

Title: Review and Correction of the Peak Flow File

Purpose and Scope

This memorandum is a formal request to the Water Science Centers (WSCs) to address erroneous, incomplete, or otherwise questionable data in the U.S. Geological Survey (USGS) National Water Information System (NWIS) Peak Flow File (PFF). These questionable data have been identified through use of new peak flow screening tools and are listed in reports (PFRReports) compiled for each state. A new PFF editor, PKEntry, has been developed to highlight data identified as questionable by PFRReports, to aid in the editing of these data, and to document the edits through comment fields and radio status buttons. Details on PFRReports and PKEntry are provided below.

Specifically, this memorandum requests that each WSC complete the following tasks by **May 1, 2009**:

1. Review PFRReports for their state (<http://nd.water.usgs.gov/internal/pfreports/>);
2. Where possible, correct erroneous or incomplete PFF data using PKEntry; and,
3. Document the correction with comments or citations in PKEntry or internal memoranda.

For the purposes of this policy memorandum, questionable data are classified as (1) data-entry errors, (2) data-processing and interpretation errors, and (3) data-collection errors. (These three error categories are defined later in this policy memorandum.) This memorandum requires WSCs to correct only the first two categories. The last category will be addressed at a later date. This effort includes review of only the annual peak flow series and does NOT include other "peaks above base". "Peaks above base" and some additional refinements of peak flow qualification data also may be addressed in the future.

Past revisions to individual peak flow data have given rise to controversy, partly as a result of the lack of a systematic process for detecting and justifying needed changes, and partly as a result of the omission of a process for documenting the change and its rationale. This memorandum describes a process whereby questionable data are (1) detected through uniform statistical tests and screens, (2) thoroughly investigated and reviewed, and (3) corrected as appropriate with corrections documented using PKEntry comment fields and memoranda to the record.

WSCs that have previously edited the peaks or codes flagged by PKReports should now use PKEntry to identify those flagged peaks or codes that were reviewed and verified as "correct" or "questionable" so that resolution of those data can be tracked. These WSCs are encouraged to use PKEntry to provide comments documenting their changes to the data.

Background

Data from the USGS PFF are used in the determination of flood-frequency estimates that aid in the design of water and transportation infrastructure, delineation of floodplain boundaries, and regulation of development and utilization of lands throughout the Nation. In addition, these flood-frequency estimates are essential to understanding the implications of climate change on flooding. This high-profile database reflects and highlights the quality of USGS water-data collection programs; as such, the accuracy, characterization, and completeness of the data are critical. Quality assurance and improvement of the PFF is an essential effort towards strengthening USGS networks and science leadership.

The PFF is a collection of instantaneous maximum (peak) stream discharge data and associated gage height measurements made at streams throughout the Nation. The PFF is organized by agency, water-measurement site, and water year. The file contains one or more years for each

site, and one or more peaks per water year. Qualification codes are an integral part of the PFF and each entry in the database may contain one or more of these codes, which help to clarify some aspect of that entry. These codes also are used by the flood-frequency program (PEAKFQ) to control the processing of records containing historic information and other data requiring special statistical interpretation. Expanded definitions and discussion of the codes are included in appendix A.

Recent work on problem flood peak flows (Costa and Jarret, 2008) and ongoing technical reviews of WSC surface water programs have raised concern about the validity of some of the flood magnitudes, dates, and associated qualification codes stored in the PFF. Reviewers have detected errors or issues that can be classified as:

1. data-entry or transcription errors - mistyped or transposed information;
2. data-processing and interpretation errors - misapplied ratings, erroneous computation of direct or indirect measurements or other estimation procedures, or inappropriate use of peak-flow qualification codes, and
3. data-collection errors -inadequate, unreliable, insufficient, or missing data needed to defend a flow or stage determination or application of a method, such as an indirect, to conditions that violate the assumptions underlying the method.

This memorandum requests that WSCs focus on correction of the first two sources of error, data-entry and data-processing errors. For the third source of error, data-collection errors, WSCs are requested as part of this exercise to only identify the suspect data as such and then, at a later time, review and correct it. (These plans do not preclude WSCs from investigating questionable peaks and correcting them at any time, if circumstances permit. But those investigations and changes must be documented using PKEEntry and internal memoranda.)

Only the annual peak flow data are being corrected at this point. Partial series or “peaks above base” also are excluded from this analysis.

Detecting suspect PFF data

Under the direction of OSW, Karen Ryberg (ND WSC) developed a set of 14 tests to aid in identifying outlying or inconsistent peak flow data. These tests, collectively identified as PFScripts, utilize ratios of peak versus high-daily-values, comparisons of dates of peaks and high daily values, analysis of residuals from stage-versus-discharge and stage-versus-peak-daily-value regressions, and changes or inconsistent usage of qualification codes to detect outliers or questionable data. The output from the tests is a series of reports collectively identified as PFReports. Table 1 lists the individual tests and summarizes their operation. The results documented in the PFReports are the principal focus of this technical memorandum.

Table 1. Description of Peak Flow Screening Scripts (PFScripts) used to generate the Peak Flow File Screening Reports (PFReports).

Script Name	Description of Checks Performed
AMV	If an annual mean discharge value is present, an annual peak discharge likely exists and is currently missing
PGTDV	Peaks are greater than or equal to daily values
GH	Missing gage height for peaks with no code or any code other than 1 or 2
LRPDV	Outliers in linear regression of peaks on daily values
LRGHP	Outliers in linear regression of gage height on peaks
DP	Dependent peaks over 2 water years (peaks in September and October of a calendar year)
DropREG	Reports potential omission of codes 5 and 6 (Once a peak-flow file indicates regulation, subsequent peaks likely will have a regulation code)
Need7	Peaks that should have a code 7, historic peak
Not7	Peaks that have code 7 and should not

DropC	Once a peak code C has been assigned, it is likely that subsequent peaks also will be assigned a code C
AB	Peaks that should have a code A or B
LRGHPA	Statistical anomalies in regression of gage height on peaks
DA	Peak-flow sites without drainage area
HUC	Peak-flow sites without a HUC code

The PFScripts were run in batch mode using data from NWIS-Web; the resulting PFReports have been posted to webpages, organized by state (URL: <http://nd.water.usgs.gov/internal/pfreports/>). For each state a "Summary of Errors" report lists peaks that require further scrutiny, sorted by site and date. The summary report includes a column for each of the PFScripts named in Table 1; an X in one of these columns indicates an apparent outlier or inconsistency was identified by that script. The columns are grouped so that the results from the peak discharge and gage height tests appear first, then the results from discharge qualification code tests, and finally the results of DA and HUC tests. The summary report is tab delimited and may be easier to view if it is imported into Excel (or similar program) and viewed with a split screen to display the column headings. More detailed descriptions of the PFReports can be found in USGS OFR 2008:1284 (URL: <http://pubs.usgs.gov/of/2008/1284/>).

Most of the PFScripts are statistical tests that have been tuned to detect significant outliers only. As a result some questionable data may have been missed while good data may have been erroneously flagged. For example, a peak might be identified as an apparent outlier, but still be technically correct. Hydrographers should consult original records, annual data reports, station descriptions, or station analyses to make an informed decision about the validity of the data. **It is most important that use of these information sources be documented in subsequent editing.**

Editing the PFF

To expedite corrections to the PFF, Burl Goree (LA WSC) has developed a new PFF editor, PKEntry. In addition to adding, editing, and deleting peaks, PKEntry provides for the printing of the peak flow summary, retrieval of peaks from ADAPS end of year summary, and integration of potential problem peaks as identified by PFScripts. The program can be downloaded and installed from the OSW scripts page (<http://water.usgs.gov/usgs/osw/adaps/scripts.html>). Electronic forms are used to display and edit the data. PKEntry also includes a HELP file that can be accessed through the program.

A major feature of PKEntry is its logging and documentation capability. Comment fields are available for site-level data, groups of peaks (for an individual site) with some similar errors, and individual peaks. The station comment should be used to document actions or findings that relate to the site (latitude, drainage area, etc.) For example, a WSC could document a correction to the drainage area.

Groups of peaks may be flagged by some tests, such as LRGHP which often flags as an outlier a peak that departs from a rating because it was referenced to a different datum. These groups can be addressed with one correct or comment entry that is applied to the entire group.

The comment field for individual peaks should contain information related to a specific correction. For example, if a new qualifier code is added, date changed, or values are changed, the hydrographer should explain the reason for the change (new data, erroneous entry, etc.). **This documentation is critical and should be of sufficient detail so as to permit WSC hydrographers in the future to explain the change and find documentation (file folders, memos, e-mails, etc. supporting that change.)**

Finally, graphical user interface option buttons enable the user to indicate if the data were verified as "correct" (even though flagged questionable by PFReports) or verified "questionable" or if the

problem is unresolved pending additional information or analysis. (The presence of changes and edits will indicate that issues concerning other flagged data have been addressed.)

It is important to note that the log maintained by PKEntry is not a formal part of the PFF in the NWIS database. It is maintained as a text file on the local NWIS server and backed up by OSW nationally. Log entries can be viewed using PKEntry but are not available in retrievals directly from the PFF using ADAPS software or on NWISWeb. The log entries will likely be added to the NWIS database at a later date as part of a larger project to enhance the overall structure of the PFF.

Revisions to the PFF

As noted previously, suspect PFF data can be classified as data entry or transcription errors, data-processing errors, and data-collection errors. These classifications will serve to guide the review process and determine the level of checking and review of a proposed change. Correcting the PFF is largely a process of investigating data identified as questionable by PFScripts; seeking confirmational information from original records, documents, or publications; and using PKEntry to make the change and document the rationale for the change. While the detection of suspect peaks and the editing of the PFF can be greatly facilitated, there is no substitute for comparing the PFF data to various WSC datasets, office files, or published information.

- **Data-entry errors**

Data-entry errors occur when the PFF data do not match data appearing in original records or publications. These are expected to be the most common source of questionable peaks. In general, data-entry errors are “errors of fact” and are the easiest to detect, confirm, and correct. If PFF data have been flagged by PFScripts, and if the data do not match the original data (as indicated by annual data reports, station analyses, flow measurement notes, indirect computation reports, or internal memoranda), then the PFF data should be replaced by the data indicated in the original records. Questionable data should be reviewed closely for misplaced decimal points, transposed numerals, incorrect dates, or repeated numbers. If the error can be attributed to these or other “clerical errors”, then the investigation need not proceed further, although the WSC should ensure that proposed corrections are checked by a second person against original records. If the change results in a need to revise published records, then the change should be documented with comments using PKEntry or an internal memo. Include in the memorandum a table listing the errors, summarizing their cause and correction, and referencing backup data (original record) supporting the correction. If the data are checked and determined to be correct, then the peak data must be “verified correct” using PKEntry.

- **Data-processing errors**

A less common source of questionable peaks arises from misinterpretation or misapplication of available data, such as a misreading of the stage-discharge rating; a faulty shift application, or a miscomputed direct or indirect measurement. In these cases, the data (flow, stage, date, and station number) within the PFF and the original record agree, but one of the PFScripts has flagged the data as an outlier. Again, resolution is possible only by investigating the original records, not to simply check for a mismatch, but to ensure that the original records were processed (computed, applied, plotted) correctly. If they were not, and the solution is clear, the data should be corrected. If the solution is not clear, then the outlier will be treated as a suspected “data-collection error” as described below. In these cases, the peak must be “verified questionable” using PKEntry.

Additional considerations for checking the validity of data-processing errors include the following:

1. If the peak was an estimate, check that it was coded as such. Obviously, estimates must have a reasonable basis, but larger uncertainty generally accompany an estimate, and therefore, some increased inconsistency with other data can be expected.

2. If the peak was derived from a rating extension, look for evidence to verify the extension and the peak.
3. Review output from the PFScripts script LRGHP. Check for indications of a datum change or incorrect shift application.

In general, corrections for data-processing errors can be made without review or comment from sources external to the WSC. However, the WSC should ensure an internal review by a hydrologist who is experienced in the methods employed to make the original PFF estimate. Comments must be entered using PKEntry sufficient to explain the processing error and how the data were corrected or alternatively citing an internal memo providing this information. Include in the memorandum a table listing the errors, summarizing their cause and correction, and referencing backup data (original record) supporting the correction.

- **Data-collection errors**

Data-collection errors include peaks for which supporting data are scant or faulty (i.e., no original data, no flow measurement or survey notes, grossly incomplete measurement or survey notes, faulty measurement or survey practices, suspected misidentification of highwater marks, or misidentification or unknown station location) or inappropriate application of estimation techniques (indirect measurements applied to conditions in which slopes were too steep or flat, or the streambed too mobile to ensure that the channel geometry present during the peak could be reliably surveyed.) There is concern that some peaks that have been computed as water flows were actually debris flows, a distinctly different phenomenon. Conventional indirect measurement procedures should not be applied to debris flows and debris flow discharges should not be entered into the PFF.

In general, the treatment of peaks involving “data-collection errors” remains a challenge. When reviewing such records, caution and a degree of humility are required. Indirect measurements, in particular, are the result of close study and interpretation of data arising from a variety of unusual circumstances. In many cases the reporting hydrographers had firsthand access to the data or study sites or were closer to the event and to the actual data-collection effort than the present analysts. Furthermore, the data were reviewed by supervisors and technical specialists. As such, in the absence of compelling information indicating that PFF data are incorrect, the existing PFF data will be retained for the present. In no case should changes be based solely on the opinion or preference of the current analyst.

In order to develop improved techniques and policy, there is need to identify data-collection error issues, classify them as to their cause, better understand their frequency, and develop systematic procedures for dealing with them. Therefore, OSW requests that each WSC include comments using PKEntry that describe the data collection issue for each peak that has been verified as questionable. The descriptive comment should identify the source of the concern. For example, the description might use one or more items such as:

1. Missing or incomplete data;
2. Inconclusive or suspect indirect; or,
3. Inappropriate application of an estimating technique.

How these peaks are addressed will depend to a large degree on their number and circumstances regarding the availability of backup information.

Externally Documenting revisions to the PFF

Where appropriate, changes to the PFF should also be recorded as “revisions” and published in the annual data reports. Revisions (Novak, 1985, p. 109) may be needed to rectify errors in peak discharge values. “Revisions could result from additional data, re-examination and re-interpretation of data, or from the discovery of errors in computation.” “Only those published records of discharge that are substantially in error should be revised and only when the revisions

are reliable. The publication of indiscriminate revisions tends to cast doubt on the reliability of all records; therefore, questionable records should be withheld from publication until additional data have been collected (Novak, 1985, p. 103).” The instructions for preparation of water-resources data reports suggest that revisions need be published only if the difference in discharge between old and new data is more than about 10 percent. Revisions may be made for errors less than 10 percent if they are needed to maintain the correct relationship between the annual maximum discharge and a supplemental peak discharge or to keep annual maximums in the proper order of magnitude (Novak, 1985, page 103). If revisions to the PFF data are needed, revisions to other NWIS data probably will be needed also. This is especially likely when correction of the PFF data is necessitated by a change to a current meter or indirect measurement on which a rating or shift application is based. Changes to ratings should cascade through the NWIS system with corrections to the peak, unit, and daily discharge data. These revisions should also be reflected in an update to the Instantaneous Data Archive as described on OSW Technical Memorandum 2006.05.

References

Costa, John E.; Jarrett, Robert D., An Evaluation of Selected Extraordinary Floods in the United States Reported by the U.S. Geological Survey and Implications for Future Advancement of Flood Science. USGS SIR 2008-5164

[Attachment](#)