



Temporal analysis of the frequency and duration of low and high streamflow: years of record needed to characterize streamflow variability

Seungho Huh^a, David A. Dickey^{b,*}, M.R. Meador^c, K.E. Ruhl^c

^aResearch Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, NC 27709, USA

^bDepartment of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203, USA

^cUS Geological Survey, 12201 Sunrise Valley Drive, MS 413, Reston, VA 20192, USA

Received 12 December 2003; revised 30 November 2004; accepted 15 December 2004

Abstract

A temporal analysis of the number and duration of exceedences of high- and low-flow thresholds was conducted to determine the number of years required to detect a level shift using data from Virginia, North Carolina, and South Carolina. Two methods were used—ordinary least squares assuming a known error variance and generalized least squares without a known error variance. Using ordinary least squares, the mean number of years required to detect a one standard deviation level shift in measures of low-flow variability was 57.2 (28.6 on either side of the break), compared to 40.0 years for measures of high-flow variability. These means become 57.6 and 41.6 when generalized least squares is used. No significant relations between years and elevation or drainage area were detected ($P > 0.05$). Cluster analysis did not suggest geographic patterns in years related to physiography or major hydrologic regions. Referring to the number of observations required to detect a one standard deviation shift as ‘characterizing’ the variability, it appears that at least 20 years of record on either side of a shift may be necessary to adequately characterize high-flow variability. A longer streamflow record (about 30 years on either side) may be required to characterize low-flow variability.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Time series analysis; Streamflow; Low flows; Floods; Trends

1. Introduction

Streamflow variability at the scale of seasons or years can be so great in many streams that it may be too difficult to detect trends even over a relatively long

period of record (Konrad and Booth, 2002). Alternatively, relatively short, multi-year sequences of wet or dry years can produce statistically significant trends, particularly when those sequences are at the start or end of a period of record (Wahl, 1998). Thus, the question of the number of years of record needed to detect a trend in streamflow variability is problematic.

Although a number of studies in the US have investigated trends in streamflow data, most have

* Corresponding author. Fax: +1 919 515 7591.

E-mail address: dickey@stat.ncsu.edu (D.A. Dickey).

focused on average monthly or annual discharge with temporal analyses of streamflow variability receiving relatively less attention (Lins and Slack, 1999; Sankarasubramanian and Vogel, 2002). Information to date suggests that the period of record recommended to detect a trend in various streamflow parameters ranges from 10 to 40 or more years. A minimum of 10 years of record has been suggested as necessary for defining the frequencies of low (Riggs, 1972) and high flows (Interagency Advisory Committee on Water Data, 1982). However, Konrad and Booth (2002) in an analysis of trends in average annual discharge and 7-day low flow in streams of western Washington concluded that analyses using a 10-year period of record could lead to spurious interpretations of trends. Richter et al. (1997) reported that the ranges of streamflow parameter estimates 'begin to narrow substantially' when based on at least 20 years of record. Many temporal studies of streamflow, typically based on average or median discharge, have been based on greater than 30 years of record (Lettenmaier et al., 1994; Lins and Slack, 1999; Douglas et al., 2000; McCabe and Wolock, 2002). Gan et al. (1991), using Colwell's indices of predictability, constancy, and contingency (Colwell, 1974) to assess trends in streamflow, suggested that about 40 years of record were needed to ensure stable characterizations of streamflow. However, relatively long periods of record may not be available for many stream-gaging stations. Because the number of sites available for studies related to characterization of streamflow variability is dependent on the length of record at each site, a better understanding is needed of the minimum number of years required to characterize streamflow variability and accurately detect trends, where they exist.

Trends in streamflow variability may vary from site to site because of environmental factors, further complicating the question of the number of years of record needed to detect a trend in streamflow variability. The influence of large-scale environmental factors such as climate on streamflow trends is well documented (Lins and Slack, 1999). Vogel et al. (1998) reported that in the relative absence of anthropogenic influences not only was streamflow variability heterogeneous among major hydrologic regions in the US, but also among many river basins. Douglas et al. (2000) suggested that geographic

variability in trends in low and high flows might be related to differences in drainage basin storage. Trends in streamflow variability have also been reported to vary with different types of streams (for example, perennial or intermittent) and physiography (Walker et al., 1995; Poff, 1996). Thus, to expand inferences from individual stream-gaged sites regarding the period of record needed to detect trends would require an understanding of environmental factors that may affect streamflow variability.

The goal of this study was to determine the minimum number of years of record required to detect a level shift in selected hydrological parameters used to characterize streamflow variability, using two statistical methods. Specifically, the following questions were addressed in this study: (1) How many years are needed to detect a certain mean change in a value if one occurs? (2) Are regional spatial patterns (among watersheds) evident in the number of years needed, such as relations with physiography or major hydrologic regions? and (3) Are the number of years needed related to natural environmental variables within watersheds, such as drainage area and elevation?

2. Streamflow data

For this study, we chose to use the frequency and duration of high- and low-flow events to characterize streamflow variability (Richter et al., 1996). Low-flow and high-flow thresholds were defined as the 25th and 75th percentiles of discharge for the entire period of continuous record, respectively. These threshold levels were then used to determine a numerical count of the number of low- (NL) and high-flow (NH) events per year, with an event defined as a period when streamflows are greater than the high-flow threshold or less than the low-flow threshold. The mean duration in days per year of the low- (DL) and high-flow (DH) events was also determined. As an illustration, Fig. 1 shows 1993 daily flows for the Salkehatchie River near Miley, South Carolina with the 25th and 75th flow percentiles indicated by horizontal lines. For example, the NH for 1993 would equal five, and consist of one high flow of long duration followed by four high flows of short duration. These streamflow parameters (NL, NH, DL,

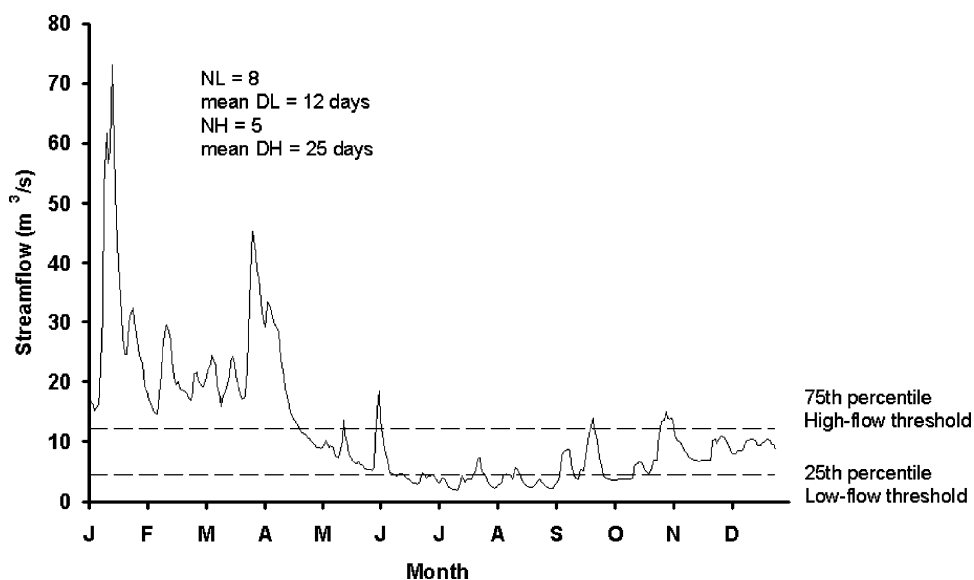


Fig. 1. Salkehatchie River data.

and DH) have been described as ecologically relevant hydrologic attributes that can be used to characterize streamflow variability (Richter et al., 1996). Ecological responses to altered patterns of these streamflow parameters are numerous (Poff et al., 1997) and include loss of riffle habitat for aquatic species, concentration or stranding of aquatic organisms, altered energy flows, and invasion and establishment of exotic species.

The streamflow data set consisted of records included as part of the Hydro-Climatic Data Network (HCDN), a database comprised of streamflow records from US Geological Survey stream gaging stations located on streams considered to be relatively free of anthropogenic influences (Slack et al., 1993). Records from 128 stations in North Carolina, South Carolina, and Virginia were examined. This three-state study area provided the opportunity to limit potential influences of broad geographic scale climatic factors on streamflow variability (Lins and Slack, 1999) while examining records from stations located within four different physiographic provinces—Coastal Plain, Piedmont, Blue Ridge, and Valley and Ridge (Fenneman, 1946)—and two major US hydrologic regions—the Mid-Atlantic and the South Atlantic-Gulf (US Water Resources Council, 1975). A final set of 50 stations was chosen for inclusion in this study based

on a minimum of 14 years of continuous streamflow record, an end-of-record date no earlier than 1985, and a drainage area less than 1036 km². Data for this study thus consisted of 200 (50 sites × 4 variables) different time series. The stations ranged in elevation from 0.6 to 490.8 m and from 62.4 to 883.2 km² in drainage area (Table 1).

The analyses that follow involve calculation of lag-one correlations, estimation of sample sizes, and examination of spatial patterns in estimated sample sizes from these records based on the longest continuous streamflow record available at a station. Begin and end dates of each record varied, and length of records ranged from 14 to 71 years (Table 1). Varying record lengths and dates do not affect the validity of sample size determination within a station, but they do affect precision as measured by the standard errors. Standard error calculations of sample size estimates are presented in Section 4.1. The standard errors are larger for the smaller numbers of observations.

2.1. Time series structure

Each of the 200 time series was evaluated to determine the appropriate Autoregressive Moving Average (ARMA) model. The first order

Table 1

Station number, physiographic province, major hydrologic region, years and period of stream flow record, 25th and 75th percentiles of flow (low-flow and high-flow thresholds, respectively), elevation, and drainage area for 50 gaged sites from Virginia, North Carolina, and South Carolina

No.	Station number	State	Physio-graphic province	Hydrologic region	Years of record	Begin date	End date	25th % (cm)	75th% (cm)	Elevation (m)	Drainage area (km ²)
1	01632000	VA	VAR	MA	71	Apr-25	Mar-95	0.42	5.04	320.6	543.9
2	01634500	VA	VAR	MA	58	Oct-37	Jul-95	0.48	2.75	197.2	266.8
3	01638480	VA	PP	MA	25	Oct-70	Jul-95	0.51	2.83	75.9	232.1
4	01643700	VA	PP	MA	29	Oct-65	Jul-95	0.62	4.48	100.3	318.6
5	01644000	VA	PP	MA	66	Apr-30	Jul-95	1.53	10.17	75.9	859.9
6	01646000	VA	PP	MA	61	Apr-35	Oct-94	0.62	1.73	46.1	149.9
7	01663500	VA	PP	MA	52	Aug-42	Oct-92	2.75	11.04	87.9	743.3
8	01665500	VA	PP	MA	53	Oct-42	Dec-94	1.27	5.04	133.9	295.3
9	01671100	VA	PP	MA	34	Oct-61	Oct-94	0.54	2.94	40.3	277.1
10	02015700	VA	VAR	MA	35	Aug-60	Sep-94	1.27	4.47	490.8	284.9
11	02017500	VA	VAR	MA	69	Oct-26	Jul-95	0.59	4.13	382.3	269.4
12	02020500	VA	VAR	MA	57	Oct-38	Oct-94	0.51	4.81	422.1	372.9
13	02027800	VA	PP	MA	36	Aug-60	Oct-94	1.76	5.41	135.5	380.7
14	02030500	VA	PP	MA	69	Apr-26	Sep-95	2.01	6.09	72.8	585.3
15	02041000	VA	PP	MA	49	Oct-46	Aug-95	1.02	4.11	54.0	409.2
16	02042500	VA	CP	MA	54	Apr-42	Aug-95	1.73	9.79	1.9	652.7
17	02044500	VA	PP	SAG	44	Oct-50	Aug-95	2.55	9.17	56.4	800.3
18	02046000	VA	PP	SAG	49	Oct-46	Aug-95	0.48	3.26	39.6	290.1
19	02052500	VA	PP	SAG	42	Oct-53	Aug-95	0.19	1.81	46.5	168.9
20	02053800	VA	BR	SAG	35	Oct-60	Jul-95	1.16	3.59	415.1	284.9
21	02061500	VA	PP	SAG	59	Apr-37	Jul-95	3.68	10.36	165.8	828.8
22	02065500	VA	PP	SAG	49	Oct-46	Aug-95	1.13	2.92	112.8	253.8
23	02069700	VA	PP	SAG	33	Oct-62	Jul-95	1.98	4.19	265.7	219.1
24	02070000	VA	PP	SAG	66	Oct-28	Aug-95	1.98	3.85	222.8	279.7
25	02082770	NC	PP	SAG	32	Aug-63	Sep-94	1.27	4.67	39.6	429.9
26	02082950	NC	PP	SAG	34	Oct-59	Sep-94	0.91	4.70	35.5	458.4
27	02083800	NC	CP	SAG	38	Dec-56	Sep-94	0.34	2.49	9.1	202.3
28	02088470	NC	CP	SAG	27	Aug-64	Oct-90	0.96	6.23	39.4	494.7
29	02091700	NC	CP	SAG	31	Oct-56	Sep-87	0.28	3.45	9.1	241.6
30	02092000	NC	CP	SAG	39	Feb-50	Sep-88	0.68	6.34	0.6	471.4
31	02092500	NC	CP	SAG	42	Jan-51	Sep-94	0.68	6.59	5.8	435.1
32	02106000	NC	CP	SAG	42	Feb-50	Sep-91	0.79	4.30	24.5	240.4
33	02112120	NC	PP	SAG	29	Apr-64	Sep-94	2.97	5.92	294.1	331.5
34	02112360	NC	PP	SAG	28	Apr-64	Sep-94	2.15	4.05	282.6	204.1
35	02113850	NC	PP	SAG	28	Apr-64	Sep-94	4.78	9.74	268.5	598.3
36	02108500	NC	CP	SAG	42	Jan-51	Sep-94	2.95	6.40	223.9	401.5
37	02129590	SC	CP	SAG	16	Oct-79	Sep-94	0.25	1.02	30.5	68.4
38	02131150	SC	CP	SAG	26	Nov-66	Sep-92	0.14	0.91	22.9	70.9
39	02131309	SC	PP	SAG	19	Aug-76	Sep-94	0.12	0.93	92.3	62.9
40	02135300	SC	CP	SAG	27	Jul-68	Sep-94	1.02	3.85	50.1	248.6
41	02143000	NC	PP	SAG	58	Aug-25	Sep-94	1.76	3.96	271.6	215.5
42	02143040	NC	PP	SAG	30	Oct-61	Sep-94	0.62	1.42	336.2	66.6
43	02149000	NC	BR	SAG	43	Jan-51	Sep-94	2.15	4.33	248.5	204.6
44	02152100	NC	PP	SAG	35	Mar-59	Sep-94	1.33	2.75	271.3	156.7
45	02153780	SC	PP	SAG	14	Oct-80	Sep-94	0.20	0.63	172.2	62.4
46	02157000	SC	PP	SAG	38	Oct-50	Sep-88	0.91	1.89	207.3	114.9
47	02175500	SC	CP	SAG	43	Feb-51	Sep-94	4.45	12.20	19.6	883.2
48	02176500	SC	CP	SAG	43	Feb-51	Sep-94	0.48	6.77	15.3	525.8
49	02197300	SC	CP	SAG	29	Jun-66	Sep-94	2.49	3.23	50.3	255.6
50	02197400	SC	CP	SAG	20	Mar-74	Sep-94	1.27	2.97	35.7	153.6

VAR, Valley and Ridge; PP, Piedmont Plateau; CP, Coastal Plain; BR, Blue Ridge; MA, Mid-Atlantic; and SAG, South Atlantic-Gulf.

autoregressive AR(1) model appeared to fit most time series (187 of 200) with no evidence of lack of fit, based on a χ^2 test on the residuals (Brocklebank and Dickey, 2003). Higher order models were indicated for 13 of the time series. However, for these time series the AR(1) captured nearly all of the correlation structure so the AR(1) model was used for all 200 time series. Lins (1985) and Vogel et al. (1998) also reported that streamflow variability time series were adequately described by an AR(1) model. Next, each time series was examined for possible linear trends. For most time series (183 of 200), the linear trend was insignificant ($P > 0.05$). The four series from a typical stream are shown in Fig. 2 and those from one of the worst ($P < 0.05$ for NH and DH) in Fig. 3. Even there, the trends did not seem dramatic and no detrending was done.

Numerous studies of streamflow series have used the approach developed by Hurst (1951) to examine

the long-term stochastic structure of streamflow. However, there is considerable debate regarding the validity and meaning of Hurst coefficients. Vogel et al. (1998) suggested that it is impossible to use these coefficients to infer long-range dependence of time series. The AR(1) model $Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t$ is considered stationary if $|\rho| < 1$. Estimates of $|\rho|$ for the four parameters ranged from 0 to 0.8 (Table 2). The mean $|\rho|$ for all 50 stations was 0.17 for NL, 0.20 for DL, 0.18 for NH and 0.14 for DH. Dickey and Fuller (1979, 1981) develop a test of the hypothesis $\rho = 1$, and rejection in favor of $\rho < 1$ is taken as evidence of stationarity. Using this, all of our series are found to be stationary. Vogel et al. (1998) examined streamflow variability using the HCDN and reported a mean $|\rho|$ regional value for the coefficient of variation of annual streamflow calculated for 18 hydrologic regions in United States of 0.22.

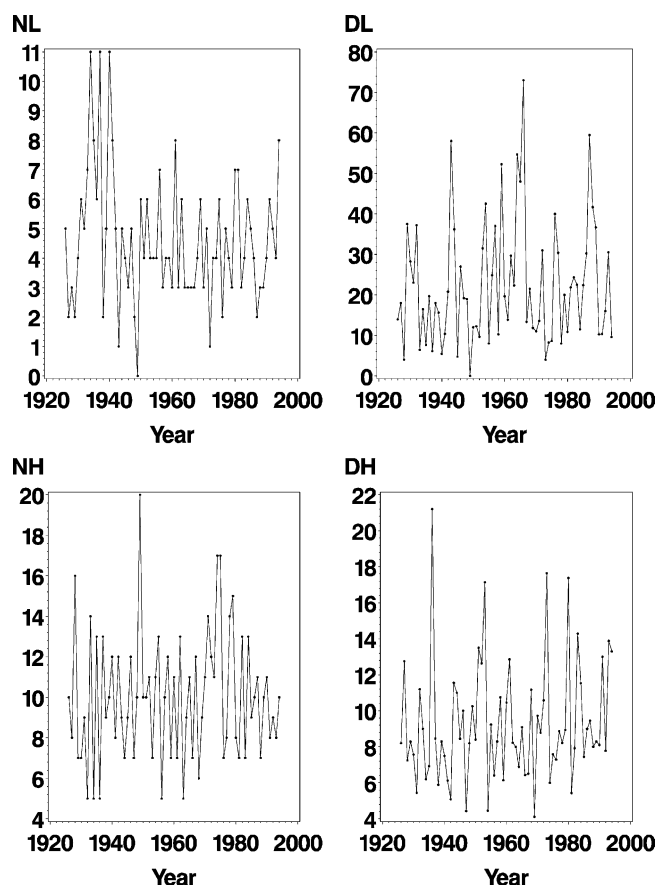


Fig. 2. NL, DL, NH, and DH for a typical stream.

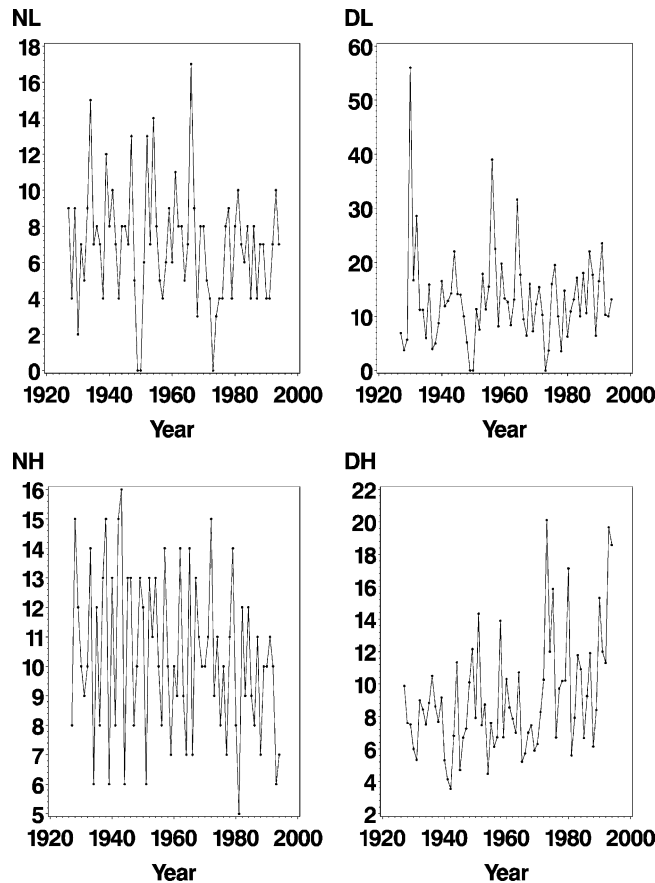


Fig. 3. NL, DL, NH, and DH for an atypical stream.

Whereas some studies have indicated that trends and non-stationarity exist in annual streamflow (e.g. Lins, 1985; Lettenmaier et al., 1994), it may not be surprising that the present study suggests stationary series for the parameters examined. First, the parameters in question are not annual streamflows, but rather exceedances. Also, the data were derived from a set of records for stations located on relatively small-sized streams in the eastern United States.

3. Sample sizes for detecting a level shift

To address the question, ‘how many samples are needed to detect a certain mean change in a value if one occurs?’ requires explanation of how one can detect level shifts in an otherwise stationary series. Studies of long-term changes in hydrologic variables

can be conducted by comparing two non-overlapping sets of data representing different time periods. The mean value of a given hydrologic variable Y can step up or down between time periods and such changes are thus known as step trends (Helsel and Hirsch, 1992) or level shifts. We consider level shifts that take place when a known event has occurred at a specific time in the streamflow record, for example, as a result of anthropogenic influences such as dam construction or a diversion, or natural factors such as climate shifts (McCabe and Wolock, 2002). The record is divided into time periods before and after this event. A two-sample t -test, accounting for autocorrelation, is conducted to compare means and the size of the level shift is determined.

In this study, stations were selected specifically because they did not have any known events that may have changed hydrologic variables over time.

Table 2
Estimates of the lag 1 autoregressive coefficient ρ for each station and target parameter

Station number	NL	DL	NH	DH	Station number	NL	DL	NH	DH
01632000	0.17	0.21	-0.22	-0.08	02082950	0.08	0.21	-0.15	-0.08
01634500	0.04	0.21	0.03	0.03	02083800	-0.02	0.08	0.11	-0.11
01638480	-0.03	0.10	0.16	0.13	02088470	-0.07	-0.39	0.01	-0.22
01643700	-0.19	0.01	0.06	-0.05	02091700	0.26	0.51	0.19	-0.16
01644000	0.16	0.02	0.15	-0.02	02092000	0.08	0.41	0.11	0.16
01646000	0.22	-0.01	0.40	0.09	02092500	0.01	0.14	0.19	-0.18
01663500	0.00	0.22	0.13	0.15	02106000	0.04	0.07	0.37	-0.12
01665500	0.01	0.07	0.05	0.12	02112120	0.27	0.24	0.17	0.23
01671100	0.26	0.16	0.08	0.19	02112360	0.33	0.16	0.22	0.30
02015700	0.23	0.71	0.16	0.26	02113850	0.11	-0.11	0.14	0.24
02017500	0.16	0.14	-0.28	0.33	02118500	0.25	0.24	0.10	0.17
02020500	-0.04	0.11	-0.10	-0.04	02129590	0.12	-0.36	-0.67	-0.19
02027800	0.29	0.30	-0.08	-0.01	02131150	0.07	-0.10	-0.40	0.08
02030500	0.32	0.21	0.10	0.23	02131309	-0.03	-0.38	0.59	-0.29
02041000	0.19	0.04	0.31	0.05	02135300	0.18	0.07	-0.32	-0.31
02042500	0.11	0.16	0.19	-0.01	02143000	0.41	0.41	0.10	0.01
02044500	0.21	-0.04	0.08	-0.06	02143040	-0.08	0.09	-0.02	-0.14
02046000	-0.10	0.04	0.05	-0.09	02149000	-0.10	0.16	0.08	0.06
02052500	-0.06	0.09	-0.10	-0.06	02152100	0.20	0.06	-0.14	-0.08
02053800	-0.05	0.09	0.09	0.17	02153780	0.12	0.41	-0.58	-0.25
02061500	0.18	0.18	-0.03	-0.04	02157000	0.28	0.17	0.37	0.21
02065500	0.50	0.45	0.06	0.21	02175500	0.30	0.12	0.14	0.08
02069700	0.25	0.35	0.24	0.24	02176500	0.18	0.07	-0.18	-0.05
02070000	0.33	0.14	-0.02	-0.03	02197300	0.80	0.58	0.12	0.39
02082770	-0.06	-0.03	0.04	0.15	02197400	0.11	0.11	0.30	0.08

NL, number of low-flow exceedences; DL, duration of low-flow exceedences; NH, number of high-flow exceedences; and DH, duration of high-flow exceedences.

These data thus provide baseline information on streams in relatively undisturbed states from which departures can be evaluated. Assuming that the variance and autocorrelation structure would not be affected by a level shift if it occurred, and given an experimental shift magnitude of interest and desired detection probability, the number of years necessary to detect a hypothetical level shift can be determined. Thus, the necessary sample size to detect a level shift of size δ with probability $1 - \beta$, can be determined for given values of δ and β . This is the problem of finding the sample size for which the hypothesis test of $H_0: \delta = 0$ has power $1 - \beta$ at a significance level α .

To solve this problem, two statistical approaches can be considered. In Section 3.1, an easy-to-use method using ordinary least squares (OLS) theory is presented. This method assumes a known error variance σ^2 . In Section 3.2, a fancier approach is presented using generalized least squares (GLS) without assuming σ^2 is known.

Besides these two approaches, we may also consider a non-parametric approach. This is a frequency domain method for which identification of the autocorrelation structure is not needed. For more details about using the frequency domain method, see Huh (2000).

Each method here considers a stationary AR(1) model that has a shift of size δ in level, and is based on the assumption that the shift occurs exactly at the midpoint of the data. This model can be represented as

$$Y_t = \beta_0 + \delta X_t + U_t \quad (1)$$

$$U_t = \rho U_{t-1} + e_t \quad |\rho| < 1, \quad t = 1, \dots, 2n$$

where $X_t = (0, \dots, 0, 1, \dots, 1)'$, $e_t \sim NI(0, \sigma^2)$, $U_1 \sim N(0, (\sigma^2 / (1 - \rho^2)))$, U_1 and e_t 's ($t > 1$) are independent, and β_0 is the level before the shift. It is not possible to catalog the effect of every possible change in a time series. We are simply assuming that the subseries, before and after the shift, differ only in their means, not in their

covariance structures. This assumption that the variance and autocorrelation are the same before and after the break is common in time series analysis. Should the variance increase after the shift, the required sample sizes would increase. Likewise a shift in the exact middle of a series would be easier to detect statistically than one near either end so again this assumptions delivers conservative estimates of sample size. The assumption of a sudden level shift rather than some more subtle and harder to detect change also produces conservative estimates of required sample sizes.

As to the assumption of a given level shift, this is necessary in any kind of power calculation. The power, or ability to detect a shift, must depend on the magnitude of the shift much as the detection of an earthquake would depend on its magnitude. We have simply computed the sample size necessary when a specified shift amount and desired detection probability are given. Knowledge of the shift magnitude is not a requirement for doing the test, but rather is used only in power computations.

3.1. Method I: ordinary least squares (OLS)

Because it is simple and well known, we first consider the ordinary least squares (OLS) approach. It is theoretically less efficient than generalized least squares for Model (1) since the U_t 's are autocorrelated, however, the difference is minor for these data.

The simple OLS estimator of δ is the difference of two means, $\hat{\delta} = \bar{Y}_2 - \bar{Y}_1$, where $\bar{Y}_1 = \frac{1}{n} \sum_{t=1}^n Y_t$ and $\bar{Y}_2 = \frac{1}{n} \sum_{t=n+1}^{2n} Y_t$. A proper test statistic for testing $H_0: \delta = 0$ is given as

$$\frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}}.$$

Thus, the per-group sample size n ($2n$ total observations) that satisfies

$$P\left(\frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}} > Z_\alpha\right) = 1 - \beta \tag{2}$$

can be determined, where Z_α is the standard Normal percentage point at a significance level α and β is a type II error probability. For independent data, the formula for n is well known. We consider modifications necessary with autocorrelated data.

First, $\text{Var}(\bar{Y}_2 - \bar{Y}_1)$ is determined. Letting $\gamma(h)$ denote the covariance between U_t and U_{t+h} , $\text{Var}(\bar{Y}_2 - \bar{Y}_1)$ can be written as

$$\text{Var}(\bar{Y}_2 - \bar{Y}_1) = \text{Var}(\bar{U}_2 - \bar{U}_1) = n^{-2} c' \Gamma c$$

where c is a column of 1s and -1 s and Γ is a matrix with (i,j) th element $[\Gamma]_{ij} = \gamma(i-j)$.

For model (1), $\text{Var}(\bar{Y}_2 - \bar{Y}_1)$ can be written as

$$\begin{aligned} \text{Var}(\bar{Y}_2 - \bar{Y}_1) &= \frac{2\gamma(0)}{n^2(1-\rho)^2} [(1-\rho^2)n - 3\rho + 4\rho^{n+1} - \rho^{2n+1}] \\ &= \frac{2\gamma(0)}{n} \left(\frac{1+\rho}{1-\rho}\right) + O(n^{-2}). \end{aligned}$$

The power under $H_a: \delta > 0$ for the one-sided test is

$$\begin{aligned} 1 - \beta &= P\left(\frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}} > Z_\alpha\right) \\ &= P\left(\frac{\bar{Y}_2 - \bar{Y}_1 - \delta}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}} > Z_\alpha - \frac{\delta}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}}\right) \\ &= P\left(Z > Z_\alpha - \frac{\delta}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}}\right). \end{aligned}$$

The values $\gamma(0)$ and ρ are assumed to be known and thus $\text{Var}(\bar{Y}_2 - \bar{Y}_1)$ is known. Therefore, to find the sample size that satisfies Eq. (2), the following equation is solved for n ,

$$Z_\alpha - \frac{\delta}{\sqrt{\text{Var}(\bar{Y}_2 - \bar{Y}_1)}} = Z_{1-\beta} = -Z_\beta$$

which becomes

$$\text{Var}(\bar{Y}_2 - \bar{Y}_1) = \left(\frac{\delta}{Z_\alpha + Z_\beta}\right)^2.$$

For a two-sided test with $H_a: \delta \neq 0$, Z_α is replaced by $Z_{\alpha/2}$ obtaining

$$\begin{aligned} &\frac{2\gamma(0)}{n^2(1-\rho)^2} [(1-\rho^2)n - 3\rho + 4\rho^{n+1} - \rho^{2n+1}] \\ &= \left(\frac{\delta}{Z_{\alpha/2} + Z_\beta}\right)^2 \tag{3} \end{aligned}$$

Letting $\delta = k\sigma$ and noting that $\gamma(0) = \sigma^2/(1 - \rho^2)$ for the AR(1) model gives

$$\frac{2}{n^2(1 - \rho)^2(1 - \rho^2)}[(1 - \rho^2)n - 3\rho + 4\rho^{n+1} - \rho^{2n+1}] = \left(\frac{k}{Z_{\alpha/2} + Z_\beta}\right)^2 \equiv V. \tag{4}$$

$$\begin{pmatrix} \sqrt{1 - \rho^2}Y_1 \\ Y_2 - \rho Y_1 \\ \vdots \\ Y_n - \rho Y_{n-1} \\ Y_{n+1} - \rho Y_n \\ Y_{n+2} - \rho Y_{n+1} \\ \vdots \\ Y_{2n} - \rho Y_{2n-1} \end{pmatrix} = \begin{pmatrix} \sqrt{1 - \rho^2}\beta_0 \\ (1 - \rho)\beta_0 \\ \vdots \\ (1 - \rho)\beta_0 \\ (1 - \rho)\beta_0 + \delta \\ (1 - \rho)\beta_0 + (1 - \rho)\delta \\ \vdots \\ (1 - \rho)\beta_0 + (1 - \rho)\delta \end{pmatrix} + \begin{pmatrix} \sqrt{1 - \rho^2}U_1 \\ e_2 \\ \vdots \\ e_n \\ e_{n+1} \\ e_{n+2} \\ \vdots \\ e_{2n} \end{pmatrix} = \begin{pmatrix} \sqrt{1 - \rho^2} & 0 \\ 1 - \rho & 0 \\ \vdots & \vdots \\ 1 - \rho & 0 \\ 1 - \rho & 1 \\ 1 - \rho & 1 - \rho \\ \vdots & \vdots \\ 1 - \rho & 1 - \rho \end{pmatrix} \begin{pmatrix} \beta_0 \\ \delta \end{pmatrix} + \varepsilon.$$

The sample estimates $\hat{\gamma}(0)$ and $\hat{\rho}$ can be substituted for the parameters $\gamma(0)$ and ρ . Eqs. (3) or (4) can then be solved for n using a search technique.

The terms ρ^{n+1} and ρ^{2n+1} are small for $|\rho| < 1$. Ignoring these, Eq. (4) reduces to a quadratic equation for n , namely

$$V(1 - \rho)^2(1 - \rho^2)n^2 - 2(1 - \rho^2)n + 6\rho = 0.$$

Solving for n ,

$$n = \frac{(1 + \rho) + \sqrt{(1 + \rho)^2 - 6V\rho(1 - \rho^2)}}{V(1 - \rho)(1 - \rho^2)}. \tag{5}$$

Eq. (5) should provide a good approximation of n without further adjustment if ρ is not too large.

3.2. Method II: generalized least squares (GLS)

In Section 3.1, we used a difference of two means adjusting the standard error for autocorrelation. We now consider a theoretically superior approach which is called generalized least squares (GLS). Here, the improvement is slight due to our small values of ρ .

In this section, GLS is implemented through the so-called ‘Cochrane–Orcutt’ transformation (Dinardo et al., 1996). Using GLS, the non-centrality parameter

of the test statistic for testing $H_0: \delta = 0$ is a function of the sample size n . Because the non-centrality parameter determines power, the sample size n can be determined once the desired value of the power is given.

Using Cochrane–Orcutt, the transformed model in matrix form is

This model is of the form

$$Z = X\beta + \varepsilon \tag{6}$$

where $\varepsilon \sim N(0, I\sigma^2)$ since $\sqrt{1 - \rho^2}U_1 \sim N(0, \sigma^2)$ and U_t is independent of the e_t 's for $t > 1$.

Now Model (6) satisfies the usual regression assumption that the error terms are independent and identically distributed as Normal random variables with mean zero and variance σ^2 . The hypothesis $H_0: \delta = 0$ can then be tested and the sample size that gives power $1 - \beta$ at significance level α can be determined. Note that the GLS estimate $\hat{\delta}_G$ is not exactly the same as the OLS estimate $\bar{Y}_2 - \bar{Y}_1$ but is asymptotically equivalent for $|\rho| < 1$. For small ρ as in our data, they are nearly equal in finite samples.

The test statistic for the hypothesis $H_0: \delta = 0$ can be shown as

$$F = \frac{\hat{\delta}_G^2/c_{11}}{\text{MSE}}$$

where c_{11} is the (2,2)th element of $(X'X)^{-1}$ and MSE is the error mean square (Rawlings et al., 1998). See the Appendix A for details of the derivation. This has a non-central F distribution with 1 and $2n - 2$ degrees

of freedom, and non-centrality parameter

$$\lambda = \frac{k^2\{(1 - \rho)^2n - \rho(\rho - 3)\}}{4}$$

for $\delta = k\sigma$.

Under the hypothesis $H_0: \delta = 0$, the test statistic F is distributed as the central F distribution with 1 and $2n - 2$ degrees of freedom. Using F_{α} , the critical value from the central F distribution at level α , the power for a certain δ is the probability that the non-central F exceeds F_{α} . Having found λ that delivers the desired power, the n that gives this value of λ can be computed. The power is a function of λ and thus in theory, of k , n and ρ . These theoretical power values are provided in Table 3. For any given value of ρ , power increases as n increases. For a fixed n , power decreases as ρ increases. Fig. 4 presents theoretical powers for some values of n and ρ with $k = 1$.

Table 3
Theoretical power values for ρ (ranging from -0.4 to 0.5), n (5, 10, 20, 30, 40 and 50), and $k = 1$ using Method II

ρ	n	Power	ρ	n	Power
-0.4	5	0.4396	0.1	5	0.2552
-0.4	10	0.8148	0.1	10	0.4915
-0.4	20	0.9886	0.1	20	0.7990
-0.4	30	0.9995	0.1	30	0.9316
-0.4	40	1.0000	0.1	40	0.9790
-0.4	50	1.0000	0.1	50	0.9940
-0.3	5	0.3978	0.2	5	0.2275
-0.3	10	0.7604	0.2	10	0.4232
-0.3	20	0.9765	0.2	20	0.7119
-0.3	30	0.9984	0.2	30	0.8709
-0.3	40	0.9999	0.2	40	0.9465
-0.3	50	1.0000	0.2	50	0.9791
-0.2	5	0.3580	0.3	5	0.2036
-0.2	10	0.6989	0.3	10	0.3595
-0.2	20	0.9551	0.3	20	0.6121
-0.2	30	0.9950	0.3	30	0.7818
-0.2	40	0.9995	0.3	40	0.8840
-0.2	50	1.0000	0.3	50	0.9409
-0.1	5	0.3207	0.4	5	0.1835
-0.1	10	0.6320	0.4	10	0.3027
-0.1	20	0.9204	0.4	20	0.5074
-0.1	30	0.9865	0.4	30	0.6674
-0.1	40	0.9980	0.4	40	0.7835
-0.1	50	0.9997	0.4	50	0.8633
0	5	0.2863	0.5	5	0.1672
0	10	0.5620	0.5	10	0.2542
0	20	0.8690	0.5	20	0.4065
0	30	0.9677	0.5	30	0.5387
0	40	0.9930	0.5	40	0.6489
0	50	0.9986	0.5	50	0.7377

4. Sample size estimates

Using Method I, the necessary per-group sample size n to detect a shift δ for a certain power $1 - \beta$ at significance level α can be obtained by using Eq. (5) when $\delta = k\sigma$. Recall there are n observations on either side of the break. Sample sizes n calculated using Method I, where $k = 1$ and $\beta = 0.2$, ranged from 9 years for NH to 368 years for NL (Table 4). Mean per-group sample sizes are $n = 29.7, 27.6, 21.2$ and 18.9 years for NL, DL, NH and DH, respectively. Mean values are 28.6 years for measures of low-flow variability (NL and DL combined) and 20.0 years for measures of high-flow variability (NH and DH combined). From Method II, the necessary per-group sample size n to detect a level shift of size $\delta = k\sigma$, where $k = 1$ and $\beta = 0.2$ ranged from 8 years for NH to 334 years for NL (Table 5). Mean sample sizes are $n = 29.7, 27.9, 21.9$ and 19.7 years for NL, DL, NH and DH, respectively. Mean values are 28.8 years for measures of low-flow variability (NL and DL combined) and 20.8 years for measures of high-flow variability (NH and DH combined). For either method, the estimate $\hat{\rho}$ is used as if it is the fixed parameter ρ . Using a non-parametric approach, the necessary sample size for $\delta = 3$, from Huh (2000), ranged from 10 years for NL, NH, and DH, to 520 years for DL.

It is seen from the tables that estimated per-group sample sizes required are very similar for most cases between OLS and GLS. Examined more closely, the estimates for GLS are slightly higher than those for OLS, which seems to contradict the superiority of GLS. The explanation for this is that the admission that σ^2 is unknown overwhelms the minor improvement in the estimation scheme. Since σ^2 would never be known in practice, we feel that sample sizes associated with GLS are more realistic and would recommend these if forced to choose between the two, almost identical, tables of sample sizes. We use the simpler OLS determined sample sizes in our later analyses, anticipating that users would prefer these. Analyses reported have been repeated with the GLS sample sizes, showing no substantial differences from the reported results. In performing a test for step change, one may as well use the simpler OLS tests regardless of how sample size was calculated,

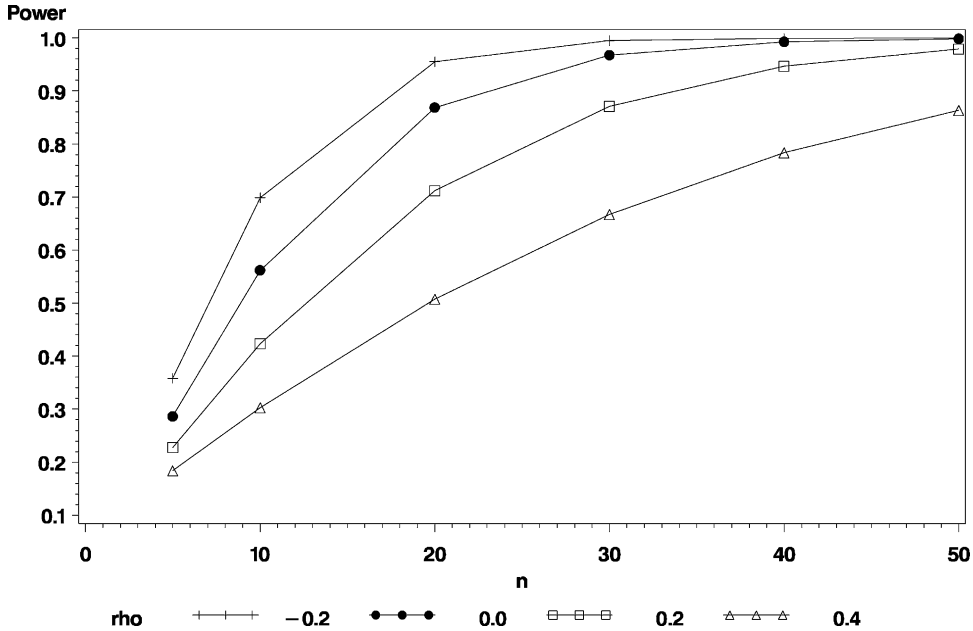


Fig. 4. Theoretical Powers for some values of n and ρ with $k=1$.

since the improvement associated with GLS appears so minor for the small ρ values we have here.

To determine sample sizes for each parameter across the entire data set of stations, a value $\hat{\rho}_0$ was used as a common estimate for all 50 stations. It was derived from the weighted average of the 50 individual estimates of $\hat{\rho}$. The number of observations in each station was used as the weighting factor and so

$$\hat{\rho}_0 = \frac{\sum_{j=1}^{50} T_j \hat{\rho}_j}{\sum_{j=1}^{50} T_j}$$

where T_j is the number of observations and $\hat{\rho}_j$ is the autocorrelation estimate for each station j .

First, the hypothesis $H_0: \rho_j = \rho_0$ is tested, where ρ_j is the lag 1 autocorrelation for station j and ρ_0 is a fixed value of ρ . In the AR(1) model, if $|\rho| < 1$,

$$\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{L} N(0, 1 - \rho^2) \text{ as } T \rightarrow \infty \quad (7)$$

or

$$\hat{\rho} \sim AN\left(\rho, \frac{1 - \rho^2}{T}\right)$$

where AN stands for Asymptotic Normal and T is the number of observations in a time series. Therefore,

under $H_0: \rho_j = \rho_0$ for $j = 1, \dots, 50$,

$$\frac{\hat{\rho}_j - \rho_0}{\sqrt{(1 - \rho_0^2)/T_j}} \quad (8)$$

is approximately distributed as $N(0,1)$ and so

$$\sum_{j=1}^{50} \frac{(\hat{\rho}_j - \rho_0)^2}{(1 - \rho_0^2)/T_j}$$

is approximately distributed as χ_{50}^2 . A weighted average $\hat{\rho}_0$ was used instead of ρ_0 in Eq. (8), to determine if

$$\hat{Z}_j \equiv \frac{\hat{\rho}_j - \hat{\rho}_0}{\sqrt{(1 - \hat{\rho}_0^2)/T_j}}$$

is $N(0,1)$ for $j = 1, \dots, 50$ or

$$\hat{H} \equiv \sum_{j=1}^{50} \frac{(\hat{\rho}_j - \hat{\rho}_0)^2}{(1 - \hat{\rho}_0^2)/T_j}$$

is distributed as χ_{49}^2 , where the degrees of freedom number is adjusted due to the constraint

$\sum_{j=1}^{50} T_j(\hat{\rho}_j - \hat{\rho}_0) = 0$. Values for ρ_j were not significantly different from $\hat{\rho}_0$ using the test statistic \hat{H} for each target variable (P -value = 0.63, 0.29, 0.05

Table 4

Sample sizes n (years of streamflow record) necessary for detecting a level shift ($k=1, \beta=0.2$, Method I) and standard errors in parentheses

Station Number	n_{NL}	n_{DL}	n_{NH}	n_{DH}	Station number	n_{NL}	n_{DL}	n_{NH}	n_{DH}
01632000	23 (6.54)	25 (7.56)	12 (2.06)	14 (2.95)	02082950	19 (6.94)	25 (10.6)	13 (3.52)	14 (4.28)
01634500	18 (4.75)	25 (8.14)	17 (4.59)	17 (4.59)	02083800	16 (4.8)	19 (6.71)	20 (7.16)	14 (3.74)
01638480	15 (6.03)	19 (8.83)	22 (10.72)	21 (9.8)	02088470	15 (5.17)	10 (2.18)	16 (6.38)	12 (3.35)
01643700	12 (3.68)	17 (6.51)	18 (7.43)	15 (5.49)	02091700	28 (13.41)	64 (41.53)	24 (10.6)	13 (3.6)
01644000	22 (6.53)	17 (4.22)	22 (6.38)	16 (3.66)	02092000	19 (6.54)	44 (22.33)	20 (7.26)	22 (8.5)
01646000	26 (8.51)	16 (3.94)	43 (17.5)	19 (5.35)	02092500	16 (5.05)	21 (7.58)	24 (8.91)	12 (2.94)
01663500	16 (4.48)	26 (9.12)	21 (6.69)	22 (7.25)	02106000	17 (5.52)	18 (6.09)	39 (18.48)	13 (3.49)
01665500	16 (4.44)	18 (5.37)	18 (5.0)	20 (6.34)	02112120	29 (15.16)	27 (13.59)	23 (10.28)	26 (12.87)
01671100	29 (13.26)	22 (8.96)	19 (6.95)	24 (10.07)	02112360	34 (18.85)	22 (9.99)	26 (12.45)	31 (16.7)
02015700	26 (11.29)	185 (158.1)	22 (9.02)	28 (12.65)	02113850	20 (8.41)	13 (4.33)	21 (9.35)	27 (13.54)
02017500	22 (6.36)	21 (5.95)	11 (1.76)	34 (11.99)	02118500	27 (11.11)	27 (10.98)	19 (6.61)	23 (8.38)
02020500	15 (3.73)	20 (5.9)	14 (3.13)	15 (3.71)	02129590	21 (11.94)	10 (2.98)	9 (1.28)	12 (4.71)
02027800	31 (14.58)	31 (14.76)	14 (4.21)	16 (5.26)	02131150	18 (7.76)	14 (4.69)	10 (2.12)	19 (8.15)
02030500	33 (11.44)	25 (7.58)	19 (5.16)	27 (8.32)	02131309	15 (6.68)	10 (2.6)	93 (89.49)	11 (3.29)
02041000	24 (8.35)	17 (5.06)	33 (13.26)	18 (5.22)	02135300	23 (11.07)	19 (7.68)	11 (2.55)	11 (2.61)
02042500	20 (6.12)	22 (7.17)	24 (8.09)	16 (4.26)	02143000	44 (18.51)	43 (18.4)	19 (5.71)	16 (4.32)
02044500	25 (9.39)	15 (4.26)	19 (6.09)	15 (3.98)	02143040	14 (4.69)	19 (7.93)	16 (5.54)	13 (3.99)
02046000	14 (3.37)	17 (5.15)	18 (5.33)	14 (3.44)	02149000	14 (3.67)	22 (8.17)	19 (6.17)	18 (5.81)
02052500	15 (4.16)	19 (6.45)	14 (3.7)	15 (4.09)	02152100	24 (10.33)	18 (6.49)	13 (3.6)	14 (4.23)
02053800	15 (4.64)	19 (7.1)	19 (7.18)	23 (9.32)	02153780	20 (12.03)	44 (37.59)	9 (1.73)	11 (4.21)
02061500	23 (7.29)	23 (7.47)	15 (3.82)	15 (3.75)	02157000	30 (13.19)	23 (8.91)	39 (18.85)	25 (9.98)
02065500	60 (30.81)	50 (24.16)	18 (5.53)	25 (9.05)	02175500	31 (13.28)	20 (7.03)	21 (7.66)	19 (6.2)
02069700	27 (12.68)	37 (19.26)	27 (12.3)	27 (12.2)	02176500	24 (8.78)	18 (5.99)	12 (2.93)	15 (4.14)
02070000	35 (12.35)	21 (6.06)	16 (3.62)	15 (3.6)	02197300	368 (418.66)	87 (64.98)	21 (8.78)	41 (23.72)
02082770	15 (4.73)	15 (5.19)	17 (6.36)	22 (8.97)	02197400	20 (10.18)	20 (10.1)	31 (19.79)	19 (9.22)

NL, number of low-flow exceedences; DL, duration of low-flow exceedences; NH, number of high-flow exceedences; and DH, duration of high-flow exceedences.

and 0.81 for NL, DL, NH and DH, respectively). In addition, the Kolmogorov test (SAS Institute Inc, 1989) was used with $\hat{\rho}_0$ in place of ρ_0 , and the hypothesis that the values for these \hat{Z}_j 's are $N(0,1)$ is not rejected. There does not appear to be any evidence against using the common $\hat{\rho}_0$ for generating sample sizes.

The weighted averages of $\hat{\rho}$'s with T_j 's as weights, for the four target variables were $\hat{\rho}_0 = \{0.14902, 0.15083, 0.055525, 0.043875\}$ for NL, DL, NH and DH, respectively. Using these $\hat{\rho}_0$ values in Eq. (5), Method I yields a set of per-group sample sizes n corresponding to NL, DL, NH and DH, respectively of 22, 22, 18, and 18 years for $k=1$. By using Method II for $k=1$, the set of sample sizes is $n=23, 23, 19,$ and 19 years for NL, DL, NH and DH, respectively.

If ρ is ignored (as if $\rho=0$) then the necessary per-group sample sizes are 16 and 17 in Method I and II, respectively, regardless of the parameters (NL, DL,

NH and DH). That is, if $\rho=0$, the sample sizes for all parameters (NL, DL, NH and DH) are equal since the difference comes from the values of ρ .

If we also assume unknown σ^2 in Method I, and use the t distribution instead of the Normal distribution, we have slightly different results. Under the same assumption, unknown σ^2 , the superiority of GLS to OLS is seen. In this new OLS table (not shown), the sample sizes required by OLS equal or exceed those of GLS but with few exceptions, the differences are very small. The only substantial differences from GLS are in stations with large ρ values (n_{DL} of Station 02015700, n_{NL} and n_{DL} of Station 02197300) as might be expected. In summary, all methods investigated lead to very similar sample size results.

It is worth mentioning that the theoretical superiority of GLS over OLS is proved assuming that ρ is known. Although the correct standard error for the OLS estimator involves ρ , the estimate itself

Table 5

Sample sizes n (years of streamflow record) necessary for detecting a level shift ($k=1$, $\beta=0.2$, Method II)

Station number	n_{NL}	n_{DL}	n_{NH}	n_{DH}	Station number	n_{NL}	n_{DL}	n_{NH}	n_{DH}
01632000	24	26	13	15	02082950	20	26	14	15
01634500	19	26	18	18	02083800	17	20	21	14
01638480	16	20	23	22	02088470	15	10	17	12
01643700	13	18	19	16	02091700	29	62	25	14
01644000	23	18	23	17	02092000	20	43	21	23
01646000	27	17	43	20	02092500	17	22	24	13
01663500	17	26	22	23	02106000	18	19	39	14
01665500	17	19	19	21	02112120	30	29	23	27
01671100	29	23	20	25	02112360	34	23	26	32
02015700	27	171	23	29	02113850	21	14	22	28
02017500	23	22	12	35	02118500	28	28	20	24
02020500	16	21	15	16	02129590	21	11	8	13
02027800	31	32	15	17	02131150	19	15	10	20
02030500	33	26	20	27	02131309	16	10	89	11
02041000	25	18	33	19	02135300	24	20	11	11
02042500	21	23	25	17	02143000	43	43	20	17
02044500	26	16	20	16	02143040	16	20	17	14
02046000	15	18	19	15	02149000	15	23	20	19
02052500	16	20	15	16	02152100	25	19	14	15
02053800	16	20	20	24	02153780	21	44	9	12
02061500	24	24	16	16	02157000	30	24	39	26
02065500	59	49	19	26	02175500	32	21	22	20
02069700	28	37	28	27	02176500	24	19	13	16
02070000	35	22	17	16	02197300	334	82	22	41
02082770	16	16	18	23	02197400	21	21	32	20

NL, number of low-flow exceedences; DL, duration of low-flow exceedences; NH, number of high-flow exceedences; and DH, duration of high-flow exceedences.

(difference of two simple averages) does not. This is in contrast to GLS, where the uncertainty in the estimate of ρ affects the estimated shift in mean as well as its standard error. This, along with the added complexity of running GLS and the relatively small improvement, even in theory, provided by GLS when ρ is small lead us to recommend OLS, the simple difference of two averages, with the appropriately adjusted standard errors.

The t test is approximately normal for reasonably large sample sizes regardless of the distribution of the original data, according to the central limit theorem. The time series results we use here are similarly based on large sample theory and do not require normality of the data. The sample sizes found here, in the neighborhood of 30, and autocorrelations well within the stationarity region should make the normal approximation reasonable. We note that Pekarova et al. (2003), using other methods, find similar sample size requirements for annual discharge series.

The choice of a one standard deviation shift is arbitrary. It is easy to use formulas (4) and (5) to investigate other k values. Over the range $k=0.1$ to 1 and using autocorrelation 0.15, the required sample sizes are between 35.5 and 38.5 percent higher than those for uncorrelated series. Thus a rule of thumb 40% increase for NL and DL due to their autocorrelations might be used. From formula (5), the required sample size on either side of the break is 87 for $k=1/2$, for example, so detecting smaller changes for NL and DL can require much larger sample sizes. A 10% increase in n over that required for independent data seems a good rule of thumb for NH and DH based on their average autocorrelations.

4.1. Measure of accuracy

To gain further insight into the accuracy of the sample sizes obtained above, a standard linearization technique was used for Method I. From the dominant

terms in Eq. (4),

$$n = \frac{2}{V(1-\rho)^2} \equiv h(\rho).$$

By Taylor series approximation,

$$\hat{n} = h(\hat{\rho}) = h(\rho) + h'(\rho)(\hat{\rho} - \rho) + R$$

where R is an appropriate remainder term and

$$h'(\rho) = \frac{4}{V(1-\rho)^3}.$$

Assuming R approximates 0,

$$\hat{n} - n = h'(\rho)(\hat{\rho} - \rho).$$

Therefore, by Eq. (7),

$$\text{Var}(\hat{n}) = [h'(\rho)]^2 \text{Var}(\hat{\rho}) = \left[\frac{4}{V(1-\rho)^3} \right]^2 \frac{1-\rho^2}{T}$$

and so

$$\begin{aligned} \text{standard error}(\hat{n}) &= \frac{4}{V(1-\rho)^3} \frac{\sqrt{1-\rho^2}}{\sqrt{T}} \\ &= \frac{4\sqrt{1-\rho^2}(Z_{\omega/2} + Z_{\beta})^2}{(1-\rho)^3 k^2} \frac{1}{\sqrt{T}}. \end{aligned}$$

That is, the standard error of \hat{n} is inversely proportional to the square root of T for given k and ρ . The standard error of \hat{n} for each station and target variable was then calculated and ranged from 1.28 (n_{NH} of station 02129590) to 418.66 (n_{NL} of station 02197300) for sample size estimates provided in Table 4. Table 4 shows the standard errors in parentheses. For $\alpha=0.05$, $\beta=0.2$ and the maximum $\hat{\rho}_0 (=0.151)$, $\text{standard error}(\hat{n}) = 50.175/(\sqrt{T}k^2)$. Furthermore,

$$\begin{aligned} \frac{\text{standard error}(\hat{n})}{\hat{n}} &= \frac{(1+\rho)}{(1+\rho) + \sqrt{(1+\rho)^2 - 6V\rho(1-\rho^2)}} \frac{4\sqrt{1-\rho^2}}{(1-\rho)\sqrt{T}} \end{aligned}$$

and standard error $(\hat{n})/\hat{n}$ varied between 0.157 (n_{NH} of station 02129590) and 1.139 (n_{NL} of station 02197300) for sample size estimates provided in Table 4. Thus, the standard errors of \hat{n} should be between 15.7 and 113.9% of \hat{n} .

As for the sample size estimates obtained from Method II, we cannot apply the linearization technique as above since there is no explicit formula for n . However, we can still get an approximate confidence interval for n .

If a 95% confidence interval for ρ is given as $\hat{\rho}_L \leq \rho \leq \hat{\rho}_U$, then the corresponding 95% confidence interval for the power can be obtained for a fixed n using powers in Table 3. Furthermore, a 95% confidence interval for a sample size n ($\hat{n}_L \leq n \leq \hat{n}_U$) can also be obtained for a fixed power. For example, suppose a 95% confidence interval for ρ is (0.3,0.4). Then for $n=20$ years before and after the break, a 95% confidence interval for the power is (0.5074, 0.6121) and, for power 0.8, a 95% confidence interval for n is available by an interpolation.

5. Sample sizes and watershed variables

Regression analysis was conducted to evaluate relations between sample sizes, using Method I, for each parameter (NL, DL, NH, and DH) and watershed variables such as elevation and drainage area. For varying values of δ or k , the P -values for the model F -test ranged from 0.08 to 0.78 and thus none were significant at $\alpha=0.05$. Therefore, there was no significant evidence to suggest that elevation or drainage area were related to the sample sizes for these stations. The same conclusion is reached if Method II sample sizes are used.

6. Sample sizes and geographic patterns

A cluster analysis of sample size estimates was conducted to assess patterns that might be associated with patterns in geographic locations of stations. The analysis was conducted using a single linkage technique based on vectors of the four sample sizes (for NL, DL, NH, and DH) calculated using Method I in Section 4. Because cluster analysis is a hierarchical procedure that continues until only one cluster is left, maximization of the 'Cubic Clustering Criterion' (CCC) (SAS Institute Inc., 1989) was used to determine the appropriate number of clusters to be examined. Patterns in descriptive statistics or sample sizes did not appear to reflect patterns related to

physiographic provinces or hydrologic regions. However, caution must be used in interpreting clustering results since the standard errors of sample size estimates are different among the stations as shown in Section 4.1. The same conclusion is reached if Method II sample sizes are used.

7. Conclusions

This study has attempted to determine the minimum number of years of record required to detect level shifts in four selected hydrological parameters used to characterize streamflow variability, using two statistical methods. The parameters selected, the number and duration of exceedences of high- and low-flow thresholds, have been considered to be more ecosystem-relevant than many traditional hydrological parameters (Richter et al., 1996). Temporal analyses of other hydrological parameters may yield different results.

A database of daily streamflows from stations relatively free of anthropogenic impacts was examined and 50 gaging stations in Virginia, North Carolina, and South Carolina were selected. The AR(1) model was found to fit well for the majority of series and linear trends did not appear to exist in most cases. All series were determined to be stationary by the Dickey–Fuller test.

The problem of finding the minimum number of years of streamflow record (sample size) needed to detect a level shift in a stationary series is discussed through two different approaches—ordinary least squares assuming a known error variance and generalized least squares without a known error variance. Measures of accuracy for the sample size estimates are determined and found to be inversely proportional to the square root of the number of observations.

Ordinary least squares with appropriately adjusted standard errors seems to be the most appealing of the methods tried for testing for a step change, based on its simplicity and the similarity of results to those of the theoretically superior GLS. We have illustrated results from OLS here, but results from GLS are substantially the same. The mean sample size estimates determined by generalized least squares were 57.6 years (28.8 before and 28.8 after the break)

for measures of low-flow variability (number and duration of low-flow exceedences combined) and $2 \times 20.8 = 41.6$ years for measures of high-flow variability (number and duration of high-flow exceedences combined). These means become 57.2 and 40.0 when OLS is used.

A cluster analysis was conducted to obtain clusters based on the sample sizes. Based on this analysis, there was no evidence of geographic patterns in the number of years related to physiography or major hydrologic regions. Regression analyses also did not detect any significant relations between number of years and elevation or drainage area.

These results have important implications for studies attempting to characterize streamflow variability. The number of sites available for future studies related to the characterization of streamflow variability is dependent on the length of streamflow record available at each site. Whereas at least 40 years of streamflow record may be necessary to adequately characterize high-flow variability for the stations examined in this study, a longer streamflow record (about 60 years) may be required to characterize low-flow variability.

Acknowledgements

We thank Gary Tasker and Jurate Landwehr whose comments helped improve the content and presentation of the paper.

Appendix A

In the theory of testing, the general linear hypothesis, the sum of squares for the linear hypothesis $H_0: K'\beta = m$ is $Q = (K'\hat{\beta} - m)'[K'(X'X)^{-1}K]^{-1}(K'\hat{\beta} - m)$. Furthermore, a proper test statistic for $H_0: K'\beta = m$ is

$$F = \frac{Q/r(K')}{\text{MSE}}$$

where r is the rank of a matrix (K') and MSE is the error mean square (Rawlings et al., 1998). Note that the test statistic F is obtained without assuming that σ^2 is known.

Thus,

$$K' = (0 \quad 1), \quad \beta = \begin{pmatrix} \beta_0 \\ \delta \end{pmatrix}$$

and

$$m = 0$$

so $r(K') = 1$ and $Q = \hat{\delta}_G^2/c_{11}$, where c_{11} is the (2,2)th element of $(X'X)^{-1}$.

The test statistic for the hypothesis $H_0: \delta = 0$, therefore, becomes

$$F = \frac{\hat{\delta}_G^2/c_{11}}{\text{MSE}}.$$

This has a non-central F distribution with 1 and $2n-2$ degrees of freedom and non-centrality parameter

$$\lambda = \frac{\delta^2}{2c_{11}\sigma^2} = \frac{k^2}{2c_{11}}$$

where $\delta = k\sigma$. By algebra,

$$c_{11} = \frac{2}{(1-\rho)^2n - \rho(\rho-3)}$$

and so

$$\lambda = \frac{k^2\{(1-\rho)^2n - \rho(\rho-3)\}}{4}$$

References

- Brocklebank, J.C., Dickey, D.A., 2003. SAS System for Forecasting Time Series. SAS Institute Inc., Cary, North Carolina.
- Colwell, R.K., 1974. Predictability, constancy, and contingency of periodic phenomena. *Ecology* 55, 1148–1153.
- Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–431.
- Dickey, D.A., Fuller, W.A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–1072.
- Dinardo, J., Johnston, J., Johnston, J., 1996. *Econometric Methods*. McGraw-Hill, New York.
- Douglas, E.M., Vogel, R.M., Kroll, C.N., 2000. Trends in floods and low flows in the United States: impact of serial correlation. *Journal of Hydrology* 240, 90–105.
- Fenneman, N.M., 1946. Physical divisions of the United States, US Geological Survey, scale 1:7,000,000.
- Gan, K.C., McMahon, T.A., Finlayson, B.L., 1991. Analysis of periodicity in streamflow and rainfall data by Colwell's indices. *Journal of Hydrology* 123 (1), 105–118.
- Helsel, D.R., Hirsch, R.M., 1992. *Statistical methods in water resources*. Elsevier, Amsterdam p. 529.
- Huh, S., 2000. Sample size determination and stationarity testing in the presence of trend breaks, Ph.D. dissertation, North Carolina State University, NC.
- Hurst, H.E., 1951. Long-term storage capacity of reservoirs. *Transactions of American Society Civil Engineering* 116, 770–808.
- Interagency Advisory Committee on Water Data, 1982. Guidelines for determining flood flow frequency, Bulletin 17-B of the Hydrology Subcommittee. US Geological Survey, Reston, VA p. 183.
- Konrad, C.P., Booth, D.B. 2002. Hydrologic trends associated with urban development for selected streams in the Puget Sound Basin, western Washington. US Geological Survey Water-Resources Investigations Report 02-4040, Tacoma, Washington.
- Lettenmaier, D.P., Wood, E.F., Wallis, J.R., 1994. Hydroclimatological trends in the continental United States, 1948–1988. *Journal of Climate* 7, 586–607.
- Lins, H.F., 1985. Streamflow variability in the United States. *Journal of Climate and Applied Meteorology* 24, 463–471.
- Lins, H.F., Slack, J.R., 1999. Streamflow trends in the United States. *Geophysical Research Letters* 26 (2), 227–230.
- McCabe, G.J., Wolock, D.M., 2002. A step increase in streamflow in the conterminous United States. *Geophysical Research Letters* 2185 (24), 2185.
- Pekarova, P., Miklanek, P., Pekar, J., 2003. Spatial and temporal runoff oscillation analysis of the main rivers of the world during the 19–20th centuries. *Journal of Hydrology* 274, 62–74.
- Poff, N.L., 1996. A hydrogeography of unregulated streams in the United States and an examination of scale-dependence in some hydrological descriptors. *Freshwater Biology* 36, 71–91.
- Poff, N.L., Allan, J.D., Bain, M.B., Karr, J.R., Prestegard, K.L., Richter, B.D., Sparks, R.E., Stromberg, J.C., 1997. The natural flow regime: a paradigm for river conservation and restoration. *BioScience* 47, 769–784.
- Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. *Applied Regression Analysis: A Research Tool*. Springer, New York.
- Richter, B.D., Baumgartner, J.V., Powell, J., Braun, D.P., 1996. A method for assessing hydrologic alteration within ecosystems. *Conservation Biology* 10, 1163–1174.
- Richter, B.D., Baumgartner, J.V., Wigington, R., Braun, D.P., 1997. How much water does a river need?. *Freshwater Biology* 37, 231–249.
- Riggs, H.C., 1972. Low-flow investigations, Techniques of Water Resource Investigation of the US Geological Survey. US Geological Survey, Reston, VA (Book 4, Chapter B1).
- Sankarasubramanian, A., Vogel, R.M., 2002. Annual hydroclimatology of the United States. *Water Resources Research* 38 (6).
- SAS Institute Inc., 1989. SAS/STAT User's Guide, Version 6, vol. 1. SAS Institute Inc., Cary, North Carolina.

- Slack, J.R., Lumb, A.M., Landwehr, J.M., 1993. Hydro-climatic data network (HCDN): streamflow data set 1874–1988, [CD-ROM], US Geological Survey. Water Resource Investigation Report 1993, 93–4076.
- US Water Resources Council, 1975. The Nation's Water resources 1975–200, Second National Assessment by the US Water Resources Council, Analytical Data, Appendix V, Streamflow Conditions. US Gov. Print. Off., Washington, DC pp. 276.
- Vogel, R.M., Yushiou, T., Limbrunner, J.F., 1998. The regional persistence and variability of annual streamflow in the United States. *Water Resources Research* 34 (12), 3445–3459.
- Wahl, K.L., 1998. Selectivity of non-parametric trend analyses to multi-year extremes, *Proceedings of the Western Snow Conference* 1998 pp. 157–160.
- Walker, K.F., Sheldon, F.F., Puckridge, J.T., 1995. A perspective on dryland river ecosystems. *Regulated Rivers: Research and Management* 11, 85–104.