WATER RESOURCES RESEARCH, VOL. 26, NO. 9, PAGES 2079–2087, SEPTEMBER 1990

Problems in Determining the Return of a Watershed to Pretreatment Conditions: Techniques Applied to a Study at Caspar Creek, California

ROBERT B. THOMAS

Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture Arcata, California

Using a previously treated basin as a control in subsequent paired watershed studies requires the control to be stable. Basin stability can be assessed in many ways, some of which are investigated for the South Fork of Caspar Creek in northern California. This basin is recovering from logging and road building in the early 1970s. Three storm-based discharge characteristics (peak discharge, quick flow, and total storm flow), daily flows, and concentration of suspended sediment were studied to see if the South Fork can be used as a control in a second experiment. Mean sediment concentration in three discharge classes and regression parameters for the other data were tested to estimate remaining treatment effects relative to the North Fork. Patterns of change were similar for most data, with rises in response followed by returns toward pretreatment conditions. The storm and sediment data showed few significant differences, but tests on daily flows indicated that differences still exist. The overall evidence suggests that the South Fork has returned to near pretreatment conditions. Better sediment data are needed for studies of the effects of land management.

INTRODUCTION

Paired experimental watersheds are widely used in forestry research to assess changes in water and sediment regimes resulting from land management treatments. Because of strong interest in determining the effects of management, and because paired studies are usually long-term, existing experimental watersheds are often used in sequences of studies. In some cases, catchments treated in one study must be used as controls for later ones.

It is difficult to decide when formerly treated watersheds are sufficiently stable to be used as controls. Responses to treatment and early recovery are fairly rapid, but subsequent recovery slows and can be hard to detect with the limited amount of data usually available. Statistical methods must form the basis for the decisions, but other factors such as management and political requirements also play a role. These problems are discussed as they pertain to the Caspar Creek North and South Fork experimental watersheds.

The Caspar Creek basins each have drainage areas of nearly 5 km^2 and are located near Fort Bragg on the California coast about 210 km north of San Francisco. The soil series in the basins are primarily Vandamme loam or Dehaven-Hotel or Irmulco Tramway complexes overlying Cretaceous sedimentary rocks. The climate typifies the northern California coast which is characterized by dry mild foggy summers and approximately 1100 mm of rainfall from October through April. In the early 1960s the basins were covered with second growth stands of redwood (*Sequoia sempervirens* (D. Don)Endl.), Douglas fir (*Pseudotsuga menziesii* (Mirb.)Franco), western hemlock (*Tsuga heterophylla* (Raf.)Sarg.), and grand fir (*Abies grandis* (Dougl. ex D. Don)Lindl.). The area had been clear-cut and burned in the late 1800s.

A study from 1962 to 1976 (referred to as the "first" study) assessed the effects of road building and selective cutting of second growth redwood forests in the South Fork on water and sediment outputs using the untreated North Fork as a control [*Rice et al.*, 1979; *Ziemer*, 1981; *Wright*, 1985; *Sendek*, 1985; *Keppeler*, 1986]. Logging part of the North Fork in 1986 initiated treatment for a second study to investigate the effects of clear-cut logging on changes in discharge and sediment yield in the North Fork [*Thomas*, 1988*b*]. One part of the second study will assess treatment effects on water and sediment production in the North Fork. The South Fork will be used as a control in this overall test. The suitability of the South Fork as a control for the second study is examined by determining whether the South Fork basin has returned to pretreatment behavior or to some predictable stationary state.

Response to treatment is time-dependent. Logging and road building can cause initial increases in discharge and sediment delivery which decreases over time as roads and harvested areas stabilize and vegetation regrows. Treatments applied at different times produce complex response patterns that depend on treatment types, their location, time of application, and the subsequent weather. When considering a previously treated basin as the control for a future study, the level of basin response and any recent changes should be known. The response may not be at pretreatment levels, but should be stable enough to serve as a control.

Several basin outputs can be used as variables to assess treatment effects. The variables differ mainly in the time periods over which water or sediment is summed or averaged. Examples are peak discharge, average daily flow, and total storm sediment discharge. The length of period over which estimates are made has a large effect on test sensitivity and, consequently, on its outcome.

Data for the first study were intended to estimate annual fluxes, while the second study focuses on peaks, runoff, and sediment loads from storm periods. Because essentially complete discharge data were collected in the earlier study, storm-based South Fork discharges will be regressed on comparable North Fork values to assess change. Sediment

Paper number 90WR00989.

TABLE 1. Treatment of South Fork of Caspar Creek, Near Fort Bragg, California, During the First Experiment, 1962 to 1985

Water Year*	South Fork Treatment	Evaluation Period
1963–1966	···	calibration
1967	6.8-km road built in summer	calibration
1968–1970	···	roads
1971	lower third logged in summer	roads
1972	middle third logged in summer	logging
1973	upper third logged in summer	logging
1974	···	logging
1975–1976	···	recovery 1
1977	record lost	···
1978–1981	···	recovery 2
1982–1985	···	recovery 3

*Water years are from August to July inclusive.

data from the first study could not be used to estimate sediment flux during storms. Mean concentrations in three discharge classes were used to measure changes in sediment production.

DATA AND METHODS

Discharge in both forks is measured in 120° V notch weirs that widen to large rectangular weirs for stages above 0.61 m (0.69 m^3/s). Data for water years 1963 to 1985 are of good quality and are essentially complete except for 1977 when all data were lost. The North Fork logging began in 1986, limiting comparisons between the two forks to data collected before then (Table 1).

Storms

Storms were defined by hydrograph responses rather than precipitation events, so the term "storm" indicates hydrograph response in this paper. Peaks in the North Fork hydrograph of at least 0.57 m^3/s (the approximate discharge that initiates bedload movement) were used to identify storms (Figure 1). Peaks of 0.57 m^3/s occur about seven times per year, on average. A subsequent peak of at least 0.57 m^3/s was considered a separate storm if its discharge was at least 50% higher than the preceding trough.

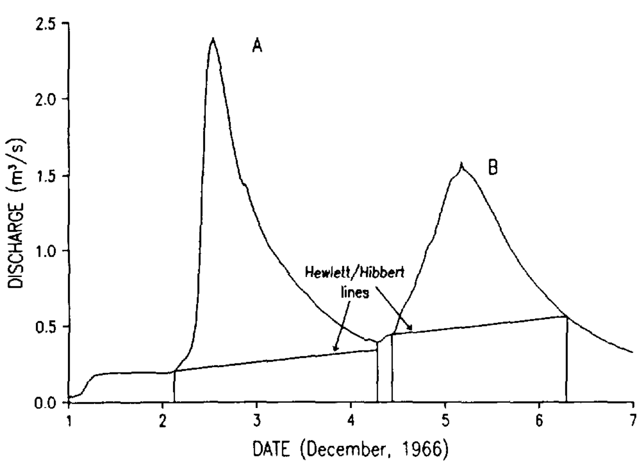

Fig. 1. Typical Caspar Creek, California, discharge hydrograph showing Hewlett/Hibbert lines used to define stormflow. Intersections of the lines with the falling limbs of hydrographs normally define ends of storms. Storms such as A with falling limbs that do not intersect the Hewlett/Hibbert lines are ended prematurely at the trough before another qualifying storm B.

Storm duration was defined and hydrographs separated using the method of *Hewlett and Hibbert* [1967]. A line with a slope of 5.5×10^{-6} m^3/s each hour was drawn through the point on the hydrograph chosen as the start of the storm. The intersection of this line with the falling limb of the hydrograph marked the end of the storm. If a qualifying subsequent storm occurred before this intersection, the storm was ended at the lowest point in the trough.

The first rise of the hydrograph before a qualifying peak was not used as the beginning of the storm. A characteristic initial rise in the hydrograph (perhaps due to channel interception) usually levels off before a steeper upturn to the peak. The beginning of the storm was taken at this second upturn. This rule nearly eliminated situations where the Hewlett/Hibbert line intersected the rising limb of the hydrograph.

The same features (i.e., recognizable hydrograph responses to the same changes in precipitation intensity) of a storm were required to be present in the hydrographs of both forks. When the storm rules did not produce this result, ad hoc decisions were made to ensure that comparable features were included for the same storm in both forks. When this process required questionable decisions, or, when hydrograph quality was poor, the storm pair was left out of the analysis. This method selected 141 storms during the 23-year period; an average of about 6.4 storms per year of record.

Quick flow, total storm flow, and peak discharge were measured for each storm. Quick flow was defined as discharge above the Hewlett/Hibbert line over the storm period. Total storm flow was calculated for the same period. The peak discharge was defined as the highest peak in the North Fork. It was matched with the corresponding feature (i.e., the peak responding to the same pulse in precipitation) in the South Fork even if the corresponding feature was not the highest peak in the South Fork.

Daily Flows

Daily flow is an alternative to storm runoff for assessing changes in the streamflow regime. I used daily flows from November through April in each water year which included just under 95% of normal annual discharge and limited the number of very low flows. Days having poor quality records in either fork were eliminated (Figure 2). With these omissions the data sets averaged just under 180 values for each year with a smallest set of 165.

Daily flows are not natural hydrologic events, which has given rise to objections to their use for comparing treatments. But defining storms is problematical as well, so comparing storm runoff characteristics is also somewhat arbitrary. Daily flows are well defined, however, and they form large linear data sets, if log-transformed. Daily flow can be a useful alternative variable under some conditions to investigate the discharge responses of forest treatments.

Sediment Data

In the first study, suspended sediment concentration was measured and temporally sampled by several methods, reflecting changing technology and attempts to improve data quality. The methods included fixed-stage, fixed-interval,

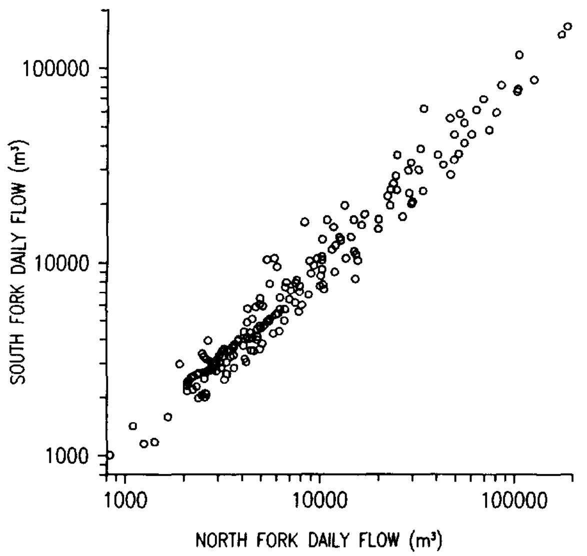

Fig. 2. Daily flows of North and South forks of Caspar Creek, California, from November 1983 through April 1984.

and discharge proportional sampling. The data were intended for use with rating curve estimators. It is now clear, however, that rating curve estimates of sediment yields are biased and depend systematically on sampling protocols [*Walling and Webb*, 1981; *Thomas*, 1988*a*]. Because each sampling method has unknown bias, there is no way to disentangle the effects of changes in sampling regime from those due to treatment.

For these reasons estimates of storm sediment yields were not used and a simpler model adopted instead. Means of (log) concentrations were estimated in "low," "medium," and "high" discharge classes. This technique is similar to the "load interval method" of calculating yields [*Verhoff et al.*, 1980], except that concentrations instead of yields are compared. Each class mean can adjust independently of the others without problems of model specification imposed by rating curves. The formal statistical tests require that data be collected randomly, which was not done with these data. It will be tacitly assumed, however, that the sediment data sets behave as random samples. The results should be interpreted with this reservation.

Test Procedures

Five characteristics were tested: quick flow, total storm flow, and peak discharge for storms; daily flows; and concentrations of suspended sediment. Tests were done among six evaluation periods defined over the 23-year period from 1963 to 1985 (Table 1). The first three periods were calibration, road building, and logging. Three recovery periods were selected to follow a possible pattern in response after treatment. Statistics were developed to relate the responses of the South Fork to the North Fork for each characteristic, and changes in the statistics were tested between selected pairs of evaluation periods.

Out of 15 pairwise comparisons possible among the six evaluation periods, the same set of eight tests was done for all five characteristics. Five sequential comparisons were made between adjacent periods; calibration and roads, roads and logging, etc. These tests were intended to determine when changes occurred during the course of the first study. Three further comparisons were made between the calibration and each recovery period to see if the system had returned to calibration levels. In all cases the undisturbed North Fork was the standard, and the disturbed South Fork was compared to it.

Distributions of the basic data are skewed, so logarithms were taken of all values. Statistical tests were performed on means (or regression parameters) of the log-transformed values, which is equivalent to comparing the medians of the untransformed data. For the regressions, the log-transformations also helped to approximate the assumptions of linearity and uniform variance.

Log-transformed South Fork storm data for each evaluation period were regressed on corresponding North Fork values to compare evaluation periods. There was one regression for each period. Differences between regression slopes and intercepts were tested to detect changes.

All evaluation periods had at least 24 storms except the first and second recovery periods which had 12 and 15, respectively. These small data sets had to be used since no other storms had acceptable records, and the results must be interpreted accordingly. This situation illustrates a general problem of analyzing storm characteristics; data for short periods may be limited due to lack of storms or incomplete records.

Logarithms of South Fork daily flows were regressed on corresponding North Fork values, one regression for each year (Figure 2). The Statistical Analysis System (SAS) program AUTOREG was used to compute regression parameters to account for serial correlation in the residuals [*SAS Institute*, 1984]. Regression slope and intercept for the evaluation periods were used to make comparisons.

Data on the concentration of suspended sediment were evaluated in three discharge classes. The low-flow class included flows from 0.44 to 1.20 m^3/s, the medium-flow class from 1.20 to 1.99 m^3/s, and the high-flow class from 1.99 to 5.40 m^3/s. The same class boundaries were used in both forks. These classes, especially the high one, span a large range of flows, but this was necessary to approximately balance the available data into three classes. The classes are adjacent and deliver about 50% of the water discharge and 80% of the suspended sediment load in the South Fork.

The sediment data for each fork were partitioned into 18 "cells" formed from all combinations of the three discharge classes and six evaluation periods, and means of the logarithms of the concentrations were taken for each cell. Each North Fork cell mean was subtracted from the corresponding South Fork mean to characterize the response of the South Fork relative to the North Fork (differences were used instead of ratios because of the logarithmic transformation). The antilogarithms of these differences are ratios of geometric means of the untransformed concentrations, or estimates of ratios of the cell medians in the South Fork to the corresponding North Fork medians.

Further differences were then taken of the cell differences to form test statistics for the five pairwise sequential tests and the three tests against the calibration period (Table 5). This statistic effectively tests the change in the ratio of the South Fork to the North Fork cell medians between pairs of evaluation periods being compared. For example, suppose X_{N1} and X_{S1} are the arithmetic means of the logarithms of concentrations in the North and South forks in the first (i.e.,

TABLE 2. Regressions on Storm Peak Discharges for the Six Treatment Periods

	Slope		Intercept			
Period	Coefficient	Standard Error	Coefficient	Standard Error	R^2, %	N
Calibration	0.967	0.0646	−0.0023	0.269	88.5	31
Roads	0.862	0.0484	0.736	0.196	90.8	34
Logging	0.954	0.0758	0.364	0.317	87.8	24
Recovery 1	0.637	0.143	1.85	0.555	66.6	12
Recovery 2	0.751	0.128	1.35	0.494	72.7	15
Recovery 3	0.796	0.0947	1.02	0.388	75.5	25

calibration) period, and X_{N6}, X_{S6} the means for the sixth (i.e., final recovery) period. Then the test statistic to compare the calibration to the final recovery period is $(X_{S6} - X_{N6}) - (X_{S1} - X_{N1})$.

The tests generally compare statistics with different variances and sample sizes. The samples were assumed to be independent, so variances of the differences were calculated by adding component variances. Approximate degrees of freedom for differences among statistics having different numbers of observations were calculated by *Satterthwaite*'s [1946] method that uses component variances and sample sizes. This procedure was carried out twice for the sediment data with differences between forks and evaluation periods.

Most multiple comparison methods set an experimentwise error rate for all possible comparisons. Any pair of means can then be tested whether planned prior to data analysis or by surveying the data. Using the data to suggest hypotheses means the tests will be less sensitive (i.e., wider intervals when fewer than all possible tests are done). Alternatively, by preselecting fewer than all possible tests, the Bonferroni procedure [*Miller*, 1980] is often more sensitive. Bonferroni was used for these data, because only eight (of the 15 possible) pairwise comparisons are of interest. Each suite of eight tests is considered an "experiment," so the stated probabilities apply in this context.

Each experiment comprising eight tests, was decided at the overall 0.05 significance level. The Bonferroni procedure requires each of the eight tests to have a two-tailed significance level of (0.05/8) = 0.00625. The probability levels in the tables and the text should be compared to 0.00625.

RESULTS AND DISCUSSION

Peak Storm Discharge

All regressions of the South Fork peak discharges on the corresponding North Fork peaks for the six evaluation periods were significant at the 0.05 level (all significance probabilities were <0.0013) (Table 2). Plots of the basic data showed reasonable agreement with the assumptions of a linear model, but there were several outliers for the first recovery period (the smallest data set) giving that regression the lowest R^2.

Inspection of the regressions suggests that changes occurred during the study (Figure 3), a pattern consistent with other studies of logging [*Harr et al.*, 1975; *Ziemer*, 1981]. The South Fork response increases relative to the North Fork during road building, logging, and the first recovery period, and then returns toward pretreatment levels. The regressions "pivot" in a region at higher peak discharges. No hydrological factor accounts for logging reducing peaks above the pivot, so this pattern probably resulted from the model used and the range of data collected. The appearance of larger differences for smaller peaks is an artifact of the logarithmic transformations.

In spite of the regression pattern, the formal tests show no significant differences for any of the 16 pairwise tests of slope and intercept (Table 3). Either no true differences exist and the apparent progression in Figure 3 is due to chance, or real differences do exist, but could not be detected because of high variance (i.e., too few storms). Further results support the latter explanation, so the "hydrological significance" of detecting changes of this magnitude should be assessed.

Increases in South Fork peaks between the calibration and third recovery periods were estimated at "low" and "high" peak levels in the North Fork. Sample percentiles were chosen from all 22 years of North Fork storm peak data below which 10% and 95% of the peaks fell. At the 10th percentile, the South Fork median peak discharge in the final recovery period was estimated to be about 63% (0.35 m^3/s) higher than the median peak during the calibration period. At

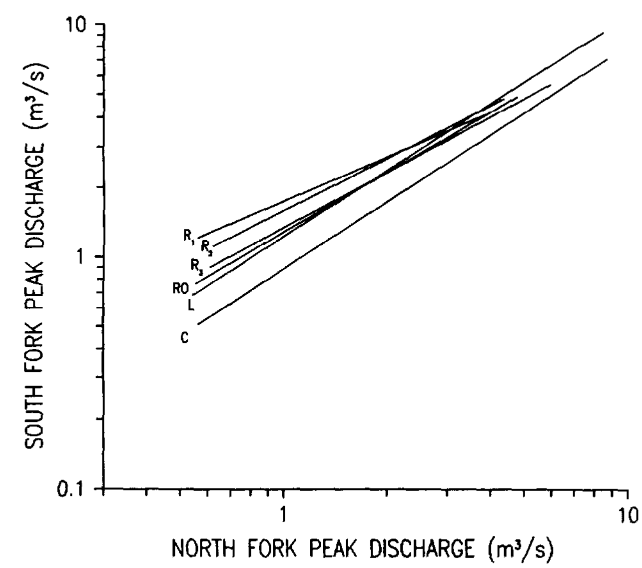

Fig. 3. Estimated regressions for peak discharges from the North and South forks of Caspar Creek for six evaluation periods: 1963–1967 (*C*, calibration), 1968–1971 (*RO*, road effects), 1972–1974 (L, logging effects), 1975–1976 (R_1, recovery), 1978–1981 (R_2, recovery), and 1982–1985 (R_3, recovery). Each line projected on the North Fork axis indicates the range of the corresponding North Fork data set.

TABLE 3. Comparisons Among Regressions on Storm Peak Discharges for the Six Treatment Periods

Comparison	Slopes: Difference	Slopes: Standard Error	Slopes: *DF**	Slopes: *SP*†	Intercepts: Difference	Intercepts: Standard Error	Intercepts: *DF**	Intercepts: *SP*†
Calibration/Roads	−0.105	0.0807	55	0.201	0.738	0.333	54	0.031
Roads/Logging	0.0923	0.0900	39	0.311	−0.372	0.372	38	0.325
Logging/R_1	−0.317	0.161	16	0.067	1.48	0.639	17	0.033
R_1/R_2	0.114	0.191	22	0.558	−0.600	0.743	22	0.428
R_2/R_3	0.0452	0.159	27	0.778	−0.331	0.593	22	0.583
Calibration/R_1	−0.329	0.157	14	0.054	1.85	0.617	15	0.009
Calibration/R_2	−0.215	0.143	20	0.148	1.35	0.563	21	0.026
Calibration/R_3	−0.170	0.115	42	0.145	1.01	0.472	43	0.037

*DF is the *Satterthwaite* [1946] degrees of freedom for testing differences.
†SP is the significance probability. No tests were significant at the 0.00625 level.

the 95th percentile, the increase was only 12%, but represented the larger absolute median increase (0.56 m^3/s). The effects of increases in peak discharge on sediment movement depend on increases in stream power and the reduction in return periods for storms of given sizes.

Ziemer [1981] found no difference in a regression test of peaks in his prelogging and postlogging periods, while *Wright* [1985] did find a significant difference in a similar test. Ziemer divided his peaks into four classes and Wright divided his into three. Although they used different classes, they both found significant increases in South Fork peaks only in the smallest classes.

Discrepant results among the present and earlier studies (even when using the "same" data) can result from several sources. These sources include differences in storm definition, the specific variables measured, the model used to test for changes, and the testing procedures employed. The outcomes may be different and still be correct; they are just answering different questions.

Quick Flow and Total Storm Discharge

Analyses parallel to that done on the peak flows were also performed on the quick flow and total storm discharge data. The regression patterns for both measures of storm discharge were similar to that for the peak data, showing initial departure from and then return toward pretreatment levels.

The quick flow data showed no significant differences in the formal tests for slopes or intercepts. The differences between the South Fork calibration and final recovery periods at the North Fork 10th and 95th percentiles were −2 and −7%, respectively. The magnitudes of these changes are probably within the limits of measurement and agree with the lack of formal significance.

The sequential test of the intercept between the logging and first recovery periods for the total storm discharge was statistically significant, and the corresponding test for slope nearly so. The test of the intercept between the calibration and first recovery periods was also significant. None of the other tests for the storm discharge data were significant, in particular, those between the calibration and final recovery periods. Changes in relative total storm discharge in the South Fork between calibration and final recovery were 19% at the North Fork 10th percentile, and −8% at the 95th percentile.

If real, the 19% increase in total storm discharge at the 10th percentile contrasts with essentially no change in quick flow. This implies that logging increased base flow, but not quick flow. *Keppeler* [1986] found an increase in summer flows and higher base flows at the start of the wet season, but these effects lasted only until 1975. Differences among studies such as those described for peak flows may account for the possibility that increased base flow has lasted until the third recovery period.

Daily Flow Data

The pattern of changes in the daily flow regressions is similar to the patterns shown by the three storm variables described in previous sections (Figures 4*a* and 5*a*). The regressions depart smoothly from calibration levels and then return toward, but not completely to those levels.

Formal tests on evaluation period average regression parameters agree with this impression (Table 4). Of the five sequential tests, two for slopes and four for intercepts are significantly different. All tests of changes in slope and intercept between calibration and the three recovery periods are significant, specifically, those between the calibration and final recovery periods, implying that daily flows have not returned to pretreatment levels. Standard errors for slopes and intercepts increase throughout the study (Figures 4*b* and 5*b*), further indicating that the response is not at calibration levels.

Several factors account for differing outcomes of the formal tests on daily flows and storm variables. Primarily, different quantities are being analyzed. Daily flows include a larger proportion of discharge (mostly low flows) than do storms, and they partition it differently. Also, the daily flow partition produces larger samples which reduces the variance, making the tests more sensitive. Although somewhat arbitrary, daily flows are well defined, and their greater sensitivity can be useful for detecting small differences when daily flows are meaningful entities for comparison.

The change in predicted South Fork daily flows at the 10th percentile of the North Fork flows is 27% and 7% at the 95th percentile. These are increases of 225 and over 4000 m^3/day, respectively, or an increase in average discharge rate of 0.00260 and 0.0463 m^3/s. This is consistent with greater increases at high discharges seen for peak flows. The increases are modest, however, suggesting that testing daily flows enables detection of small differences.

Sediment Concentration

Sediment was the most critical quantity measured, but its quality is poor. Sample sizes were small (especially at high

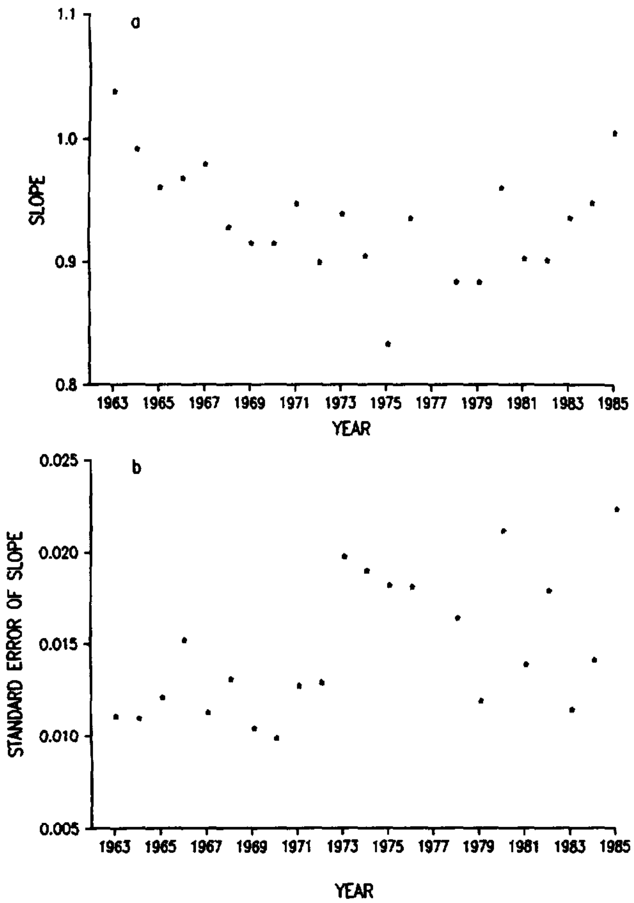

Fig. 4. (*a*) Estimated slopes and (*b*) standard errors of slopes for annual regressions of South Fork peak discharges on corresponding North Fork peaks for water years 1963 through 1985. Data for 1977 are missing.

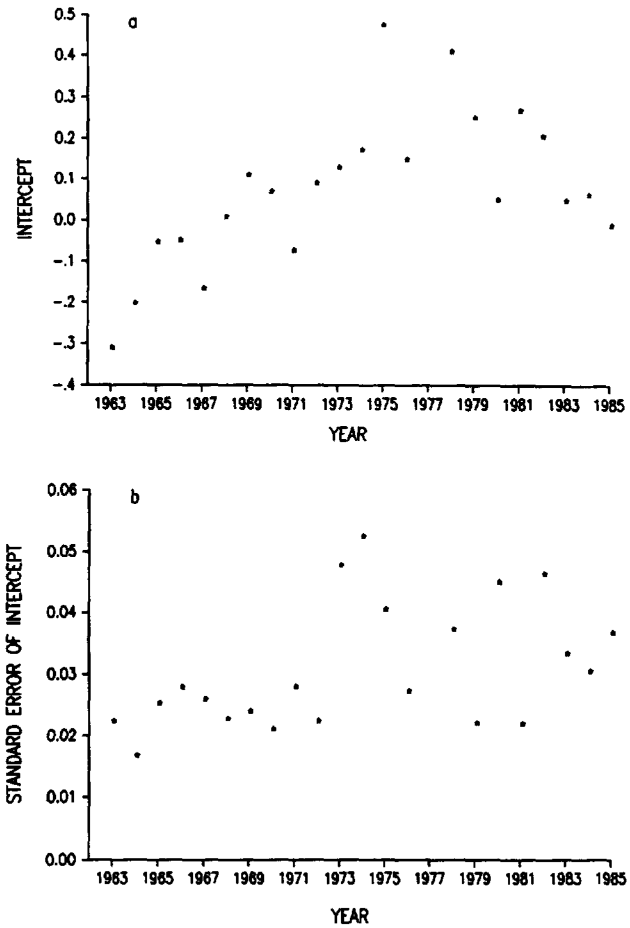

Fig. 5. (*a*) Estimated intercepts and (*b*) standard errors of intercepts for annual regressions of South Fork peak discharges on corresponding North Fork peaks for water years 1963 through 1985. Data for 1977 are missing.

flows) and data collection was nonrandom. Changes in instrumentation and sampling protocol over the course of the study may also have influenced results.

The test differences indicate changes similar to those for the streamflow data with concentrations generally rising through the first recovery period and then falling toward pretreatment levels (Table 5). The drop between the first and second recovery periods is the only change that is statistically significant for all discharge classes and implies rapid reduction in increased sediment delivery soon after cutting. There are no significant differences between the calibration and final recovery periods.

Changes in the ratios of South to North Fork median concentrations between the calibration and final recovery periods for low, medium, and high discharge classes are 17, −22, and 15%, respectively (Table 6). Increased levels of median concentration in the low and high discharge classes, although not statistically significant, may be of concern for some beneficial uses in a watershed. Reduced median concentration in the middle flow class cannot be explained hydrologically and may result from poor sampling protocol.

A plot of corresponding North and South Fork means provides another analysis (Figure 6). Movement parallel to the dashed "line of equal response" shows changes in overall climatic stress in the basins, while perpendicular movement indicates relative changes between basins. The mean pairs for each class follow approximate counterclockwise loops, with the points for the third recovery periods tending to return to the calibration points. This pattern implies relatively larger changes in the South Fork early in the study with a return to original conditions toward the end of the study.

SUMMARY

All variables showed similar patterns of change during the study. The peak, quick flow, and total discharge from storms, daily flows, and suspended sediment concentrations in three flow classes generally increased during the early stages of the study. The variables reached maxima either during logging or during the following two years and then returned toward pretreatment conditions. For all three storm variables, however, there were only two statistically significant differences out of 48 tests, and both related to total storm discharge. Neither significant test was between the calibration and last recovery periods.

The apparent conflict between the pattern and the formal tests was investigated by calculating changes in the South Fork response between the calibration and last recovery periods for the 10th and 95th percentiles in the North Fork. Percent changes were generally modest except for a substan-

TABLE 4. Comparisons Among Regressions of Daily Flows for the Six Treatment Periods

	Slopes				Intercepts			
Comparison	Statistics*	Standard Error†	*DF*‡	*SP*§	Statistics*	Standard Error†	*DF*‡	*SP*§
Calibration/Roads	−0.0613	0.00797	1463	0.000	0.184	0.0161	1439	0.000
Roads/Logging	−0.0129	0.0116	794	0.267	0.101	0.0276	615	0.000
Logging/R_1	−0.0294	0.0163	704	0.072	0.182	0.0348	712	0.000
R_1/R_2	0.0234	0.0151	613	0.124	−0.0675	0.0295	585	0.022
R_2/R_3	0.0399	0.0117	1176	0.001	−0.168	0.0249	1174	0.000
Calibration/R_1	−0.104	0.0139	471	0.000	0.467	0.0267	433	0.000
Calibration/R_2	−0.0803	0.00976	1107	0.000	0.399	0.0197	960	0.000
Calibration/R_3	−0.0404	0.0101	1038	0.000	0.231	0.0215	1051	0.000

*Statistic is the difference between mean annual slopes or intercepts in the indicated evaluation periods.
†Standard error is the standard error of the statistic.
‡*DF* is the *Satterthwaite* [1946] degrees of freedom for the test.
§*SP* is the significance probability which should be compared to 0.00625. An *SP* of 0.000 indicates significance probability <0.0005.

tial increase in the small South Fork peaks in the 10th percentile. Although this was not significant, it suggests agreement with *Wright* [1985] and *Ziemer* [1981] if the lack of significance is due to small sample size. Although the change in peaks at the 95th percentile was a smaller percentage, it did represent a larger absolute increase in discharge, which may be responsible for more geomorphic work being done. This conclusion differs from the earlier studies.

In contrast to the storm flow parameters, 12 out of the 16 tests done on the daily flows were statistically significant, including the tests between the calibration and each of the three recovery periods. The daily flows evidently partition and utilize available discharge information in ways that allow detection of smaller differences than do variables based on storms.

Logging and road building probably altered discharge response somewhat as indicated by similar patterns of change for all variables and tests of the daily flow data. Storm characteristics, however, are apparently not sensitive to differences of this size with the number of storms available in this study. Because the second Caspar Creek study is based on storm variables, the lack of statistical significance between the calibration and last recovery periods is adequate evidence that the system has returned to pretreatment conditions.

Results for the suspended sediment data were more ambiguous. Eight out of the 24 tests done on median concentrations in the three discharge classes showed statistically significant differences. The only consistent pattern was significant negative differences in all three classes (Table 5) between the first and second recovery periods, showing rapid reduction in median concentration soon after treatment. No test between the calibration and last recovery periods for any of the three discharge classes was significant.

The apparent confusion is reduced if all tests are considered together. The pattern of change is widespread and points toward recovery, and formal tests of differences for the critical sediment and peak variables between the calibration and third recovery periods are not significant. In general, the measures of difference are large for small storms, and small for larger ones, suggesting only changes in evapotranspiration. A possible exception is the peak discharge data in which an important increase for large storms is

TABLE 5. Test Statistics for Comparing Suspended Sediment Concentration in Three Discharge Classes Among Six Treatment Periods

	0.44–1.20 m^3/s				1.20–1.99 m^3/s				1.99–5.40 m^3/s			
Comparison	Test Difference*	Standard Error†	*DF*‡	*SP*§	Test Difference*	Standard Error†	*DF*‡	*SP*§	Test Difference*	Standard Error†	*DF*‡	*SP*§
Calibration/Roads	0.833	0.205	191	0.000	0.178	0.260	80	0.497	0.260	0.232	37	0.510
Roads/Logging	0.254	0.215	188	0.239	0.110	0.220	200	0.618	0.181	0.190	118	0.343
Logging/R_1	−0.070	0.203	153	0.730	0.212	0.163	214	0.195	0.489	0.150	163	0.001
R_1/R_2	−0.806	0.172	152	0.000	−0.491	0.134	312	0.000	−1.29	0.131	134	0.000
R_2/R_3	−0.053	0.168	573	0.752	−0.260	0.139	428	0.062	0.501	0.143	189	0.001
Calibration/*R*1	1.02	0.191	152	0.000	0.500	0.214	42	0.024	0.930	0.201	23	0.000
Calibration/R_2	0.211	0.157	171	0.182	0.009	0.208	38	0.966	−0.360	0.212	28	0.100
Calibration/R_3	0.160	0.188	278	0.401	−0.251	0.217	45	0.256	0.141	0.211	28	0.509

*Test difference is the difference between the respective treatment period differences between the North and South forks.
†Standard error is the standard error for the test difference.
‡*DF* is the *Satterthwaite* [1946] degrees of freedom for the test.
§*SP* is the significance probability which should be compared to 0.00625. An *SP* of 0.000 indicates a significance probability of <0.0005.

TABLE 6. Estimates of Median Suspended Sediment Concentrations, Standard Deviations, and Sample Sizes for All Evaluation Periods in the North and South Forks of Caspar Creek

	0.44–1.20 m^3/s			1.20–1.99 m^3/s			1.99–5.40 m^3/s		
Evaluation Period	Median	Standard Deviation	*n*	Median	Standard Deviation	*n*	Median	Standard Deviation	*n*
				North Fork					
Calibration	59.4	6.36	44	143.	25.7	20	867.	137.	10
Roads	66.7	6.20	55	186.	24.1	54	493.	35.3	22
Logging	84.5	10.7	48	300.	31.5	60	924.	99.8	53
Recovery 1	81.2	4.50	73	228.	14.1	61	373.	25.8	67
Recovery 2	22.8	1.67	161	83.3	5.20	92	547.	39.0	66
Recovery 3	39.5	4.30	178	135.	11.1	114	500.	29.2	164
				South Fork					
Calibration	70.8	4.84	59	193.	13.6	41	701.	72.1	16
Roads	183.	24.4	68	300.	38.2	62	517.	64.9	34
Logging	299.	20.6	70	540.	39.9	54	1160	76.2	94
Recovery 1	267.	36.0	59	508.	41.0	84	764	33.7	139
Recovery 2	33.5	1.99	213	113.	6.94	120	309.	23.1	70
Recovery 3	55.1	4.86	156	142.	10.1	132	465.	37.8	142

Medians and standard deviations are in milligrams per liter.

indicated if the lack of statistical significance is due to large variance.

If a previously treated control is still recovering, the effect will be to overemphasize the measured effect (Figure 7). However, the return to calibration should slow over time, making any adjustments smaller than those measured here. Large effects should not be influenced by small departures of the control watershed from undisturbed conditions, but care should be exercised when interpreting small differences.

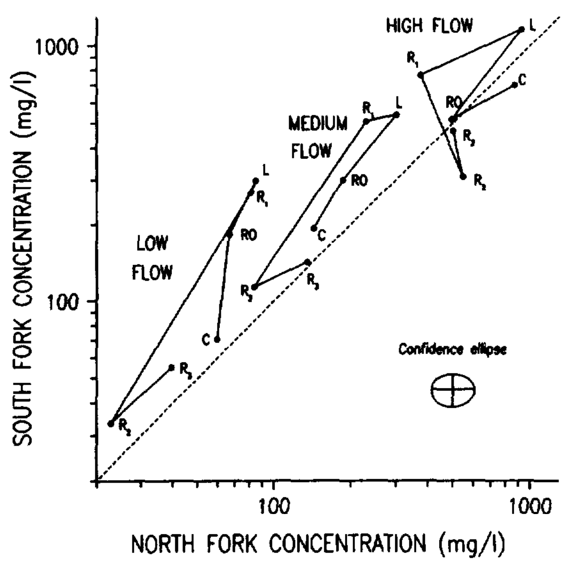

Fig. 6. Corresponding pairs of North and South Fork mean concentrations of suspended sediment for three flow classes within the six treatment periods: 1963–1967 (*C*, calibration), 1968–1971 (*RO*, road effects), 19072–1974 (*L*, logging effects), 1976–1976 (R_1, recovery), 1978–1981 (R_2, recovery), and 1982–1985 (R_3, recovery). The low-flow class includes discharges from 0.44 to 1.20 m^3/s, the medium-flow class from 1.20 to 1.99 m^3/s, and the high-flow class from 1.99 to 5.4 m^3/s. The dashed line shows equal response in both forks. The translated confidence ellipse, when centered on any pair of means, indicates two standard deviations in any direction.

The decision to proceed with the next experiment in Caspar Creek seems reasonable because of the asymmetry between the statistical significance and the hydrological importance of the variables tested. All tests indicate a return to pretreatment conditions, so the expense of delaying experimental logging on the North Fork is deemed unjustified.

Several methodological problems arising in the study are worth reviewing. The outcomes of formal tests depend on factors other than real differences in quantities being tested. Definition of variables, selection of test procedures, and sample size can all affect the results of a particular study. Variables should be selected that measure real quantities of interest and adequate size samples collected to detect important hydrological differences. Storm data may not detect small differences due to lack of sensitivity stemming from limited sample sizes. Other variables such as daily flows may be preferred in some cases when more subtle effects are studied.

The quality and quantity of sediment data need improvement. Traditional methods of measurement and particularly sampling and estimation have been shown to give biased results. Interest in the effects of forest treatments on sediment loads in streams justifies allotting a larger portion of study resources to that effort.

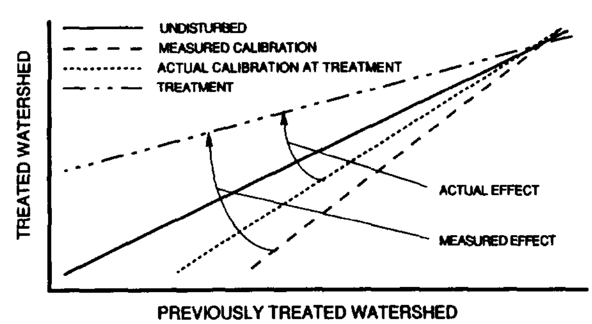

Fig. 7. Typical relationships for a study using a previously treated control watershed that is still returning to undisturbed conditions.

Acknowledgments. Data collection for both Caspar Creek studies has been carried out in cooperation with the California Department of Forestry and Fire Protection in the Jackson State Forest, Fort Bragg, California.

REFERENCES

Harr, R. D., W. C. Harper, J. T. Krygier, and F. S. Hsieh, Changes in storm hydrographs after road building and clear-cutting in the Oregon Coast Range, *Water Resour. Res.*, *11*(3), 436–444, 1975.

Hewlett, J. D., and A. R. Hibbert, Factors affecting the response of small watersheds to precipitation in humid areas, in *Forest Hydrology*, edited by W. E. Sopper and H. W. Lull, pp. 275–290, Pergamon, New York, 1967.

Keppeler, E. T., The effects of selective logging on low flows and water yield in a coastal stream in northern California, M.A. thesis, 137 pp., Humboldt State Univ., Arcata, Calif., 1986.

Miller, R. G., *Simultaneous Statistical Inference*, 2nd ed., 299 pp., Springer-Verlag, New York, 1980.

Rice, R. M., F. B. Tilley, and P. A. Datzman, A watershed's response to logging and roads: South Fork of Caspar Creek, California, 1967–1976, *Res. Pap. PSW-146*, 12 pp., Pac. Southwest Forest and Range Exp. Sta., Berkeley, Calif., 1979.

SAS Institute, *SAS/ETS Users Guide*, version 5 ed., 738 pp., Cary, N. C., 1984.

Satterthwaite, F. E., An approximate distribution of estimates of variance components, *Biometrics*, *2*, 110–114, 1946.

Sendek, K. H., Effects of timber harvesting on the lag time of Caspar Creek watershed, M.A. thesis, 46 pp., Humboldt State Univ., Arcata, Calif., 1985.

Thomas, R. B., Monitoring baseline suspended sediment in forested basins: The effects of sampling on suspended sediment rating curves, *Hydrol. Sci. J.*, *33*(5), 499–514, 1988*a*.

Thomas, R. B., Measuring the sediment yields of storms using PSALT, Proceedings of the Porto Alegre Symposium on Sediment Budgets, *IAHS Publ.*, *174*, 101–109, 1988*b*.

Verhoff, F. H., S. M. Yaksich, and D. A. Melfi, River nutrient and chemical transport estimation, *J. Environ. Eng. Div. ASCE*, *10*(6), 591–608, 1980.

Walling, D. E., and B. W. Webb, The reliability of suspended sediment load data, Proceedings of the Florence Symposium, June 1981, *IAHS Publ.*, *122*, 34–48, 1981.

Wright, K. A., Changes in storm hydrographs after roadbuilding and selective logging on a coastal watershed in northern California, M.A. thesis, 55 pp., Humboldt State Univ., Arcata, Calif., 1985.

Ziemer, R. R., Storm flow response to road building and partial cutting in small streams of northern California, *Water Resour. Res.*, *17*(4), 907–917, 1981.

R. B. Thomas, Pacific Southwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture, 1700 Bayview Street, Arcata, CA 95521.

(Received January 24, 1989;
revised October 24, 1989;
accepted April 26, 1990.)