IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 35-45

Climatological series shift test comparison on running windows

José A. Guijarro

State Meteorological Agency, Moll de Ponent s/n, Portopí,
07015-Palma de Mallorca, Spain
jguijarrop@aemet.es

(Manuscript received in final form October 3, 2012)

ABSTRACT–The detection and correction of inhomogeneities in the climate series is of paramount importance for avoiding misleading conclusions in the study of climate variations. One simple way to address the problem of multiple shifts in the same series is to apply the tests on windows running along the series of anomalies. But it is not clear which of the available tests works better. 500 Monte Carlo simulations have been done for the ideal case of a 600 normally distributed terms (a 50 years series of monthly differences), with a single shift in the middle and magnitudes of 0 to 2 standard deviations (*s*) in steps of 0.2 *s*. The compared tests have been: 1) classical t-test; 2) standard normal homogeneity test; 3) two-phase regression; 4) Wilcoxon-Mann-Whitney test; 5) Durbin-Watson test (lag-1 serial correlation), and 6) squared relative mean difference (simpler than t-test and hence faster to compute). The criterion for qualifying the performance of each test was the ability to detect shifts without false alarms and to locate them at the correct point. Results indicate that, under these precise simulated conditions, the best test are the classical t-test, Alexandersson's SNHT and SRMD, with almost identical results, followed by the Wilcoxon-Mann-Whitney test, while two phase regression and Durbin-Watson performances are very poor.

Key-words: homogenization, shift tests comparison, climatological series.

1. Introduction

Climatological series are very important for studying climate variability at all scales, but the climate signal is too often merged with unwanted variations due to changes in the type or exposure of the instruments, methods of observation, relocations of the stations, or changes in their surroundings.

Many methodologies have been proposed so far to detect and correct these inhomogeneities, which commonly appear as either sudden shifts or smooth trends in relative series. These relative series are usually computed as difference or ratio series between the problem series and a reference, that can be an observed trusted homogeneous series or a synthetic one compiled from a selection of the nearest or more correlated stations. Reviews of the different methods can be seen in *Easterling* and *Peterson* (1992), *Peterson et al.* (1998), *Aguilar et al*. (2003), and *Beaulieu et al*. (2008).

Several comparisons of shift detection methods have been undertaken so far (*Easterling* and *Peterson*, 1995; *Bosshard* and *Baudenbacher*, 1997; *Ducre-Robitaille et al*., 2003; *Beaulieu et al*., 2008), their results being influenced by the type (shifts and/or local trends), number and position of the simulated inhomogeneities, differences in station variance and between-station correlation structure, series length, autocorrelation, and nonstationarity.

The frequent concurrence of several jumps in the same series makes their detection problematic. One simple way to address this problem is to apply the test on time moving windows. During the development of an automated homogenization function for the CLIMATOL R contributed package (*Guijarro*, 2011a), the chosen approach for the detection of multiple change points was the application of a two-sample t-test for equal means to windows running along the series of anomalies (differences between the tested series and a synthetic reference series computed from neighboring stations). At this point, the question whether there were better detection tests emerged, but the available reviews are not fully conclusive, since the performance of the tests depends on the particular settings of the simulations and the significance threshold values chosen in each case, as it happens in the differing results of *Ducre-Robitaille et al*. (2003) and *Beaulieu et al*. (2008).

Therefore, new Monte Carlo experiments were designed to test the sensitivity and correctness of several algorithms in detecting and locating a shift in repeated series of white noise that simulate the ideal case of series of differences between a tested series with a single abrupt change in the mean and a homogeneous well correlated reference series. In this way we avoid the problems of simulating networks of observation or pairs of tested and reference stations as in the aforementioned evaluation exercises. Moreover, no a priori level of significance will be imposed, and location errors of the break point will be studied with no established thresholds of good/bad location. Next sections will explain this methodology, and the results of the tested algorithms will be discussed.

2. Methodology

500 series of 600 normally distributed terms (equivalent to tested minus reference monthly series of 50 years) were generated with the help of the *R*

function *rnorm* (*R Development Core Team*, 2010). Single shifts were added to all of them just in the middle (from term 301) with magnitudes from 0 to 2 standard deviations (*s*) in steps of 0.2 *s*, yielding a total of 5500 testing series. Six shift detection algorithms were applied on them, but not over the whole series, but on fixed width windows running along them. Different sample sizes were tried, from n=1 to 5 years (12 to 60 terms in steps of 12), and since two samples were involved in the shift tests, window widths of 2, 4, 6, 8, and 10 years were used. In this way, for n years sample size, every algorithm was tested $600–24 \cdot n+1$ times in each of the 5500 series (from 577 times with 1 year samples to 481 for samples of 5 years). *Fig. 1* shows an example series with a 0.8 *s* shift.

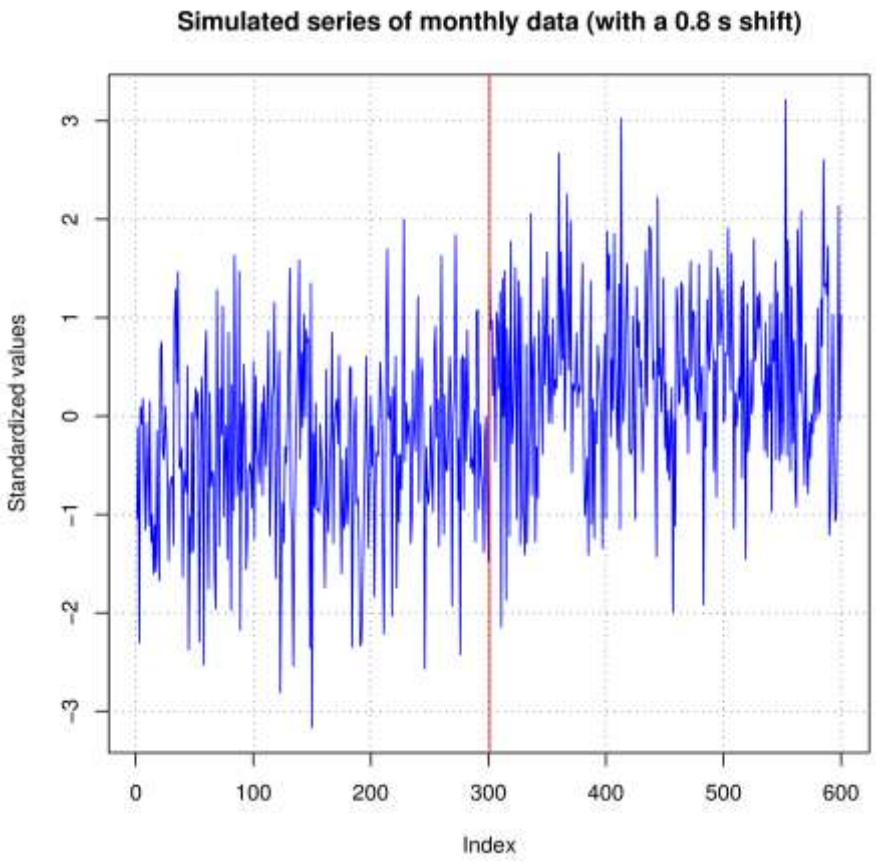

Fig. 1. Example of white noise difference series of 600 terms with a shift of 0.8 standard deviations in term 301.

The six algorithms tested were the following:

1. t-test: the classical test of mean differences of two samples.
2. SNHT: *Alexandersson*‘s (1986) algorithm, but modified to test the middle point of the window only.

3. TPR: two-phase regression, as formulated by *Easterling* and *Peterson* (1995).

4. WMW: Wilcoxon-Mann-Whitney test, which is similar to the Wilcoxon rank sums applied by *Karl* and *William*s (1987) but as formulated by *Gérard-Marchant* and *Stooksbury* (2008), and divided by the number of terms to make it less dependent on the sample size.

5. DW: lag-1 Durbin-Watson test for serial correlation.

6. SRMD (squared relative mean difference): $z = [\ (m_1\text{-}m_2)\cdot s^{-1}]^2$, where m_1 and m_2 are the sample means and s is the standard deviation of the whole window.

The reference values of DW and t-test were their returned p-values, but $\log_{10}$ transformed and sign reversed to allow more friendly figures (they are called *pV*, by analogy with the alkalinity index *pH* used in chemistry). *Fig. 2* displays the values returned by the six algorithms after being applied to a series similar to that in *Fig. 1* on running windows of 10 years (sample sizes of 5 years, i.e., 60 terms). Only the maximum value reached along the series, and its location (the middle point of the window giving that value) were retained for the statistical analysis of the results.

3. Results and discussion

The frequencies of the maximum values returned by the tests on each series and the errors of their corresponding locations (diagnosed break term minus 301) were analyzed statistically, and the results are shown graphically in form of boxplots, where each box summarizes 500 results. *Fig. 3* shows the influence of window size on the results yielded by the t-Test. It is clear that sample sizes of 12 terms are too small to allow the detection of shifts. If we take the value of the top whisker of the first box (homogeneous series) as a reasonable threshold to avoid false break detection, only roughly half of the 2 *s* shifts would be identified. With wider windows the power of detection improves: the half of the breaks detection reference is achieved with 0.8 and 0.6 *s* shifts for samples of 3 and 5 years respectively. (The intermediate 4 year sample graph can be seen in Figure 4). These results are in accordance with those of *Beaulieu et al*. (2008), who found that shifts under 1 *s* were difficult to identify, while all techniques tested by them worked well for breaks greater than 2 *s*.

The performance of the six tests with samples of 4 years can be seen in *Fig. 4*. As every test has its own metric, the units displayed in the vertical axis are all different, but it is easy to see that some tests reach higher values quicker than others as the shift magnitude increases, showing their greater power of detection.

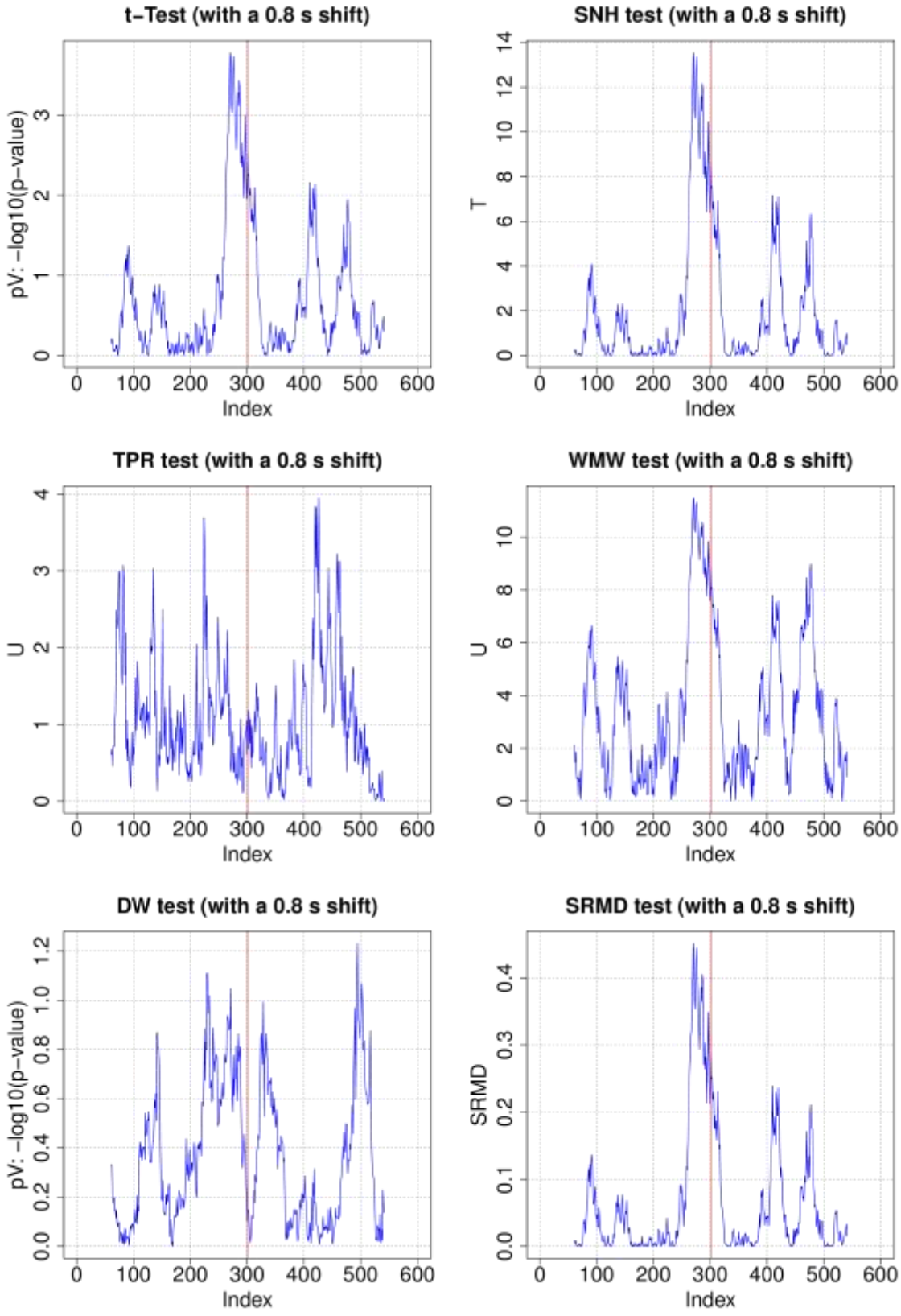

Fig. 2. Graphs of the values returned by the tests when applied to a series similar to that in *Fig 1* on running windows of 120 terms (two samples of 5 years). The vertical bar in the middle of the series indicates the true possition of the shift.

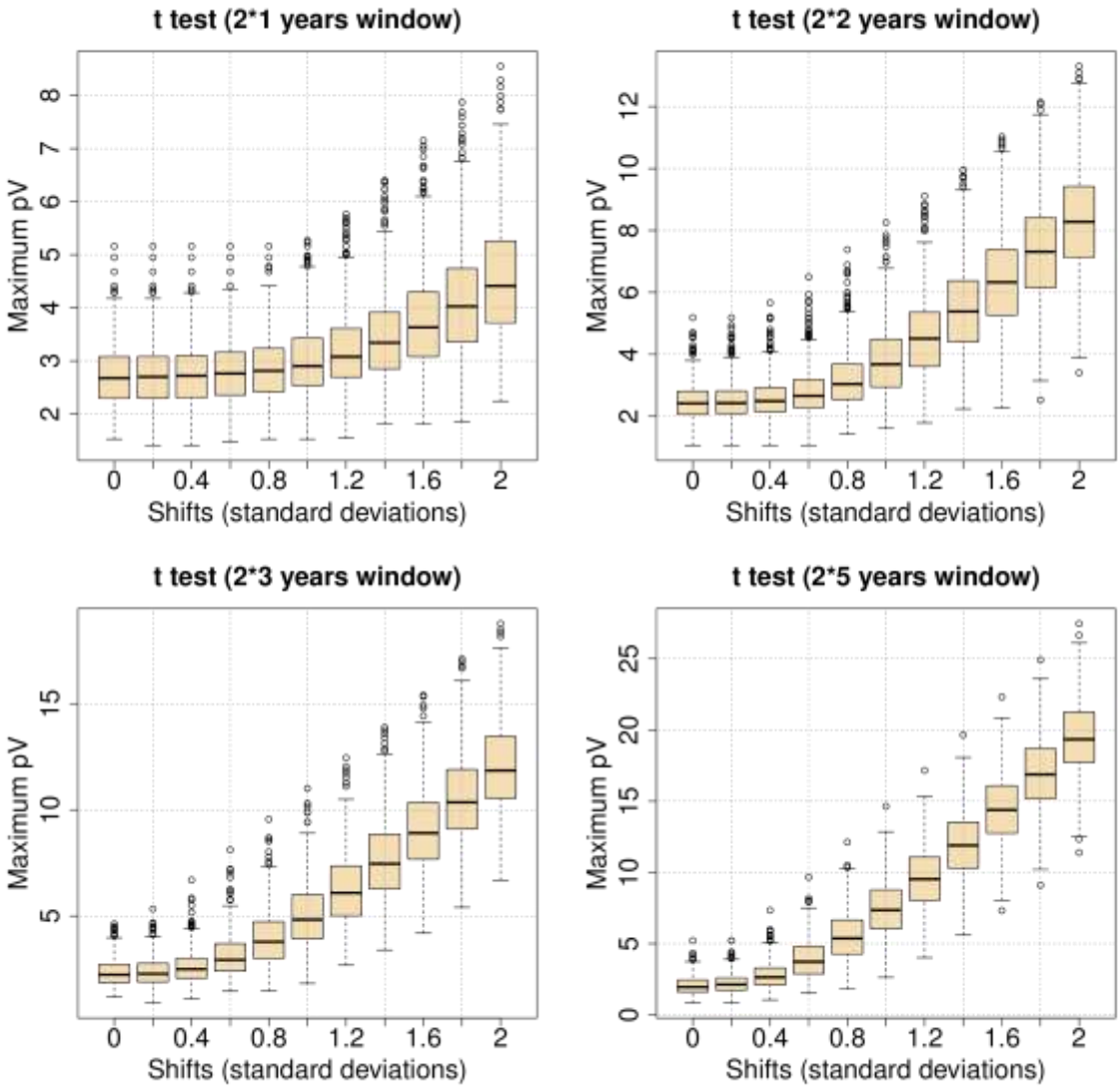

Fig. 3. Influence of window size on the results yielded by the t-test.

Table 1 presents the percentage of shift detection of every algorithm for each shift, for the 5 years samples, when the threshold detection is placed: a) at the maximum value obtained with the homogeneous series (no false detection is allowed); b) at the 99 percentile of the homogeneous values (permitting 1% of false detection). The best performances correspond to t-test, SNHT and SRMD, that give almost identical results, showing that they belong to the same family of tests. WMW follows, with good results form 1 *s* shift onwards, while DW and TPR both yield similar discouraging scores. Note that the thresholds of any test applied hundreds of times on every series through such a running window procedure, must be higher than their corresponding significant levels when applied only once on each series. E.g., the 14.23 of SNHT allowing 1% of false detection is higher than the 13.813 published by *Khaliq* and *Ouarda* (2007) for a 99% confidence level and sample size of 600 values (the whole simulated series).

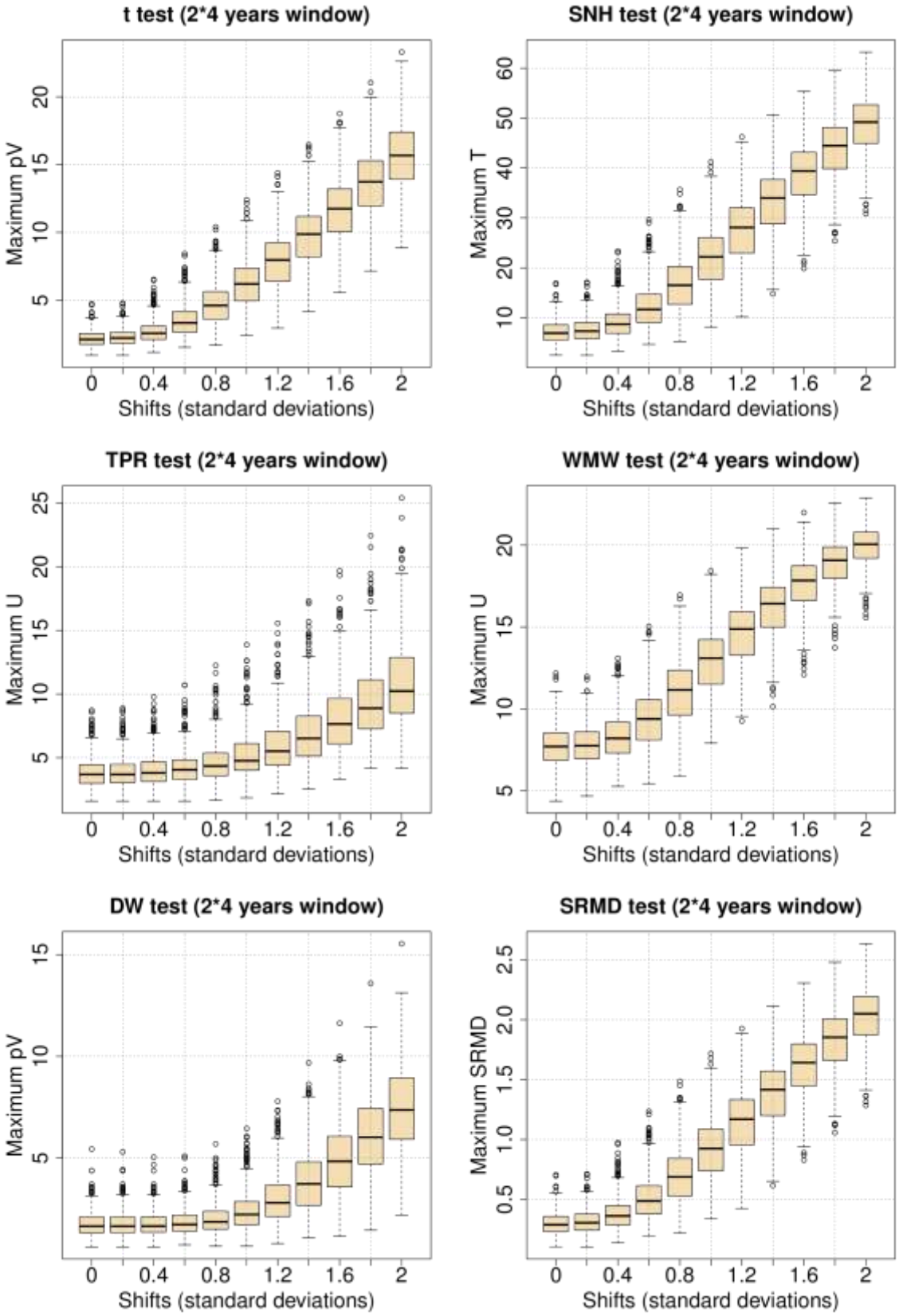

Fig. 4. Values of the six algorithms for shifts ranging from 0 to 2 standard deviations.

Table 1. Threshold values and percentage of shift detection in the cases of no false detection and allowing 1% of false detections, for a 5 years sample size (running windows of 10 years, i.e., 120 terms)

		Shift (standard deviations)									
Test	Thresh.	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
					No false detection:						
t-test	5.22	0.0	2.4	16.6	54.4	88.8	98.8	100.0	100.0	100.0	100.0
SNHT	19.06	0.0	2.4	16.8	54.6	88.8	98.8	100.0	100.0	100.0	100.0
TPR	8.69	0.0	0.6	1.6	2.8	7.8	16.6	34.4	55.0	74.8	88.4
WMW	13.78	0.0	2.0	12.6	47.4	82.8	98.4	100.0	100.0	100.0	100.0
DW	4.98	0.0	0.0	0.0	0.6	3.6	14.8	37.8	65.2	86.0	95.8
SRMD	0.635	0.0	2.4	16.6	54.4	88.8	98.8	100.0	100.0	100.0	100.0
					1% false detection:						
t-test	3.96	2.2	13.2	44.6	81.4	97.8	100.0	100.0	100.0	100.0	100.0
SNHT	14.23	2.4	13.2	44.4	81.4	97.8	100.0	100.0	100.0	100.0	100.0
TPR	6.87	1.6	3.0	4.6	11.6	20.4	41.0	62.2	79.4	91.0	97.4
WMW	12.04	0.8	8.6	35.0	73.4	95.2	99.8	100.0	100.0	100.0	100.0
DW	3.59	1.0	1.2	1.8	5.4	17.4	41.2	66.4	87.0	96.4	99.6
SRMD	0.474	2.4	13.2	44.4	81.4	97.8	100.0	100.0	100.0	100.0	100.0

With respect to the location errors, *Fig. 5* shows the corresponding box plots for the 4 years sample size (running windows of $2 \cdot 4 \cdot 12 = 96$ terms). Again, the t-test family (including SNHT and SRMD) reaches the best results, with small location errors for shifts greater than 0.6 standard deviations. Location errors of WMW are only slightly higher, but those of DW and specifically TPR are very big.

As CLIMATOL must apply the chosen test many times in iterative runs during the homogenization of a climatological network, computing efficiency is also important, and therefore, the time used by each of the tests was accounted for. Those adjusting regression models (TPR and DW) were the most time consuming using the R *lm* function. The R implementation of the t-test is much faster, but at the same time much slower than SNHT, probably due to its higher complexity and the inherent computation of p-values and other statistical parameters. This is why SRMD was introduced, achieving identical results as SNHT (in this two sample version), but at 20% higher speed. If TPR or DW had given better results, rewriting the regression algorithm to shorten their computing time would have been explored.

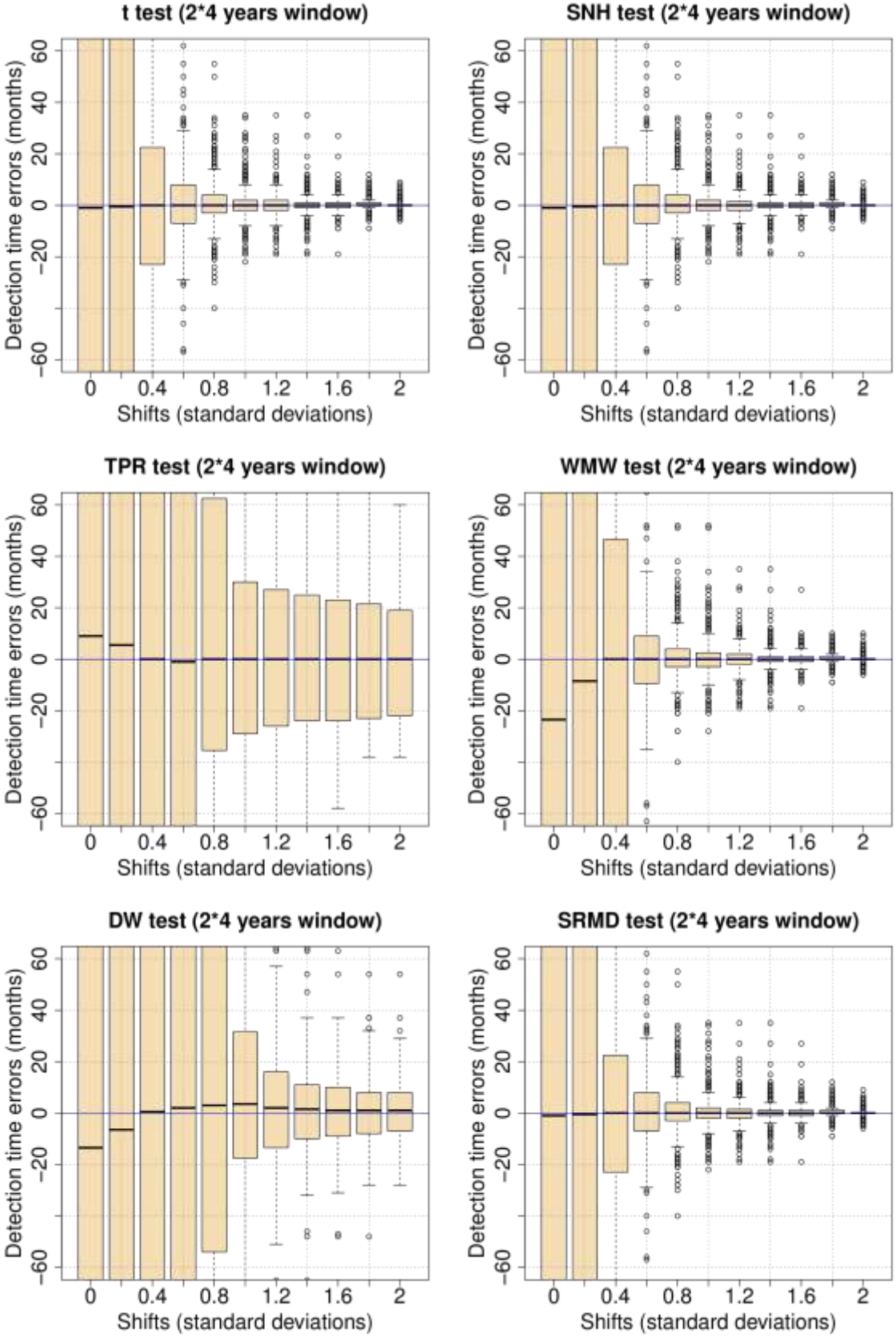

Fig. 5. Location errors of the six algorithms for shifts ranging from 0 to 2 standard deviations.

The combination of several of these tests was also tried, but when the best algorithm is used, there is no advantage in adding the results of any others. Therefore, CLIMATOL 2.0 implemented SRMD on running windows (4 years samples by default). Nevertheless, practical applications of that version showed that clear inhomogeneities spanning less than 3 years are common in real climatological series, and they were difficult to correct automatically due to the constraint of the minimum 3 years sample size required by the algorithm. Hence, the following 2.1 version dropped SRMD in favor of the popular and well tested SNTH which, freed from the window size restriction, is able to resolve close shifts. To avoid possible masking effects when multiple shifts are present in the same series, this test was implemented in two stages. In the first stages SNHT is applied on shifted windows of user defined width, and when significant shifts detected in this way have been corrected, SNHT is applied to the whole series in the second stage (*Guijarro*, 2011b).

4. Conclusions

The results of the simulations performed in this work indicate that, under these precise conditions of detection of a single shift in the middle of the series by means of fixed width windows running along the series, the best tests are the classical t-test and SNHT. SMRD is a simple derivative of the t-test with the same performance. The Wilcoxon-Mann-Whitney test yields acceptable results, but the two-phase regression and Durbin-Watson performances are very poor (although they can be better in other situations, e.g., in detecting local trends).

Nonetheless, windows need to have a minimum width of 6 years (two samples of 3 years), and that restrains the time resolution at which two close shifts can be identified. As a result, the t-test procedure of comparing the means of two samples was abandoned in favor of the standard formulation of SNHT, applied on stepped windows to avoid misleading results in the presence of multiple breaks in a first stage, then followed by an application on the whole series.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.,* 2003: *Guidelines on climate metadata and homogenization*. WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organization, Geneve.

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol., 6*, 661–675.

Beaulieu, C., Seidou, O., Ouarda, T.B.M.J., Zhang, X., Boulet, G., and *Yagouti, A.,* 2008: Intercomparison of homogenization techniques for precipitation data. *Water Resour. Res, 44*, 20.

Bosshard, W. and *Baudenbacher, M.,* 1997: Evaluation of various homogeneity tests by simulation of climatological time series. In: *Proceedings of the First Seminar for Homogenization of Surface Climatological Data*, Budapest, 6–12 October 1996, Hungarian Meteorological Service, 19–34.

Ducré-Robitaille, J.F., Vincent, L.A., and *Boulet, G.,* 2003: Comparison of techniques for detection of discontinuities in temperature series. *Int. J. Climatol. 23*, 1087–1101.

Easterling, D.R. and *Peterson, T.C.*, 1992: Techniques for detecting and adjusting for artificial discontinuities in climatological time series: a review. *5th International Meeting on Stat. Climatology*, June 22–26, 1992, Toronto.

Easterling, D.R. and *Peterson, T.C.*, 1995: A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol. 15*, 369–377.

Gérard-marchant, P.G.F. and *Stooksbury, D.E.*, 2008: Methods for Starting the Detection of Undocumented Multiple Changepoints. *J. Climate 21*, 4887–4899.

Guijarro, J.A., 2011a: http://cran.r-project.org/web/packages/climatol/index.html.

Guijarro, J.A., 2011b: User's guide to Climatol. 40 pp. http://webs.ono.com/climatol/climatol-guide.pdf

Karl, T.R. and *Williams, C.N.,* 1987: An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Clim. Appl. Meteor. 26*, 1744–1763.

Khaliq, M.N. and *Ouarda, T.B.M.J.*, 2007: On the critical values of the standard normal homogeneity test (SNHT). *Int. J. Climatol. 27*, 681–687.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E., Hanssen-bauer, I., Alexandersson, H., Jones, P., and *Parker, D.*, 1998: Homogeneity Adjustments of „In Situ" Atmospheric Climate Data: A Review. *Int. J. Climatol. 18*, 1493–1518.

R Development Core Team, 2010: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.