Responses to Comment #47 from [redacted]

There are three major contentions in this comment and responses are provided below.

1. The MGB substitute for the standard Grubbs Beck test violates government policy.

Response:

The standard Grubbs-Beck (GB) test was developed in 1972 and represents a reasonable approach for identifying one or more low outliers in a set of normally distributed data. It was adopted in Bulletin 17B as an outlier detection approach and is an ASTM standard because it was the best available science at the time. The recent work of Cohn et al. (2013) extended or generalized the Grubbs-Beck for detecting potentially influential low floods (PILFs) in a flood series. The new test is more applicable for flood frequency analyses and represents an improvement in scientific methods for detecting outlier or PILFs.

2. The "PILF" concept at the heart of MGB is fatally flawed.

Response:

The major concern about the MGB test is that it sometimes identifies several low floods that are not really outliers in the context of Bulletin 17B (i.e., data points that depart from the trend of the rest of the data). PILFs are potentially influential low floods that can have a significant effect on estimating the high discharge end of the frequency curve. In effect, inclusion of PILFs in the flood frequency analysis can cause the frequency curve to depart from the high magnitude flood data at the high end of the curve. The smallest observations in the data set do not convey meaningful or valid information about the magnitude of major floods that are most often of greatest interest to the analyst. The smaller observations do convey valid information about the frequency of significant flooding even when they are censored (as done by EMA). Based on testing on observed and simulated data, the censoring of low floods is a reasonable approach for defining the magnitude and frequency of major floods. The analyst can choose to lower the MGB censoring threshold (using PeakFQ) and to censor fewer floods if the analyst believes this is a more reasonable approach.

3. The MGB procedure produces absurd results when applied to actual data.

Response:

The "absurd" results likely refers to concern over the number of low floods that are often censored using MGB. The MGB test can censor up to half of the flood observations but these floods generally are less than a 2-year recurrence interval event. As discussed earlier, such small floods do not yield much information about the magnitude and frequency of the larger floods. Based on testing of data for 82 long term gaging stations, the EMA/MGB procedure was shown to provide

reasonable estimates of extreme floods like the 1-percent chance flood. These results are summarized in the May 3, 2015 HFAWG Testing Report that is on the Bulletin 17C web site. Flood frequency graphs from the HFAWG Tester Report are provided later in Comment #47 and results indicate that the frequency curves based on the EMA/MGB more aggressive censoring are more consistent with the plotting positions for the highest flood data than the Bulletin 17B procedure that used the GB outlier detection method.

Attachment to Comment #47 – The MGB Disaster – circulated in 2014 to the HFAWG by Jerry Coffey
(Responses are given in **Bold Type**)

A lot of subjective arguments have been offered in defense of the EMA/MGB procedure on the theory that we don't know what the truth is so we hang everything on the outcomes we prefer. My first set of comments are based on four of the thirteen examples recently tested where the true facts are crystal clear. We will see that MGB as proposed [MGBT(50)] performs very badly when we know the correct answer.

Response:

The Monte Carlo simulations performed by [redacted] for the HFAWG Testing Report are quantitative experiments where he assumed a given frequency distribution and simulated 10,000 samples of given size from that distribution. For those samples, we know the true 1-percent chance flood. These analyses are described in the May 3, 2015 HFAWG Testing Report that is on the Bulletin 17C web site. These analyses demonstrate that the EMA with the MGB test provides more accurate estimates of the 1-percent chance flood than does the current Bulletin 17B technique.

The performance of the MGB test is discussed in the following peer-reviewed paper by Lamontagne and others (2016):

[redacted] **2016, Robust flood frequency analysis: Performance of EMA with multiple Grubbs-Beck outlier tests: Water Resources Research, Vol. 52, 3068-3084, April 2016.**

The theoretical basis for the MGB test is described in the following peer-reviewed paper published in Water Resources Research in August 2013:

[redacted] **and** [redacted]**. (2013) A Generalized Grubbs-Beck Test Statistic for Detecting Multiple Potentially-Influential Low Outliers in Flood Series, Water Resources Research, Vol 49, doi:10.1002/wrcr.20392, pp. 5047-5058.**

Taken together, these papers and reports constitute more than subjective arguments.

We do not know the correct answer (the true 1-percent chance flood) when using observed data. We can make judgments about whether there are low outliers in an observed sample.

In the next section we explore some reasons that may have contributed to the bad results, particularly incorrect assumptions and the multiple comparisons problem. If you don't like these reasons, come up with your own.

Finally we look at biases in the EMA/MGB estimates and the several different sources for these biases.

Part 1 -- Four Examples Where We Know the Truth.

Response:

The "truth" referenced here is based on the observation that many of the PILFs identified as being influential are not low outliers in the traditional sense. That is, these values do NOT depart significantly from the other observations. The MGB test is identifying low floods that have a large INFLUENCE on the upper tail of the frequency distribution. This concept is different from low outliers in the Bulletin 17B sense.

Let's take a look at the record from the Quinault River -- Station 12039500. This is one of the examples tested at the insistence of the SOH. The MBG procedure [MGBT(50)] sweeping out (down) from the median with a small p threshhold and back in (up) with a larger one. I draw your attention to the plot (these are the order statistics). Note that there are NO LOW OUTLIERS, NONE, ZERO! All the points track the curve quite closely, except possibly for one point on the HIGH side (see the blue circle) . But when the MGBT(50) procedure is applied, 47 low outliers are identified. And according to some this procedure involves what is described as a "correct" and accurate probability calculation. Note that ALL the points at or below the lower quartile are identified as low outliers based on this preferred calculation.

Now look what happens when the exact same procedure is applied beginning at the lower quartile rather than the median (I guess this would be MGBT(25) to follow the same notation). We use the identical "correct" and accurate probability calculations as above and the result is that NONE of the points at or below the lower quartile is a low outlier. So it appears that the same "correct" and accurate probability calculations are incredibly sensitive to the starting point for the sweep. Two such contradictory results cannot both be true.

So let's examine the standard Grubbs-Beck (GB) test as a tie breaker. Remember that the latest HFAWG draft report (April 2014) states that "The Bulletin 17B Grubbs-Beck (GB) test provides an objective and defensible recommendation as to which values should be treated as outliers". The complaint against the GB test is that it finds the lowest outlier but then misses higher points that may also be outliers. Fortunately in this case the GB test finds NO outliers so the criticism does not apply.

So here is the score --

Inspection of the curve shows NO low outliers.

The GB test shows NO low outliers.

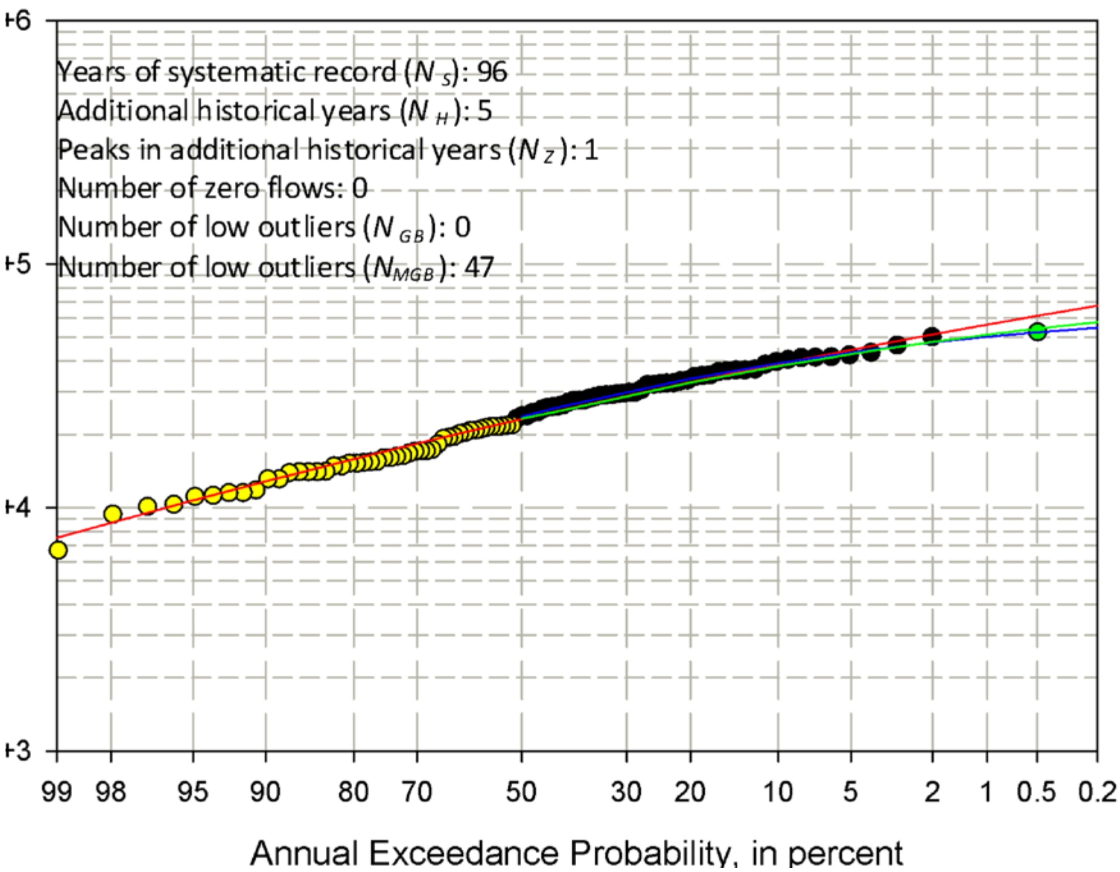
The complex MGB procedure started at the lower quartile - MGBT(25) - shows NO low outliers.

But the same procedure and same calculations started at the median - MGBT(50) - shows 47 low outliers.

Response:

The number of outliers or PILFs is not the critical issue. The critical issue is which method is giving the most reasonable estimate of the 1-percent chance flood or whatever the design discharge. As shown in the following graph for the Quinault River from the HFAWG Testing Report, the blue frequency curve based on the EMA/MGB [50] test is closer to the plotting positions of the highest floods than the red frequency curve based on the B17B/GB) detection test. If the analyst believes the MGB [50] test is identifying too many low floods, then the censoring threshold can be lowered to achieve what the analyst believes is more reasonable estimates of the 1-percent chance or x-percent chance flood.

Quinault River at Quinault Lake, WA
(Station 12039500)



Let's look at another example. This one is the Beaver River in Utah - Station 10234500.

-- Once again the points closely track the line fitted to the sample -- there are NO LOW OUTLIERS!
-- The GB test shows NO LOW OUTLIERS.
-- The lower quartile MGB test - MGBT(25) - finds NO LOW OUTLIERS.

-- But MGBT(50) using the same probability calculations as MGBT(25) detects 45 low outliers.

[see Exhibit A for the plot of the record at this station]

Here is a third example -- the Embarras River at Ste. Marie, IL Station 03345500

-- Inspection of the curve shows exactly one low outlier.
-- The GB test finds exactly one low outlier.
-- MGBT(25) finds exactly one low outlier.

-- MGBT(50) finds 34 low outliers!

[see Exhibit B for the plot of the record at this station]

And a fourth example -- Arroyo Mocho near Livermore CA Station 11176000

-- Inspection of the curve shows exactly one low outlier.
-- The GB test shows one low outlier.
-- MGBT(25) finds one low outlier.

-- MGBT(50) finds 19 low outliers!

[see Exhibit C for the plot of the record at this station]

These four examples show the extreme sensitivity of the MGB calculation to the starting point. For readers who may not deal with sensitivity issues every day -- "sensitivity" is the antithesis of "robustness" so in this case the robustness claims made by some are simply BS (bad science).

Response:

The hydrologic basis for starting at the median for the MGB test is that for many stations in the western US, it is necessary to censor up to 50 percent of the sample to get reasonable estimates of the design flood discharges. Starting at the median for censoring gives a test applicable nationwide.

See the following report for examples of frequency curves in the arid west:

[redacted] (2014). Evaluation of the expected moments algorithm and a multiple low-outlier test for flood frequency analysis at streamgaging stations in Arizona. Scientific Investigations Report 2014-5026, U.S. Geological Survey, Reston, VA. http://pubs.usgs.gov/sir/2014/5026/

In each of the first two cases, the correct result is that there are NO LOW OUTLIERS, In the third and fourth cases there is exactly ONE LOW OUTLIER. In all four cases, inspection, the GB test, and the MGBT(25) give the correct result. But in each case the result given by MGBT(50) is absurd. Mathematicians (and some philosophers) actually consider such demonstrations to be a form of proof -- the technique is called "reductio ad absurdum."

Response:

The MGB test is not being used to identify low outliers but rather PILPs, potentially influential low floods. This is a different concept.

In each of the four cases, MGBT(50) finds ALL the points at or below the 25th percentile to be low outliers, while MGBT(25) -- the same calculation starting at the lower quartile rather than the median -- finds zero or one low outlier in this range for each case. This is a rather stark contradiction.

The extreme sensitivity to the starting point argues that MGBT(25) may still be suspect. In these four cases MGBT(25) happens to give the correct number of low outliers. In seven other cases MGBT(25) identifies fewer low outliers in the bottom quarter of the data set than MGBT(50). But in two cases MGBT(25) gives a number of outliers roughly consistent with the results of MGBT(50). It is not clear whether there may be an optimum starting point to tame MGBT, but it is clear that the median is way too high.

Response:

Remember there are many stations in the arid west where the logical starting point for the MGB test is the median.

Part 2 -- What is wrong?

The examples above are completely transparent and clear cut -- the correct results are obvious and unambiguous. And MGBT(50) gives incorrect results, in fact, absurd results. So what is wrong with the MGB calculation? I have been asking this question for several years and the EMA/MGB developers have declined to address it before now.

Response:

The obvious results are that for the four stations noted above, there are one or no low outliers in the Bulletin 17B sense. The real question is what method provides the more reasonable estimate of the 1-percent chance flood. Since we have samples generally less than size 100, it is sometimes difficult to make this subjective decision. That is why the simulated results using an assumed distribution are important to consider.

There are some possibilities that would have been considered if the developers had been inclined toward research rather than advocacy. Here is a brief list:

-- Risky Assumptions

The calculation (described as "correct" in the April 2014 draft report) assumes that observations of annual maximum flow are 1) "independent" and 2) normally distributed.

NEITHER ASSUMPTION IS TRUE.

We touched on the reasons in the stationarity discussion last year and the relationship between LPIII and other distributions. For details see Exhibit D.

-- Multiple Comparisons

The calculation also may suffer from an issue statisticians describe as the "multiple comparisons" problem. If you test how close one order statistic is to an expected value, you can calculate (or at least estimate) a probability or "p-value" for some difference. But when you perform the test 40 or 50 times the p-values are no longer valid. The probability that at least one of the tests exceeds the criterion is much larger than the p-value used in the multiple tests. Furthermore an insignificant ripple in the sample curve that tests as an exceedance can occur just as easily in the first few tests as in the 40th, but the inferences drawn are radically different -- in the first case, one would infer a large number of low outliers and in the second case one would infer a much smaller number of low outliers.

Response:

The p-values are in fact valid, because they are estimated at each sample observation in the distribution. The order statistic is omega (w), which is a function of the kth observation and the total sample size n. See equation (5) from Cohn et al. (2013), as well as the text in Section 4.1 (page 5050, paragraph [37]) that states "*That distribution can be used to determine quantitatively if the kth smallest observation in a sample of size n is, in fact, unusually small*".

[redacted] **and** [redacted] **(2013) A Generalized Grubbs-Beck Test Statistic for Detecting Multiple Potentially-Influential Low Outliers in Flood Series, Water Resources Research, Vol 49, doi:10.1002/wrcr.20392, pp. 5047-5058.**

As we noted in the four cases above, the MGBT(25) calculation happened to give the correct result even when MGBT(50) gave absurd results. In theory the MGBT(25) calculation is not as vulnerable as the MGBT(50) because there are only half as many chances for error, but even this does not assure that the lower quartile is a "safe" starting point for the sweep. (In fairness, this multiple comparison effect has been something of an embarrassment to statisticians because many of the corrections for the problem are so conservative that they are not very useful.)

Perhaps this motivates some of the criticisms of the Rosner procedure that MGB claims to emulate. In 1983 [redacted] noted that the Rosner procedure tends to "identify more outliers than actually exist in the sample." Thirteen years later, Spencer and [redacted] cited the problem noted by [redacted], asserting that it can occur whenever the "k" potential outliers used in the Rosner calculation exceeds the actual number of outliers. Their solution -- "if a step-forward consecutive test is to be used ***[as with Rosner and MGB]***, one should have ***accurate a priori knowledge*** of the potential number of outliers" [emphasis added]. With MGBT(50), on the contrary, we have a bloated estimate of the number of outliers from the downward sweep -- exactly the opposite of the necessary condition identified by Spencer and McCuen.

Note that accurate a priori knowledge of the number of outliers tends to blunt some of the effect of multiple comparisons (permitting a more realistic correction to the probability calculations.)

Another point to note is that the radical choice of the median as the starting point for the sweep produces another absurdity in the original Rosner procedure. That procedure allows trimming of the data up to the selected p (.50 in this case) on either side. This would leave as an admissible procedure trimming ALL the data (if an even number of points) or all except one point (if an odd number of points). Rosner avoids this absurd situation by limiting the size of p, a strategy we should adopt here.

Part 3 -- Other Considerations

Comparing the four clear cut cases [e.g. MGBT(25) vs MGBT(50)], we find that the median starting point alone introduces a consistent downward bias of at least 10% (weighted average), 15% (median), or 18% (unweighted average). This is in addition to the tendency of EMA to tilt in the direction of the last few points on the high end (a bias that is exacerbated by censoring out large numbers of low points.)

Unfortunately some advocates of EMA/MGB have treated this downward bias as an "advantage" of EMA/MGB making the new procedure more desirable to parties who have an interest in smaller flood plains. This is one of the problems when objective research is replaced by advocacy.

Response:

For the four stations described above, it is true that MGB [50] test gives a lower 1-percent chance discharge than the MGB [25] and Bulletin 17B. Therefore, it is incumbent on the hydrologist or engineer to make a decision as to whether the censoring threshold should be lowered. No test or analysis techniques can be used in the absence of engineering judgment.

The new material added to the draft report of April 2014 is full of spin not appropriate to a research report. As we shall see below it is also full of nonsense (perhaps motivated by the spin). As several have noted privately, it is an advocacy document not a research document.

It characterizes the plots that fill the report as "samples" which is a good analogy if not entirely correct. In fact the points plotted to represent annual peak flows are the logs of the order statistics of the "sample" if the "sample" is regarded as a random realization of the underlying continuous curve (the "fitted" curve, in this case LPIII). Note that the recorded flows (and their logs and their order statistics) are sets of discrete points, finite samples if you will.

A lot of the "subjective" [the adjective actually used] arguments revolve around perceived vertical gaps between consecutive points. There is also some BS (bad science) about "discontinuities" and perceived "breaks" or "kinks" in the plots. Why is this BS? While the underlying distribution is smooth and continuous, the plots themselves are sets of discrete points -- every single point is a "discontinuity" and between any consecutive pair of points there is always a "break" or "kink." What the advocates are drawing attention to are vertical gaps between consecutive pairs of points. And such gaps are there between ANY pair of points in a discrete sample.

The important fact is the magnitude of such gaps and whether their magnitude differs from what might be expected at random. There is a substantial statistical literature on this and related subjects ([redacted] and many others come to mind) which would tell you that gaps of the size observed here are very common in samples. In other words, the slight-of-hand in the report that draws attention to gaps actually deals with gaps that are not significant. Thus the labored inferences that are based on these insignificant gaps are not science but spin.

Response:

The purpose of the MGB test is to identify what portions of the left-hand tail (Potentially-Influential Low Floods) influence the right-hand tail (big floods we are attempting to estimate). The gaps and their magnitudes result from streamflow - the physical process. The magnitudes of these gaps are not important. What is important is the leverage and influence these PILFs may exert on the right-hand tail. Additional examples - 328 sites in Arizona (Paretti et al., 2014) show that the gaps are not common.

[redacted]**. (2014). Evaluation of the expected moments algorithm and a multiple low-outlier test for flood frequency analysis at streamgaging stations in Arizona. Scientific Investigations Report 2014-5026, U.S. Geological Survey, Reston, VA. http://pubs.usgs.gov/sir/2014/5026/**

What we have known about EMA for six years or more is its extreme sensitivity to the points in the tail on BOTH ends of the distribution -- whichever way the points bend, EMA follows. In early tests of EMA, the lack of robustness and bias on both ends was quite obvious. While the B17B procedure is relatively robust to outliers at both ends of the curve, EMA has never been.

Response:

The "outliers" on the upper end of the frequency curve are large floods that need to be retained in the sample to obtain reasonable estimates of the large flood discharges like the 1-percent chance discharge. The EMA historical weighting procedure is more logical than the historical weighted moments approach in Bulletin 17B. So an improvement has been achieved with EMA with respect to adjusting for large floods independent of the MGB test.

With respect to low outliers, the current Bulletin 17B Grubbs-Beck test is NOT robust. The Bulletin 17B threshold often has to be increased to censor more low floods in order to obtain reasonable estimates of the design discharges (such as the 1-percent chance flood).

The motivation for the Multiple Grubbs Beck procedure was to remove as many points as possible to blunt the sensitivity of EMA to low points (ignoring the also problematic sensitivity of EMA on the high end). Note that B17B does not need this intervention -- with its supplementary procedures, it is already robust with respect to both of these problems. So EMA already tracked deviations on the high end because of its inherent lack of robustness, and with wholesale censoring of low end points (via MGB), this sensitivity was exacerbated. If the last few points turn down then the EMA estimate is biased downward. Choice of the 50% start produces absurd numbers of low outliers and imposes an ADDITIONAL 10% or so downward bias relative to the 25% start.

Response:

The motivation for the MGB test was to have an objective test for identifying PILPs so the analyst did not have to make too many subjective decisions. In some cases, like the four stations discussed here, the MGB [50] censoring threshold may have to be lowered to achieve more reasonable estimates of the design discharges.

Bulletin 17B may have given reasonable estimates of the design discharges for the four stations discussed. However, when applied to more arid streams in the western US, Bulletin 17B often needs intervention by raising the low censoring threshold.

Another reminder that the MGB test is not identifying low outliers but it is identifying PILPs.

Section 4.4 of the April 2014 draft contains about 1.7 false or misleading statements per paragraph. [I'll buy a modest lunch for each person who can identify them all.] Here are a few freebies to start you off.

1) The test comparing 25% and 50% start points was not an effort to cap the PILFs at 25% -- it actually displayed the unstable behavior of the MGB test. In the examples discussed above MGBT(50) classified ALL points below 25% as low outliers. Actually starting the sweep at 25% found at most ONE low outlier (which was the correct result). The behavior of MGBT(50) differed radically from the behavior of MGBT(25) with respect to the points below 25%. Not only was the behavior of MGBT(25) NOT predictable from the behavior of MGBT(50), the two sweeps contradicted each other over the set of points that were common to both tests. Also it is not clear or even likely that the number of low outliers below 25% found by MGBT(50) would be the same as those found in a sweep starting at 25% (claimed in another paragraph.) In 11 of the 13 cases actually tested, the count of low outliers BELOW 25% was less starting the sweep at 25% than starting the sweep at 50%. So about 85% of the tested cases contradicted the assumption made about the untested cases.

Response:

The highlighted statement is not clear. The testing showed that starting at the median or the 25th percentile would give different results. Where is the claim that the results would be the same.

2) Repeating just one example -- station 10234500 which we worked out above. Here we have NO low outliers. GB gets it right and even MGBT(25) gets it right. In these circumstances the B17B estimate is correct, and the EMA/MGBT(50) estimate is biased downward by at least 18% [comparing to EMA/MGBT(25)]. Given the facts, claiming that B17B is biased upward is absurd if not dishonest.

Response:

There are no claims that Bulletin 17B is biased upward. "Getting it right" would mean getting 1-percent chance estimates closer to the true (unknown) value. If the analyst believes the MGB [50] discharges are not reasonable, then the censoring threshold should be lowered.

Exhibit A Beaver River - station 10234500

(p.145 April 2014, p.141 April 2013)

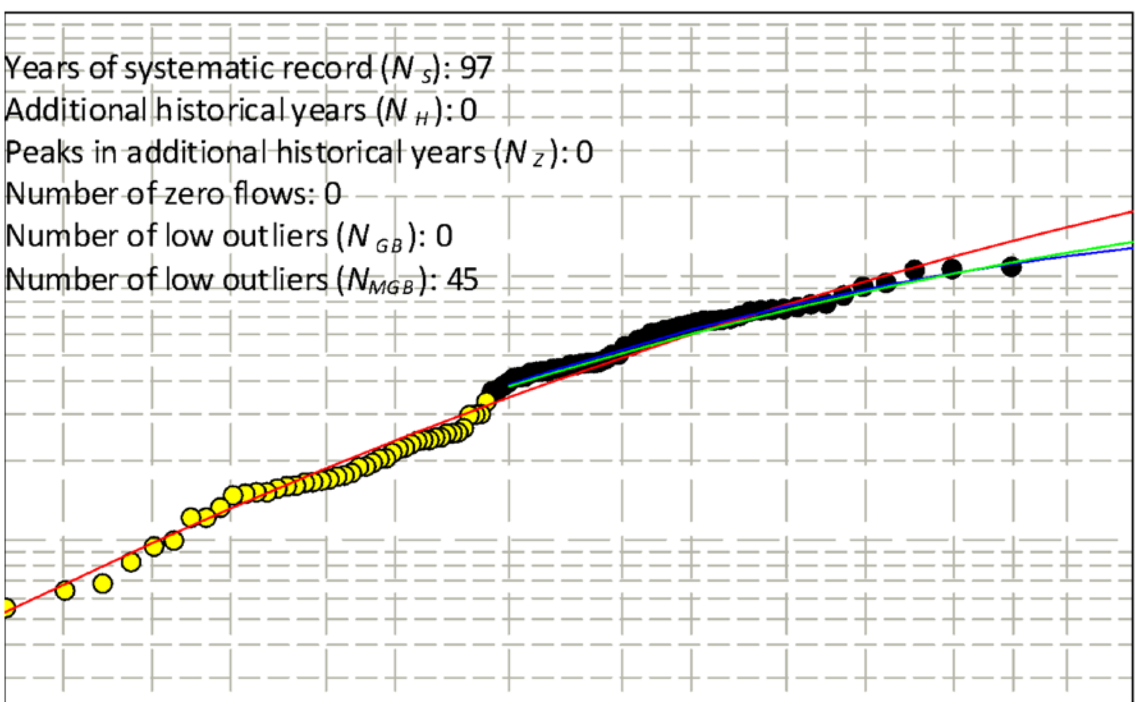
Once again the points closely track the line fitted to the sample -- there are NO LOW OUTLIERS!
The GB test shows NO LOW OUTLIERS.
The lower quartile MGB test - MGBT(25) - finds NO LOW OUTLIERS.

But MGBT(50) using the same probability calculations detects 45 low outliers.

Response:

The number of outliers or PILFs is not the critical issue. The critical issue is which method is giving the most reasonable estimate of the 1-percent chance flood or whatever the design discharge. As shown in the following graphs from the HFAWG Testing Report, for two of the three stations, the blue frequency curves based on the EMA/MGB [50] test are closer to the plotting positions of the highest floods than the red frequency curve based on the B17B/GB) detection test. If the analyst believes the MGB [50] test is identifying too many low floods, then the censoring threshold can be lowered to achieve what the analyst believes is more reasonable estimates of the 1-percent chance or x-percent chance flood.

Beaver River near Beaver, UT
(Station 10234500)



Note that again (as in the first example) the largest deviation from the curve is the highest point and (as was noted several years ago) the EMA estimate is dragged down by this excursion. That may be why the EMA estimates appear to be biased downward, but since the EMA and MGB developers never investigated this problem, we cannot know with any confidence what may be wrong. Careful observers may also note that there appears to be a strong serial correlation between the order statistics. This may offer a clue why the MGB probability calculations may be less precise and accurate than some expected.

Exhibit B Embarras River at Ste. Marie, IL Station 03345500

(p.136 April - 2014 or p.132 April - 2013)

-- Inspection of the curve shows exactly one low outlier.
-- GB test finds exactly one low outlier.
-- MGBT(25) finds exactly one low outlier.

-- MGBT(50) finds 34 low outliers!

Note that MGBT(50) says that all points at or below the lower quartile are low outliers, while the same routine with the same probability calculation starting at the 25th percentile, i.e. MGBT(25), finds only one low outlier. Again MGBT(50) contradicts MGBT(25) exhibiting the extreme sensitivity of the MGB calculation to the starting point and the absurd result of MGBT(50).

Embarras River at Ste. Marie, IL
(Station 03345500)

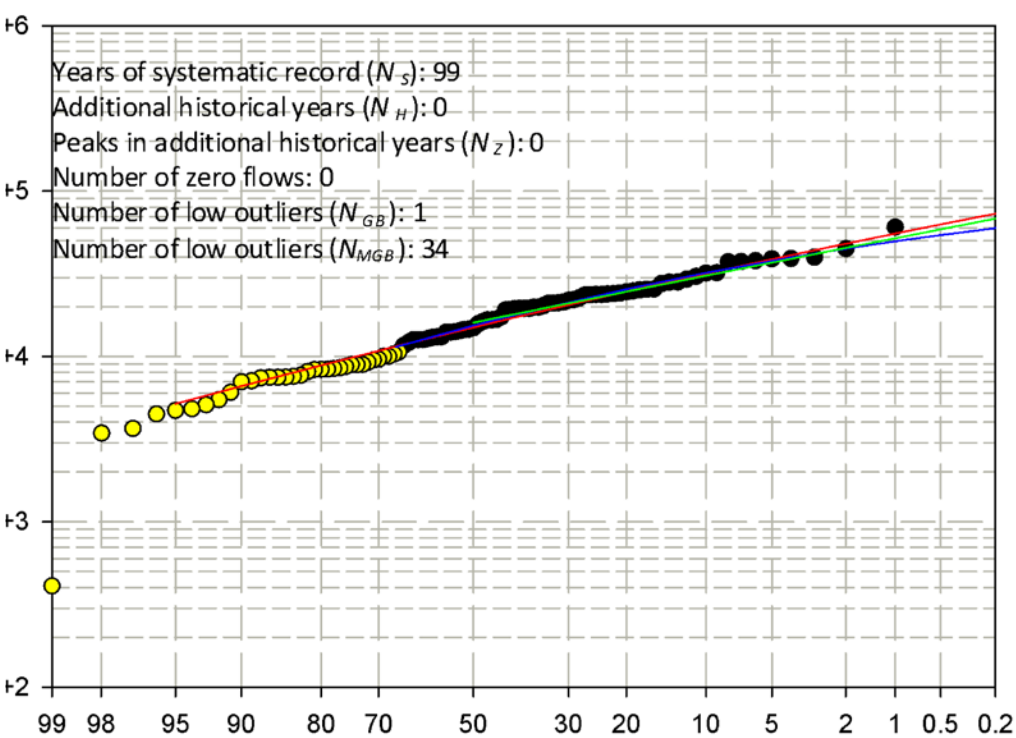


Exhibit C Arroyo Mocho near Livermore CA Station 11176000

(p.168 April 2014 or p.164 April 2013)

-- Inspection of the curve shows exactly one low outlier.
-- GB test shows one low outlier.
-- MGBT(25) finds one low outlier.

-- MGBT(50) finds 19 low outliers!

Note again the extreme sensitivity of the MGB calculation to the starting point, the absurdity of the MGBT(50) result, and the contradiction between MGBT(25) and MGBT(50) estimates at or below the lower quartile.

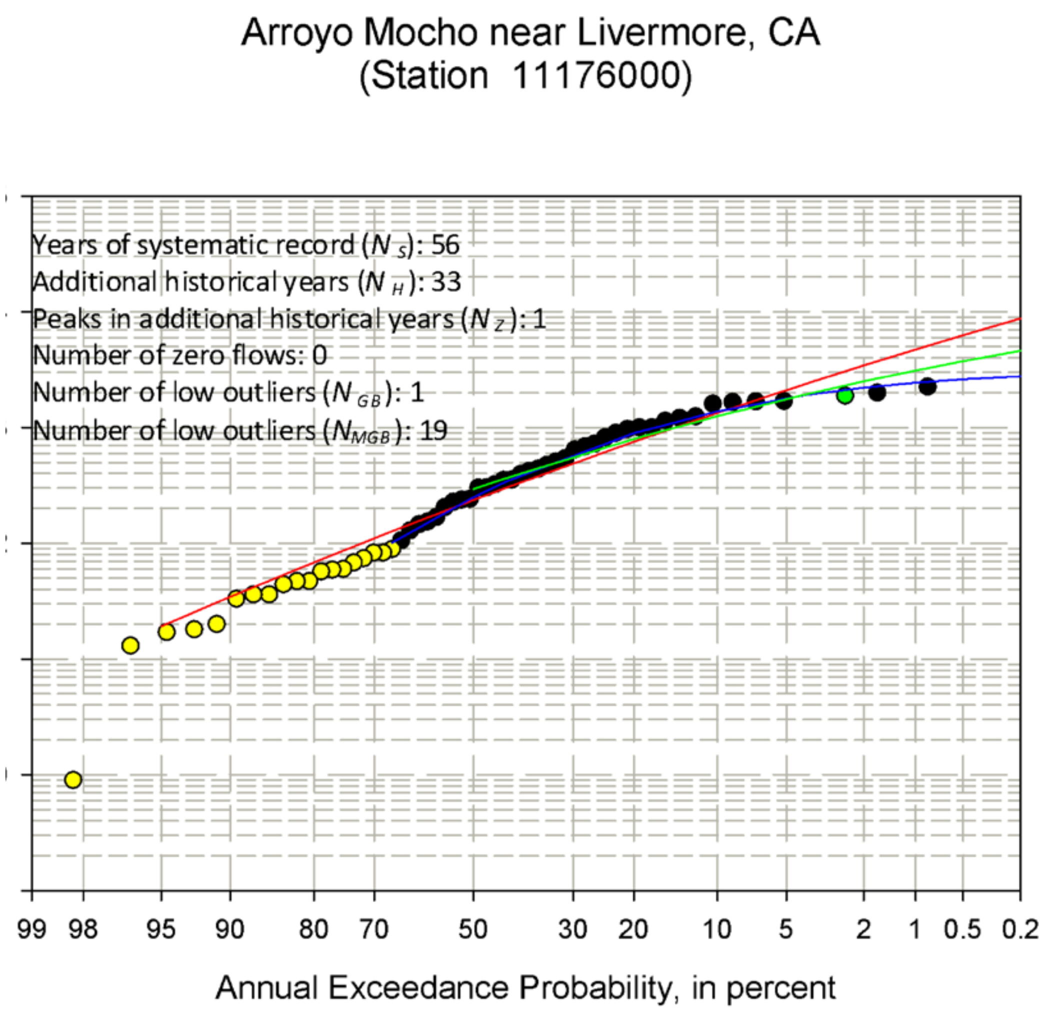


Exhibit D Bad Assumptions

1) As we saw in the stationarity discussion last year, these observations (peak annual flows) are approximately cyclo-stationary (i.e. there is little or no trend but probably some cyclic relationships). Observations from such processes are NOT INDEPENDENT. Also note that the order statistics (the points actually plotted in the many diagrams of the report) are never independent.

Response:

A simple lag-1 serial correlation estimate, or full correlogram on annual peak-flow data, routinely shows rho(1) is close to 0. Thus, it is reasonable to assume that peak-flow observations are independent in flood frequency. See Chapters 18 (page 18.2) and 19 (page 19.3) in the Handbook of Hydrology ([redacted] 1993, McGraw-Hill). The order statistics shown on the graphs are in fact an empirical CDF, and used in comparisons of fitting procedures. There is no issue with these order statistics as they are in fact independent. Log(Xi) of Qi, where Qi {.} are a sample of peak flows from an individual site, are independent. See, for a brief summary, http://en.wikipedia.org/wiki/Order_statistic#Probabilistic_analysis or the Order statistics book by David.

2) The assumed underlying distribution of peak flows is LP III (Log Pearson type III), a three parameter distribution related to the Gamma distribution. When the degrees-of-freedom parameter (i.e. the sample size) goes to infinity, the distribution described by Pearson converges to the normal. The Log Pearson III goes to the Log Normal when the skew (in the log domain) is zero. In the cases we have tested, the samples are not very large and the skew is rarely zero (check the test results), so normality is a demonstrably RISKY ASSUMPTION.

Response:

There is no assumption of normality when fitting the logarithms of the annual peak flows to the Pearson Type III distribution. The station or weighted skew along with the mean and standard deviation are used to estimate the design discharges. The assumption of normality is only for the MBG test. [redacted] et al. (2013) define the p values of the kth-order statistic as a function of random standard normal variables.

[redacted] and [redacted] (2013) A Generalized Grubbs-Beck Test Statistic for Detecting

Multiple Potentially-Influential Low Outliers in Flood Series, Water Resources Research, Vol 49, doi:10.1002/wrcr.20392, pp. 5047-5058.

The same assumption of normality was made for the Bulletin 17B Grubbs-Beck test. The one-sided 10-percent chance significance K values in Appendix 4 of Bulletin 17B are for normal samples of size N. The tables and test results available when drafting Bulletin 17B were for the normal distribution. The test for outliers was independent of skew and that was one attractive feature of the adopted test.

References

[redacted]. (2013) A Generalized Grubbs-Beck Test Statistic for Detecting Multiple Potentially-Influential Low Outliers in Flood Series, Water Resources Research, Vol 49, doi:10.1002/wrcr.20392, pp. 5047-5058.

[redacted] 2016, Robust flood frequency analysis: Performance of EMA with multiple Grubbs-Beck outlier tests: Water Resources Research, Vol. 52, 3068-3084, April 2016.

[redacted]. (2014). Evaluation of the expected moments algorithm and a multiple low-outlier test for flood frequency analysis at streamgaging stations in Arizona. Scientific Investigations Report 2014-5026, U.S. Geological Survey, Reston, VA. http://pubs.usgs.gov/sir/2014/5026/.