Advances in Water Resources 111 (2018) 132–155

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology

Francesco Serinaldi[*,a,b], Chris G. Kilsby[a,b], Federico Lombardo[c]

[a] School of Engineering, Newcastle University, Newcastle Upon Tyne NE1 7RU, UK
[b] Willis Research Network, 51 Lime St., London, EC3M 7DQ, UK
[c] Dipartimento di Ingegneria, Universitá degli Studi Roma Tre, Via Vito Volterra 62, Rome 00146, Italy

ARTICLE INFO

Keywords:
Trend hypothesis tests
Nonstationarity
Ergodicity
Hypothesis test interpretation
Trend attribution
Stream flow trends in the United States

ABSTRACT

The detection and attribution of long-term patterns in hydrological time series have been important research topics for decades. A significant portion of the literature regards such patterns as 'deterministic components' or 'trends' even though the complexity of hydrological systems does not allow easy deterministic explanations and attributions. Consequently, trend estimation techniques have been developed to make and justify statements about tendencies in the historical data, which are often used to predict future events. Testing trend hypothesis on observed time series is widespread in the hydro-meteorological literature mainly due to the interest in detecting consequences of human activities on the hydrological cycle. This analysis usually relies on the application of some null hypothesis significance tests (NHSTs) for slowly-varying and/or abrupt changes, such as Mann-Kendall, Pettitt, or similar, to summary statistics of hydrological time series (e.g., annual averages, maxima, minima, etc.). However, the reliability of this application has seldom been explored in detail. This paper discusses misuse, misinterpretation, and logical flaws of NHST for trends in the analysis of hydrological data from three different points of view: historic-logical, semantic-epistemological, and practical. Based on a review of NHST rationale, and basic statistical definitions of stationarity, nonstationarity, and ergodicity, we show that even if the empirical estimation of trends in hydrological time series is always feasible from a numerical point of view, it is uninformative and does not allow the inference of nonstationarity without assuming a priori additional information on the underlying stochastic process, according to deductive reasoning. This prevents the use of trend NHST outcomes to support nonstationary frequency analysis and modeling. We also show that the correlation structures characterizing hydrological time series might easily be underestimated, further compromising the attempt to draw conclusions about trends spanning the period of records. Moreover, even though adjusting procedures accounting for correlation have been developed, some of them are insufficient or are applied only to some tests, while some others are theoretically flawed but still widely applied. In particular, using 250 unimpacted stream flow time series across the conterminous United States (CONUS), we show that the test results can dramatically change if the sequences of annual values are reproduced starting from daily stream flow records, whose larger sizes enable a more reliable assessment of the correlation structures.

Faced with a sample of unknown origin, many applied statisticians working in economics, meteorology and the like, hasten to decompose it into a trend and an oscillation (and added periodic terms). They assume implicitly that the addends are attributable to distinct generating mechanisms, and are statistically independent. This last implicit assumption is quite unwarranted, except when the sample is generated by Brownian motion. (B.B. Mandelbrot, The Fractal Geometry of Nature, p. 352, 1982)

1. Introduction

Due to the complexity of hydrological systems, their analysis and modeling heavily rely on historical records as theoretical reasoning and deduction are often inadequate. This analysis is even more problematic when we depart from the hypothesis of stationarity to embrace that of nonstationarity. Even though the two notions of stationarity and nonstationarity should apply to models and not to the real-world processes themselves (see Section 4 below), considerable literature assumes that the observed time series generated by the real world seldom appear to

* Corresponding author at: School of Engineering, Newcastle University, Newcastle Upon Tyne NE1 7RU, UK.
E-mail address: francesco.serinaldi@ncl.ac.uk (F. Serinaldi).

https://doi.org/10.1016/j.advwatres.2017.10.015
Received 6 June 2017; Received in revised form 4 October 2017; Accepted 13 October 2017
Available online 18 October 2017

be stationary but exhibit more complicated nonstationary behavior. In many cases, conclusions on nonstationarity are based on the outcome of trend tests applied to finite-size time series covering relatively short periods of record.

A change of paradigm from stationary to nonstationary can be claimed to account for human activities producing predictable changes, such as land-use and land-cover changes, and water resources exploitation, or more complex but less predictable phenomena such as the worldwide hydrologic change ascribed to anthropogenic climate change (ACC) (Milly et al., 2015). In this respect, in the last three decades, a huge number of studies have investigated possible human-driven changes in the form of slowly-varying trends or abrupt changes in time series of hydrological variables across different regions of the world. Broadly speaking and taking for granted unavoidable differences, the aim of these studies has been to understand if these changes are detectable, what is their pattern, and ultimately, to infer nonstationarity, thus promoting the implementation of nonstationary models to support new design and planning strategies (e.g., Ouarda and El-Adlouni, 2011; Rootzén and Katz, 2013; Cheng et al., 2014, among many others).

Therefore, we believe there exists a need for careful inspection of the basic concepts of null hypothesis statistical tests (NHSTs) for trends and their application to hydrological problems. Following Serinaldi and Kilsby (2015), this paper is an attempt to meet this need. In fact, the purpose of our work is neither to review the state of the art of the research related to trend analysis, nor to give examples of the problems discussed thereto. Rather, we summarize and attempt to shed some light on the reasons for contradictory results encountered in the literature, and discuss widespread practices that can easily be identified in many studies. Therefore, the conceptual perspective of this study should be seen as a guideline in agreement with the general but scientifically based and widely applicable statement by Mandelbrot in the opening motto of this paper. On the other hand, we attempt to confute a certain mechanistic approach often characterizing the literature on the topic. We highlight that conceptual arguments and mathematical definitions are necessary to provide practical advice to identify trends, to interpret results, and to avoid misleading usage and conclusions.

The paper is structured as follows. By using a simple example, Section 2 introduces the discussion and research questions, and summarizes our conclusions in order to provide the reader with a clear outline of what will follow. Section 3 reviews the role of trend testing and some problems related to historical derivation, epistemological reasons, and detection and attribution of changes under temporal persistence. Then, in Section 4, we discuss the importance of clear terminology corresponding to well-defined concepts to avoid misunderstandings relying on different interpretation of the same terms. Section 5 gives an overview of the properties of some commonly used trend tests, namely, Mann-Kendall (MK) and Pettitt. In Section 6, we analyze 250 unimpacted stream flow time series across the conterminous United States spanning the period 1950-2011. Discussion and final remarks are given in Section 7.

2. NHST for trends: overview of key ideas

2.1. Setting the scene with a simple example

We start our discussion by a simple example of typical trend testing exercise familiar to practitioners as we want to highlight the basic concepts behind trend testing procedures. Fig. 1(a) and (b) shows the average annual discharge of two nearby rivers that we will call the Nera River and Velino River, referring to the following discussion for an in-depth description of these data. Both time series, ranging from 1916 to 2015, show an apparent change point around the years 1974–1975 as well as statistically significant and similar autocorrelation functions (Fig. 1(c) and (d)); the Kendall correlation coefficient between the two time series is $\tau_K = 0.32$. Since we do not know if the autocorrelation is a consequence of a possible *deterministic* change of regime or the effect of the dependence structure, we use statistical tests accounting for the latter. Therefore, we apply both the classic Pettitt tests and four additional versions accounting for possible first-order Markov autocorrelation structure and fractional Gaussian noise (fGn; also known as Hurst–Kolmogorov process) (Serinaldi and Kilsby, 2016a). Following the common interpretation of trend tests, the tests unanimously lead to the conclusion that a possible deterministic change around 1974–1975 at the 5% significance level. After splitting the series into two sub-series (before and after the change point) we find that their autocorrelation is not significant (Fig. 1(e)–(h)). Therefore, we next apply standard Pettitt and MK tests to the two sub-series for both rivers (Villarini et al., 2009a). Since no significant trend or change point was found in the sub-series, we can conclude that autocorrelation is reasonably the effect of an abrupt change occurred around 1974–1975. Based on these results, the common approach attempts to explain such changes by some anthropogenic activities, including some more easily recognizable (i.e. river training, water abstraction, dam construction, etc.) and some less (i.e., ACC, climate teleconnections, etc.). In the latter case, the attribution is performed by some further statistically-based analysis (see e.g., Merz et al., 2012; Viglione et al., 2016, and references therein, for an overview of the attribution problem and examples).

However, the nature of the time series analyzed above is completely different. In fact, the true nature of these data is that they are nothing more than artificially generated series designed to be a complete contrast. The Nera River sequence is a step-wise signal superimposed on a sample of independent pseudo-random realizations drawn from a Gaussian distribution (Fig. 1(i)), whereas the Velino River sequence is a sub-set extracted from a longer time series of size 2000, which is a realization of a discrete-time fGn with unit variance and Hurst parameter $H = 0.8$ (Fig. 1(j)).

This synthetic experiment highlights that even procedures specifically devised to account for the interplay between possible deterministic trends and/or change points are not able to discriminate, and can easily lead to incorrect conclusions. In fact, persistence generates local trends and abrupt changes, and deterministic changes result in artificial persistence. Notice that sequences similar to Velino time series can easily be extracted by a quick visual inspection of the entire time series, since these local trends and step changes are a characteristic of persistent processes.

Since abrupt changes can be seen as a special (limit) case of monotonic slowly-varying trend, in the following discussion we generally indicate both types with the term ‘trend’, unless otherwise specified. The null hypothesis is defined as {H_0: there is no deterministic trend}, while the alternative hypothesis as {H_1: there is deterministic trend}. Further specifications are given according to the specific context throughout the text.

2.2. Forgotten questions whose answers are often taken for granted

Some research questions arise from the example in the previous section:

1. What is the origin of slowly-varying trends (or step changes) in hydro-meteorological time series? Are they due to external drivers (e.g., well-defined human interventions) or are they related to intrinsic persistence or other causes? Alternatively, is measured persistence a spurious effect of trends induced by external forcing, or are observed trends spurious effects of persistence or other generating mechanisms?
2. Can null hypothesis statistical tests (NHSTs) for trends answer the above questions? Which information can trend NHSTs provide?
3. Under the assumption that trend NHSTs can provide information about trends in recorded series, can one draw conclusions about nonstationarity, thus justifying, for instance, the use of nonstationary modeling in hydrological frequency analysis?

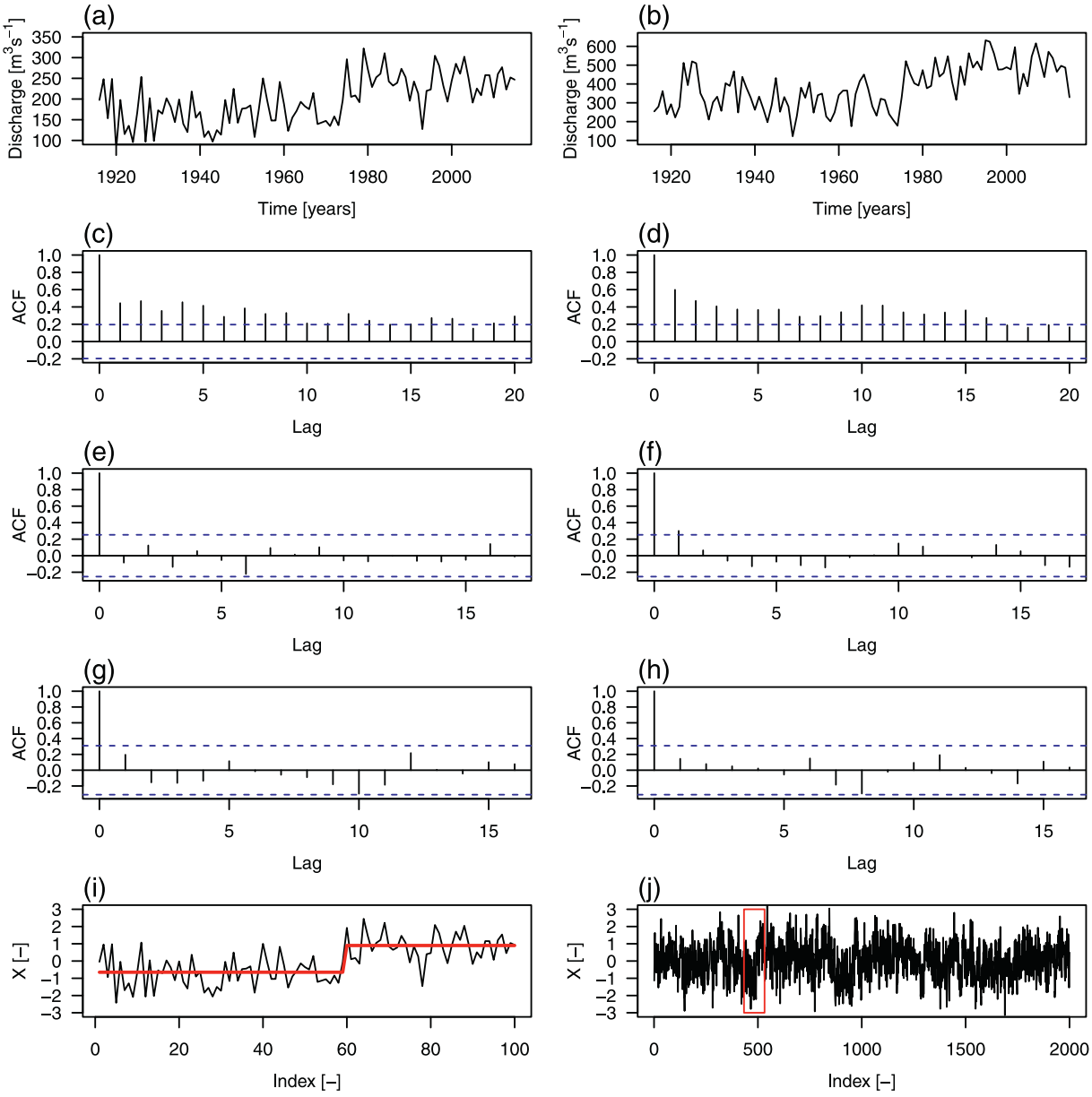

Fig. 1. Panels (a) and (b) respectively depict time series of the average annual discharge of the Nera River and Velino River. In (c) and (d), ACFs of the time series shown in panels (a) and (b). In (e) and (f), ACFs of the first part of the time series shown in panels (a) and (b), before the change point. In (g) and (h), ACFs of the second part of the time series shown in panels (a) and (b), after the change point. Panels (i) and (j) show the original nature of time series plotted in panels (a) and (b), respectively. See text for further details.

2.3. Resetting some beliefs concerning NHST for trends

Since answers to the above questions require an extensive discussion of arguments often neglected in hydro-meteorological applications, in this section we firstly summarize the main conclusions, and then we present in full detail the reasoning leading to them in the remainder of the paper. Broadly speaking, searching for answers to the above questions reveals critical aspects related to trend NHST that can be classified as empirical, methodological, and theoretical. The first refer to the nature of the data, the second to the models used to make inference, and the latter to logical foundations and semantics (i.e. the link between symbol, concept, and referent; Eco (1976)). All these aspects are already known and discussed but spread out in various research areas; however, they are often overlooked, and their impact on results dramatically underestimated in many hydro-meteorological studies.

Following the structure of the subsequent sections, these arguments can be summarized as follows:

1. NHSTs have a logically flawed rationale coming from ill-posed and theoretically unfounded hybridization of Fisher significance tests and Neyman-Pearson hypothesis tests; they do not provide the information that scientists need (i.e., the likelihood of H_0 given the data and/or physical significance), do not allow conclusions about the truth of falsehood of any hypothesis, and do not apply to exploratory non-randomized studies. Trend tests share the general problems of NHST procedures. Such issues are concerned with the inverse probability problem, the confusion between substantive and statistical hypotheses, and the fact that NHSTs are not devised for exploratory studies. In fact, hydro-meteorological data are unique in the sense that every record is the only available realization or

trajectory of the underlying process. Since alternative experiments cannot be performed, these observations do not provide the type of independent information that would be obtained by observing the same variables over a period of similar length at another point in time. Even though hypothesis testing falls in the realm of so-called confirmatory analysis, its nature is basically dissenting as its outcomes can only be rejection or no rejection, and both cases reflect lack of knowledge about the null hypothesis H_0 and the alternative hypothesis H_1: formal acceptance is not a contemplated option. Moreover, statistical significance does not imply physical significance because the former depends on the sample size, and almost every test assigns statistical significance to physically negligible differences for very large samples (see Sections 3.1).

2. Hydro-meteorological data are commonly characterized by spatial and temporal dependence. This property can greatly help to interpret and account for many features of hydro-meteorological records such as apparently unexpected variability. Dependence is usually incorporated into the null hypothesis H_0 in order to compare the assumption H_1 of deterministic trend with a more realistic H_0 relaxing the assumption of independence. Nevertheless, dependence can be strongly underestimated due to the limited extent and uniqueness of the hydro-meteorological data, which should therefore be carefully taken into account. For example, this study highlights that even more refined statistical techniques accounting for dependence can be not enough. In fact, we show that the nature, quantity, and quality of some annual summary statistics are not sufficient to infer the dependence (and thus the variability) resulting from the entire daily process (see Sections 3.2 and 6).
3. Trend tests are widely used to assess the effect of known external forcings (e.g., land cover change) on hydro-meteorological records (e.g., flow peaks) in order to explore inhomogeneity or trends or nonstationarity (e.g., McCuen, 2003). Such procedures often result in circular reasoning because if we assume that the forcing process is changing according to some deterministic function of time - and thus it is nonstationary - and it affects in some way a target process of interest (e.g., flood intensity or frequency), then we already know that the target process is nonstationary. In these cases, we are interested in the size of the effects and not in the presence/absence of generating mechanism, which is already known (see Section 3.3).
4. The outcome of trend NHSTs cannot support and justify the use of nonstationary models. As a deterministic trend is a systematic change reflecting a time-dependent process, the mathematical rule describing the evolution of this change should be established by deductive reasoning (a priori; see e.g., Poppick et al., 2017) or assumed as a working hypothesis but cannot be inferred solely from the data without external information, because, without attribution, new data might easily change the nature and shape of the supposed trend. Therefore trends cannot result for instance from fitting arbitrary parametric curves or applying smoothing filters to observed records. Despite the possible goodness of fit, these pseudo-trends might yield completely unreliable predictions. This lack of reliability reveals the actual nature of such data-driven trends, i.e. that they refer to the time series and not to the underlying process, and thus are affected by sampling uncertainty and can change as additional data become available (Luke et al., 2017; Serinaldi and Kilsby, 2015). Therefore, even though nonstationary modeling is legitimate, every step should be approached with great care in order to be logically and scientifically correct, bearing in mind the underlying assumptions of procedures, methods, and models used in each stage of the analysis. Beside possible ill-posed selection of nonstationary models yielding unreliable predictions, overlooking theoretical assumptions generates misconceptions such as the incorrect belief of the existence of temporally varying return periods and corresponding return levels and their confusion with time varying probabilities of exceedance and quantiles, whereby the mathematical definitions of return period yield unique and comparable values in stationary and nonstationary contexts (Cooley, 2013; Olsen et al., 1998; Salas and Obeysekera, 2014; Serinaldi, 2015; Serinaldi and Kilsby, 2015; Volpi et al., 2015) (see Sections 3.3 and 4.2).
5. Another consequence of the limited extent and uniqueness of the hydro-meteorological data is that one needs to make a number of implicit but strong assumptions in order to treat these records as outcomes of deterministic, stochastic, or some mixed processes. In this respect stationarity and ergodicity play a key role in statistical inference. Ergodic theory deals with the relationship between statistical averages and sample averages, which is a central problem in the estimation of statistical parameters in terms of real data. For example, empirical summary statistics (e.g., moments) are informative only under the assumption that the process is stationary and ergodic. For example, even if the sample mean of an observed time series can always be estimated and does not change irrespective of stationary or nonstationary assumptions, in the first case it is assumed to be representative of the process thanks to ergodicity, while in the latter it is not (if one does not account for the source of nonstationarity). This means that other realizations of the same nonstationary process can have completely different sample averages, none of which can give insight into the actual population mean of the process, if any. Therefore, assuming nonstationarity requires great care in order to understand what we can really infer from data under lack of ergodicity. Without supporting the nonstationary choice with deductive (top-down) arguments identifying the mechanism generating the time-dependent behavior of the process, the modeling procedure reduces to a mechanistic numerical exercise attempting to minimize some performance criterion with the aim to follow local patterns of fitting data sets. As mentioned above, this approach yields models that reveal the weakness of their derivation and justification when predictions are compared with new (future) observations in validation data sets (see Sections 4.1 and 4.2).
6. Nonstationarity is a very stringent assumption as it implies that one or more characteristics of the distribution of a system depend on time by a deterministic function d_t. As the term deterministic implies being free of uncertainty, nonstationarity cannot be claimed from the data only without an attribution identifying the source of the deterministic dependence on time. Therefore, Koutsoyiannis and Montanari (2015) noted that “Because it explains in deterministic terms part of the variability, a nonstationary description is associated with reduced uncertainty. Hence unjustified or inappropriate claim of nonstationarity results in underestimation of variability, uncertainty and risk”. Here, uncertainty does not refer to specific parametrization but to the existence and overall behavior of time-dependent deterministic processes. For example, the existence and general evolution of seasonal behavior is deduced from arguments independent from data (i.e., planetary dynamics), while its parametrization varies for each specific data set. Excluding spurious local trends characterizing stationary stochastic processes, trends of interest in hydro-meteorology are those related to mechanisms generating departures from the so-called natural variability (which is implicitly assumed to be stationary). Such trends are therefore a form of nonstationarity, which implies the existence of a deterministic function of time d_t requiring detection and attribution by combining deductive reasoning, which supports and justifies the existence of d_t, and inductive inference, which provides preliminary knowledge and quantication/parametrization of d_t. The definition of deterministic trend has direct practical consequences (see Sections 4.3 and 4.4):
   (a) The commonly used approach of comparing nested models with time-varying and constant parameters by using some performance criterion is not sufficient to infer nonstationarity if d_t does not result from deductive reasoning, but results from simple fitting procedures. General poor performance in prediction confirms the weakness of such a bottom-up procedure (Serinaldi and Kilsby, 2015).

(b) Replacing the dependence on time (i.e., d_t) with dependence on teleconnection indices or other environmental variables showing clear stochastic behavior does not make models nonstationary, but simply doubly stochastic stationary. This replacement makes models nonstationary only if such auxiliary environmental variables are themselves nonstationary, and thus time-dependent according to a well-defined function d_t. This is particularly important for a correct application and interpretation of frequency analysis based on generalized linear models (GLMs), generalized additive models (GAMs), or similar.

7. Trend NHSTs suffer logical flaws and some of them are also incorrect or incorrectly applied. For example, the still widely applied so-called trend-free prewhitening (TFPW) (Yue et al., 2002) was shown to be theoretically flawed (Serinaldi and Kilsby, 2016a), as its original version does not address the variance inflation related to dependence, which can be even exacerbated. This explains the contradicting results reported in the literature concerning the outcome of this test compared with alternative procedures. The correctness of applied tests (or methodology in general) should not be taken for granted, and a preliminary check of their performance under H_0 (controlled conditions) should be performed (by simulation) before their application, especially if the methodology was not developed by statisticians (see Section 5).

   The choice between stationary and nonstationary depends on a stringent process of attribution supported by deductive arguments, which come before and go beyond statistical inference techniques (see Section 6).

3. The (non)logic of trend hypothesis tests: what they cannot say about trends and nonstationarity

3.1. The consequences of a difficult birth: NHSTs logical flaws and misinterpretations

In several fields of applied science, NHSTs have been widely discussed and criticized for a long time (Cohen, 1994; Gill, 1999; Johnson, 1999; Levine et al., 2008; Beninger et al., 2012; Ellison et al., 2014; Nuzzo, 2014; Briggs, 2016; Greenland et al., 2016; Wasserstein and Lazar, 2016, and references therein), but to our knowledge, the problems concerning NHSTs received little attention in hydrological sciences (McBride et al., 1993; Nicholls, 2001; Clarke, 2010). NHST is a synthesis of the Fisher test of significance, developed as a general approach to scientific inference, and the Neyman-Pearson hypothesis test, designed for applied decision making and quality control (Levine et al., 2008). These methods are conceptually different and imply different interpretations of their outcomes. Neyman and Pearson believed they had made Fisher's theory of significance testing more complete and consistent, whereas Fisher never perceived the emerging Neyman-Pearson theory as correcting and improving his own work on tests of significance (Gigerenzer et al., 1989, pp. 98 and 102). A heated controversy followed, and "although the debate continues among statisticians, it was silently resolved in the 'cookbooks' written in the 1940s to the 1960s, largely by non-statisticians, to teach students in the social sciences the 'rules of statistics' " (Gigerenzer et al., 1989, p. 106). The result was a so-called 'hybrid system', i.e. NHST (Beninger et al., 2012), merging "Fisher's easy-to-calculate P value into Neyman and Pearson's reassuringly rigorous rule-based system" (Nuzzo, 2014). Overlooking the great differences in conceptual interpretation, this seemed perfectly acceptable to statistics end-users, partly because often the same formulas were used and the same numerical results obtained (Gigerenzer et al., 1989, p. 106). This has led to an enormous confusion about, for instance, the meaning of a significance level, coining the well-known expression 'the null hypothesis is rejected at the α level', which occurs neither in Fisher nor in the writings of Neyman and Pearson. Moreover, the neglect of controversial issues and alternative theories, and the anonymous presentation of an apparently monolithic body of statistical techniques often turned the hybrid theory into a mechanical ritual, even though Fisher, and Neyman-Pearson had all warned against drawing inferences from tests without judgment (Gigerenzer et al., 1989, p. 107 and 209). This historical digression confirms how damaging a mechanistic approach can be through overlooking subtle, but fundamental, theoretical concepts.

The differences between Fisher and Neyman-Pearson systems highlight their incompatibility and the problems affecting the NHST synthesis. With Fisher significance testing, no explicit alternative hypothesis H_1 to the null H_0 is identified, and the p-value that results from the *model* and the *data* is evaluated as the strength of the evidence for the research hypothesis. Therefore there is no notion of 'power of test' nor of accepting alternative hypothesis H_1 in the final interpretation. Conversely, Neyman-Pearson tests identify complementary hypotheses, H_0 and H_1, in which rejection of one implies acceptance of the other, and this rejection is based on a predetermined significance level α. Neyman-Pearson hypothesis test defines the significance level α a priori as a *function of the test* (i.e., *before even looking at the data*), whereas Fisher's test of significance defines the significance level afterwards as a *function of the data*. The NHST synthesis pretends to select α a priori, but actually using a posteriori p-values to evaluate the strength of the evidence. This allows inclusion of the alternative hypothesis but removes the search for a more powerful test (Gill, 1999). The power of a test is actually a problematic issue in hybrid NHST as it is most often undefined. The sampling distributions of both H_0 and H_1 are specified in Neyman-Pearson theory and an effect size or point prediction must be specified for H_1 in order for the concept of power to be meaningful and for defining the sample size required to obtain the required power. Conversely, in hybrid NHST, H_1 is simply specified to be not H_0 and vice versa (e.g., {H_0: there is no deterministic trend} and {H_1: there is deterministic trend}), i.e., H_0 and H_1 are written such that they are mutually exclusive and exhaustive (Levine et al., 2008). Moreover, one of the NHST hypotheses is always labeled as the null hypothesis as in the Fisher test, whereas Fisher intended the null hypothesis simply as something to be 'nullified' or falsified in agreement with (and influenced by) the contemporary 1935 Karl Popper's *Logic of Scientific Discovery*. NHST partially uses the Neyman-Pearson decision process except that failing to reject the null hypothesis is treated as a 'modest' support for the null hypothesis (Gill, 1999).

Leaving aside problems related to some abuses and misinterpretations that can be partially corrected, the hybrid NHST suffers some logical flaws that cannot be overcome:

1. *Converse inequality argument or inverse probability problem:* p-values do not and cannot assess the strength of evidence supporting a hypothesis or model. In fact, a p-value is simply the probability of obtaining the result (data or evidence $\mathscr{D}$) if H_0 were true, $P[\mathscr{D}|H_0]$, while the researcher is interested in the probability of the null hypothesis, $P[H_0]$, or the probability of the null hypothesis given the data, $P[H_0|\mathscr{D}]$. Of course, in general, $P[H_0|\mathscr{D}] \neq P[\mathscr{D}|H_0]$, and they are related by the Bayes theorem $P[H_0|\mathscr{D}] = (P[\mathscr{D}|H_0]P[H_0])/P[\mathscr{D}]$. Interpreting p-values as $P[H_0|\mathscr{D}]$ rather than $P[\mathscr{D}|H_0]$ corresponds to switching, for instance, the statements: "(1) Most people who face a firing squad die from bullet wounds, and (2) Most people who die from bullet wounds have received them from a firing squad!" (Beninger et al., 2012). The (flawed) logic of NHST is as follows: (i) if H_0 is true then the data are highly likely to follow an expected pattern, (ii) the data do not follow the expected pattern, (iii) therefore H_0 is highly unlikely. This can translate to statements such as: (i) if a person is an American then it is highly unlikely she/he is a member of Congress, (ii) the person is a member of Congress, (iii) therefore it is highly unlikely she/he is an American. In other words, a low p-value, i.e. $P[Congress|American]$, does not imply a low $P[American|Congress]$ (Pollard and Richardson, 1987; Gill, 1999). Thus, p-value says nothing about the truth or otherwise of H_0 or H_1 or the strength of evidence for or against either one. In this respect,

Neyman and Pearson were very clear "...as far as a particular hypothesis is concerned, no test based on the (*objective*) theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis" (Neyman and Pearson, 1933).

2. *Substantive theories vs. statistical hypotheses*: In hybrid NHST, the statistical null hypothesis and the statistical alternative hypothesis are written such that they are mutually exclusive and collectively exhaustive. Therefore, if we accept the incorrect assumption that one could reject H_0 on the basis of a small p-value, then H_1 is inferred to be probably true since no other alternatives (besides H_0 and H_1) are logically possible (Levine et al., 2008). However, H_1 can result from multiple and conflicting substantive theories. For example, local step changes in Fig. 1(a) and (b) can result from fluctuations of a persistent process or the superposition of uncorrelated random noise and a deterministic stepwise signal. Accepting a substantive theory on the basis of results concerning a statistical hypothesis relies on the formal fallacy of 'affirming the consequent' (i.e., 'If p, then q; q; therefore, p'), which is the form of all scientific inference aimed at supporting a theory by *verifying* its observational consequences. Statistical hypotheses are numerical consequences of the substantive theories, not their semantic equivalents (Meehl, 1997).
   It should be noted that Fisher did not distinguish between substantive hypotheses and statistical hypotheses (Gigerenzer et al., 1989, p. 97). However, the p-value was intended simply as an informal way to judge whether evidence is worthy of a second look (Nuzzo, 2014), and 'rejecting H_0' does not mean a categorical adoption of the belief that it is false. In fact, according to Fisher, "in learning by experience, ... conclusions are always provisional and in the nature of progress reports, interpreting and embodying the evidence so far accrued" (Fisher, 1935, p. 25). On the other hand, Neyman and Pearson introduced their hypothesis test as a 'rule of behavior' to make decisions accounting for possible consequences "without hoping to know whether each separate hypothesis is true or false" (Neyman and Pearson, 1933). In other words, both Fisher and Neyman and Pearson were well aware of the fallacy of 'affirming the consequent' and the impossibility for inductive inference to make conclusions about the truth or falsehood of a scientific hypothesis. Statistical considerations alone cannot lead to a decision.
3. *Classic NHST does not apply to exploratory studies*: Most research studies can be generally classified as either experimental or observational (Flueck and Brown, 1993). The major distinction is that the former requires the ability of the scientist to control the principal inputs in order to assess the effects on the outputs. Therefore, studies of trends in hydro-meteorological variables can be classified as observational because there is no scope for controlling the inputs (e.g., the researcher cannot control the amount of rainfall), thus making such studies more difficult to plan and analyze than experimental ones. Both experimental and observational studies usually have three stages denoted as preliminary, exploratory, and confirmatory, even if the third stage can or should actually aim to falsify/disprove the scientific hypothesis, according to the so-called 'modus tollens' of deductive inference (i.e., 'If p, then q; no q; therefore, no p') (Meehl, 1997). Leaving aside the preliminary stage concerning general insights into questions about the research topic (e.g., which measurements are useful and can be made, amount of available or collectable data in the study period, etc.), exploratory studies aim to define claims about foreseen or unforeseen relations on the basis of a plausible conceptual model (i.e., a researcher's description of the process of interest) and appropriate scientific evidence, whereas confirmatory studies are specifically defined processes focused on replicating or disproving a result while minimizing sampling and non-sampling errors (Flueck and Brown, 1993) with small probability that results can come from causes different from the tested theory.
   Exploratory studies are flexible in their research of evidence (e.g., variables to be included), but this flexibility should not be confused with a superficial treatment of the data and methods. Focusing on data, analyses can rely on randomized or non-randomized samples. NHST requires randomized samples as it involves three steps often overlooked but fundamental: (i) the choice of the probabilities of occurrence, α and β, of Type I and Type II errors (not only the significance level α); (ii) *random selection* of only n samples from the designed population, whereby n is related to the sampling distribution of the test statistic, α, and β, in order to guarantee the desired test's significance and power; and (iii) the test must be performed only once. All these steps should be performed *before collecting data*. Therefore, logical flaws apart, NHST yields 'valid' results only if these steps are followed, thus justifying the definition of 'design based' inference. When the above steps do not apply, NHST is out of context because of lack of a priori basis.
   In the case of research related to ACC detection, for instance, "most detection studies apply NHST to a sample of data, and determine whether to reject the null hypothesis of zero trend in the atmospheric variable under consideration" (Nicholls, 2001). These studies typically use all the available records and these data are far from being randomly selected samples with size fulfilling the requirements in terms of α and β (Prosdocimi et al., 2014). Moreover, hydro-meteorological observations usually exhibit serial and spatial correlation, and other properties that can be accounted for but make the outcomes further uncertain. Based on the above remarks, it follows that such studies fall in the non-randomized exploratory family, thus excluding confirmatory tools such as NHST (assuming that this is a logically coherent procedure), and requiring in turn the use of split samples or future data subsets to provide confirmatory/disproving information (Flueck and Brown, 1993). Moreover, in the exploratory stage, one is really not interested in finding a statistically significant effect or trend (which can always emerge by increasing the sample size) but in physically significant effects.

Assuming that one overlooks these aspects, how should the outcome of NHSTs be interpreted? Rejection of H_0 does not necessarily imply the acceptance of H_1, as the discrepancy of the observations from the conditions corresponding to H_0 can actually result from factors not included in the formulation of H_0 (e.g., larger variability due to lack of independence) and different from H_1. It is also less legitimate to accept substantive hypotheses owing to the formal logic fallacy of 'affirming the consequent'. On the other hand, if H_0 is not rejected, then this does not mean that it can be concluded that H_0 is true, but only that experimental evidence does not support the rejection of the null hypothesis. Unfortunately, the intricacy of such reasoning is once again a result of the hybrid nature of NHST. In fact, Fisher intended significance tests as tools for screening situations worthy of deeper study (without H_1), while Neyman-Pearson hypothesis tests were proposed as rules of action implicitly accounting for the consequences (quantified a priori by α and β) of choosing between two competing alternatives.

Therefore, even overlooking logical flaws, trend NHSTs can only reveal possible changes which are not compatible with random fluctuations corresponding to very specific reference processes (e.g., independent and identically distributed (*iid*) random variables), thus requiring further investigation. We can then extend to general trend NHSTs what Busuioc and von Storch (1996) recommended for Pettitt test: trend/change tests should be used not as *tests* but as mere *tools for preliminary screening*. Small (large) values of the test statistics should be taken as indications for possible upward or downward changes. Such changes should be accepted as physically meaningful if they can be related with a predictable process based on theoretical models (e.g., logistic models describing population growth under limited resources) and/or well identified physical dynamics justifying causality (e.g., dam building, river training, etc.).

3.2. Hidden dependence: the limits of short time series and the role of reference models

Once clarified what conclusions can(not) be drawn from trend NHSTs and in which context, we can better discuss the role of persistence in trend detection. The example in Section 2.1 shows that the underestimation of persistence plays a key role and should be accounted for. The estimation of Hurst parameter involved in Pettitt tests adapted for fGn yields H values of 0.5 and 0.66, for the Nera River and Velino River, respectively. However, while $H = 0.5$ is consistent with the fact that the Nera River data are actually independent, the value $H = 0.66$ underestimates the actual value 0.8 even though the estimator is corrected for the bias by the method described in Appendix A. This confirms that model identification under scarce observations (i.e. short time series) is a difficult statistical task, subject to large uncertainty and bias. Koutsoyiannis and Montanari (2007) have already investigated this aspect showing that very long time series (thousands of observations) are required to correctly recognize fGn. Therefore, underestimation of persistence is an aspect that should not be overlooked when using whatever trend NHST involving a correction procedure for this property. Note that the underestimation of persistence might lead to consider data approximately independent, and thus applying standard tests with no corrections. This choice can inflate the number of detected significant *deterministic* trends.

One should also account for the underestimation of variance, which is another well-known phenomenon related to the persistence of some stochastic processes (e.g., Koutsoyiannis, 2011; Tyralis and Koutsoyiannis, 2011; Koutsoyiannis and Montanari, 2015) arising from the fact that the process stays in a given subset of the state space for several time steps, thus requiring much time to explore the entire state space. For example, even though the fGn time series in Fig. 1(j) has theoretical unitary variance, its 100-size subsets have an average variance equal to 0.9. Note that the larger uncertainty related to persistence should not be confused with "the assumption of a deterministic temporal change in the pdf (specifically, the second moment) of a random process (evidently, from historical observations to future analysis period)" (Milly et al., 2015) (see Section 4 for further details). In other words, persistence inflates the overall variance, which is larger than that corresponding to the independent case. Of course, as for persistence, estimating the variance from short time series can yield substantial underestimation, with similar consequences on the outcome of trend NHSTs (i.e., inflated number of detected *deterministic* trends).

Dependence is introduced in trend NHSTs to build a more realistic H_0 relaxing the assumption of independence in the *iid* model, whereas deterministic trends relax the hypothesis of identical distribution in the *iid* model in the H_1 side. Since long-term patterns in finite samples can result from (be effects of) both persistence and deterministic changes in distribution, in trend NHSTs we attempt to compare two hypotheses that can produce comparable effects, knowing a priori that they can. Thus, why is dependence considered a 'null' condition, while deterministic changes in distribution are assumed to produce an effect? The difference between the two schemes is that the deterministic trends require attribution whereas persistence is compatible with pure stochastic processes, implicitly assuming that persistence provides a realistic description of natural systems.

As mentioned above, NHSTs cannot tell us which model is the most credible, and they cannot be used for such exploratory studies but only in a confirmatory/disproving stage by using independent data and a well-specified model reproducing properties that are unlikely to be reproduced by other competitors. Therefore, in absence of physical theory, both options (persistence and deterministic changes) are legitimate, but the main issue concerns their ability to describe variations in the wider population. This is usually only achievable when there are additional sources of data against which each model can be judged.

3.3. Distinguishing processes and time series: a matter of attribution

Often trend NHST is the first step to infer systematic changes of the studied process over time and thus its nonstationarity, eventually justifying the adoption of nonstationary models. However, this procedure is logically flawed, and the opposite should be done. Exploratory analysis should suggest a set of theories/models. These models should be used to reproduce challenging properties of the observed data, and then confirmatory/disproving analysis in terms of prediction should be applied. A successful model/theory can provisionally be retained until disproved by further applications on new data. The common inversion of reasoning is partly related to the confusion between processes and time series.

The problem of contaminated data series with trends and seasonal effects has been a matter of common experience for hydrologists. The traditional way of dealing with such an issue is to produce a new time series (the output of a certain filter or adjusting procedure) which represents in some sense an estimate of what the real series would be if the contaminating effect were absent. According to Jaynes (2003, p. 536): "Then choice of the 'best' physically realizable filter is a difficult and basically indeterminate problem; fortunately, intuition has been able to invent filters good enough to be usable if one knows in advance what kind of contamination will occur". When a data set is filtered according to incorrect assumptions, detrending may introduce spurious artifacts that distort the information that statistics and probability theory could have extracted from the raw data; so, caution is advisable especially with refined filters giving a false sense of reliability, whereby this can come only from reasoned judgment. Hence, testing trends on finite and short time series can easily be inconclusive and/or misleading because of the intrinsic difficulty, if not impossibility, of detecting nonstationarity (of a process) solely from data *without exogenous information*, as is discussed later (leaving aside the logical arguments discussed above).

Let us suppose that a dam was built along a river, thus influencing its regime according to the dam operation rules. If we know a priori the existence of the dam, we do not need to perform a trend analysis because we already know that the flow regime has been changed by dam construction; we can study how the dam impacted on specific characteristics of the flow regime (i.e., the effect size), if this information is not already included in dam design specifications. On the other hand, if we do not have a priori information on the dam existence, trend NHST can only tell us that some source of discrepancy from pure randomness is present; however, this does not allow one to infer nonstationarity of the underlying process without additional information identifying a clear causality rule. In fact, as shown in Sections 2.1, 3.1, and 3.2, multiple factors can generate such discrepancy in finite time series, and trend NHST does not allow one to draw conclusions on the substantive causes. In these circumstances we should propose a set of theories based on plausible reasoning, develop suitable models, and compare their prediction performance with independent observations. However, these models will be credible only if they incorporate rules describing the dynamics of the process (e.g., dam's effects), thus making its evolution predictable (i.e., river flow will follow a given regime until the dam operates).

Therefore, without clear attribution via exogenous information, trend NHST can only provide a generic indication that further investigation is required (according to the rationale of Fisher and Neyman-Pearson original methodologies). In this respect, such attribution cannot be vague or based on some kind of statistical analysis affected by its own uncertainty, because what is needed is not some sort of statistical correlation but a (substantive) causal physical relationship that should be general and valid beyond the period of the observed records. Therefore, even sophisticated regression models (e.g., GLMs, GAMs, etc.) do not fulfill these requirements as they fall in the class of analog models (Flueck and Brown, 1993) for which extrapolation is not advisable (Cooley, 2013) and easily leads to physically inconsistent

predictions (Serinaldi and Kilsby, 2015; Luke et al., 2017). Non-stationarity requires the postulation of a law of temporal evolution of the process, and this law should be based upon substantive hypotheses in order to be general and valid for prediction of still unobserved data (Poppick et al., 2017). Using the example of the dam, GLMs fitted on the data can incorporate time dependent terms but these data-driven regression laws do not say anything about the dam operation rules and their effect, and their extrapolation in time is not supported by any reasoned judgment about the causes of the observed patterns (are they real or spurious? how will they evolve?). On the other hand, additional information on the dam existence and operation and its mathematical formalization can justify the introduction of nonstationary models (e.g., Ayalew et al., 2017). Thus, nonstationarity and corresponding modeling strategies are allowed only if we make (a priori) assumptions about the processes, and the causes of nonstationarity are clearly identified and formalized via *deductive* reasoning about e.g., the effects of a dam on the river regime. Nonstationarity cannot result from *inductive* inference from data only, as the observed patterns can be the effect of various unknown causes (persistence, nonlinearity, nonstationarity, etc.), which cannot be discriminated in exploratory studies or misusing questionable confirmatory tools.

It should be noted that these remarks are well-known in climatology (Hasselmann, 1997), for instance, but seem to be overlooked in many hydro-meteorological studies relying almost exclusively on trend testing to draw conclusions. Indeed, according to Mitchell et al. (2001, p. 700), "Detection is the process of demonstrating that an observed change is significantly different (in a statistical sense) than can be explained by natural internal variability. However, the detection of a change in climate does not necessarily imply that its causes are understood... from a practical perspective, attribution of observed climate change to a given combination of human activity and natural influences requires another approach. This involves statistical analysis and the careful assessment of multiple lines of evidence to demonstrate, within a pre-specified margin of error, that the observed changes are: unlikely to be due entirely to internal variability; consistent with the estimated responses to the given combination of anthropogenic and natural forcing; and not consistent with alternative, physically plausible explanations of recent climate change that exclude important elements of the given combination of forcings... Detection (ruling out that observed changes are only an instance of internal variability) is thus one component of the more complex and demanding process of attribution".

These recommendations are fully general and not restricted to the problem of ACC detection and attribution. They highlight the importance of defining the magnitude of internal variability (space-time covariance and dependence structure (Hasselmann, 1993; 1997; Poppick et al., 2017)), which is a challenging task (as discussed in Section 3.2 and further in Section 6), as well as the need of jointly using deductive and inductive methods, and excluding other physically reasonable explanations before arriving at a clear attribution.

4. *Voces significant res mediantibus conceptis*[1]: missing interpretant generates a hiatus between sign and object

4.1. Stationarity

In the previous sections, we discussed some theoretical and practical limits of trend testing, including the problems posed by the intrinsic nature of the hydro-meteorological data, the misuse of confirmatory tools in exploratory analyses, and the influence of dependence, as well as the basis and logic of NHSTs and their interpretation. All these aspects raise serious questions regarding the actual information and conclusions that can be drawn from trend NHSTs and exploratory studies relying on them. However, to better understand why nonstationarity cannot be inferred from this analysis we need to go back to basic concepts and definitions.

As the title of this section suggests, we put the emphasis on the meaning of common terms in the context of trend testing, as the semantics of those terms is often confusing in the literature. Although there is ongoing debate about this issue (Koutsoyiannis and Montanari, 2015; Milly et al., 2015), we believe it is worth recalling and expanding it, where necessary, because of its importance for correctly setting up and interpreting data analysis. Unless stated otherwise, throughout this paper we use upper case letters for random variables and lower case letters for values, parameters, or constants.

Referring to Koutsoyiannis and Montanari (2015) for a rigorous presentation of the formal definitions of stationarity and non-stationarity, we recall that "... a stationary stochastic process in the sense of Khinchin ... is a set of random variables X_t depending on the parameter t, $-\infty < t < +\infty$, such that the distributions of the systems $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ and $(X_{t_1+\tau}, X_{t_2+\tau}, ..., X_{t_n+\tau})$ coincide for any n, $t_1, t_2, ..., t_n$, and τ". This definition has been translated in various ways such as: "Stationarity means that hydrological variables fluctuate randomly within an unchanging envelope of variability" (Bayazit, 2015), or "... stationarity, or temporally stable probability distribution functions (PDF)," (Rice et al., 2015). Even though such definitions are acceptable in informal discussions, the actual meaning of Khinchin's definition merits some further discussion to avoid misunderstandings. Assuming that t denotes time, Khinchin's definition means that the n-dimensional joint distribution of n random variables is identical independently of their location along the time axis. However, since the mathematical definition refers specifically to random variables X_t, the sets of realizations $(x_{t_1}, x_{t_2}, ..., x_{t_n})$ and $(x_{t_1+\tau}, x_{t_2+\tau}, ..., x_{t_n+\tau})$ are unavoidably different. By the way, Mandelbrot (1982) (p. 384) emphasized that: "When mathematicians first encountered stationary processes having extremely erratic samples, they marvelled that the notion of stationarity could encompass such wealth of unexpected behavior. Unfortunately, this is a kind of behavior that many practitioners insist is *not* stationary".

So, the actual problem in inductive exploratory analysis of trends is to understand if such fluctuations are consistent with a unique n-dimensional joint distribution or they come from different distributions. Given the uniqueness of observed hydro-meteorological records and the well-known uncertainty in making inference from very short time series (the most common case in hydro-meteorology), the problem is challenging. In order to make the problem easier to treat, one often focuses only on the first moments (actually, up to second order because of the high uncertainty in estimating higher-order moments (Lombardo et al., 2014)), thus introducing the concept of weak stationarity, where Khinchin's definition reduces to identity of *population* means $\mathrm{E}[X_i]$, *population* variances $\mathrm{Var}[X_i]$, and *population* covariances over n time steps independently of their location along the time axis.

4.2. Ergodicity

We cannot emphasize too strongly the clear distinction between the population properties that are deduced logically from the theory and the sample properties that are determined empirically from observations. Sample estimates are derived from time averages whose relationship to the statistical parameters of the theoretical process must be established only in the form of ergodicity. In order to highlight the importance of ergodicity, it is worth recalling that a stochastic process $X(t, \zeta)$ is a collection of time functions depending on the outcome ζ of an experiment $\mathscr{E}$, or a collection of random variables over a parametric support t (time, space, etc.) (Papoulis, 1991, pp. 285–286). Fig. 2 helps to clarify these definitions. For a fixed outcome ζ^* (i.e., a fixed coordinate along the 'Event space' axis), $X(t, \zeta^*)$ is a single time function (or trajectory) describing a *sample* (or realization) of the given process. One of these realizations can be the sequence of the truly observed records, while the others are possible outcomes that did not occurred

[1] Signs correspond to objects through interpretants (Eco, 1976).

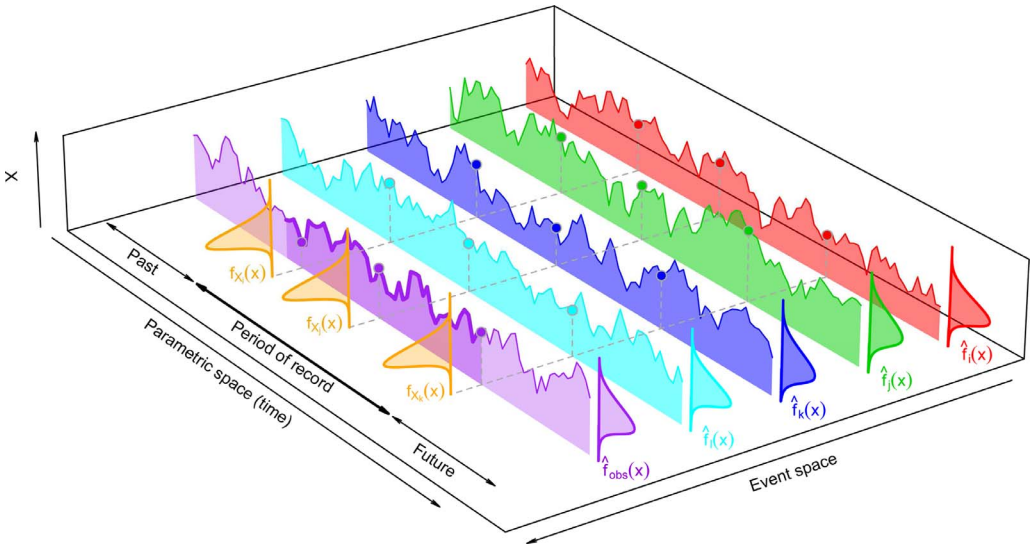

Fig. 2. Explanatory sketch of a stochastic process $X(t, \zeta)$. For a given outcome ζ^* of an experiment, the trajectory $X(t, \zeta^*)$ denotes a sample of the process. For example, the 'obs' case refers to a possible observed time series, while the other patterns are alternative samples associated with other possible results of the experiment. For fixed t^*, the set of values along the 'Event space' axis describe the state space of a random variable $X(t^*, \zeta)$, i.e. the given process at time t^*. For fixed t and ζ, bullet points denotes the specific value assumed by the process at a specific time. $\hat{f}_{obs}, \hat{f}_i, \hat{f}_j, \hat{f}_k$, and $\hat{f}_l$ are the empirical probability density functions of various samples of $X(t, \zeta)$ corresponding to some given values of ζ, while f_{X_i}, f_{X_j}, and f_{X_k} are the probability density functions of the random variables $X(t_i, \zeta)$, $X(t_j, \zeta)$, and $X(t_k, \zeta)$ describing the state of the process at time t_i, t_j, and t_k, respectively. See Section 4.2 for further details.

but could. On the other hand, if t is fixed as t^* and ζ varies, then $X(t^*, \zeta)$ is a random variable describing the *state* of the given process at time t^*. If both t and ζ are fixed, $X(t^*, \zeta^*)$ is a number, i.e. the specific value assumed by the process at the specific time.

In real world applications, we often know only a single finite-size sample of $X(t, \zeta)$ (e.g., a sequence of daily stream flow values between year t and $t + \tau$). So, a central problem is to infer the parameters of the underlying stochastic process from such sample. This is possible only if the process is ergodic, meaning that the time average of any (integrable) function $g(X(t, \zeta))$ equals the true (ensemble) expectation $\mathrm{E}[g(X(t, \zeta))]$, as the size of the available sample tends to infinity (Papoulis, 1991, pp. 427–428). Clearly, this is not possible if $\mathrm{E}[g(X(t, \zeta))]$ depends on t. Therefore, we must assume a stationary underlying process. Focusing on the mean of the process, ergodicity implies that

$$\lim_{\tau\to\infty} \overline{X}_\tau(t, \zeta_{obs}) := \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{s=t}^{t+\tau} x(s, \zeta_{obs}) = \mathrm{E}[X(t, \zeta)] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x. \quad (1)$$

More generally, ergodicity allows the use of the empirical probability density function $\hat{f}_{obs}$ (or $\hat{f}_i, \hat{f}_j$, etc.) of a sample of $X(t, \zeta)$ as an estimate of the probability density functions f_{X_i} (f_{X_j}, ...) of the random variables $X(t_i, \zeta)$ ($X(t_j, \zeta)$, ...) describing the state of the process at time t_i (t_j, ...). If a process is nonergodic, then statistical inference from data is not allowed because sample averages, variances, and distributions are not representative of their population counterparts. Moreover, we should consider that stationarity is a necessary (but not sufficient) condition to ergodicity of stochastic processes (Koutsoyiannis and Montanari, 2015). Therefore, a nonstationary process is nonergodic; thus, estimates from data are not representative of the process when we claim nonstationarity. In fact, nonstationarity implies that the population distributions f_{X_i}, f_{X_j}, and f_{X_k} in Fig. 2 are not identical to each other, and thus $\hat{f}_{obs}$ is no longer representative of any of them. In fact "The histogram is assumed (at least implicitly) to be an estimate of the marginal STATIONARY distribution. Note that second-order stationarity, or one of the other forms of weak stationarity, is not sufficient; strong stationarity at least must exist for the special case of $n = 1$ (the number of points, not the dimension of the space). If the random function is not stationary, at least to this extent, then the histogram is not an estimate of a distribution related in a known way to the random function" (Myers, 1989). Similarly, if f_{X_i}, f_{X_j}, and f_{X_k} have different moments (e.g., mean and/or variance changing in time), the empirical sample moments are not representative of any of these local population moments.

This can be surprising in light of the extensive use of nonstationary models such as GLMs and GAMs with time-dependent parameters in hydro-meteorological frequency analysis, for instance. Of course, the problem is not about these models by themselves, but their misuse. In fact, these models actually fit local trends (observed in the period of record) that can be due to multiple factors (anthropogenic activity, persistence, nonlinearity, etc.), which in turn cannot be identified by data alone. Moreover, they are often justified owing to the better performance compared with *iid* versions (which are not challenging competitors), and overlooking more realistic options that can yield patterns close to the observed ones. Nonstationary models are legitimate when there is additional information on the cause of time-dependent behavior. The identification of the cause of local trends is paramount for extrapolation in order to be sure that the nonstationary effects continue beyond the period of record. Without this additional information, prediction based on pure data-driven time-dependent patterns easily yields physically inconsistent results when extrapolating into the future or past (Luke et al., 2017; Serinaldi and Kilsby, 2015; Villarini et al., 2009b). Actually, low reliability and high uncertainty in predictions of evolution of nonstationary patterns might be an index of the little evidence supporting the nonstationary choice (Serinaldi and Kilsby, 2015).

4.3. Nonstationarity

At this point, a definition of nonstationarity is required. In order to illustrate nonstationary processes, Koutsoyiannis and Montanari (2015) considered the decomposition $X_t = d_t + \mathscr{E}_t$, where $\mathscr{E}_t$ is a stationary stochastic process and d_t is a *deterministic function of time* $d_t \equiv d(t)$. Milly et al. (2015) proposed a similar representation, namely, $X_t = a_t + b_t \mathscr{E}_t$, where a_t and b_t are deterministic and $\mathscr{E}_t$ is stochastic. We slightly generalize the decomposition suggested by Koutsoyiannis and Montanari (2015) as follows

$$\mathrm{G}[X_t] = d_t + \mathrm{G}[\mathscr{E}_t], \quad (2)$$

where $\mathrm{G}[\cdot]$ is a generic operator (some examples are given below).

According to Koutsoyiannis and Montanari (2015), a deterministic function of time is "precisely known and perfectly predictable" meaning that a system input corresponds to a single system response, contrasting stochastic dynamics where a single input could result in multiple outputs. Since every inductive analysis based on observed data is always affected by uncertainty, a deterministic function cannot be inferred from the data only, but it should result from deductive reasoning and be validated by data which were not used in the model construction. Notice that this definition is consistent with the idea which became famous as Laplace's demon, i.e. the classical definition of strict physical

determinism. According to Laplace, the demon is indeed a superhuman intelligence that could know and model all details of the universe to infinite precision: "for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes" (Laplace, 1814). In other words, if all changes in nature are expressible through mathematical functions of time, complete and precise knowledge of the initial conditions at a certain moment allows one to perfectly predict the conditions at all later (and earlier) times.

However, predictability and determinism are also easy to disentangle in practical applications. As shown in many studies on deterministic chaos, the approximate character of scientific knowledge renders dynamical systems unpredictable even though they are fully governed by underlying deterministic laws (Sivakumar, 2016; Yevjevich, 1974). Actually, determinism is a matter of spatio-temporal scales; in fact, even if the process (i.e., vector function representing the deterministic dynamics) is perfectly known, perfect predictability beyond a given temporal horizon can completely be lost owing to very small uncertainty in initial conditions (Berliner, 1992; Koutsoyiannis, 2010; Lorenz, 1963) that is magnified by possible nonlinearity leading to emergence of (deterministic) chaotic behavior (e.g., von Storch and Zwiers, 2003, pp. 1–2). While chaos theory explains unpredictability of a deterministic system in practice, Laplace's demon assumes perfect predictability under ideal, complete and precise knowledge of the system (including initial conditions). Therefore, the two ideas are compatible with each other, as already recognized by Laplace himself, who wrote "All these efforts in the search for truth tend to lead it (human mind) back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed" (Laplace, 1814).

The decomposition in Eq. (2) is a rather general description of nonstationarity. In fact, if $\mathrm{G}[\cdot]$ is the identity operator, Eq. (2) describes the (simplest) decomposition of the process itself, $X_t = d_t + \mathscr{E}_t$. If $\mathrm{G}[\cdot]$ is the expectation, we have $\mathrm{E}[X_t] = d_t + \mathrm{E}[\mathscr{E}_t]$, referring to a process that is nonstationary in mean (first moment). When $\mathrm{G}[\cdot]$ denotes the variance, then $\mathrm{Var}[X_t] = d_t + \mathrm{Var}[\mathscr{E}_t]$, describing a process whose variance depends on a deterministic function of time, such as the Brownian motion (Bm) with $\mathrm{Var}[X_t] \propto t$. In other words, nonstationarity implies that the distribution of the system $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ depends on time by a deterministic function, which can however refer to one or more characteristics of the distribution (e.g., mean, variance, higher-order moments, autocorrelation, etc.). For Bm, as well as fractional Bm and autoregressive integrated moving average processes (ARIMA), nonstationarity refers to deterministic functions of statistical moments, and inference is performed on the increment process $Y(t) = X(t+1) - X(t)$ because taking the first difference yields a stationary process by removing the dependence of the moments on time. Whatever is the specific form of nonstationarity (in mean, variance, etc.), statistical inference (e.g., calculation of moments) only applies to a corresponding stationary process obtained by suitable transformations (e.g., differencing) *under the assumption* that the original process has a specific form of nonstationarity. However, without an attribution identifying the source of the deterministic dependence on time, nonstationarity cannot be claimed from the data only. Claiming nonstationarity (i.e., the existence of a deterministic function of time for some statistical properties of X_t) on the basis of the outcome of NHSTs such as MK, Pettitt, or unit root NHSTs such as Dickey and Fuller (1979) and Kwiatkowski–Phillips–Schmidt–Shin (Kwiatkowski et al., 1992) is simply not possible owing to the problems discussed in Section 3 and the definitions given above.

4.4. What is a trend?

We argue that some widespread misconceptions concerning trend detection and attribution result from a lack of definition of 'trend'. Thus, attempting a reasonable definition and opening the debate about this point seems useful. First of all, the concept of trend should be related to nonstationarity. This seems a reasonable assumption, as we are not interested in local but long-term patterns spanning for instance the entire series of records and resulting from e.g. persistence of underlying stationary processes. In this case, the stochastic nature of persistent (quasi-periodic or monotonic) patterns make their magnitude, onset, and end unpredictable, and a pure stochastic stationary description is sufficient. Thus, we should conclude that a trend of true interest (which is the focus of the largest part of hydro-meteorological literature on the topic) should strictly be related to a form of nonstationarity.

If so, recalling Khinchin's definition of stationarity and the discussion in Section 4.3, a stochastic process has a trend if one or more of its statistical properties vary in time according to a *deterministic* law of time d_t. The function d_t can be monotonic, non-monotonic, periodic, and can refer to the average, variance, or other statistical properties of the process (see Section 4.3). In this respect, there is no difference, for instance, among (i) trends defined as smooth, long-range changes in some moment/parameter of the time-varying distribution (as used e.g., in GLM/GAM modeling), (ii) stochastic trends captured by random walk-type processes (i.e., Bm, fBm, ARFIMA, etc.), or (iii) trend described by physical equations in processes involving stochastic differential equations or different types of physical-statistical models. In all of these cases, for the parameters of GLMs/GAMs, for the variance of Bm and fBm processes, and for the specific characteristics described by the physical part of physical-statistical models, there is a function d_t accounting for deterministic time-dependent evolution of the system. Concerning d_t, a misconception widespread in GLM/GAM-based hydrological frequency analysis is the belief that replacing t with other variables makes the model nonstationary. This is true only if such variables are themselves nonstationary, and thus time-dependent according to a well-defined function. Replacing t with teleconnection indices (e.g., North Atlantic Oscillation index) or other variables showing clear stochastic behavior simply yields stationary doubly stochastic models.

If a trend is identified with the existence of a deterministic function of time d_t, and thus with nonstationarity, remarks on detection and attribution provided in Section 4.3 apply, and in particular we should exclude the possibility to make inference from data only. For example, seasonal cycles are forms of d_t resulting from a fundamental deductive reasoning (exogenous information) concerning planetary dynamics and corresponding mathematical theory. This deductive step allows for the choice of inference tools and then a quantitative evaluation of seasonal components from data for each specific case. Seasonal cycles are predictable with negligible uncertainty as we are almost sure of the occurrence of equinoxes and solstices in the next decades or even centuries, unless the occurrence of unpredictable catastrophic events acting at the solar system or galaxy scale. Here, lack of uncertainty refers to the existence of the seasonal d_t, and not to its contingent parametrization, which varies in each specific case.

Under these premises, introducing other forms of trend should rely on the same approach, merging deductive and inductive reasoning. For example, a commonly used approach of comparing nested models with time-varying and constant parameters by using some performance criterion is not enough if the time-dependent function d_t does not result from deductive reasoning, but results from simple fitting procedures. In these cases, higher parametrized (time-dependent) models might simply account for local apparent trends, giving very poor performance in prediction owing to lack of identification of substantial causes acting beyond the period of record (Serinaldi and Kilsby, 2015).

Trends of interest in hydro-meteorology are often monotonic or low-frequency type spanning over the period of record (possibly related to anthropogenic activity). We argue that 'frequency' is actually the main difference between seasonal trends and other forms of deterministic time-dependent trends. Seasonal trends look like monotonic or half-wave trends if the focus is on sub-annual time windows, because the process is not fully developed at such a time scale, and one cannot

retrieve signal components with frequencies lower than half period of record, as described by the Nyquist–Shannon sampling theorem. Therefore, even the use of effective filtering methods such as singular spectrum analysis, wavelet analysis, or empirical mode decomposition cannot help in trend identification if we are not able to arrive at a clear attribution of the patterns described by the lowest frequency components resulting from filtering.

Being a form of nonstationarity or its expression, trends are allowed only if they rely on exogenous knowledge involving theoretical arguments or empirically well-defined processes (in agreement with McCuen (2003)). Without such an additional information, trends cannot be inferred from the data only, because they refer to the underlying process $(X_{t_1}, X_{t_2}, ...,X_{t_n})$ and not to its realizations $(x_{t_1}, x_{t_2}, ...,x_{t_n})$ (i.e., observed time series; in agreement with the starting point of Chandler and Scott (2011)). Without attribution to unique substantive cause and exclusion of any other possible cause, exploratory tools, filtering, or model selection can only highlight local (low-frequency, persistent) fluctuations but they do not allow one to make conclusions on stationarity or nonstationarity. This is in agreement with von Storch and Zwiers (2003) (p. 9) who stated: "Trends in the large-scale state of the climate system may reflect systematic forcing changes of the climate system (such as variations in the Earths orbit, or increased CO_2 concentration in the atmosphere) or low-frequency internally generated variability of the climate system. The latter may be deceptive because low-frequency variability, on short time series, may be mistakenly interpreted as trends. However, if the length of such time series is increased, a metamorphosis of the former 'trend' takes place and it becomes apparent that the trend is a part of the natural variation of the system". These remarks are general and hold true not only for climate but also in every context exhibiting large uncertainty about the number, type, and effect of the acting physical processes.

Based on the discussion above, we can then provide an unambiguous definition of trend as: time-dependent deterministic and therefore predictable change d_t of the properties of a process X_t, where the term "deterministic" implies prediction variance equal to zero (one-to-one relationship). This definition highlights that trends (and nonstationarity) refer to the underlying process, and attempting to infer nonstationarity requires both detection and attribution based on a combination of deductive reasoning, which supports and justifies the existence of a time-dependent deterministic function (i.e., trends and nonstationarity) and inductive reasoning, which provides (i) preliminary knowledge by exploratory data analysis, and (ii) quantification/parametrization of d_t by confirmatory/disproving analysis and modeling. We stress that 'prediction variance equal to zero' does not refer to the specific parametrization of d_t but its existence and its overall evolution. For example, the parametrization of seasonal trends (components) obviously varies even for the same process observed at different locations, but the existence of the seasonal cycle and its effects in terms of alternation of wet/dry and cold/warm conditions along the calendar year are predictable with (almost) no uncertainty. Other forms of trend/nonstationarity are allowed only if they are supported by the same kind of deductive and inductive arguments.

5. Trend and abrupt change tests: an overview of overlooked critical aspects in practical applications

In this section, we discuss some problems concerning the practical application of two NHSTs for trends, i.e. the Mann-Kendall (MK) (Mann, 1945; Kendall, 1970) and Pettitt tests (Pettitt, 1979). The following remarks apply under the assumption that we disregard logical arguments in Sections 3 and 4, still apply these tests for exploratory analysis, and use them to make conclusions on trends/nonstationarity.

Among many available statistical testing procedures devised for assessing the significance of a change (e.g., Kundzewicz and Robson, 2004), the MK and Pettitt tests are widely used rank-based nonparametric tests to check the presence and timing of slowly-varying and abrupt changes in the mean or median of hydro-meteorological variables such as rainfall, runoff, and temperature (e.g., Villarini et al., 2009a; 2011b; Ferguson and Villarini, 2012; Rougé et al., 2013; Tramblay et al., 2013; Guerreiro et al., 2014; Sagarika et al., 2014; Rice et al., 2015; 2016; Mallakpour and Villarini, 2015; 2016; Ahn and Palmer, 2016; Archfield et al., 2016; Do et al., 2017, among others).

The popularity of these tests is related to their simplicity in terms of implementation, their robustness against outliers or measurement errors (as they are rank-based), and the availability of exact or asymptotic distributions of their test statistics under null hypothesis {H_0: no trend/no change} and independence, i.e. *iid* conditions. Moreover, being based on the so-called Mann-Whitney statistic, the Pettitt test and the MK test are formally related to each other, thus highlighting that distinguishing between slowly-varying and abrupt changes is only a matter of scale and time of evolution of the change (Rougé et al., 2013; Serinaldi and Kilsby, 2016a).

Even though these tests are used to check changes in the mean or median, the first myth to dispel is that MK and Pettitt tests are devised to detect changes in the central tendency summary statistics. They actually check a wider hypothesis called stochastic ordering. Given a sequence of random variables $\{X_i\}_{i=1,...,n}$, with cumulative distribution functions F_i, MK checks against the alternative hypothesis H_1: $F_i(x) \gtrless F_{i+k}(x)$, for every i, every x, and every $k > 0$ (Mann, 1945), while Pettitt checks against H_1: $F_b(x) \gtrless F_a(x)$, where F_b (F_a) is the common distribution of the m ($n - m - 1$) random variables before (after) the change point (Pettitt, 1979). Therefore, even though these hypotheses are commonly restricted to a shift in the location parameter, these tests are sensitive to all possible conditions resulting in stochastic ordering (Serinaldi and Kilsby, 2016a).

Based on the above belief, when a change point or a monotonic trend is detected, often the magnitude of the abrupt change is quantified by the difference in mean or median between the sub-series before and after the change, while the trend by the so-called Sen's slope (e.g., Khaliq et al., 2009b; Rice et al., 2015; 2016; Nilsen et al., 2016; Tananaev et al., 2016). In light of the actual nature of MK and Pettitt tests, such quantification is not justified, especially for MK. In fact, even if we assume that MK only checks for changes in mean/median, it refers to monotonic changes that can be linear or nonlinear (stepwise, S-shaped, or abrupt as a limiting case) resulting from more general changes in the overall shape of the distribution. Of course, the choice of linear trends reflects practical requirements; indeed, assuming more complicated (higher-parametrized) patterns can be unjustified for usually short time series because of the additional uncertainty affecting the estimation. Since Sen's estimator for the slope of a linear trend is rank-based (nonparametric), it is considered more robust than classical mean square error (MSE); however, its nonparametric nature does not make it more coherent with MK outputs than MSE estimates of linear trends. Even though the need of quantifying a possible change is understandable, reducing the indication of possible monotonic trends given by MK to that of a linear trend is too restrictive, and does not reflect the rationale and outcome of MK test. Since perfectly linear trends rarely describe realistic evolution patterns of complex hydro-meteorological processes (even under actual deterministic forcings), such a kind of quantification should be at most purely qualitative, and possibly avoided in order to provide a correct communication. In any case, it cannot be considered an actual trend in light of discussion on detection and attribution in Section 4.4.

Of course, trend tests can only detect inhomogeneities within the time interval covered by the observed records. This also explains why they cannot be used to infer nonstationarity: stationarity is a property of the theoretical process X_t, for $-\infty < t < +\infty$, and concerns the identity of *population* statistical properties for every subset of random variables in every point of the time line, while trend tests can only check possible changes in finite and usually short time windows where observed fluctuations might easily be spurious. Since we cannot extrapolate conclusions beyond the period of records without identifying

a deterministic and predictable cause of such inhomogeneities, the outcome of trend tests cannot be used to justify the application of nonstationary models for frequency analysis. Such a usage is inappropriate and might lead to unrealistic predictions (Serinaldi and Kilsby, 2015).

The previous remarks have important consequences on the procedures used to account for the effects of the temporal correlation. The effect of the autocorrelation on tests devised for independent data is a general increase of the rejection rate of the null hypothesis {H_0: no trend} of the statistical test, even if the underlying process is stationary. This is due to the information redundancy that makes the effective sample size smaller than the observed size, thus implying that the effective variance of the test statistics to be used in the testing procedure under serial dependence is larger than that provided by standard results obtained under the hypothesis of independence (e.g., Bayley and Hammersley, 1946; Koutsoyiannis and Montanari, 2007). This phenomenon is known as variance inflation and has been accounted for using three general approaches: the explicit calculation of the inflated variance (e.g., Hamed and Rao, 1998; Koutsoyiannis, 2003; Matalas and Sankarasubramanian, 2003; Yue and Wang, 2004; Hamed, 2008; 2009b), prewhitening procedures (e.g., Katz, 1988; Kulkarni and von Storch, 1995; von Storch, 1999; Yue et al., 2002; Yue and Wang, 2002; Bayazit and Önöz, 2007; Hamed, 2009b), and bootstrap techniques (Khaliq et al., 2009a; Kundzewicz and Robson, 2004).

Referring to Khaliq et al. (2009b) and Bayazit (2015) for a review, we focus on some aspects that are generally overlooked:

1. Firstly, all tests involving the *iid* hypothesis should be corrected for the effect of autocorrelation. Neglecting this aspect might lead to contradictory results further discussed in Section 6.3.
2. Since some procedures involve trend removal (e.g., Yue et al., 2002; Hamed, 2008), this is usually supposed to be linear. As mentioned above this choice is understandable from a practical point of view but less defensible if it is interpreted as a deterministic evolution of some physical (hydro-meteorological) process. Linear trends cover a very limited subset of the actual hypothesis tested by MK and Pettitt tests as well.
3. The procedures proposed in the literature consider corrections based on the autocorrelation values estimated on the data themselves. This poses two problems: (i) for short time series, autocorrelation is generally underestimated (e.g., Koutsoyiannis, 2003; Koutsoyiannis and Montanari, 2007), where the bias is larger if the underlying process exhibits long range dependence (LRD); thus, when the correction procedure involves a specific dependence structure (often Markovian), autocorrelation should be adjusted (see Serinaldi and Kilsby (2016a) and Appendix A); and (ii) it is taken for granted that the dependence structure of the underlying process can be retrieved by the analyzed summary statistics (usually, annual minima, averages, maxima, etc.). The point (ii) is subtle but critical; in fact, the behavior of summary statistics can be strongly influenced by the nature of the underlying process; for example, processes with LRD yields maximum values over blocks of observations that tend to cluster in time (e.g., Bunde et al., 2005; Eichner et al., 2011). This results in apparent trends in terms of frequency and magnitude if the analysis relies on short series of such maxima, even though these summary statistics might easily show no or very weak autocorrelation. Since this behavior is found in stream flow time series (Serinaldi and Kilsby, 2016c), we show in the case study that it might have a dramatic effect on trend NHST outcomes.
4. In some cases, correction procedures are flawed, failing to provide any adjustment. As an example among others, the so-called trend-free prewhitening (TFPW) (Yue et al., 2002) was shown to be theoretically flawed (Serinaldi and Kilsby, 2016a), as its original version does not address the variance inflation problem, which can be even exacerbated. Since it has been widely applied thanks to its relative simplicity, results of several analyses reported in the literature should be taken with great care and possibly revised.

Two aspects characterizing several published trend analyses based on NHST need to be mentioned: (i) the correctness of applied tests (or methodology in general) is almost always taken for granted, while a preliminary check of their performance under H_0 (controlled conditions) should be performed (by simulation) before their application, especially if the tests were not developed by statisticians and result from some empirical reasoning without a necessary mathematical proof (as shown for the hybridization of Fisher and Neyman-Pearson methods in Section 3); (ii) empirical results are often interpreted without the necessary rigor, thus resulting in misleading conclusions, confusing artifacts with meaningful results.

6. Case study

In this section, we investigate the consequences of the above discussion on data analysis and its interpretation. To this aim, we use data already analyzed in the literature to show how results and conclusions can remarkably change if we account for logical, methodological, and practical issues discussed in previous sections. Note that our analysis is not exactly a study of reproducibility because data and some methods are not precisely equal to those applied in previous studies. However, the use of MK and Pettit is justified for sake of comparison with previous studies, and key general results are reproduced and then compared with new findings relying on more realistic null hypotheses.

6.1. Observational data

Long term trends in stream flows over the conterminous United States (CONUS) have been extensively studied. Referring to Sagarika et al. (2014) for a recent review, we recall that the past studies focused on various summary statistics and/or data sets, including peak discharge records (Barrett and Salis, 2016; Hirsch and Ryberg, 2012; Lins and Cohn, 2011; Mallakpour and Villarini, 2015; Villarini et al., 2009a; Villarini and Smith, 2010; Villarini et al., 2011a; Vogel et al., 2011), monthly data (Kalra et al., 2008; Lettenmaier et al., 1994), and mean daily observations (Ahn and Palmer, 2016; Lins and Slack, 1999; McCabe and Wolock, 2002; Rice et al., 2016; 2015; Sagarika et al., 2014). The interest for such an area is not only practical, but is also related to the great variety of hydrologic regimes/conditions across CONUS, as well as the availability of data and metadata, which allow for more accurate studies than in other parts of the globe.

Since the trend analysis described below (Section 6.2) requires both daily data and summary statistics (i.e., maxima or averages) on a seasonal and annual basis, in this study, mean daily flow records are used. The data set is extracted from the Hydro-Climatic Data Network (HCDN-2009) (Lins, 2012), which comprises 743 stations maintained by the U.S. Geological Survey (USGS). HCDN-2009 is a subset of the wider USGS GAGES-II (Geospatial Attributes of Gages for Evaluating Streamflow, version II) reference stations providing geospatial data and classifications for 9 322 stream gages. HCDN-2009 provides a stream flow data set suitable for analyzing hydrologic variations and trends in a climatic context, as it includes quality-controlled time series from stations that were screened to exclude sites where human activities or other activities affect the natural flow, and with sample size sufficiently large for analysis of patterns in stream flow over time (Lins, 2012). A list of HCDN-2009 stations along with basic attributes can be found at the web site http://water.usgs.gov/osw/hcdn-2009/, while the data set is freely available at http://waterdata.usgs.gov/nwis/sw.

This study focuses on 250 stations having continuous and simultaneous observations with no missing values between the water years 1951 and 2011 included (i.e., October 1950 to September 2011). Following Sagarika et al. (2014), the data set comprises only one station on a particular stream within each U.S. hydrologic unit code (HUC), to reduce spatial bias in the results. Moreover, even though some stations

have continuous data spanning longer periods, we selected only simultaneous data from 1951 to 2011 to guarantee temporal homogeneity and to allow some remarks on spatial correlation discussed later in Section 6.3. For seasonal analysis, seasons are defined as autumn (October–December), winter (January–March), spring (April–June), and summer (July–September).

6.2. Methodology

6.2.1. Testing local significance

In this study, possible slowly-varying trends and abrupt changes of some stream flow properties are analyzed by MK and Pettitt tests in four different settings:

1. Classical versions devised for independent random variables. Hereinafter they are denoted as 'standard MK' and 'standard Pettitt'.
2. A corrected and unbiased TFPW (TFPWcu) version of both tests accounting for first-order autoregressive AR(1) dependence (i.e., classic Markovian dependence) and bias correction for ACF underestimation (denoted as AR(1)-TFPWcu MK and AR(1)-TFPWcu Pettitt). TFPWcu procedure is applied to show that a correct TFPW procedure yields results different from those of the classical setting, highlighting that the similarities of results often recognized in the literature are actually artifacts (see Section 5). The reader is referred to Serinaldi and Kilsby (2016a) for further details.
3. A prewhitening version accounting for fGn dependence proposed by Hamed (2008) for MK test and adapted by Serinaldi and Kilsby (2016a) for Pettitt. This version allows one to account for long range dependence (LRD) and improves the original prewhitening procedure by introducing bias corrected estimates of the Hurst parameter H (characterizing the fGn ACF) based on the formulas provided in Appendix A. These tests are denoted as fGn-CPW MK and fGn-CPW Pettitt, where CPW indicates 'conditional prewhitening', meaning that the prewhitening procedure is applied only if H is found significantly different from 0.5 at the 5% significance level. These versions are detailed in Hamed (2008) and Serinaldi and Kilsby (2016a).
4. The last version is based on Monte Carlo simulation of daily stream flow sequences in order to check the impact of daily dynamics on the annual/seasonal statistics and trend test outcomes. This way, we introduce a more realistic null scenario in terms of dependence structures that is built by exploiting the whole available information instead of few tens of annual/seasonal summary statistics. In more detail, each daily stream flow time series is deseasonalized (following the procedure described by e.g., Serinaldi and Kilsby (2016c) and Serinaldi and Kilsby (2016b)), and residuals are resampled by the iterative amplitude adjusted Fourier transformation (IAAFT) method (Schreiber and Schmitz, 1996; Kugiumtzis, 1999; Schreiber and Schmitz, 2000; Venema et al., 2006a; 2006b; Serinaldi and Lombardo, 2017), which allows for simulation of surrogate data preserving almost exactly the empirical distribution function and power spectrum (ACF) of the observations. IAAFT surrogates are stationary (Franzke, 2013) by construction because of randomization of Fourier phases. Combining surrogate residuals and seasonal components yields synthetic daily stream flow time series under the null hypothesis preserving almost exactly both the marginal distribution and correlation structure of the observed ones. Small discrepancies in marginal distributions do not matter as the used trend tests are rank based. Therefore, summary statistics of interest (here, averages and maxima on a seasonal and annual basis) are extracted and standard MK and Pettitt tests are applied. The procedure is repeated many times to obtain the sampling distribution function of MK and Pettitt test statistics accounting for the dependence properties of the daily process under stationary conditions. In other words, our new null hypothesis is 'observed trends in annual/seasonal values are consistent with patterns coming from a stationary daily process with given (observed) marginal distribution and dependence structure'.

6.2.2. Testing field significance

When data from multiple stations are analyzed, one can ask whether the results imply that there is a significant effect when considering the entire group of stations, i.e. the so-called field significance (Daniel et al., 2012; Katz and Brown, 1991; Livezey and Chen, 1983; Wilks, 1997; 2006). This recognizes that, when performing multiple tests, it is more likely to detect significant changes by chance. This probability increases if the data are spatially correlated. In this case, spatial correlation acts similarly to temporal correlation, introducing information redundancy owing to possible similar patterns across spatially correlated sequences. Referring to Khaliq et al. (2009b) for an overview of methods to treat field significance, we recall that they fall in two categories (Daniel et al., 2012). One controls the false discovery rate (FDR, i.e., the expected fraction of local null hypothesis rejections that are incorrect) (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) and is nearly equivalent to the Walker test (Fisher, 1929; Katz, 2002; Wilks, 2006). The other relies on counting the number of rejections at local level and then comparing these values with the selected critical values obtained from the empirical distribution of number of rejections resulting from bootstrap procedures preserving (approximately) spatial or spatio-temporal correlation (Khaliq et al., 2009b; Wilks, 1997).

Daniel et al. (2012) highlighted that the choice of method depends on the spatial nature of the studied effect (e.g., trend). If it is expected that the effect is widespread (local), it would be preferable to use the count-based (FDR approach), while a combination of both approaches could be applied if there is no a priori expectation. In this study, we focus on the Walker test because (i) it is easy to implement and robust to cross correlations (Khaliq et al., 2009b; Wilks, 2006), and (ii) the counting method requires intensive and different simulation procedures for IAAFT-based trend tests and the other tests (Standard, AR(1)-TFPWcu, and fGn-CPW). We recall that the Walker test consists of comparing the smallest of p-value corresponding to the test statistics computed on k time series with the critical value $p_W = 1 - (1 - \alpha_{global})^{1/k}$: the global null hypothesis {H_0: no global trend} may be rejected at the a global level α_{global} if the smallest of k independent local p-values is less than or equal to p_W (Wilks, 2006). The counting method is only applied to check the variability of field significance for the lag-1 ACF term ρ_1, and the Hurst coefficient H in the series of annual/seasonal summary statistics (see Khaliq et al., 2009b, p. 121, for a detailed description). All tests are performed at the 5% local and global significance level.

6.3. Results

6.3.1. Temporal dependence of maximum and mean flows

Given the influence of autocorrelation on detection of possible trends, the dependence structure of annual and seasonal maximum and mean flows are preliminarily investigated. Owing to the limited sample size of such a type of data, only parsimonious models such as AR(1) and fGn are usually considered, even though the latter would require very long time series for a reliable inference. Figs. 3 and 4 show the spatial distribution of the sites where ρ_1 and H are found to be locally and globally significant, along with the sites where both models are locally significant. Global significance is assessed at the HUC scale. Of course, the number of sites showing possible significant persistence are higher for mean values than for maxima owing to the higher variability of the latter. No particular spatial patterns are evident.

These results are not surprising if we consider the persistence of the underlying flow process and recognize that maxima are values sampled from the daily process, while mean values result from averaging all daily data in each season or year. Nonetheless, Figs. 3 and 4 allow some methodological remarks. When we test significance for ρ_1 and H, we are

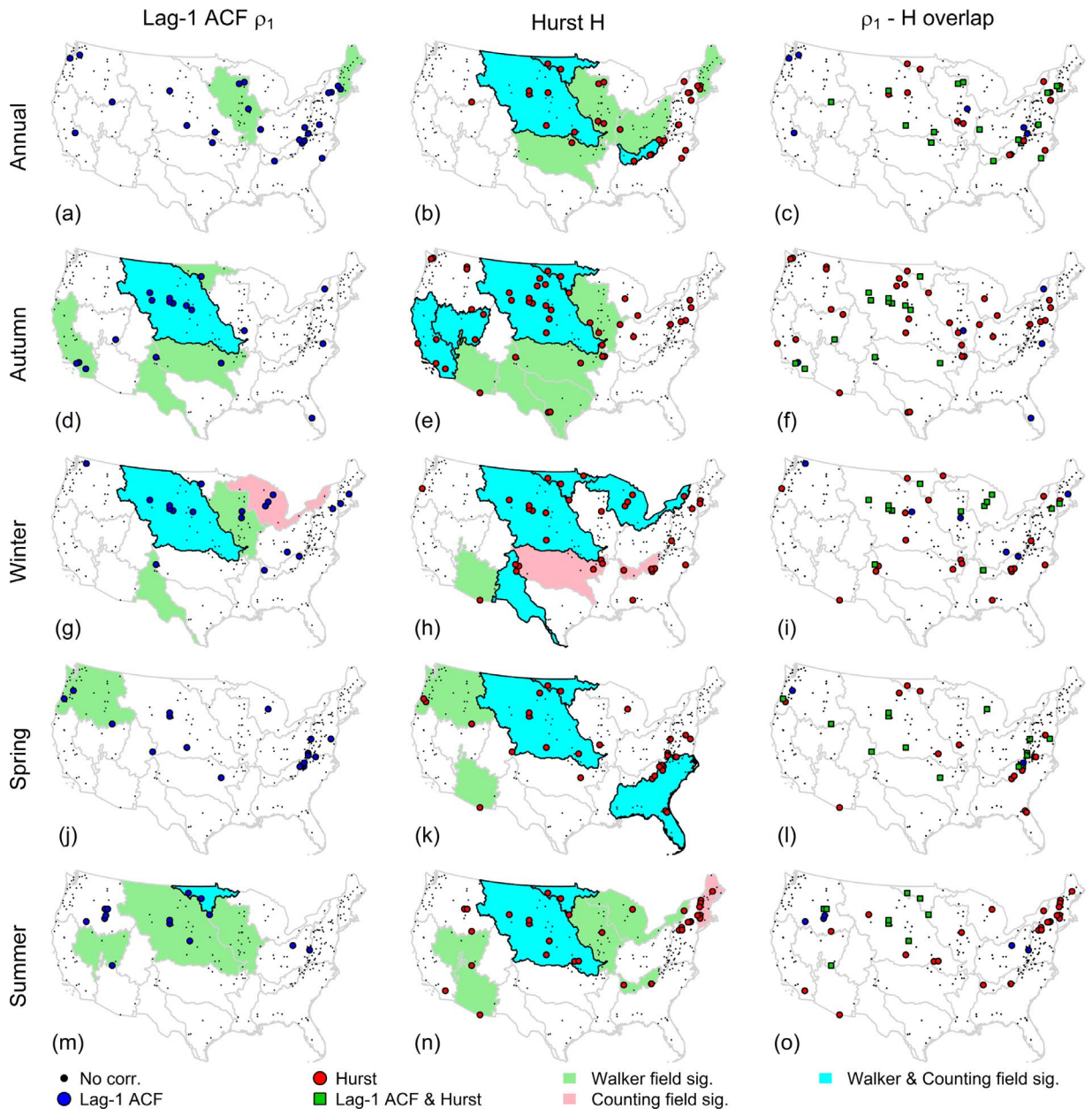

Fig. 3. Spatial patterns of significant values (at the 5% significance level) of ρ_1 and H for the annual flow maxima on an annual and seasonal basis. Left column of panels refers to ρ_1, middle column to H, while the right column highlights the sites where both ρ_1 and H are found significant. Maps in left and middle columns also show hydrologic units where ρ_1 and H are found to be field significant according to Walker test, bootstrap test, and both. See text for further details.

implicitly assuming an underlying model (AR(1) or fGn) as alternative hypothesis H_1; this implies that the estimates should be corrected accordingly for possible bias. For example, testing ρ_1 under the assumption that the underlying process is fGn requires a bias correction procedure different to that used under the assumption of AR(1) process. This aspect is often neglected and ρ_1 is commonly tested using biased estimates yielded by default estimators working under *iid* assumptions (Koutsoyiannis, 2016).

Another remark concerns the output of the Walker test and counting method for global significance. When the two approaches yield different results, the Walker test tends to identify more areas with significant results; according to the nature of the Walker test (which is sensitive to local effects), in some of these cases, field significance results from a single locally significant station falling in the HUC area, especially if the number of stations in that area is small. Daniel et al. (2012) provided some insight into selecting suitable sub-regions to assess field significance. To avoid bias, sub-regions should be identified *a priori* (before performing the analysis) and domain limitation should be made as a result of some physical insight. In this respect, HUCs are a credible choice. However, Daniel et al. (2012) also stress that "If some region in a domain is seen to contain a large number of significant stations, it is certainly inappropriate to apply the field significance test over just this limited domain without a physically based justification". Also in this case, the clustering of significant trends might be due to the spatial correlation of (large scale) meteorological variables driving the flow processes in a specific area. In other words, local clusters in space are the natural effect of spatial correlation, as local clusters of high/low values in a time series might reflect temporal correlation. Since it is easy to confuse spatio-temporal correlation with deterministic trends, these remarks highlight the importance of using clear definitions in order to distinguish between stochastic fluctuations and deterministic changes whose attribution should be based on *a priori* information and causality.

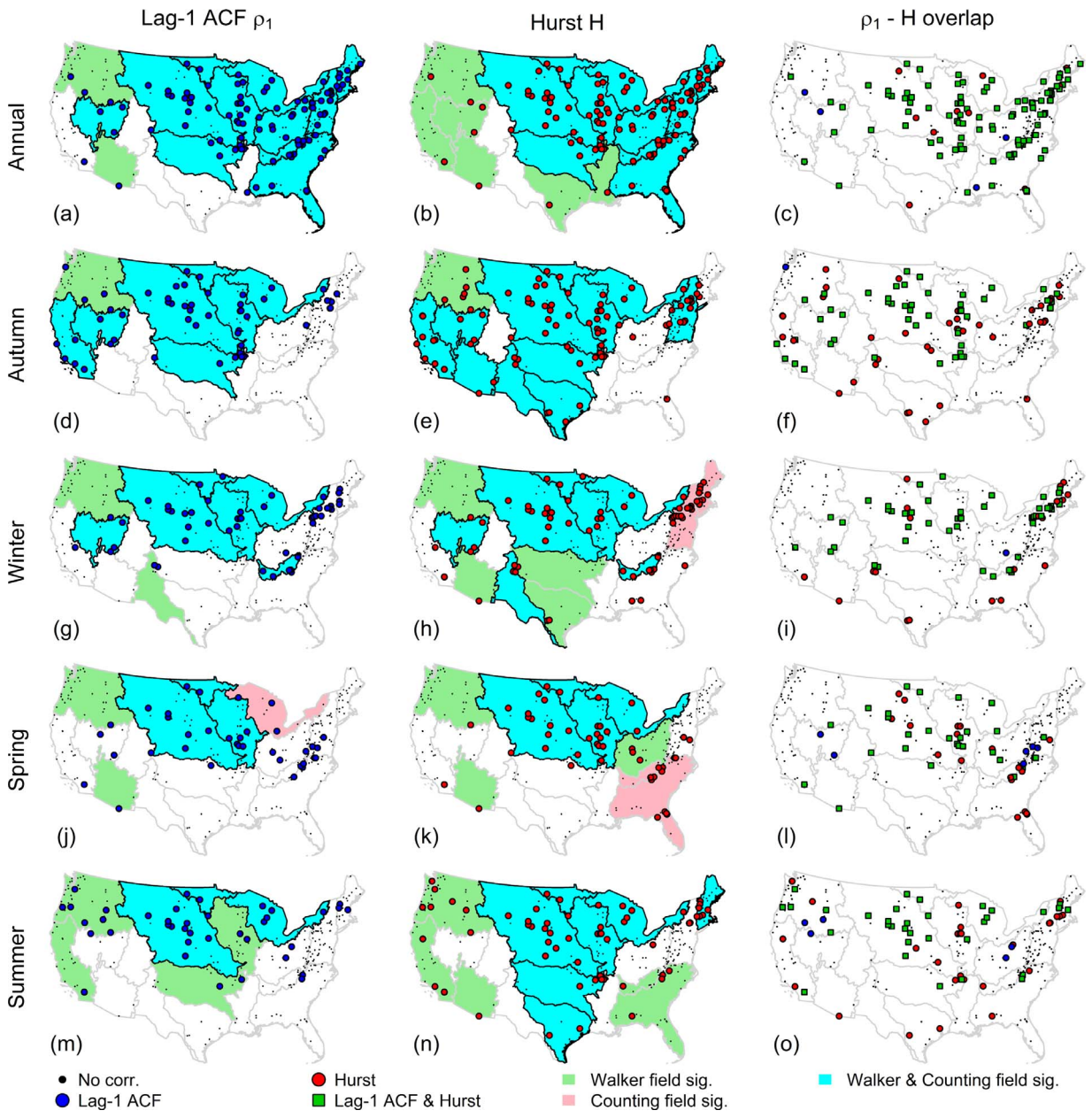

Fig. 4. As for Fig. 3, but for annual average flow values. The same interpretation applies.

6.3.2. Trend analysis of maximum and mean flows

Results of trend analysis are shown in Figs. 5–8. According to the aim of this study, as for correlation, the following remarks focus on methodological aspects. In fact, the spatial patterns of significant trends and their sign are similar to those published in the literature (e.g., Sagarika et al., 2014), while their interpretation deserves some remark. For example, AR(1)-TFPWcu MK and fGn-CPW MK tests on maxima (Fig. 5) tend to give similar number of rejections, which is smaller than that of standard MK. This contrasts with results often reported in the literature and confirm that they are related to the ineffectiveness of the original TFPW procedure. As a variety of hypotheses has been proposed to explain the disagreement between TFPW and other prewhitening procedures in terms of possible physical causes, it is therefore instructive to discover that they are pure speculations, and results are simply artifacts.

Another important result is the strong decrease of significant trends obtained by IAAFT MK. The annual/seasonal maxima extracted from (stationary) series reproducing the observed persistence of the daily sequences show apparent trends that are stronger than those resulting from persistent models directly fitted on annual/seasonal maxima, thus reducing the evidence for deterministic trends in such summary statistics. In other words, focusing on annual/seasonal maxima might lead to underestimate variability and temporal clustering of high/low values.

Similar remarks hold for annual/seasonal average values (Fig. 6). It should be noted that the residual clusters of positive trends in the New England (north west) for summer maximum and mean values (Figs. 5(t) and 6(t)) are coherent with the spatial correlation and climate dynamics of that area (Kingston et al., 2007). This does not mean that the New England rivers have not witnessed a possible increase in the last 60 years; however, it might be explained in terms of spatial correlation as a shared behavior depending on a common meteorological driver acting in an area characterized by quite a uniform response. Obviously, from a practical point of view, such an increase raises management problems, whose solutions however change if we assume that these changes are deterministic (in the sense specified in Section 4) or stochastic. In fact,

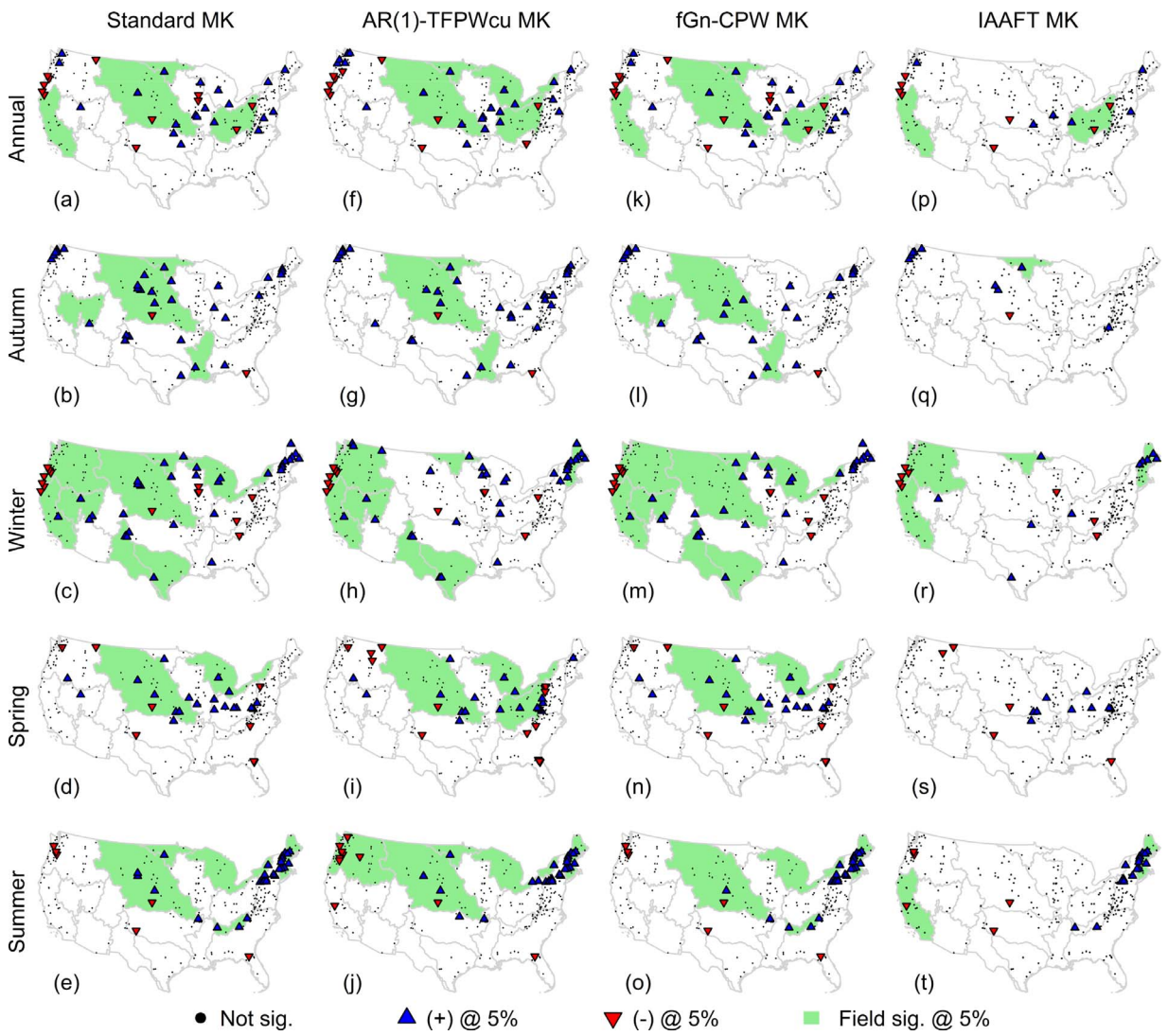

Fig. 5. Spatial distribution sites showing significant monotonic trends (at the 5% significance level) according to MK test for the annual flow maxima on an annual and seasonal basis. Panels in each row, from left to right, show results for standard, AR(1)-TFPWcu, fGn-CPW, and IAAFT versions of the MK test. Hydrologic units exhibiting field significance according to Walker test are also highlighted.

in the first case we should identify a predictable evolution law (making attribution), which is unlikely linear or polynomial, and cannot be deduced from the data themselves in the form of some arbitrary smoothing function.

Results for the Pettitt test in Figs. 7 and 8 highlight another aspect often overlooked in the literature and related to the fact that both tests rely on the Mann-Whitney statistic and they are theoretically related to each other (Rougé et al., 2013). This implies that standard MK and Pettitt often yield similar results in terms of significant changes in a given direction (upward/downward), even if such results are obviously interpreted in a different manner (slowly varying monotonic changes and abrupt changes) (e.g., Sagarika et al., 2014; Pathak et al., 2016). The link between the two tests also implies that both are sensitive to autocorrelation (Serinaldi and Kilsby, 2016a), and they should yield coherent results when autocorrelation is accounted for. A comparison of Figs. 5 and 6 with Figs. 7 and 8 supports this conclusion showing that both tests yield similar spatial patterns in terms of significant upward/downward changes. In this context, distinguishing and interpreting slowly varying and abrupt changes can be secondary as it is simply a matter of scales (e.g., Rougé et al., 2013). This highlights once again the need for attribution based on exogenous information which is not derived exclusively on purely statistical non-causal relationships. Results for IAAFT-based tests also show the dramatic decrease of evidence for deterministic changes when fluctuations of seasonal/annual averages and maxima are influenced by the entire flow process at daily scale. Therefore, even if the autocorrelation of the observed seasonal/annual averages and maxima is properly accounted for in AR(1)-TFPWcu and fGn-CPW trend tests, the above results reveal that it is not sufficient and might strongly underestimate the actual persistence of the underlying processes. When this is taken into account, apparently strong trends/changes might become coherent with the intrinsic variability characterizing stationary but persistent processes.

Recalling the discussion about the natural local clustering of significant results due to spatial correlation and testing multiplicity (Daniel et al., 2012), we studied how the number of significant trends globally changes across the CONUS for the different types of tests, i.e. the different treatment of the autocorrelation. Fig. 9 summarizes the global number of rejections for all combinations of data and tests. Standard tests (especially Pettitt) generally yield higher number of rejections for maximum values. For mean flows, there is not much

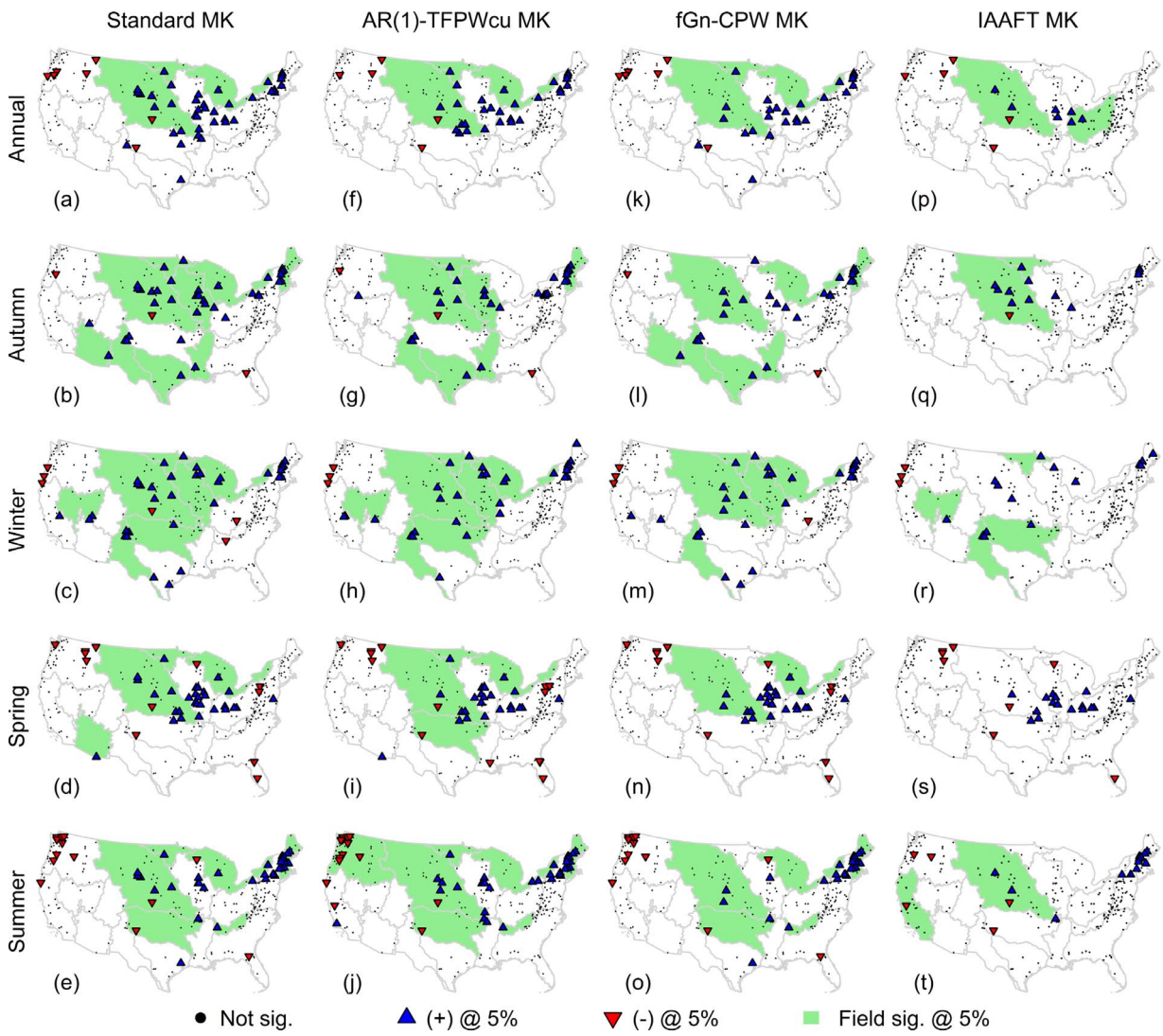

Fig. 6. As for Fig. 5, but for annual average flow values. The same interpretation applies.

difference between standard tests and AR(1)-TFPWcu and fGn-CPW versions. In all cases, IAAFT version yields a dramatic decrease of significant outcomes. Fig. 9 also reports the 95% prediction bands (continuous red lines) of the number of rejections expected over 250 independent trials (i.e., performed tests). Prediction limits correspond to the 2.5th and 97.5th percentiles of a binomial distribution with parameters 250 and 0.05 (i.e., the number of trials and the rate of successes). The diagrams show that IAAFT-based tests yield a number of occurrences which is almost always consistent with what is expected when a purely random experiment with 5% of probability of success is performed. In some cases the outcome is also close to the expected value $\lfloor 250 \cdot 0.05 \rfloor = 12$.

However, the binomial distribution describes the outcome of independent experiments, whereas stream flow time series are spatially correlated, especially in some specific areas reacting in a similar way to common (large-scale) meteo-climatic dynamics. Cross-correlation can be accounted for by simulation; however, a fast computation, which is very accurate in several cases, can be performed by using the beta-binomial distribution (see Appendix B). Considering the 2.5th and 97.5th percentiles of the beta-binomial distribution (red dot-dashed lines in Fig. 9), the number of significant outcomes always fall within the prediction bands for IAAFT-based tests, while the rate of rejection for AR(1)-TFPWcu and fGn-CPW tests is less unexpected than under spatial independence. In this respect, it is worth recalling that the pairwise spatial correlation terms involved in the beta-binomial parametrization refer to the spatial correlation of seasonal/annual averages and maxima; therefore, stronger correlation and over-dispersion is expected if the spatial correlation of daily series is taken into account. It should also be noted that the estimation of the spatial (cross) correlation is affected by temporal correlation (Katz and Brown, 1991; Hamed, 2009a; 2011). Accounting for these aspects results in stronger spatial correlation and further increase of over-dispersion. This implies wider prediction intervals of the number of significant outcomes, and thus even less evidence for global field significance. These results confirm the dramatic impact of the spatial correlation on field significance (Douglas et al., 2000) and the double effect of the autocorrelation on local trend detection and estimation of spatial correlation. If we also consider that autocorrelation is a non optimal measure of dependence implying systematic underestimation of the actual intensity of persistence (as shown above), it can be concluded that we often strongly underestimate the actual spatio-temporal variability of the processes generating the analyzed data.

We stress once again that all the above analyses do not allow any conclusion about the actual stochastic or deterministic nature of the

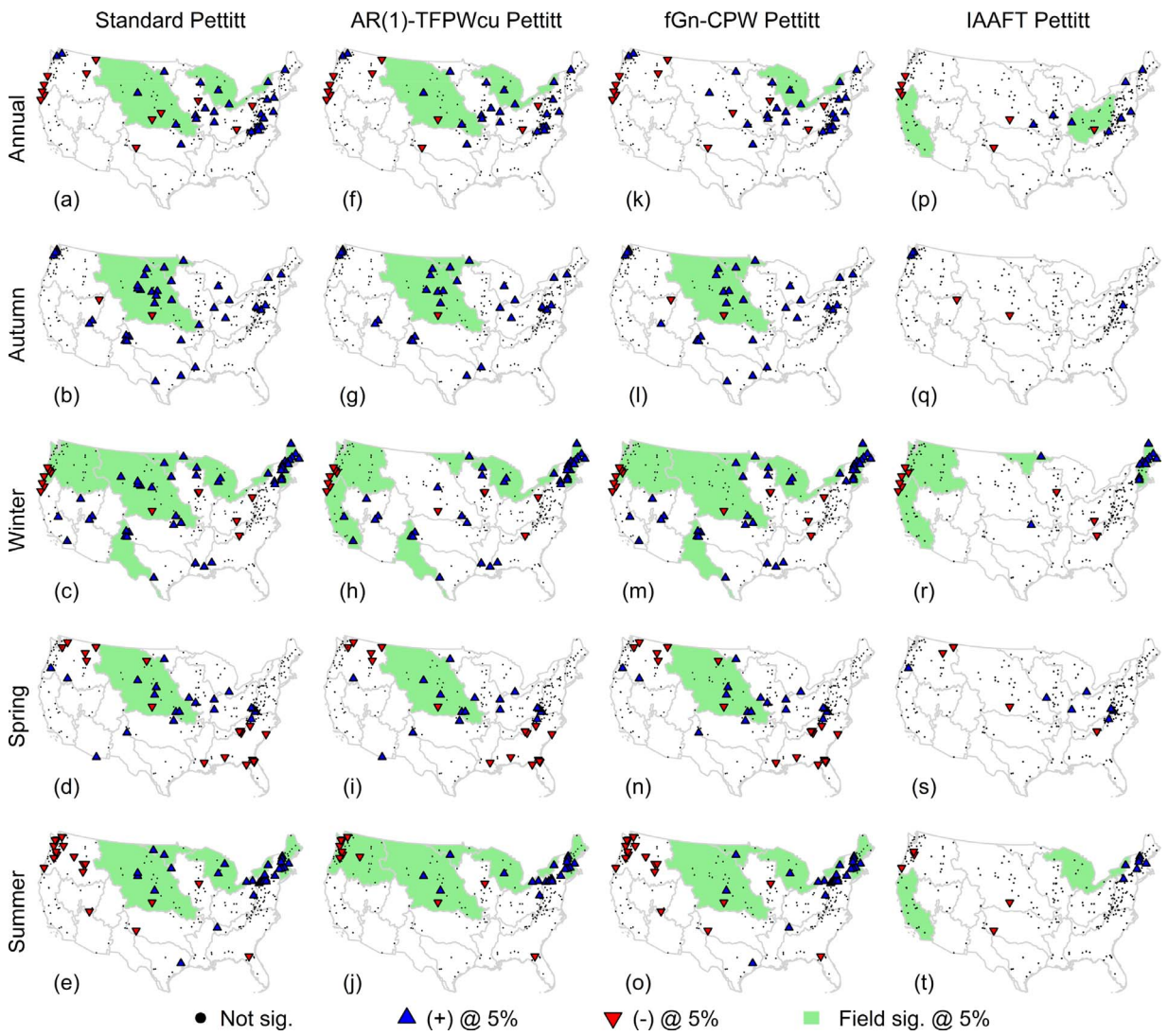

Fig. 7. Spatial distribution sites showing significant abrupt change (at the 5% significance level) according to Pettitt test for the annual flow maxima on an annual and seasonal basis. Panels in each row, from left to right, show results for standard, AR(1)-TFPWcu, fGn-CPW, and IAAFT versions of the Pettitt test. Hydrologic units exhibiting field significance according to the Walker test are also highlighted.

observed trends. Results only tell us that a stochastic stationary representation cannot be excluded, and several results reported in the literature simply depend on the choice of an unrealistic stationary option (*iid*) as a null hypothesis. In other words, observed trends are consistent with both a stationary and nonstationary assumptions, when we choose a suitable and more realistic stationary benchmark. The choice between the two modeling options depends upon a stringent process of attribution supported by additional information and the clear identification of a cause excluding any reasonable alternative explanation. This process goes beyond the application of trend tests or detection of statistically significant correlation with other hydro-meteorological variables by NHSTs, which, we recall, suffer logical flaws and are not devised for exploratory studies.

7. Discussion and conclusions

Published trend analysis in hydro-meteorology often consists of a superficial application of statistical tools, such as NHSTs, as cookbook recipes. This attitude is further spread by the availability of powerful statistical software implementing the state of the art of statistical methodologies developed by statisticians but also questionable procedures developed by practitioners. As already highlighted by von Storch and Zwiers (2003), this approach is particularly dangerous for anyone who is not sufficiently acquainted with the basic concepts of statistics. Moreover, software availability also promotes the tendency to jointly apply combinations of sophisticated techniques (often obscure, redundant or even contrasting with each other) that compound and amplify the problems caused by the indiscriminate use of recipes (von Storch and Zwiers, 2003, p. 97).

Since every hydro-meteorological record is the only available realization or trajectory of the underlying process, this prevents the application of confirmatory analysis because any statement, or null hypothesis, cannot be contested with a statistical test since independent data are unavailable. No statistical test, regardless of its power or complexity, can overcome this problem because it is not a matter of methodology but of available information. The only possible solution is extending this information by additional data going backward (i.e. collecting paleo data sets) or forward (i.e. awaiting the availability of new data to test theories). In this respect, the use of physics-based models can partly help (e.g., Poppick et al., 2017). However, even such

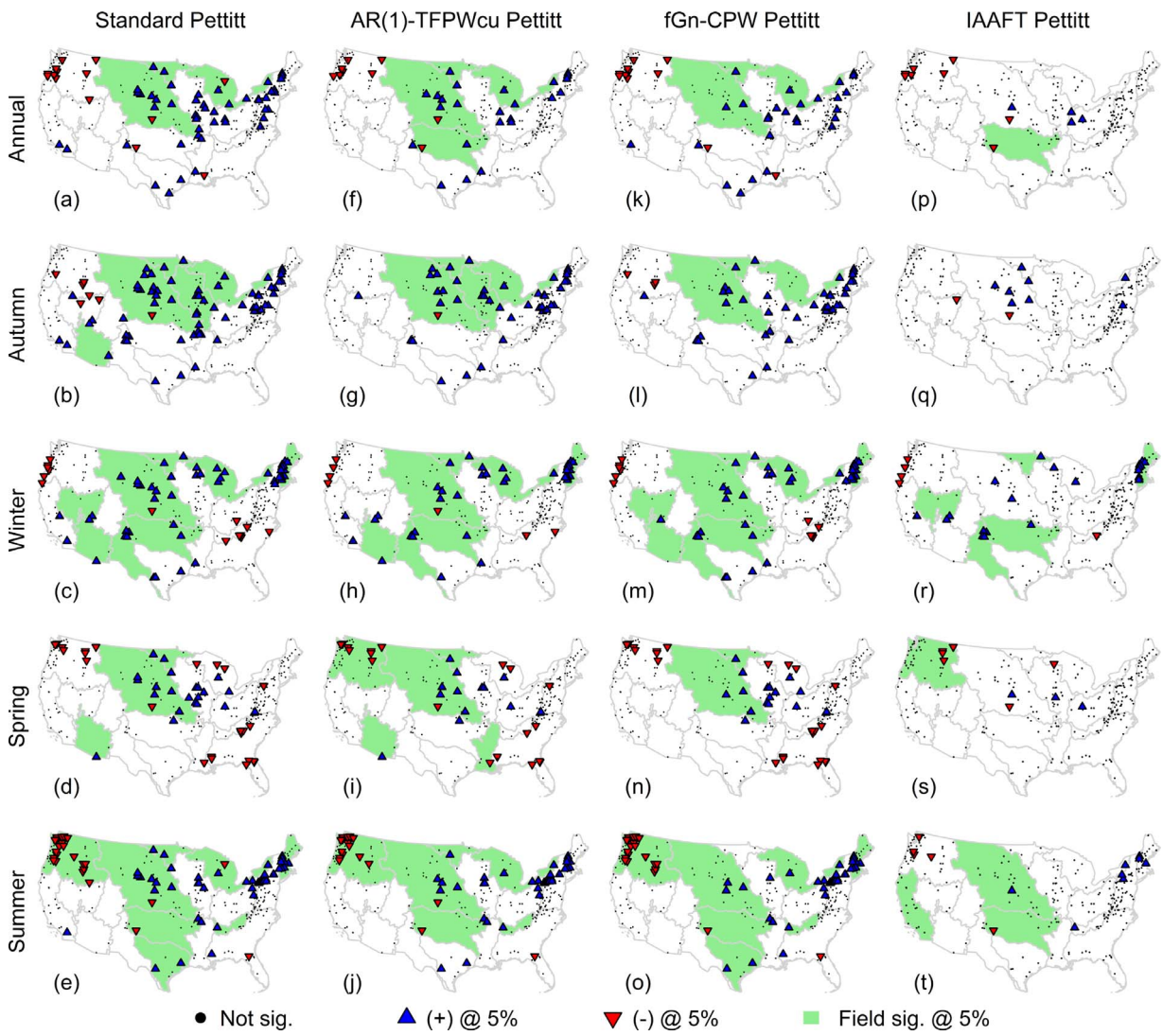

Fig. 8. As for Fig. 7, but for annual average flow values. The same interpretation applies.

models can never be fully validated/disproved as we do not know if they capture all the important properties of the physical processes, and thus the answers given by such models could simply be spurious. A rigorous attribution is required to attempt the identification of unique causes and exclusion of any other plausible alternative.

Based on the above discussion, the actual meaning, interpretation, and limits of trend tests should be recovered. As every statistical test, trend tests can be valuable tools in appropriate contexts, while they cannot be an appropriate tool to infer nonstationarity. They can at most be used as mere tools for preliminary screening whose outcome should be carefully checked and complemented with exogenous information. If a clear physical mechanism related to a predictable evolution of the properties of the process at hand is not identified, we cannot make conclusions about the reason of rejection or lack of rejection, since multiple factors not included in the null and alternative hypotheses can actually play a role.

The case study presented in this paper shows the dramatic difference resulting from the use of trend tests involving different dependence structures, namely, (unrealistic) independence, AR(1) and HK dependence estimated from the target annual summary statistics, and empirical dependence of target variables resulting from that of the parent daily process. We stress that our discussion and these results do not support stationarity versus nonstationarity. Dependence is only used as a more realistic and challenging alternative to deterministic trends in order to show that NHSTs are actually inconclusive when we compare two options yielding similar observations, and a decision concerning the generating mechanism of the studied process and its modeling cannot rely on data only. Since trend tests are also used as automatic criteria to justify the use of nonstationary models for frequency analysis, we discourage such a kind of cookbook recipe usage, as it is inappropriate and might lead to unreliable and paradoxical predictions (Serinaldi and Kilsby, 2015). Of course, nonstationary modeling is still legitimate when it is justified by preliminary attribution relying on additional deductive information on the cause of time-dependent behavior.

It should also be noted that this study does not focus on the interpretation of the physical meaning of probability (e.g., frequentist or Bayesian approach to data analysis) but on the role of inductive and deductive information and reasoning in the scientific inferential procedure. Following Christakos (2011, pp. 177–181), the knowledge bases can be classified between two major categories: *general* (or core), denoted by KB-G, and *specificatory* (or site-specific), KB-S, whereby the former may include scientific theories, natural laws, phenomenological models, cultural relations, and long-established worldviews, while the

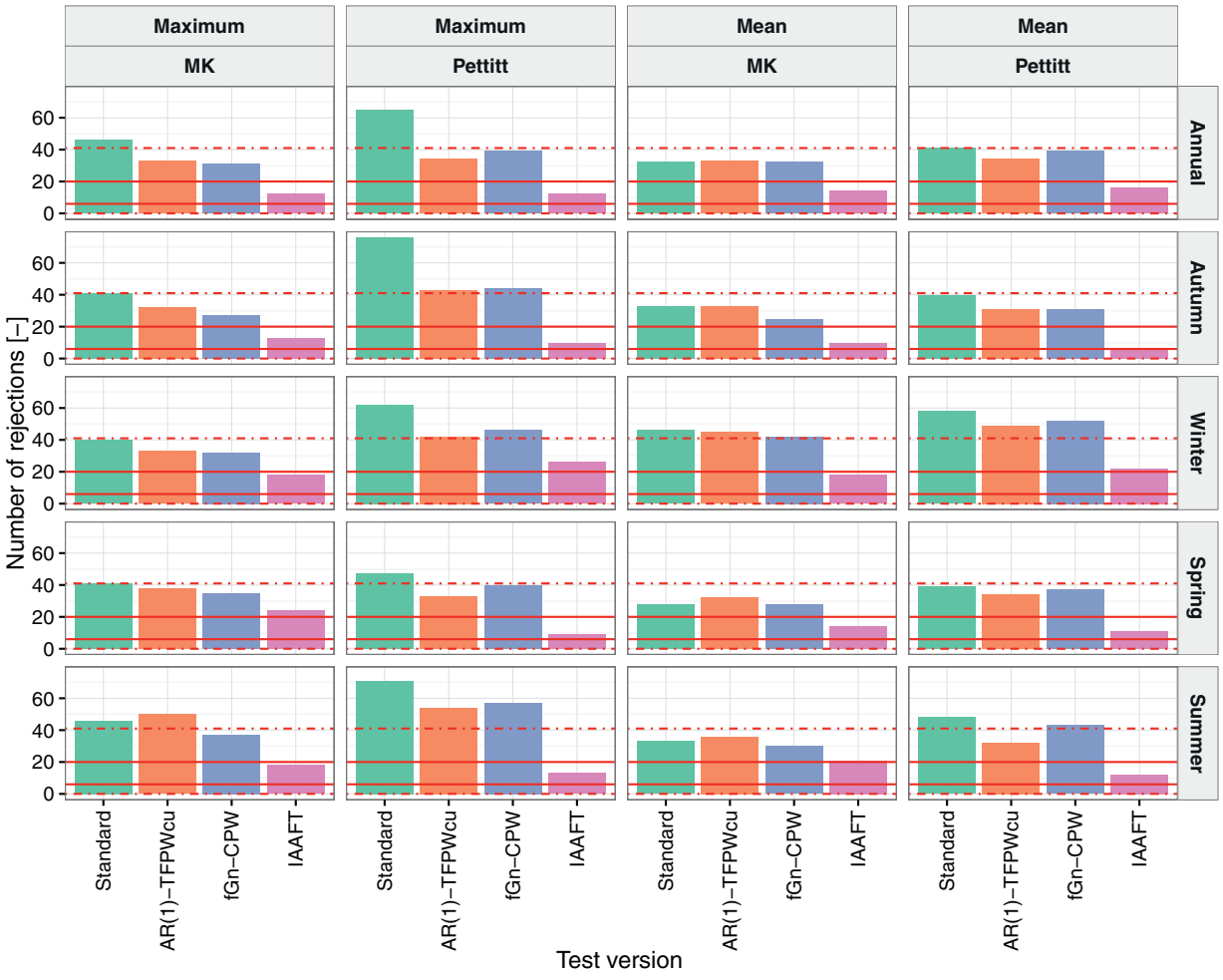

Fig. 9. Overall number of rejections for all tests across the CONUS. Each panel shows the number of rejections for the four versions of a specific test (MK or Pettitt) applied to a given variable (maxima or averages) in a given scale (annual or seasonal). Continuous lines denote the 95% prediction bands for the binomial distribution, while dashed lines indicate the 95% prediction bands for the BB distribution accounting for the spatial correlation.

latter considers different sources of evidence that are tied to the particular local situation and may be not of general validity (i.e., exact and inexact (uncertain) measurements and records). This study shows how a mechanistic application of tests (and models) involving only KB-S is not sufficient to draw conclusions on stationarity or nonstationarity (regardless of the complexity or refinement of the analysis tools) but seem to be the only source of information used in much literature. On the other hand, we emphasize the fundamental role of KB-G that, being based on wisdom of the past, seems to be "irretrievably lost in the postmodern world" (Christakos, 2011, p. 179). In this respect, how KB-G and KB-S are blended is a secondary aspect, although Bayesian framework offers attractive tools to perform a synthesis (Christakos, 2011, pp. 375–380).

Trend tests may be tools for very preliminary screening, for example, in large scale analyses (involving e.g., large number of time series) concerning data quality control, where we are interested in detecting time series affected by systematic instrumental errors. In these cases, we know a priori the mechanism generating 'non-stationarity' (i.e., instrumental malfunction) and its possible effects, and an unambiguous attribution can be made from the knowledge of the instrument specifics. In other cases, rejections can be used to identify primary sub sets deserving further investigations for a clear attribution of the origin of the detected inhomogeneities. However, lack of rejection does not authorize the conclusion that nothing is happening in the remaining subset, which should be further analyzed in any case. Other conclusions in this exploratory context go beyond what the trend tests can tell us.

Acknowledgements

FS and CGK were supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/K013513/1 "Flood MEMORY: Multi–Event Modelling Of Risk and recoverY", and Willis Research Network. Information on HCDN-2009 data used in this study can be found at the web site http://water.usgs.gov/osw/hcdn-2009/, while the data set is freely available at http://waterdata.usgs.gov/nwis/sw. The authors wish to thank three anonymous reviewers for their remarks and constructive criticisms, and Prof. Hans von Storch (Helmholtz-Zentrum Geesthacht, Germany) for his useful comments on an earlier version of this paper. The analyses were performed in R Development Core Team (2016).

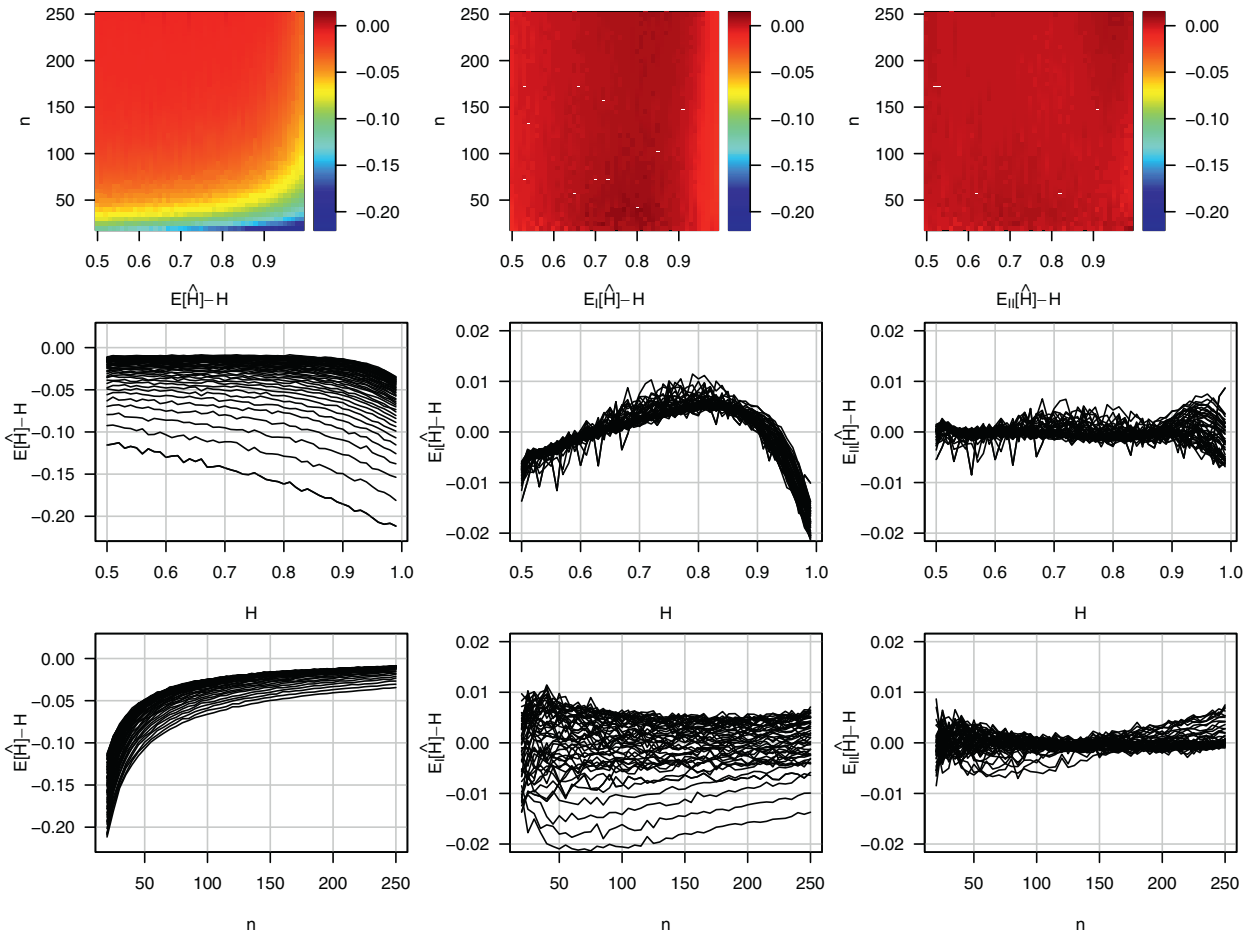

Fig. 10. Level plots and projections showing the original bias of the ML estimator of H (panels in the left side), the effect of the first stage of the bias correction procedure described by Eq. (A.2) (panels in the middle), and the effect of the second stage of the bias correction procedure described by Eq. (A.1) (panels in the right side).

Appendix A. Bias correction for the maximum likelihood estimates of H

Tests involving prewhitening based on fGn correlation structure require the estimation of the Hurst parameter H. It is well known that H is difficult to estimate especially for short time series (Koutsoyiannis and Montanari, 2007). Following Hamed (2008), for fGn-CPW and fGn unconditional prewhitening (fGn-UPW) versions of MK and Pettitt tests (Serinaldi and Kilsby, 2016a), H is estimated by the maximum likelihood (ML) method (McLeod and Hipel, 1978; McLeod et al., 2007), which was proven to be more accurate (Tyralis and Koutsoyiannis, 2011) than other estimators relying on graphical diagnostic plots (see e.g., Serinaldi, 2010). Nonetheless, residual bias affect ML estimates as well. Even though we showed that the direct assessment of persistence on summary statistics, such as seasonal/annual averages and maxima might lead to underestimation of persistence in any case, it is worth adjusting the ML estimation bias in order to reduce the problem. Bias correction was defined based on an extensive Monte Carlo simulation using 10,000 random samples drawn from a fGn process for each combination of H between 0.5 and 0.99 by steps equal to 0.01, and sample size n ranging from 20 to 250 by steps equal to 5. The left side panels on Fig. 10 show a map and two projections of the surface describing the bias of the ML estimator as a function of H and n. The correction involves two stages: first, the dependence of bias of H and n is adjusted by a rational polynomial of the first order, and second, the residual dependence on H (middle column of panels in Fig. 10) is removed by a four-order polynomial. The right side panels in Fig. 10 show that procedure yields a final residual bias lower than 0.01 (i.e., one order of magnitude lower than the original bias) for all combinations of H and n. The resulting bias correction formulas are

$$H_{\text{unb}} = -0.826 + 3.889H_{\text{I}} - 10.94H_{\text{I}}^2 + 10.91H_{\text{I}}^3 - 4.06H_{\text{I}}^4, \tag{A.1}$$

where

$$H_{\text{I}} = \frac{-3.40n + 92.674}{271.609n + 77.280} + \frac{98.470n + 156.590}{96.862n - 240.783}H_{\text{ML}}, \tag{A.2}$$

in which, H_{ML} is the original ML estimate, and is the corrected output.

Appendix B. Beta-binomial distribution

The beta-binomial (BB) distribution is a compound distribution resulting for the ordinary binomial distribution $f(k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $p \in [0, 1]$ represents the constant probability of a success in n trials, when p is assumed to be no longer constant but fluctuating according to a beta

distribution function $f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$, where B denotes beta function, and α and β are two positive shape parameters. The BB density function can be written as (Skellam, 1948)

$$f(k) = \binom{n}{k}\frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha,\beta)}, \tag{B.1}$$

Letting $\pi = \alpha/(\alpha+\beta)$, the mean and variance of beta-binomial can be written as (Ahn and Chen, 1995)

$$\mathrm{E}[K] = n\pi, \tag{B.2}$$

and

$$\mathrm{Var}[K] = n\pi(1-\pi)[1+(n-1)\rho_{\mathrm{BB}}], \tag{B.3}$$

where $\rho_{\mathrm{BB}} = 1/(\alpha+\beta+1)$ is known as the "intra class" or "intra cluster" correlation. Being positive by definition, ρ_{BB} gives rise to over-dispersion as it inflates the variance $n\pi(1-\pi)$ of the original binomial distribution with constant p. On the other hand, ρ_{BB} does not affect the expected value, which is identical for both BB and standard binomial. For correlated experiments, $\rho_{\mathrm{BB}} = \frac{\sum\sum_{j\neq k}\rho_{\mathrm{jk}}}{n(n-1)}$, where ρ_{jk} denotes the pairwise correlation between experiment (site) j and k. In the context of this study, ρ_{BB} is therefore the average cross-correlation between the time series recorded across the CONUS area, which is ≈ 0.04. This is the value used to compute the prediction limits reported in Fig. 9. It is worth noting the impact of ρ_{BB} despite its apparently small value.

References

Ahn, H., Chen, J.J., 1995. Generation of over-dispersed and under-dispersed binomial variates. J. Comput. Graph. Stat. 4 (1), 55–64.

Ahn, K.-H., Palmer, R.N., 2016. Trend and variability in observed hydrological extremes in the United States. J. Hydrol. Eng. 21 (2), 3518–3532.

Archfield, S.A., Hirsch, R.M., Viglione, A., Blöschl, G., 2016. Fragmented patterns of flood change across the United States. Geophys. Res. Lett 2016GL070590.

Ayalew, T.B., Krajewski, W.F., Mantilla, R., Wright, D.B., Small, S.J., 2017. Effect of spatially distributed small dams on flood frequency: Insights from the soap creek watershed. J. Hydrol. Eng. 22 (7). http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0001513. 04017011–1–04017011–11.

Barrett, K.R., Salis, W., 2016. Prevalence and magnitude of trends in peak annual flow and 5-, 10-, and 20-year flows in the northeastern United States. J. Hydrol. Eng. 04016059. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0001474.

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.

Bayazit, M., Önöz, B., 2007. To prewhiten or not to prewhiten in trend analysis? Hydrol. Sci. J. 52 (4), 611–624.

Bayley, G.V., Hammersley, J.M., 1946. The "effective" number of independent observations in an autocorrelated time series. Suppl. J. R. Stat. Soc. 8 (2), 184–197.

Beninger, P.G., Boldina, I., Katsanevakis, S., 2012. Strengthening statistical usage in marine ecology. J. Exper. Marine Biol. Ecol. 426-427, 97–108. http://dx.doi.org/10.1016/j.jembe.2012.05.020.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological) 57 (1), 289–300.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals Stat. 29 (4), 1165–1188.

Berliner, L.M., 1992. Statistics, probability and chaos. Statist. Sci. 7 (1), 69–90. http://dx.doi.org/10.1214/ss/1177011444.

Briggs, W., 2016. Uncertainty: The Soul of Modeling, Probability & Statistics. Springer, New York, USA. http://dx.doi.org/10.1007/978-3-319-39756-6.

Bunde, A., Eichner, J.F., Kantelhardt, J.W., Havlin, S., 2005. Long–term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. Phys. Rev. Lett. 94, 048701.

Busuioc, A., von Storch, H., 1996. Changes in the winter precipitation in Romania and its relation to the large–scale circulation. Tellus A 48 (4), 538–552.

Chandler, R., Scott, M., 2011. Statistical methods for trend detection and analysis in the environmental sciences. Wiley, Chichester, UK.

Cheng, L., AghaKouchak, A., Gilleland, E., Katz, R.W., 2014. Non-stationary extreme value analysis in a changing climate. Climatic Change 127 (2), 353–369.

Christakos, G., 2011. Integrative Problem-Solving in a Time of Decadence. Springer, London, UK.

Clarke, R.T., 2010. On the (mis)use of statistical methods in hydro-climatological research. Hydrol. Sci. J. 55 (2), 139–144. http://dx.doi.org/10.1080/02626661003616819.

Cohen, J., 1994. The Earth is round ($p < .05$). American Psychologist 997–1003.

Cooley, D., 2013. Return periods and return levels under climate change. In: AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., Sorooshian, S. (Eds.), Extremes in a Changing Climate. Water Science and Technology Library 65. Springer Netherlands, pp. 97–114. http://dx.doi.org/10.1007/978-94-007-4479-0_4.

Daniel, J.S., Portmann, R.W., Solomon, S., Murphy, D.M., 2012. Identifying weekly cycles in meteorological variables: The importance of an appropriate statistical analysis. J. Geophys. Res.: Atmos. 117 (D13), D13203.

Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 74 (366), 427–431.

Do, H.X., Westra, S., Leonard, M., 2017. A global-scale investigation of trends in annual maximum streamflow. J. Hydrol. 552, 28–43.

Douglas, E., Vogel, R., Kroll, C., 2000. Trends in floods and low flows in the United States: impact of spatial correlation. J. Hydrol. 240 (1–2), 90–105.

Eco, U., 1976. Peirce's notion of interpretant. MLN 91 (6), 1457–1472.

Eichner, J.F., Kantelhardt, J.W., Bunde, A., Havlin, S., 2011. The statistics of return intervals, maxima, and centennial events under the influence of long-term correlations. In: Kropp, J., Schellnhuber, H.-J. (Eds.), In Extremis. Springer, Berlin, Heidelberg, pp. 2–43.

Ellison, A.M., Gotelli, N.J., Inouye, B.D., Strong, D.R., 2014. P values, hypothesis testing, and model selection: it's déjà vu all over again. Ecology 95 (3), 609–610.

Ferguson, C.R., Villarini, G., 2012. Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States. J. Geophys. Res. Atmosph. 117 (D5), D05123.

Fisher, R.A., 1929. Tests of significance in harmonic analysis. Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci. 125 (796), 54–59.

Fisher, R.A., 1935. The design of experiments, ninth. Oliver and Boyd, Edinburgh, Scotland.

Flueck, J.A., Brown, T.J., 1993. Criteria and methods for performing and evaluating solar–weather studies. J. Climate 6 (2), 373–385. http://dx.doi.org/10.1175/1520-0442(1993)006<0373:CAMFPA>2.0.CO;2.

Franzke, C., 2013. A novel method to test for significant trends in extreme values in serially dependent time series. Geophys. Res. Lett. 40 (7), 1391–1395.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance: How Probability Changed Science and Everyday Life, New York, USA.

Gill, J., 1999. The insignificance of null hypothesis significance testing. Political Res. Q. 52 (3), 647–674. http://dx.doi.org/10.2307/449153.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. 31 (4), 337–350.

Guerreiro, S.B., Kilsby, C.G., Serinaldi, F., 2014. Analysis of time variation of rainfall in transnational basins in Iberia: abrupt changes or trends? Int. J. Climatol. 34 (1), 114–133.

Hamed, K.H., 2008. Trend detection in hydrologic data: The Mann–Kendall trend test under the scaling hypothesis. J. Hydrol. 349 (3–4), 350–363.

Hamed, K.H., 2009a. Effect of persistence on the significance of Kendall's tau as a measure of correlation between natural time series. Eur. Phys. J. Special Top. 174 (1), 65–79.

Hamed, K.H., 2009b. Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. J. Hydrol. 368 (1–4), 143–155.

Hamed, K.H., 2011. The distribution of Kendall's tau for testing the significance of cross-correlation in persistent data. Hydrol. Sci. J. 56 (5), 841–853.

Hamed, K.H., Rao, A.R., 1998. A modified Mann–Kendall trend test for autocorrelated data. J. Hydrol. 204 (1–4), 182–196.

Hasselmann, K., 1993. Optimal fingerprints for the detection of time-dependent climate change. J. Climate 6 (10), 1957–1971. http://dx.doi.org/10.1175/1520-0442(1993)006<1957:OFFTDO>2.0.CO;2.

Hasselmann, K., 1997. Multi-pattern fingerprint method for detection and attribution of climate change. Climate Dyn. 13 (9), 601–611. http://dx.doi.org/10.1007/s003820050185.

Hirsch, R.M., Ryberg, K.R., 2012. Has the magnitude of floods across the USA changed with global CO2 levels? Hydrol. Sci. J. 57 (1), 1–9.

Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press.

Johnson, D.H., 1999. The insignificance of statistical significance testing. J. Wildlife Manag. 63 (3), 763–772. http://dx.doi.org/10.2307/3802789.

Kalra, A., Piechota, T.C., Davies, R., Tootle, G.A., 2008. Changes in U.S. streamflow and western U.S. snowpack. J. Hydrol. Eng. 13 (3), 156–163.

Katz, R.W., 1988. Statistical procedures for making inferences about climate variability. J. Climate 1 (11), 1057–1064.

Katz, R.W., 2002. Sir Gilbert Walker and a Connection between El Nio and Statistics. Stat.

Sci. 17 (1), 97–112.
Katz, R.W., Brown, B.G., 1991. The problem of multiplicity in research on teleconnections. Int. J. Climatol. 11 (5), 505–513.
Kendall, M.G., 1970. Rank Correlation Methods, 4th. Griffin, London.
Khaliq, M.N., Ouarda, T.B.M.J., Gachon, P., 2009a. Identification of temporal trends in annual and seasonal low flows occurring in Canadian rivers: the effect of short- and long-term persistence. J. Hydrol. 369 (12), 183–197.
Khaliq, M.N., Ouarda, T.B.M.J., Gachon, P., Sushama, L., St-Hilaire, A., 2009b. 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.
Kingston, D.G., McGregor, G.R., Hannah, D.M., Lawler, D.M., 2007. Large-scale climatic controls on New England river flow. J. Hydrometeorol. 8 (3), 367–379.
Koutsoyiannis, D., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. Hydrol. Sci. J.l 48 (1), 3–24.
Koutsoyiannis, D., 2010. HESS Opinions "A random walk on water". Hydrol. Earth Syst. Sci. 14 (3), 585–601.
Koutsoyiannis, D., 2011. Hurst-Kolmogorov dynamics and uncertainty. J. Am. Water Resourc. Assoc. 47 (3), 481–495.
Koutsoyiannis, D., 2016. Generic and parsimonious stochastic modelling for hydrology and beyond. Hydrol. Sci. J. 61 (2), 225–244.
Koutsoyiannis, D., Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. Water Resources Res. 43 (5), W05429.
Koutsoyiannis, D., Montanari, A., 2015. Negligent killing of scientific concepts: the stationarity case. Hydrol. Sci. J. 60 (7-8), 1174–1183.
Kugiumtzis, D., 1999. Test your surrogate data before you test for nonlinearity. Phys. Rev. E 60, 2808–2816.
Kulkarni, A., von Storch, H., 1995. Monte Carlo experiments on the effect of serial correlation on the Mann–Kendall test of trend. Meteorologische Zeitschrift 4 (2), 82–85.
Kundzewicz, Z.W., Robson, A.J., 2004. Change detection in hydrological records–a review of the methodology. Hydrol. Sci. J. 49 (1), 7–19.
Kwiatkowski, D., Phillips, P., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J. Econ. 54 (1–3), 159–178.
Laplace, P.S.M., 1814. A Philosophical Essay on Probabilities, 1. Dover Publications, New York, US. 1995, translated from the sixth French edition.
Lettenmaier, D.P., Wood, E.F., Wallis, J.R., 1994. Hydro-climatological trends in the continental United States, 1948-88. J. Climate 7 (4), 586–607.
Levine, T.R., Weber, R., Hullett, C., Park, H.S., Lindsey, L.L.M., 2008. A critical assessment of null hypothesis significance testing in quantitative communication research. Human Commun. Res. 34 (2), 171–187. http://dx.doi.org/10.1111/j.1468-2958.2008.00317.x.
Lins, H., 2012. Hydro-Climatic Data Network 2009 (HCDN-2009). U.S. Geological Survey Fact Sheet 2012–3047.
Lins, H.F., Cohn, T.A., 2011. Stationarity: wanted dead or alive? J. Am. Water Resourc. Assoc. 47 (3), 475–480.
Lins, H.F., Slack, J.R., 1999. Streamflow trends in the United States. Geophys. Res. Lett. 26 (2), 227–230.
Livezey, R.E., Chen, W.Y., 1983. Statistical field significance and its determination by monte carlo techniques. Monthly Weather Rev. 111 (1), 46–59. http://dx.doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.
Lombardo, F., Volpi, E., Koutsoyiannis, D., Papalexiou, S.M., 2014. Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology. Hydrol. Earth Syst. Sci. 18 (1), 243–255.
Lorenz, E.N., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20 (2), 130–141. http://dx.doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
Luke, A., Vrugt, J.A., AghaKouchak, A., Matthew, R., Sanders, B.F., 2017. Predicting nonstationary flood frequencies: evidence supports an updated stationarity thesis in the United States. Water Resources Research 53.
Mallakpour, I., Villarini, G., 2015. The changing nature of flooding across the central United States. Nat. Climate Change 5 (3), 250–254.
Mallakpour, I., Villarini, G., 2016. Analysis of changes in the magnitude, frequency, and seasonality of heavy precipitation over the contiguous USA. Theor. Appl. Climatol. 1–19.
Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. Macmillan, New York, US.
Mann, H.B., 1945. Nonparametric tests against trend. Econometrica 13 (3), 245–259.
Matalas, N.C., Sankarasubramanian, A., 2003. Effect of persistence on trend detection via regression. Water Resources Res. 39 (12), 1342. 1342
McBride, G.B., Loftis, J.C., Adkins, N.C., 1993. What do significance tests really tell us about the environment? Environ. Mana. 17 (4), 423–432.
McCabe, G.J., Wolock, D.M., 2002. A step increase in streamflow in the conterminous United States. Geophys. Res. Lett. 29 (24), 38.1–38.4. 2185
McCuen, R., 2003. Modeling Hydrologic Change: Statistical Methods. CRC Press.
McLeod, A.I., Hipel, K.W., 1978. Preservation of the rescaled adjusted range: 1. A reassessment of the Hurst Phenomenon. Water Resources Res. 14 (3), 491–508.
McLeod, A.I., Yu, H., Krougly, Z.L., 2007. Algorithms for linear time series analysis: with R package. J. Stat. Softw. 23 (5), 1–26.
Meehl, P.E., 1997. The Problem is Epistemology, Not Statistics: Replace Significance tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. Erlbaum, Mahwah, NJ (USA), pp. 393–425.
Merz, B., Vorogushyn, S., Uhlemann, S., Delgado, J., Hundecha, Y., 2012. HESS Opinions "More efforts and scientific rigour are needed to attribute trends in flood time series". Hydrol. Earth Syst. Sci. 16 (5), 1379–1387.
Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., Dettinger, M.D., Krysanova, V., 2015. On critiques of "Stationarity is dead: Whither water management?". Water Resources Res.h 51 (9), 7785–7789.
Mitchell, J.F.B., Karoly, D.J., Hegerl, G.C., Zwiers, F.W., Allen, M.R., Marengo, J., 2001. Detection of Climate Change and Attribution of Causes. Cambridge University Press, Cambridge, UK, pp. 393–425.
Myers, D.E., 1989. To be or not to be... stationary? That is the question. Math. Geol. 21 (3), 347–362.
Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. A: Math., Phys. Eng. Sci. 231 (694–706), 289–337. http://dx.doi.org/10.1098/rsta.1933.0009.
Nicholls, N., 2001. commentary and analysis: The insignificance of significance testing. Bull. Am. Meteorol. Soc. 82 (5), 981–986. http://dx.doi.org/10.1175/1520-0477(2001)082<0981:CAATIO>2.3.CO;2.
Nilsen, I.B., Stagge, J.H., Tallaksen, L.M., 2016. A probabilistic approach for attributing temperature changes to synoptic type frequency. Int. J. Climatol. http://dx.doi.org/10.1002/joc.4894.
Nuzzo, R., 2014. Statistical errors: P-values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 506 (7487), 150–152.
Olsen, J.R., Lambert, J.H., Haimes, Y.Y., 1998. Risk of extreme events under nonstationary conditions. Risk Anal. 18 (4), 497–510. http://dx.doi.org/10.1111/j.1539-6924.1998.tb00364.x.
Ouarda, T.B.M.J., El-Adlouni, S., 2011. Bayesian nonstationary frequency analysis of hydrological variables. J. Am. Water Resources Assoc. 47 (3), 496–505.
Papoulis, A., 1991. Probability, Random Variables, and Stochastic Processes, Third. 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 Resources Manag.t 30 (13), 4649–4665.
Pettitt, A.N., 1979. A non-parametric approach to the change-point problem. J. R. Stat. Soc. Series C (Appl. Stat.) 28 (2), 126–135.
Pollard, P., Richardson, J.T., 1987. On the probability of making Type I errors. Psychol. Bull. 102 (1), 159–163. http://dx.doi.org/10.1037/0033-2909.102.1.159.
Poppick, A., Moyer, E.J., Stein, M.L., 2017. Estimating trends in the global mean temperature record. Adv. Stat. Climatol., Meteorol. Oceanogr. 3 (1), 33–53. http://dx.doi.org/10.5194/ascmo-3-33-2017.
Prosdocimi, I., Kjeldsen, T.R., Svensson, C., 2014. Non-stationarity in annual and seasonal series of peak flow and precipitation in the UK. Nat. Hazards Earth Syst. Sci. 14 (5), 1125–1144.
R Development Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0
Rice, J.S., Emanuel, R.E., Vose, J.M., 2016. The influence of watershed characteristics on spatial patterns of trends in annual scale streamflow variability in the continental U.S. J. Hydrol. 540, 850–860.
Rice, J.S., Emanuel, R.E., Vose, J.M., Nelson, S.A.C., 2015. Continental U.S. streamflow trends from 1940 to 2009 and their relationships with watershed spatial characteristics. Water Resour. Res. 51 (8), 6262–6275.
Rootzén, H., Katz, R.W., 2013. Design Life Level: Quantifying risk in a changing climate. Water Resour. Res. 49 (9), 5964–5972.
Rougé, C., Ge, Y., Cai, X., 2013. Detecting gradual and abrupt changes in hydrological records. Adv. Water Resour. 53, 33–44.
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.
Salas, J.D., Obeysekera, J., 2014. Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. J. Hydrol. Eng. 19 (3), 554–568. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000820.
Schreiber, T., Schmitz, A., 1996. Improved surrogate data for nonlinearity tests. Phys. Rev. Lett. 77, 635–638.
Schreiber, T., Schmitz, A., 2000. Surrogate time series. Phys. D: Nonlinear Phen. 142 (3–4), 346–382.
Serinaldi, F., 2010. Use and misuse of some Hurst parameter estimators applied to stationary and non-stationary financial time series. Physica A: Stat. Mech. Appl. 389 (14), 2770–2781.
Serinaldi, F., 2015. Dismissing return periods!. Stoch. Environ. Res. Risk Assess. 29 (4), 1179–1189. http://dx.doi.org/10.1007/s00477-014-0916-1.
Serinaldi, F., Kilsby, C.G., 2015. Stationarity is undead: Uncertainty dominates the distribution of extremes. Adv. Water Resourc. 77, 17–36.
Serinaldi, F., Kilsby, C.G., 2016a. The importance of prewhitening in change point analysis under persistence. Stochastic Environmental Research and Risk Assessment 30 (2), 763–777.
Serinaldi, F., Kilsby, C.G., 2016b. Irreversibility and complex network behavior of stream flow fluctuations. Phys. A: Stat. Mech. Appl. 450, 585–600.
Serinaldi, F., Kilsby, C.G., 2016c. Understanding persistence to avoid underestimation of collective flood risk. Water 8 (4), 152.
Serinaldi, F., Lombardo, F., 2017. General simulation algorithm for autocorrelated binary processes. Phys. Rev. E 95, 023312. http://dx.doi.org/10.1103/PhysRevE.95.023312.
Sivakumar, B., 2016. Chaos in Hydrology: Bridging Determinism and Stochasticity. Springer, Dordrecht, Netherlands.
Skellam, J.G., 1948. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. J. R. Stat. Soc. Ser. B (Methodol.) 10 (2), 257–261.
von Storch, H., Zwiers, F.W., 2003. Statistical Analysis in Climate Research. Cambridge University Press, New York, US.
Tananaev, N.I., Makarieva, O.M., Lebedeva, L.S., 2016. Trends in annual and extreme

flows in the Lena River basin, Northern Eurasia. Geophys. Res. Lett. 43 (20), 10764–10772.
Tramblay, Y., El Adlouni, S., Servat, E., 2013. Trends and variability in extreme precipitation indices over Maghreb countries. Nat. Hazards Earth Syst. Sci. 13 (12), 3235–3248.
Tyralis, H., Koutsoyiannis, D., 2011. Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process. Stoch. Environ. Res. Risk Assess. 25 (1), 21–33.
Venema, V., Ament, F., Simmer, C., 2006a. A stochastic iterative amplitude adjusted fourier transform algorithm with improved accuracy. Nonlinear Process. Geophys. 13 (3), 321–328.
Venema, V., Bachner, S., Rust, H.W., Simmer, C., 2006b. Statistical characteristics of surrogate data based on geophysical measurements. Nonlinear Processes in Geophysics 13 (4), 449–466.
Viglione, A., Merz, B., Viet Dung, N., Parajka, J., Nester, T., Blschl, G., 2016. Attribution of regional flood changes based on scaling fingerprints. Water Resourc. Res. 52 (7), 5322–5340.
Villarini, G., Serinaldi, F., Smith, J.A., Krajewski, W.F., 2009a. On the stationarity of annual flood peaks in the continental United States during the 20th century. Water Resour. Res. 45 (8), W08417.
Villarini, G., Smith, J.A., 2010. Flood peak distributions for the eastern United States. Water Resour. Res. 46 (6), W06504.
Villarini, G., Smith, J.A., Baeck, M.L., Krajewski, W.F., 2011a. Examining flood frequency distributions in the Midwest U.S. J. Am. Water Resourc. Assoc. 47 (3), 447–463.
Villarini, G., Smith, J.A., Serinaldi, F., Bales, J., Bates, P.D., Krajewski, W.F., 2009b. Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. Adv. Water Resourc. 32 (8), 1255–1266. http://dx.doi.org/10.1016/j.advwatres.2009.05.003.
Villarini, G., Smith, J.A., Serinaldi, F., Ntelekos, A.A., 2011b. Analyses of seasonal and annual maximum daily discharge records for central Europe. J. Hydrol. 399 (3-4), 299–312.
Vogel, R.M., Yaindl, C., Walter, M., 2011. Nonstationarity: flood magnification and recurrence reduction factors in the United States. J. Am. Water Resourc. Assoc. 47 (3), 464–474.
Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., Koutsoyiannis, D., 2015. One hundred years of return period: Strengths and limitations. Water Resourc. Res. 51 (10), 8570–8585.
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.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: context, process, and purpose. Am. Stat. 70 (2), 129–133.
Wilks, D.S., 1997. Resampling hypothesis tests for autocorrelated fields. J. Climate 10 (1), 65–82.
Wilks, D.S., 2006. On "field significance" and the false discovery rate. J. Appl. Meteorol. Climatol. 45 (9), 1181–1189.
Yevjevich, V., 1974. Determinism and stochasticity in hydrology. J. Hydrol. 22 (3), 225–238. https://doi.org/10.1016/0022-1694(74)90078-X.
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.
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., Wang, C.-Y., 2004. The Mann–Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. Water Resour. Manag. 18 (3), 201–218.