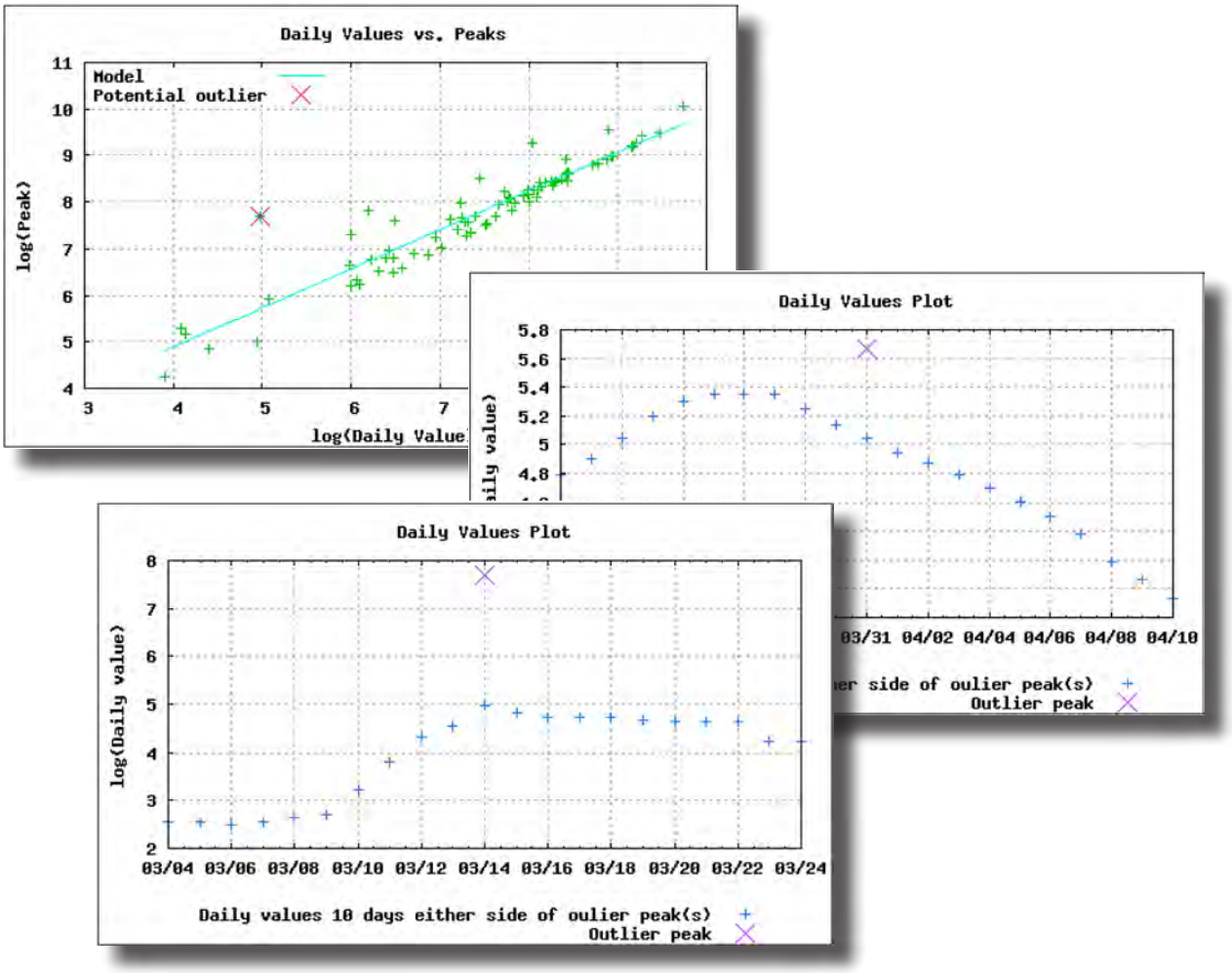
USGS
science for a changing world

Office of Surface Water

PFReports: A Program for Systematic Checking of Annual Peaks in NWISWeb



Open-File Report 2008–1284

U.S. Department of the Interior
U.S. Geological Survey

PFReports: A Program for Systematic Checking of Annual Peaks in NWISWeb

By Karen R. Ryberg

Office of Surface Water

Open-File Report 2008–1284

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia: 2008

For product and ordering information:
World Wide Web: *http://www.usgs.gov/pubprod*
Telephone: 1–888–ASK–USGS

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: *http://www.usgs.gov*
Telephone: 1–888–ASK–USGS

Suggested citation:
Ryberg, K.R., 2008, PFReports: A program for systematic checking of annual peaks in NWISWeb: U.S. Geological Survey Open-File Report 2008–1284, 17 p.

Contents

Introduction....1
Acknowledgements....1
Limitations....2
Data....2
Qualification Codes....2
Process Overview....2
Tests....4
AMV....4
PGTDV....4
GH....4
DP....4
DropREG....5
Need7....5
Not7....6
DropC....6
AB....6
DA....6
HUC....6
REGRESSION TESTS....6
LRGPDV....6
Regression details....6
LRGHP....7
Regression details....7
LRGHPA....12
Information Reports....14
Computer Code....16
Summary....16
References Cited....17

Figures

1. Screenshot of a summary of errors report....4
2–16. Graphs showing:
2. Data for a site with a strong correlation between the daily value and the peak streamflow....8
3. Data for a site with a relatively high degree of variability between the daily value and the peak streamflow....8
4. Residuals by water year for site referenced in figure 2....9
5. Daily values surrounding the peak identified as an outlier in figure 2....9
6. Semi-studentized residuals by water year for site referenced in figure 3....10
7. Plot of the daily values surrounding the peak identified as an outlier in figure 3....10

8. Data for a site with a strong linear correlation between peak streamflow and gage height 11
9. Data for a site with multiple linear relations between peak streamflow and gage height 11
10. Semi-studentized residuals from regression of gage height on peak streamflow ordered by water year 12
11. Outlier indicated on plot of regression of gage height on peak streamflow 13
12. Outlier indicated on plot of semi-studentized residuals from regression of gage height on peak streamflow ordered by water year 13
13. Peaks versus gage height for a site identified as having a statistical anomaly 14
14. Semi-studentized residuals from regression of gage height on peak streamflow ordered by water year for a site identified as having a statistical anomaly 15
15. Regression of gage height on streamflow for a site identified as having statistical anomalies 15
16. Semi-studentized residuals from regression of gage height on peak streamflow ordered by water year for a site identified as having statistical anomalies 16

Tables

1. Data-qualification codes 3
2. List of peak flow file tests and their descriptions 5

PFReports: A Program for Systematic Checking of Annual Peaks in NWISWeb

By Karen R. Ryberg

Introduction

The accuracy, characterization, and completeness of the U.S. Geological Survey (USGS) peak-flow data drive the determination of flood-frequency estimates that are used daily to design water and transportation infrastructure, delineate flood-plain boundaries, and regulate development and utilization of lands throughout the Nation and are essential to understanding the implications of climate change on flooding. Indeed, this high-profile database reflects and highlights the quality of USGS water-data collection programs. Its extension and improvement are essential to efforts to strengthen USGS networks and science leadership and is worthy of the attention of Water Science Center (WSC) hydrographers.

This document describes a computer program, PFReports, and its output that facilitates efficient and robust review and correction of data in the USGS Peak Flow File (PFF) hosted as part of NWISWeb (the USGS public Web interface to much of the data stored and managed within the National Water Information System or NWIS). Checks embedded in the program are recommended as part of a more comprehensive assessment of peak flow data that will eventually include examination of possible regional changes, seasonal changes, and decadal variations in magnitude, timing, and frequency. Just as important as the comprehensive assessment, cleaning up the database will increase the likelihood of improved WSC regional flood-frequency equations. As an example of the value of cleaning up the PFF, data for 26,921 sites in the PFF were obtained. Of those sites, 17,542 sites had peak streamflow values and daily values. For the 17,542 sites, 1,097 peaks were identified that were less than the daily value for the day on which the peak occurred. Of the 26,921 sites, 11,643 had peak streamflow values, concurrent daily values, and at least 10 peaks. At the 11,643 sites, 2,205 peaks were identified as potential outliers in a regression of peak streamflows on daily values.

Previous efforts to identify problems with the PFF were time consuming, laborious, and often ineffective. This new suite of checks represents an effort to automate identification of specific problems without plotting or printing large amounts of data that may not have problems. In addition, the results of the checks of the peak flow files are delivered through the World Wide Web with links to individual reports so that WSCs can focus on specific problems in an organized and standardized fashion.

Over the years, technical reviews, regional-flood studies, and user inquiries have identified many minor and some major problems in the PFF. However, the cumbersome nature of the PFF editor and a lack of analytical tools have hampered efforts at quality assurance/quality control (QA/QC) and subsequently to make needed revisions to the database.

This document is organized to provide information regarding PFReports, especially those tests involving regression and to provide an overview of the review procedures for utilizing the output. It also may be used as a reference for the data qualification codes and abbreviations for the tests. Results of the checks for all peak flow files (March 2008) are available at *http://nd.water.usgs.gov/internal/pfreports/*.

Acknowledgements

The program to perform these checks began with version 0.96 of the script “peakcheck” by Joseph Nielsen, Office of Surface Water. Karen Ryberg, North Dakota WSC, wrote the program to perform the checks, plot data, and output the reports.

Retired USGS employee Lamar Sanders contributed ideas and much enthusiasm to this project.

Many members of OSW, in particular Joseph Nielsen and Robert Mason, contributed discussion and suggestions. Robert Mason compiled the information for the headings in the error and information reports.

Richard Verdi, Florida Integrated Science Center, thoroughly examined the reports for Florida and found bugs that affected particular cases, contributing greatly to the improvement of the code. Gregg Wiche and Steve Robinson, North Dakota WSC, reviewed the reports for North Dakota early in the process and contributed ideas, in particular the daily values plot to help examine outliers from the regression of peak flow on daily values. Several hydrologists in other WSCs have reviewed or utilized some of the programs, and their comments and suggestions led to a greatly improved program.

Burl Goree, Louisiana WSC, developed a program that will simplify the entry of peak-flow data. He reviewed results

of the tests throughout the process and suggested the summary of errors report that may interface with his program.

Robert Mason, Joseph Nielsen, and Larry Bohman (Surface Water Specialist, Southeast Area) greatly improved this document through their insightful reviews and helpful comments.

Limitations

PFReports operates within certain limitations that include:

- All checks were performed on annual peaks only. Secondary (partial) peaks were not included.
- Peaks were retrieved from NWISWeb. It is thus conceivable that an error found by a check is due to a problem in the transfer of peaks from NWIS to NWISWeb or within NWISWeb and not in the peaks in the PFF in NWIS. However, there are no known NWISWeb bugs that would affect these checks.
- The tests described provide an automated method to check values in the PFFs. However, the tests may be supplemented by additional checks at the WSCs. Suggested manual checks include checks of urban sites that should be coded as urbanized (code C) but are not and checking that the peak of record in the PFF matches the manuscript in the site data sheets for the Annual Data Report.

Data

The source for all of the peak flow data checked was NWISWeb, specifically the Uniform Resource Locators (URLs) listed below, where $fips is the Federal Information Processing Standards (FIPS) code for the State or territory and $site is the site number.

For a list of peak flow sites and their hydrologic unit codes (HUCs) and drainage areas, *http://nwis.waterdata.usgs.gov/nwis/peak?state_cd=$fips&sort_key=site_no&format=sitefile_output&sitefile_output_format=rdb&column_name=site_no&column_name=station_nm&column_name=huc_cd&column_name=drain_area_va&rdb_compression=file&hn2_compression=file&list_of_search_criteria=state_cd*.

For the peak data, *http://nwis.waterdata.usgs.gov/nwis/peak?site_no=$site&format=rdb*.

For a list of daily value sites, *http://nwis.waterdata.usgs.gov/nwis/dv?state_cd=$fips&index_pmcode_00060=1&sort_key=site_no&format=sitefile_output&sitefile_output_format=rdb&column_name=site_no&column_name=station_nm&rdb_compression=file&list_of_search_criteria=state_cd*.

For daily values, *http://nwis.waterdata.usgs.gov/nwis/dv?cb_00060=on&format=rdb&begin_date=1850-01-01&end_date=2020-09-30&site_no=$site*.

For annual-mean values, *http://waterdata.usgs.gov/nwis/annual/?site_no=$site&PARAmeter_cd=00060&year_type=W&format=rdb&submitted_form=parameter_selection_list*.

Qualification Codes

Data-qualification codes used to characterize the peak data or the conditions under which they were collected are listed in table 1 along with their descriptions. These codes are important in documenting USGS data for internal use and for the public and are used by the flood-frequency program (PEAKFQ) to control the processing of records.

Process Overview

PFReports was run in Unix and the output was posted to a Web site. The output lists and summarizes peaks failing various tests. No attempt has been made to correct any of the data.

Each State or territory has a home page accessible from *http://nd.water.usgs.gov/internal/pfreports/*, that begins with a summary of the data used. For example, the top of the page for Louisiana showed the following on January 23, 2008:

1,011 peak flow sites
381 peak flow sites have streamflow values
170 sites have a peak flow file and a daily values file
166 sites have both daily values and streamflow values in the peak flow file
166 sites have streamflow values in the peak flow file, a daily values file, and annual-mean values

This listing indicates that of the 1,011 sites that had peak flow files, only 381 had peak streamflow values in them. Therefore, for the majority of sites, 630, most of the tests were not performed. In Louisiana, there were 170 sites with daily values. Daily values data allowed additional tests to be performed on those 170 sites. For some sites across the Nation, the period of record may be short or have 1 or more breaks. Tests involving daily values were still performed, but only on the portion of the record with concurrent peaks and daily values. All tests were performed for sites with at least 10 peak streamflow values (the regression of peaks on daily values required at least 10 peaks with concurrent daily values), daily values, and annual-mean values.

After the data summary, there is a link to the qualification code list and descriptions provided for easy access. Next is a "Summary of Errors" that lists peaks with potential problems, sorted by site and date. A column in the report represents each test and the heading for that column is an abbreviation for the

Table 1. Data-qualification codes.

Code	Description
	Peak Streamflow-Qualification Codes (peak_cd)
1	Discharge is a maximum daily average
2	Discharge is an estimate
3	Discharge affected by dam failure
4	Discharge less than indicated value, which is minimum recordable discharge at this site
5	Discharge affected to unknown degree by regulation or diversion
6	Discharge affected by regulation or diversion
7	Discharge is an historic peak
8	Discharge actually greater than indicated value
9	Discharge due to snowmelt, hurricane, ice-jam or debris dam breakup
A	Year of occurrence is unknown or not exact
B	Month or day of occurrence is unknown or not exact
C	All or part of the record affected by urbanization, mining, agricultural changes, channelization, or other
D	Base discharge changed during this year
E	Only annual maximum peak available for this year
	Gage height qualification codes (gage_ht_cd, ag_gage_ht_cd)
1	Gage height affected by backwater
2	Gage height not the maximum for the year
3	Gage height at different site and (or) datum
4	Gage height below minimum recordable elevation
5	Gage height is an estimate
6	Gage datum changed during this year

name of the corresponding test. An X appears in each column for which a peak has a potential error. The columns are grouped so that the results of the peak tests appear first, then the results of code tests, and finally the results of site tests. An example of the summary of errors is shown in figure 1.

Users may find that the review (and perhaps the correction process) is more efficiently done by checking peaks down each column before proceeding to the next column. However, users may want to investigate all the errors reported for a peak (moving across the columns). Regardless, they need to realize that even if an apparent error in one column is corrected or resolved, that peak may be flagged in subsequent columns for different reasons and, thus, may not be addressed by any one correction.

Some errors are not associated with a particular peak, but with the site as a whole. These include missing HUC, missing drainage area, and statistical anomalies in the regression of gage height on peaks. Therefore, a site with one of these errors is listed with all peak dates in the summary because a missing HUC or drainage area affects all the peaks and the statistical anomalies may be caused by a range of peaks from one outlier to the majority of the peaks.

The summary report is a tab-delimited text file, and the columns in the report may not line up properly in a Web browser. The text file may be imported into Excel or another program that recognizes tab delimiters and the output grouped, sorted, and tracked by WSC preference.

Once the errors for the initial columns, AMV, PGTDV, and GH, have been resolved, it may be desirable to rerun the tests because the corrections may reduce subsequent errors and will affect the regression relations used for some tests. Users may request that the program be rerun by sending an e-mail to

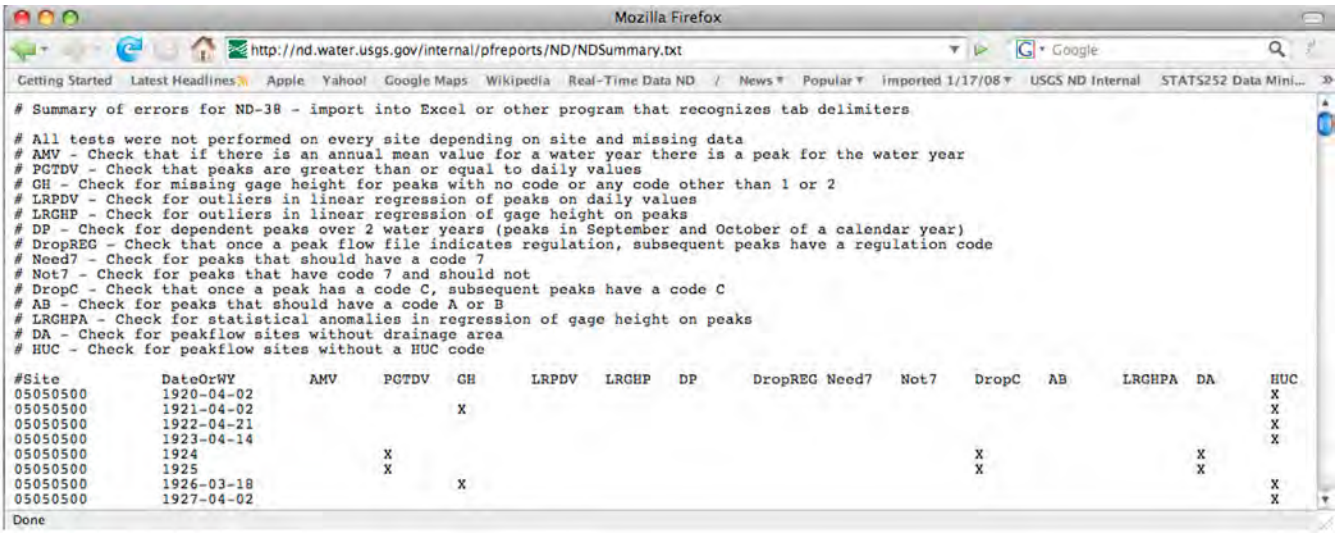


Figure 1. Screenshot of a summary of errors report. The file is a tab-delimited text file accessed by a Uniform Resource Locator (URL). The column headings and X codes, indicating potential errors, may not line up on screen. The file should be imported into Excel or another text editor that recognizes tab delimiters.

kryberg@usgs.gov, but they must understand that rerunning the program requires hours of computer time for each State and a lengthy queue is likely.

After the summary of errors, links to the individual tests are available on the State home page. In addition to the output from the screening programs, various informative reports are posted on the home page. The results of these information reports are not included in the summary report.

Tests

The tests for which the output is summarized in the "Summary of Errors" report are listed in table 2 and are described in the following subsections. The title of each subsection corresponds to the abbreviation for each test, and the titles are the same as the column headings listed in the summary report. Each test report has a description of the test, considerations regarding restraints or limitations of the test, and suggestions for appropriate corrective actions at the top of the file. That same information is repeated in this document with additional information for some tests, in particular, more details on the use of linear regression and residual analysis.

The tests are described in the order of the summary of errors and table 2 with the exception of the tests based on regression, which are described in a separate section of this report.

AMV

This test checks for the absence of peaks for sites and years having complete daily values (DV) record for a water year. A complete DV record is usually accompanied by the presence of an annual mean value in the database. Peaks should be entered in the PFFs for all discharge sites for which sufficient records exits to compute annual statistics with complete DV records. The test does not apply for sites with less than 5 years of record.

PGTDV

This test finds and lists peaks that do not meet or exceed the DVs for the date of the peak. This test applies only to those sites/peaks for which DV data exist.

GH

This test lists peaks lacking a gage height. Generally, a gage height serves as the basis for a peak-flow determination. Peaks with code 1 (discharge is a maximum daily average) were not checked and are not listed. Peaks with code 2 (discharge is an estimate) are listed in a separate report in the information reports section.

Table 2. List of peak flow file tests and their descriptions.

Abbreviation used in summary of errors	Description
AMV	Check that if there is an annual-mean value for a water year there is a peak for the water year
PGTDV	Check that peaks are greater than or equal to daily values
GH	Check for missing gage height for peaks with no code or any code other than 1 or 2
LRPDV	Check for outliers in linear regression of peaks on daily values
LRGHP	Check for outliers in linear regression of gage heights on peaks
DP	Check for dependent peaks over 2 water years (peaks in September and October of a calendar year)
DropREG	Check that once a peak flow file indicates regulation, subsequent peaks have a regulation code
Need7	Check for peaks that should have a code 7, historic peak
Not7	Check for peaks that have code 7 and should not
DropC	Check that once a peak has a code C, subsequent peaks have a code C
AB	Check for peaks that should have a code A or B
LRGHPA	Check for statistical anomalies in regression of gage heights on peaks
DA	Check for peak-flow sites without drainage area
HUC	Check for peak-flow sites without a HUC code

DP

This test checks to see if two peaks in September of one water year and October of the next water year are from the same flood event. According to WRD Data Reports Preparation Guide (Novak, 1985), two peaks are considered independent only when a well-defined trough between them is equal to or less than 75 percent of the instantaneous discharge of the lower peak. The test includes checks for dependent peaks spanning two water years, but applies only to those sites for which daily values exist. Because long-term instantaneous values are not available in NWISWeb, DV data (instead of instantaneous) were utilized to approximate the trough; thus the test may be insensitive for some sites.

For peaks identified as dependent, examine the instantaneous data and, if peaks still meet the dependent criteria, remove the lower of the two peaks from the peak-flow file and replace it with the next highest peak for that water year.

DropREG

This test checks for inadvertent omission or change of codes for regulation and diversion. Once a peak code of 5 is first used, any subsequent peak that does not have a code of 5 or 6 will be flagged. Once a peak code of 6 is used, any subsequent peak that does not have a code of 6 will be flagged. After a site is qualified with a code 5 or 6, the test flags any subsequent peaks that omit the regulation or diversion code, even though it is possible that regulation may end or decrease in effect.

Peaks affected by regulation or diversion, but to an unknown or insignificant degree (less than 15–20 percent of the discharge), should be qualified with code 5. Peaks affected by known or planned regulation should be qualified with code 6. Regulation of one peak discharge does not necessarily imply regulation of succeeding peaks; the occurrence of significant regulation or diversion should be verified independently for each peak. However, once a stream is regulated or diverted sufficiently to affect the peak flow, peaks generally continues to be affected. In rare cases, if a source of regulation has been removed or ceases to be effective, the coding could be discontinued or a code 6 replaced with a code 5 if residual regulation from unidentified sources remains.

Need7

This test checks for peaks missing historic-peak code 7. Peaks observed during non-systematic periods of record are identified as “historic,” a term referring to the reason the peak was recorded rather than to either a period of time or to the relative magnitude of a peak. Peaks may be nonsystematic (thus historic) if they occur before systematic streamgaging begins, after it ends, or during a break in systematic gaging unless the peak was observed and recorded in anticipation of near-term initiation of streamgaging or its resumption. This test was performed on sites with at least 5 streamflow values to avoid recommending code 7 for new gage sites.

Historic peaks include nonsystematic peaks recorded solely because the peak was very large even if they occurred

in the years just before streamgaging begins or the years just after it ends. The term "historic peak" does not refer to the largest peak in a record.

Not7

This test checks for peaks incorrectly qualified with historic-peak code 7. Historic peaks include nonsystematic peaks recorded solely because the peak was very. The term "historic peak" does not refer to the largest peak in a record. Peaks qualified with code 7 but occurring at the beginning of, during, or at the end of what appears to be a systematic period of data collection are listed as peaks that may be incorrectly qualified with code 7.

DropC

This test checks for inadvertent omission of a code C, identifying unusual land use or channel characteristics. Peaks affected by urbanization, mining, agriculture, or channelization are qualified with code C. Once a site is affected by these conditions, the condition usually stays effective.

Add code C in years subsequent to the introduction of factors previously described unless the code was inappropriately applied. If it was inappropriately applied, remove it.

AB

This test checks for omitted or inconsistent A or B qualification codes that should accompany peaks with inexact dates. Peaks with unknown or inexact dates should be qualified with code A (year is unknown or inexact) or B (month or day are unknown or inexact). This test lists peaks that lack code A or B but have an invalid or missing month or day (examples include peak dates of 1978, 1978–04, 1978–04–00, or 1978–00–00), and peaks that have code A but have a valid month and day.

Review listed peaks against original records. Peaks with code A and a valid date most likely should be recoded with code B or the month and day should be removed. Peaks with a year but no month or day should be reviewed to determine if a code A or B is appropriate. Document the reason for any changes.

DA

This test checks sites for an omitted drainage area. Drainage areas should accompany peak flows unless the drainage area is indeterminate.

HUC

This test checks sites for an omitted hydrologic unit code (HUC). HUCs should be included in the PFFs for all sites.

REGRESSION TESTS

The tests described in this section were done by using linear regression and residual analysis. The tests are described in general first. Then more detailed information about the regression techniques used is provided along with sample plots illustrating the types of problems identified.

LRGPDV

This test identifies suspected outlier peaks from a regression of peak streamflow on DV. Some of the outliers may already have been corrected by fixing the errors listed in PGTDV. For example, a peak mistyped as 100 rather than 1,000 will be identified as a peak that is less than the corresponding DV in PGTDV and as an outlier in the LRGPDV report. Corrections made for PGTDV, additions made on the basis of AMV, and changes made on the basis of this report will affect subsequent linear regression relations and change the outcome of future applications of this test, including possibly identifying outliers previously unidentified.

Outliers are listed by station and date on the online report (the site number and name are a clickable link to the PFF in NWISWeb) and indicated as points with a large red X through them on plots of daily values versus peaks; water year versus semi-studentized residual; log predicted peak versus semi-studentized residual; and a hydrograph of daily flows 10 days before and after the identified outlier(s). The test is performed only on sites that have DV data and at least 10 peaks. To minimize false reports, the DV with the DV/Peak ratio closest to 1 from either the day before the peak, the day of the peak, or the day after the peak is used in the regression.

For indicated outliers, check the magnitude and date of peaks. DVs may also be incorrect. Examination of the DV hydrographs will help determine if the initial focus of the data checking should be on the peak value, the peak date, or the daily value.

Regression details

The independent, or explanatory, variable is the DV (from the day before the peak, the day of the peak, or the day after the peak, whichever has a DV/Peak ratio closest to 1) and the dependent, or response, variable is peak streamflow. These two values have a positive linear relation; large daily values correspond to large peak values. The values were transformed by taking their logarithm to make them approximately normal and meet one of the assumptions necessary to perform linear regression (Neter and others, 1996).

Linear regression was performed and the results were examined for outliers, or points that do not fit the regression model well, the regression model being

$$\log(peakQ) = \log(DV) + e,$$

where

e is the difference between the peak streamflow predicted by the model and the observed peak streamflow, called the error or the residual.

Outliers are defined as those points with a large, statistically improbable residual. The definition of "large" could vary from site to site depending on the variability inherent to that site. For some sites, there is a very strong correlation between the daily value and the peak (fig. 2). Other sites with flashier streamflow characteristics have greater variability between the daily value and the peak (fig. 3). The residuals in figure 2 are relatively small compared to those in figure 3. Therefore, the residuals were adjusted, or rescaled, to account for differing variability by using the equation for semi-studentized residuals (Jennifer A. Hoeting, Ph.D., Colorado State University, oral and written commun., 2004):

$$e_i^* = \frac{e_i}{\sqrt{MSE}},$$

where

e_i^* is the adjusted or semi-studentized residual,

e_i is the residual (Helsel and Hirsch, 1995) for each data point used, and

MSE is the mean-squared error (Helsel and Hirsch, 1995) or estimated variance of the residuals.

The semi-studentized residuals were then examined for outliers. Outliers were defined as data points with semi-studentized residuals greater than 4 or less than -4. For sites with at least 20 peaks, the outliers have a probability of occurrence of less than 0.001, given the linear relation between daily values and peaks. The outliers may indicate that a different mechanism was at work for that particular peak (backwater, extremes) than for the majority of the other peaks or that the data was miscoded. For many outliers, miscoding the peak streamflow value or the date of occurrence is the most likely problem. If the date is miscoded, the program is comparing the peak to a daily value other than one associated with the peak event.

In figure 2, the problem with the identified outlier is not easily seen but becomes obvious when the residual plot (fig. 4) and the plot of daily values near the peak (fig. 5) are examined. Figure 4 shows that despite the outlier peak being fairly close to the regression line in figure 2, once adjusted for the low variability at the site, the peak stands out as being very different from the rest of the data. The fact that the peak plots 4 days after the peak of the DV hydrograph in figure 5 suggests that the date of the peak should be verified.

Figures 6 and 7 show the outlier peak also identified in figure 3, the site with a greater degree of variability between peaks and daily values. Examination of figure 7 suggests that the peak may have been entered incorrectly. At a log scale, there is a very large difference between the peak and the daily value. Upon further inspection of the PFF, it was found that the daily value was 146 cubic feet per second (cfs) and the peak was entered as 2,175 cfs.

LRGHP

This test identifies suspected outlier peaks from a regression of gage height on peak streamflow. The sensitivity of this test is diminished by changes in the datum, the stage-discharge rating, or regulation. These changes are apparent in plots of residuals versus date. Hence, outliers are initially identified through statistical tests, but must be checked against residual plots and the data itself (there may be a code that explains the outlier). Outliers are listed by station and date on the online report and indicated as points with a red X through them on plots of peak versus gage height; water year versus (semi-studentized) residual; and log predicted gage height versus residual. The test is performed only on sites that have at least 10 peaks.

Regression details

The independent variable is log(peak streamflow) and the dependent variable is log(gage height). The residuals were adjusted for variance (semi-studentized) so that the test applies to sites with differing degrees of variability between streamflow and gage height. The residuals were examined for behavior that would be statistically improbable given a stable relation between peak streamflow and gage height. Residuals in the residual plots should be randomly scattered above and below a centerline (y=0). Approximately the same number of residuals should be above the line as below, and the majority should be within ±3 studentized residual units of the centerline. The magnitude of the adjusted residuals was checked and any greater than 4 or less than –4 were identified as potential outliers. Adjusted residuals greater, in absolute value, than 4 are unlikely given the linear relation between peak flow and gage height.

For some sites, there is a very strong linear relation between peak streamflow and gage height (fig. 8). For other sites, there appear to be multiple linear relations (fig. 9). Multiple relations are usually caused by a change in gage datum or regulation.

Multiple relations represent different populations of data, with different regression relations. Normally, different populations are not mixed in regression analysis. However, the differences identified can be used to our advantage, especially when examining the adjusted residuals. The time of this change is usually obvious in the water year versus residual plot (fig. 10). Any plot like figure 10 should be checked for a qualification code explaining the change, in this case the obvious difference in the peak streamflow/gage height relation between 1965 and 1970. For sites with multiple changes in gage datum and/or changes in regulation, differences like those in figures 9 and 10 may not be as dramatic because of the overall greater degree of variability in the peak streamflow/gage height relation at the site.

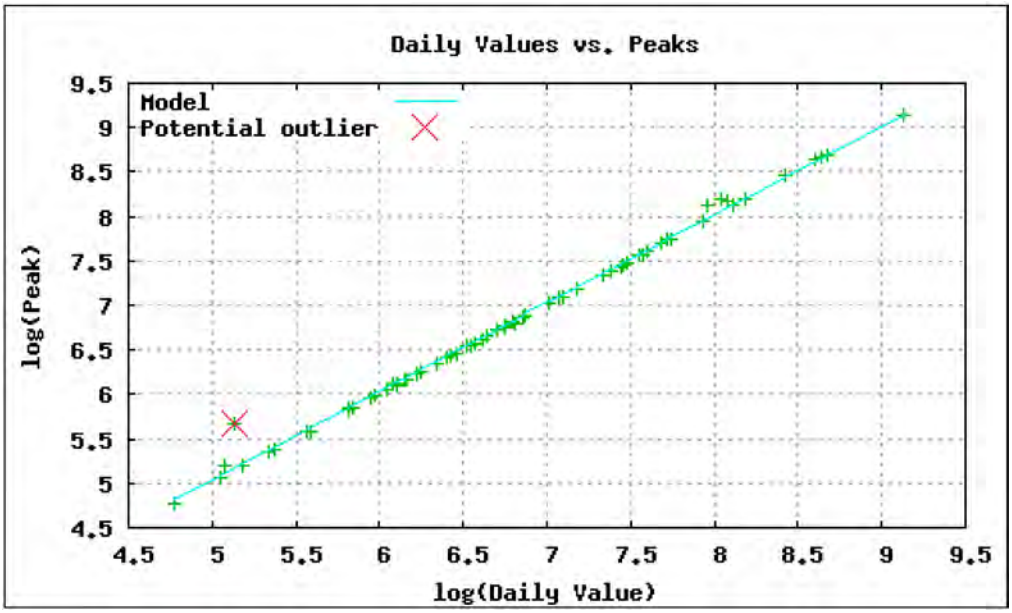


Figure 2. Data for a site with a strong correlation between the daily value and the peak streamflow.

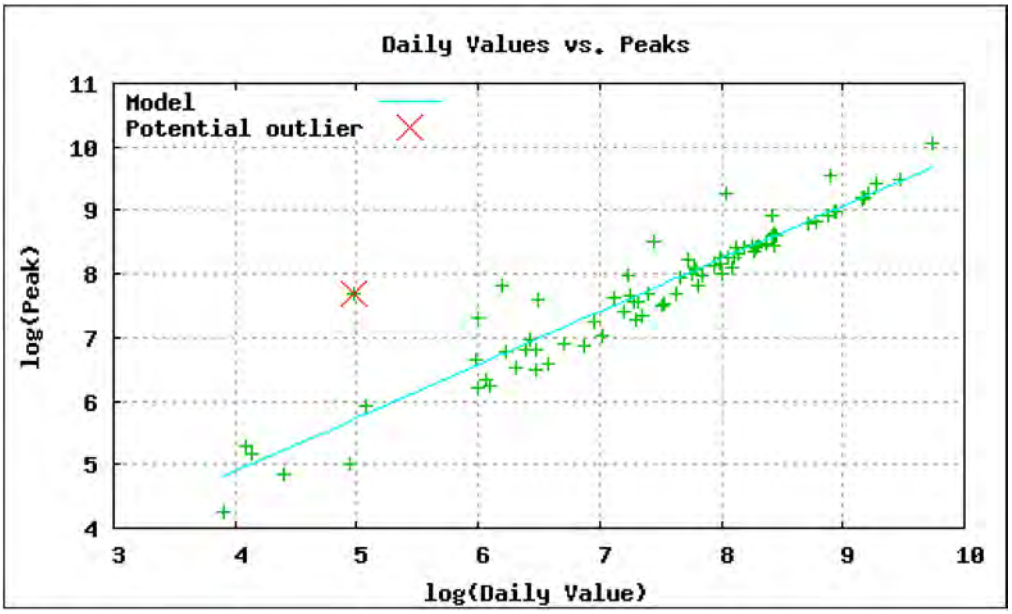


Figure 3. Data for a site with a relatively high degree of variability between the daily value and the peak streamflow.

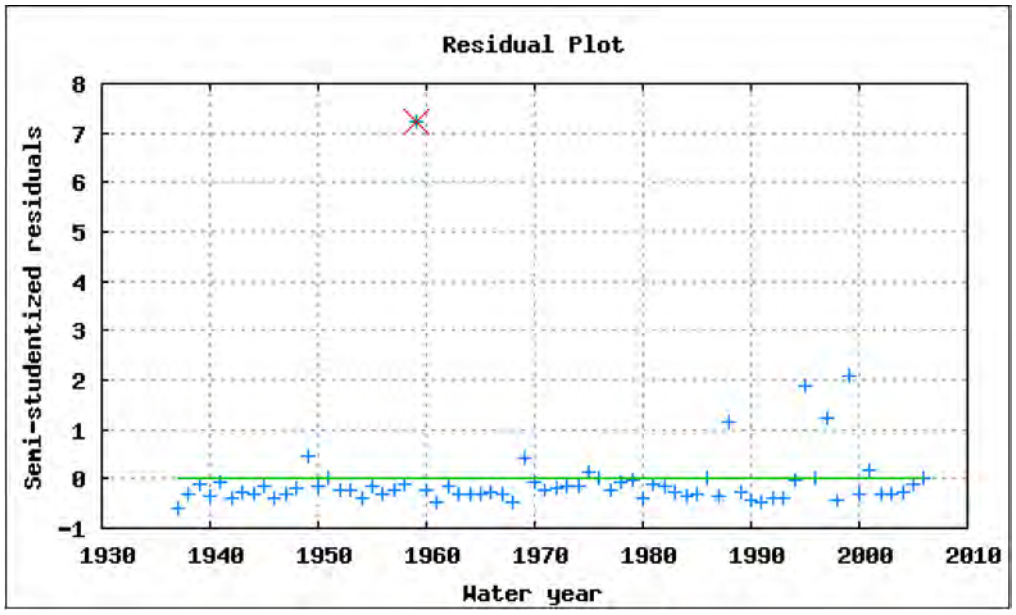


Figure 4. Plot of residuals by water year for site referenced in figure 2.

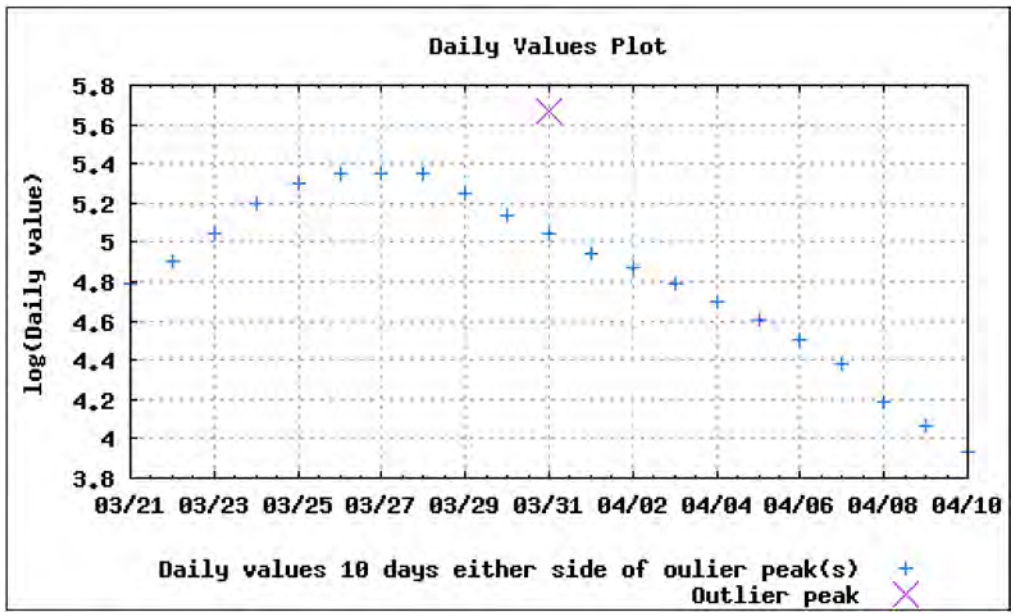


Figure 5. Plot of the daily values surrounding the peak identified as an outlier in figure 2.

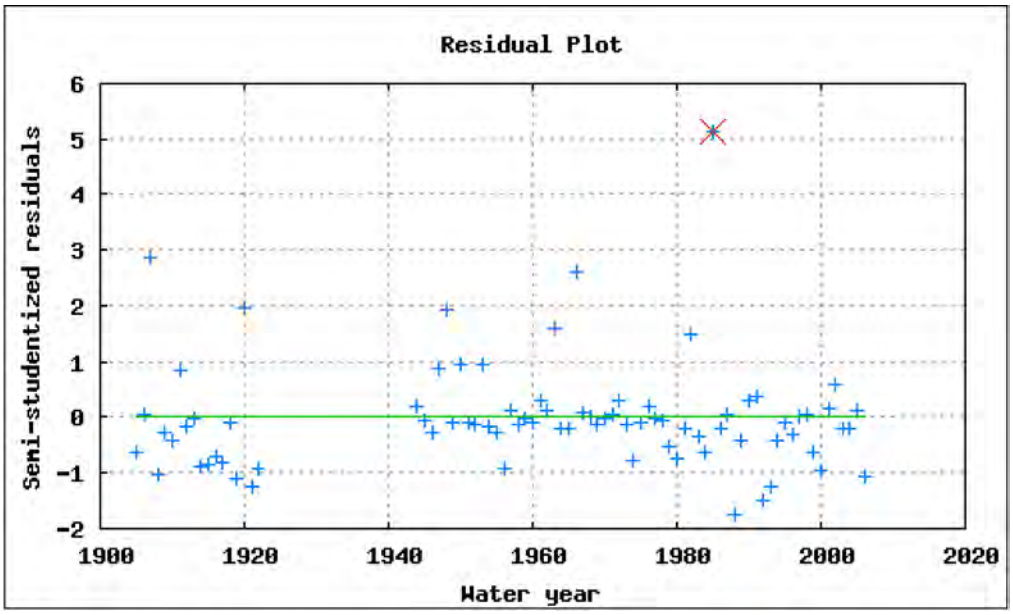


Figure 6. Plot of semi-studentized residuals by water year for site referenced in figure 3.

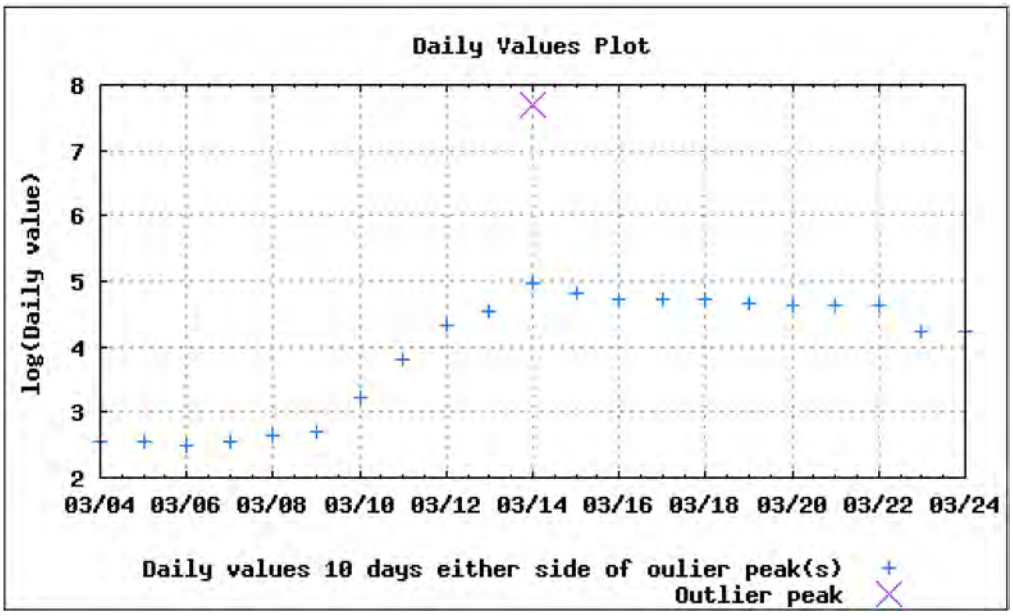


Figure 7. Plot of the daily values surrounding the peak identified as an outlier in figure 3.

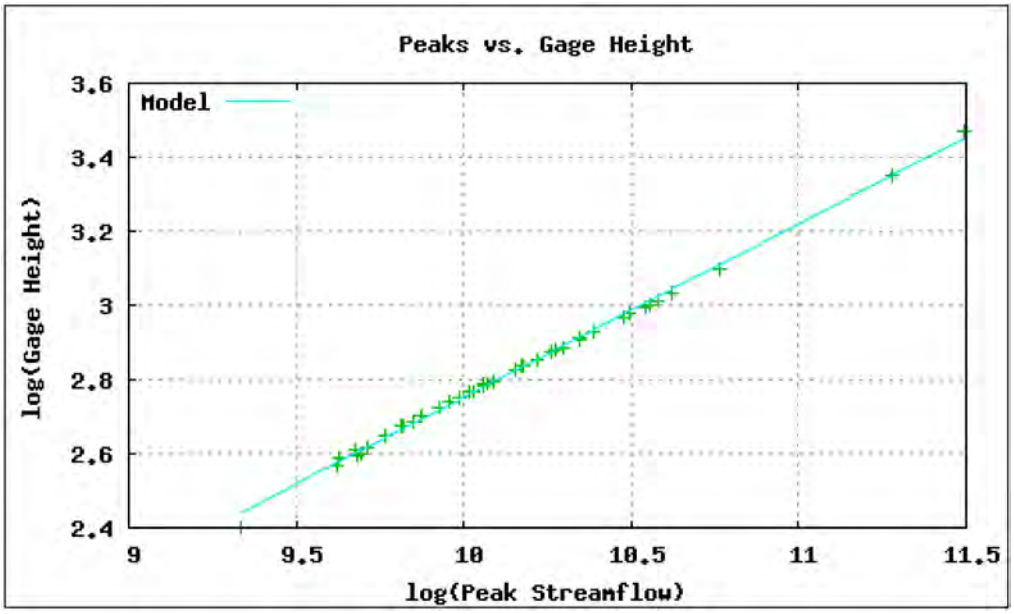


Figure 8. Data for a site with a strong linear correlation between peak streamflow and gage height.

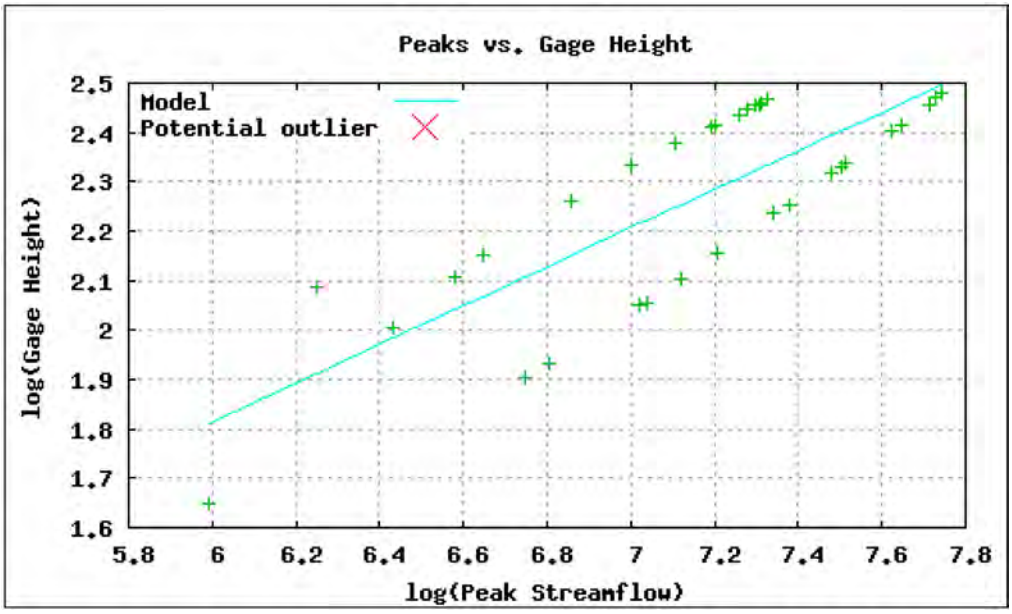


Figure 9. Data for a site with multiple linear relations between peak streamflow and gage height.

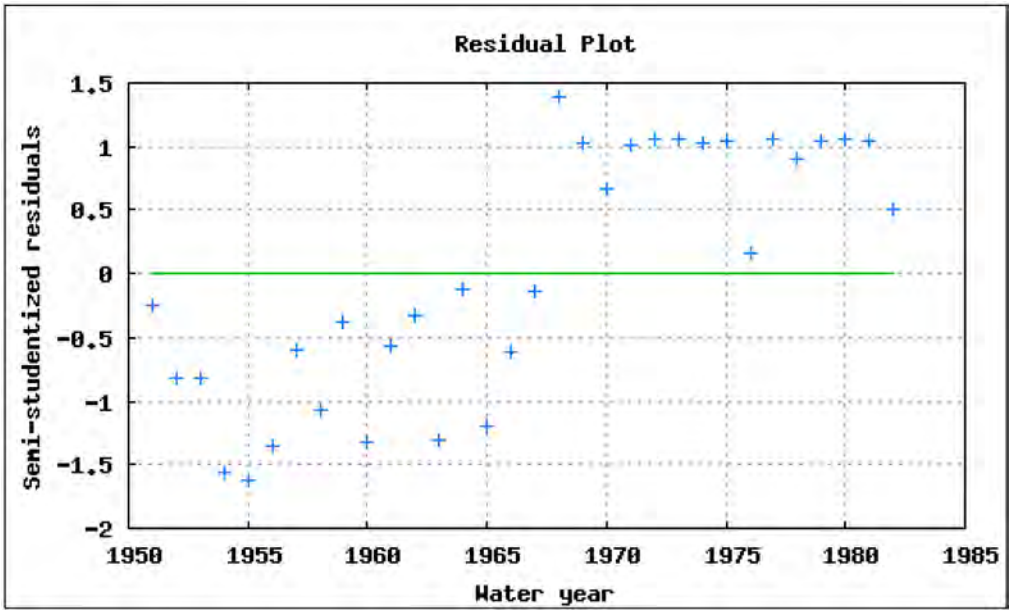


Figure 10. Semi-studentized residuals from regression of gage height on peak streamflow ordered by water year.

Figures 11 and 12 show data for a peak-flow site with no codes indicating changes in datum or regulation. The site exhibits a strong correlation between streamflow and gage height, with more variability at high and low streamflow (which is often the case). However, there is one point very different from the others and identified as an outlier in the upper middle of both plots. In examining the data in the PFF, it is suspected that the actual peak was 13,500 cubic feet per second (cfs) but was entered as 1,350 cfs. A review of the original records (such as the station descriptions and the report of the indirect measurement of discharge) might confirm this error. Once an outlier such as this is corrected, the regression relation may change if the test is rerun. The relation may improve or other points may be identified as outliers once an extreme outlier is corrected.

LRGHPA

This test identifies statistical anomalies from a regression of gage height on peak streamflow. This test may be the most time consuming for WSCs to address. Many of the events causing anomalies may be indicated in the qualification codes; however, spot checks of sites identified as having anomalies have discovered many without qualification codes. Identifying causes of anomalies not indicated by codes may require significant research, such as reading the station manuscripts. Addressing the LRGHPA sites will greatly improve the qualification information we provide with our data. However, the station anomalies, unless they are outliers that are corrected, will not go away and will be identified again in future runs of this test.

The linear regression and residual adjustment were done at the same time and in the same manner as explained in the LRGHP section. Further analysis of the residuals was performed to search for statistical anomalies, behavior that would be statistically improbable given a stable relation between peak flow and gage height. Outliers are also statistical anomalies, but were addressed in the LRGHP test.

The peak streamflow/gage height pairs of data from prior to the datum or regulation change and after the change are from separate populations. Normally, different populations are not mixed in developing a regression relation. However, that difference can be used to our advantage when examining the adjusted residuals.

Residuals in the residual plots should be randomly scattered above and below a centerline (y=0). Approximately the same number of residuals should be above the line as below, and the vast majority should be within ±3 studentized residual units of the centerline. Sites with 8 or more points in a row (consecutive years in a time series record of annual peaks) on one side of the centerline or with 6 or more points in a row steadily increasing or decreasing were identified as having statistical anomalies and were plotted. The test was performed only on sites that had at least 10 peaks.

The anomalies identified are usually caused by a change in datum or regulation, and the WSC should verify that the appropriate codes have been used for the peaks. Anomalies may also be caused by bed scour, by an outlier influencing the regression relation, or by changes in the stage-discharge rating.

Figure 13 shows data for a site where the gage datum was changed, resulting in two different populations of peak

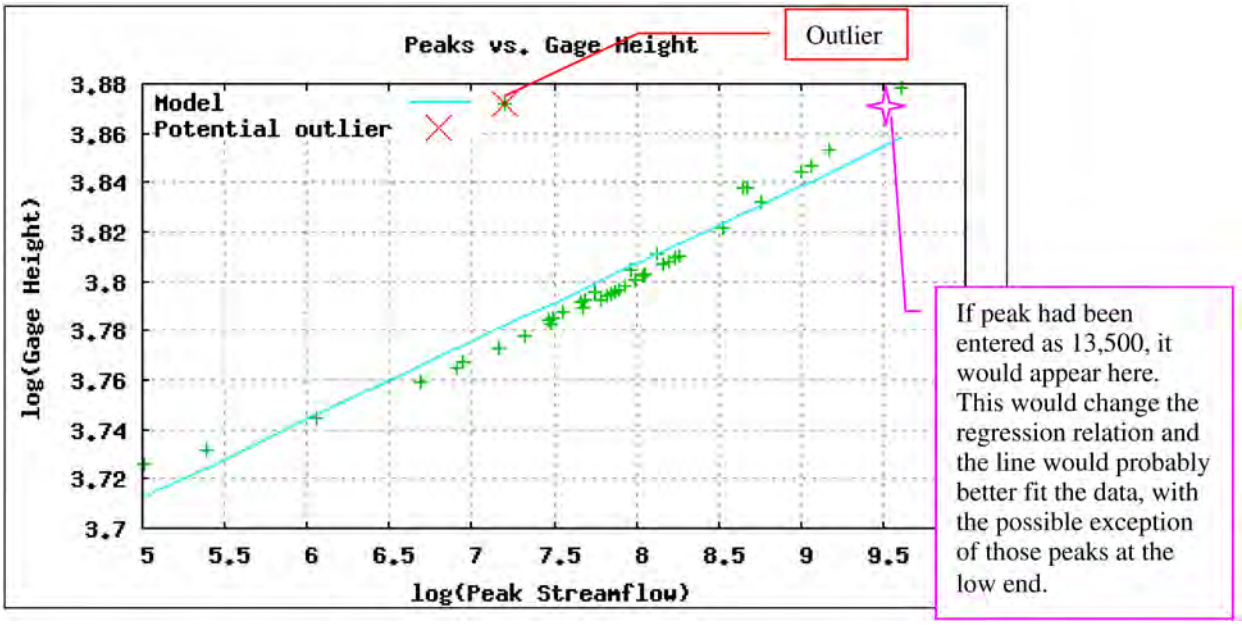


Figure 11. Outlier indicated on plot of regression of gage height on peak streamflow.

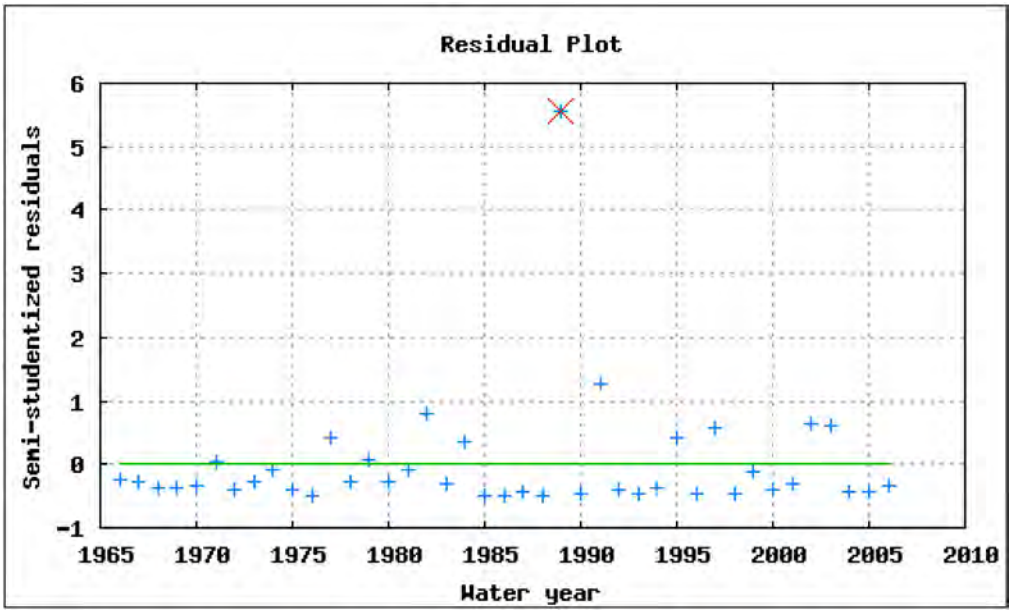


Figure 12. Outlier indicated on plot of semi-studentized residuals from regression of gage height on peak streamflow ordered by water year.

streamflow and gage height data pairs. The regression model line is an average of the two relations and does not represent either relation well. However, the plot of the residuals by water year (fig. 14) shows the time when the separation into two different populations of data occurred. The PFF should have code 6 for the gage height (gage datum changed during this year) for water year 1986—it did not in this case.

Figure 15 suggests that there may be more than two relations between peak streamflow and gage height over the history of the gage. Examination of the residual plot by water year (fig. 16) confirms this. There appears to be a change between 1950 and 1960 and an abrupt change in 1974. The station manuscript was examined and the following information explains figures 14 and 15 (U.S. Geological Survey, 2008).

> GAGE.--Water-stage recorder. Datum of gage is 930.00 ft above National Geodetic Vertical Datum of 1929. Prior to July 10, 1954, nonrecording gage 1.2 mi downstream at datum 30.00 ft lower. July 23, 1954, to Dec. 19, 1973, water stage recorder 2.7 mi downstream at datum 9.10 ft lower.
>
> REGULATION.--Flow regulated by temporary retention in ten retarding basins beginning 300 ft above station, four of which have slow release outlet structures to regulate the flow. Retarding basins were completed during the period l955 to l961 and have a combined capacity of l9,245 acre-ft.

The PFF codes for this site were examined. Regulation is indicated by code 6 applied in water year 1955 and continuing through the period of record. However, the gage datum changes in 1954 and 1973 are not indicated by the required gage-height qualification codes.

Information Reports

These reports are provided to give information about the data in the peak flow files. The information cannot be verified programmatically, so these are not considered "tests."

The first report lists peaks that do not occur within 3 days of the maximum daily value. This started out as a test but it quickly became evident that there were many, many peaks that did not occur within 3 days of the maximum daily value. Fixing some of the errors identified in other reports may result in changes to dates, peaks, and daily values that ultimately reduce the number of sites identified in this report.

The second report lists those peaks with qualification code 2 that are missing gage height. The remaining reports simply list which peaks were qualified by codes 3, 4, 8, 9, A, D, and E (table 1). WSCs should use their institutional knowledge to verify that peaks caused by known events like dam failures (code 3) and hurricanes (code 9) are properly coded.

It is important that the codes in the information reports are correct because the codes control what peaks are used in the statistical flood frequency analyses in PEAKFQ. Any changes should be documented.

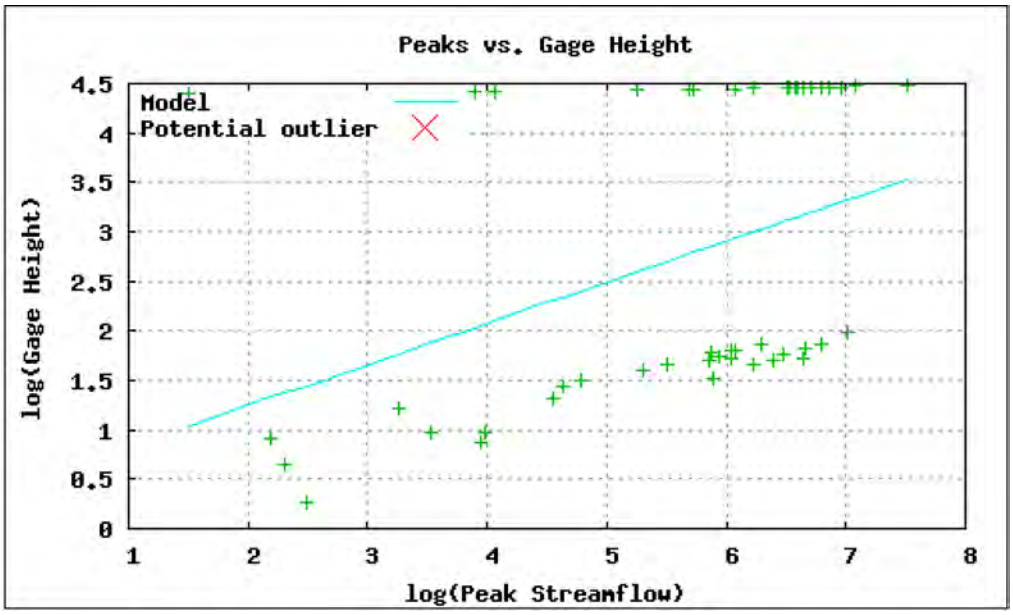


Figure 13. Plot of peaks versus gage height for a site identified as having a statistical anomaly (8 or more points in a row on one side of the centerline of the residual plot).

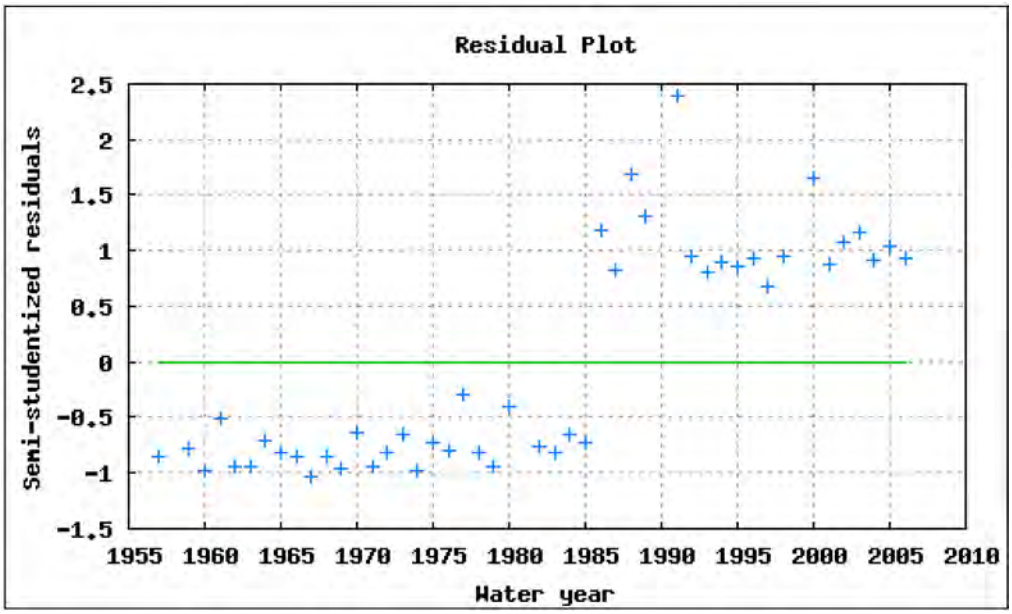


Figure 14. Plot of semi-studentized residuals from regression of gage height on peak streamflow ordered by water year for a site identified as having a statistical anomaly (8 or more points in a row on one side of the y=0 centerline of the residuals).

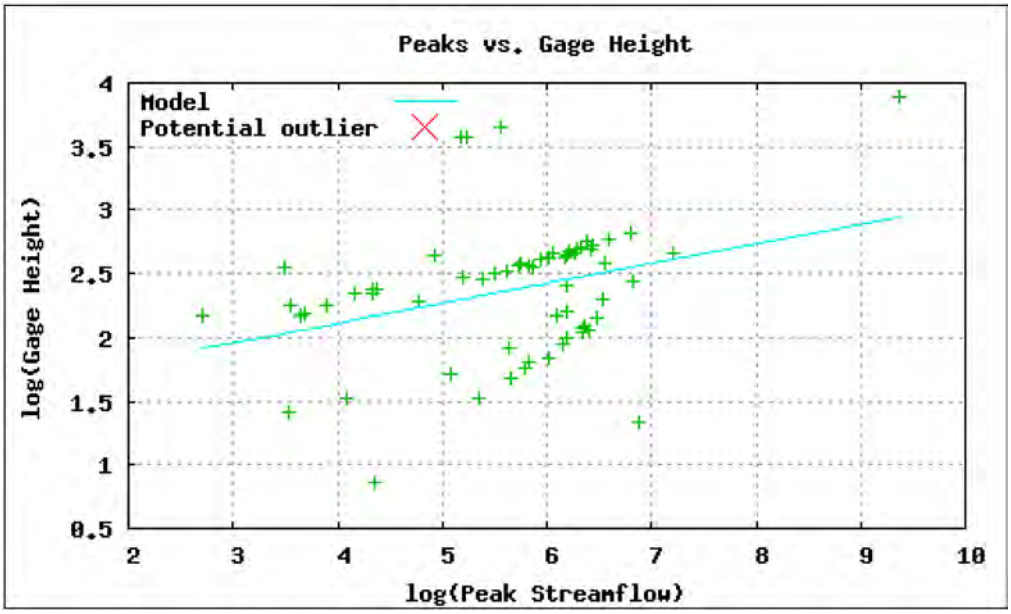


Figure 15. Plot of regression of gage height on streamflow for a site identified as having statistical anomalies (8 or more points in a row on one side of the y=0 centerline of the residuals and 6 or more adjusted residuals in a row increasing or decreasing).

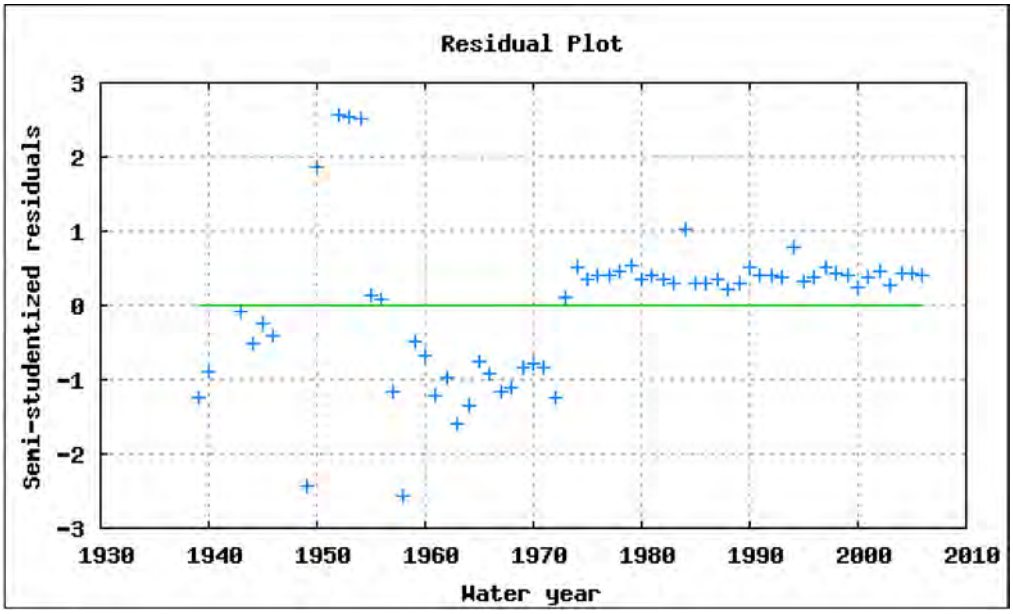


Figure 16. Plot of semi-studentized residuals from regression of gage height on peak streamflow ordered by water year for a site identified as having statistical anomalies (8 or more points in a row on one side of the y=0 centerline of the residual plot and 6 or more adjusted residuals in a row increasing or decreasing).

Computer Code

PFReports, the code to perform the tests and produce the reports was written in Perl and runs in Unix. Perl uses gnuplot (the proper name gnuplot is spelled with a lowercase g; *http://www.gnuplot.info/*) to produce the graphs. PFReports uses the following Perl modules,

- LWP::UserAgent,
- Date::Calc,
- Time::Local,
- Time::localtime,
- Statistics::Descriptive, and
- Statistics::LineFit,

and uses the utility ascii2pdf. Documentation for the Perl modules may be found by searching for them on the CPAN (Comprehensive Perl Archive Network) Search Site, *http://search.cpan.org/*. Regression tests were performed and compared in both S-Plus and in the Perl program using Statistics::LineFit to ensure consistent results.

Summary

The accuracy, characterization, and completeness of the U.S. Geological Survey peak-flow data drive the determination of flood-frequency estimates that are used daily to design water and transportation infrastructure, delineate flood-plain boundaries, and regulate development and utilization of lands throughout the Nation and are essential to understanding the implications of climate change on flooding. This document describes a computer program and its output that facilitates efficient and robust review of data in the USGS Peak Flow File (PFF) hosted as part of NWISWeb.

Previous efforts to identify problems with the PFF were time consuming, laborious, and often ineffective. This new program represents an effort to automate identification of specific problems without plotting or printing large amounts of data that may not have problems. In addition, the results of the checks of the peak-flow files are delivered on the World Wide Web with links to individual reports so that Water Science Centers can focus on specific problems in an organized and standardized fashion. Results of the checks for all peak-flow files are available at *http://nd.water.usgs.gov/internal/pfreports/*.

References Cited

Helsel, D.R., and Hirsch, R.M., 1995, Statistical methods in water resources: New York, Elsevier Science B.V., 529 p.

Neter, John, Kutner, M.H., Nachtsheim, C.J., and Wasserman, William, 1996, Applied linear statistical models (4th ed.): Boston, WCB/McGraw-Hill, 1,408 p.

Novak, C.E., 1985, WRD Data Reports Preparation Guide: U.S. Geological Survey Open-File Report 85–480, 331 p., accessed January 9, 2008, at *http://pubs.er.usgs.gov/usgspubs/ofr/ofr85480*

U.S. Geological Survey, 2008, Water-resources data for the United States, water year 2007: Water-Data Report US–2007, accessed January 9, 2008, at *http://wdr.water.usgs.gov/wy2007/search.jsp*

Publishing support provided by:
Helena Publishing Service Center

For more information concerning this publication, contact:
Director, USGS North Dakota Water Science Center
821 E. Interstate Ave.
Bismarck, ND 58503
(701) 250–7400

Or visit the North Dakota Water Science Center Web site at:
http://nd.water.usgs.gov