Responses to Comment #45: Management Issues for SOH – a comment from [redacted] circulated in 2013

The following comments by [redacted] were circulated to members of the Hydrologic Frequency Analysis Work Group (HFAWG) in 2013. [redacted] is a member of the HFAWG.

A: Technical

There are four serious technical issues that have not yet been addressed

1) EMA Sensitivity - low flows

EMA has shown unusual sensitivity and instability relative to small flows. The problem with low flows was obvious during the early testing of EMA when the new procedure failed its first series of tests (2008-9). The testing team has NOT investigated EMA's low-outlier sensitivity, focusing its attention on censoring data that makes EMA perform badly (even though B17B handles that same data with little censoring and little drama). Without fuller understanding of the sensitivity issue with low flows, it is very risky to use EMA on any series that contains low flows. Note that for simple gage series, B17B and EMA should produce identical results. The radical MGB fix is no substitute for investigating the problem.

Response: The Multiple Grubbs-Beck (MGB) test for censoring low floods was revised in 2013 and is not the same test as was used in the 2008-2009 time period. The MGB test used in Bulletin 17C is documented in a peer-reviewed journal article in Water Resources Research (Vol. 49, 5047-5058, August 2013). The low-outlier sensitivity was investigated in the HFAWG report "Evaluation of Recommended Revisions to Bulletin 17B", dated May 3, 2015, that is on the Bulletin 17C web site at (http://acwi.gov/hydrology/Frequency/b17c/index.html).

2) EMA Sensitivity - high flows

Shortly after the EMA sensitivity to low flows was documented, a similar problem was suggested for high flows (Minority Report Appendix F, item F-2 in 2010). There has been no effort to determine if EMA has some similar (and as yet also unknown) weakness relative to high flows. If EMA displays unusual sensitivity and/or instability relative to high flows (see the erratic behavior in the Nancy Barth example cited below), then the MGB fix (censoring low flows) may well exacerbate the issue.

Response: The EMA procedure as implemented in Bulletin 17C was compared to Bulletin 17B procedures at 82 long-term gaging stations and simulated data in the HFAWG report "Evaluation of Recommended Revisions to Bulletin 17B", dated May 3, 2015, that is on the Bulletin 17C web site at (http://acwi.gov/hydrology/Frequency/b17c/index.html). Of the 82 long-term stations, 36 stations had historical floods or high outliers. Historical floods or high flows were also simulated as part of a Monte Carlo analysis. These comparisons indicated that overall the EMA/MGB procedures provided more reasonable or accurate estimates of flood discharges than the Bulletin 17B procedure.

3) Invalid comparisons

Many of the comparisons between B17B and EMA in 2011 through 2013 were performed with input data sets that differed between the procedures being compared. Such tests are likely to be misleading. Since the variation present in the input data is no longer constant, the estimated variance of estimators is no longer comparable. Only one such comparison (performed by [redacted] during a two day meeting in 2012) has been released to HFAWG and it showed erratic behavior of EMA/MGB relative to B17B at the same levels of censoring -- a very different picture than the comparison published in the testing report. The testing team has asserted that the non-constant inputs make little difference, but has refused to actually perform constant input comparisons requested by members of HFAWG (including [redacted], SOH member) .

Response: The EMA and Bulletin 17B procedures were applied to the same observed and simulated data as is discussed in the HFAWG report "Evaluation of Recommended Revisions to Bulletin 17B", dated May 3, 2015, that is on the Bulletin 17C web site at (http://acwi.gov/hydrology/Frequency/b17c/index.html). The reference above to different input data sets relates to the number of low floods that were censored by the MGB test and GB test in Bulletin 17B. The resultant frequency analyses were based on a different number of censored low floods. During the June 12-13, 2013 meeting of the HFAWG, [redacted] discussed two stations in Texas where EMA and Bulletin 17B procedures differed significantly and compared results for the same number of censored low floods. The rationale for the results are documented in the minutes of the June 12-13, 2013 meeting (http://acwi.gov/hydrology/Frequency/minutes/HFAWG_mins_june12-13_2013_meeting_revised_071213.pdf). The critical issue is the reasonableness or accuracy of the flood discharges and not how many low floods are censored.

4) Capping MGB censoring to reduce the censoring of good data

Proposals have been made to cap the radical MGB procedure at a lower quantile so as to reduce the number of good data points censored by the current procedure capped at the median. This would make the procedure more consistent with the likely effects of interleaved mixtures. The usual "significance levels" accepted in the scientific community are generally 1% to 5%. As [redacted] points out, a 10% standard was accepted in B17B for good reason (the reason was the likelihood of mixtures in some data sets). This standard permits limited censoring in B17B (despite claims to the contrary). MGB permits censoring up to the median of the distribution (50%) and requires that the ***average*** significance level for outliers be relaxed to over 20%. Members of the testing group have offered the unsupported opinion that this would not make much difference. It is not clear whether this opinion means that MGB with a lower cap would still censor large amounts of good data, or if EMA/MGB would produce similar results with MGB capped at a lower level. If the latter is true then there is be no reason not to cap the procedure at a lower quantile (say between 10% and 25%). If the former is true then any use of the MGB procedure becomes difficult to defend. The single test performed by [redacted] suggests that neither interpretation is true. That example showed that when both B17B and EMA were applied to data sets censored to the same level, B17B produced superior results.

Response: Of the 82 long-term gaging station records, the MGB test censored more than 25 percent of the sample for 13 stations. For those 13 stations, the 1-percent chance discharge were compared between the MGB test censoring at the 50th percentile and limiting the censoring to the 25th percentile. These results are discussed in the HFAWG report on the Bulletin 17C web site at (http://acwi.gov/hydrology/Frequency/b17c/index.html). The differences in the 1-percent chance discharges were small for 10 of the 13 stations and for the other three stations, the MGB test censoring at the median provided the most reasonable results.

B. Management Issues

1) Lack of Independent Review

Only the initial review of EMA was conducted according to the protocols agreed to prior to testing, which protocols were designed to assure the objectivity of the process to any critic of the results. The initial protocols required that the development of EMA be performed without knowledge of the test data sets that would be used to challenge the procedure. After EMA failed this first round of tests, the challenge data sets were known to those who were "fine-tuning" the EMA/MGB fix and these data were in fact used to select the latest variant of the multiple pass MGB procedure. Some members of the development/testing team took umbrage at any suggestion that such a protocol was ever needed -- asserting that such "blind" procedures were an attack on their integrity.

Any attorney worth his fees would make a meal of this ego trip and thereby challenge the legitimacy of any "tests" in which the known test data were used to "fine-tune" the performance of the procedure after it failed the "blind" test. And since the testing group never ascertained the reasons for the failure, their efforts to fix the data rather than the procedure could easily be turned against us in court.

The writing of Bulletin 17C should be based on closing the gaps in the testing process. And any of the principal authors of papers advocating EMA and/or MGB should be removed from any influence over further testing.

Response: The 82 long-term stations represent a range of data sets with historical data, low outliers and high outliers for which Bulletin 17C will be applied in the future. These data sets provide a good test for the Bulletin 17C procedures. In addition, simulated data were used to evaluate the Bulletin 17C procedures. All these tests are documented in the HFAWG report on the Bulletin 17C web site at (http://acwi.gov/hydrology/Frequency/b17c/index.html).

2) The "All or Nothing" mentality

Too much effort has been spent by the testing team trying to turn EMA into a magic bullet. The revision of B17B is seen by many as a set of incremental improvements. EMA does not have to replace all B17B procedures to be helpful, and there are other issues beside the main algorithm. But the opportunity to produce incremental improvements has been compromised by EMA advocates in an effort to produce a single elegant equation to replace the messy cookbook that B17B uses to deal with the messy details of reality.

Response: The new procedures in Bulletin 17C do provide incremental improvements in several areas over Bulletin 17B. The new EMA procedures provide for a single, uniform and consistent framework for analyzing data sets with historical information, zero flows or low outliers; for identifying potentially influential low floods, for estimating more accurate confidence limits; and for developing new procedures for estimating regional skew.

3) Assuming what you can't prove

Actual testing has shown that stationarity is a reasonable approximation out to the length of most gage records (100 years or so). For 200 years or beyond, the evidence for stationarity and/or cyclostationarity is mixed, and data that has become available more recently tends to support the latter. Perhaps both assumptions should be considered in light of particular problems and issues and some judgement applied in each case. This would mean that calculations that rely on stationarity assumptions (e.g., EMA treatment of historical or paleo floods) should be offered only for comparison purposes (and with proper caveats). It may also be useful to add dates to flood records in case there is some breakthrough in using cyclic properties.

Response: The user of Bulletin 17C needs to evaluate whether the data series is stationary and whether Bulletin 17C procedures are applicable. The procedures described in Appendix 3 provide statistical test to determine whether the time series is stationary. In the section “Climate Variability and Change”, several references are provided that describe procedures for analyzing nonstationary data sets.