Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-576-SC1, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management" *by* Louise J. Slater et al.

Francesco Serinaldi

francesco.serinaldi@ncl.ac.uk

Received and published: 15 December 2020

"*Are hydroclimatic extremes stationary or nonstationary?*" Well, to provide a scientific answer to this question (...and any scientific question), we should start from agreeing about a formal and rigorous definition of stationarity and nonstationarity, and therefore logically deduce its consequences.

The quite surreal debate generated by the foregoing question is actually a sort of Babelian confusion affecting the hydro-climatological community and resulting from the fact that different people assign different meaning to the same words. The Authors themselves seem not to recognize that they use the terms 'nonstationarity' rather

loosely throughout the paper with different meanings, overlooking the (il)logical and practical (negative) consequences of basing a discussion on vague concepts.

Since the Authors are so kind to cite my work both in the introduction and conclusion to compare (or contrast) Khinchin's formal (mathematical) definition of stationarity used in my works with the the so-called 'functional' definition given by my friend Gabriele, please let me clarify why the latter is not a definition and it is technically and practically unusable in the context of statistical inference.

To do that, it is worth recalling the basic concepts of statistical inference, which seem to be missed by most people playing with hydro-meteorological data. I will not refer to (too) theoretical statistics references, but to von Storch and Zwiers (2003), which is one of the most cited book on statistics applied to climatology, but also the less read by those who cite it, I guess. Indeed, the book start with this caveat:
"*Cookbook recipes for a variety of standard statistical situations are not offered by this book because they are dangerous for anyone who does not understand the basic concepts of statistics.*"
Concerning statistical inference, von Storch and Zwiers (2003, p. 69) state: *"The word inference is central in statistical analysis. A dictionary definition of inference rephrases 'to infer' as 'to conclude by reasoning from something known or assumed.' A broad definition of statistical inference could be 'the procedure that involves extracting information from data about the process underlying the observations.' There are two central steps in this process.*

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

2. The observations are analysed in the context of the adopted statistical model"

This is consistent with the definitions given by statisticians and probabilists, such as Aitken (1947) or Cramer (1946), and summarized for instance in Serinaldi et al. (2020; Appendix 1), and references therein.

How do these 'boring' and apparently detached concepts apply to our discussion? Well, all the statistical methods mentioned in the manuscript (from trend tests to GAMLSS, Bayesian techniques, etc.) are devised to draw conclusions on the **population properties** starting from the study of the **sample properties**, whose interpretation is bounded to the features of the "*models implicit in the chosen statistical method*". Let me further clarify: when we apply a trend test (e.g. standard MK), we study the data to draw conclusions on a specific property of the population, i.e. the stationarity of the underlying process according to Khinchin's mathematical definition, and the results must be interpreted in light of the assumed (zero) model, which is usually 'independent and identically distributed random variables' (iid rv's). This explains why standard MK (without corrections for serial dependence) cannot be used in presence of serial correlation, which indeed prevents the correct interpretation of standard MK results, as it is in conflict with the model assumption of independence of the test itself.

If we accept the 'functional' definition (which is not a definition as it cannot be translated into a usable mathematical formalization), we have two main problems: (i) it cannot be used to build any statistical inference because the "*shifts in the probability distribution of a given dataset*" are only sampling fluctuation, and we cannot deduce any statistical method from them, and (ii) the existing methods (tests and estimation methods) do not provide information about this supposed type of stationarity, as they are devised to deal with **population** stationarity, not sample shifts. In trend analysis, sample shifts/fluctuations are used to infer population properties, and their results must be interpreted in light of the (theoretical/population) models implicit in the method (such as iid model, or serially correlated identically distributed rv's, etc.). Similarly, when we

apply a test for the difference between two means, we use sample means $\bar{x}_1$ and $\bar{x}_2$ to infer the differences between the population means μ_1 and μ_2. The null hypothesis is $\mu_1 - \mu_2 = 0$ and not $\bar{x}_1 - \bar{x}_2 = 0$; different symbols are not used to make books esthetically fancy via elegant Greek letters, but to denote different things with very different meaning. Checking stationarity in a supposed 'functional' world is something like using sample means to infer the differences between the sample means with test devised for the population means!

Now, if the Authors do not like Khinchin's definition of stationarity, they can replace it, but only with another formal mathematical definition that enables the deduction of mathematical tools (tests and estimation procedures), which can therefore be used to analyse data. Using an analogy, we can build geometric theories different from the Euclidean one, but they has to be mathematical/theoretical. In this context, Gabriele's definition sounds like: 'As I do not like the mathematical definition of 'rectangle', I define a rectangle as the surface of my desk'. With this 'functional' definition, we cannot deduce any theory. What is the consequence? Well, we cannot deduce the formulas to compute the perimeter and area of an objects with rectangular shape because the 'functional' definition does not allow us deducing them, and because any object different from Gabriele's desk is not a rectangle by 'functional' definition!

Based on the above remarks also the following sentence makes little sense:
"*the issue is not whether observations arise from a long-term excursion from some underlying stationary process but rather whether the probability distribution of future (events) will resemble the distribution that is obtained from fitting a probability distribution to observations over a historical record"*
In fact, the (population) probability distribution of future (events) will be identical to the distribution that is obtained from fitting a probability distribution to observations over a historical record **if and only if** observations come from a long-term excursion (whatever 'excursion' means) from some underlying stationary process. I used 'identical to' instead of 'resemble' because the latter denotes the confusion existing in the original

sentence between empirical distribution and theoretical/population distribution.
Fitting a parametric distribution is not only a numerical exercise (minimizing some distance, metric, or criterion), but means inferring the hypothetical 'true' population distribution under the assumption that the observations are enough representative of the entire population. Once we get the fitted model, the (theoretical) distribution of the future data is that one by (mathematical) definition. In his sentence, Gabriele confuses the empirical distributions of subsamples (which can obviously fluctuate) with the population distribution. In doing this, he also reduces the inference of the population distribution to the a simple numerical exercise, where the population distribution is no longer the model generating past and future data, but only an analytical curve smoothing the empirical distributions and potentially different for every new sample. This approach and interpretation are totally uninformative in the context of statistical inference.
Recall that under nonstationarity (in formal Khinchin's sense), each observation comes from its own distribution: this means that the sample is not representative of any specific population, as we have as many populations as the number of observations, and we have no idea if the single observation falls in the body or in the tail of the distribution of its own population (these concepts are better discussed in Serinaldi et al. (2018; Sec. 4.2), which I invite to read more carefully).

The ambiguous use of the term '(non)stationarity' without a technical definition confused also Referee #1. Commenting P3L13, they write "*'Climate is non-stationary by definition': or does it depend on the record length as you have stated? That is, if you pick the correct length of record it will be stationary? This statement probably just needs explaining to fit in the context here.*". Let me try to clarify: Climate cannot be either stationary or nonstationary 'by definition' because (i) climate is not a mathematical process, and (ii) there is not any formal definition for climate. What people call climate is a sequence of observations coming from an unknown natural process; only the mathematical models used to describe this unknown process can be stationary, nonstationary, linear, nonlinear, etc. Once a model option is chosen (based on *a priori*

and *a posteriori* criteria), that model describes the entire process (past and future, and all possible states allowed by the model itself). And if the model captures all the key dynamics of the unknown process, therefore, it will reproduce what the Authors loosely call (apparent/local/functional) 'nonstationarities' or (apparent/local/functional) 'stationarities', which are actually local/transient steady or unsteady states assumed by the process at given spatio-temporal scales. Such a kind of behavior (local trends, persistence, clustering, scaling, seasonal cycles, etc.) can result from both (cyclo)stationary, nonstationary, linear or nonlinear models (stationary processes can produce signals that are much more complex than the iid sequence shown in Fig. 1a, as mentioned by Referee #2). The choice of the best modeling strategy depends on multiple (theoretical and empirical) considerations. If the chosen model is (non)stationary (in Khinchin's sense), we can only say that the climate behavior can be reasonably and provisionally modeled by that (non)stationary model, until further analysis of the phenomenon suggests updates or revision.

To summarize:

- If we assume a unique formal definition of stationarity and we agree on it, nonstationarity can only be assumed *a priori*, and cannot be detected from data only.
- If we decide to introduce a different definition, this must be rigorous, unambiguous, and mathematically clear and well defined, in order to avoid confusion and allow the deduction of a full set of consistent formal tools enabling consistent inference. This is fundamental to guarantee that people talk about the same object.
- If we accept the rationale and logic of statistical inference, we cannot introduce spurious ill-defined concepts confusing sample and population properties, and we must interpret the results according to the models implicit in the statistical methods. 'Functional' shortcuts make little sense, if we want to maintain the discussion in a scientific context.

If the foregoing discussion looks too 'philosophical' or 'a matter of semantics', let me further discuss the practical consequences of a bit 'too loose' approach to statistical inference.

One of the available methods to (attempt to) deal with the effects of serial correlation on standard MK test is the so-called trend-free prewhitening (TFPW; Yue et al., 2002). Along the years, several people recognized that TFPW always generated results close to the original MK tests, and these results disagree with those yielded by other methods devised for correcting variance inflation due to serial correlation. They attempted various (incorrect) interpretations, such as the supposed effect of correlation structures more persistent than first-order autoregressive (supposedly) removed by TFPW. However, the actual cause of this behavior is that the TFPW procedure does not perform a proper prewhitening and does not preserve the required nominal significance level, thus giving an incorrect rejection rate under persistence similar to that of the original MK test. This explains why TFPW almost always yields results similar to standard MK. Serinaldi and Kilsby (2016) discussed such technical flaws of TFPW by showing the mathematical incosistencies of its formulation. Nonetheless, TFPW is still applied (see e.g. Ayers et al. (2019), for one of the latest applications), and it will be again and again. As a result, there are tens of published papers in which the hypothesis of non-stationarity is supported by a flawed technique. Those who apply TFPW neglecting its theoretical inconsistencies, actually perform a trend test knowing neither the actual significance level nor the power of the test!
This happens because too many people miss the structure of the statistical inference. TFPW is only an example. Many other methods listed in the manuscript are routinely applied without knowing their theoretical properties. Thus, the most imaginary conclusions are drawn because results are not interpreted in light of those properties, and numerical artifacts are interpreted as supposed physical properties.
So, when the Authors, in their conclusion, talk about Type I and Type II errors in relation to the 'functional' definition of (non)stationarity, they should bear in mind that assuming this definition means discarding the formal definition, and they use the 'functional'

definition to justify the application of methods that however focus on the type of formal stationarity that they dismissed! And they also miss that in the 'functional' world, Type I and Type II are not even defined!

The list of questionable statistical methods characterizing hydro-meteorological literature would be long. For example:

- What the Authors call "*Soft statistical attribution approaches*" is nothing but the so-called 'data-dredging', i.e. a practice widely deprecated in applied statistics.
- "*Regional coherence*" generally *does not* provides greater confidence in spatial patterns of change, as it simply reflects information redundancy due to the fact that the stations in a given area record the effects of the same weather systems; one thinks to have more evidence thanks to multiple records, while the information content is often not much different from that of a single station (see e.g. Douglas et al. (2000) for an example of what I mean). Actually, "*Regional coherence*" comes from a questionable approach usually referred to as "selling space for time", which is even more questionable when dealing with nonstationary processes.
- "*One simulation study assessed the ability of the Pettitt test to detect shifts and found that the test performs best when the step-change is located centrally in the distribution*" Correct, but why? Well, the answer cannot be provided by MC simulations for sure. Instead, it is sufficient to recall the steps of statistical inference mentioned above, and spend few minutes with the original Pettitt's paper to understand its theoretical properties, which allow the correct interpretation of results. Indeed, the Pettitt test statistic is the maximum of the absolute value of a sequence of Mann-Whitney statistics. These form a Brownian bridge, which is a Brownian motion constrained to assume zero values at the boundaries by definition (... theoretical definition, not 'functional' definition!). Therefore, the probability to observe a maximum decreases as the bridge approaches the boundaries,

> where it is constrained to converge to zero. Thus, the generally neglected theoretical properties of Pettitt test explain why the step changes reported in the literature are almost never located near the boundaries of a time series... Anyway, after many years in this business, I guess that almost all of those who have applied Pettitt or MK have never read the original papers, and if they did, I guess they missed their meaning.

To conclude, perhaps the Authors will disagree with me, and for sure my opinion is very different from that of the Referees. However, I think that this manuscript only summarizes mainstream confusion on these topics. Of course, this is legitimate. Nonetheless, I would avoid to compare definitions and studies based on clear formalism and mathematically/inferentially sound technical arguments with improvised 'definitions' and procedures not supported by any theoretical and formal argument, as the latter cannot be put on the same level of the former. I invite the Authors to reflect on the fact that being 'mainstream', 'widely accepted' or 'widely applied' does not mean necessarily being right. As mentioned above, applying a technically flawed method 1000 times does not make it (and results) correct; it only produces a proliferation of mistakes, and reflects a widespread superficial approach characterizing this time of decadence. The causes of such a big confusion and general low quality in 'statistical hydrology/hydro-climatology' are many, but this is not the right place to discuss them.
As an engineer, I know that questionable shortcuts and rough methodologies are often motivated by need of quick-and-dirty solutions, but science requires more rigor, more reasoning, and more time than running R packages or Matlab toolboxes: "*The problems caused by the indiscriminate use of recipes are compounded when obscure sophisticated techniques are used. It is fashionable to surprise the community with miraculous new techniques, even though the statistical model implicit in the method is often not understood.*" (von Storch and Zwiers, 2003, p. 97).

Sincerely,

Francesco Serinaldi

References

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

Ayers, JR, Villarini, G, Jones, C, Schilling, K. Changes in monthly baseflow across the U.S. Midwest. Hydrological Processes. 2019; 33: 748-758.

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

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

Serinaldi, F., Chebana, F. Kilsby, C.G. Dissecting innovative trend analysis. Stoch Environ Res Risk Assess 34, 733–754 (2020). https://doi.org/10.1007/s00477-020-01797-x

Serinaldi, F., Kilsby, C.G. The importance of prewhitening in change point analysis under persistence. Stoch Environ Res Risk Assess 30, 763–777 (2016). https://doi.org/10.1007/s00477-015-1041-5

Serinaldi F, Kilsby CG, Lombardo F. Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. Advances in Water Resources 2018, 111, 132-155, https://doi.org/10.1016/j.advwatres.2017.10.015

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

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-576, 2020.