Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR038981

Rethinking Paired-Catchment Studies: Should We Be Replicating Our Controls?

Steve Wondzell[1], Sherri Johnson[1], Gordon Grant[1], Don Henshaw[1], and Adam Ward[2]

[1]USDA Forest Service, Pacific Northwest Research Station, Corvallis, OR, USA, [2]Department of Biological and Ecological Engineering, Oregon State University, Corvallis, OR, USA

Key Points:

- Paired-catchment studies are seldom replicated making it difficult to determine a minimum detectable effect size (MDES)
- We used bootstrapped sampling to make hundreds-of-thousands of reference-by reference comparisons (RxR) with a paired-catchment approach
- 95% confidence intervals from the frequency distribution of RxR comparisons estimated the MDES resulting from errors and random effects

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Wondzell,
steven.wondzell@usda.gov;
swondzell@gmail.com

Citation:

Wondzell, S., Johnson, S., Grant, G., Henshaw, D., & Ward, A. (2025). Rethinking paired-catchment studies: Should we be replicating our controls? *Water Resources Research*, *61*, e2024WR038981. https://doi.org/10.1029/2024WR038981

Received 25 SEP 2024
Accepted 5 SEP 2025

Author Contributions:

Conceptualization: Steve Wondzell, Gordon Grant, Adam Ward
Data curation: Sherri Johnson, Don Henshaw
Formal analysis: Steve Wondzell
Methodology: Steve Wondzell
Visualization: Steve Wondzell
Writing – original draft: Steve Wondzell

Abstract

Paired-catchment studies are widely used to examine the effects of land management practices ("treatments") on hydrologic processes. Catchments are matched and a pretreatment calibration regression is used to identify the hydrological relationship between the reference and treated catchments. This method assumes that the calibration regression represents the actual relationship between the catchments (assumption of representativeness) and that the relationship will remain stable over time (assumption of stability). Errors are assumed to be small and similar between reference and treated catchments. Thus, observed differences between the catchments following treatment are assumed to result from that treatment alone. However, calibration periods are often short and it is impossible to know if the calibration period is representative. Further, because the study is unreplicated, it is impossible to determine if stability is maintained. Consequently, it is difficult to determine a minimum detectable effect sizes (MDES) below which estimates of changes in streamflow are statistically uncertain. Here, we use bootstrapped sampling from reference-by-reference (RxR) comparisons in a paired-catchment study design to evaluate the MDES. We generate frequency distributions of the potential changes in flow—changes that cannot be caused by treatment effects. From these, we estimate bootstrapped ±95% confidence intervals encompassing the non-treatment effects which we use as the MDES. We apply this method to long-term paired-catchment studies and reexamine changes in both annual water yields and late summer low flows at the HJA Experimental Forest. This bootstrapping method is widely transferable to any long-term paired catchment study sites where multiple reference catchments exist.

Plain Language Summary

To determine the effect of land management activities on stream flows, hydrologists often study a pair of side-by-side watersheds, treating one (e.g., 100% clear-cut logging) and keeping the other untouched as a reference. These studies are usually unreplicated, with one treated and one reference watershed. Without replication, it is impossible to tell if errors or unexpected factors caused the changes in stream flow. In fact, the effect of errors will be included in the "treatment effect." To identify the effect of errors on study results, we made comparisons between pairs of reference watersheds from the H. J. Andrews Experimental Forest. Our analysis showed that there were long-term changes in measured stream flows from these reference watersheds. The size of the changes resulting from errors creates a threshold, and for a treatment effect from a logged watershed to be real, it has to be larger than this threshold. The influence of errors and other effects limited our ability to assess the effect of land-management activities on stream flows, especially in late summer when flows are very low. In the future, these studies need multiple reference watersheds if we really want to understand the effect of land management on stream flow.

1. Introduction

Paired-catchment studies have been an important tool to examine whole-catchment responses to changes in forest cover since the first study, using a reference catchment, was established at Wagon Wheel Gap, Colorado, USA in 1910 (Bates & Henry, 1928). Since then, these studies have provided insight into a variety of catchment processes (see reviews by Hibbert, 1967; Bosch & Hewlett, 1982; Brown et al., 2005). For example, these studies provided many of the earliest insights into the effects of forest harvest on stream flows—including total annual yields, summer low flows (Harr et al., 1982; Hoover, 1944; Rothacher, 1965, 1970), and peak flows (Harr & McCorison, 1979). The paired-catchment approach has also been used to examine the effects of forest harvest on stream temperatures (Gomi et al., 2006; Janisch et al., 2012), stream nutrient budgets (Campbell et al., 2016; Webster et al., 2022), and erosion and delivery of both suspended and bedload sediment to the mouth of the catchment (Udawatta et al., 2002; Wright, 2023). Paired-catchment studies have also been used to examine effects of riparian vegetation (Dunford & Fletcher, 1947; Rowe, 1963), road building (Rothacher, 1970), changing forest

Writing – review & editing: Sherri Johnson, Gordon Grant, Don Henshaw, Adam Ward

composition (Hornbeck et al., 1993, 1997), forest regrowth (Hicks et al., 1991; Perry & Jones, 2017; Segura et al., 2020), natural disturbance (Amatya et al., 2021), and different patterns and extents of forest harvest among others.

Paired-catchment studies often use a before-after, control-impact (BACI) design (Moore & MacDonald, 2024). First, a pair of catchments are selected with similar physical characteristics, including proximity, size, soils and geology, drainage pattern, aspect, slope, and vegetation. Researchers assume stream flows from pairs of physically similar catchments located close together will be highly correlated. However, this assumption cannot be tested prior to selection because gages are established after the catchments are selected. After gages are established, streamflow is measured in both catchments over a multi-year calibration period to develop a pre-treatment relationship between the two catchments. Then, one catchment is treated leaving the other as a reference. After treatment, measurements from both catchments continue over periods that may extend from years to decades. Data are analyzed by using observations from the reference catchment, along with the pre-treatment relationship, to predict the expected behavior of the treated catchment as if it had not been treated. The treatment effect is simply the difference between the observed and expected behavior of the treated catchment and the variation in differences over the calibration period is used to estimate uncertainty.

The most typical approach used for developing the pre-treatment or calibration relationship is a linear regression between the reference (predictor) and treated (response) catchments. However, the calibration data are a time series of repeated measures over time. As such, the data can be temporally correlated and the regression residuals are often heteroscedastic (Watson et al., 2001). These factors can create problems for statistical analyses and have thus received much attention in the hydrology literature (Bren & Lane, 2014). However, another issue—the absence of any meaningful replication—has typically been ignored in paired-catchment studies. This stands in stark contrast to other disciplines where the specific experimental design of BACI studies have been closely examined. For example, Underwood (1991) identified potential problems in BACI studies lacking replicated controls, because any change between the control and treatment pair will be identified as an impact. Underwood argued for multiple reference sites so that differences between reference sites over time could be used to assess uncertainty. Conversely, Stewart-Oaten and Bence (2001) argued that, since sites were not selected randomly and treatments not imposed randomly, control sites can only be used to "*reduce extraneous variation*" and cannot be used to assess uncertainty. Nevertheless, BACI designs necessarily assume that the pre-treatment relationship between treatment and reference sites would remain unchanged over time. We follow Gomi et al. (2006) and call this the assumption of "stability." However, it is not possible to ensure this assumption is met in studies with only one reference site and one treatment site.

The BACI design, as applied to paired-catchment studies, can be effective at reducing extraneous variation. Typically, catchments are physically matched, gaged using identical methods, and studied by the same group of observers. Thus, if the calibration data are reasonably well correlated, then whatever errors, sources of uncertainty, or "drift" from external factors such as climate change might be present, it is likely that they would affect both the reference and treated catchments in roughly similar ways (Bren & Lane, 2014). In that case, post-treatment observations from the reference catchment should account for any changes that would affect both catchments over time so that any post-treatment changes between the catchment pair can be confidently ascribed to the treatment itself (Webster et al., 2022). Note that BACI designs do not assume that the reference catchments be changeless, but rather, that, if the treated catchment had been left untreated, it would have changed in the same way as the reference catchment so that the original pre-treatment relationship would continue to provide an unbiased estimate of the relationship between the two catchments (Wicht, 1943).

It seems reasonable that stability could be maintained over short post-treatment periods of a few years in paired-catchment studies. And if the harvest treatment is sufficiently severe, then we would reasonably expect that immediate post-treatment responses ought to be much larger than any errors resulting from small losses in stability. Some paired-catchment studies, however, continue to collect data many decades after the initial treatments. For example, measurements from paired-catchment studies at the HJA that were established in the 1950s and 1960s to examine impacts of logging on post-harvest stream flows continue to this day. Because of the length of record, comparisons between reference and treated catchments have been used to examine questions not posed in the original study, for example, to study the effects of post-harvest forest regrowth on late summer deficit flows as much as 45–50 years after the initial treatments (Gronsdahl et al., 2019; Perry & Jones, 2017; Segura et al., 2020).

Long-term paired-catchment studies rely on the pre-treatment relationship developed from data collected at the beginning of the study. Those long study periods, however, provide ample opportunity for loss of stability. Again using the HJA as an example, changes over the duration of the 70-year long studies included: (a) changing technologies used for measuring discharge and for recording continuous stage height data; (b) the gages have been abraded by bedload sediment during floods and occasionally destroyed by debris flows during large floods; (c) gages have been vandalized, gage designs changed, and gages rebuilt; (d) changes in staff and the ways in which they collect and process field data; and (e) the catchments were similar but not identical at the beginning of each study and have continued to change through time in response to disturbance, plant succession, and stochastic events that are unique to each of the catchments. Collectively, these events may have changed the original relationship between the catchments, but, since one catchment was treated and has been moving along a unique trajectory of change it is no longer possible to test whether or not stability has been preserved.

How then might researchers reliably estimate the minimum detectable effect size (MDES) in paired-catchment studies? Several authors have used ±95% prediction intervals (±95% PIs) around the regression fit to the calibration data and extended the ±95% PI across the post-treatment observations (Harris, 1977; Keppeler & Ziemer, 1990; Moore & MacDonald, 2024; Som et al., 2012). If treatment effects are larger than the ±95% PIs, they judge that the treatment response is likely to be statistically significant. We call this approach the *regression-based ±95% PI* approach. However, this approach presupposes that the calibration regression is representative. That is, the regression through the calibration data accurately measures both the relationship between the reference and treated catchments and the variability in the time series data upon which it is based. We call this the assumption of representativeness.

Unfortunately, calibration periods are usually quite short. For example, calibration periods for paired-catchment studies at the HJA were either 6, 9, or 10 years in length, for studies that have lasted from 55 to over 70 years. What are the chances that a calibration period of 6 years, or even 10 years, drawn from a 70-year long time series, accurately captures both the long-term relationship and the variance around that relationship for any pair of catchments? This question is not easily answered. We simply have no way of knowing if the calibration regression accurately represents the actual relationship between reference and treated catchments over long time periods.

Comparisons among multiple reference catchments provide a potential alternative to assessing representativeness and stability. In this approach, pairs of reference catchments are compared in exactly the same way that reference-by-treated catchment pairs are analyzed in paired catchment studies. We call this a reference-by-reference (RxR) comparison. Clearly, there cannot be a treatment effect in RxR comparisons. Further, if data from the catchments were perfectly correlated, then the expected and observed values would always be exactly equal so that the differences (observed minus expected) would always equal zero. Because of inherant variability between catchments, combined with any measurement errors, there will always be some differences between expected and observed values. In traditional paired catchment studies this difference is considered to be the treatment effect. However, because reference catchments are not treated, any changes over time can only result from non-treatment effects. To avoid confusion, we call the difference between observed and expected values in RxR comparisons a RxR response.

A small number of studies have used RxR comparisons in paired-catchment studies. For example, Webster et al. (2022) used RxR comparisons when examining effects of forest harvest on stream water chemistry to ensure that treatment effects were not confounded with long-term trends in stream chemistry resulting from climate change and recovery from acidification. Webster et al. (2022), however, did not use comparisons among reference catchments to determine a MDES. Gomi et al. (2006) pioneered the use of RxR comparisons to determine MDES in a paired-catchment study of stream temperature responses to forest harvest. In their case, data were available for three reference catchments. For their RxR comparisons, they used the same time periods used for calibration in their reference-by-treated comparison. They then assumed that RxR responses would be normally distributed, allowing them to calculate the standard deviations (SD). And because 95% of observations fall between ±1.96 SDs for normally distributed data, the 95% prediction intervals could be calculated as ±1.96*SD. We call this approach the *reference-based 1.96*SD* approach.

While the RxR comparison provides an interesting approach to determine MDES in paired-catchment studies, it can be difficult to apply to hydrologic studies. Gomi et al.'s (2006) analyses used daily temperature data so calibration periods of only a few years provided a large number of observations. In contrast, when evaluating the effect of forest harvest on total annual yield, a full year's worth of stream gage data results in only a single observation. Collecting sufficient data to implement this approach would take many years. Further, the RxR comparison as implemented by Gomi et al. (2006), Janisch et al. (2012), and Leach et al. (2022) all assume that the calibration period is representative of the full study period. While this assumption might be reasonable for a short-term stream temperature study it may not be reasonable for a multi-decade study of stream flow, as noted above.

Considering the approach developed by Gomi et al. (2006), we note that the choice of the "calibration period" is entirely arbitrary. If relationships between the reference catchments are truly stable over time, one could pick any time period for calibration data and use it to estimate non-treatment variability from the data not used in the calibration. For example, Ssegane et al. (2017) used block-bootstrap sampling to estimate the frequency distributions of regression coefficients fit to their calibration data and also examine their stability. This approach suggests an alternative method for evaluating the MDES. Bootstrap sampling without replacement could be used to randomly select calibration data which could be used to generate an RxR regression, and that regression could be used to calculate a RxR-response from data not used in the RxR calibration. Further, because the number of possible combinations is very large, the RxR time series can be resampled thousands of times to generate a very close approximation of the true underlying frequency distribution of the RxR responses. And with potentially 10s-of-thousands of observations in the frequency distribution, bootstrapped confidence intervals will be robust, regardless of the shape of the underlying frequency distribution.

In this paper, we reexamine the results of paired-catchment studies of whole-catchment responses to forest harvest at the HJA Experimental Forest. First, we use a graphical analysis of mean August discharge to illustrate issues with both representativeness and stability that are readily apparent in these paired-catchment studies. We next develop the bootstrapped-sampling method for RxR catchment comparisons that demonstrates how bootstrapped sampling can capture a frequency distribution of non-treatment effects. We then use the bootstrapped ±95% confidence intervals (CIs) as an estimate of the MDES and compare that to the size of the treatment effects observed from reference-by-treated catchment pairs. We examined three separate paired-catchment studies that were implemented at the HJA in past decades, examining their results over two different time periods. We examined the August low-flow period because regional studies have reported low-flow deficits in late summer driven by enhanced evapo-transpiration from regrowing plantation forests (Perry & Jones, 2017; Segura et al., 2020). These low-flow deficits have been identified as a potential concern because they could exacerbate existing management issues during annual low flow periods. We also examined total annual yields because this is a traditional measure often reported from paired catchment studies.

2. Methods

2.1. Study Site Description

The 64 km^2 H. J. Andrews Experimental Forest was established in 1948. It is located in conifer dominated forest on the west slope of the Cascade Mountain, Oregon, USA (44.2324N, 122.1921W). The area has a Mediterranean climate, with wet cool winters and hot dry summers. Elevations within the HJA range from 410 to 1,630 m. Annual precipitation averages around 2,200 mm, with slightly more precipitation falling at higher elevations (Daly et al., 2019). Seventy percent of the precipitation falls between November and March, falling as a mix of rain and snow at the highest elevations but as rain at the lowest elevations.

At the time of establishment, the HJA was densely forested with primary (never-logged) coniferous forests and few roads. During the post-World War II period, the region was experiencing a major expansion of logging into -lands managed by the US Forest Service along the western slope of the Cascade Mountains. These new road and logging entries into primary forests raised many questions which the HJA was established to answer. Initial research focused on the effects of road building and forest harvest on stream flows and sediment production. Road building within the HJA began around 1950 and the entire logging road network had been built out by about 1980. Today, about 25% of the HJA area has been logged, and some 40% remains in primary old-growth forest established ~500 years ago (Figure 1).

Three sets of paired-catchment studies have been established at the HJA over time (Table 1; Supporting Information S1). The first set was established in 1952, encompassing WS1, WS2, and WS3 (Rothacher, 1970). These catchments are located in the rain-dominated elevations where snowpacks seldom accumulate in winter and when snowfall does occur, it typically melts within a few weeks. Forests in all three catchments were dominated by ~450-year old Douglas-fir (*Pseudotsuga menziesii*). WS1 was 100% clearcut but without roads being built into

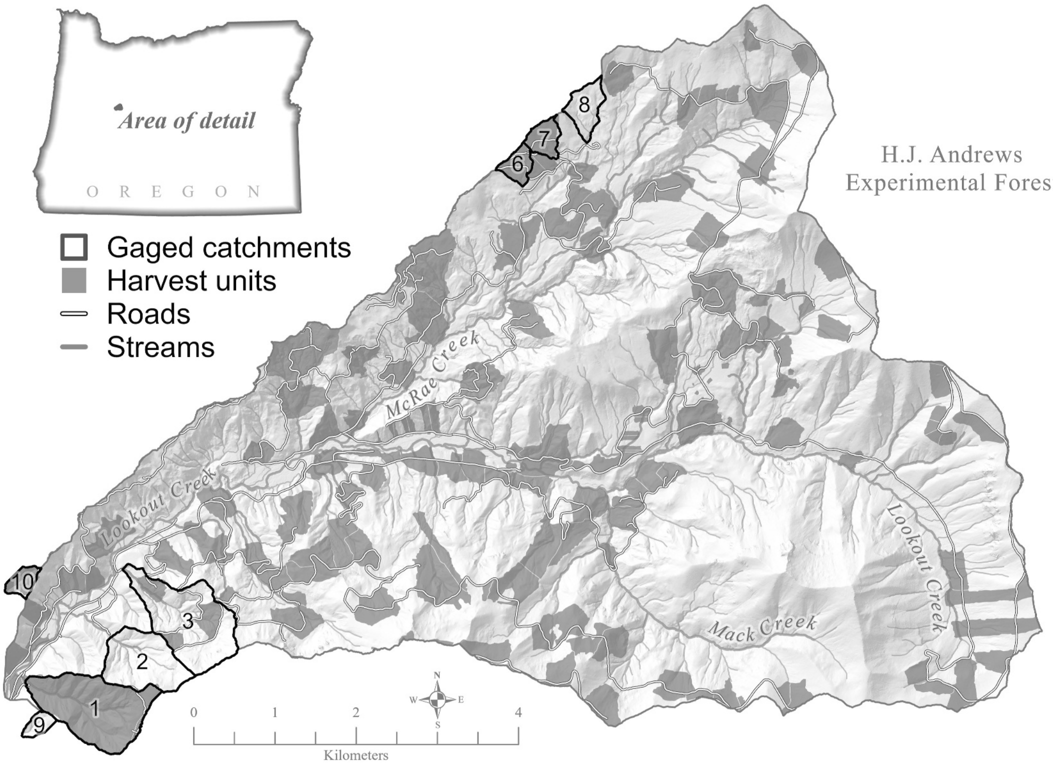

Figure 1. The H. J. Andrews Experimental Forest showing the locations of gaged paired-catchment studies (catchments with bold-black outlines) along with shaded polygons showing the locations of all harvest units within the Experimental Forest. Note that catchments 4 and 5 were never established. See Table 1 for details of treatments in the paired-catchment studies.

the catchment. WS2 served as the reference whereas roads were built into WS3 to access three smaller patch cuts that collectively covered ~25% of the catchment; 6% of the basin was roaded. The second paired-catchment study was established in 1963 (Harr et al., 1982). These catchments (WS6, WS7, and WS8) are smaller and located at an elevation where snowpacks persist throughout the winter. Prior to logging, the original forests were a mixture of Douglas-fir, western hemlock (*Tsuga heterophylla*), and Pacific silver fir (*Abies amabilis*) ranging in age from 100 to 250 years with some patches of older trees. At the start of the study, a permanent gravel road was already present in WS6, covering about 9% of the catchment area (1.2 ha). WS6 was 100% clearcut. Harvest treatments in WS7 were more complicated. It was first logged as a shelterwood cut in 1974 with the remaining overstory trees harvested in 1984 and portions of the catchment were thinned in 2001. WS8 serves as the reference catchment (Harr et al., 1982). A final paired-catchment study was established in 1968 in two small low elevation catchments. WS9 serves as the reference catchment and WS10 was 100% clearcut without roads being built into the catchment (Harr, 1986).

The length of the planned reference-by-treatment calibration periods varied among catchment pairs (Table 1). There were 10 calibration years for WS2-WS1 for both August and annual data. The WS2-WS3 catchment pair was unusual. The first full water year of data starts in 1953, roads were built into the catchment during the spring and summer of 1959, and the catchment was logged from August 1962 through February 1963. Road building had little effect on discharge, therefore water year 1963 was used as the first treatment year resulting in a 9-year calibration period for August and 10 years for annual data. There were 9 calibration years for WS8-WS6 for both August and annual data sets, whereas the WS8-WS7 had 9 calibration years for August and 10 for the annual data sets. The WS9-WS10 comparison had 6 calibration years for both August and annual data sets.

2.2. Paired-Catchment Analysis

Discharge data for our analyses were downloaded from the HJA databank (data set HF004; Johnson et al., 2023). We used mean unit-area discharge (mm/day) as our discharge metric because differences in the sizes of catchments meant that their absolute discharges were quite different. Further, expressing values in units of mm/day

Table 1
Descriptive Details of Gaged Catchments at the H. J. Andrews Experimental Forest (Modified From Johnson et al., 2021)

Watershed	Gaged area (ha)	Gage elevation (m)	Max elevation (m)	Forest origin	History	Stream gage start	August calibration years	Annual calibration years
WS1	96	439	1027	~1500 AD	100% clear cut 1962–1966; burned 1967; no roads	1952	10	10
WS2	60	545	1079	~1500 AD	Reference, no harvest	1952	REF	REF
WS3	101	476	1080	~1500 AD	6% roads 1959; 25% patch clearcut 1963	1952	9	10
WS6	13	878	1029	~1700–1850 AD	100% clearcut 1974; 9% roads	1963	9	9
WS7	15.4	918	1102	~1700–1850 AD	60% overstory harvest 1974; remaining trees harvested 1984; 12% non-commercial thin 2001	1963-1987; Restart 1995	9	10
WS8	21.4	962	1182	70% ~1850 AD; 30% ~1500 AD	Reference, no harvest	1963	REF	REF
WS9	8.5	426	731	~1500 AD	Reference, no harvest	1968	REF	REF
WS10	10.2	461	679	~1500 AD	100% clearcut 1975	1968	6	6

means that differences in the lengths of months or years have no influence on our analyses. Data downloaded from the H. J. Andrews database include flags inserted during the quality control processing, identifying observations that include estimated values as well as missing values for which no data is available. We used these flags to filter the data for analysis as described below.

The low-flow analyses were based on August monthly discharge from data set HF004 (Entity 3; Johnson et al., 2023). We used the data quality flags to delete months with less than 27 days of observations and months in which stream discharge was estimated for 10 or more days. Filtering removed between 1.74% and 4.41% of the monthly observations from the record, except for WS7 where the gage was shut down for water years 1988 through 1994 and is thus missing data from 14.58% of its period of record. Also, this choice of data filtering would have eliminated several of the calibration years for the WS9 versus WS10 analysis so for this comparison we accepted months with as many as 31 days of estimated discharge so that all six calibration years could be used in our analysis.

The analysis of total annual yields was based on data from HF004 (Entity 4; Johnson et al., 2023). We again used the data quality flags to delete years with less than 365 days of observations, and years in which stream discharge was estimated for 50 or more days. However, these filters would have eliminated three of the six calibration years for the WS9 versus WS10 comparison, so, for this catchment pair, we accepted years with as many as 90 days of estimated discharge so that all six calibration years could be used in our analysis.

Our analyses followed the methods developed by Watson et al. (2001) and Gomi et al. (2006) for the analysis of data from paired-catchment studies. We developed regression relationships between discharge from the treated and corresponding reference catchments of the general form:

$$Q_{\text{expt}} = \beta_0 + \beta_1 Q_{\text{ref}} + \epsilon \tag{1}$$

where Q_{expt} is the expected discharge in the treated catchment given an observed discharge in the reference catchment Q_{ref}, and β_0 and β_1 are regression coefficients. Errors, ϵ (i.e., residuals), were modeled with an iterative autoregressive generalized least squares (GLS) procedure (Proc AutoReg; SAS v. 9.4, SAS Institute Inc., Cary, NC, USA) with a backwards stepwise variable selection procedure to test for significant ($p < 0.05$) autocorrelation at time lags of 1, 2, and 3 years. Both the regression and the Durbin-Watson test suggested that autocorrelation was not present so simple linear regression was used throughout. Regression equations were used to predict expected discharge in the treated catchments during the post-logging period. Expected discharges were subtracted from observed discharges to calculate the change in discharge resulting from the logging treatment (i.e., $Q_{\text{obs}} - Q_{\text{expt}}$), hereafter referred to as the treatment effect.

Our analyses were based on comparisons among the three reference catchments instrumented at the HJA. We followed the same basic procedure to examine pairs of reference catchments to evaluate the MDES associated with the BACI study designs. However, because there are no treatments in any of the RxR comparisons, we refer to changes in observed versus expected flows as a RxR-response, which is analogous to the treatment effect estimated from reference-by-treated catchment comparisons but must result from some combination of measurement error, lack of representativeness, or loss of stability.

The reference catchments were never intended to function as matched catchment pairs as they differ in size and physiography, encompass different elevational ranges from rain-dominated low elevation to snow-dominated higher elevations, and were established at different times. Despite these limitations, linear fits of regressions between pairs of reference catchments for mean annual discharge have $r^2 > 0.90$, similar to pre-treatment regressions of intended reference versus treated catchment pairs (Figures S1, S2, and S3 in Supporting Information S1), giving some confidence that comparisons among pairs of reference catchments are reasonable. In contrast, r^2s for RxR comparisons of mean August discharge were lower, ranging from 0.71 to 0.32, whereas r^2s for the intended reference versus treated catchment pairs ranged from 0.92 to <0.01 (Figures S1, S2, and S3 in Supporting Information S1). However, the data collected with the V-notch weirs installed is tightly correlated (Figure 3 and S4 in Supporting Information S1), suggesting that, while the catchments might be reasonably well matched, the flat-bottomed trapezoidal flumes were poorly designed for measuring late summer low flows. In fact, the flume designs were a compromise, allowing measurement of a wide range of stream discharges while

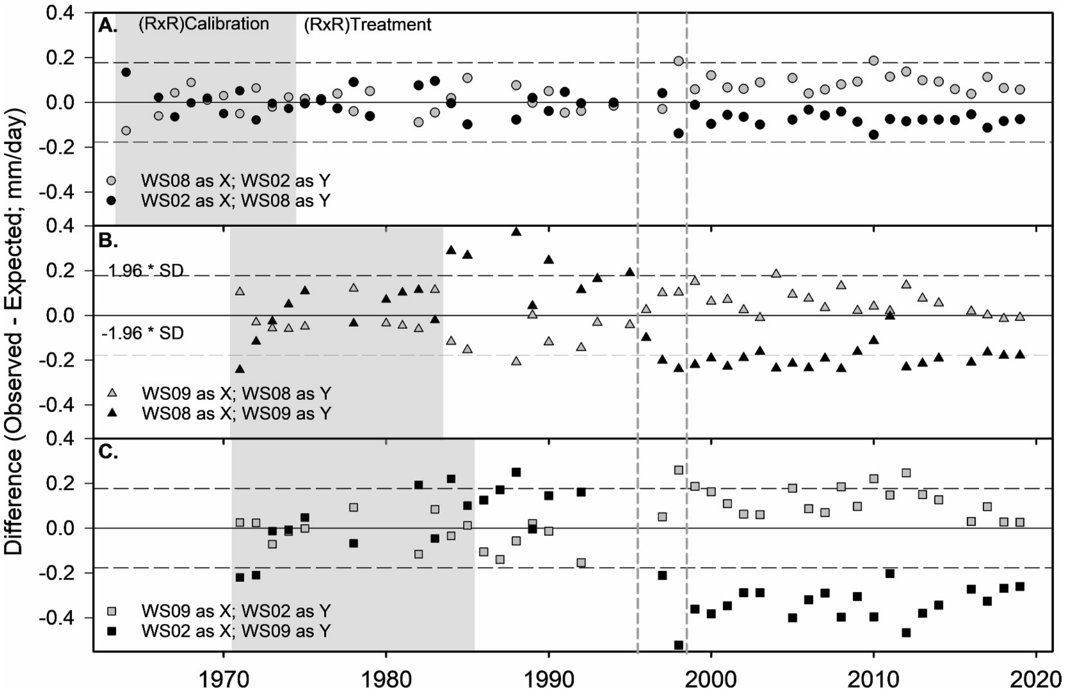

Figure 2. Results of paired-RxR-catchment analyses for all possible reference pairs from the H. J. Andrews Experimental Forest for mean August discharge (mm/day). Shaded areas show the first 10-year with non-missing data used for the calibration regression. Solid horizontal lines denote no difference between observed and expected discharge (mean difference = 0). Dashed horizontal lines show reference-based ±1.96*SD PIs (±0.177 mm/day) calculated from the combined residuals from all six pairwise RxR comparisons. Vertical dashed lines show the times of a major flood (1996) and the year in which the gages were outfitted with removable V-notch weir plates (summers since 1999).

also passing bedload sediment downstream to be captured in sediment settling ponds so that post-logging changes in erosion and sediment production could also be measured.

Reference catchments do not have predetermined reference (predictor) and treated (dependent) catchments. Therefore, we let each catchment in the reference pair serve as both the predictor and dependent catchment. For example, first using discharge data from WS2 as the x-variable and data from WS8 as the y-variable and then switching these so that WS8 is the x-variable and WS2 the y-variable. With three reference catchments (WS2, WS8, and WS9), we have six possible pairings for this analysis. Because reference catchments were established at different times, RxR comparisons are not available for the specific calibration periods used in the actual paired-catchment experiments. Thus, analyses comparing paired reference catchments selected the number of calibration observations to match the number of calibration years used in the planned reference-by-treatment paired-catchment comparisons, so n = 10, 9, or 6 depending on the catchment pair (Table 1).

2.3. Deriving ±95% Bootstrapped Confidence Intervals

We calculated the bootstrapped CIs as the ±95% confidence intervals estimated from bootstrapped samples of calibration years selected at random from the entire period of record between pairs of reference catchments. Programs were written in SAS v 9.4 (SAS Institute Inc., Cary, NC, USA) and have been archived (Wondzell et al., 2025). We followed Miller (2004), using bootstrapped sampling, without replacement, to draw 100,000 random samples from each of the six possible pairs of RxR comparisons to generate a total of 600,000 calibration data sets. We developed calibration regressions for each data set and then calculated the RxR response effect for each regression. We combined RxR-responses from all six RxR catchment pairs to generate a frequency distribution from which we calculated ±95% confidence intervals. Because the bootstrapped approach draws a random sample, rerunning the bootstrapped sampling with a different random number seed generated slightly different confidence intervals. Therefore, we ran 20 iterations of the bootstrapped sampling with 6, 9, or 10 RxR calibration years to estimate the mean and standard deviation of the ±95% CIs. Because the bootstrapped sampling captured the actual shape of the frequency distribution there was no need to assume that it followed a normal distribution. Nor was it necessary to test our bootstrapped calibration regressions for autocorrelation in the residuals. Samples were picked at random from the entire time series and will almost never represent a true time sequence. Thus, no true temporal autocorrelation can be present in the regression residuals. To ensure that the ±95% confidence intervals calculated in the bootstrapped RxR comparison were applicable to the real paired-catchment experiments, we made sure that the number of bootstrapped calibration observations exactly matched the number of calibration observations used in the actual paired-catchment experiment. Finally, the observations for the evaluation of the RxR responses were drawn randomly from years not used in the RxR regression. We will refer to these as *bootstrapped CIs* throughout the remainder of the manuscript.

2.4. Other Methods to Derive Evaluation Intervals

First, note that we use “evaluation intervals” as a general term to refer to intervals for assessing uncertainty in effect sizes. We use the term “confidence intervals” to specifically refer to the bootstrapped CIs generated through our bootstrapped analysis. The following two approaches generate evaluation intervals that are formally called “prediction intervals.” In all cases these evaluation intervals are used to estimate the MDES in paired-catchment analyses.

Many studies use prediction intervals derived from the pre-treatment calibration regressions of the actual treated and reference pairs to assess uncertainty around future observations (Harris, 1977; Moore & MacDonald, 2024; Som et al., 2012). As noted above, we refer to these as *regression-based ±95% PIs* throughout the manuscript. Note that regression-based ±95% PIs have a curved, or hour-glass shape, around the regression line. Their width is narrowest at the mean of the predictor variable (*x*-axis) and increases in width with distance from the mean. As such, the regression-based ±95% PIs will not plot as a uniform straight line across a time series of discharges. Rather, their width will be a function of the value of the predictor variable (*x*-axis; mean August discharge or mean annual discharge) which will be different each year. This results in a stepped function in which the value of the regression based ±95% PI changes every year. Further, results from paired-catchment studies are usually reported as the difference between the observed and expected values, which makes it difficult to directly plot the associated regression based ±95% PIs. We solved this problem by calculating the difference between each year's change in discharge and its accompanying prediction interval so that the distance between each value and the prediction interval accurately reflects the predictive uncertainty derived from each calibration regression.

We also calculated ±95% prediction intervals following Gomi et al. (2006) as ±1.96*SD from RxR calibration regressions. Specifically, Gomi et al. (2006) chose to use the single largest SD from multiple RxR comparisons as the base for calculating prediction intervals. However, we were concerned that selecting the largest SD could over-emphasize potential errors in the catchment data. Instead, we calculated a pooled-SD combining all residuals from all six possible pairwise comparisons. We matched the number of years used in the RxR calibration with the number of years in the actual calibrations used in the actual reference-by-treated catchment comparison. In this case, we used the first 6, first 9, or first 10 years of common record for each comparison pair. We will refer to these as *reference-based 1.96*SD PIs* throughout the remainder of the manuscript.

2.5. The Effect of Calibration Length

We explored the likelihood that calibration regressions from different length calibration periods would accurately represent the relationship observed between paired catchments over the full period of record. We limited this analysis to just the WS2 versus WS8 reference pair and compared representativeness between both mean annual discharge and mean August discharge with calibration periods of 6 and 10 years. We used the normalized delta root mean square error (ND.RMSE) as the basis for evaluating the representativeness of each calibration regression from the 100,000 bootstrapped calibration data sets:

$$\text{ND.RMSE} = \frac{\text{RMSE}_{\text{calibration}} - \text{RMSE}_{\text{full}}}{\overline{Q}} * 100 \tag{2}$$

where $\text{RMSE}_{\text{calibration}}$ is calculated from the full data set, but using parameters from the calibration regression fit to either the 6- or 10-year calibration data randomly selected from the full data set via bootstrapped sampling;

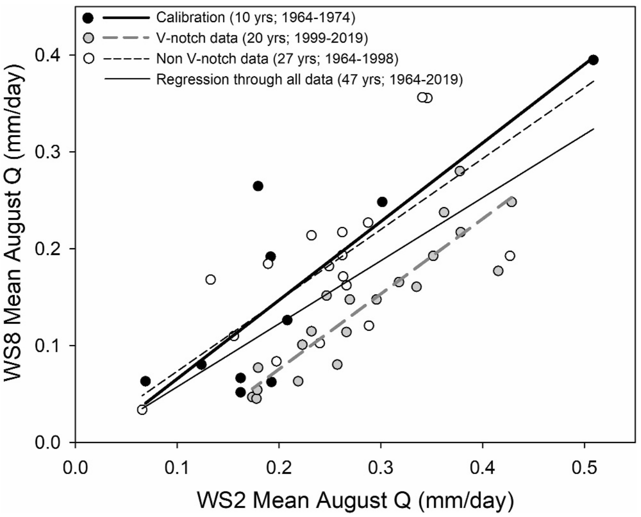

Figure 3. Relation between WS2 and WS8 in mean August discharge over all years of record and regression lines fit to specific proportions of the data: (1) The first 10 years used as a calibration period for this RxR analysis; (2) The 20 years from 1999 through 2019 when summer discharge was measured using V-notch weirs; (3) The 27 calibration and post-calibration years without V-notch weirs; and finally, (4) The regression line fit through all 47 years of data.

$\text{RMSE}_{\text{full}}$ is the RMSE determined from a regression fit to the full 52-year data set; and $\overline{Q}$ is the mean discharge for either annual or August time periods.

We suggest that the ND.RMSE is a useful metric to evaluate representativeness. First, we know that a linear ordinary least-squares regression minimizes the sums of squares of regression residuals. Thus, a linear regression through all the data from the full observation period ($n = 52$) will have the lowest possible RMSE when compared to any calibration regression fit to a subset of calibration points if the parameters from the calibration regression are used to calculate a RMSE-like statistic ($\text{RMSE}_{\text{calibration}}$) calculated from the full data set. In this case, we evaluate calibration regressions based on either 6 or 10 randomly selected calibration years. If by chance, the parameters of the calibration regression are identical to the full regression, then $\text{RMSE}_{\text{calibration}}$ will exactly equal $\text{RMSE}_{\text{full}}$. As the calibration regression becomes less and less representative of the full data set, $\text{RMSE}_{\text{calibration}}$ will increase. Therefore, the difference between these two RMSEs will be a measure of how well the calibration regression represents the full data set. We also know that the magnitude of the RMSE depends on the magnitude of the underlying data and, at the HJA, mean annual discharge (WS8, 3.22 mm/day) is much greater than mean August discharge (WS8, 0.16 mm/day). Therefore, we divide the difference in RMSE by the mean discharge to get a normalized-delta-RMSE which we abbreviate as ND.RMSE throughout the remainder of the manuscript. We arbitrarily selected an ND.RMSE of 5.0% as a threshold for accepting a calibration regression as representative. This threshold indicates that the overall errors in the estimated discharges in the RxR-treated catchment would be less than or equal to 5% of the mean discharge.

We also explored the influence that the length of the calibration period could have on the MDES using bootstrapped confidence intervals combined from all six possible pairs of RxR comparisons. We ran the bootstrapping analysis 5 times for calibration periods of 3, 4, 5, 6, 7, 9, 10, 12, and 15 years to estimate the upper and lower ±95% bootstrapped confidence intervals. Each bootstrapped run drew 100,000 random samples from each of the six possible pairs of RxR comparisons to generate a total of 600,000 calibration data sets. We then averaged the upper and lower confidence intervals across the 5 analyses. We used this analysis to examine how changing the number of calibration observations would change the width of confidence intervals for well correlated discharge data (annual discharge, Figure 8) versus poorly-correlated discharge data (August discharge, Figure 3). Direct comparisons of the CIs were not possible because the upper and lower limits vary as a function of the mean discharge. Consequently, we normalized the values, dividing CIs by the mean discharge. The resulting CIs are expressed as a proportion of the long-term mean discharge.

3. Results

3.1. Illustrating Problems With Representativeness and Stability

We examined mean August discharge from all possible pairs of reference catchments established at the HJA (Figure 2). If our reference catchments satisfied assumptions for both representativeness and stability over the long term, we would expect that the differences between observed and expected discharge over the post-RxR calibration period would have a mean of zero and that interannual variation around the mean would be similar to that observed for the RxR calibration period with most points falling between the reference-based 1.96*SD PIs. The expected pattern was evident for comparisons of WS2 and WS8 between 1975 and 1996 (Figure 2a). All other comparisons, however, appeared problematic. Most notable was the fact that after 1996 or 1999, all comparisons showed persistent runs, with differences either positive or negative relative to the expected mean (i.e., observed minus predicted = zero). Further, comparisons between WS8 and WS9, and WS2 and WS9 (Figures 2b and 2c) showed that the immediate post-RxR calibration period tended to have persistent runs in the data and the pattern flipped sign after 1996 or 1999. The fact that such striking patterns were evident in our RxR comparisons strongly suggested that some combination of underlying errors, lack of representativeness, and/or stability were influencing our data.

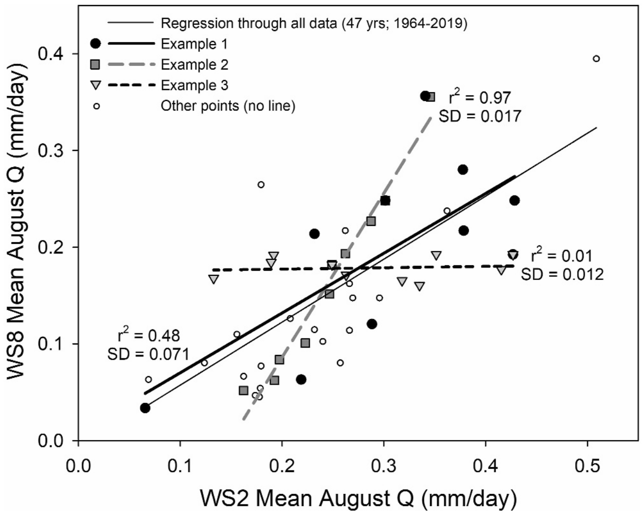

Figure 4. Three potential calibration regressions, each with 10 points (Examples 1–3) intentionally selected to illustrate a wide range in the slopes and intercepts potentially resulting from a selection of 10 calibration years demonstrating that neither high r^2 nor low standard deviation (SD) ensure that the calibration is representative of the full data set. For reference, the regression line fit to the full 47-year time series (as in Figure 3) and the other data points not included in the example regressions are shown.

The problems noted with the time-series analysis (Figure 2) stemmed, at least in part, from underlying problems in the calibration regressions. We used mean August discharge from the WS2 versus WS8 reference comparison to illustrate these problems (Figure 3). Clearly, the August data were quite noisy with substantial variation between the two catchments. The arbitrary selection of the first 10 data points from the common record, by chance, resulted in a regression slope that was much steeper than the regression fit to the full 47-year data set (Figure 3). Consequently, expected discharges in WS8 predicted from the calibration regression tended to be higher than observed (i.e., observed minus expected is negative). However, regression lines fit to the first 10 RxR calibration years were quite representative of all data collected before 1999, which explained the close fit of the differences between the observed and expected discharge in both WS2 and WS8 between 1975 and 1996 as shown in Figure 2a, The non-representativeness of the 10-year calibration to the full data set stemmed from data collected after 1999. After 1999 discharge was measured using V-notch weir plates bolted to the flumes. Both the nature of these data and the regression line fit to them were markedly different than the earlier data and their respective regression lines (Figure 3). While the slope of the regression line was similar to the calibration regression, the intercept is much lower (−0.080 vs. −0.015 mm/day). The standard deviation of the residuals was also much smaller (0.029 vs. 0.063). Clearly, the measurements of summer low-flow discharge using the V-notch weirs were more precise than comparable measurements from the trapezoidal flumes and substantially reduced the variability in the data.

To illustrate that calibration data may poorly represent overall relationships between paired catchments, we intentionally selected three such sets of calibration points and compared the resulting regressions to the regression fit to the full 47 years of data (Figure 4). These regressions give widely varying results and illustrate that neither the r^2 of the regression nor the standard deviation of the regression residuals can determine the accuracy in the prediction of the expected discharge. In fact, the regression with an $r^2 = 0.97$ and a small SD poorly matched the overall regression fit to all the data. The regression with $r^2 = 0.01$ provided no predictive power, whereas the regression line that best matched the full 47-year regression, only had an $r^2 = 0.48$ and a large SD (Figure 4).

3.2. Bootstrapped Confidence Intervals

The potential number of unique combinations of calibration data sets that can be drawn for our RxR comparisons is very large. The comparison with the smallest number is between WS2 and WS9, which has only 37 years of observations without missing data. For analyses with six calibration years, $2.32*10^6$ unique data sets are possible. For the WS2 and WS8 comparison, with a 47-year long record for catchments with 10 years of calibration data, $5.18*10^9$ unique data sets are possible. We used bootstrapped sampling to draw a random sample of 100,000 regressions from each of the six possible pairs of RxR comparisons for a total of 600,000 regressions. While drawing such a large number of bootstrapped samples poses some risk that we will include identical combinations of years in our sample, the effect of this on our results should be very small. For example, a sample of 100,000 from the WS2-WS9 RxR comparison with six calibration years only samples 4.3% of all possible combinations. Further, the RxR response is based on a single year, drawn at random from the 31 years not used in the calibration regression.

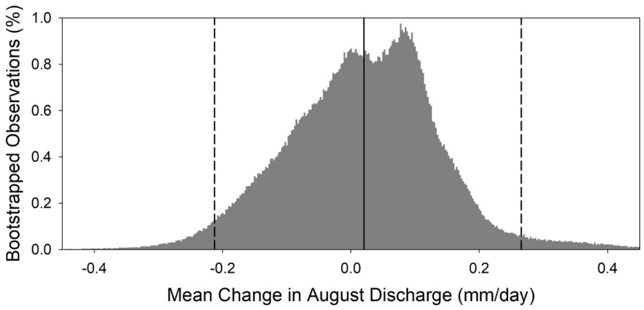

Figure 5. Frequency distribution of 600,000 estimates of RxR responses, pooled from 100,000 iterations in each of the 6-possible pairwise comparisons among the three HJA reference watersheds showing two-tailed, 95% bootstrapped confidence intervals (vertical dashed lines) resulting from RxR calibration regressions based on a calibration period of 10 years.

We generated frequency distributions from the bootstrapped RxR responses for mean August discharge (Figure 5). These frequency distributions capture the effects of both representativeness and stability. In fact, the bimodal shape of the distribution results from the loss of stability apparent in our data after 1999 (Figure 3). We then identified boundaries around the mean RxR response that contained 95% of all the bootstrapped responses. That is, non-treatment effects lead to an apparent difference between pre- and post-

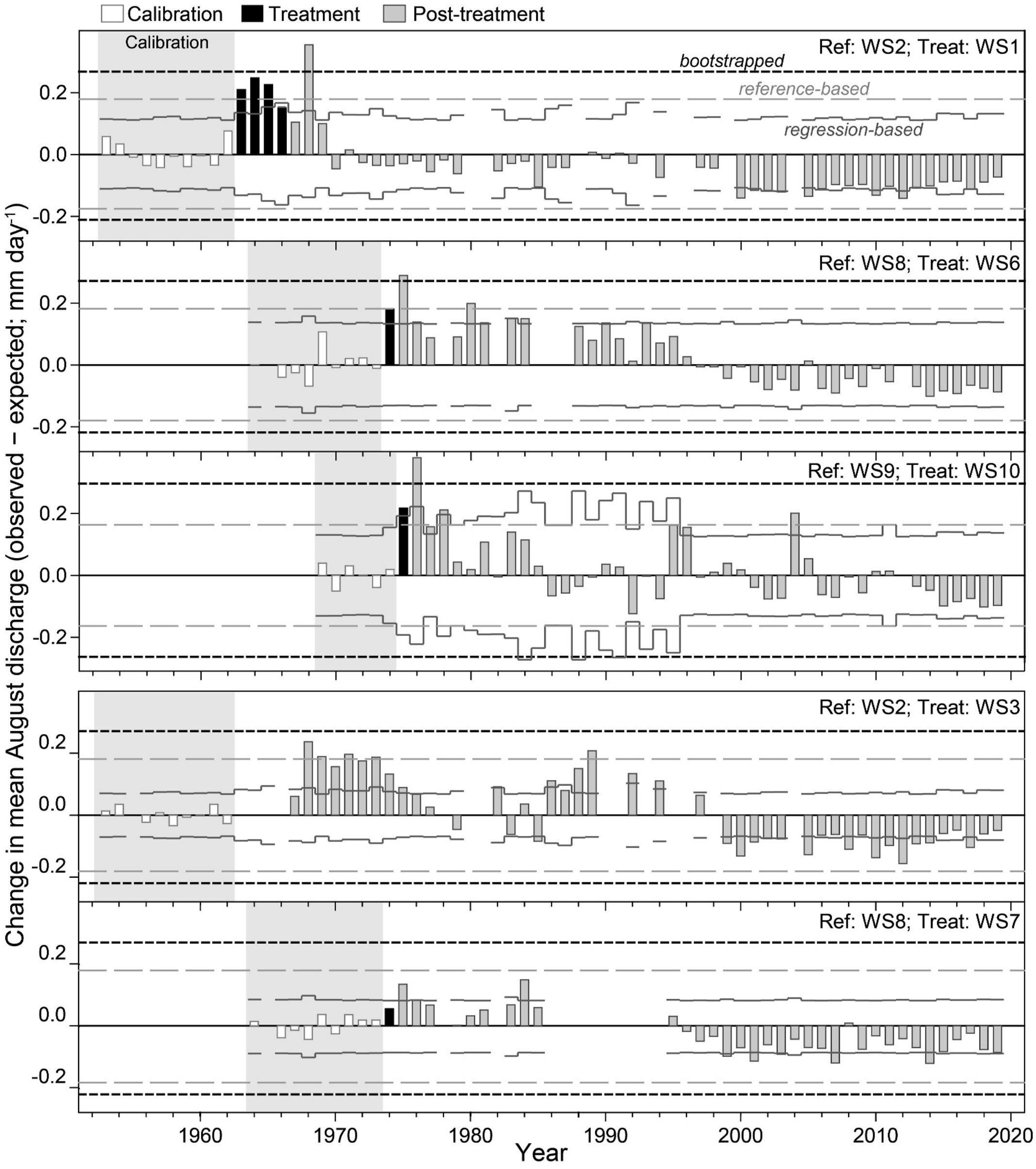

Figure 6. Changes in mean August discharge from treated catchments at the H. J. Andrews Experimental Forest relative to ±95% evaluation intervals (horizontal lines) calculated from three different methods: regression-based ±95% PIs (solid dark gray line but with gaps where data are missing); reference-based 1.96*SD PIs (light gray long dashes); bootstrapped CIs (black short dashes). WS1, WS6, and WS10 were 100% clearcut; WS3 and WS7 were partially cut. White bars with gray background indicate the calibration period, black bars indicate the treatment year(s), and gray bars indicate post-treatment years.

calibration periods that are larger than these boundaries in 5% of all RxR comparisons. These boundaries serve as ±95% confidence intervals. However, note that the bootstrapped approach draws a random sample, and even with a sample size of $6*10^5$ regressions, rerunning the bootstrapped sampling with a different random number seed resulted in tiny differences in the confidence intervals. However, the standard deviation around the mean ±95% CI (−0.21 and 0.27 mm/day) from 20 iterations of the bootstrapped sampling with 10 RxR calibration years was ≤0.001 mm/day. Some paired catchment studies at the HJA have shorter calibration periods of 9 or 6 years leading to wider confidence intervals, however, the SD around the mean of the bootstrapped CIs remained ≤0.001 mm/day. Thus, the bootstrapping approach generated robust estimates of the ±95% CIs which could be

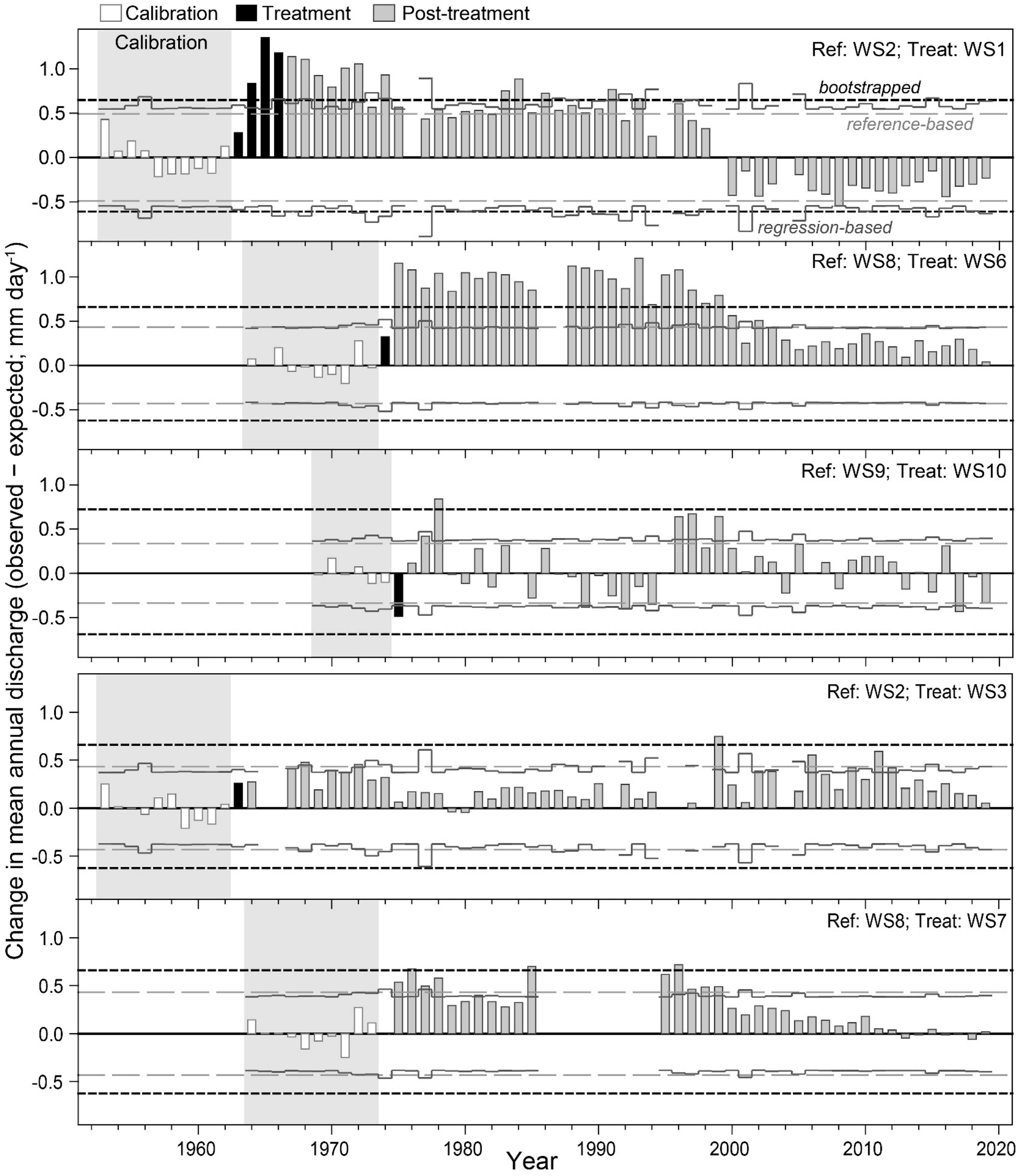

Figure 7. Changes in mean annual discharge from treated catchments at the H. J. Andrews Experimental Forest. See captions of Figure 6 for details.

applied to the real paired-catchment experiments where August monthly surpluses and deficits ranged between 0.37 mm/day and −0.27 mm/day, respectively.

3.3. Post-Treatment Changes in Low-Flow Discharges

Our re-analyses of mean August discharges from all five reference-by-treated catchment pairs from the H. J. Andrews (Figure 6) showed that increases in discharge exceeded the MDES in only a single post-harvest year in each of the three 100% clearcut catchments based on the bootstrapped CIs. In contrast, evaluations using the other ±95% evaluation intervals consistently showed significant surpluses in the early post-treatment years and frequently showed significant deficits in the last two decades of the catchment records. These differences in results were a consequence of the width of the evaluation intervals generated from each approach. In general, the regression-based ±95% PIs were the narrowest, the reference-based 1.96*SD PIs were slightly wider, and the

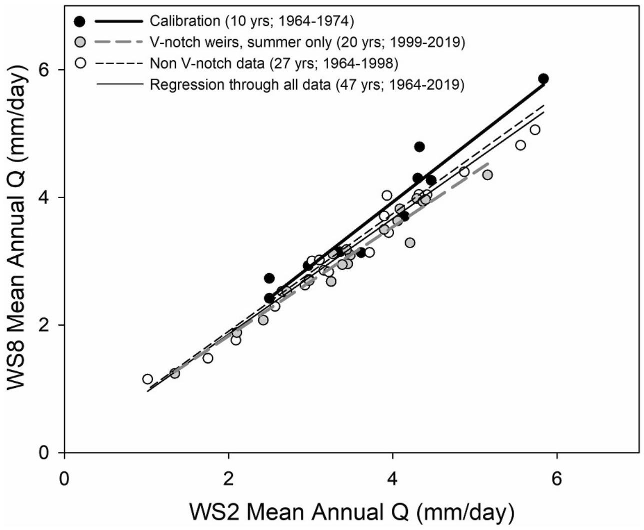

Figure 8. Relation between WS2 and WS8 in mean annual discharge over all years of record and regression lines fit to specific proportions of the data: (1) The first 10 years used as a calibration period for this RxR analysis; (2) The 20 years from 1999 through 2019 when summer discharge was measured using V-notch weirs; (3) The 27 calibration and post-calibration years without V-notch weirs; and finally, (4) The regression line fit through all 47 years of data.

bootstrapped CIs were the widest. Also, note that the time series data in the WS9 versus WS10 comparison are effectively uninterpretable. The calibration with 6 years of pretreatment data for mean August discharge had a slope of −0.01 and $r^2 < 0.01$ and provided no predictive power (intercept = 0.14 mm/day).

3.4. Post-Treatment Changes in Total Annual Yields

Post-treatment annual yields showed expected patterns with surplus runoff from most catchments in the years immediately after logging. In two of the three 100% clear-cut catchments, these surpluses exceeded the MDES from all three methods (Figure 7). While there were clear trends toward surplus runoff from the partially harvested catchments, the significance of these increases differed according to the method used to estimate the MDES. The regression-based ±95% PIs and the reference-based 1.96*SD PIs were similar and suggested that post-treatment surpluses would be significant in several years. In contrast, the bootstrapped CIs were much wider and in only one (WS2 vs. WS3) and three (WS8 vs. WS7) post-treatment years did changes in discharge exceed the MDES. Among the treated catchments, only WS1 showed consistent deficits in total annual discharge in the last two decades, but these deficits were smaller than the associated ±95% evaluation intervals except in a single year. In general, the regression-based ±95% PIs and the reference-based 1.96*SD PIs were similar in magnitude, and both were narrower than the bootstrapped CIs except for the WS2-vs-WS1 catchment comparison where all three methods returned similar CIs.

The post-treatment period shown for the WS2-vs-WS1 catchment pair is unusual with an abrupt change from surpluses to deficits in 1999. We think the abruptness of this change is an artifact of gaging problems. Each of the paired-catchment studies at the HJA has its own unique history of instrumentation. In general, stage height—discharge data were collected from the gages over the first several years after gages were first installed. These data were used to construct rating curves for each gage. Once rating curves were finalized, no further stage—discharge data were collected. Following a major flood in 1996, the HJA embarked on an effort to validate the rating curves for all catchments. For WS1 specifically, no stage—discharge data had been collected for the 40 years from 1959 through 1999. New data, collected after 2000 did not fit the rating curve developed in the early years of the study. We have not been able to identify a specific time or event that would have so changed the stage height—discharge relationship. Therefore, we decided to correct gage data back to 2000 and republished the corrected WS01 discharge data to the HJA databank. We flagged earlier data as questionable, but, without validation data from that 40-year long period, we do not know how far backwards in time we should push our corrections. We expect that the rating curve was accurate for the early post-treatment years when the paired-catchment study was actively underway, and since 2000 we have continued to collect validation data and are certain that post-2000 data are reliable.

3.5. How Long Does the Calibration Period Need to Be?

The degree to which data from the two catchments in a paired-catchment study are correlated is a critical determinant of the MDES for the treatment effect. In the examples illustrated here, the correlation in the mean annual discharges used in the RxR comparisons (Figure 8) were much higher than on the mean August discharge (Figure 3). For the annual data, all regression lines fit to different subsets of the data were similar to the regression fit to all the data over the full time period (Figure 8). Also, the annual data appear to satisfy the assumption of stability. The August data are included in the annual data, but because summer discharges are very low, the small differences in estimated discharge with and without the v-notch weirs had little effect on the total annual yield.

The difference in correlations of these two data sets had a large influence on the representativeness of our bootstrapped calibration regressions. Longer calibration periods improved the chances that the calibration period would be representative of the overall regression between catchments (Figure 9). However, there were very large differences between pairs of catchments in which discharge data were well correlated (mean annual discharge;

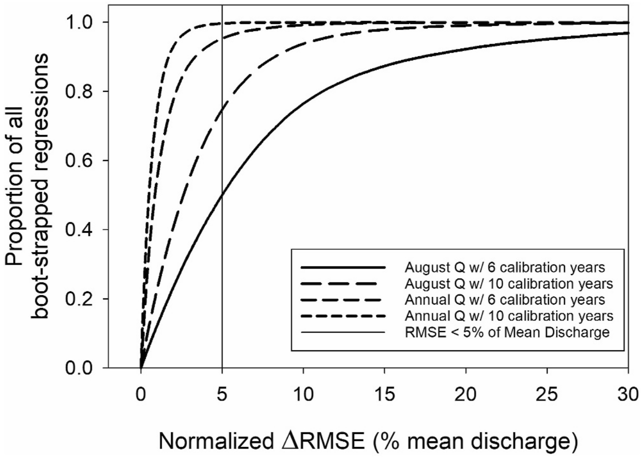

Figure 9. Cumulative frequency distribution of the ND.RMSE comparing bootstrapped calibration regressions to a single regression fit to the entire 52-year data set from the RxR-comparison between WS2 and WS8. The vertical line denotes RxR-calibration regressions that fit the full 52-year regression with less than a 5% error in the expected discharge and capture 50.05%, 74.75%, 95.30%, and 99.68% of the 100,000 bootstrapped regressions from August Q, with 6 and 10 calibration years through Annual Q with 6 and 10 calibration years, respectively.

Figure 8) versus catchment pairs that were poorly correlated (mean August discharge; Figure 3). We chose a 5% threshold for the ND.RMSE as an indicator that the calibration regression fits the overall regression determined from the full 52-year regression. In well correlated catchment pairs, 95% of randomly selected 6-year calibration periods had MDES equal to or smaller than this threshold and increasing the calibration period to 10 years showed that nearly 100% of the calibration regressions would be judged as representative (Figure 9). The situation was quite different for the poorly correlated August discharge data. Here, only half of the calibration regressions would be judged as representative with 6 years of calibration data, increasing to 75% when the calibration period was extended to 10 years (Figure 9), and to 92% when the calibration was extended to 15 years (data not shown). However, because the August discharge data are noisy, the r^2 of regressions with 15 years of calibration data are low, with a median of 0.53 and 5- and 95-percentiles of 0.23 and 0.76, respectively.

We also examined the relationship between the r^2 derived from the calibration period and the ND.RMSE calculated over the entire data set using the calibration regression. Specifically, we examined the 100,000 bootstrapped calibration regressions between WS2 and WS8, with either 6- or 10-calibration years for both mean annual and mean August discharge. For the well-correlated annual data, nearly all regressions closely fit the overall regression and had high r^2. Thus, it would be highly unlikely to get an unrepresentative calibration regression, even with calibration periods as short as 6 years. That was not the case for the poorly correlated August data. With six calibration years 7.4% of the calibration regressions had high r^2 but would not be judged as representative ($r^2 > 0.85$; ND.RMSE >5.0%). If a lower r^2 threshold was deemed acceptable, 14.8% of the calibrations would not be considered representative ($r^2 > 0.75$; ND.RMSE >5.0%). Conversely, 21.7% of representative calibration regressions (ND.RMSE <5.0%) had $r^2 < 0.500$. For mean August discharge, increasing the length of the calibration period improved the representativeness of the calibration regressions, however most of these regressions had low r^2. For example, with 15 calibration years, 40.2% of the calibration regressions had $r^2 < 0.500$ but with ND.RMSE <5.0%.

The degree of correlation in data from paired catchments also influences the width of the resulting evaluation intervals. For example, the width of the 10-calibration year bootstrapped CIs (upper limit minus lower limit) for annual RxR comparisons (Table 2) was 37% of the mean annual discharge (3.40 mm/day) when averaged over all six possible pairwise comparisons between the three reference catchments. In comparison, the width of the bootstrapped CIs for August RxR comparisons (Table 2) was more than 200% of the mean August discharge (0.23 mm/day).

Table 2
Evaluation Intervals Generated From Calibration Periods of 6, 9, and 10 Years Using Three Methods

		August			Annual		
Number of calibration years	Bounds	Regression-based ±95% PIs	Reference-based 1.96*SD PIs	Boot-strapped 95% CIs	Regression-based ±95% PIs	Reference-based 1.96*SD PIs	Boot-strapped 95% CIs
6	Lower	±0.1694	±0.1631	−0.2627	±0.3853	±0.3357	−0.6884
	Upper			0.2960			0.7241
9	Lower	±0.1168	±0.1806	−0.2188	±0.4752	±0.4325	−0.6242
	Upper			0.2701			0.6607
10	Lower	±0.1226	±0.1768	−0.2124	±0.6108	±0.4910	−0.6123
	Upper			0.2655			0.6499

Note. There is not a unique solution for generating ±95% prediction intervals from the calibration regressions across all 5 possible paired-catchment comparisons. Therefore, we only show the comparison between the WS2 versus WS1 catchment pair as an example. See text for further explanation. Units are mm/day.

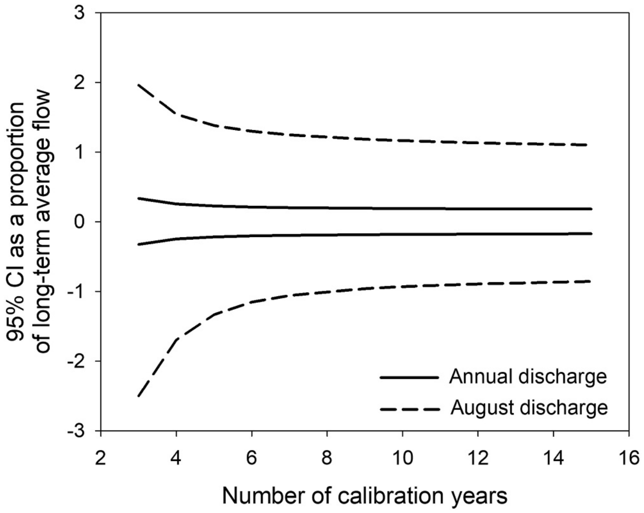

Figure 10. Change in width of bootstrapped confidence intervals with increasing number of calibration years used in a paired-catchment study, contrasting data with high correlation between catchments (annual discharge, Figure 8) versus low correlation and loss of stability (August discharge, Figure 3). Data are normalized by the long-term average flows from the three reference catchments by plotting the confidence limits as a proportion of flow.

The three different approaches for calculating the ±95% evaluation intervals showed differential sensitivity to increasing or decreasing the length of the calibration period. In general, the bootstrapped confidence intervals were widest and the regression-based ±95% PIs were narrowest (Table 2; Figures 6 and 7). Among the three methods, only the bootstrapped CIs consistently increased in width as the length of the calibration period was shortened (Table 2). In contrast, neither the width of the regression-based ±95% PIs nor the reference-based 1.96*SD PIs changed systematically with changes in the length of the calibration period. We note that it is difficult to compare ±95% evaluation intervals with different length calibration periods because the calibration periods and resulting regressions are unique to each catchment pair. Only the WS1 and WS2 catchment pair have a 10-year calibration period allowing a comparison by truncating the calibration period. Therefore, we used the WS2 versus WS1 comparison as an example, using the full 10-year period and then truncated the first year to create a 9-year calibration period and then truncating the first 4 years of data to create a 6-year calibration period.

The length of the calibration period is also a critical determinant of the power of the paired-catchment approach to identify significant treatment effects. However, only the bootstrapped CIs were robust to changes in the length of the calibration period. Therefore, we calculated the width of the bootstrapped CIs resulting from calibration periods of different lengths for both mean August and mean annual discharge (Figure 10). Bootstrapped CIs from both data sets rapidly converge to a relatively constant value with as few as 6 to 8 calibration years. Decreases in the CIs beyond 10 calibration years were very gradual. The inherent interannual variability between the data sets, however, leads to large differences in the size of the bootstrapped CIs. For example, with six calibration observations, mean August discharge would have to change by more than ±122% to be significant ($p < 0.05$) whereas for mean annual discharge, a 21% change in the mean would be significant. It is also clear that substantially lengthening the calibration period for catchment pairs with poorly correlated data (i.e., the August data set) will not substantially narrow the confidence interval.

4. Discussion

4.1. Gaging Uncertainty

Paired-catchment studies have been important to the advancement of catchment hydrology. However, their utility is challenged by difficulties in establishing clear statistical measures with which to evaluate uncertainty and to clearly quantify the minimum detectable effect sizes in post-treatment discharge. It is intuitively clear that pretreatment data from well correlated catchments should have little variability in the relationship between a reference and treated catchment. Consider a hypothetical catchment pair with a perfect relationship such that all observations over time fit perfectly on to the long-term regression line (i.e., $r^2 = 1.0$). Any selection of two or more calibration years taken from such a data set would return exactly the same regression coefficients and therefore, any of the three approaches used here to evaluate the MDES would give identical results. It is, of course, unreasonable to expect any catchment pair to have a perfect relationship. In fact, there can be large differences in the degree to which data from catchment pairs are correlated—both among different catchment pairs as well as among seasons or specific time periods of interest, as was illustrated in this study by comparing the relationships between WS2 and WS8 for annual discharge versus August discharge.

Scientists starting a new paired-catchment experiment face a dilemma. What proportion of resources should be devoted to the calibration period before starting the treatment? Bren and Lane (2014) were one of the few studies to attempt to answer this question. They examined data from two paired-catchment studies in Australia. They subdivided the calibration observations into a calibration data set and a verification data set against which calibration relationships using different length calibration periods could be tested. They used the Nash-Sutcliffe coefficient of efficiency (NSE) to evaluate the quality of the calibration relationship. They determined that little further improvement in the calibration relationships occurred by extending the calibration period beyond three years, with similar results for data analyzed on a daily ($n = 365$ per year), monthly ($n = 12$ per year), quarterly ($n = 4$ per year), or annual ($n = 1$ per year) basis. However, they worked with relatively short data sets of 11 calibration years and 3 verification years in one case and only 3 calibration and 3 verification years in the second case. Their technique then optimizes for a high-quality calibration relationship without an overly long calibration period. However, their calibration relationship, like in most other paired-catchment studies, is based on only a few years of data. As we have shown with RxR comparisons, the quality of the calibration relationship (NSE, r^2, RMSE) will not effectively evaluate the representativeness and stability of a calibration relationship over a lengthy post-treatment monitoring period. For example, a calibration regression with a high r^2 might be interpreted as suggesting that discharge from the reference catchment explains most of the variability in the discharge from the treated catchment and thus, the catchments are both well correlated and that the calibration regression is likely to be representative. We tested this by comparing the r^2 from the 6- or 10-year calibration regression with the ND.RMSE for the entire data set using the 100,000 bootstrapped calibrations between WS2 and WS8. Our results clearly showed that there were high probabilities of making both type I and type II errors when using short calibration periods with catchments with poorly correlated discharge data. And while the probability of getting a calibration regression with a high r^2 that was poorly representative (high ND.RMSE) was low, our analysis still suggests that this might occur from 7% to 15% of the time—and without additional data, there is no way to know that the catchments are poorly correlated and the calibration is non-representative.

Our analysis of different methods of estimating ±95% evaluation intervals suggested that only the bootstrapped CIs were robust. It is clear from our analyses that the width of the evaluation intervals should narrow systematically as the length of the calibration period is increased. This was not always observed in the different methods we examined. The calibration data sets from which the prediction intervals are generated from real paired-catchment studies are effectively a single systematic random sample of all possible calibration data sets. When the calibration period begins, there is no a priori way to know if the calibration data will be representative of both the long-term relationship between the catchments and the interannual variability around this relationship. Thus, decisions on when to start and end the calibration period act like a random sample, and as shown from the bootstrapped sampling, there are many possible calibration data sets. Only the bootstrapped sampling approach explores the full range of variability within the possible calibration data sets. As a consequence, only the bootstrapped approach generates robust evaluation intervals. Unfortunately, the bootstrapping approach developed here cannot be applied at the beginning of a study. It relies on long-term data collected from multiple reference catchments. Thus, the bootstrapped analyses can only be applied decades after the study was established.

The rapid convergence of the bootstrapped CIs to a stable CI should not be interpreted as support for the use of short calibration intervals in paired-catchment studies that use either the regression-based ±95% PIs or the reference-based 1.96*SD PIs to establish a MDES. The regression-based ±95% PI approach only selects a single calibration data set; the reference-based 1.96*SD PI approach selects one calibration data set for each of the possible reference pairs. CIs estimated using these approaches with small sample sizes are subject to substantial sampling artifacts. For example, choosing to increase or decrease the calibration period by just 1 year adds or deletes a data point with unknown fit to the regression line for a catchment pair. If the calibration period is short, a single data point can have a large relative influence on the calibration regression, either improving it or making it worse. It is likely that longer calibration periods would be helpful if using either the regression-based ±95% PIs or the reference-based 1.96*SD PIs. However, our work highlights problems with unreplicated paired-catchment studies and we cannot recommend either of these approaches as a reasonable way to estimate the MDES.

4.2. The HJA Paired-Catchment Studies Revisited

Our reanalysis of the HJA paired-catchment results using bootstrapped ±95% CIs suggests that only the increases in mean annual discharge in the years immediately following logging in 100% clearcut catchments were larger than the MDES. These results affirm some of the earliest published results from the scientists who originally established these studies (Harr et al., 1982; Rothacher, 1970). Our results also show that effect sizes observed from partially harvested catchments (Harr et al., 1982; Rothacher, 1970) were smaller than our estimate of the MDES. This reanalysis of the HJA's paired-catchment results poses a dilemma. A long history of paired-catchment studies provides definitive support for a relationship between the amount of catchment area logged and the magnitude of the post-harvest surplus in annual average discharge. Further, these results are strongly supported by our mechanistic understanding of catchment hydrology. In this case, then, the issue appears to result entirely from the analytical power of the paired-catchment approach relative to the magnitude of the treatment effect generated by partial harvest. Clearly, it would be unreasonable to use a threshold-based p-value to conclude that logging had no effect on stream discharge (Goodman, 2016), but at the same time, it is important to recognize problems inherent in paired-catchment studies that make it difficult to accurately measure small changes over very long time periods.

The bootstrapped confidence intervals for mean August discharge are much larger than the late-summer low-flow deficits observed in the treated watersheds. The large MDES result from the poor correlation of mean August discharges among catchments combined with the loss of stability after the V-notch weir plates were installed on the gages (Figures 3 and S4 in Supporting Information S1). As described earlier, the flat-bottomed trapezoidal flumes used in the H. J. Andrews paired-catchment studies were never designed to accurately and precisely measure late summer low-flow discharges. Thus, the moderate to low correlations among these data might suggest that they should not be used in paired catchment studies to evaluate the effects of forest regrowth on stream flow. However, we include them here for two reasons. First, they enable us to compare the effects of well correlated data (mean annual discharge) versus less well correlated data (mean August discharge) on MDES. Secondly, the low flow data from the HJA catchments have been previously published and used to highlight potential environmental effects of plantation forests on late summer low flows. Both Hicks et al. (1991) and Perry and Jones (2017) showed that conversion of old-growth forests to plantation forests resulted in late-summer low-flow deficits and suggested that these deficits could be detrimental to cold-water dependent salmonids and might exacerbate issues between in-stream flows and other beneficial uses. As such, previous analyses of the H. J. Andrews low-flow data are influencing land management within the region. Given that context, it is important to realistically evaluate the underlying uncertainty in estimated changes in discharge spanning periods of 50–70 years.

The problems with poorly correlated late-summer discharges and loss of stationarity once V-notches were installed suggest, perhaps, that the post-harvest deficits observed decades after treatment may be artifacts of some change in the way discharge was measured. However, the analyses of Hicks et al. (1991) and Perry and Jones (2017), as well as our analysis showed that the deficits were observed consistently in most of the treated catchments. It is difficult to imagine a methodological problem that would consistently generate deficits in only the treated catchments. We would expect that sampling error, problems with representativeness, and stability would have a somewhat random effect and generate surpluses for some treated catchments and deficits for others. Thus, the consistent deficits suggest that the late-summer low-flow deficits are real but that the underlying sources of uncertainty suggest that we are unable to realistically estimate their magnitude. It appears that, over time, hydrologic recovery combined with problems stemming from measurement error, representativeness of the calibration period, and loss of stability limited our ability to interpret changes in discharge occurring many decades after the initial logging treatments.

4.3. Implications for Future Catchment Studies

The high variability in the results of paired-catchment studies globally has been known since Hibbert's 1967 review. He found broad general agreement among the 39 studies he reviewed that (a) reducing forest cover increased annual water yields, (b) that increased forest cover reduced yields, and that the (c) "*response to treatment is highly variable and, for the most part, unpredictable*" (Hibbert, 1967; emphasis ours). Since then, further studies have suggested that the variability in treatment responses is due to a wide variety of factors, including differences in "*annual rainfall, vegetation type, ET regime, aspect and slope, leaf area reduction, geology, soil type, soil moisture, and soil depth*" (Neary, 2016). However, the potential for non-treatment effects—that is measurement error, lack of representativeness, or loss of stability—to influence results is seldom considered. It was only possible to estimate the potential influence of non-treatment effects and establish rigorous estimates of the MDES for the HJA catchments because three reference catchments had been established and maintained over the long term. Based on our results, we join Wicht (1943) and Underwood (1991) and recommend that all future paired-catchment studies include multiple reference catchments.

5. Conclusions

Gaged, long-term catchments are "outdoor laboratories" critical for the continued advancement of hydrology and related disciplines (Tetzlaff et al., 2017). Here, we investigated one aspect of studies conducted in these catchments. Many of today's long-term research catchments were first established as paired-catchment studies and there is great temptation to continue comparisons between reference and treated catchments. However, these comparisons rely on discharge data collected during the calibration period as much as 50- to 70-year previously. Unfortunately, most paired-catchment studies lack methods to ensure that the calibration relationship was initially representative and that the relationship remains stable over long time periods. Thus, these studies cannot determine the minimum detectable effect size. This will be especially problematic if data from paired catchments are poorly correlated, in which case the calibration regression may not be representative of the overall long-term relationship between catchment pairs. Unfortunately, it will always be difficult to determine if catchments are well or poorly correlated from short calibration periods. If replicated reference catchments are available from long-running paired-catchment studies, reference-by-reference comparisons, using bootstrapped sampling, can be used to establish robust confidence intervals.

Our bootstrapping analysis clearly demonstrated that paired-catchment studies, at least as implemented at the HJA, have the statistical power to identify large changes in discharge resulting from whole catchment disturbances in the years immediately following major disturbances (i.e., 100% clearcut logging). However, the analyses presented here suggest that the paired-catchment approach lacks the analytical power needed to statistically resolve small changes in discharge resulting from less extreme disturbances (i.e., partial cutting). Similarly, the paired-catchment approach would appear to lack the power to statistically resolve changes in discharge resulting from changes in forest age, composition, or structure during forest regrowth and vegetation succession. We recognize that these findings are specific to the HJA; further research would be needed to confirm that long-term paired-catchment studies at other sites face similar issues with representativeness and stability. This study does suggest, however, that the effects of processes that might influence discharge many decades after the initial treatment may not be measurable in paired-catchment studies. Further, because paired-catchment studies rely on calibration periods from the early years of the study, long-running studies risk loss of stability. Without replication that loss of stability will be interpreted as a treatment effect, potentially leading to a misinterpretation of the influence of forest harvest and regrowth on catchment hydrology. We reemphasize that long-term studies are valuable in their own right, however, we also urge caution when evaluating treatment effects from unreplicated, paired-catchment studies.

Data Availability Statement

Data from this study are archived in the H. J. Andrews Databank, including meteorological data (Daly et al., 2019) at (http://andlter.forestry.oregonstate.edu/data/abstract.aspx?dbcode=MS001) (https://doi.org/10.6073/pasta/c021a2ebf1f91adf0ba3b5e53189c84f) and stream discharge data (Johnson et al., 2023) at http://andlter.forestry.oregonstate.edu/data/abstract.aspx?dbcode=HF004 (https://doi.org/10.6073/pasta/0066d6b04e736af5f234d95d97ee84f3). These data were provided by the H. J. Andrews Experimental Forest and Long-Term Ecological Research program, administered cooperatively by the United States Department of Agriculture Forest Service, Pacific Northwest Research Station, Oregon State University, and the Willamette National Forest. This material is based upon work supported by the National Science Foundation under the Grant LTER8 DEB-1440409. The programs used for the data analyses described in the paper have been archived and are available to download at: https://zenodo.org/records/16921946 (Wondzell et al., 2025). The files include the authors' versions of the input data sets, the analysis programs, and examples of program outputs, along with descriptions of the materials in each subdirectory.

Acknowledgments
We thank all those who originally designed and implemented, and subsequently maintained the H. J. Andrews Experimental Forest's stream gages, other hydrologic measurements, and managed the data for the Watershed Program over the last 7 decades, including Jack Rothacher, Richard Fredriksen, Ross Mersereau, Dennis Harr, Fred Bierlemeir, Al Levno, Terry Cryer, Craig Creel, Greg Downing, John Moreau, Suzanne Remillard, Jay Sexton, Mark Schulze, Adam Kennedy, and Ben Nash, and in recent years, the wildland fire fighters who have worked tirelessly to protect the HJA's research facilities from wildfires that have burned over two-thirds of the HJA since 2020. We thank Adam Price for collegial review and comments on the paper. We also thank the Journal's Associate Editor, and 2 anonymous reviewers for edits and comments that helped improve the paper. Funding for this project was provided by the USDA Forest Service, Pacific Northwest Research Station and the USDI Bureau of Land Management.

References

Amatya, D. M., Herbert, S., Trettin, C. C., & Hamidi, M. D. (2021). Evaluation of paired watershed runoff relationships since recovery from a major hurricane on a coastal forest—A basis for examining effects of Pinus palustris restoration on water yield. *Water*, *13*(21), 3121. https://doi.org/10.3390/w13213121

Bates, C. G., & Henry, A. J. (1928). Second phase of streamflow experiment at Wagon Wheel Gap, Colo. *Monthly Weather Review*, *56*(3), 79–80. https://doi.org/10.1175/1520-0493(1928)56<79:sposea>2.0.co;2

Bosch, J. M., & Hewlett, J. D. (1982). A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. *Journal of Hydrology*, *55*(1–4), 3–23. https://doi.org/10.1016/0022-1694(82)90117-2

Bren, L. J., & Lane, P. N. J. (2014). Optimal development of calibration equations for paired catchment projects. *Journal of Hydrology*, *519*(2014), 720–731. https://doi.org/10.1016/j.jhydrol.2014.07.059

Brown, A. E., Zhang, L., McMahon, T. A., Western, A. W., & Vertessy, R. A. (2005). A review of paired catchment studies for determining changes in water yield resulting from alterations in vegetation. *Journal of Hydrology*, *310*(1–4), 28–61. https://doi.org/10.1016/j.jhydrol.2004.12.010

Campbell, J. L., Yanai, R. D., Green, M. B., Likens, G. E., See, C. R., Bailey, A. S., et al. (2016). Uncertainty in the net hydrologic flux of calcium in a paired-watershed harvesting study. *Ecosphere*, *7*(6), e01299. https://doi.org/10.1002/ecs2.1299

Daly, C., Schulze, M., & McKee, W. (2019). Meteorological data from benchmark stations at the HJ Andrews Experimental Forest, 1957 to present. Long-Term Ecological Research [Database]. *Forest Science Data Bank*. https://doi.org/10.6073/pasta/c021a2ebf1f91adf0ba3b5e53189c84f

Dunford, E. G., & Fletcher, P. W. (1947). Effect of removal of stream-bank vegetation upon water yield. *EOS: Transactions American Geophysical Union*, *28*(1), 105–110.

Gomi, T., Moore, R. D., & Dhakal, A. S. (2006). Headwater stream temperature response to clear-cut harvesting with different riparian treatments, coastal British Columbia, Canada. *Water Resources Research*, *42*(8), W08437. https://doi.org/10.1029/2005WR004162

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science*, *352*(6290), 1180–1181. https://doi.org/10.1126/science.aaf5406

Gronsdahl, S., Moore, R. D., Rosenfeld, J., McCleary, R., & Winkler, R. (2019). Effects of forestry on summertime low flows and physical fish habitat in snowmelt-dominant headwater catchments of the Pacific Northwest. *Hydrological Processes*, *33*(25), 3152–3168. https://doi.org/10.1002/hyp.13580

Harr, D. R. (1986). Effects of clearcutting on rain-on-snow runoff in western Oregon: A new look at old studies. *Water Resources Research*, *22*(7), 1095–1100. https://doi.org/10.1029/wr022i007p01095

Harr, D. R., Levno, A., & Mersereau, R. (1982). Streamflow changes after logging 130-year-old Douglas-fir in two small watersheds. *Water Resources Research*, *18*(3), 637–644. https://doi.org/10.1029/wr018i003p00637

Harr, D. R., & McCorison, M. F. (1979). Initial effects of clearcut logging on size and timing of peak flows in a small watershed in western Oregon. *Water Resources Research*, *15*(1), 90–94. https://doi.org/10.1029/wr015i001p00090

Harris, D. D. (1977). *Hydrologic changes after logging in two small Oregon coastal watersheds*. Geological Survey Water Supply Paper 2037. US Dept. of Interior, U.S. Geological Survey. https://pubs.usgs.gov/wsp/2037/report.pdf

Hibbert, A. R. (1967). Forest treatment effects on water yield. In W. E. Sopper & H. W. Lull (Eds.), *International Symposium For. Hydrology* (pp. 527–543). Pergamon.

Hicks, B. J., Beschta, R. L., & Harr, R. D. (1991). Long-term changes in streamflow following logging in western Oregon and associated fisheries implications. *JAWRA Journal of the American Water Resources Association*, *27*(2), 217–226. https://doi.org/10.1111/j.1752-1688.1991.tb03126.x

Hoover, M. D. (1944). Effect of removal of forest vegetation upon water yields. *EOS: Transactions of the American Geophysical Union*, *25*, 969–975.

Hornbeck, J. W., Adams, M. B., Corbett, E. S., Verry, E. S., & Lynch, J. A. (1993). Long-term impacts of forest treatments on water yield: A summary for northeastern USA. *Journal of Hydrology*, *150*(2–4), 323–344. https://doi.org/10.1016/0022-1694(93)90115-p

Hornbeck, J. W., Martin, C. W., & Eagar, C. (1997). Summary of water yield experiments at Hubbard Brook experimental forest, New Hampshire. *Canadian Journal of Forest Research*, *27*(12), 2043–2052. https://doi.org/10.1139/x97-173

Janisch, J. E., Wondzell, S. M., & Ehinger, W. J. (2012). Headwater stream temperature: Interpreting response after logging, with and without riparian buffers. *Forest Ecology and Management*, *270*, 302–313. https://doi.org/10.1016/j.foreco.2011.12.035

Johnson, S., Wondzell, S., & Rothacher, J. (2023). Stream discharge in gaged watersheds at the HJ Andrews Experimental Forest, 1949 to present. Long-Term Ecological Research [Database]. *Forest Science Data Bank*. https://doi.org/10.6073/pasta/0066d6b04e736af5f234d95d97ee84f3

Johnson, S. L., Henshaw, D., Downing, G., Wondzell, S., Schulze, M., Kennedy, A., et al. (2021). Long-term hydrology and aquatic biogeochemistry data from H. J. Andrews experimental forest, Cascade Mountains, Oregon. *Hydrological Processes*, *35*(5), e14187. https://doi.org/10.1002/hyp.14187

Keppeler, E. T., & Ziemer, R. R. (1990). Logging effects and streamflow: Water yield and summer low flows at Caspar Creek in northwestern California. *Water Resources Research*, *26*(7), 1669–1679. https://doi.org/10.1029/WR026i007p01669

Leach, J. A., Hudson, D. T., & Moore, R. D. (2022). Assessing stream temperature response and recovery for different harvesting systems in northern hardwood forests using 40 years of spot measurements. *Hydrological Processes*, *36*(11), e14753. https://doi.org/10.1002/hyp.14753

Miller, D. P. (2004). Bootstrap 101: Obtain robust confidence intervals for any statistic. In *SUGI 29 Proceedings: SAS Users Group International Conference, May 9-12, 2004, Palais Des Congrès de Montréal, Montréal, Canada*. Paper 193-29.

Moore, R. D., & MacDonald, R. J. (2024). Quantifying the influence of forestry and forest disturbance on stream temperature: Methodologies and challenges. *Hydrological Processes*, *38*(7), e15223. https://doi.org/10.1002/hyp.15223

Neary, D. G. (2016). Long-term Forest paired catchment studies: What do they tell US that landscape-level monitoring does not? *Forests*, *7*(8), 164. https://doi.org/10.3390/f7080164

Perry, T. D., & Jones, J. A. (2017). Summer streamflow deficits from regenerating Douglas-fir forest in the Pacific Northwest, USA. *Ecohydrology*, *10*(2), e1790. https://doi.org/10.1002/eco.1790

Rothacher, J. (1965). Streamflow from small watersheds on the western slope of the Cascade Range of Oregon. *Water Resources Research*, *1*(1), 125–134. https://doi.org/10.1029/wr001i001p00125

Rothacher, J. (1970). Increases in water yield following clear-cut logging in the Pacific Northwest. *Water Resources Research*, *6*(2), 653–658. https://doi.org/10.1029/wr006i002p00653

Rowe, P. B. (1963). Streamflow increases after removing woodland-riparian vegetation from a Southern California watershed. *Journal of Forestry*, *61*, 365–370.

Segura, C., Bladon, K. D., Hatten, J. A., Jones, J. A., Hale, V. C., & Ice, G. G. (2020). Long-term effects of forest harvesting on summer low flow deficits in the Coast Range of Oregon. *Journal of Hydrology*, *585*, 124749. https://doi.org/10.1016/j.jhydrol.2020.124749

Som, N. A., Zégre, N. P., Ganio, L. M., & Skaugset, A. E. (2012). Corrected prediction intervals for change detection in paired watershed studies. *Hydrological Sciences Journal*, *57*(1), 134–143. https://doi.org/10.1080/02626667.2011.637494

Ssegane, H., Amatya, D. M., Muwamba, A., Chescheir, G. M., Appelboom, T., Tollner, E. W., et al. (2017). Calibration of paired watersheds: Utility of moving sums in presence of externalities. *Hydrological Processes*, *31*(20), 3458–3471. https://doi.org/10.1002/hyp.11248

Stewart-Oaten, A., & Bence, J. R. (2001). Temporal and spatial variation in environmental impact assessment. *Ecological Monographs*, *71*(2), 305–339. https://doi.org/10.1890/0012-9615(2001)071[0305:tasvie]2.0.co;2

Tetzlaff, D., Carey, S. K., McNamara, J. P., Laudon, H., & Soulsby, C. (2017). The essential value of long-term experimental data for hydrology and water management. *Water Resources Research*, *53*(4), 2598–2604. https://doi.org/10.1002/2017WR020838

Udawatta, R. P., Krstansky, J. J., Henderson, G. S., & Garrett, H. E. (2002). Agroforestry practices, runoff, and nutrient loss: A paired watershed comparison. *Journal of Environmental Quality*, *31*(4), 1214–1225. https://doi.org/10.2134/jeq2002.1214

Underwood, A. J. (1991). Beyond BACI: Experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Australian Journal of Marine & Freshwater Research*, *42*(5), 569–587. https://doi.org/10.1071/mf9910569

Watson, F., Vertessy, R., McMahon, T., Rhodes, B., & Watson, I. (2001). Improved methods to assess water yield changes from paired-catchment studies: Application to the Maroondah catchments. *Forest Ecology and Management*, *143*(1–3), 189–204. https://doi.org/10.1016/s0378-1127(00)00517-x

Webster, K. L., Leach, J. A., Hazlett, P. W., Buttle, J. M., Emilson, E. J. S., & Creed, I. F. (2022). Long-term stream chemistry response to harvesting in a northern hardwood forest watershed experiencing environmental change. *Forest Ecology and Management*, *519*, 120345. https://doi.org/10.1016/j.foreco.2022.120345

Wicht, C. L. (1943). Determination of the effects of watershed management on mountain streams. *EOS: Transactions of the American Geophysical Union*, *2*, 594–608.

Wondzell, S., Johnson, S., Gordon, G., Henshaw, D., & Ward, A. (2025). SAS analysis programs documenting the analyses reported in the paper: "Rethinking paired-catchment studies": Should we be replicating our controls? [SAS programs and datasets]. *Water Resources Research*, *61*, e2024WR038981. Zenodo. https://doi.org/10.5281/zenodo.16921946

Wright, C. M. (2023). *Effects of forest harvest, floods, and wildfire on bedload export from headwater catchments in the H.J. Andrews experimental forest, 1957–2022* (MS Thesis) (p. 115). Oregon State University.