Available online at www.sciencedirect.com

Journal of Hydrology 278 (2003) 253–266

Journal of Hydrology

www.elsevier.com/locate/jhydrol

On detecting anomalous behaviour in runs

Roger L. Hughes*

The Department of Civil and Environmental Engineering, The University of Melbourne, Parkville, Vic. 3010, Australia

Received 29 October 2002; accepted 14 April 2003

Abstract

Suppose that values in a finite sequence of data are labelled as either $+1$ or -1 depending, respectively, on whether they are either above or below the median. Using this sequence of $+1$ and -1 s a unit-lag autocorrelation coefficient may be determined. The present study establishes the probability distribution for the number of runs of a fixed length of -1 s (or $+1$ s) occurring in the labelled data. This distribution is calculated both with and without the constraint of the sample autocorrelation. The distribution is compared with observed stream flow data to illustrate its use in detecting both deficits and excesses of low or high flows of a given duration. The use of this distribution, which is not restricted to stream flow data, provides an extremely convenient alternative to the more traditional methods of detecting anomalous behaviour and avoids requiring knowledge of the form of the parent distribution.

Keywords: Runs-analysis; Time series analysis; Distribution-free; Poisson distribution

1. Introduction

The objective of the present study is to determine the probability distribution governing the occurrence of runs of random values above, or below, the median in data of finite record length. This distribution is determined for the cases in which the binary autocorrelation of the record length is included and not included. The study is independent of the parent distribution of the data. As such it may be used either in the calculation of the probability of a specific number of runs (either side of the median) of given length in a record or the detection of unusual temporal behaviour in a record without the need for knowledge of the parent distribution.

The identification of unusual behaviour in an hydrological time sequence is often of value in understanding hydrological processes. Various methods of identifying anomalous behaviour of a sequence are available, such as variable-lag autocorrelation methods and Fourier methods, see for example Mills (1965) and Hannan (1960). These methods are often dependent on the form of the underlying distribution of the given sequence and are often indirect, sometimes having difficulty in determining unusual behaviour that has no dominant periodic structure. No attempt is made here to survey the vast literature on the topic. The above cited books review this literature better than any note can here.

* Tel.: +61-3-8344-4793; fax: +61-3-8344-4616.
E-mail address: rlhughes@unimelb.edu.au (R.L. Hughes).

0022-1694/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0022-1694(03)00156-2

The present study, seeks to present another technique. This technique is independent of the form of the underlying distribution and falls within the special category of tests discussed by Sprent and Smeeton (2001), Conover (1999), Neave and Worthington (1988), Maritz (1981), and Hollander and Wolfe (1973) amongst others. Essential to the technique is the use of the median (rather than the mean) and the consideration of runs of a specified length. It contrasts with standard runs-analysis, which often weights all runs equally irrespective of their length, see Feller (1968).

In a particular observed time sequence, it is convenient to label terms with values above the median $+1$ and the terms with values below the median -1. Thus a sequence of $+1$ and -1 s is generated. This sequence is then compared with the theoretical probability distribution of ± 1 values. Various standard statistical tests to determine the significance of various features of the data are then available.

Allowance is made for any unit-lag autocorrelation in the data. The unit-lag autocorrelation coefficient, r, used here, is defined by

$$r = \frac{1}{D-1}\sum_{j=1}^{D-1} S_j S_{j+1} \tag{1.1}$$

where S_j denotes the value of ± 1 that is associated with the jth term in the sequence, and D denotes the number of terms in the sequence. (Note that the values of S_j^2 are always 1 and hence factors involving their mean value have been neglected from the definition of the autocorrelation coefficient.) In implementing the present study, the implementer must decide if there is physical reason to regard the value of r as a constraint.

Peel et al. (2003) investigated drought lengths using a runs methodology in a complementary manner with the present study. They analysed both annual precipitation and runoff data, which they found to be well described by a unit-lag autoregressive process. The present study obtains the full distribution of runs of any given length from which the expected number of runs may be determined, if desired.

In Section 2 a probability distribution is introduced that ignores any autocorrelation. In Section 3 a similar distribution is introduced that includes the influence of the unit-lag autocorrelation. With these distributions it is a simple matter to compare any given record with the theoretical behaviour and conduct a statistical test on the significance of any deviation between the two records as explained later. The conclusions are given in Section 4.

The present manuscript is concerned with the variation of some quantity with time. The analysis fully generalises to include variations with time and space by the use of empirical orthogonal functions, see Kantz and Schreiber (1997). The orthogonality of the resulting time functions enables the autocorrelation of each function to be considered separately. The different time sequences involved in such an expansion will, in general, have different unit-lag autocorrelations. The extension of the present study to include variations in both time and space is not considered further here.

2. Correlation absent

As already noted in a particular time sequence, it is convenient to label terms with values above the median with $+1$ and the terms with values below the median with -1. Thus generating a sequence of $+1$ and -1 s. A derivation of the probability distribution for the number of runs of a given length in data of these $+1$ and -1 s with no autocorrelation is derived in Appendix A. The reader is referred to this appendix for details. For present purposes the resulting probability distribution for exactly N runs of length M in the record of length D is

$$\begin{aligned}\Pr(D,M,N)=&\frac{1}{D!}(T(N)-(N+1)T(N+1)\\&+\frac{1}{2}(N+1)(N+2)T(N+2)\\&-\frac{1}{6}(N+1)(N+2)(N+3)T(N+3)\ldots)\end{aligned} \tag{2.1}$$

where $T(N)$ is defined by

$$\begin{aligned}T(N)=&\frac{((D/2)!)^2}{(D/2-NM)!N!}\frac{(D/2+1)!}{(D/2+1-N)!}\\&\times\frac{(D-N-NM)!}{(D/2-N)!}\end{aligned} \tag{2.2}$$

and the numerical coefficient of the nth term, $n=1,2,\ldots$, involving $T(N-1+n)$ in Eq. (2.1) is given by

$$a_n=(-1)^{n-1}\frac{1}{(n-1)!}\text{ for } n\geq 2 \tag{2.3}$$

Thus the coefficient of $T(N-1+n)$ in Eq. (2.1) is

$$(-1)^n\binom{N}{n-1}.$$

The sum in Eq. (2.1) is over all terms for which $N< D/(1+M)$.

It is useful to consider the form of the distribution (2.1) in the case of $D/2 \gg NM$. In this case Eq. (2.2) may be approximated by

$$T(N) \approx \frac{D!}{N!}2^{-2N-NM}D^N \tag{2.4}$$

and so Eqs. (2.1) and (2.3) imply

$$\Pr(D,M,N) \approx \frac{1}{N!}2^{-2N-NM}D^N\Big(1-(2^{-2-M}D) + \frac{1}{2}(2^{-2-M}D)^2 - \cdots\Big) \approx \frac{1}{N!}\eta^N\,\mathrm{e}^{-\eta} \tag{2.5}$$

where

$$\eta = 2^{-(2+M)}D \tag{2.6}$$

Thus $\Pr(D,M,N)$ is approximately distributed as a Poisson distribution with mean (and hence variance) η as given by Eq. (2.6). In the limit of large $D/2$, Eq. (2.5) is exact. As a check on the formulation of this distribution, we note that in the limit of large $D/2$ the expected number of terms involved in runs is

$$\sum_{M=0}^{\infty}\frac{D}{2}\frac{M}{2^{M+1}} = \frac{D}{2}\frac{1}{4}\sum_{M=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\lambda}\lambda^M\bigg|_{\lambda=1/2}$$
$$= \frac{D}{2}\frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}\lambda}\sum_{M=0}^{\infty}\lambda^M\bigg|_{\lambda=1/2}$$
$$= \frac{D}{2}\frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{1}{1-\lambda}\bigg|_{\lambda=1/2} = \frac{D}{2} \tag{2.7}$$

as expected because the sum of all terms labelled either ±1 must be half the record length.

To gauge the error in using Eq. (2.5) we consider a hypothetical data record 12 years long ($D=12$) and we are interested in runs of length 3 years ($M=3$) below the median. Then the full distribution, Eq. (2.1), predicts that there are no runs of this length with a probability of 0.60, one run with a probability of 0.38 and two runs with a probability of 0.02. (Even for this modest example the numerical values of $T(N)$ in Eq. (2.2) are of the order of 10^8. However their calculation is rapid and is easily done with a hand held calculator.) By comparison, the asymptotic form of the distribution, Eq. (2.5), yields for no runs a probability of 0.69, one run a probability of 0.26 and for two runs a probability of 0.05. However, for a data record of 50 years ($D=50$), there is no substantial discrepancy between the full solution and the approximate solution.

To illustrate the above distribution we consider annual flow in the Atbara (a Sudanese tributary of the Nile). The record is 52 years long (with, of course, 26 years of flows above and below the median). These values are clustered and produce 14 runs below the median and 14 above. Thus assuming no autocorrelation in the data, the distribution of runs of flows either above or below the median is given by the Poisson distribution, Eqs. (2.5) and (2.6). We consider the predicted 95% mode-centred interval (or the one sided 95% interval if the mode is too low to allow mode-centring of the interval) of the number of runs of length M at Atbara. This interval covers all N below 12 for run length $M=1$; below 7 for $M=2$; below 5 for $M=3$; below 3 for $M=4$; and below 3 for $M=5$ as shown in Fig. 1 by crosses. From Fig. 1 it can be seen that the observations of runs, both above and below the mode, lie within these bounds with the bounds being very conservative for runs of small length. Of importance here is that the conservative nature of these estimates make the above test of limited value.

Generally hydrologists, and other earth scientists, are interested in records for which there is a physical reason to believe there is a correlation. The above distribution has been included here for completeness.

3. Correlation present

In Section 2, no allowance was made for any autocorrelation in the data. In many situations of hydrological interest autocorrelation is important. It is a simple matter to determine the unit-lag autocorrelation in any particular record. The present section presents the probability distribution of runs (above and below the median) when the observed unit-lag autocorrelation in the data is imposed.

The cases where the two ends of the sequence of data are (i) either both labelled $+1$ or -1 and (ii) of different labels, that is one is $+1$ and the other is -1, must be treated separately. These two cases are treated in Appendices B and C, respectively. The two cases have mutually exclusive possibilities for their autocorrelation coefficients.

For illustrative purposes only, we consider three rivers in Africa, the Atbara (considered in Section 2 without imposing the constraint of its lack of autocorrelation), the Zaire, and the Zambezi. Figs. 1–3 plot the occurrence of specified lengths of flows above and below the median for these three rivers. All three rivers have impoundments, whether controlled or uncontrolled. Hence the examples considered are illustrative only.

The first case considered here is that of the Atbara. As seen from Fig. 1, the 26 values occur above the median and the 26 values below the median in the record of length 52 years. Labelling values above and below the median $+1$ and -1, respectively, and determining the unit-lag autocorrelation coefficient of this record of $+1$ and -1 s yields -0.059, compared with a conventional product-moment correlation of -0.091 for the original data. (Note the small autocorrelation was why this river was chosen for analysis in Section 2 where the autocorrelation was ignored.)

Using Appendix C (because the end values have different values) the 95% mode-centred confidence interval is covered by the range 5–9 for runs of length 1 year. As the observed occurrence of runs, both above and below the median that is 8 and 7, respectively, is well within both the 95% covered modally centred range, we conclude that the behaviour is as expected. A similar calculation may be done for each of the record lengths in Fig. 1.

Of interest also is the maximum length of runs in the data. These are runs of 5 years duration for above median flows and 4 years duration for below. Again using the distribution derived in Appendix C, the probability of having a run exceed 4 years is 0.50. (The probability of no 4 year, 5 year, 6 year and 7 year runs being 0.40, 0.69, 0.87 and 0.95, respectively.) Such behaviour is consistent with the occurrence of maximum run lengths of 5 and 4 years for above and below median runs, respectively.

We conclude that no anomalous behaviour appears to be occurring in the Atbara. However, the tests applied to the above data are clearly much more stringent when the autocorrelation is a constraint, as here, than when it is not used as in Section 2.

The second case considered is that of the Zaire, shown in Fig. 2. The record length for the Zaire is of odd length. As such the median must occur in the record. The theory presented in Appendices B and C

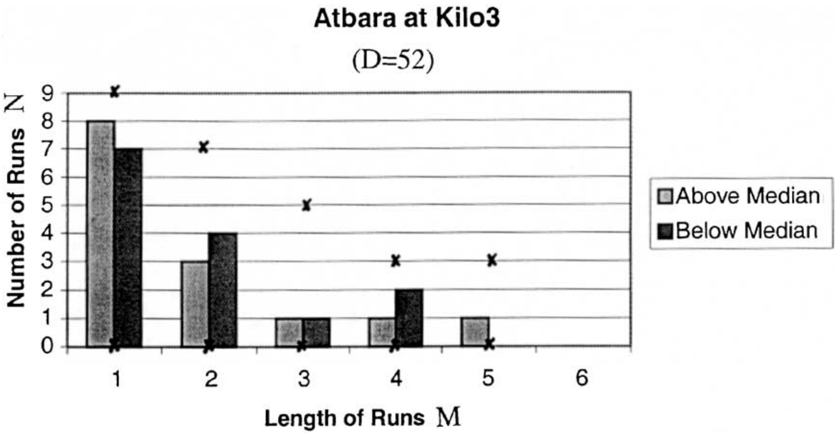

Fig. 1. Occurrence of runs of flow values either above or below the median for the Atbara in north-eastern Africa. Run lengths are in years, and X marking the 95% confidence interval (without autocorrelation).

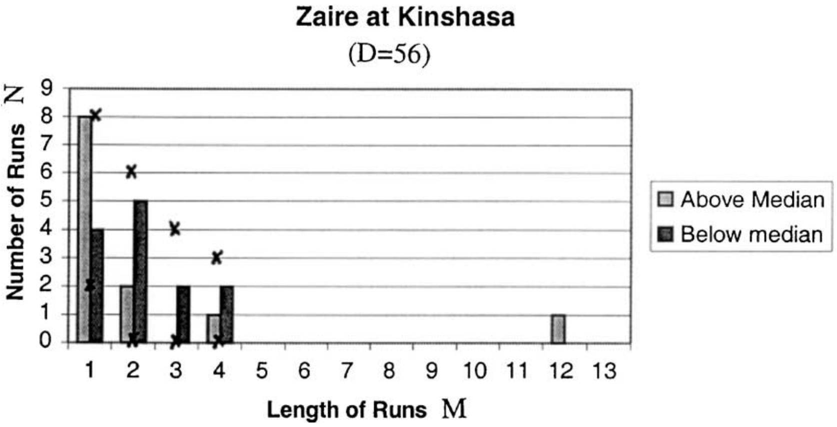

Fig. 2. Occurrence of runs of flow values either above or below the median for the Zaire in central Africa. Run lengths are in years, and X marking the 95% confidence interval (with autocorrelation).

assumes that records are of even length. In the example the value represented by the median is neglected from the analysis. Thus the record is 56 years in length, rather than 57 years, and there are 28 values above and below the median.

Again labelling values either +1 or −1 depending on whether the values are above or below the median, respectively, and determining the unit-lag autocorrelation coefficient of this newly labelled sequence gives a correlation coefficient, using Eq. (1.1), of 0.13 (compared with a product moment correlation of 0.54 for the original flow record). Using Appendix B, because the end values are of the same type, the mode of runs of length 1 is 5. The 95% modally centred range is 2–8 (a tighter non-centred range of 3–8 is also acceptable at the 95% level). Therefore the behaviour of runs of length 1 year is within the limits of normal behaviour. Similarly for runs of length 2 years giving 0–6 with mode 3, 3 years giving 0–4 with mode 1, and 4 years giving a range 0–3 with mode 1. Runs of length 5 years or more have a mode of 0 and the probability of having a run of 12 as observed is 0.0019. Thus, unless the present case has been biased by its choice for discussion here, we conclude that the presence of a single run of length 12 years is anomalous with respect to a binary correlation of 0.13 at the 95% level.

The final case considered here is that of the Zambezi in Fig. 3. In this case there are 54 years of observations. Labelling the data +1 or −1 depending

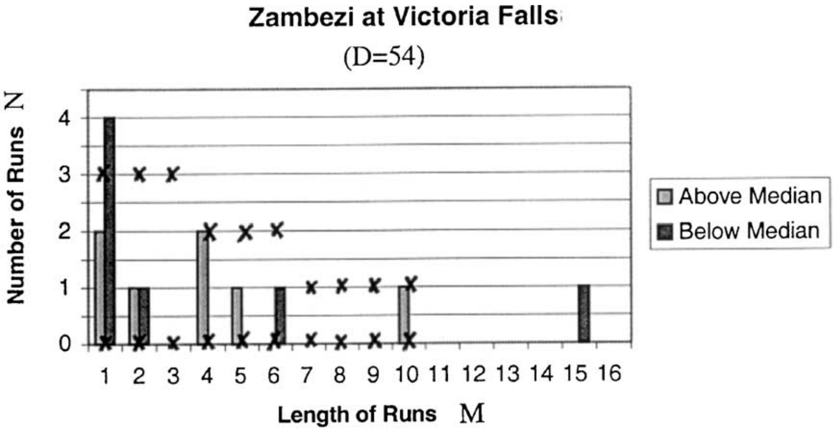

Fig. 3. Occurrence of runs of flow values either above or below the median for the Zambezi in southern Africa. Run lengths are in years, and X marking the 95% confidence interval (with autocorrelation).

on its position relative to the median, Eq. (1.1) gives the binary lag-1 autocorrelation coefficient of the data as 0.51 (compared with the product-moment correlation of the raw data of 0.46, a result that is proportionately much closer than in the previous two cases). Like the Atbara, but unlike the case of the Zaire, the case of the Zambezi has $+1$ at one end and -1 at the other end of the record. Thus the calculation in Appendix C is appropriate.

For runs of length 1 year the 95% predicted range is 0–3 runs. Thus the occurrence of four runs below the median is anomalous. The maximum number of runs at the 95% confidence level for runs of length 2 and 3 years is three, for runs of length 4–6 years is two, and for runs of length 7–12 years is one. Thus the remaining behaviour observed in the record is acceptable at the 95% level with the exception of the run below the median of length 15 years. This latter run has a probability of only 0.014 of occurring. Furthermore, the probability of having a run of length 15 years or greater is predicted to be 0.028 which is unacceptable at the 95% level.

These few examples illustrate the ways in which the probability distribution presented here may be used. The distribution is extremely valuable where there is reason to believe that a record contains unit-lag autocorrelation and one is searching for anomalous behaviour.

In passing it is noted that a simple expression for the probability, similar to the Poisson distribution (2.5), may be obtained from Appendices B and C in the case of large D such that both $D/2 \gg MN$ and $D/2 \gg \frac{1}{4}(D-1)(1-r)$ which are the conditions of a long record with high autocorrelation. In this case

$$\Pr(D,M,N) \approx \frac{1}{N!}\xi^N \mathrm{e}^{-\xi} \tag{3.1}$$

where $\xi = D(1-r)^2/8$. For the Zambezi, where the binary autocorrelation coefficient is 0.51, the mean ξ is equal to 1.8 irrespective of M. Such behaviour is consistent with that of Fig. 3 although there is a slow decay (slower than in Figs. 1 and 2 for which the correlation coefficient was lower) with M. As noted earlier, Eq. (3.1) is only valid when $D/2 \gg MN$ that is when $M \ll D/2\xi \sim 15$ here. As we have seen the distribution is no longer independent of M but decays rapidly with M when M is just less than $M \sim 15$ making the occurrence of an event of length 15 in Fig. 3 is anomalous.

4. Conclusions

The distributions developed here are applicable to the analysis of a wide variety of discrete time series with significant applications in hydrology. Provided only a few calculations are required the calculation of probabilities in the case where the autocorrelation is ignored may be performed simply on a hand calculator. However, although the distributions developed here are conceptually simple their implementation is tedious if autocorrelation is included. If a machine is to be used, the implementation requires significant computer programming because of the large numbers generated by the factorials generated by combinatorial mathematics. Once programmed, the use of these distributions provides a very simple means of determining anomalous behaviour.

The distributions discussed here are derived in their entirety in each of the three appendices, Appendices A, B and C. In Appendix A no account is made of the unit-lag autocorrelation coefficient. The resulting distribution is less constrained and less likely to yield significant results than when the autocorrelation is included. Such an allowance is made in Appendices B and C. The distributions presented in these latter appendices refer, respectively, to the two cases in which the two ends are either the same or opposite with respect to being above or below the mean.

River flow data for three rivers in Africa are considered. These three rivers were loosely selected for illustrative purposes only. The three rivers considered are studied for illustrative purposes only. They are not intended as a substantial study.

It is noted that by the use of statistics based on the median (in contrast with the use of the mean) abnormal statistics associated with times of low flow do not bias the finding of abnormal statistics associated with high flows, and visa versa. The present analysis can be used to analyse either high or low flows. (In the case of high flows by multiplying the sequence of flows by negative one.) Only anomalous lengths of high or low flows influence the analysis. The techniques described here are not affected by anomalous extreme events in the size of the flow. This behaviour limits the use of this tool to the analysis of the length of anomalous events. However, it implies that the technique is not influenced by extreme events that may result from incorrectly recorded data and more importantly it implies that the results of the technique are independent of the form of the underlying distribution.

A method dependent on the distributions derived and discussed here provides a useful alternative to the current methods of time series analysis for anomalously long or short phenomena, with the appealing property that the analysis is direct. Unusual behaviour need not be inferred indirectly as is often the case in other methods of analysis.

5. Programme availability

As Sprent and Smeeton (2001, p31) state "appropriate software is virtually essential for the implementation of all but the simplest (non-parametric) methods". The computer programmes used to implement Appendices B and C are freely available on the World Wide Web at http://www.civenv.unimelb.edu.au/~rogerh/.

In running these programmes, the programme will ask for D, M and $\sum S_j S_{j+1}$, respectively, and deliver probabilities for all N. The probability when $N = 0$ is inferred from the sum of the probabilities for all other N, and may be in error if the probability of this outcome is extremely small.

Acknowledgements

The author would like to thank Prof. T.A. McMahon (Melb.) for noting, in casual conversation, the need for this work and encouraging it, and Dr M.C. Peel (Melb.) for converting the Fortran version of the programme developed to visual basic, for running that programme and selecting the three data sets discussed in Section 3 from the hundreds of cases he has tested. Ms P. Doughty very kindly established a new web page for the computer programmes generated in this project and Dr W.W. Wood (Melb.) very helpfully checked much of the derivations. The enthusiastic encouragement of Prof. G.G.S. Pegram (Natal) was greatly appreciated. Two anonymous reviewers made very useful comments to improve the readability of the manuscript, based on their extremely careful reading, for which the author is very grateful.

Appendix A. Probability distribution for runs uniformly below the median in a finite sequence without serial correlation

The present appendix ignores any measured autocorrelation in the record considered. This simplification has a dramatic effect on the complexity of the expression for the probability distribution. We consider an ordered record of D discrete events where each event is measured by a real number. Each event in this record, say event j, with a value less than the median of the record is labelled $S_j = -1$ and similarly each event greater than the median is labelled $S_j = +1$. For brevity, attention is restricted to records of even length. (Generally it is sufficiently accurate to truncate a record by one element or by ignoring a predetermined median if its sequence has an odd number of values.) Thus, there are $D/2$ events with $S_j = -1$ and $D/2$ events with $S_j = +1$. The present appendix is concerned with determining the probability that exactly N runs of exactly M events have $S_j = -1$. Clearly,

$$\Pr(D, M, N) = \frac{\text{number of combinations with } D, M, N}{\text{total number of combinations with } D} \tag{A1}$$

for ergodic behaviour. The determination of both the denominator and the numerator of the right hand side of this equation need to be considered in turn.

To obtain the numerator of Eq. (A1), we consider the number of combinations in which D events can be arranged with *at least* N runs of length exactly M. We denote this number by $T(N)$. Thus, the numerator of Eq. (A1) is equal to $t(N) = \Xi(T)$ where Ξ is a functional that converts the number of combinations in which there are at least N runs into the number of combinations in which there are exactly N runs. The total number of combinations possible, without regard to the number or length of runs of $S_j = -1$ in denoted by S.

The form of $T(N)$ is

$$T(N) = H_1 H_2 H_3 H_4 \tag{A2}$$

where H_1, H_2, H_3, H_4 are factors as defined below.

The factor H_1 represents the number of ways in which the MN events involved in runs can be drawn from the $D/2$ events with $S_j = -1$ to form N unlabeled ordered groups. Thus

$$H_1 = \frac{(D/2)!}{(D/2 - NM)!N!} \tag{A3}$$

The factor H_2 represents the number of ways in which the $S_j = +1$ events can be ordered. There are $D/2$ such events. Thus

$$H_2 = (D/2)! \tag{A4}$$

Into the array of $S_j = +1$ events must be imbedded the N groups that could be constructed in H_1 ways and the $D/2 - NM$ remaining single $S_j = -1$ events. The factors H_3 and H_4 represent the number of locations for imbedding the N runs of $S_j = -1$ events and the remaining $S_j = -1$ events not in runs in the $D/2$ events with $S_j = +1$, respectively.

There are $D/2 + 1$ locations where the first of the N runs may be imbedded including the ends. This number decreases by one after each run is imbedded. The factor H_3 is given by

$$H_3 = \frac{(D/2 + 1)!}{(D/2 + 1 - N)!} \tag{A5}$$

However, possible locations for imbedding each of the $D/2 - NM$ remaining $S_j = -1$ events increases after each imbedding. Thus

$$H_4 = \frac{(D - N - NM)!}{(D/2 - N)!} \tag{A6}$$

Thus by Eq. (A2), $T(N)$ may be determined.

As stated earlier the numerator of Eq. (A1) is given by $\Xi(T)$. It is tempting to suppose that $\Xi(T)$ is given by $T(N) - T(N+1)$. However out of a single realisation of $(N+1)$ runs it is possible to choose N runs in $(N+1)$ ways and similarly if $(N+2)$ runs occur, it is possible to choose N runs in $(N+1)(N+2)/2$ ways. Hence

$$\begin{aligned} T(N) = t(N) &+ t(N+1)\frac{(N+1)}{1!} \\ &+ t(N+2)\frac{(N+2)(N+1)}{2!} \\ &+ t(N+3)\frac{(N+3)(N+2)(N+1)}{3!}\cdots \end{aligned} \tag{A7}$$

where the sum extends to the maximum number of runs of length M that is possible. Setting

$$\hat{T}(N) = N!T(N) \tag{A8}$$

$$\hat{t}(N) = N!t(N) \tag{A9}$$

Eq. (A7) takes the form

$$\hat{T}(N) = \hat{t}(N) + \frac{1}{1!}\hat{t}(N+1) + \frac{1}{2!}\hat{t}(N+2) + \frac{1}{3!}\hat{t}(N+3) + \cdots \tag{A10}$$

Hence writing

$$\hat{t}(N) = a_1\hat{T}(N) + a_2\hat{T}(N+1) + a_3\hat{T}(N+2) + \cdots \tag{A11}$$

we have

$$\begin{aligned} \frac{1}{1!}\hat{t}(N+1) &= \frac{1}{1!}a_1\hat{T}(N+1) + \frac{1}{1!}a_2\hat{T}(N+2) + \frac{1}{1!}a_3\hat{T}(N+3) + \cdots \\ \frac{1}{2!}\hat{t}(N+2) &= \frac{1}{2!}a_1\hat{T}(N+2) + \frac{1}{2!}a_2\hat{T}(N+3) + \cdots \\ \frac{1}{3!}\hat{t}(N+3) &= \frac{1}{3!}a_1\hat{T}(N+3) + \cdots \end{aligned} \tag{A12}$$

Adding Eq. (A11) to the components of Eq. (A12) and comparing the result with Eq. (A10) yields

$$\begin{aligned} a_1 &= 1 \\ a_2 &= -\frac{1}{1!}a_1 = -1 \\ a_3 &= -\frac{1}{1!}a_2 - \frac{1}{2!}a_1 = \frac{1}{2} \\ a_4 &= -\frac{1}{1!}a_3 - \frac{1}{2!}a_2 - \frac{1}{3!}a_1 = -\frac{1}{6} \end{aligned} \tag{A13}$$

Thus using Eqs. (A8) with (A9), (A11) and (A13) the numerator of Eq. (A2), that is $t(N)$, is obtained. By induction (a delightful proof involving much cancellation), it may be shown that in general

$$a_n = (-1)^{n-1}\frac{1}{(n-1)!} \tag{A14}$$

It follows that the functional Ξ is given by

$$\Xi(T) = (N!)^{-1}\sum a_n(N+n-1)!T(N+n-1) \tag{A15}$$

The factor $S(N)$ represents the number of ways in which the all events can be ordered. There are D such events. Thus

$$S = D! \tag{A16}$$

Thus by Eq. (A1)

$$\Pr(D,M,N) = \frac{\Xi(T)}{S} = \frac{\Xi(H_1H_2H_3H_4)}{S} = \Xi\left(\frac{H_1H_2H_3H_4}{S}\right) \tag{A17}$$

because Ξ is a linear functional and S is independent of N.

Appendix B. Probability distribution for runs uniformly below the median in a serially correlated finite sequence (same behaviour at both ends)

Consider all ordered records of D discrete events where each event is measured by a real number. Each event, say event j, is either greater than the median of D in which case we set $S_j = +1$, or less than the median in which case we set $S_j = -1$. (In the case of a record of odd length, one event is the median and this event is randomly given a S_j value of $S_j = -1$ or $S_j = +1$.) We define a median autocorrelation coefficient, r, for a particular record by Eq. (1.1) in the text

$$r = \frac{1}{D-1}\sum_{j=1}^{D-1} S_jS_{j+1}. \tag{B1}$$

This appendix is concerned with determining the probability that a record drawn at random from D discrete events has exactly N runs of M events with $S_j = -1$, given r, the autocorrelation coefficient of the record. Note that by Eq. (B1) r can only take certain discrete values.

The present appendix is concerned only with the cases where D is even which by Eq. (B1) requires $(D-1)r$ to be odd. The alternate cases, where D is odd can be treated similarly. However, simple interpolation is probably adequate in most hydrological cases of interest. These alternate cases become indistinguishable, in practice, for large D.

If the record is drawn without bias, then by definition of bias, the probability is

$$\Pr(D,M,N,r) = \frac{\text{number of combinations with } D,M,N,r}{\text{total number of combinations with } D,r} \tag{B2}$$

The determination of both the denominator and the numerator of the right hand side of this equation need to be considered in turn.

To obtain the numerator of Eq. (B2), we consider the number of combinations in which D events can be arranged with *at least* N runs of exactly length M. We denote this number by $Q(N)$. The numerator of Eq. (B2) is of the form $q(N) = \Xi(Q)$ where the functional Ξ is defined later. The number $Q(N)$ takes the form

$$Q(N) = E_1E_2(F_3G_{3c}H_c + F_3G_{3e}H_e + F_4G_4H_e + F_5G_5H_e) \tag{B3}$$

where E_1, E_2; F_3, F_4, F_5; G_{3c}, G_{3e}, G_4, G_5; H_c and H_e are factors as defined below.

The factor E_1 represents the number of combinations in which the $S_j = +1$ events can be ordered. There are $D/2$ such events. Thus

$$E_1 = (D/2)! \tag{B4}$$

The factor E_2 represents the number of ways in which the MN events involved in runs can be drawn from the $D/2$ events with $S_j = -1$ to form N unlabeled ordered groups. Thus

$$E_2 = \frac{(D/2)!}{(D/2-MN)!N!} \tag{B5}$$

The factors F_3, F_4 and F_5 represent the number of possible locations for imbedding the N runs of $S_j = -1$ events in the $D/2$ events with $S_j = +1$.

The factor F_3 corresponds to the case when $S_1 = +1$ and $S_D = +1$ after the runs have been imbedded.

In this case, there are $(D/2 - 1)$ possible sites between the already imbedded events for the imbedding and so

$$F_3 = \frac{(D/2 - 1)!}{(D/2 - 1 - N)!} \tag{B6}$$

The factor F_4 corresponds to the case when $S_1 = -1$ or $S_D = -1$ (but not both) after the imbedding of the N runs of $S_j = -1$ events. In each case, there are $(D/2 - 1)$ possible sites for the imbedding plus one site that must be used, and so

$$F_4 = 2N\frac{(D/2 - 1)!}{(D/2 - N)!} \tag{B7}$$

The factor F_5 corresponds to the case when $S_1 = -1$ and $S_R = -1$ after the imbedding of the runs. Thus, there are $(D/2 - 1)$ possible sites for the imbedding plus two sites that must be used. Hence

$$F_5 = N(N-1)\frac{(D/2 - 1)!}{(D/2 + 1 - N)!} \tag{B8}$$

The factors G_{3c}, G_{3e}, G_4 and G_5 represent the number of possible locations for imbedding the runs of the remaining $(D/2 - MN)$ events of type $S_j = -1$. There are $(D/2 + 1 - N)$ sites available to be filled with these events of type $S_j = -1$.

In the case where $S_1 = +1$ and $S_D = +1$ after the runs considered earlier were added, there are two possibilities, either $S_1 = +1$ and $S_D = +1$, or $S_1 = -1$ and $S_D = -1$ after placing the remaining events. (The value of r excludes the case of these remaining events occupying only one of the end sites as noted earlier.) If the correlation coefficient is to be r, then these events must occur in $N_c = \frac{1}{4}(D-1)(1-r) - N$ runs within the available sites excluding end sites (leaving only the central sites) or in $N_e = \frac{1}{4}(D-1)\times(1-r) - N + 1$ runs if the end sites are included. It follows that for these two cases

$$G_{3c} = \frac{(D/2 - 1 - N)!}{(D/2 - 1 - N - N_c)!N_c!} \tag{B9}$$

$$G_{3e} = \frac{(D/2 - 1 - N)!}{(D/2 + 1 - N - N_e)!(N_e - 2)!}, \tag{B10}$$

respectively.

In the case when $S_1 = -1$ or $S_D = -1$ (but not both) after the imbedding of the N runs of $S_j = -1$ events but before the remaining $S_j = -1$ events have been imbedded, one imbedding must be used to make $S_1 = -1$ and $S_D = -1$. Thus

$$G_4 = \frac{(D/2 - N)!}{(D/2 + 1 - N - N_e)!(N_e - 1)!} \tag{B11}$$

In the remaining case of $S_1 = -1$ and $S_D = -1$ before the imbedding of the remaining $S_j = -1$ events all further imbedding must occur away from the edges. Hence

$$G_5 = \frac{(D/2 + 1 - N)!}{(D/2 + 1 - N - N_e)!N_e!} \tag{B12}$$

These sites must be filled with the remaining $(D/2 - MN)$ events of type $S_j = -1$. If N_c sites are available, this can be done in H_c ways, where

$$H_c = \frac{(D/2 - MN)!(D/2 - MN - 1)!}{(D/2 - MN - N_c)!(N_c - 1)!} \tag{B13}$$

Similarly, if N_e sites are available

$$H_e = \frac{(D/2 - MN)!(D/2 - MN - 1)!}{(D/2 - MN - (N_e)!(N_e - 1)!} \tag{B14}$$

Hence by Eq. (B3), $Q(N)$ can be determined.

As stated earlier the numerator of Eq. (B2) is given by $\Xi(Q)$. Out of a single realisation of $(N+1)$ runs it is possible to choose N runs in $(N+1)$ ways and similarly if $(N+2)$ runs occur, it is possible to choose N runs in $(N+1)(N+2)/2$ ways. Hence

$$\begin{aligned} Q(N) = {} & q(N) + q(N+1)\frac{(N+1)}{1!} \\ & + q(N+2)\frac{(N+2)(N+1)}{2!} \\ & + q(N+3)\frac{(N+3)(N+2)(N+1)}{3!}\cdots \end{aligned} \tag{B15}$$

where the sum extends to the maximum number of runs of length M that is possible. Setting

$$\hat{Q}(N) = N!Q(N) \tag{B16}$$

$$\hat{q}(N) = N!q(N) \tag{B17}$$

Eq. (B15) takes the form

$$\begin{aligned} \hat{Q}(N) = {} & \hat{q}(N) + \frac{1}{1!}\hat{q}(N+1) + \frac{1}{2!}\hat{q}(N+2) \\ & + \frac{1}{3!}\hat{q}(N+3) + \cdots \end{aligned} \tag{B18}$$

Inverting these equations (by, for example, following the argument leading from Eqs. (A.10)–(A.14)) we have

$$\hat{q}(N) = \hat{Q}(N) - \hat{Q}(N+1) + \frac{1}{2}\hat{Q}(N+2) - \frac{1}{6}\hat{Q}(N+3)... \tag{B19}$$

Hence using Eqs. (B16) with (B17), (B19) and (B21) the numerator of Eq. (B2), that is $q(N)$, is obtained. It follows that the functional Ξ is given by

$$\Xi(Q) = (N!)^{-1}\sum a_n(N+n-1)!Q(N+n-1) \tag{B20}$$

where

$$a_n = (-1)^{n-1}\frac{1}{(n-1)!} \tag{B21}$$

To obtain the denominator of Eq. (B2), we consider the number of combinations in which D events can be arranged without regard to the length of runs but with an autocorrelation of r. We denote this number by P. However, P takes the form

$$P = J_1(K_2L_2 + K_3L_3) \tag{B22}$$

where J_1; K_2, K_3 and L_2, L_3 are factors as defined below.

The factor J_1 represents the number of ways in which the events $S_j = +1$ can be arranged. Thus

$$J_1 = (D/2)! \tag{B23}$$

Note that J_1 cancels E_1 in Eq. (B2).

The factors K_2 and K_3 represent the number of ways in which the runs of $S_j = -1$ events, as determined by the discrete autocorrelation coefficient, can be embedded in the possible locations between $S_j = +1$ events. The factor K_2 corresponds to the number of ways in which the runs can be embedded in the possible sites when the end sites are excluded. The factor K_3 corresponds to the case when the end sites are definitely included. Thus putting $\hat{N}_c = N + N_c$

$$K_2 = \frac{(D/2-1)!}{(D/2-1-\hat{N}_c)!\hat{N}_c!} \tag{B24}$$

for the case where the end sites are not included, and

$$K_3 = \frac{(D/2-1)!}{(D/2+1-\hat{N}_e)!(\hat{N}_e-2)!} \tag{B25}$$

for the case where the end sites are included.

Associated with these factors are the factors L_2 and L_3, that represent the number of ways in which the $S_j = -1$ events may be positioned within the sites just identified. In the case of the end sites being excluded, the events can be arranged in $(D/2)!$ ways and these arrangements must be partitioned into N_c non-empty combinations by $(N_c - 1)$ partitions in $(D/2 - 1)$ possible locations. Thus

$$L_2 = \frac{(D/2)!(D/2-1)!}{(D/2-\hat{N}_c)!(\hat{N}_c-1)!} \tag{B26}$$

Similarly,

$$L_3 = \frac{(D/2)!(D/2-1)!}{(D/2-\hat{N}_e)!(\hat{N}_e-1)!} \tag{B27}$$

Thus P is determined by Eq. (B22) and so by Eq. (B2) is $\Pr(D,M,N,r)$. Note that P is independent of N although in some problems allied to the present problem it may be a function of N. The expression for $\Pr(D,M,N,r)$. is complicated but it may be easily evaluated by machine. It is not constructive to present it in its canonical form. Instead, it is best to assemble it from its parts as required. Note that some cancellation occurs but this is slight. Most of the complication occurs because of the correlation that may occur between S_j and S_{j+1} as measured by r.

Thus

$$\Pr(D,M,N,r) = \frac{\Xi(Q)}{P} = \frac{\Xi(E_1E_2(F_3G_{3c}H_c + F_3G_{3e}H_e + F_4G_4H_e + F_5G_5H_e))}{J_1(K_2L_2 + K_3L_3)} \tag{B28}$$

Appendix C. Probability distribution for runs uniformly below the median in a serially correlated finite sequence (opposite behaviour at the ends)

We again consider all ordered record of D discrete events where each event is measured by a real number. Each event, say event j, is either greater than the median of D in which case we set $S_j = +1$, or less than the median in which case we set $S_j = -1$. (In the case of a record of odd length, one event is the median and this event is randomly given a S_j value of $S_j = -1$ or $S_j = +1$.) Consistent with the notation of Appendix B, we define a median autocorrelation coefficient, r, for a particular record by Eq. (1.1) in the text

$$r = \frac{1}{D-1}\sum_{j=1}^{D-1} S_j S_{j+1}. \tag{C1}$$

This appendix is concerned with determining the probability that a record drawn from random from D discrete events has exactly N runs of M events with $S_j = -1$, given the autocorrelation coefficient of the record. Note that by Eq. (C1) r can only take certain discrete values. These discrete values are different to those allowed in Appendix B because of the different behaviour at the ends of the sequences involved.

As in Appendix B, the present appendix is concerned only with the cases where D is even which by Eq. (C1) requires $(D-1)r$ to be odd. Simple interpolation is probably adequate in most hydrological cases of interest where the record is of odd length.

If the record is drawn without bias, then by definition of bias, the probability is

$$\Pr(D,M,N,r) = \frac{\text{number of combinations with } D,M,N,r}{\text{total number of combinations with } D,r} \tag{C2}$$

As in Appendix B, the determination of both the denominator and the numerator of the right hand side of this equation need to be considered in turn.

To obtain the numerator of Eq. (C2), we consider the number of combinations in which D events can be arranged with *at least* N runs of exactly length M. We denote this number by $Q(N)$. The numerator of Eq. (C2) is of the form $q(N) = \Xi(Q)$ where the functional Ξ is defined later but it is the same as in Appendix B. By contrast with Appendix B, the number $Q(N)$ now takes the form

$$Q(N) = E_1E_2(F_3G_{3h}H_h + F_4G_{4h}H_h) \tag{C3}$$

where E_1, E_2; F_3, F_4; G_{3h}, G_{4h} and H_h are factors as defined below.

The factor E_1 represents the number of combinations in which the $S_j = +1$ events can be ordered. There are $D/2$ such events. Thus

$$E_1 = (D/2)! \tag{C4}$$

The factor E_2 represents the number of ways in which the MN events involved in runs can be drawn from the $D/2$ events with $S_j = -1$ to form N unlabeled ordered groups. Thus

$$E_2 = \frac{(D/2)!}{(D/2-MN)!N!} \tag{C5}$$

The factors F_3 and F_4 represent the number of possible locations for imbedding the N runs of $S_j = -1$ events in the $D/2$ events with $S_j = +1$.

The factor F_3 corresponds to the case when $S_1 = +1$ and $S_D = +1$ after the runs have been imbedded. In this case, there are $(D/2-1)$ possible sites between the already imbedded events for the imbedding and so

$$F_3 = \frac{(D/2-1)!}{(D/2-1-N)!} \tag{C6}$$

The factor F_4 corresponds to the case when $S_1 = -1$ or $S_D = -1$ (but not both) after the imbedding of the N runs of $S_j = -1$ events. In each case, there are $(D/2-1)$ possible sites for the imbedding plus one site that must be used, and so

$$F_4 = 2N\frac{(D/2-1)!}{(D/2-N)!} \tag{C7}$$

The above factors are identical to those by the same name in Appendix B. However, the remaining factors vary from those given there.

The factors G_{3h} and G_{4h} represent the number of possible locations for imbedding the runs of the remaining $(D/2-MN)$ events of type $S_j = -1$. There are $(D/2+1-N)$ sites available to be filled with these events of type $S_j = -1$.

In the case where $S_1 = +1$ and $S_D = +1$ after the runs considered earlier were added, $a - 1$ must be produced at one but not both of the ends after any further imbedding. (The value of r imposes the case of these remaining events occupying only one of the end sites as noted earlier. Such cases are the subject of Appendix B.) If the correlation coefficient is to be r, then these events must occur in

$N_h = \frac{1}{4}(D-1)(1-r) - N + \frac{1}{2}$ runs within the available sites. It follows that for these two cases

$$G_{3h} = 2\frac{(D/2-1-N)!}{(D/2-N-N_h)!(N_h-1)!} \tag{C8}$$

If, however, $S_1 = +1$ and $S_D = -1$, or vice versa, after the runs considered earlier were added, all remaining events must be imbedded away from the ends and so

$$G_{4h} = \frac{(D/2-N)!}{(D/2-N-N_h)!N_h!} \tag{C9}$$

Either way, these sites must be filled with the remaining $(D/2-MN)$ events of type $S_j = -1$. If N_h sites are available, this can be done in H_h ways, where

$$H_h = \frac{(D/2-MN)!(D/2-MN-1)!}{(D/2-MN-N_h)!(N_h-1)!} \tag{C10}$$

Hence by Eq. (C3), $Q(N)$ and $Q(N+1)$ can be determined, and so $Q(N) - Q(N+1)$ can be determined.

As stated earlier the numerator of Eq. (C2) is given by $q = \Xi(Q)$. Out of a single realisation of $(N+1)$ runs it is possible to choose N runs in $(N+1)$ ways and similarly if $(N+2)$ runs occur, it is possible to choose N runs in $(N+1)(N+2)/2$ ways. Hence, as in Appendix A,

$$\begin{aligned} Q(N) = q(N) + q(N+1)\frac{(N+1)}{1!} \\ + q(N+2)\frac{(N+2)(N+1)}{2!} \\ + q(N+3)\frac{(N+3)(N+2)(N+1)}{3!}\cdots \end{aligned} \tag{C11}$$

where the sum extends to the maximum number of runs of length M that is possible. Setting

$$\hat{Q}(N) = N!Q(N) \tag{C12}$$

$$\hat{q}(N) = N!q(N) \tag{C13}$$

Eq. (B11) takes the form

$$\begin{aligned} \hat{Q}(N) = \hat{q}(N) + \frac{1}{1!}\hat{q}(N+1) + \frac{1}{2!}\hat{q}(N+2) \\ + \frac{1}{3!}\hat{q}(N+3) + \cdots \end{aligned} \tag{C14}$$

Hence inverting gives

$$\begin{aligned} \hat{q}(N) = \hat{Q}(N) - \hat{Q}(N+1) + \frac{1}{2}\hat{Q}(N+2) \\ - \frac{1}{6}\hat{Q}(N+3)... \end{aligned} \tag{C15}$$

Hence using Eqs. (C12) with (C13), (C15) the numerator of Eq. (C2), that is $q(N)$, is obtained. It follows that the functional Ξ is given by

$$\Xi(Q) = (N!)^{-1}\sum a_n(N+n-1)!Q(N+n-1) \tag{C16}$$

where

$$a_n = (-1)^{n-1}\frac{1}{(n-1)!} \tag{C17}$$

To obtain the denominator of Eq. (C2), we consider the number of combinations in which D events can be arranged without regard to the length of runs but with an autocorrelation of r. We denote this number by P. However, P takes the form

$$P = J_1K_4L_4 \tag{C18}$$

where J_1, K_4 and L_4 are factors as defined below.

The factor J_1 represents the number of ways in which the events $S_j = +1$ can be arranged. Thus, as in Appendix B,

$$J_1 = (D/2)! \tag{C19}$$

Note that J_1 cancels E_1 in Eq. (C2).

The factor K_4 represents the number of ways in which the runs of $S_j = -1$ events, as determined by the discrete autocorrelation coefficient, can be embedded in the possible locations between $S_j = +1$ events. The factor K_4 corresponds to the number of ways in which the runs can be embedded in the possible sites when one and only one of the end sites is included as in this appendix. Thus

$$K_4 = 2\frac{(D/2-1)!}{(D/2-\hat{N}_h)!(\hat{N}_h-1)!} \tag{C20}$$

where $\hat{N}_h$ for $N + N_h$.

Associated with this factor is the factor L_4, that represent the number of ways in which the $S_j = -1$ events may be positioned within the sites just identified. Thus

$$L_4 = \frac{(D/2)!(D/2-1)!}{(D/2-\hat{N}_h)!(\hat{N}_h-1)!} \tag{C21}$$

Thus P is determined by Eq. (C18) and so by Eq. (C2) is $\Pr(D,M,N,r)$. Note that, as in Appendix B, P is independent of N but, as in the case considered in that appendix, in some problems allied to the present problem it may be a function of N. Again, the expression for $\Pr(D,M,N,r)$ is complicated but it may be easily evaluated by machine, and it is not constructive to present it in its canonical form. Instead, it is best to assemble it from its parts as required. Note that some cancellation occurs but this is slight. As in Appendix B, most of the complication occurs because of the correlation that may occur between S_j and S_{j+1} as measured by r.

Thus

$$\Pr(D,M,N,r) = \frac{\Xi(Q)}{P} = \frac{\Xi(E_1E_2(F_3G_{3h}H_h + F_4G_{4h}H_h))}{J_1K_4L_4}. \tag{C22}$$

References

Conover, W.J., 1999. Practical Nonparametric Statistics, Third ed, Wiley, New York, 584 pp.

Feller, W., 1968. An Introduction to Probability Theory and its Application, Third ed, Wiley, New York, 62 pp.

Hannan, E.J., 1960. Time Series Analysis, Chapman & Hall, London, 151 pp.

Hollander, M., Wolfe, D.A., 1973. Nonparametric Statistical Methods, Wiley, New York, 503 pp.

Kantz, H., Schreiber, T., 1997. Nonlinear Time Series Analysis, Cambridge University Press, Cambridge, pp. 135–139; 304 pp.

Maritz, J.S., 1981. Distribution-free Statistical Methods, Chapman & Hall, New York, 264 pp.

Mills, F.C., 1965. Statistical Methods, Sir Isaac Pitman and Sons, London, 842 pp.

Neave, H.R., Worthington, P.L., 1988. Distribution-free Tests, Unwin Hyman, London, 430 pp.

Peel, M.C., Pegram, G.G.S., McMahon, T.A., 2003. Global analysis of runs of annual precipitation and runoff equal to or below the median. Submitted for publication.

Sprent, P., Smeeton, N.C., 2001. Applied Nonparametric Statistical Methods, Third ed, Chapman & Hall/CRC, Boca Raton, 461 pp.