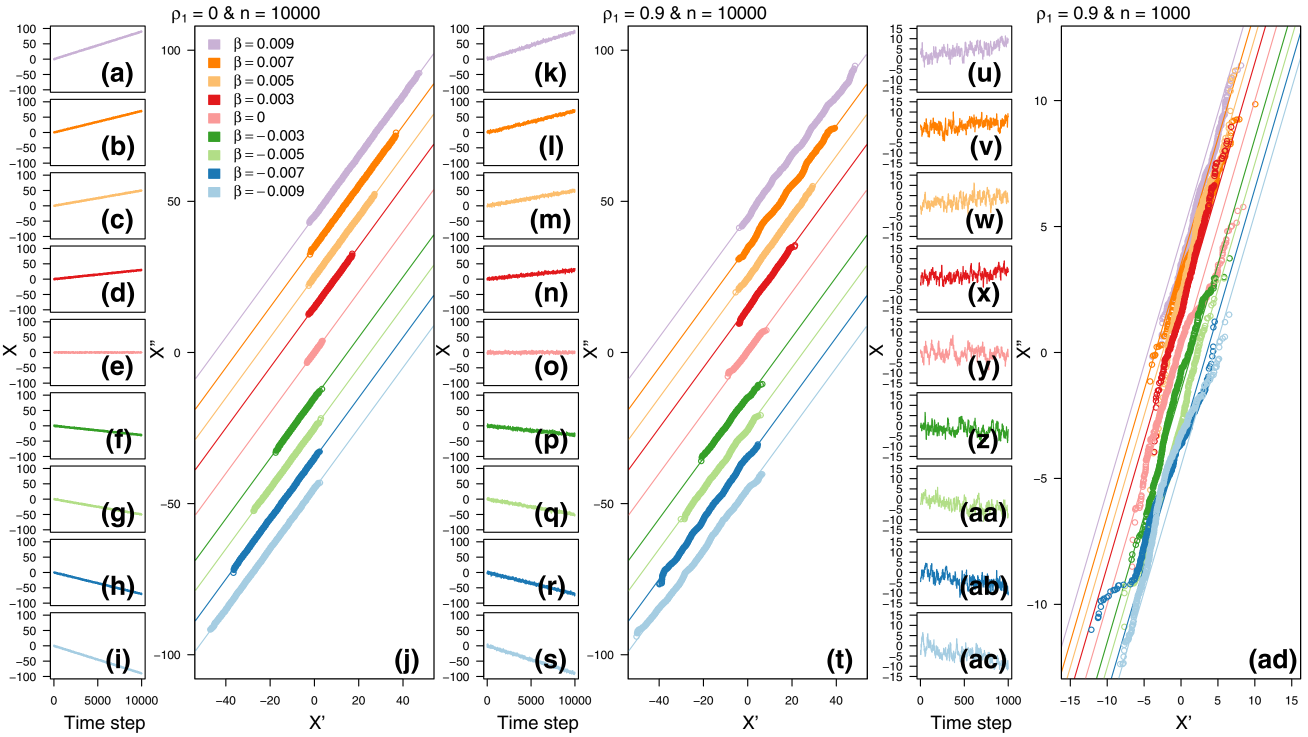

Fig. 5 Simulations from model M1 with varying trend slope β. The diagrams highlight the effect of serial correlation and sample size on the uncertainty of ITA plots (two-sample q–q plots)

operational standpoint (i.e. considering finite-size sample properties), for combinations of β, ρ_1 and n yielding sequences with small r_{SN}, the empirical ITA plots fluctuate, introducing departures from the expected patterns that can be incorrectly interpreted as systematic trends.

3.2 Do ITA diagrams depend on autocorrelation? The role of model assumptions

As discussed above, the theoretical ITA plot patterns (say, for $n \rightarrow \infty$) are invariant to serial correlation for the model M1. However, in a further discussion, Şen (2017b) [pp. 194–198 and figures 5.19, 5.20 and 5.21, which are identical to figures 4, 5 and 6 in Şen (2014)] concludes that "*as the absolute value of the serial correlation coefficient increases the trend representing lines get away from 1:1* (45°) *straight-line basic line*". Therefore, does the slope of the trend line, and thus the shift in the ITA plot, depend or not on serial correlation from a theoretical point of view? To answer this question, we reproduced Figures 5.19 and 5.20 reported by Şen (2017b, p. 197) in Fig. 6a–b. The patterns of time series and ITA plots shown in Fig. 6a–b cannot be produced by model M1 (Eq. 1) as they correspond to the following one (hereinafter, M2):

$$y_t = \rho_1 y_{t-1} + \beta t + \varepsilon_t. \tag{2}$$

Figure 6c–d shows results for M1 (with fixed β and varying ρ_1) for the sake of comparison. Model M2 is not mentioned in any Şen's works, which exclusively refer to M1. The fundamental missing information in Şen (2014, 2017b) is that the results concerning the (theoretical) dependence of trend slope on serial correlation (actually ρ_1), are not general but model-dependent, and cannot be mechanistically applied to real-world data for at least two reasons: (1) observed data do not come for sure from such models, which are only approximations, and (2) the interpretation of ITA diagrams depends on the assumed model. Therefore, β is theoretically dependent of ρ_1 only under model M2.

Using Monte Carlo simulations, Şen (2017b) studied the relationship between β, ρ_1 and the shift of the ITA plots from the 1:1 line [Figure 5.18 and table 5.1 in Şen

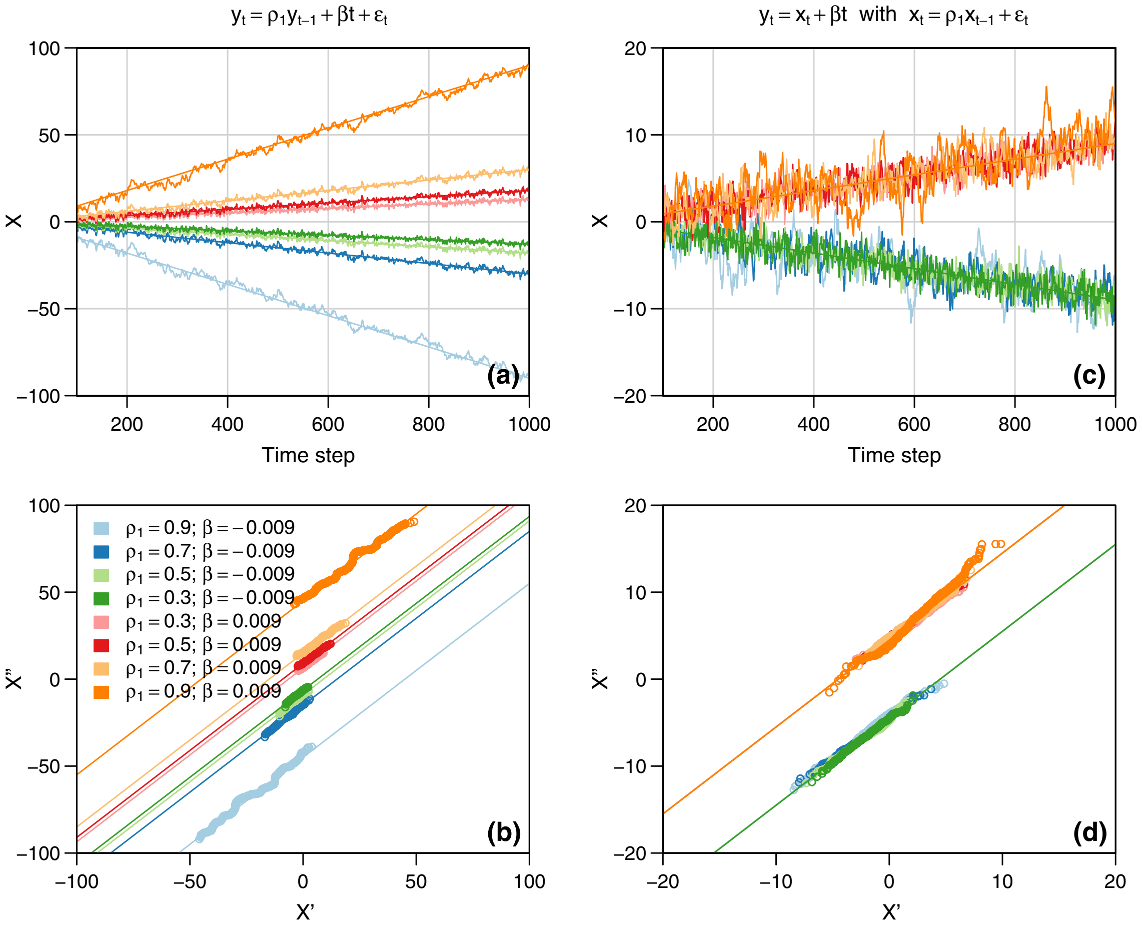

Fig. 6 **a**, **b** Time series and ITA plots (two-sample q–q plots) of time series simulated from model M2 with fixed β and varying ρ_1. **c**, **d** Similar to panels (**a**) and (**b**) but for time series drawn from model M1

(2017b)], concluding that "*This table [5.1] can be used to determine the magnitude of monotonic trend in any time series provided that the serial correlation coefficient and the slope on the square area template are determined*". However, this statement holds true only for time series coming from model M2 (which is not mentioned in any Şen's work), while it does not if we assume M1 or other models combining linear trends and correlated process in different ways.

Moreover, the properties of model M2 are well known (van Giersbergen 2005; Hamed 2009). In particular, recalling that $n' = n/2$, under M2, the theoretical relationships between β, ρ_1 and the shift of the ITA plots from the 1:1 line is (see "Appendix 2")

$$\Delta\mu \cong \frac{\beta n'}{1 - \rho_1}. \tag{3}$$

Equation (3) yields the exact values corresponding to the approximate Monte Carlo $\Delta\mu$ reported in table 5.1 of Şen (2017b) for $n = 1000$ (see Table 1 for a comparison) and shows that $\Delta\mu$ depends on n. Therefore, Şen's numerical results are already known from theoretical standpoint and are not general, as they strictly depend on sample size and the specific model assumed to describe the observed time series. Neglecting these issues and claiming that those results are valid with no assumptions or restrictions can lead to incorrect conclusions.

3.3 Can ITA diagrams reveal the sign of autocorrelation? A matter of incorrect labeling

Another apparent property of ITA diagrams should be their capability to distinguish between positive and negative correlation (Şen 2017b, p. 194). This conclusion is based on Şen's interpretation of Figure 5.17 of Şen (2017b), which is reproduced in Fig. 7a to support our discussion. The ITA plots in Fig. 7a can be obtained only if the corresponding time series look like those shown in Fig. 7b. This would mean that negative values of ρ_1 should be able to invert the sign of the observed trend, which is not possible. In fact, for model M2, the mean depends on time according to the relationship $\mu_t \cong \frac{\beta t}{1-\rho_1}$ (Eq. 14 in "Appendix 2"). Therefore, the effective trend of the process is greater (smaller) than β for positive (negative) values of ρ_1, but it is always positive with a minimum equal to $\beta/2$ for $\rho_1 = -1$. For model M1, β does not depend on ρ_1. Figure 7c–f shows correct ITA plots and time series corresponding to models M1 and M2 for fixed β and varying ρ_1, and confirms the foregoing theoretical remarks.

Figure 5.17 in Şen (2017b) (here, Fig. 7a), which is the support of Şen's conclusions, does not report results for fixed β and varying ρ_1. Actually, it refers to model M2 with positive ρ_1 values and $\beta \in \{-0.09, 0.09\}$, which is indeed similar to Fig. 6a for $\beta \in \{-0.009, 0.009\}$. Therefore, the supposed ability of ITA plots to highlight positive and negative correlation results from a speculation around a diagram with incorrect labels [Figure 5.17 in Şen (2017b)] that does not show what is supposed to do.

4 ITA test for trends: mathematical inconsistencies and equivalence to standard parametric test for the difference between two means

The ITA plots come with a formal ST (Şen 2017c). Similar to ITA diagrams, this ST is introduced claiming that it "*has non-parametric basis without any restrictive assumption,*

Table 1 Effective trend slope of time series drawn from model M2 for $n = 1000$

Trend slope β	AR(1) ρ_1					
	$0^{\dagger}$	0.1	0.3	0.5	0.7	0.9
− 0.09	− 45 (− 45)	− 50.048 (− 50)	− 64.343 (− 64.286)	− 90.08 (− 90)	− 150.133 (− 150)	− 450 (− 450)
− 0.07	− 35 (− 35)	− 38.934 (− 38.889)	− 50.058 (− 50)	− 70.08 (− 70)	− 116.8 (− 116.667)	− 350 (− 350)
− 0.05	− 25 (− 25)	− 27.824 (− 27.778)	− 35.772 (− 35.714)	− 50.08 (− 50)	− 83.465 (− 83.333)	− 250 (− 250)
− 0.03	− 15 (− 15)	− 16.713 (− 16.667)	− 21.486 (− 21.429)	− 30.078 (− 30)	− 50.131 (− 50)	− 150 (− 150)
0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
0.03	15 (15)	16.624 (16.667)	21.372 (21.429)	29.92 (30)	49.871 (50)	150 (150)
0.05	25 (25)	27.736 (27.778)	35.658 (35.714)	49.921 (50)	83.205 (83.333)	250 (250)
0.07	35 (35)	38.846 (38.889)	49.944 (50)	69.922 (70)	116.538 (116.667)	350 (350)
0.09	45 (45)	49.957 (50)	64.223 (64.286)	89.922 (90)	149.872 (150)	450 (450)

Şen's values obtained by simulation (Şen 2017b, table 5.1) are compared with those resulting from Eq. (3) (in parentheses)

$^{\dagger}\rho_1 = 0$ corresponds to the *i/id* case

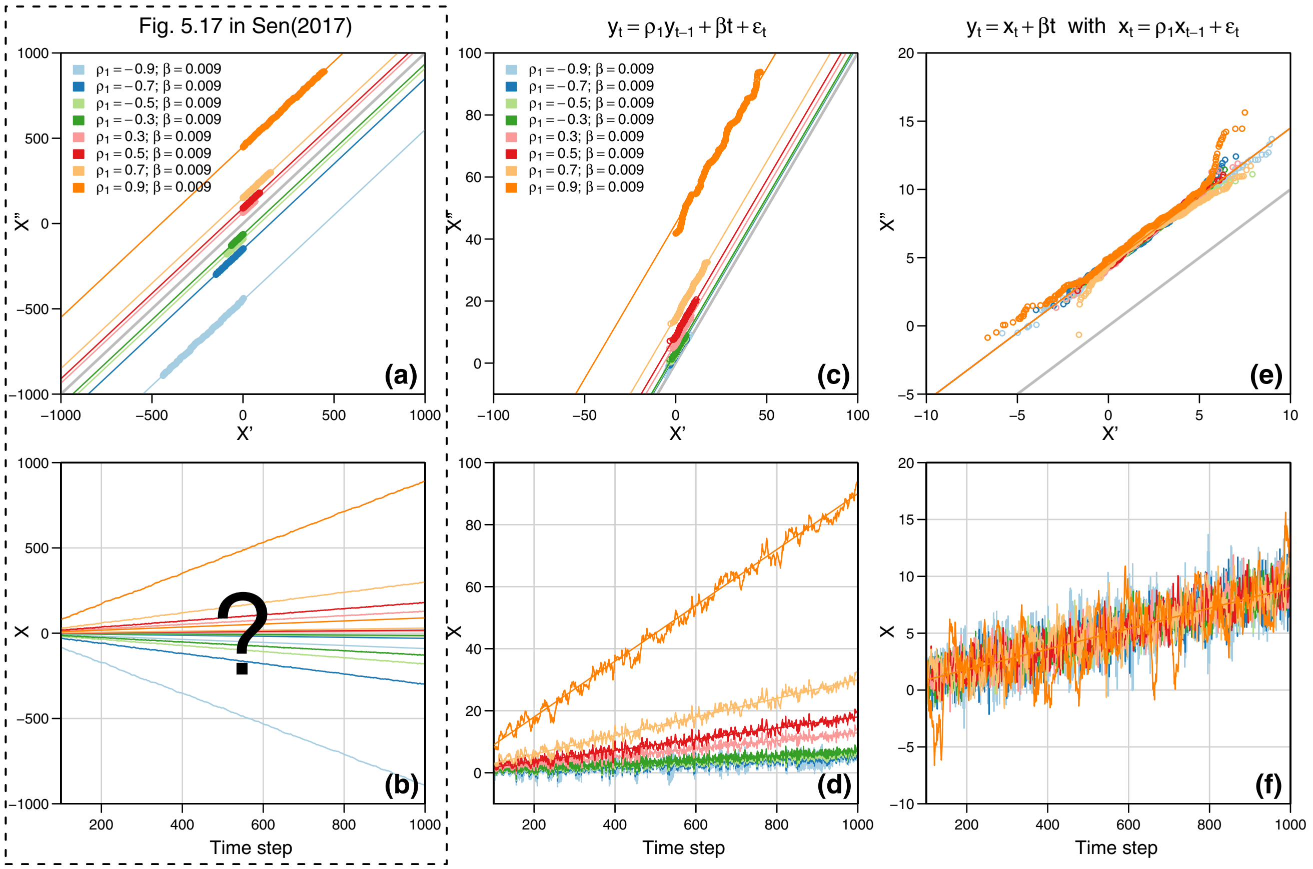

Fig. 7 Effect of negative serial correlation. Panel (**a**) reproduces ITA plots (two-sample q–q plots) of figure 5.17 in Şen (2017b), while panel (**b**) shows the corresponding time series (note that variability around the trend lines appears very small because of the high signal-to-noise ratio). Panels (**c**) and (**d**) depict the correct results for the model M2. **e**, **d** Similar to panels (**c**) and (**d**) but for the model M1

and its application is rather simple with the concept of sub-series comparisons that are extracted from the main time series... The suggested methodology is valid even for time series with serial correlation structure" (Şen 2017c). In this section, we double check Şen's test formalism and verify if this test is really assumption-free.

The first (rather strong) assumption is that this formal test only deals with linear trends, while true rank-based ('non-parametric') tests, such as MK, deal with more general monotonic trends. In fact, Şen's test aims at establishing the statistical significance of the slope parameter β of a linear trend $x = \alpha + \beta t$ (Şen 2017b, p. 200), where β is estimated by the sampling averages, m' and m'', of $\mathbf{x}'$ and $\mathbf{x}''$, respectively (Şen 2017b, p. 201)

$$\hat{\beta} = \frac{2(m'' - m')}{n} = \frac{m'' - m'}{\frac{3n}{4} - \frac{n}{4}} = \frac{m'' - m'}{\tau'' - \tau'}, \tag{4}$$

where $\tau' = \frac{n}{4}$ and $\tau'' = \frac{3n}{4}$ are the averages of the sequences of time steps $\{1, 2, \ldots, n/2\}$ and $\{n/2 + 1, n/2 + 2, \ldots, n\}$ in the first and second half of the time series $\{x_1, \ldots, x_n\}$.

Generally, there is neither empirical nor theoretical argument justifying the supposed evolution of natural processes according to straight lines (see Serinaldi et al. 2018, for a discussion). Moreover, systematic deviations of ITA diagrams from the 1:1 line do not necessarily correspond to linear trends. In fact, sequences of observations exhibiting a monotonic trend in the mean (or whatever central tendency index) yield a shift in ITA plots, which therefore do not allow for distinguishing linear or nonlinear trends in the original time series. Figure 8 shows that time series with linear trend, abrupt change or S-shaped trend can yield indistinguishable ITA plots. Since ITA plots cannot provide any evidence about the existence of a linear trend, testing the statistical significance of β is arbitrary.

Let μ' and μ'' be the population means corresponding to the sample means m' and m''. Testing $\beta = \frac{2(\mu'' - \mu')}{n}$ means testing the difference between two means, and ITA plots do not play any role in this formulation. In fact, Şen test is only the most common test for the difference between two means for two samples of the same size $n/2$, where the two populations are Gaussian with known and identical standard deviation $\sigma' = \sigma'' = \sigma$. This test is reported in every statistical handbook as one of the simplest examples of ST,

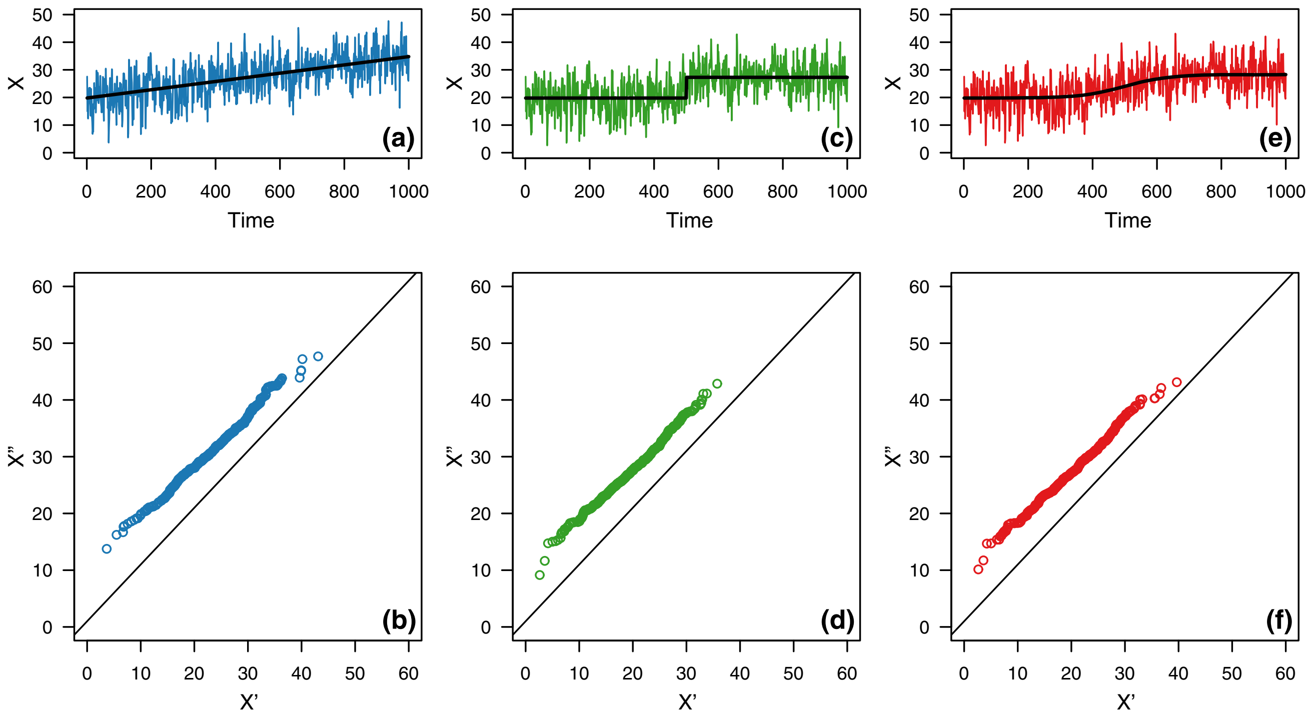

Fig. 8 Time series and ITA plots (two-sample q–q plots) for three different types of monotonic deterministic trends (linear (**a**), step-wise (**c**), and nonlinear S-shaped (**e**)). Panels (**b**), (**d**) and (**f**) show that different trends can correspond to similar ITA plots

and it is also fully parametric and (obviously) affected from serial dependence.

Under the null hypothesis, $H_0 : \beta = 0$, Şen's test assumes that the following test statistic has standard Gaussian distribution

$$t_{\mathrm{ITA}} = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} = \frac{2}{n} \frac{(m'' - m') - (\mu' - \mu'')}{\sigma_{m''-m'}} = \frac{2}{n} \frac{\Delta m - \Delta \mu}{\sigma_{\Delta m}} = \frac{2}{n} t_{\mathrm{standard}}, \tag{5}$$

in which $\mathrm{E}[\hat{\beta}] = \beta = 0$ and

$$\sigma^2_{\hat{\beta}} = \mathrm{Var}[\hat{\beta}] = \frac{8}{n^2} \frac{\sigma^2}{n} (1 - \rho_{m'm''}), \tag{6}$$

where $\rho_{m'm''}$ is the cross-correlation coefficient of the sample means m' and m'', while t_{standard} is discussed later.

Despite the claims about the lack of assumptions of this test, it actually implies a number of strong assumptions:

1. The test statistic in Eq. (5) is normally distributed if the sampling distribution of m' and m'' is Gaussian. For small samples, this property requires that $\mathbf{x}'$ and $\mathbf{x}''$ are normally distributed as well. This assumption can be relaxed for large samples sizes ($n \to \infty$) according to the central limit theorem (Mood et al. 1974, p. 234-236), bearing in mind that the convergence of the sampling distribution of m' and m'' to Gaussian can be very slow when the distribution of the parent process X is skewed and/or heavy tailed.
2. The derivation of the variance in Eq. (6) requires that the two samples $\mathbf{x}'$ and $\mathbf{x}''$ are homoscedastic (Şen 2017b, p. 205), and the variance of the parent process, σ^2, is known. In fact, if σ^2 is unknown and estimated from the sample standard deviations s' and s'', the test statistic is no longer normally distributed but follows a Student distribution with $n-2$ degrees of freedom (Mood et al. 1974, p. 432-435).

These assumptions are the same characterizing the standard test for differences between two means (*using known variances*) relying on the test statistic t_{standard} (Kottegoda and Rosso 2008, pp. 252), which is identical to Şen's t_{ITA} up to the factor $2/n$ (Eq. 5). Note that the expression of t_{ITA} also neglects that the variances are actually unknown and estimated on the data. The direct comparison of the two methods also reveals that Eq. (5) is incorrect, as it assumes that $\sigma_{m'} = \sigma_{m''} = \sigma/\sqrt{n}$ (see Şen (2017b) p. 205 and Şen (2017c) p. 946), while the variances of the sample means over samples of size $n/2$ are equal to $\sigma/\sqrt{n/2}$, resulting in the corrected expression

$$\sigma^2_{\hat{\beta}} = \frac{16}{n^2} \frac{\sigma^2}{n} (1 - \rho_{m'm''}), \tag{7}$$

which returns indeed the variance of t_{standard}, $4\sigma^2/n$, when we remove the nuisance factor $2/n$ in Eq. (5) and set

$\rho_{m'm''} = 0$. Therefore, Şen's expression in Eq. (6) underestimates the actual variance of t_{ITA} of a factor two.

However, the main theoretical inconsistency in Şen's formulation is not the foregoing multiplicative factor but the interpretation and estimation of $\rho_{m'm''}$. At its first appearance in the derivation of $\sigma^2_{\hat{\beta}}$, $\rho_{m'm''}$ is correctly introduced as the "*cross-correlation coefficient between the ascendingly sorted two-halves–arithmetic averages*" (Şen 2017b, p. 205). However, in the subsequent paragraph, Şen (2017b, p. 205) states that "*the most significant point in the application of this formulation is that the cross-correlation is between the two-sorted half time series*". This statement is also repeated by Şen (2017c, p. 246), and this definition is used in the applications, resulting in very high correlation values (reflecting the alignment of the points in the ITA diagram), and thus very low values of $\sigma^2_{\hat{\beta}}$ [see lines 6 and 7 in table 5.3 of Şen (2017b)]. Şen (2017b) describes these values saying that "*one of the important points in this table is high cross-correlation values in row 6* [of Table 5.3], *because they are calculated depending on the ordered sequence in each half series*".

Firstly, it is (or should be) obvious that the sample means of the two sub-series $\mathbf{x}'$ and $\mathbf{x}''$ do not change if the two samples are sorted or not, and thus Şen's test statistic is not related in any way to ITA plots. Moreover, if the data are uncorrelated, the sample means m' and m'' are uncorrelated as well, i.e. $\rho_{m'm''} = 0$. Secondly, the correlation $\rho_{m'm''}$ between the sample means m' and m'' of the two samples $\mathbf{x}'$ and $\mathbf{x}''$ *is not* the correlation $\rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}$ between the pairs of sorted values, $(x'_{(i)}, x''_{(i)})$, $i = 1, \ldots, n/2$, reported in ITA plots. A *reductio ad absurdum* argument can prove the theoretical inconsistency of switching $\rho_{m'm''}$ with $\rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}$. Under *i*/*id* conditions (i.e. lack of trend and persistence), for large n and neglecting the sampling uncertainty, the points of the ITA plot are approximately well aligned along the 1:1 line and $\rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}} \cong 1$. Replacing $\rho_{m'm''}$ with $\rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}$ into Eq. (7), it follows that $\sigma^2_{\hat{\beta}} \cong 0$ as $(1 - \rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}) \cong 0$. In this case, every empirical estimate of t_{ITA} that is not almost exactly equal to zero indicates a significant trend. In other words, $\sigma^2_{\hat{\beta}} \cong 0$ with or without the presence of trends. On the other hand, under *i*/*id*, the estimates of the sample means from two samples are uncorrelated with $\rho_{m'm''} \cong 0$ (the values of $\rho_{m'm''}$ under *i*/*id* and serial correlation are further investigated by Monte Carlo simulations in "Appendix 3"). Therefore, $\rho_{m'm''} \cong 0 \neq 1 \cong \rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}$. The unjustified (and theoretically unjustifiable) replacement of $\rho_{m'm''}$ with $\rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}$ strongly deflates the variance of the test statistics, thus leading to an incorrect and dramatically high rate of rejection of the null hypothesis when it is true, i.e. an effective level of significance much higher than the desired target level (see Monte Carlo simulations and additional discussion in "Appendix 4").

5 Theoretical inconsistency of confidence intervals of ITA plots and corresponding significance test

Even though ITA plots are introduced as diagnostic tools that are not affected by sample size, serial correlation, and distributional assumptions, Şen (2017b, pp. 314–317) suggests quantifying their sampling uncertainty by confidence intervals (CIs) describing the expected fluctuations of the pairs of order statistics $(x'_{(i)}, x''_{(i)})$ around the 1:1 line in the ITA plots under the assumption of no trend. As usual, the distribution used to build CIs is also used to introduce a formal ST on the significance of the departures of ITA plots from the 1:1 line (Şen 2017b, pp. 297–304). We note some logical contradiction of suggesting statistical tests (as those in Sect. 4) and CIs to complement a method that is supposed to be inherently free from sample size effects. However, this contradiction can be due to the lack of distinction between population and sample properties in the original description of these methods. Nonetheless, Şen (2017b, pp. 297–304) introduced such a test and CIs as follows.

5.1 Reviewing ITA test for departures from the 1:1 line

Under the assumption of no trend (in the central tendency measures such as the mean), we expect that the difference between the sample means of $\mathbf{x}'$ and $\mathbf{x}''$ has expected value $\mathrm{E}[\mu'' - \mu'] = 0$. We also expect that the pairs of order statistics $(x'_{(i)}, x''_{(i)})$ in the ITA plots are aligned along the 1:1 line with small departures. Even though we have shown in Sect. 2 that the latter condition is neither necessary nor sufficient in empirical analysis of finite-size samples, such departures are quantified by the "*square root of square deviation summation (SRSDS),* s_{d}, *between the two half series scatter points from the 1:1 line as*"

$$s_{\mathrm{d}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n/2}(X_i - X_{n/2+1})^2}. \tag{8}$$

Therefore, according to Şen (2017b, p. 303), "*in order to convert this information into an objective form the division of the mean difference,* $(m_2 - m_1)$ [i.e., $(m'' - m')$ in the present notation] , *by the SRSDS in Eq. 7.16* [i.e. Eq. (8) in this paper], *leads to the definition of trend test statistic,* t_{s}, *as*"

$$t_{\mathrm{s}} = \frac{m'' - m'}{s_{\mathrm{d}}}. \tag{9}$$

Finally, Şen (2017b, p. 303) provides the following interpretation: "*The small values of this test statistics, t_{s}, imply that there is trend and variability, which is regarded as the null hypothesis, H_{o}. On the contrary, the big values corresponds to the alternative hypothesis, H_{a}, where there is no trend or variability. Theoretically, t_{s} has zero mean and unit variance, and hence, the standard normal pdf can be used for the significance test*".

Focusing on the analytical and conceptual inconsistencies, firstly but least, (1) $X_{n/2+1}$ should be $X_{n/2+i}$; (2) using this correction, X_i and $X_{n/2+i}$ should be $x_{(i)}$ and $x_{(n/2+i)}$ as Eq. (8) refers to differences between corresponding order statistics; (3) the factor $1/n$ should be $2/n$ because the sum is taken over $n/2$ terms; and (4) the suggested interpretation is incorrect, as the null hypothesis of 'no trend' corresponds to $t_{\mathrm{s}} \rightarrow 0$; in fact, if the test statistic in Eq. (9) is standard normal under the null, it means that $(m'' - m') \rightarrow 0$, and this can happen only if $m'' \cong m'$, i.e. if the null hypothesis is 'no trend', as usual in standard statistical testing.

Secondly and most important, the statistic t_{s} has neither unit variance nor Gaussian distribution because the expression of the sample variance s_{d}^2 is not consistent with the numerator in Eq. (9) and does not provide a valid standardization factor. In fact, generally speaking, formulas yielding standardized statistics with zero mean and unit variance require subtraction of the expected value (here, $\mathrm{E}[\mu'' - \mu'] = \mathrm{E}[\Delta\mu] = 0$) and division by the standard deviation of the variable of interest, which is $\Delta m = (m'' - m')$ in the present case. However, the standard deviation of Δm is not s_{d} but $\sigma_{\Delta m}$ in Eq. (5). Using $\sigma_{\Delta m}$, t_{s} becomes identical to the statistic t_{original} in Eq. (5), which is actually distributed as $\mathcal{N}(0, 1)$, thus revealing that also this test is once again nothing but the standard test for the difference between two means reported in every and handbook of applied statistics (see e.g. Kottegoda and Rosso 2008, pp. 252–253). The corrected statistic t_{s} is also identical to t_{ITA} up to the factor $2/n$. Moreover, such tests rely on several assumptions and depend on the preliminary knowledge (or lack of knowledge) of the population variances and serial dependence. In fact, the expression of $\sigma_{\Delta m}$ assumes different forms according to the specific case at hand. For example, under serial independence, homoscedasticity (i.e. $\sigma' = \sigma'' = \sigma$), and same sample size $n' = n'' = n/2$, if σ is known (not estimated from the same sample), we have (see e.g. Kottegoda and Rosso 2008, pp. 252) $\sigma_{\Delta m} = \frac{2}{\sqrt{n}}\sigma$, while

$$\sigma_{\Delta m} = \sqrt{\frac{2}{n}\left(s'^2 + s''^2\right)} \tag{10}$$

if the population standard deviations σ' and σ'' are unknown but equal, and s' and s'' are their sample version. Under serial dependence, $\sigma_{\Delta m}$ should be multiplied by a correction factor, f_{corr}, to account for the variance inflation effects yielding

$$\begin{cases} \sigma^*_{\Delta m} = \sigma_{\Delta m}\sqrt{1 + \left(\frac{n}{2} - 1\right)\bar{\rho}} = \sigma_{\Delta m} f_{\mathrm{corr}} \\ \bar{\rho} = \dfrac{\sum\sum_{j \neq l} \rho_{ij}}{\frac{n}{2}\left(\frac{n}{2} - 1\right)} \end{cases} \tag{11}$$

where $\bar{\rho}$ is the average of the off-diagonal elements of the correlation matrix of $n/2$ variables, and $\rho_{ij} = \mathrm{Corr}\,[X_i, X_j]$ denotes the pairwise correlation of X_i and X_j (Matalas and Langbein 1962).

Some Monte Carlo experiments further clarify the above criticisms. We simulated 1000 time series of size $n = 100$ from the *i*/*id* model and an AR(1) process with $\rho_1 = 0.9$. For each series, we computed t_{s} according to Eqs. (8) and (9) (corresponding to Eqs. 7.16 and 7.17 in Şen (2017b, p. 303)), and t_{standard} using the variances in Eq. (10) for the *i*/*id* case and Eq. (11) for the AR(1) process. Figure 9 shows that the distribution of t_{s} is far from being $\mathcal{N}(0, 1)$ and its shape depends on the serial correlation as expected, while the empirical probability density function of t_{original} is close to $\mathcal{N}(0, 1)$ for both processes, thus confirming the effectiveness of the correction factor f_{corr}. Note that different values of n and ρ_1 yield similar results (not shown). Since Şen's t_{s} is not Gaussian and depends on serial correlation, it follows that $\mathcal{N}(0, 1)$ cannot be used to compute valid critical values to perform a statistical test.

5.2 Reviewing the CIs of ITA diagrams

Even though s_{d} cannot be used to describe the variance of Δm, thus making the test based on t_{s} invalid, one can think that s_{d} can be applied at least to build CIs around the 1:1 line. Indeed, in principle s_{d} should describe the variance of the fluctuations of the order statistics of $\mathbf{x}''$ with respect to those of $\mathbf{x}'$ (after correcting the expression in Eq. (8) for the formal errors mentioned above). Therefore, if such fluctuations are approximately Gaussian, we can define confidence limits from the distribution $\mathcal{N}(0, s_{\mathrm{d}}^2)$. However, this is not correct either, because each order statistic appearing in the ITA plot has its own distribution, which is a beta of the form $F_{X_{(i)}} = \mathcal{B}(F_X(x_{(i)}); \nu p, \nu(1 - p))$, where F_X is the parent distribution of X, $\nu = n' + 1$, and $p \in [1/\nu, n'/\nu]$ (Stigler 1977; Hutson 1999; Nadarajah and Gupta 2004; Serinaldi 2009). Therefore, the ensemble of fluctuations of

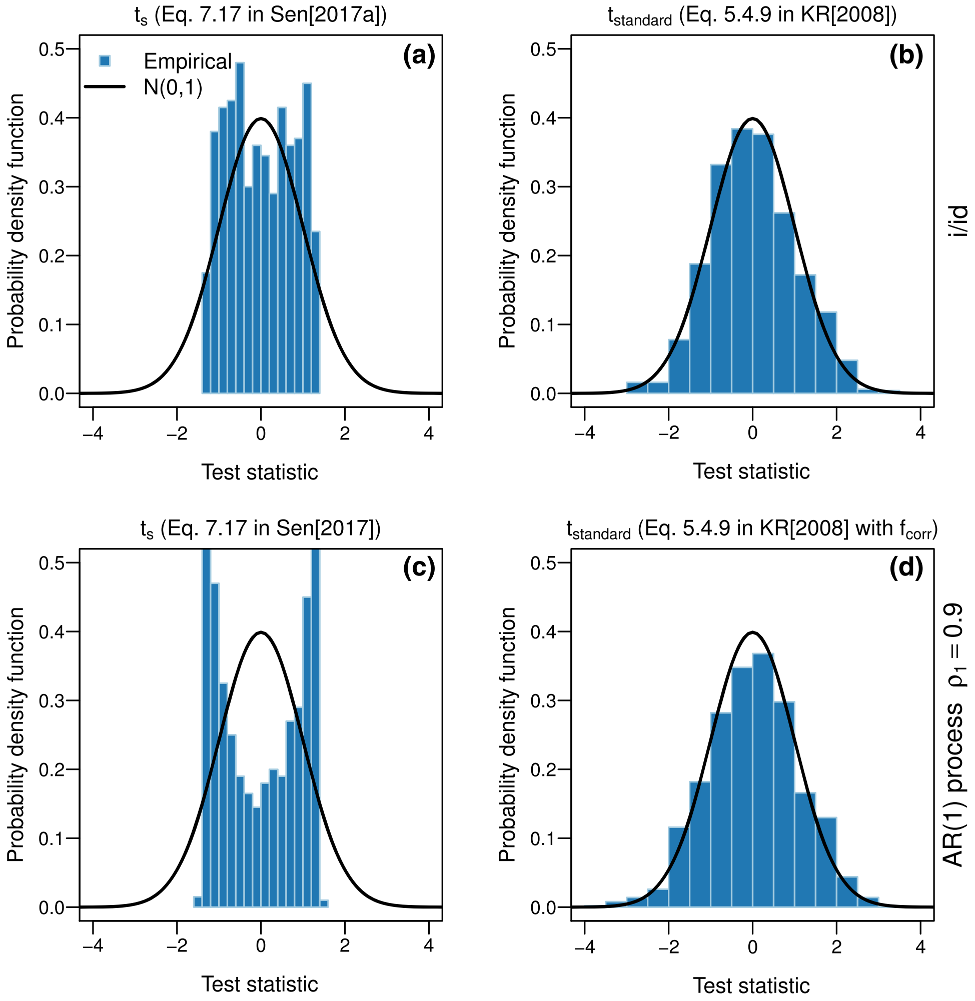

Fig. 9 Sampling distributions of the test statistics t_s and t_{standard} for the *i*/*id* process and AR(1) process with $\rho_1 = 0.9$. KR[2008] = Kottegoda and Rosso (2008)

a set of order statistics does not necessarily converge to a Gaussian distribution and this hypothetical distribution does not describe the uncertainty of each order statistic, meaning that we cannot define a unique CI with constant width for all data points reported in ITA plots. A proper Monte Carlo simulation reported in "Appendix 5" can provide a visual assessment of these remarks.

This behavior is further illustrated in Fig. 10 where constant-width ITA CIs (at the 95% confidence level) are reported along with true point-wise CIs for order statistics computed by two different methods: (1) from simulated samples, and (2) by using the theoretical distribution $F_{X_{(i)}}$. Both methods yield almost identical CIs summarizing the different degree of uncertainty characterizing extreme and non-extreme order statistics. Especially for skewed distributions (i.e. exponential and Gumbel), the upper tails of the ITA plots might substantially depart from the expected 1:1 line and fall outside Şen's (supposed) CIs. Therefore, splitting the ITA plot in three areas corresponding to low, medium, and high values, and thus studying their alignment with 1:1 line separately, as suggested by Şen (2012), is generally misleading as this suggestion overlooks the different uncertainty affecting central and extreme order statistics related to sample size and shape of the parent distribution.

6 Building on the sand: ITA follow-ups

Taking the correctness of ITA for granted without any independent preliminary check led not only to mechanistic applications of ITA diagnostics but also to attempts of improvement whose outcome should be interpreted according to the foregoing discussion. For example, Güçlü (2018b) suggested the so-called multiple ITA, consisting of splitting the time series of size n in k (= 3,4,...) non-overlapping sub-sets of size $\lfloor n/k \rfloor$, and then applying ITA to subsequent pairs of sub-sets, thus obtaining $k-1$ ITA diagrams (e.g., for $k=3$, there are two diagrams of the pairs of sorted values $(x'_{(i)}, x''_{(i)})$ and $(x''_{(i)}, x'''_{(i)})$, $i = 1, \ldots, \lfloor n/3 \rfloor$). Such a procedure increases the uncertainty of each ITA plot, as the diagrams rely on smaller

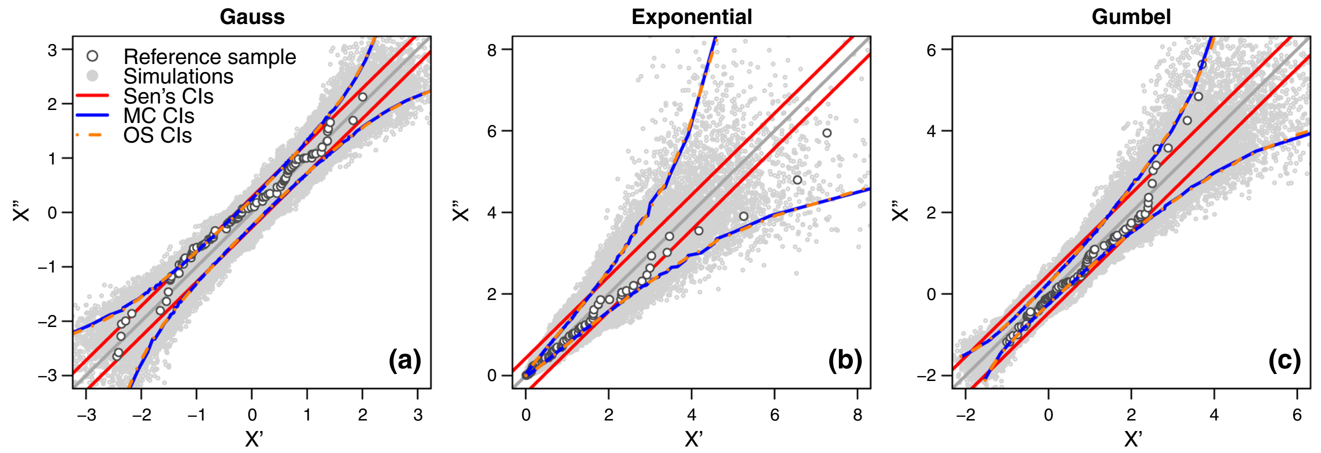

Fig. 10 Comparison of Şen's CIs and true point-wise CIs of order statistics for the three models $\mathcal{N}(0,1)$, $\mathcal{E}(1)$, and $\mathcal{G}(0,1)$. True CIs are computed via Monte Carlo simulation ('MC CIs') and theoretical formulas ('OS CIs') at the 95% confidence level. CI = confidence interval

samples. Moreover, this segmentation has a two-fold negative effect: (1) it can conceal possible serial correlation, which is already under-represented for instance in the usually short hydro-climate time series (Serinaldi and Kilsby 2016b; Iliopoulou and Koutsoyiannis 2019); and (2) it emphasizes spurious trends resulting from (concealed) serial correlation, thus leading to incorrect conclusions about the presence of deterministic trends. Generally speaking, focusing on small subsets always reveals some trend since the straight lines usually fitted to time series have never zero slope; however, such trends are statistically and physically less and less significant because they rely on a smaller and smaller amount of information.

McCuen (2018) explored the problem of statistical significance investigating how much deviation from the 1:1 line can be expected because of sampling uncertainty. McCuen's approach consists of testing the slope of the zero-intercept regression line fitted to the ITA plot, i.e. $\mathbf{x}''_{(\bullet)} = \beta_{\mathrm{M}}\mathbf{x}'_{(\bullet)}$, where $\mathbf{x}'_{(\bullet)} = \{x'_{(i)}\}_{i=1}^{n/2}$ and $\mathbf{x}''_{(\bullet)} = \{x''_{(i)}\}_{i=1}^{n/2}$. Using Monte Carlo simulations, McCuen (2018) computed the critical values of β_{M} under the null hypothesis $H_0 : \beta_{\mathrm{M}} = 1$ (and *i/id* and $X \sim \mathcal{N}(0,1)$), and concluded that these critical values (obtained for a Gaussian distribution) hold approximately true for uniformly distributed data but not for data following an exponential distribution. These results are expected if we recognize the identity of ITA plots and two-sample q–q plots, and recall their properties discussed in Sect. 2. Firstly, $\beta_{\mathrm{M}} \gtrless 1$ does not necessarily correspond to changes/trends but can be related to fluctuations in the second moment (or scale parameter), i.e. possible heteroskedasticity [see Sect. 2, Fig. 1, and examples in D'Agostino and Stephens (1986, pp. 24–57)]. Secondly, McCuen's critical values do not hold for the exponential distribution because this distribution is right skewed and lower bounded to zero. Therefore, under *i/id* (no trends), the lower part of the ITA plots always converges to zero (see Fig. 3), the bundle of ITA plots resulting from sampling uncertainty has a fan shape, and each ITA plot is generally well fitted by zero-intercept regression line with $\beta_{\mathrm{M}} \neq 1$. In other words, for the exponential distribution, β_{M} estimates are almost always different from the unity even if data are *i/id*, and this does not depend on trends but on the shape of the distribution. Similar remarks hold for other skewed families. Parallelism with the 1:1 line under sampling uncertainty holds approximately only for symmetric distributions such as the uniform or Gaussian (see e.g. Fig. 2) mentioned by McCuen (2018). Therefore, a closer preliminary consideration of the nature and meaning of ITA plots reveals that testing the slope of a zero-intercept regression line is not an optimal strategy to obtain a general purpose test identifying deviations from *i/id* (in terms of step changes and/or trend in the mean levels) via ITA plots.

Similar remarks hold for other works as well. For example, Şen (2017a) used ITA to analyze time series preprocessed by the so-called over-whitening procedure, without accounting for the effect of sample size and distribution shape on ITA diagrams. In other cases, the term ITA was used even if the methodology is weakly if not related to ITA construction. For example, Şen et al. (2019) proposed the so-called 'Innovative Polygon Trend Analysis' (IPTA) that is based on a diagram plotting the summary statistics of the two halves of the twelve monthly series $\{x_{j,i}\}$, $j = 1,\ldots,12$ and $i = 1,\ldots,n$. For instance, focusing on the mean values m, IPTA diagrams report the twelve points m'_j versus m''_j, where m'_j and m''_j denote the average values of the first and second half of the observed monthly series. In this case, according to Şen et al. (2019), the presence of a possible trend for a specific month should be based on a single point in IPTA diagrams, which however provide only a visualization of the differences $\Delta m_j = (m_j'' - m_j')$ and do not add any additional information compared with Δm_j. On the other hand, as for the original ITA diagrams, IPTA interpretation overlooks the effect of sample size, distribution shape, and serial correlation on the sample differences Δm_j and their departures from the expected value zero (under 'no trend' assumption). We also stress that Δm_j are routinely analyzed by existing standard tests for the difference between two means discussed in Sect. 4 and 5.1, accounting for the above mentioned factors as well as the additional effect of multiple testing (e.g. Katz and Brown 1991; Wilks 2006).

These studies show the possible negative consequences of taking the validity of new techniques for granted without performing a necessary assessment against benchmark and/or challenging conditions. Especially when new methods promise paramount results under minimal or no assumptions, these techniques should be carefully validated/falsified against the supposed conditions that they should be independent of, and these *neutral* validation studies should be performed by independent experts (other than the developers) to avoid biases in favor of the new methods (Boulesteix et al. 2018). Moreover, the seemingly widespread difficulty to distinguish names and their meaning (Klemeš 1986), and thus recognizing that different names refer to the same (often known) concept, exacerbates the proliferation of questionable methods.

7 Conclusions

When dealing with observations of complex hydro-climatic processes, whose dynamics are not fully known, statistical techniques play a key role to retrieve and summarize information, and enable analysis and prediction (Cramér 1946, pp. 146–148) (see also Shmueli 2010). They are often the only feasible approach to get insights, and therefore are often abused and misused as well. Even though the problem of misusing statistics is not new and is widely documented in the applied statistical literature, it is exacerbated when supposed 'innovative' techniques are developed overlooking basic literature, elementary principles of statistical inference, and necessary careful checks under a reasonable spectrum of different (and possibly challenging) controlled conditions.

The lack of independent validation is mainly due to the fact that the so-called *neutral* comparison and validation studies may be time consuming and difficult to both organize and perform (Boulesteix et al. 2018). They also require the involvement of authors with enough experience, and are often more difficult to publish as "*most high-ranking statistical journals mainly focus on the development of new methods and on innovative applications... As a consequence of the lack of comparison studies, end-users' decisions for or against application of particular methods are often consciously or subconsciously driven by arguments that are to some extent independent of the performance of the method, such as the charisma and marketing strategy of its developers, its use in similar previous studies, the method's fancy name that is easy to remember when heard at a conference, or the availability of user-friendly software*" (Boulesteix et al. 2018). In this study, we used Şen's ITA as a paradigmatic example (among many others) involving all these concerns, and performed a *neutral* validation study to independently check theoretical basis, methodological aspects, mathematical formulation, and consequent interpretation of ITA diagnostic diagrams and formal tests for trend detection.

Referring to the main text for the detailed discussion of the results of our inquiry, we showed that this method

- Cannot discriminate between deterministic trends and spurious trends resulting for instance from serial dependence, when it is applied to finite-size samples (i.e. in real-world applications);
- Overlooks the existing literature, thus neglecting the equivalence of ITA plots and well-known two-sample q–q plots and their intepretation;
- Is characterized by extensive mathematical inconsistencies affecting the formulation of ITA statistical tests. Once these theoretical inconsistencies are corrected, ITA tests are equivalent to well-known classical parametric tests for the difference between two means reported in standard handbooks of applied statistics;
- Contradicts the basic principles of statistical inference, as it is supposed to be free from any assumption while its finite-sample properties strongly depend on sample size and characteristics of the underlying data generating process, such as the shape of the marginal distributions and particularly the autocorrelation.

Overall, ITA suffers from a number of theoretical inconsistencies affecting its derivation, formulas and interpretation. Thus, this study shows the importance of avoiding mechanistic application of new methods taking them for granted, and performing *neutral* validation/falsification analysis to recognize possible methodological problems affecting new methodologies. Therefore, we recommend to reconsider ITA tools (once corrected for mathematical inconsistencies) in light of their equivalence to existing techniques, thus recognizing their actual purpose, correct interpretation, advantages, disadvantages, and limits. As for the TFPW method mentioned in the introduction, empirical results obtained by ITA should be called into question and double checked.

As a more general recommendation to end-users, we suggest bearing in mind the very general principles of statistical inference and mathematical modeling well synthesized for instance by Cramér (1946), Aitken (1947), von Storch and Zwiers (2003), Papoulis (1991), Morrison (2008) and Shmueli (2010) and summarized in "Appendix 1". We also suggest carefully checking every new technique before using it to study real-world data. Such a cautionary approach can help avoiding misleading conclusions, which are often used as a support decision making, thus causing (costly) errors in design and planning.

Acknowledgements Francesco Serinaldi and Chris G. Kilsby acknowledge the support from the Willis Research Network. The authors thank four anonymous reviewers for their critical remarks that helped improve mathematical notation and overall presentation. The analyses were performed in R (R Development Core Team 2019).

Appendix 1

Proposing methodologies that should be model-free, applicable with no assumptions, and unaffected by the sample size, and thus uncertainty-free, contradicts the basic principles of statistical science. Aitken (1947, pp. 2–3) well summarized such principles recalling that every science relies on three main stages:

1. Examination of data collected in a particular field of inquiry to disclose elements of regularity suggesting a law or laws. This is the stage of *inductive synthesis* (see also Cramér 1946, pp. 141–144).
2. Expression of these laws, if possible, in the form of logical axioms such as those characterizing the Euclidean geometry or Newtonian mechanics. This is the stage of *deductive synthesis* and relies on the methods of logic and mathematics, which are used to develop the consequences of the axioms, producing an ensemble of theorems or propositions. In statistics, this *pure* branch consists of the framework provided by probability and statistical mathematics. When the discrepancies between theory and facts are too great to be explained in some way, observations invalidate the applicability of the axioms, and a new set of axioms should be found for the description and explanation of the investigated phenomena. However, "*these axioms and the deductions based on them would still have an abstract validity, as a logical structure of propositions exempt from self-contradiction*" (see also Cramér 1946, pp. 145–146).
3. Interpretation of the abstract functions, equations, constants, etc., "*which occur in the pure formulation, as measures and measurable relations of actual phenomena. This interpretative stage constitutes the applied branch of the science*" (see also Cramér 1946, pp. 146–148).

The foregoing principles are fully general and well-known in applied disciplines as well. Specializing them in the statistical context, Papoulis (1991, p. 4) stresses that "*In the application of probability to real problems, the following steps must be clearly distinguished*

1. *Step 1 (physical) We determine by an inexact process the probabilities* $\mathbb{P}[\mathcal{A}_i]$ *of certain events* $\mathcal{A}_i$*...*
2. *Step 2 (conceptual) We assume that probabilities satisfy certain axioms, and by deductive reasoning we determine from the probabilities* $\mathbb{P}[\mathcal{A}_i]$ *of certain events* $\mathcal{A}_i$ *the probabilities* $\mathbb{P}[\mathcal{B}_i]$ *of certain events* $\mathcal{B}_i$*...*
3. *Step 3 (physical) We make a physical prediction based on the numbers* $\mathbb{P}[\mathcal{B}_i]$ *so obtained*".

Likewise, in the context of modeling dynamic systems, Morrison (2008, p. 7) states "*The next hurdle* [to get over in undergraduate mathematics] *is the differences among observed reality, mathematical models, and computational realizations of mathematical models. Even a lot of accomplished scientists are not clear on these points... learning to cope with three things makes up the basics of a liberal scientific education: facts, abstractions, and the comparison of facts with abstractions... Understanding and ultimately research occurs only when facts are reduced to abstraction, the abstractions manipulated to make predictions, and the prediction compared with new facts*". The practical use of a mathematical theory is not restricted to prediction but includes description and analysis (Shmueli 2010). In particular, concerning the descriptive purposes, "*a large set of empirical data may, with the aid of the theory, be reduced to a relatively small number of of characteristics which represent, in a condensed form, the relevant information supplied by the data*" (Cramér 1946, p. 147). Therefore, every statistical analysis, including descriptive statistics, relies on a mathematical theory with its axioms, assumptions and theorems. For example, while we can numerically compute the sample mean for whatever sequence of real numbers, it represents an estimate of a corresponding population mean only under the assumption that the observations come from a sequence of identically distributed random variables, i.e. for stationary and ergodic random processes. As every statistical analysis relies on some assumptions, this explains why there cannot be diagnostic plots or supposed innovative methods that are assumption-free. This is also well known in statistics applied to climatology. von Storch and Zwiers (2003, p. 69) state:

1. "*A statistical model is adopted that supposedly describes both the stochastic characteristics of the observed process and the properties of the method of observation. It is important to be aware of the models implicit in the chosen statistical method and the constraints those models necessarily impose on the extraction and interpretation of information.*"
2. "*The observations are analysed in the context of the adopted statistical model.*"

Sometimes, some assumptions can be relaxed. For example, "*non-parametric approaches to statistical inference are distinguished from parametric methods in that the distributional assumption is replaced by something more general. For example, instead of assuming that data come from a distribution having a specific form, such as the normal distribution, it might be assumed that the distribution is unimodal and symmetric.*" (von Storch and Zwiers 2003, p. 76). However, "*While they allow us to relax the distributional assumption needed for parametric statistical inference, these procedures rely more heavily upon the sampling assumptions than do parametric procedures*" (von Storch and Zwiers 2003, p. 76). In other words, statistical inference does not allow for 'free lunches' and what we gain in terms of flexibility by relaxing some assumption is paid in terms of power of discriminating among different options. It follows that the primary inherent conceptual flaw of ITA is to present it as something which is presumed to be valid albeit it clearly contradicts basic scientific principles.

Appendix 2

For the model M2, we have

$$\mathrm{E}[x_t] = \rho_1 \mathrm{E}[x_{t-1}] + \mathrm{E}[\beta t] + \mathrm{E}[\varepsilon_t] \tag{12}$$

$$\mu_t = \rho_1 \mu_{t-1} + \beta t + 0. \tag{13}$$

Since we can often assume $\mu_{t-1} \cong \mu_t$ for $\beta \ll 1$, it follows

$$\mu_t \cong \frac{\beta t}{1-\rho_1}. \tag{14}$$

Therefore

$$\begin{aligned}\Delta\mu = & \mu'' - \mu' = \mu_{\frac{3n}{4}} - \mu_{\frac{n}{4}} \cong \frac{\beta\left(\frac{3n}{4} - \frac{n}{4}\right)}{1-\rho_1} \\ & = \frac{\beta\frac{n}{2}}{1-\rho_1} = \frac{\beta n'}{1-\rho_1}.\end{aligned} \tag{15}$$

Appendix 3

To show the correlation between the means in two samples, we simulated 5000 time series with size $n = 100$ from three different processes: (1) *i/id* with standard Gaussian distribution, i.e. $y_t = \epsilon_t$ with $\epsilon \sim \mathcal{N}(0,1)$, (2) *i/id* process with superimposed linear trend ($y_t = \beta t + \epsilon_t$ with $\beta = 0.1$ and $\epsilon \sim \mathcal{N}(0,1)$), and (3) a discrete-time AR(1) process with parameter $\rho_1 = 0.95$ ($y_t = \rho_1 y_{t-1} + \epsilon_t$ with $\epsilon \sim \mathcal{N}(0,1)$). We computed the means of the two halves of each time series, thus obtaining 5000 pairs (m', m'') and drew the scatter plots of these estimates (Fig. 11). These diagrams describe the empirical joint density functions of m' and m'' and show that such sample means are uncorrelated for the first two models ($\rho = 0.02$ and -0.03, respectively), while relatively weak correlation ($\rho = 0.28$) emerges only for highly correlated AR(1) processes (ρ

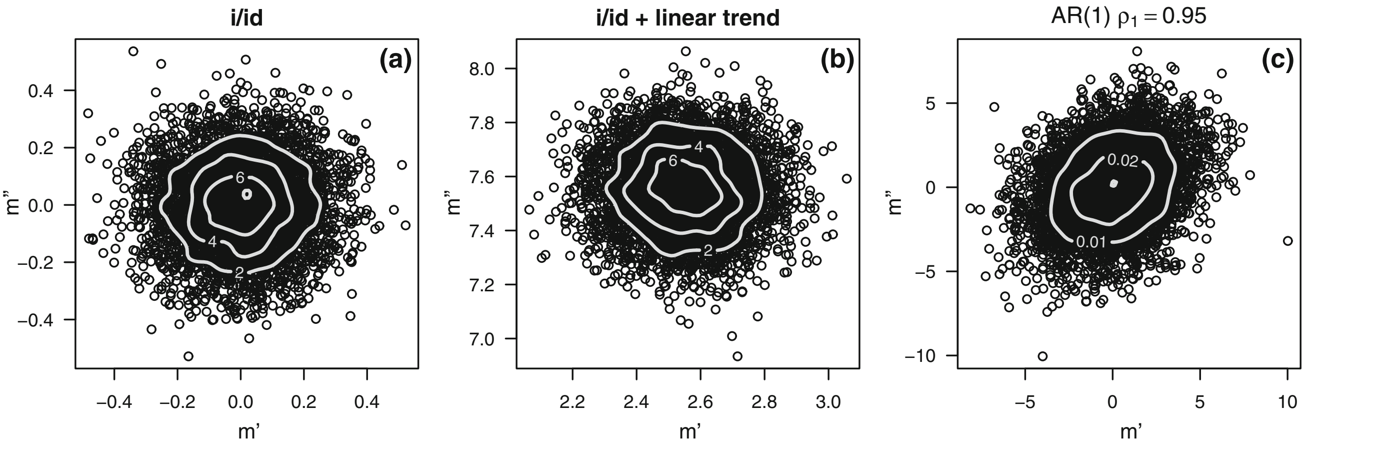

Fig. 11 Sampling joint distribution of m' and m'' for the *i/id* process (**a**), *i/id* with superimposed linear trend (**b**), and AR(1) process with $\rho_1 = 0.95$ (**c**)

assumes values close to zero for $\rho_1 \le 0.8$, which still denotes remarkable serial correlation). The marginal distributions of m' and m'' are close to Gaussian (not shown) with standard deviations close to the expected theoretical values, namely $0.14 \cong \sqrt{\frac{2}{n}}\sigma$ for the first two models and $2.25 \cong \sqrt{\frac{2}{n(1-\rho_1^2)}\left[1+\left(\frac{n}{2}-1\right)\bar{\rho}\right]}\sigma$ for the AR(1) model, where $\sigma = 1$ and $\bar{\rho}$ is the average of the off-diagonal elements of the correlation matrix of $n/2$ variables, i.e. $\bar{\rho} = \frac{\sum\sum_{j\neq l}\rho_{ij}}{\frac{n}{2}\left(\frac{n}{2}-1\right)}$, in which $\rho_{ij} = \mathrm{Corr}\left[X_i, X_j\right]$ denotes the pairwise correlation of X_i and X_j (Matalas and Langbein 1962).

Appendix 4

In order to show the effect of switching $\rho_{m'm''}$ and $\rho_{\mathbf{x}'_{(i)}\mathbf{x}''_{(i)}}$ on the significance of the two-mean tests, we simulated 1000 time series of size $n \in \{20, 40, 60, 80, 100, 150, 200, 250\}$ from two processes: (1) an AR(1) process with ρ_1 ranging between 0 and 0.9 by 0.1 steps, and (2) a fractional Gaussian noise (fGn) with Hurst parameter $H \in \{0.5, 0.55, \ldots, 0.95\}$. The time series are kept trend-free to check the effect of the autocorrelation of the parent processes on the effective rejection rate of the tests (applied at the 5% nominal significance level). Note that the AR(1) and fGn processes with $\rho_1 = 0$ and $H = 0.5$, respectively, yield the *i*/*id* process, for which the effective significance level is expected to be equal to the nominal level. We compared the Şen test with the standard tests for two means with known or unknown variances. Figure 12 shows that the Şen test always yields a rejection rate greater than the 40% for both processes, every degree of serial correlation (including the *i*/*id* case), and every sample size. On the other hand, the two standard tests yield effective significance levels that are close the nominal level (5%) under *i*/*id* (as expected) and gradually increase because of the variance inflation effect of the increasing serial correlation. These conclusions further stress the importance of performing suitable Monte Carlo analysis and comparisons with other available tests when a new test is proposed, in order to check its properties under the null hypothesis and the effect of assumptions and factors such as serial dependence and sample size.

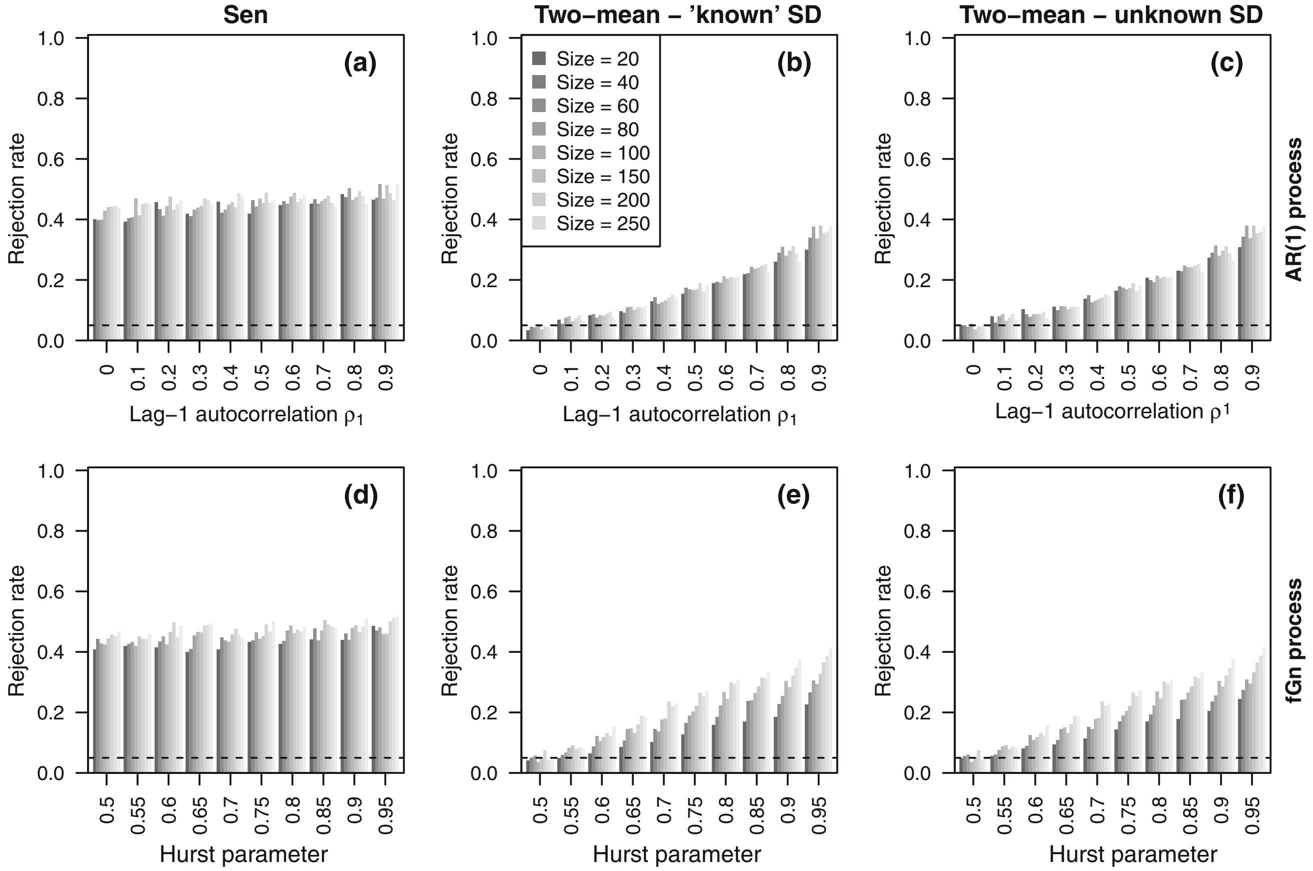

Fig. 12 Effective significance of three formulations of the hypothesis test for the difference between two means (Şen's version and standard test with known or unknown standard deviation (SD)). The effect of the serial correlation is shown for AR(1) and fGn processes

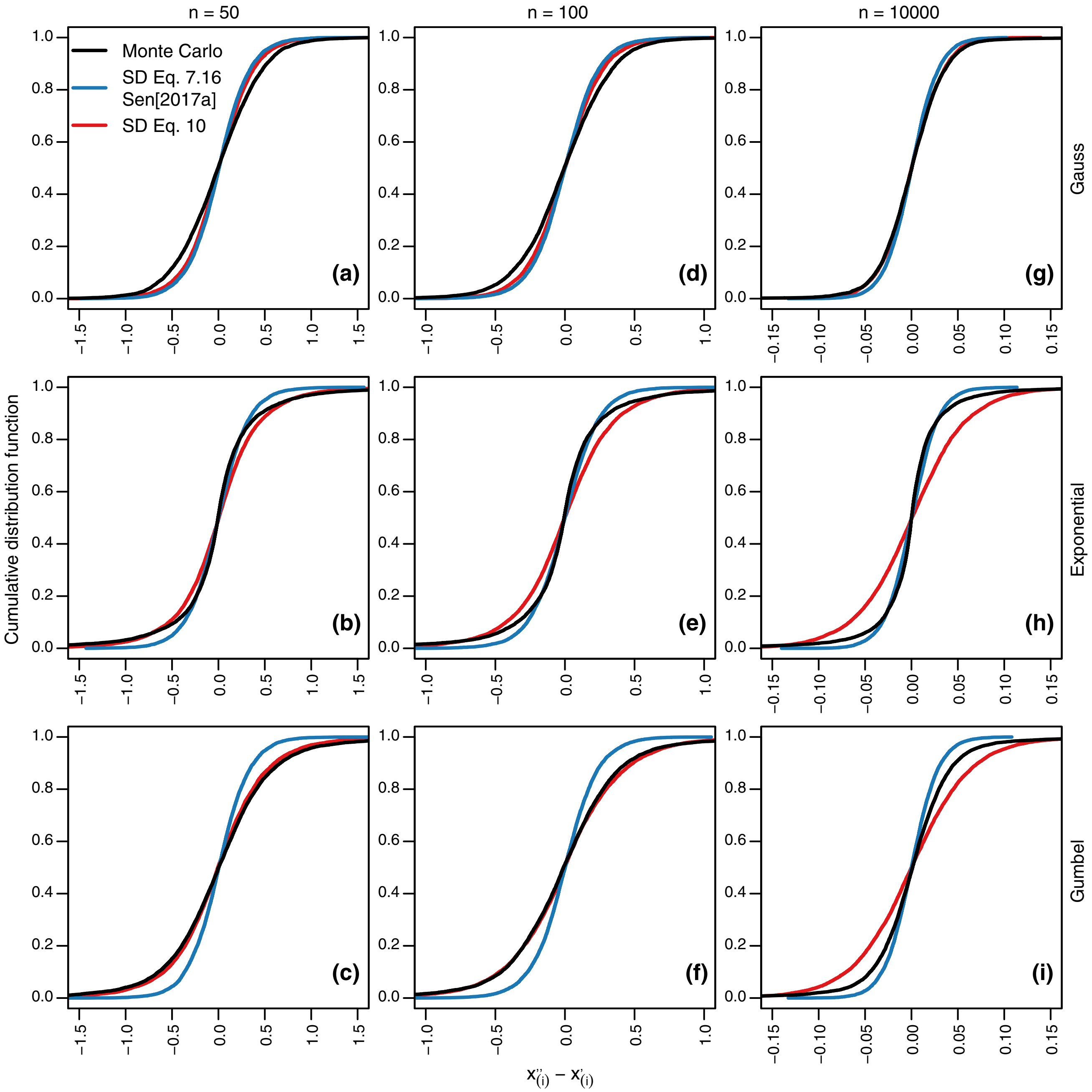

Fig. 13 Sampling distribution of $d_i = x''_{(i)} - x'_{(i)}$ and corresponding δ_i for three different distributions ($\mathcal{N}(0,1)$, $\mathcal{E}(1)$, and $\mathcal{G}(0,1)$) and $n \in \{50, 100, 10,000\}$. Sampling distributions are obtained by empirical distribution from Monte Carlo simulation and from Gaussian distributions with zero mean and standard deviation as for Eq. (8) [i.e. equation 7.16 in Şen (2017b)] and Eq. (10). See text for further details

Appendix 5

We simulated samples of size $n \in \{50, 100, 10,000\}$ from three different distributions (Gaussian, Exponential, and Gumbel). The smaller sample sizes ($n \in \{50, 100\}$) cover the typical range of hydro-climatic observations such as annual means or maxima, while $n = 10,000$ was chosen to check results for relatively large samples. For each sample, we selected $\mathbf{x}'$ and $\mathbf{x}''$ and then we computed the differences between the order statistics, $d_i = x''_{(i)} - x'_{(i)}$, and their variance by using Şen's equation and the standard formula for the variance (which is corrected for the errors mentioned in Sect. 5.1). The variances are used to compute the Gaussian quantiles, δ_i, corresponding to the empirical frequencies i/v, under the assumption that d_i values follow a Gaussian distribution. This experiment was repeated $q =$

1000 times, and the 1000 sequences of d_i (δ_i) were merged obtaining samples of size $q \cdot v$. Such samples allow for building and comparing the sampling distributions of d_i and δ_i. Figure 13 shows that the distributions of δ_i are generally different from those of d_i irrespective of the formula used to compute the sample variances of d_i. As expected, discrepancies depend on the shape of the parent parent distribution of X and they are smaller when F_X is Gaussian. The main reason of these discrepancies is that the distribution of d_i is not unique but depends on the rank of the order statistics. In fact, as already shown in Fig. 2, the uncertainty of ITA plots is generally smaller for central order statistics and larger on the (unbounded) tails.

References

Aitken AC (1947) Statistical mathematics, 5th edn. Oliver and Boyd Interscience Publishers, New York

Alashan S (2018) Data analysis in nonstationary state. Water Resour Manag 32(7):2277–2286

Basarin B, Lukić T, Pavić D, Wilby RL (2016) Trends and multi-annual variability of water temperatures in the river Danube, Serbia. Hydrol Process 30(18):3315–3329

Bayazit M (2015) Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review. Environ Process 2(3):527–542

Bennett ND, Croke BFW, Guariso G, Guillaume JHA, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LTH, Norton JP, Perrin C, Pierce SA, Robson B, Seppelt R, Voinov AA, Fath BD, Andreassian V (2013) Characterising performance of environmental models. Environ Modell Softw 40:1–20

Boulesteix A, Binder H, Abrahamowicz M, Sauerbrei W, for the Simulation Panel of the STRATOS Initiative (2018) On the necessity and design of studies comparing statistical methods. Biom J 60(1):216–218

Caloiero T (2018) SPI trend analysis of New Zealand applying the ITA technique. Geosciences 8(3):101

Caloiero T, Coscarelli R, Ferrari E (2018) Application of the innovative trend analysis method for the trend analysis of rainfall anomalies in Southern Italy. Water Resour Manag 32(15):4971–4983

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman & Hall, London, England

Cramér H (1946) Mathematical methods of statistics. Princeton Landmarks in Mathematics. Princeton University Press, New Jersey, USA

Cui L, Wang L, Lai Z, Tian Q, Liu W, Li J (2017) Innovative trend analysis of annual and seasonal air temperature and rainfall in the Yangtze River Basin, China during 1960–2015. J Atmosph Solar Terr Phys 164:48–59

D'Agostino RB, Stephens MA (eds) (1986) Goodness-of-fit techniques. Marcel Dekker Inc, New York

Flueck JA, Brown TJ (1993) Criteria and methods for performing and evaluating solar-weather studies. J Clim 6(2):373–385

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31(4):337–350

Güçlü YS (2018a) Alternative trend analysis: half time series methodology. Water Resour Manag 32(7):2489–2504

Güçlü YS (2018b) Multiple Şen-innovative trend analyses and partial Mann–Kendall test. J Hydrol 566:685–704

Hamed KH (2009) Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. J Hydrol 368(1–4):143–155

Hutson AD (1999) Calculating nonparametric confidence intervals for quantiles using fractional order statistics. J Appl Stat 26:343–353

Iliopoulou T, Koutsoyiannis D (2019) Revealing hidden persistence in maximum rainfall records. Hydrol Sci J 64(14):1673–1689

Katz RW, Brown BG (1991) The problem of multiplicity in research on teleconnections. Int J Climatol 11(5):505–513

Khaliq MN, Ouarda TBMJ, 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. J Hydrol 368(1–4):117–130

Klemeš V (1986) Dilettantism in hydrology: transition or destiny? Water Resour Res 22(9S):177S–188S

Kottegoda NT, Rosso R (2008) Applied statistics for civil and environmental engineers, 2nd edn. Wiley-Blackwell, Chichester

Kumar S, Merwade V, Kam J, Thurner K (2009) Streamflow trends in Indiana: effects of long term persistence, precipitation and subsurface drains. J Hydrol 374(1–2):171–183

Li J, Wu W, Ye X, Jiang H, Gan R, Wu H, He J, Jiang Y (2019) Innovative trend analysis of main agriculture natural hazards in china during 1989–2014. Nat Hazards 95(3):677–720

Matalas NC, Langbein WB (1962) Information content of the mean. J Geophys Res (1896-1977) 67(9):3441–3448

McCuen RH (2018) Critical values for Şen's trend analysis. J Hydrol Eng 23(11):06018-005

Mood AMF, Graybill FA, Boes DC (1974) Introduction to the theory of statistics, 3rd edn. McGraw-Hill, New York

Morbidelli R, Saltalippi C, Flammini A, Corradini C, Wilkinson SM, Fowler HJ (2018) Influence of temporal data aggregation on trend estimation for intense rainfall. Adv Water Resour 122:304–316

Morrison F (2008) The art of modeling dynamic systems: forecasting for chaos, randomness and determinism. Dover Books on Computer Science Series. Dover Publications Incorporated, Mineola

Nadarajah S, Gupta AK (2004) Characterizations of the Beta distribution. Commun Stat Theory Methods 33:2941–2957

Papoulis A (1991) Probability, random variables, and stochastic processes. McGraw-Hill, New York

Pathak P, Kalra A, Ahmad S, Bernardez M (2016) Wavelet-aided analysis to estimate seasonal variability and dominant periodicities in temperature, precipitation, and streamflow in the midwestern United States. Water Resour Manag 30(13):4649–4665

R Development Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

Sagarika S, Kalra A, Ahmad S (2014) Evaluating the effect of persistence on long-term trends and analyzing step changes in streamflows of the continental United States. J Hydrol 517:36–53

Şen Z (2012) Innovative trend analysis methodology. J Hydrol Eng 17(9):1042–1046

Şen Z (2014) Trend identification simulation and application. J Hydrol Eng 19(3):635–642

Şen Z (2017a) Hydrological trend analysis with innovative and over-whitening procedures. Hydrol Sci J 62(2):294–305

Şen Z (2017b) Innovative trend methodologies in science and engineering. Springer, Cham

Şen Z (2017c) Innovative trend significance test and applications. Theoret Appl Climatol 127(3):939–947

Şen Z, Şişman E, Dabanli I (2019) Innovative Polygon Trend Analysis (IPTA) and applications. J Hydrol 575:202–210

Serinaldi F (2009) Assessing the applicability of fractional order statistics for computing confidence intervals for extreme quantiles. J Hydrol 376(3–4):528–541

Serinaldi F, Kilsby CG (2016a) The importance of prewhitening in change point analysis under persistence. Stoch Environ Res Risk Assess 30(2):763–777

Serinaldi F, Kilsby CG (2016b) Understanding persistence to avoid underestimation of collective flood risk. Water 8(4):152

Serinaldi F, Kilsby CG, Lombardo F (2018) Untenable nonstationarity: an assessment of the fitness for purpose of trend tests in hydrology. Adv Water Resour 111:132–155

Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310

Stigler SM (1977) Fractional order statistics, with applications. J Am Stat Assoc 72(359):544–550

Tananaev NI, Makarieva OM, Lebedeva LS (2016) Trends in annual and extreme flows in the Lena River basin, Northern Eurasia. Geophys Res Lett 43(20):10,764–10,772

Tosunoglu F, Kisi O (2017) Trend analysis of maximum hydrologic drought variables using Mann–Kendall and Şen's innovative trend method. River Res Appl 33(4):597–610

van Giersbergen NPA (2005) On the effect of deterministic terms on the bias in stable AR models. Econ Lett 89(1):75–82

von Storch H (1999) Misuses of statistical analysis in climate research. In: von Storch H, Navarra A (eds) Analysis of climate variability. Springer, Dordrecht, pp 11–26

von Storch H, Zwiers FW (2003) Statistical analysis in climate research. Cambridge University Press, New York

Wang XL, Swail VR (2001) Changes of extreme wave heights in northern hemisphere oceans and related atmospheric circulation regimes. J Clim 14(10):2204–2221

Wasserstein RL, Lazar NA (2016) The ASA statement on p-values: context, process, and purpose. Am Stat 70(2):129–133

Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond $p < 0.05$. Am Stat 73(1):1–19

Wilk MB, Gnanadesikan R (1968) Probability plotting methods for the analysis of data. Biometrika 55(1):1–17

Wilks DS (2006) On "Field Significance" and the false discovery rate. J Appl Meteorol Climatol 45(9):1181–1189

Wu H, Qian H (2017) Innovative trend analysis of annual and seasonal rainfall and extreme values in Shaanxi, China, since the 1950s. Int J Climatol 37(5):2582–2592

Xiao M, Zhang Q, Singh VP (2017) Spatiotemporal variations of extreme precipitation regimes during 1961–2010 and possible teleconnections with climate indices across China. Int J Climatol 37(1):468–479

Yue S, Wang C (2002) Applicability of prewhitening to eliminate the influence of serial correlation on the Mann–Kendall test. Water Resour Res 38(6):41–47

Yue S, Pilon P, Phinney B, Cavadias G (2002) The influence of autocorrelation on the ability to detect trend in hydrological series. Hydrol Process 16(9):1807–1829

Zhang X, Zwiers FW (2004) Comment on "Applicability of prewhitening to eliminate the influence of serial correlation on the Mann–Kendall test" by Sheng Yue and Chun Yuan Wang. Water Resour Res 40(3):W03–805

Zhang X, Vincent LA, Hogg WD, Niitsoo A (2000) Temperature and precipitation trends in Canada during the 20th century. Atmosp Ocean 38(3):395–429

Zhou Z, Wang L, Lin A, Zhang M, Niu Z (2018) Innovative trend analysis of solar radiation in China during 1962–2015. Renew Energy 119:675–689

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.