Journal of Hydrology 530 (2015) 117–126

Contents lists available at ScienceDirect

Journal of Hydrology

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

Hypothesis tests for hydrologic alteration

Charles N. Kroll [a,*], Kelly E. Croteau [a,b], Richard M. Vogel [c]

[a] Environmental Resources Engineering, SUNY ESF, Syracuse, NY, United States
[b] Windward Environmental, Syracuse, NY, United States
[c] Civil and Environmental Engineering, Tufts University, Medford, MA, United States

ARTICLE INFO

Article history:
Received 26 June 2015
Received in revised form 30 August 2015
Accepted 23 September 2015
Available online 28 September 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Felix Frances, Associate Editor

Keywords:
Flow duration curves
Hydrologic alteration
Hypothesis testing

SUMMARY

Hydrologic systems can be altered by anthropogenic and climatic influences. While there are a number of statistical frameworks for describing and evaluating the extent of hydrologic alteration, here we present a new framework for assessing whether statistically significant hydrologic alteration has occurred, or whether the shift in the hydrologic regime is consistent with the natural variability of the system. Four hypothesis tests based on shifts of flow duration curves (FDCs) are developed and tested using three different experimental designs based on different strategies for resampling of annual FDCs. The four hypothesis tests examined are the Kolmogorov–Smirnov (KS), Kuiper (*K*), confidence interval (CI), and ecosurplus and ecodeficit (Eco). Here 117 streamflow sites that have potentially undergone hydrologic alteration due to reservoir construction are examined. 20 years of pre-reservoir record is used to develop the critical value of the test statistic for type I errors of 5% and 10%, while 10 years of post-alteration record is used to examine the power of each test. The best experimental design, based on calculating the mean annual FDC from an exhaustive jackknife resampling regime, provided a larger number of unique values of each test statistic and properly reproduced type I errors. Of the four tests, the CI test consistently had the highest power, while the *K* test had the second highest power; KS and Eco always had the lowest power. The power of the CI test appeared related to the storage ratio of the reservoir, a rough measure of the hydrologic alteration of the system.

1. Introduction

River systems provide an array of services to humans and the environment. These systems are used to meet human needs, including domestic, industrial, and agricultural water supplies, waste disposal, hydropower, and recreational activities. They also provide critical habitat for many aquatic and non-aquatic species. Anthropogenic activities and climatic influences can alter the hydrology of river systems. Activities such as damming a river, water discharges and withdrawals, and regional variations in climate can alter the hydrologic system and impact freshwater biodiversity and ecosystem services (Bunn and Arthington, 2002; Magilligan and Nislow, 2005; Gao et al., 2009). There is clearly tremendous interest in understanding ecological responses to altered flow regimes as evidenced by the hundreds of citations to the recent review article on this topic by Poff and Zimmerman (2010).

A wide variety of metrics has been developed to assess changes in hydrologic systems (Olden and Poff, 2003; Gao et al., 2009). A common set of metrics is the Nature Conservancy's Indicators of Hydrologic Alteration (IHA), which describe changes in 33 hydrologic statistics that characterize a wide array of hydrologic function (Richter et al., 1996). The IHA are often used to assess the impact of human activities on hydrology and to determine environmental flow recommendations for water managers. The ecological limits of hydrologic alteration (ELOHA) provide a framework for linking statistics such as those in the IHA to critical ecological responses (Poff et al., 2010). In the ELOHA framework, relationships between altered flow and ecological characteristics are empirically developed using existing and newly collected field data (Arthington et al., 2006). Similarly, numerous empirical multivariate relationships have been developed which characterize the impact of various anthropogenic influences on streamflow regimes ranging from flood regimes (Fitzhugh and Vogel, 2011) and low flow regimes (Homa et al., 2013) to the entire flow regime (McManamay, 2014). Fitzhugh (2014, page 826) reviews numerous recent studies which have sought to characterize alteration of a streamflow regimes over regions of the U.S.

Even with the broad suite of metrics of hydrologic change which have been introduced, as well as numerous empirical multivariate statistical models of the relationship between streamflow

* Corresponding author.
E-mail addresses: cnkroll@esf.edu (C.N. Kroll), kmetz878@gmail.com (K.E. Croteau), richard.vogel@tufts.edu (R.M. Vogel).

http://dx.doi.org/10.1016/j.jhydrol.2015.09.057
0022-1694/

regimes and anthropogenic factors, it is often difficult to assess whether changes to the hydrologic system are significant or are instead simply a result of the natural variability of streamflow under stationary conditions (Burn and Hag Elnur, 2002). While one could perform a hypothesis test to determine if there were significant changes in particular IHA statistics (Magilligan and Nislow, 2005) or other relevant hydrologic statistics, it is more challenging to assess changes to the complete streamflow regime. While many previous studies have examined tests of trends (e.g. Douglas et al., 2000; Burn and Hag Elnur, 2002) and shifts (e.g. Salas and Boes, 1980; Buishand, 1984) in hydrologic series, here we explore tests to assess the significance of an alteration to the complete hydrologic series.

One tool utilized in this experiment is the river's flow duration curve (FDC) (Foster, 1924; Searcy, 1959). The FDC is a plot of typically mean daily streamflow versus the probability of exceeding that streamflow. The FDC covers the entire range of streamflow magnitudes, and thus integrates the complete streamflow regime into a single tool. FDCs have been employed for a wide range of applications, including hydropower design, habitat assessment, flood abatement, water quality evaluation and for comparative hydrologic assessments (Vogel and Fennessey, 1995; Castellarin et al., 2013). An FDC is a convenient tool for observing and understanding hydrologic change. For instance, if a reservoir was placed in a river, one might expect a flattening of the FDC, where the higher flows are reduced (due to flood storage) and the lower flows are increased (by augmenting low flows with reservoir releases). Water withdrawals or reduction in precipitation would generally result in a lowering of the entire FDC. Botter et al. (2008) provide a theoretical linkage between the structure of FDC's and underlying ecohydrological, climatic and other watershed processes. Castellarin et al. (2013) provide a detailed review of the influence of a variety of natural and anthropogenic influences on FDC's and associated hydrologic processes.

FDCs and FDC statistics can be employed for assessing hydrologic alteration. Using shifts in FDCs, Vogel et al. (2007) defined the ecodeficit and ecosurplus as the percent loss or gain in streamflow due to flow regulation. Gao et al. (2009) employed a principal component analysis to examine how IHA statistics were related to ecodeficit and ecosurplus. Homa et al. (2013) developed regional regression models for quantiles of an FDC at altered streamflow sites in Massachusetts. Similarly, Mejia et al. (2014) derived a stochastic model of FDC's suitable for 11 urbanizing Washington, DC–Baltimore basins based on the stochastic properties of rainfall and various watershed properties.

Here we develop four new hypothesis tests of hydrologic alteration based on shifts in FDCs. Developing a hypothesis test based on the complete FDC, instead of a single hydrologic statistic, poses a unique challenge requiring a sampling strategy to implement each hypothesis test. Here we are faced with evaluating both the effectiveness of a number of different sampling strategies in addition to the power of the resulting hypothesis tests. Each hypothesis test is defined by its 'test statistic', whereas the sampling strategies are held fixed across all hypothesis tests considered. The hypothesis tests are developed similarly to common tests of a change in the probability distribution of a series. Two of the tests are based on deviations between the cumulative distribution functions (cdfs) of the FDCs. The test statistics for these tests are similar to those employed in the well-known Kolmogorov–Smirnov (Smirnov, 1948) and Kuiper's (Kuiper, 1960) tests. One of the other two tests is based on exceeding confidence interval-type bounds on the FDC, while the final test is based on the combined ecosurplus and ecodeficit, which has been termed ecochange.

To develop such tests, we could develop a hydrologic model for a basin, perturb a parameter or management scheme in the model, and assess the significance of the change in the FDCs due to the magnitude of the perturbation. While this would be controlled experiment, it would be reliant on how well the model represents reality, and how well the model perturbation represented a change in the hydrologic system. Instead we choose a method similar to Burn and Hag Elnur (2002) and Douglas et al. (2000), where measured streamflow sequences are employed along with a resampling strategy to assess the significance of the test. Unlike those studies which examined the significance of regional trends, here we develop a test to assess the significance of hydrologic alteration at a single streamflow site. Our tests rely upon having a period of record (here 20 years) during which it is assumed there is no alteration in the hydrologic series, and some periods of record (here assumed 5 years) after which a potential alteration has occurred to evaluate the significance of the alteration.

For each hypothesis test considered, 20 years of unaltered daily streamflow are used to develop the critical value of the test statistic corresponding to type I error probabilities of 5% or 10%. This is done by using either annual FDCs, or median or mean annual FDCs (see Vogel and Fennessey, 1994, for definitions of mean and median annual FDC's) obtained via an exhaustive jackknife resampling of the 20 year record in 5 year increments. Once each test is developed, the power of the test $(1 - \beta)$ is assessed, where β is the probability of a type II error. Power is assessed by an exhaustive jackknife resampling of 10 years of annual or jackknifed median or mean FDCs of potentially altered streamflows in 5 year increments. The significance of the alteration could be assessed with one 5 year post-alteration sequence. In practice, both type I and type II errors are of concern. Type I errors correspond to overprotecting the environment; type II errors, which are potentially worse than type I errors, correspond to not protecting the environment when we really should have.

For a case study, streamflow alteration in this experiment is due to the construction of a reservoir (Magilligan and Nislow, 2005). A subset of the reservoir sites employed by Poff et al. (2007) and Gao et al. (2009) that have a relatively long historic daily streamflow record both before and after construction of the reservoir are analyzed. It is assumed that the construction of a reservoir will produce a significant alteration of the streamflow record (which may not be true at all sites), and that no other forms of hydrologic alteration are impacting these records. The proposed hypothesis test framework provides a framework to assess hydrologic alteration, which could then inform water management decisions.

2. Development of test statistics

In this section, we describe four FDC-based hypothesis tests of hydrologic alteration. The resampling scheme to determine the critical values of the test statistic, significance and power of each test is held fixed across tests. For each test, the null hypothesis (H_o) is that there is no hydrologic alteration, and the alternative hypothesis (H_a) is that there is hydrologic alteration. Two of the tests are based on common hypothesis tests of distributional change (Kolmogorov–Smirnov and Kuiper), while the other two are based on observed shifts in the FDC. In this section (Section 2), the test statistic for each hypothesis test is described. In the following section (Section 3), the methodology to develop the critical values of each test statistic and the power of each test is discussed, and the reservoir sites employed in this analysis are presented.

2.1. Test 1: Kolmogorov–Smirnov test (KS)

The 2-sample KS test is a non-parametric hypothesis test where the null hypothesis is that two samples are drawn from the same distribution. The KS test compares the cumulative distribution function (cdf) of two data sets, and computes a test statistic based on the largest discrepancy between the distributions (Smirnov, 1948). There have been a number of applications of the KS test in hydrology, such as for distributional goodness of fit of floods (Chowdhury et al., 1991) and verifying water quality models (Reckhow et al., 1990).

Here, a version of the KS test is applied to the quantiles of the FDC for the two streamflow series: a baseline FDC (b) and a sampled FDC (s). While any baseline FDC could be employed, here we choose the median annual FDC of the unaltered streamflow series. The median annual FDC represents the distribution of daily streamflow in a "typical" year, and is not affected by abnormally wet or dry periods during the period of record (Vogel and Fennessey, 1994). The sampled FDC is the FDC constructed from some portion of the unaltered or potentially altered streamflow series. The test statistic, D_{KS}, is a function of the maximum deviation in the cdf, $F(x)$, between the two data sets:

$$D_{KS} = \max|F_b(x) - F_s(x)|$$

where $F_b(x)$ and $F_s(x)$ are the cdfs of the baseline and sampled FDCs, respectively, evaluated at x, as illustrated in Fig. 1, and the maximization is over all values of x. The site presented in Fig. 1 (USGS# 09288000) is representative of how a reservoir can impact the FDC at a site.

2.2. Test 2: Kuiper's Test (K)

The 2-sample K test is similar to the KS test, but instead of developing a test statistic as the maximum absolute difference between the distributions, it instead uses the maximum positive and negative differences (Kuiper, 1960). As illustrated in Fig. 2, this test statistic is defined by:

$$D_K = D^+ + D^- = \max(F_s(x) - F_b(x)) + \max(F_b(x) - F_s(x))$$

Since hydrologic alteration often is observed in the tails of the FDC and the difference in the FDCs can be both positive and negative, the K test may provide a more complete measure of hydrologic alteration than the KS test. For instance, a reservoir typically reduces maximum discharges while increasing minimum discharges. Such combined changes in the FDC should be better captured by this test.

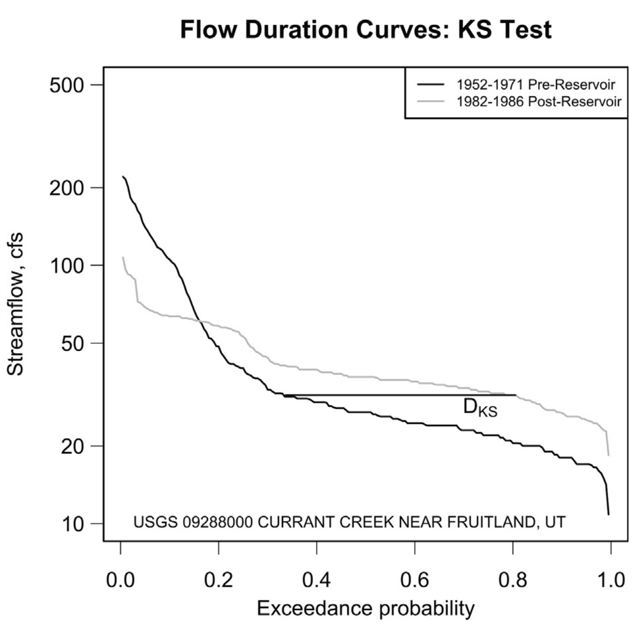

Fig. 1. Flow duration curves: KS test.

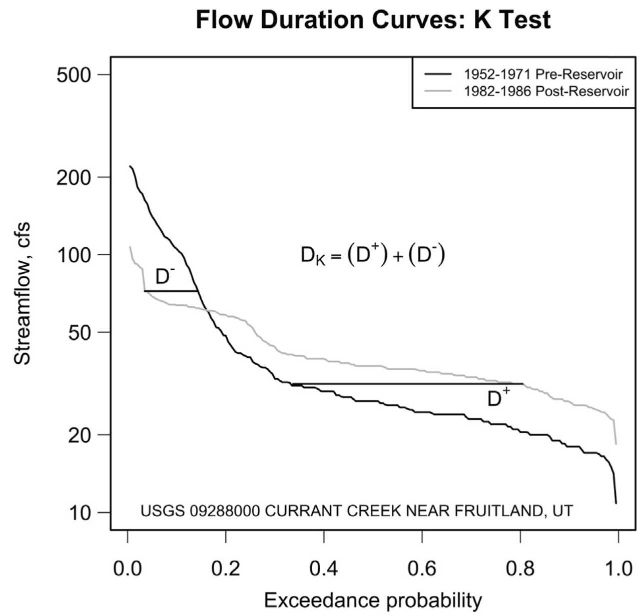

Fig. 2. Flow duration curves: K test.

2.3. Test 3: FDC confidence interval (CI)

The procedure for the CI test involves developing confidence intervals for the unaltered median annual flow duration curve, and determining the smallest confidence which would be necessary so that a specific percentage, $100 \times (1 - \text{type I error})\%$ = $100 \times (1 - \alpha)\%$, of the sampled FDCs from the unaltered streamflow series are completely contained within the confidence interval. As shown in Fig. 3, this is often a confidence interval that is larger than the $100 \times (1 - \alpha_{CI})\%$ confidence interval since the $100 \times (1 - \alpha_{CI})\%$ confidence interval is developed such that $100 \times \alpha_{CI}\%$ of the years *corresponding to each quantile* associated with the pre-alteration FDC fall outside of the confidence interval. While Vogel and Fennessey (1994) present a methodology to develop confidence intervals for the median FDC based on interpolating between quantiles of FDCs for individual years, developing confidence intervals can be challenging when one has limited data as one may be forced to extrapolate beyond the largest and smallest observed quantiles. This is true for the annual flow duration curve experiment (Section 3.1.1) where the number of quantiles is limited to the number of pre-disturbance years (here 20). In this case we assumed the annual FDC quantiles at each site are normally distributed; a probability plot correlation coefficient test of normality (Filliben, 1975) resulted in acceptance of this hypothesis approximately 80% of the time across all quantiles at our 117 sites. When using the jackknifed median or mean FDCs to develop confidence intervals (Sections 3.1.2 and 3.1.3), the assumption of normality was not required as a larger number of FDCs (and thus quantiles) were available to determine the critical value of the test statistic. In either case, the test statistic associated with each year is calculated as:

$$D_{CI} = -\log(\alpha_{CI})$$

where α_{CI} is from the $100 * (1 - \alpha_{CI})\%$ confidence interval that would completely contain $100 * (1 - \alpha)\%$ of the sampled FDCs (annual or jackknifed) (e.g. $D_{CI} = 2$ for a 99% confidence interval). The negative log transformation provides a more convenient scale for the test statistic that was similar to that of our other tests. The values of this test statistic can range from 0 to infinity (although the maximum value here was approximately 4), where a higher

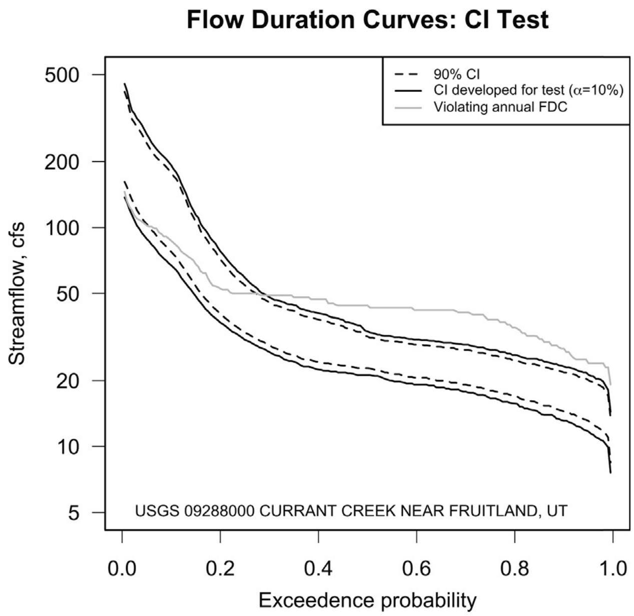

Fig. 3. Flow duration curves: CI test.

value indicates that the sampled FDC is more dissimilar to the unaltered median annual FDC.

2.4. Test 4: Ecochange = ecodeficit and ecosurplus (Eco)

This method is based on the sum of the ecosurplus and ecodeficit of the sampled FDC compared to the unaltered median annual FDC, which Gao et al. (2009) termed ecochange. As shown in Fig. 4, the ecosurplus is the percent area between the FDCs where the sampled FDC is below the unaltered median annual FDC, while the ecodeficit is the percent area where the sampled FDC is above the unaltered median annual FDC (Gao et al., 2009). For this method, the ecochange (Eco) is computed as the sum of ecosurplus and ecodeficit):

$$\text{Ecochange} = \text{Eco} = \text{Ecosurplus} + \text{Ecodeficit}$$

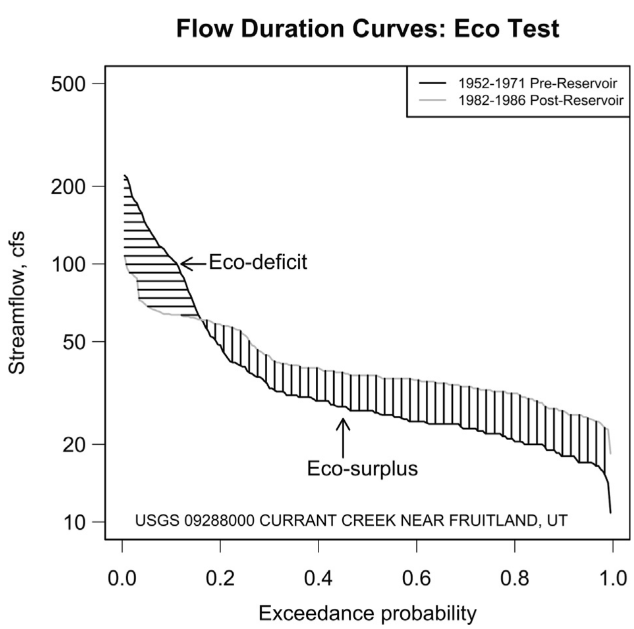

Fig. 4. Flow duration curves: Eco test.

is used as a test statistic in the hypothesis test of hydrologic alteration. Here ecodeficit and ecosurplus are standardized by the mean annual streamflow as recommended by Gao et al. (2009) so that the magnitudes are physically meaningful and comparable across sites.

3. Development of critical values of test statistics, power, and study sites

For each of the hypothesis test proposed in Section 2, the test statistic is estimated by comparing a sampled FDC to a baseline FDC. As mentioned previously, in this experiment the baseline FDC is the median annual FDC of the unaltered hydrologic series. Here estimator 2 from Vogel and Fennessey (1994, eqn. 3) is employed to estimate the median annual FDC. This method develops FDCs for each year, estimates the quantiles from the annual FDCs at specific exceedance probabilities, and determines the median daily flow across all annual quantiles at each exceedance probability.

In this experiment the critical value of each test statistic is empirically estimated using the unaltered streamflow series so that the type I error (α) is reproduced by each test. This is done by repeatedly comparing FDCs derived from a sequence of the unaltered flow series (the sampled FDC) to the median annual FDC (the baseline FDC). Once the critical values of each test statistic are developed, FDCs derived from a sequence of the potentially altered flow series are repeatedly compared to median annual FDC (the same baseline FDC) to assess the power of each hypothesis test, i.e. the probability of accepting the null hypothesis (i.e. no change in flow series) given that the null hypothesis is incorrect (i.e. there is a hydrologic alteration of the flow series).

3.1. Development of critical values of each hypothesis test statistic

To estimate the rejection region corresponding to the critical values of each test statistic, three different sampling strategies are employed: one based on annual FDCs, one based the median annual FDC from an exhaustive jackknife resampling, and one based on the mean annual FDC from an exhaustive jackknife resampling. Below, each of these methods is discussed.

3.1.1. Annual FDCs

In this experiment, to develop the rejection region for each hypothesis test for each method, annual FDCs from the unaltered flow series (here of length $N = 20$ years) are employed. For the KS test, the quantiles for the unaltered median annual FDC were first estimated for every exceedance probability from 0.5% to 99.5% using a step size of 0.5%. Each annual FDC of the unaltered flow series is then compared to the median annual FDC at each quantile, and the D_{KS} (the maximum deviation between the cdfs of the FDCs) is estimated. This results in N values of D_{KS}. The N values of D_{KS} are then ranked and, using a Weibull plotting position, the critical value of the hypothesis test is obtained by interpolating the D_{KS} values with an exceedance probability equal to the type I error α. A Weibull plotting position is suitable because it is known to yield unbiased estimates of exceedance probabilities, regardless of the distribution of the observations, an important property for the repeated implementation of the KS test.

For the K test, a similar procedure is employed as for the KS test, where instead of estimating the D_{KS} as the maximum difference in the cdf, D_K is estimated as the sum of the maximum positive and negative differences between the cdfs. Similar to the KS Test, the critical value of the K Test is obtained by interpolating the D_K value with an exceedance probability equal to α.

The CI test is based on developing confidence intervals (CIs) for the median annual FDC. Typically CIs for the median FDC are developed from annual FDCs by interpolating between the quantiles at a specific exceedance probability (Vogel and Fennessey, 1994). Difficulty arises when one has a short record so that there are only a few annual FDC's to interpolate between. To avoid this, here it is assumed that the quantiles at a specific exceedance probability are normally distributed, and then the CIs are obtained from a fitted normal distribution at each quantile. Recall earlier that we found the normal distribution to be a good approximation for annual FDC quantiles at a specific exceedance probability. The critical value is obtained by interpolating the D_{CI} value for which $100 - \alpha\%$ of the values exceeds it. This is analogous to the CI where $100 - \alpha\%$ of the annual FDCs have at least one quantile that falls outside of it.

For the Eco test, each annual FDC is compared to the median annual FDC, and the ecochange (Eco) is computed as Eco = ecodeficit + ecosurplus. The *N* values of Eco are then ranked, and using a Weibull plotting position, the critical value of the hypothesis test is obtained by interpolating the Eco value with an exceedance probability equal to the value of α under consideration.

3.1.2. Jackknifed median FDCs

One problem with developing the rejection region for the hypothesis tests based on annual FDCs (Section 3.1.1) is that if an exceptionally wet or dry year occurs, one might always conclude that hydrologic alteration has occurred. To address this concern, an exhaustive jackknife resampling is employed. Here FDCs from 5 years are selected (without replacement), and the median FDC from the 5 years is determined. 5 years was chosen as a convenient period after which one might assess whether a significant hydrologic alteration has occurred within a watershed; the period of resampling of the unaltered flow series should be the same as the period of record after the potential alteration. In a manner similar to the annual FDCs in Section 3.1.1, the median FDC from the 5 years is then compared to the unaltered median annual FDC (our baseline case). This procedure is reproduced $\binom{20}{5} = 15{,}504$ times for every possible combination of 5 years from the 20 year unaltered record. For the KS, *K*, CI, and Eco tests, this results in 15,504 values of D_{KS}, D_K, D_{CI}, and Eco. These values are then ranked, and using a Weibull plotting position the critical value for the hypothesis test is obtained by interpolating the value with an exceedance probability equal to our type I error α.

3.1.3. Jackknifed mean FDCs

When using jackknifed median FDCs (Section 3.1.2) with 20 years of record, a problem arises that at any given quantile there are only 16 unique values that the median FDC quantile could be (the highest 2 and lowest 2 values at a quantile could never be the 5-year median). Because of this, the type I error is not reproduced for the hypothesis test. To address this issue, the same procedure as used in Section 3.1.2 was used, except that the randomly selected 5 years were averaged at each quantile rather than taking the median. This produces 15,504 unique FDCs to test against the median FDC at each site. All test statistics and critical values were calculated in the same way as for the jackknifed medians (Section 3.1.2).

3.2. Estimation of the power of the hypothesis tests

After the critical values were developed from the unaltered streamflow series for each method and at each site, the power of the hypothesis tests were determine by resampling 10 years of potentially altered (post-alteration) streamflow data. All test statistics are developed so that each test is an upper-tailed test; if a test statistic is greater than the critical value, the null hypothesis is rejected. Below the estimation of the power for the annual FDCs and median and mean jackknifed FDCs are discussed.

3.2.1. Annual FDCs

For the annual FDC experiment, test statistics for the KS, *K*, and Eco methods were determined by comparing the 10 annual post-alteration FDCs to the unaltered median annual FDC. For the CI method, the 10 annual post-alteration FDCs were compared to the CIs developed from the unaltered median annual FDCs. The power of the test $(1 - \beta)$ was determined by counting the number of post-alteration annual FDCs that were correctly detected as altered, where β is the probability of a type II error.

3.2.2. Jackknifed median FDCs

For the jackknifed median FDC experiment, 5 of the 10 annual FDCs were selected without replacement and the median was taken at each quantile. The 5 years were taken to parallel the development of the critical value of the hypothesis test (Section 3.1.2). The interpretation of this hypothesis test is whether the potentially altered 5 year record is typical of a randomly chosen 5 year period of record from the unaltered streamflow series. This results in $\binom{10}{5} = 252$ FDCs to test against the critical value at each site. The test statistics for the KS, *K*, and Eco methods were determined by comparing each of the 252 post-alteration jackknifed median FDCs to the unaltered median annual FDC at each site. For the CI method, each of the 252 post-alteration jackknifed median FDCs are compared to the CIs developed from the unaltered flow series to determine the smallest CI necessary to contain each post-alteration FDC, and the test statistic was then calculated as described in Section 2.3. The power was determined by counting the number of post-alteration test statistics that exceeded the critical value estimated by the method described in Section 3.1.2.

3.2.3. Jackknifed mean FDCs

For the jackknifed mean FDC experiment, again 5 of the 10 annual post-alteration FDCs (252 unique sets) were selected and the mean was taken at each quantile. Calculation of the test statistics and power are the same as with the jackknifed median FDC experiment, except with jackknifed mean FDCs instead of jackknifed median FDCs.

3.3. Test sites

In this experiment, 117 USGS streamflow sites that have been subject to potential hydrologic alteration due to the installation of an upstream dam are analyzed. These sites are a subset of the 189 sites employed by Gao et al. (2009) and Poff et al. (2007). These sites had no pre-existing upstream mainstem dam prior to the installation of the dam, no more than two tributary inputs between the upstream dam and the gauge, and no dams on tributaries with an estimated drainage area larger than the mainstem river of the candidate dam (Gao et al., 2009). For this analysis, each site needed to have 20 years of pre-alteration (pre-dam, i.e. unaltered) record and 10 years of post-alteration (post-dam, i.e. potentially altered) record. FDCs are construction from mean daily streamflow records. Five-year periods before and after the dam installation were excluded from the analysis to remove a period of record when the dam was being installed and the reservoir filled, thus requiring a total continuous record of 20 + 5 + 5 + 10 = 40 years. The particular 117 sites selected matched all of these criteria, were spatially distributed throughout the US, and spanned a range of storage ratios (storage/mean annual flow) from 0.004 to 2500 days so as to capture a broad range of levels of alteration which can be expected in practice. Table 1 contains the USGS gauging station number, State, year of dam completion, and the contributing drainage area for each of the sites used in this study.

All rivers considered here had dams installed, though no information was available regarding the degree of alteration corresponding to each dam site. The installation of some dams might not produce a significant hydrologic alteration (such as run-of-river dams), though one would expect many of the dam sites would, especially those with large storage ratios. Hypothesis tests that result in higher percentage of detection of significant alterations exhibit higher power and thus would generally be considered to be better tests, since all of the tests are expected to reproduce their type I errors.

4. Results and discussion

4.1. Reproduction of type I error α

The only test that did not reproduce the type I error α, was the jackknifed median FDCs experiment, which produced higher values of α with the KS and CI methods in particular, as shown in Fig. 5. This is due the limited number of unique values of the test statistics to determine the critical value of the hypothesis test. This leads to situations where the critical value is nearly identical when α is set to 5% and 10%. This is most pronounced in the KS and CI methods because the test statistics only rely on the part of the FDC that is most dissimilar, which would be more likely to be repeated as the jackknifed median FDC value. The *K* method is less affected because there are two parts of the FDC that are considered, and the Eco method is much less affected because the entire FDC influences the test statistic value. Because the annual FDC and jackknifed mean FDC experiments had all unique values at each quantile, this problem did not occur for these experiments and the type I error α was exactly reproduced.

4.2. Power of hypothesis tests for hydrologic alteration

Fig. 6 presents box-plots of the power of each of the hypothesis tests across all 117 sites for the annual FDC experiment. Fig. 6 indicates a general lack of power for all of the hypothesis tests based on the annual FDC, though the CI test performs better than the other methods. The especially low power associated with hypothesis tests based on annual FDC's is due to the variability of annual FDCs, a feature that is smoothed out with a median or mean FDC. Because of this, it can be difficult for the hypothesis tests to distinguish between differences due to annual variability and differences due to alteration.

As seen in Fig. 7, the power of the tests is considerably improved (increased) when comparing the 5-year jackknifed median FDCs to the unaltered median annual FDC. Although there is still some variability in the jackknifed median FDCs, the annual variability is smoothed out, so that one extreme event will not have as large of an effect on the FDC being compared. With this experiment, as with the annual FDC experiment, CI and *K* are the

Table 1
Description of streamflow sites.

USGS number	State	Reservoir date	Drainage area (km^2)	USGS number	State	Reservoir date	Drainage area (km^2)	USGS number	State	Reservoir date	Drainage area (km^2)
01162000	MA	1977	213	03366500	IN	1975	755	09218500	WY	1971	357
01181000	MA	1964	243	03374000	IN	1963	28,809	05592000	IL	1970	2742
01315500	NY	1948	2059	03380500	IL	1965	1204	01459500	PA	1973	254
01413500	NY	1973	424	03212500	KY	1980	5553	09119000	CO	1978	2744
09239500	CO	1973	1460	05369000	WI	1957	4627	09124500	CO	1966	879
01467000	NJ	1947	324	03438000	KY	1966	635	09288000	UT	1976	364
01534000	PA	1974	1017	03406500	KY	1969	1564	09096500	CO	1959	208
01545500	PA	1962	7710	03325000	IN	1966	4697	06891500	KS	1977	1104
01568000	PA	1970	534	03575000	AL	1966	888	08102500	TX	1954	9278
01632000	VA	1980	543	04148500	MI	1973	2525	07339000	OK	1968	2072
01648000	DC	1965	137	04221500	NY	1951	800	06919000	MO	1969	2977
02021500	VA	1960	852	09132500	CO	1962	1363	09342500	CO	1974	727
02036500	VA	1974	58	03216500	KY	1968	1037	09426000	AZ	1968	11,992
02073000	VA	1955	983	05464000	IA	1979	13,349	08383500	NM	1979	10,463
02192000	GA	1978	3673	05506500	MO	1977	860	10128500	UT	1932	420
02198500	GA	1968	25,531	09241000	CO	1966	561	10130500	UT	1957	1108
02223500	GA	1965	11,408	06016000	MT	1964	7071	10137500	UT	1965	356
02233500	FL	1965	678	06207500	MT	1964	2985	10237000	UT	1983	780
02329000	FL	1958	2967	06345500	ND	1970	3238	10261500	CA	1971	1356
02368000	FL	1964	1643	06352000	ND	1976	1442	11043000	CA	1973	572
02450000	AL	1956	930	06410500	SD	1981	763	11231500	CA	1954	236
02472500	MS	1969	790	06609500	IA	1964	2252	11274000	CA	1966	29,065
09110000	CO	1962	1237	06625000	WY	1967	678	11317000	CA	1939	178
03016000	PA	1970	9480	06710500	CO	1962	426	11418000	CA	1969	2874
03024000	PA	1972	2763	06714000	CO	1975	10,012	11454000	CA	1959	1491
03103500	PA	1965	1523	06760000	CO	1953	43,451	12148500	WA	1962	210
03159500	OH	1972	2447	06877600	KS	1984	49,592	12394000	ID	1978	1570
09430500	NM	1963	4805	03269500	OH	1973	1266	12479000	WA	1933	524
03214000	WV	1960	3077	03208000	KY	1969	1018	13077000	ID	1978	40,152
07340000	AR	1977	6943	06922000	MO	1955	4309	13148500	ID	1962	802
09326500	UT	1978	359	07111000	CO	1959	196	14034500	OR	1982	252
03277500	KY	1975	1206	07123000	CO	1975	31,209	14046500	OR	1971	13,313
03284000	KY	1935	10,245	07176000	OK	1963	16,709	14145500	OR	1962	1017
03425000	TN	1973	27,685	03533000	TN	1936	7544	14151000	OR	1965	481
03329000	IN	1969	9912	03014500	NY	1979	499	14191000	OR	1961	18,822
03340500	IN	1969	28,917	08246500	CO	1951	730	14203500	OR	1970	325
03347000	IN	1965	627	08146000	TX	1960	7892	14209000	OR	1956	321
03360500	IN	1962	12,139	08153500	TX	1972	2334	14309000	OR	1985	202
03362000	IN	1972	260	04261000	NY	1956	377	14362000	OR	1980	580

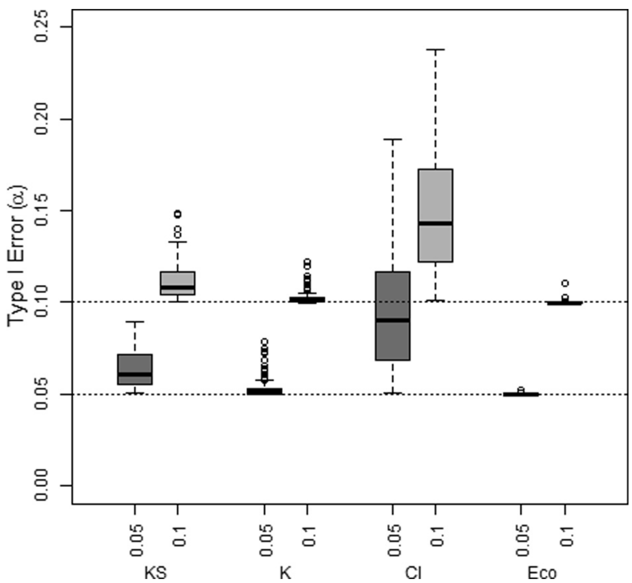

Fig. 5. Type I error (α) from jackknifed median FDC experiment.

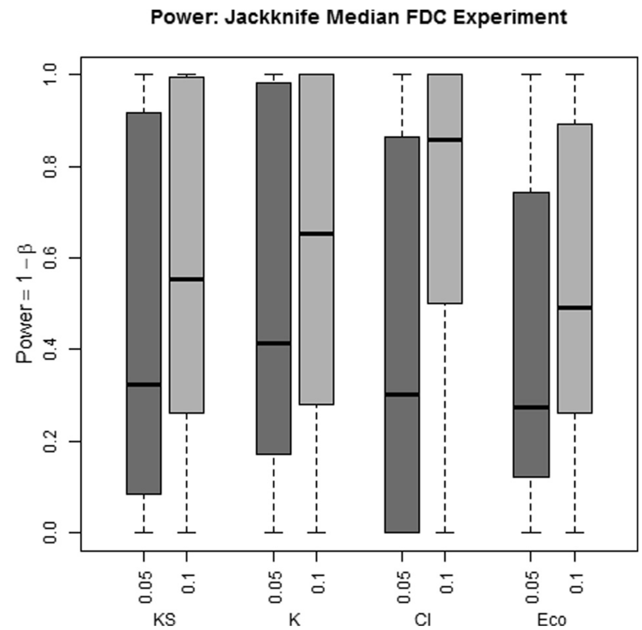

Fig. 7. Power of jackknifed median FDC experiment.

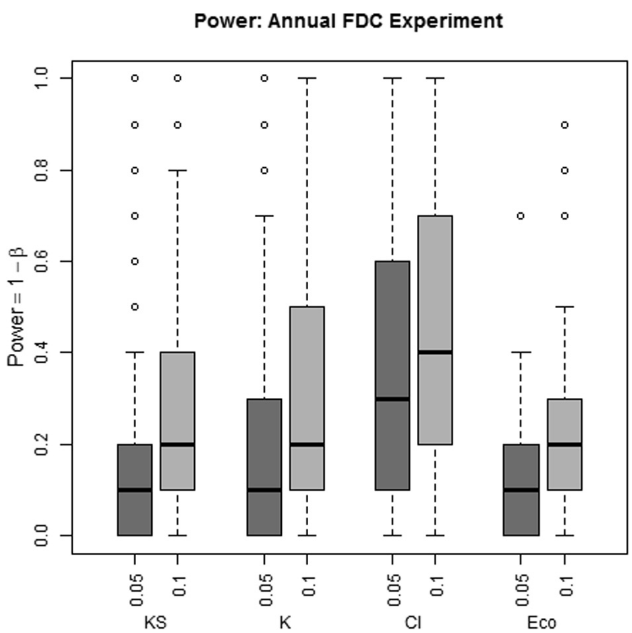

Fig. 6. Power of annual FDC experiment.

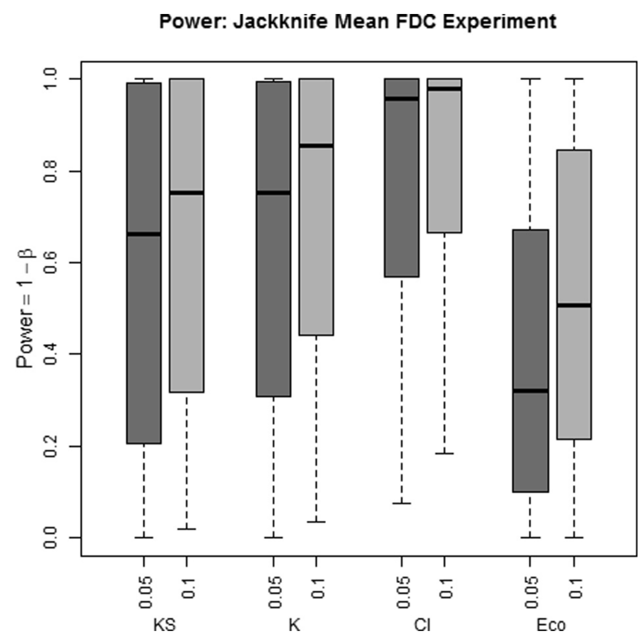

Fig. 8. Power of jackknifed mean FDC experiment.

best methods, although CI saw the most issues in reproducing the type I error due to having identical critical values with different values of α. By looking at Fig. 5 in conjunction with Fig. 7, it can be seen that there is a noticeable trade-off between type I error and type II error which is expected for any hypothesis test (see Fig. 3 in Vogel et al., 2013). The Eco method had the lowest median power for both values of α.

As shown in Fig. 8, the jackknifed mean FDC experiment showed that the CI test performed the best, followed once again by the *K* method. The Eco method had the lowest power. Overall, the CI method from the jackknifed mean experiment had the highest mean power across sites (0.96 for $\alpha = 5\%$ and 0.98 for $\alpha = 10\%$), likely because it used 5-year mean FDCs, which smoothed out the extreme events that cause the majority of variability in annual FDCs. In addition, using jackknifed mean FDCs also performed more consistently than the other experiments since it had many unique values, as opposed to the annual FDCs experiment that had few unique values, and the jackknifed median FDCs experiment, which had many non-unique (duplicated) values. These results indicate that the CI method performs best, followed by the *K* method, and that the jackknifed mean FDC experiment provides the best technique among those considered to distinguish hydrologic alteration. Importantly, among all of our comparisons, the results in Fig. 8 indicate that the CI method implemented using jackknifed mean FDC's is a promising approach for providing the ability to discriminate between alternative hydrologic flow regimes because it was the only method with which nearly always led to power above 0.5, and usually much larger than 0.5.

4.3. Power and magnitude of alteration

An effort was made to assess whether the performance of the hypothesis tests was related to the level of hydrologic disturbance. Since the alteration under consideration here is the implementation

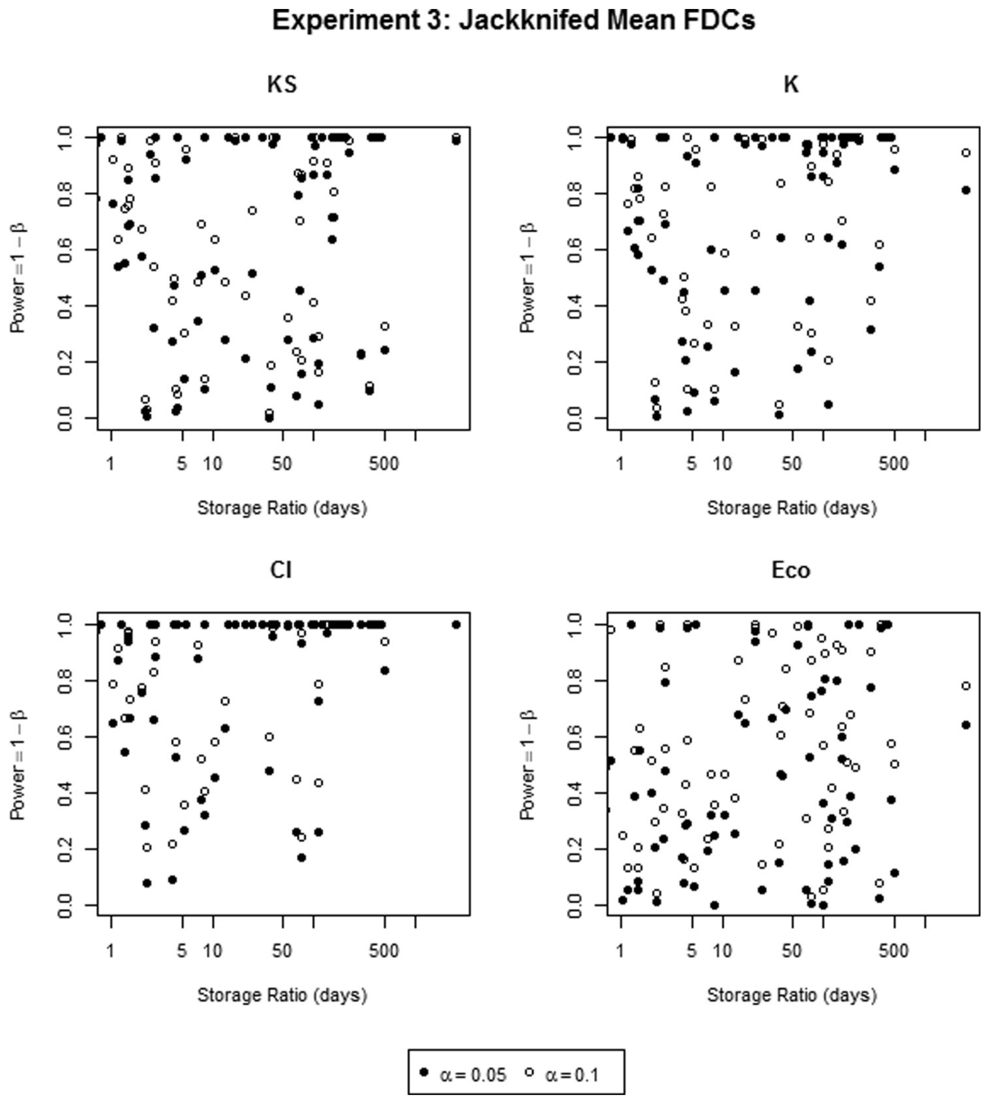

Fig. 9. Storage ratio vs. power for jackknifed mean FDCs experiment.

of a dam and reservoir upstream of the stream gauge, the storage ratio was used as a metric of the magnitude of alteration. The storage ratio, in days of available storage, is defined as:

$$\text{Storage Ratio(days)} = \frac{\text{Storage of Reservoir(ac:ft)}}{\text{Mean Annual Flow(ac:ft=yr)}} * \left(365\,\frac{\text{days}}{\text{yr}}\right)$$

where storage of reservoir is the volume of storage in the reservoir just upstream of the gauge and mean annual flow is during the 20 year pre-reservoir record. The storage ratio is a measure of the average residence time of water in the reservoir, which one might expect to be related to the level of hydrology alteration.

Fig. 9 presents the power of the hypothesis test versus storage ratio for the jackknifed mean experiment for the four methods. While for KS, *K*, and Eco tests there does not appear to be any relationship between the power of the test and the storage ratio, for the CI test, which had the highest overall power among the tests considered, at higher storage ratios a higher percentage of sites generally exhibited high power. A similar analysis of power versus the hydrologic disturbance index, a broad measure of hydrologic alteration which is part of the USGS's GAGES database (Falcone et al., 2010), showed no relationships for any of the tests. No effort was made in this experiment to understand the type of reservoir regulation at the study sites, and reservoirs which are ''run-of-river'' may have a minimal change in the FDC before or after reservoir construction, especially when the reservoir is near full storage. An analysis of pre- and post-alteration FDCs confirmed this observation for some of the sites considered. Further assessment of the impact of reservoir regulation, as well as other forms of hydrologic alteration, on the performance of the proposed hypothesis tests is warranted.

5. Conclusions

This experiment develops and assesses four new flow duration curve (FDC) based hypothesis tests of hydrologic alteration. It is assumed that an unaltered period of record exists (here assumed to be 20 years) prior to a potential hydrologic alteration, and a shorter streamflow record after a potential hydrologic alteration. Resampling the unaltered record, the critical value of each hypothesis test is developed to ensure that the probability of a type I error is reproduced. With the post-alteration record the significance of the alteration can be determined, and by resampling the post-alteration record the power of each hypothesis test can be estimated.

All hypothesis tests and sampling designs were examined using 117 gauged streamflow sites spatially distributed across the US where a reservoir has been constructed. Reservoir systems tend to alter streamflow series, producing a ''flatter'' flow duration curve with smaller maximum streamflows and larger minimum streamflows. The hypothesis tests are general, and could be applied to any hydrologic alteration as long as a pre- and post-alteration streamflow record was available. It would also be possible to apply these tests for other river hydraulic properties including river stage and/or velocity, as well as to other environmental variables with a temporal signature (such as temperature and dissolved oxygen concentrations). See Vogel and Fennessey (1995) for a review of the myriad of applications of FDC's for which these hypothesis tests may be relevant.

Three different experimental designs were explored, each with a different resampling of the record to determine the distribution of the test statistic. An initial experiment compared individual annual FDCs to the unaltered median annual FDC. In a second experiment, the median of 5 random years was taken in an exhaustive jackknife resampling to yield 15,504 FDCs that were compared to the unaltered median annual FDC. The 5 years was chosen as a period after which one might assess whether a potential alteration has occurred to a watershed; the resampling of 5 random years allows the experimenter to assess whether the potentially altered 5 year period is significantly different than a random 5 year period from the unaltered record. The final experiment increased the number of unique quantile values in the resampled FDCs, where the mean FDC of 5 random years in an exhaustive jackknife resampling was compared to the unaltered median annual FDC. The result from the experimental design indicated that:

- While the Annual FDC experiment provides an easy method to understand and interpret, there are limited unique values of a test statistic.
- The Jackknife Median FDC experiment provides a large number of simulations, but produces limited unique values of the test statistic.
- The Jackknife Mean FDC experiment provides both a large number of simulations and a large number of unique values of the test statistic. Here, the Jackknife Mean FDC experiment is the best technique explored.

Four different FDC-based hypothesis tests were performed for each of the experimental designs. Two were based on established tests of distributional change: the Kolmogorov–Smirnov (KS) and the Kuiper (K) tests. Another test (CI) was based on adjusting the confidence intervals of the unaltered median annual FDC to completely contain $100 \times (1-\alpha_{CI})$% of the annual or jackknifed FDCs and comparing each sampled post-alteration annual or jackknifed FDC to see if it deviated from this confidence interval. The final test (Eco) was based on the combined ecodeficit and ecosurplus between the unaltered median annual FDC and the post-alteration annual or jackknifed FDC. The results from the hypothesis tests indicated that:

- The CI test consistently outperforms the other methods, producing the highest power for each experiment.
- The K test was the second best method.
- The KS and Eco tests always performed worst.
- While the CI test was best, this technique is complicated to understand, which may limit its use in practice.
- Due to its ease of use, the K test should also be further explored.
- A slight relationship between the power of the CI test and the storage ratio, a measure of the residence time within a reservoir, was observed. No relationship was found for the other tests.

This study is a preliminary investigation using FDC-based hypothesis tests to detect hydrologic alteration. Some ways in which this research can be expanded upon includes:

- Examine the performance of methods when sites are partitioned based on magnitude of alteration.
- Assess how reservoir systems are managed, and how this management is related to the power of the hypothesis tests.
- Explore the sensitivity of methods to other types of hydrologic alteration, such as water withdrawals, land use change, and climatic variation.
- Examine the impact of experimental design assumptions such as the number of pre- and post-alteration years of record, the number of resampled years, and comparisons to other FDC statistics, such as the mean annual FDC.

Acknowledgements

The authors would like to acknowledge the Environmental Resources Engineering Department at SUNY ESF for support of this project. The authors would also like to acknowledge the two anonymous reviewers who provided insightful comments and suggestions that greatly improved this manuscript. All data employed in this analysis is publicly available from either the USGS National Water Information System (http://waterdata.usgs.gov/nwis), the National Inventory of Dams (https://catalog.data.gov/dataset/national-inventory-of-dams), or the USGS Gages II database (http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml).

References

Arthington, A.H., Bunn, S.E., Poff, N.L., Naiman, R.J., 2006. The challenge of providing environmental flow rules to sustain river systems. Ecol. Appl. 16, 1311–1318.
Botter, G., Zanardo, S., Porporato, A., Rodriguez-Iturbe, I., Rinaldo, A., 2008. Ecohydrological model of flow duration curves and annual minima. Water Resour. Res. 44, W08418. http://dx.doi.org/10.1029/2008WR006814.
Buishand, T.A., 1984. Tests for detecting a shift in the mean of hydrological time series. J. Hydrol. 73, 51–69.
Bunn, S.E., Arthington, A.H., 2002. Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. Environ. Manage. 30, 492–507.
Burn, D.H., Hag Elnur, M.A., 2002. Detection of hydrologic trends and variability. J. Hydrol. 255, 107–122.
Castellarin, A., Botter, G., Hughes, D.A., Liu, S., Ouarda, T.B.M.J., Parajka, J., Post, D.A., Sivapalan, M., Spence, C., Viglione, A., Vogel, R.M., 2013. Prediction of Flow Duration Curves in Ungauged Basins. Chapter 7 in Prediction in Ungauged Basins. Cambridge University Press, 496 pages.
Chowdhury, J.U., Stedinger, J.R., Lu, L.H., 1991. Goodness-of-fit tests for regional generalized extreme value flood distributions. Water Resour. Res. 27, 1765–1776.
Douglas, E.M., Vogel, R.M., Kroll, C.N., 2000. Trends in flood and low flows in the United States: impact of spatial correlation. J. Hydrol. 240, 90–105.
Falcone, J.A., Carlisle, D.M., Wolock, D.M., Meador, M.R., James, A., 2010. GAGES: a stream gage database for evaluating natural and altered flow conditions in the conterminous United States. Ecology 91, 621–622.
Filliben, J.J., 1975. The probability plot correlation coefficient test for normality. Technometrics 17, 111–117.
Fitzhugh, T.W., Vogel, R.M., 2011. The impact of dams on flood flows in the United States. River Res. Appl. 27, 1192–1215.
Fitzhugh, T.W., 2014. EFCAM: a method for assessing alteration of environmental flow components. River Res. Appl. 30, 825–844.
Foster, H.A., 1924. Theoretical frequency curves and their application to engineering problems. Am. Soc. Civ. Eng. Trans. 87, 142–203.
Gao, Y., Vogel, R.M., Kroll, C.N., Poff, N.L., Olden, J.D., 2009. Development of representative indicators of hydrologic alteration. J. Hydrol. 374, 136–147.
Homa, E.S., Brown, C., McGarigal, K., Compton, B.W., Jackson, S.D., 2013. Estimating hydrologic alteration from basin characteristics in Massachusetts. J. Hydrol. 503, 196–208.
Kuiper, N.H., 1960. Tests concerning random points on a circle. Proc. Koninklijke Nederlandse Akademie van Wetenschappen, Ser. A 63, 38–47.
Magilligan, F.J., Nislow, K.H., 2005. Changes in hydrologic regime by dams. Geomorphology 71, 61–78.
McManamay, R.A., 2014. Quantifying and generalizing hydrologic responses to dam regulation using a statistical modeling approach. J. Hydrol. 519, 1278–1296.
Mejıa, A., Daly, E., Rossel, F., Jovanovic, T., Gironas, J., 2014. A stochastic model of streamflow for urbanized basins. Water Resour. Res. 50, 1984–2001. http://dx.doi.org/10.1002/2013WR014834.
Olden, J.D., Poff, N.L., 2003. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Res. Appl. 19, 101–121.
Poff, N.L., Olden, J.D., Merritt, D.M., Pepin, D.M., 2007. Homogenization of regional river dynamics by dams and global biodiversity implications. Proc. Natl. Acad. Sci. U.S.A. 104, 5732–5737.

Poff, N.L., Richter, B.D., Arthington, A.H., Bunn, S.E., Naiman, R.J., Kendy, E., Acreman, M., Apse, C., Bledsoe, B.P., Freeman, M.C., 2010. The ecological limits of hydrologic alteration (ELOHA): a new framework for developing regional environmental flow standards. Freshw. Biol. 55, 147–170.
Poff, N.L., Zimmerman, J.K.H., 2010. Ecological responses to altered flow regimes: a literature review to inform the science and management of environmental flows. Freshw. Biol. 55, 194–205.
Reckhow, K., Clements, J., Dodd, R., 1990. Statistical evaluation of mechanistic water-quality models. J. Environ. Eng. 116, 250–268.
Richter, B.D., Baumgartner, J.V., Powell, J., Braun, D.P., 1996. A method for assessing hydrologic alteration within ecosystems. Conserv. Biol. 10, 1163–1174.
Salas, J.D., Boes, D.C., 1980. Shifting level modeling of hydrologic series. Adv. Water Res. 3, 59–63.
Searcy, J.K., 1959. Flow-duration curves. In: Manual of Hydrology: Part 2. Low-Flow Techniques. USGS Water-Supply Paper 1542-A.
Smirnov, N., 1948. Table for estimating goodness of fit of empirical distributions. Ann. Math. Stat. 2, 279–281.
Vogel, R.M., Fennessey, N., 1995. Flow duration curves II: a review of applications in water resources planning. J. Am. Water Resour. Assoc. 31, 1029–1039.
Vogel, R.M., Fennessey, N.M., 1994. Flow duration curves I: new interpretation and confidence intervals. J. Water Resour. Plan. Manage. 120, 485–504.
Vogel, R.M., Sieber, J., Archfield, S.A., Smith, M.P., Apse, C.D., Huber-Lee, A., 2007. Relations among storage, yield and instream flow. Water Resour. Res. 43, W05403. http://dx.doi.org/10.1029/2006WR005226.
Vogel, R.M., Rosner, A., Kirshen, P.H., 2013. Likelihood of societal preparedness for global change – trend detection. Nat. Hazard. Earth Syst. Sci. Brief Commun. 13, 1–6.