Comment on "Forest and floods: A new paradigm sheds light on age-old controversies" by Younes Alila et al.

Jack Lewis,[1,2] Leslie M. Reid,[1] and Robert B. Thomas[1,2]

Received 12 October 2009; revised 26 February 2010; accepted 16 March 2010; published 22 May 2010.

Citation: Lewis, J., L. M. Reid, and R. B. Thomas (2010), Comment on "Forest and floods: A new paradigm sheds light on age-old controversies" by Younes Alila et al., *Water Resour. Res.*, *46*, W05801, doi:10.1029/2009WR008766.

1. Introduction

[1] The paper by *Alila et al.* [2009, hereafter referred to as AKSH] presents a technique for analyzing altered peak flow frequencies after logging. The paper suggests that the established method of chronologically pairing peak flows by corresponding hydrologic input at control and treated watersheds is inappropriate, leading to irrelevant research hypotheses and impeding scientific progress. In general, we agree that analyses of changes in flood frequency are useful for evaluating the effects of watershed disturbance, and that simple regression models often provide inadequate descriptions of posttreatment peak flow responses. However, the proposed method and accompanying discussion have several problems that undercut the strength of the paper's conclusions:

[2] 1. The recovery adjustment used by the method augments the effect the analysis is attempting to detect.

[3] 2. Even in the absence of impacts, the frequency distribution of observed peaks is expected to have greater variance than that of predicted peaks, thereby introducing an artificial shift when comparing upper quantiles of observed and expected frequency distributions.

[4] 3. A more appropriate analysis of uncertainty is needed if the utility of the method is to be validly assessed.

[5] 4. Frequency pairing does not overcome the problem of low power in testing for changes in very large events prior to forest regrowth.

[6] 5. The relative merits of alternative statistical approaches are mischaracterized.

2. Use of the Recovery Adjustment

[7] AKSH compare analyses based on frequency pairing (FP) and chronological pairing (CP) of data for a 48 year record from the Fool Creek watershed. However, the comparison is inequitable because different data sets were used in the FP and CP analyses. Comparison of Figures 3a and 3b illustrates the inconsistency: for example, in the FP analysis (Figure 3b) the two smallest "observed" posttreatment peaks of the CP analysis (Figure 3a) are plotted at about twice their original magnitudes. AKSH adjusted the data used in the FP analysis for recovery to avoid the problem of lack of stationarity introduced by the changing distribution of peak flows as the forest regrows. They used the CP data set to calculate an average recovery rate for all peaks, then augmented all postlogging peak flows in the FP data set, with the amount of shift increasing with time after logging; the largest storm was spared significant adjustment because it occurred immediately after logging. The conclusion that only the frequency paired approach revealed that "all peak flows save the largest event were shifted upward" (AKSH, paragraph 29) reflects the fact that the AKSH procedure itself shifted the peaks used in the FP analysis upward. Figure 7b, showing the unadjusted analysis, is the appropriate figure for comparison to Figure 3a; both reveal a more modest upward shift converging at the two largest events.

[8] For any given time after logging, peaks of all sizes at Fool Creek were augmented by adding a constant discharge, and those at H.J. Andrews were multiplied by a constant; the adjustments do not take into account results of the studies cited by AKSH suggesting that large peak flows respond differently to logging than small peak flows. Augmenting peak flows without regard to possible differences in response as a function of peak flow magnitude is unwarranted in an analysis intended to characterize changes in large peak flows. Because recovery was assumed to apply to peaks of all sizes, regardless of whether all sizes were actually affected, adjusting for assumed recovery created or inflated the very effect the study sought to measure.

[9] Because of the problem of nonstationarity, frequencies of altered peak flows have generally been characterized by reference to the recurrence interval of corresponding events in a control watershed, or to that of expected events in the treated watershed (predicted by pretreatment calibration with a control). For example, change can be described by statements such as "under the pretreatment peak flow distribution, the recurrence interval of the observed peak flow was 2 years while that of the expected peak was only 1 year." This effectively conveys the idea that large flows are more frequent under the altered regime, even though the continually evolving posttreatment frequency distribution is difficult to define for any specific point in time.

3. Comparing Observed and Expected Frequency Distributions

[10] The AKSH method assumes that the frequency distribution of observed peaks in an unimpacted watershed is identical to that of the peaks predicted through calibration with a control watershed. To see that the assumption is

[1]Pacific Southwest Research Station, U.S. Forest Service, Arcata, California, USA.
[2]Retired.

0043-1397/10/2009WR008766

incorrect, consider the analysis of variance table associated with the pretreatment calibration regression. The sum of squares, $\sum(y_i - \bar{y})^2$, in peak flows from the watershed that is to be treated can be partitioned as the sum of variation due to the model, $\sum(\hat{y}_i - \bar{y})^2$, and that due to residuals, $\sum(y_i - \hat{y}_i)^2$. The variance of the observed responses is thus always greater than that of the predicted responses, while the means of observed and predicted values in regression are always identical. Consequently, the upper quantiles of the observed distribution will always be greater than those of the predicted distribution (as AKSH Figure 1b illustrates for preharvest and postharvest observations). The exact partitioning of sums of squares depends on the fact that regression residuals always sum to zero. If there are no effects on peak flows, the postlogging regression should be very similar to the prelogging regression, so the same partitioning of sums of squares should hold approximately true for postlogging peak flows predicted by the prelogging regression. Total variation will be greater than that explained by the regression line, so the variation of observed values will be greater than that of predicted values. This difference in variation needs to be accounted for before the effects of logging can be evaluated.

[11] AKSH (paragraph 36) may have been addressing this issue when they referred to "loss of variability associated with the use of the pretreatment calibration equation." A correction for "loss of variability" was applied to expected peaks before comparing distributions, but because the correction method was not described, the reader cannot evaluate the implications or appropriateness of the method or ascertain what issue it was designed to address. The expected difference in variation between observed and predicted values therefore remains an unresolved issue in the analysis and may contribute to differences between estimated upper quantiles of the observed and predicted distributions.

[12] Regression models with log-transformed responses predict the mean of the logarithm of y for a given x. Where such models are appropriate, the distribution of y for a given x is lognormal or at least has positive skew, so the antilog of the mean is less than the mean of the antilogs. Thus the antilog of the prediction underestimates the mean of y given x, and a bias correction is needed [e.g., *Cohn et al.*, 1989]. AKSH apparently did not correct for bias, so the predicted values for the WS1 and WS3 analyses may be biased downward. In the absence of a treatment effect, the mean of the distribution of observed peaks should therefore be greater than that of the predicted peaks. As a result, both the mean and variance of the observed peaks distribution are expected to be greater than those of the predicted distribution for WS1 and WS3, even if no logging effect exists. The required bias correction is typically small, but given the sensitivity of upper quantiles to a shift in both mean and variance, the differences reported cannot necessarily be attributed to logging.

4. Assessment of Uncertainty

[13] Objective procedures are needed to assess the information content of experimental data; that is, to distinguish signal from random noise, and AKSH used a Monte Carlo simulation to calculate confidence intervals for that purpose. However, the simulation did not account for stochasticity of the observed postlogging peaks and the resulting uncertainty of the estimated quantiles. Both sets of quantiles are statistics that vary as a result of sampling, so a valid analysis of uncertainty would require that confidence limits be shown for frequency distributions of both the expected and observed peak flows. Properly calculated confidence intervals for both sets of quantiles would reveal the extent of overlap in the distributions, allowing assessment of the proposed method's ability to distinguish the two distributions. Confidence limits for quantiles can be estimated using either the binomial distribution [*Conover*, 1999] or bootstrapping [*Efron and Tibshirani*, 1993].

[14] The AKSH Monte Carlo simulation employed unnecessary and questionable parametric assumptions (normally distributed regression error with constant variance, GP and GEV frequency distributions for expected peaks, and normality of simulated values). The most widely accepted and well understood simulation technique for estimating uncertainty in complex analyses is bootstrapping. A bootstrap would resample the joint (CP) distributions of both pretreatment and posttreatment flows. Each bootstrap iteration would include computation of the pretreatment regression, prediction of expected posttreatment peak flows, and estimation of quantiles for both the observed and expected peaks, followed by calculation of any statistics for which confidence intervals are sought.

[15] Each bootstrap iteration precisely mimics the analysis process, thus ensuring that uncertainty is properly characterized. In contrast, the AKSH simulation process seems unrelated to the actual analysis. The reasoning behind the procedure of iteratively resampling and reestimating parameters of an assumed frequency distribution is unclear. It would seem that the simulated distributions could diverge arbitrarily from the observed distribution in 10,000 iterations. Finally, uncertainty estimates due to prediction and quantile estimation were simulated separately and the two variances summed to create confidence limits for estimated quantiles, as though quantiles were sums of independent random variables.

[16] AKSH argue that undue reliance on significance testing may lead researchers to overlook subtle effects, and it may be for this reason that confidence intervals are mentioned only once in the discussion. Notwithstanding issues regarding the utility of hypothesis testing or the method used to calculate confidence intervals, it is instructive to note that, in every case illustrated, the observed frequency distributions drop inside the calculated confidence bands for events with long return periods, suggesting that the FP method has produced results that are rather consistent with those from past CP-based analyses.

5. Detecting Changes in Very Large Events

[17] AKSH suggest that comparison of upper quantiles of frequency distributions is the correct approach for assessing effects of logging on large peak flows and that CP analysis is ill suited for such assessments. However, statistical detection of changes in upper quantiles is no easier than detection of changes in the magnitude of chronologically matched large floods of similar frequency, even if distributions are stationary. Evaluation of the strength of FP-based conclusions regarding large peak flows would require a statistical test comparing the tails of expected and observed frequency distributions. The quantile test [*Johnson*

et al., 1987] is a simple nonparametric two-sample rank test that can be calculated using the reported sample sizes and observed rankings from Figures 2, 5, 6, and 7. The quantile test fails to show any differences in the upper quantiles (even at a relaxed significance level, $\alpha = 0.10$) for any of the watersheds examined, without adding a recovery adjustment to the postlogging peaks. Even with the recovery adjustment, changes in the upper quantiles are not significant ($\alpha = 0.05$) for return periods greater than 3 years at Fool Creek (27 year data set) and 1 year (or less) for the three other data sets.

[18] Lack of statistical power will always be a hindrance to detecting changes in unusual events. A prudent course of action when faced with nonsignificant results is (1) to note the apparent direction of change, regardless of statistical significance, (2) to determine whether the power of the test was sufficient to detect a change of minimal importance, and (3) to conduct metastudies to investigate whether analogous changes have repeatedly been measured but declared insignificant in the absence of sufficient statistical power.

[19] The paper cautions that, because of uncertainties around the upper tails of distributions, the reader should avoid concluding that the largest floods did not increase. However, if the purpose of the research is to decide whether experimental results contain enough information to cast doubt on conventional scientific wisdom, the null hypothesis must be that the largest floods are unchanged by logging. If the available data are uninformative, the reader should avoid conclusions of any kind. In any case, statistical power analysis can be used to ascertain the minimum detectable change afforded by the data, allowing evaluation of whether the available data are capable of disclosing changes of a magnitude considered to be operationally meaningful.

[20] AKSH (paragraph 36) point out the potential "errors in predicting large events, particularly if they extend beyond the range of the pretreatment calibration data," and warn that "the convergence (or lack thereof) of the observed and expected cdfs [cumulative distribution functions] may be an artifact of the pretreatment calibration model." They note that some of the largest events at WS3 required extrapolation of the pretreatment calibration regression, which was not shown. Despite these warnings, the divergence of the CDFs at the upper extreme is interpreted as evidence that "the biggest floods…can even be more affected than the small and medium floods" (paragraph 50). Consideration of confidence intervals would be particularly useful in determining whether conclusions such as this are supported by the available data.

6. Relative Merits of CP, FP, and Associated Statistical Methodologies

[21] AKSH blame CP for "stifling the progress of science" (paragraph 67), "cloud[ing] our view of the more general relation between forest land use and the biophysical environment" (paragraph 66), creating "persistent disagreement between forest hydrologists" (paragraph 66), and misdirecting the scientific community by "irrelevant hydrological research hypotheses, flawed statistical methods, and their misleading outcomes for over 50 years" (paragraph 72). Yet CP is simply a form of blocking in experimental designs. Blocking is used in nearly every field of science for reducing the errors contributed by extraneous factors. In hydrology, chronologically pairing storm events from similar watersheds reduces confounding influences from storm event intensity and duration, vegetation, topography, and subsurface drainage patterns. It is a critical element of repeated measures, paired BACI (before-after control-impact) designs, the strategy most commonly used in well-planned paired watershed studies. CP is most effective in small watershed studies where spatial variability in climatic inputs is limited. The AKSH analysis itself depends heavily upon CP to determine expected peak flows in the treated watersheds (from the pretreatment calibration) and to adjust peaks for recovery. In addition, while not matching individual peaks, the comparison of expected and observed frequency distributions benefits from the fact that the observed postlogging peaks were generated from the same set of hydrologic inputs as the posttreatment control peaks that were used to estimate the expected peaks.

[22] In any case, while a primary objective of the paper was to show the advantages of FP over CP, a direct comparison was made for only one of the four data sets presented (Fool Creek 48 year data). It would have been instructive to carry out similar comparisons for data sets reflecting the shorter record lengths more typical of those generally available, such as those from WS1 and WS3, and for the 27 year Fool Creek data set.

[23] AKSH state (paragraph 59) that "frequency-paired analysis exhibits a stronger relation with less scatter than the chronologically paired analysis,…which makes the former statistically more powerful in detecting peak flow changes." CDFs are smoother than related regression analyses only because the data are sorted to create a nondecreasing display. The uncertainty of a q quantile displayed by an empirical CDF is a function of only the sample size, the value of q, and the ordered sample values [*Conover*, 1999]. The smoothness of CDFs in no way implies low uncertainty in the tails and implies nothing about statistical power relative to CP analyses. Statistical power cannot be compared without reference to specific tests and effect sizes. Hypothesis tests for CP and FP have entirely different null hypotheses, so direct comparisons of statistical power may not be possible. AKSH go on to state (paragraph 59) that "chronological pairing introduces an artificial level of effects variability." The opposite statement is perhaps more compelling: frequency pairing artificially reduces the variability by forcing a monotonic relationship on both data sets so that adjacent points are no longer independent.

[24] While FP has its appropriate uses, the same is true of CP. For example, CP permits an analysis of recovery by looking at the trend of residuals with time; this, in fact, was the analysis used by AKSH to adjust for recovery, as indicated by their reference in section 3.5 to *Thomas and Megahan* [1998]. There are no residuals in frequency pairing, and differences between expected and observed values for the same frequency are not associated with a particular time, so FP cannot be easily used to evaluate recovery.

[25] CP also provides an advantage over FP in the evaluation of individual events, as illustrated by consideration of the 1964 flood at WS3. A 90% confidence interval for the largest flood in a 25 year record is 9 to 487 years, so the actual position of the 1964 flood on Figure 5b is highly uncertain. On the other hand, a regression plot analogous to Figure 3a would reveal that this largest peak on record was

nearly 4 times higher than expected based on the chronologically matched peak in the control. Since it was only a 1 year peak in the control (dates were shown in a prepublication draft of Figure 5b), the prediction was made near the mean of the pretreatment data and should be reliable.

[26] AKSH state that CP conceals changes in frequency, but observing that many of the largest postlogging peaks were expected to be midsized peaks on the basis of the pretreatment regression (Figure 3a) does reveal that the frequency of large peaks increased after logging. Additional calculations could be used to quantify those changes. Frequencies and the conversion of medium peaks to large peaks may indeed deserve more attention, but that is no reason to abandon methods utilizing CP.

[27] More sophisticated models based on CP data can be used directly to compare frequency distributions of undisturbed and postlogging peak flows at any time after logging. A nonlinear regression model employing the response in a control watershed, proportion of area logged, antecedent wetness, and time since logging explained 95% of the variation in logarithms of peak flows at 10 treated watersheds in the Caspar Creek Experimental Watershed [*Lewis et al.*, 2001; *Lewis* 2006]. Such a model could be used to analyze changes in flood frequency by comparing predictions for the uncut condition in a particular watershed to those for the same set of events at a hypothetical time after logging. Unlike the AKSH procedure, predictions need not be restricted to postlogging events at the control watershed. The entire set of responses from the control can be used to predict the frequency distribution of any prelogging or postlogging condition that the model can represent through its parameters. The method depends on having a model, such as the one cited above, for which residual variance for postlogging data is no greater than that for prelogging data. Otherwise, the logging-induced variability will not be fully represented in the predictions. Predicted frequency distributions will exhibit reduced variability relative to distributions of observed responses for reasons discussed earlier. But this method, which employs both CP and FP, compares distributions on equal terms, while utilizing the entire period of record and accounting for nonstationary responses.

[28] AKSH state that inferences about treatment effects in ANOVA and ANCOVA (used loosely to include regression methods) are invalid for flood events smaller and larger than an average peak flow. Although no statistical justification is given for the statement, support is drawn from a reference to *Harris* [1977, p. 7], who stated without explanation that "the mean of the individual postlogging values provides the only valid comparison with the prediction limit." In actuality, prediction limits are a standard tool in statistics, and if the regression assumptions are met, valid inferences can be made throughout the range of the regression data [e.g., *Miller*, 1981]; inferences away from the mean are simply subject to greater variance. Prediction limits define the range within which individual future observations from the same sampling population will fall with a specified degree of certainty, and inference for multiple future observations can be made using simultaneous inference methods. So what is the source of the notion that inferences are limited to the mean? Perhaps it is based on the concern that CP often "reaffirms the perception of a rapidly vanishing treatment effect with event size" (AKSH, paragraph 29). But the FP analysis of Figure 7b equally reaffirms that perception with regard to the Fool Creek data. An unexpected conclusion only implies a flawed method if the conclusion is incorrect.

7. Conclusion

[29] The peak flow controversy has persisted primarily because reliably characterizing temporary impacts on relatively rare events is a very difficult problem. While attention to flood frequencies is merited and may shed light on the issue, a consideration of problems with the AKSH analysis suggests that methods based on FP are unlikely to resolve the controversy, particularly if other useful approaches are abandoned. Like all statistical techniques, ANOVA and ANCOVA based upon CP have their limitations and may be misused, but there is nothing inherently "inappropriate" about these techniques. Even the analysis set forth by AKSH required the use of CP-based regression. Methods based on FP and CP should be complementary tools that together produce a more complete understanding of experimental results than either class of method used alone.

References

Alila, Y., P. K. Kuras, M. Schnorbus, and R. Hudson (2009), Forests and floods: A new paradigm sheds light on age-old controversies, *Water Resour. Res.*, *45*, W08416, doi:10.1029/2008WR007207.

Cohn, T. A., L. L. DeLong, E. J. Gilroy, R. M. Hirsch, and D. K. Wells (1989), Estimating constituent loads, *Water Resour. Res.*, *25*(5), 937–942, doi:10.1029/WR025i005p00937.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, 3rd ed., chap. 3, pp. 143–144, John Wiley, New York.

Efron, B., and R. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman and Hall, Boca Raton, Fla.

Harris, D. D. (1977), Hydrologic changes after logging in two small Oregon coastal watersheds, *U.S. Geol. Surv. Water Supply Pap.*, *2037*, 31 pp.

Johnson, R. A., S. Verrill, and D. H. Moore (1987), Two-sample rank tests for detecting changes that occur in a small proportion of the treated population, *Biometrics*, *43*, 641–655, doi:10.2307/2532001.

Lewis, J. (2006), Fixed and mixed-effects models for multi-watershed experiments, paper presented at 3rd Federal Interagency Hydrologic Modeling Conference, Subcomm. on Hydrol., Reno, Nev., 2–6 April. (Available at http://acwi.gov/hydrology/mtsconfwkshops/conf_proceedings/3rdFIHMC/third_fihmc_nevada-2006.pdf)

Lewis, J., S. R. Mori, E. T. Keppeler, and R. R. Ziemer (2001), Impacts of logging on storm peak flows, flow volumes and suspended sediment loads in Caspar Creek, California, in *Land Use and Watersheds: Human Influence on Hydrology and Geomorphology in Urban and Forest Areas*, *Water Sci. Appl.*, vol. 2, edited by M. S. Wigmosta and S. J. Burges, pp. 85–125, AGU, Washington, D. C.

Miller, R. G. (1981), Regression techniques, in *Simultaneous Statistical Inference*, 2nd ed., chap. 3, pp. 114–116, Springer, New York.

Thomas, R. B., and W. F. Megahan (1998), Peak flow responses to clear-cutting and roads in small and large basins, western Cascades, Oregon: A second opinion, *Water Resour. Res.*, *34*(12), 3393–3403, doi:10.1029/98WR02500

J. Lewis, 647 Elizabeth Dr., Arcata, CA 95521, USA. (jacklewis@suddenlink.net)

R. B. Thomas, 69254 Whippletree, Sisters, OR 97759, USA.

L. M. Reid, Pacific Southwest Research Station, U.S. Forest Service, 1700 Bayview Dr., Arcata, CA 95521, USA.