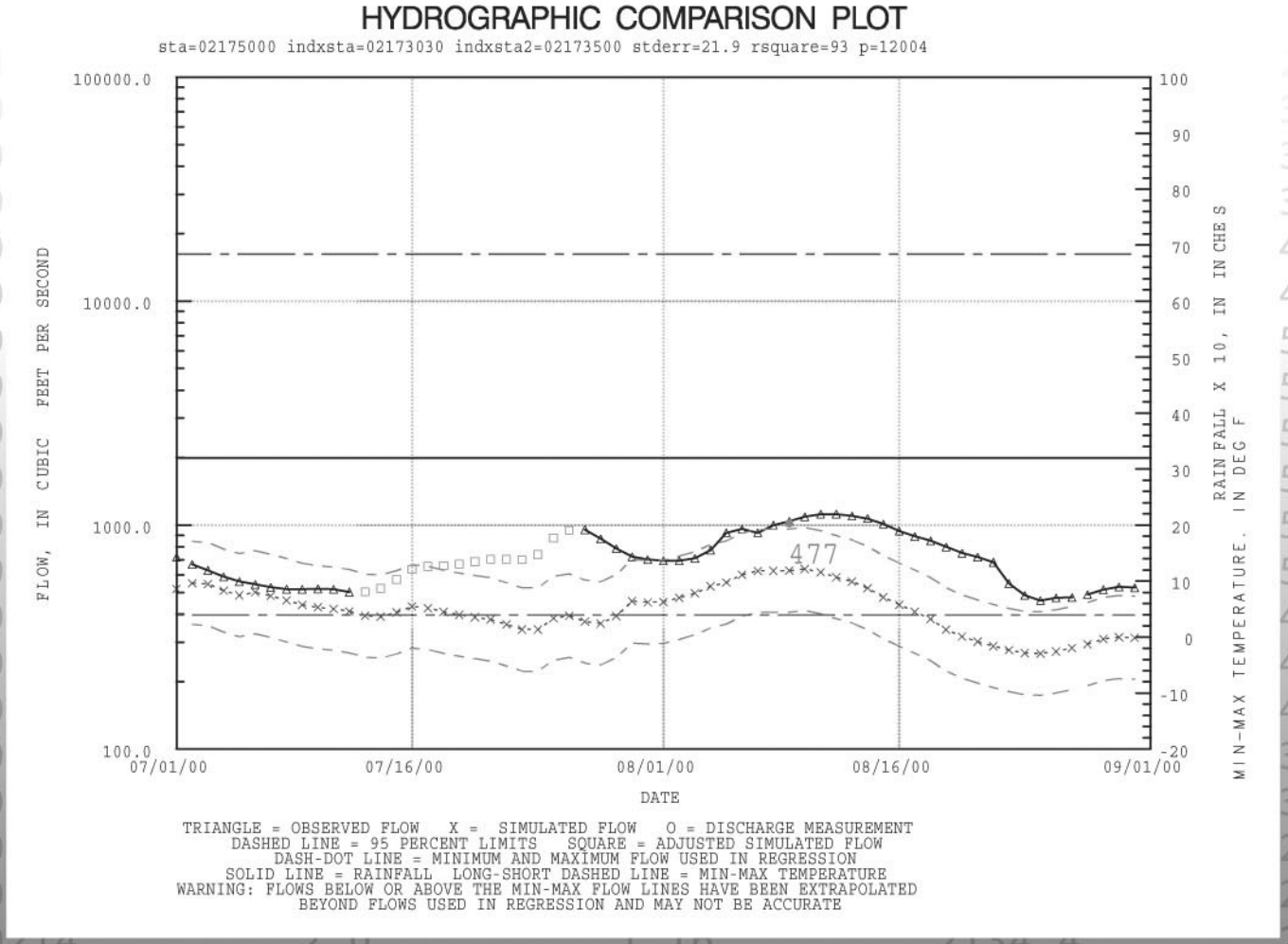
USGS
science for a changing world

Daily Values Flow Comparison and Estimates Using Program HYDCOMP, Version 1.0



U.S. GEOLOGICAL SURVEY
Open-File Report 02-286

Daily Values Flow Comparison and Estimates Using Program HYDCOMP, Version 1.0

By Curtis L. Sanders

U.S. Geological Survey

Open-File Report 02-286

Columbia, South Carolina
2002

U.S. DEPARTMENT OF THE INTERIOR
GALE A. NORTON, Secretary

U.S. GEOLOGICAL SURVEY
Charles G. Groat, Director

For additional information write to:

District Chief
U.S. Geological Survey
Suite 129
720 Gracern Road
Columbia, SC 29210-7651

Copies of this report can be purchased from:

U.S. Geological Survey
Branch of Information Services
Box 25286
Denver, CO 80225-0286
888-ASK-USGS

Additional information about water resources in South Carolina is available on the internet at **http://sc.water.usgs.gov**

CONTENTS

Abstract 1
Introduction 2
Purpose and scope 4
Statistical methods 4
Program description 8
Program execution 10
Selection of the five most-correlated index stations for every station in the state using Option 4 12
    Methods for determining the most-correlated index stations 16
    Data entry for Option 4 17
Displaying the five most-correlated stations for each station in the state using Option 5 19
Methods for Selecting the best index stations from the most-correlated index stations 20
Data entry, hydrographic comparison, and estimates of missing data using Options 1, 2, and 3 21
    How to enter data on screen 1 for Options 1, 2, and 3 22
    How to enter data on screen 2 for Options 1 and 2 28
    How to analyze tabular and graphic output from Options 1 and 2 29
Summary 46
References 46
Appendix 1: Program software components 48
Appendix 2: Program installation 51
Appendix 3: Program maintenance 52

FIGURES

1. Illustration of the HYDCOMP menu 9
2-4. Illustrations of OUT.FILE file reports produced by Option 4:
    2. Missing-data report in OUT.FILE file produced by Option 4 13
    3. Report of index station numbers, ADAPS data-descriptor numbers, standard errors of estimate, R-square, and regression coefficients in OUT.FILE file produced by Option 4 14
    4. Report of station numbers, index station numbers, ADAPS data-descriptor numbers, standard errors of estimate, and R-square in OUT.FILE produced by Option 4 15
5-6. Illustration of the data input form for:
    5. Option 4 18
    6. Option 5 19
7-8. Illustration of the output by Option 5 to the:
    7. Computer screen, tabulating station numbers, ADAPS DATA descriptor numbers, standard errors of estimate, and R-squares 19
    8. Printer, tabulating station numbers, ADAPS data descriptor numbers, standard errors of estimate, and R-squares 20
9-11. Illustration of:
    9. Screen 1 of the data input forms for Options 1, 2, and 3 21
    10. Screen 2 of the data input forms for Options 1-3 22
    11. The help information for Options 1-3 23-25
12-15. Illustration of a report listing:
    12. Date and time of run, station and ADAPS data descriptor numbers, station name, begin and end dates of regressions and plot periods, seasonal limits of regression, regression limits in terms of flow, and if flows were transformed to logarithms for Options 1 and 2 29
    13. Selected index station numbers, standard error of estimate, R-square, and regression coefficients for the regressions done by Options 1 and 2 30
    14. Date, log-bias, Kalman algorithm adjustment factor, unadjusted simulated flow, and adjusted simulated flow output by Options 1 and 2 to the printer 31

15. Date, index station numbers, log-bias, Kalman algorithm adjustment factor (adjust), unadjusted simulated flow (regressq), and adjusted simulated flow (final_q) output by Options 1 and 2 to the terminal screen 32
16-18. Illustration of plots of residuals output by Options 1 and 2 against:
16. Time period 34
17. Day of year 35
18. Best index station flow 36
19-25. Illustration of hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for:
19. September-October, 1999 39
20. November-December, 1999 40
21. January-February, 2000 41
22. March-April, 2000 42
23. May-June, 2000 43
24. July-August, 2000 44
25. September-October, 2000 45

TABLES

1. Guidelines for interpretation of regression standard error of regression 6

CONVERSION FACTORS AND ABBREVIATIONS

Multiply	By	To obtain
foot (ft)	0.3048	meter
cubic feet per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

ADAPS = Automated Data Processing System

OLS = Ordinary least-square

USGS = U.S. Geological Survey

Daily Values Flow Comparison and Estimates Using Program **HYDCOMP,** Version 1.0

By Curtis L. Sanders, Jr.

ABSTRACT

A method used by the U.S. Geological Survey for quality control in computing daily value flow records is to compare hydrographs of computed flows at a station under review to hydrographs of computed flows at a selected index station. The hydrographs are placed on top of each other (as hydrograph overlays) on a light table, compared, and missing daily flow data estimated. This method, however, is subjective and can produce inconsistent results, because hydrographers can differ when calculating acceptable limits of deviation between observed and estimated flows. Selection of appropriate index stations also is judgemental, giving no consideration to the mathematical correlation between the review station and the index station(s).

To address the limitation of the hydrograph overlay method, a set of software programs, written in the SAS macrolanguage, was developed and designated Program HYDCOMP. The program automatically selects statistically comparable index stations by correlation and regression, and performs hydrographic comparisons and estimates of missing data by regressing daily mean flows at the review station against -8 to +8 lagged flows at one or two index stations and day-of-week. Another advantage that HYDCOMP has over the graphical method is that estimated flows, the criteria for determining the quality of the data, and the selection of index stations are determined statistically, and are reproducible from one user to another.

HYDCOMP produces a file and list of the five most-correlated index stations for each station in the State, prioritized by standard error of estimate of the regression. Initially, HYDCOMP will load the most-correlated index stations into another file containing the "best-index stations," but will not overwrite stations already in the file. A knowledgeable user should delete unsuitable index stations from this file based on standard error of estimate, hydrologic similarity of candidate index stations to the review station, and knowledge of the individual station characteristics. Also, the user can add index stations not selected by HYDCOMP, if desired.

Once the file of best-index stations is created, a user may do hydrographic comparison and data estimates by entering the number of the review station, selecting an index station, and specifying the periods to be used for regression and plotting. For example, the user can restrict the regression to ice-free periods of the year to exclude flows estimated during iced conditions. However, the regression could still be used to estimate flow during iced conditions.

HYDCOMP produces the standard error of estimate as a measure of the central scatter of the regression and R-square (coefficient of determination) for evaluating the accuracy of the regression. Output from HYDCOMP includes plots of percent residuals against (1) time within the regression and plot periods, (2) month and day of the year for evaluating seasonal bias in the regression, and (3) the magnitude of flow. For hydrographic comparisons, it plots 2-month segments of hydrographs over the selected plot period showing the observed flows, the regressed flows, the 95 percent confidence limit flows, flow measurements, and regression limits. If the observed flows at the review station remain outside the 95 percent confidence limits for a prolonged period, there may be some error in the flows at the review station or at the index station(s). In addition, daily minimum and maximum temperatures and daily rainfall are shown on the hydrographs, if available, to help indicate whether an apparent change in flow may result from rainfall or from changes in backwater from melting ice or freezing water.

HYDCOMP statistically smooths estimated flows from non-missing flows at the edges of the gaps in data into regressed flows at the center of the gaps using the Kalman smoothing algorithm. Missing flows are automatically estimated by HYDCOMP, but the user also can specify that periods of erroneous, but nonmissing flows, be estimated by the program.

INTRODUCTION

Daily value flow records at stations being reviewed (hereafter referred to as "the review station") are generally verified and missing data are estimated by graphical comparison of logarithmic flow hydrographs within the U.S. Geological Survey (USGS). These comparisons are usually made at the time computations are being finalized, but could be done any time during the year. Traditionally, hydrographs are produced and placed on top of each other as "overlays" on a light table for comparison and so data can be estimated. Although valid, this method is time

consuming. Estimated flows are subjectively determined and can produce inconsistent results, because users may differ on the acceptable limits of deviation between observed flows and estimated flows. Selection of index stations is completely judgemental, without the added benefit of comparative statistics, such as the squared correlation coefficient (R-square) or a standard error of estimate. Once the hydrograph overlays are removed from the light table, there is no record of the hydrographic comparison.

To address the limitation of the graphical method, a set of software programs, using the SAS statistical package (SAS, 1993), was developed and designated Program HYDCOMP. The program automatically selects statistically comparable index stations and makes hydrographic comparisons and estimates of missing data by regression. Statistically similar index stations are selected using R-square values and standard errors of estimate from regression analyses. The selected index stations are automatically stored in the HYDCOMP database by the program.

Hydrographs are compared by relating flows at a station to -8 to +8-day lagged flows at one or two selected index stations and days of the week by regression, and by producing plots comparing the observed flows to the regressed flows and the 95 percent confidence limits of the regression. HYDCOMP uses a Kalman smoothing method to adjust the regressed flows to match non-missing flows at the edges of the gaps in flow data and simultaneously approach regressed flows at the center of large gaps of missing flow data. Flow-measurement, rainfall, and minimum and maximum temperature data are also plotted on the hydrographs. In addition, HYDCOMP plots percent residuals against time, season of year, and index station flow to assist in the analysis of the regression itself. For hydrographic comparisons and estimates of missing flow data, the user can limit regressions to seasonal periods of the year and/or upper and lower flow limits. For selection of index stations, the user can limit correlations to seasonal periods of the year (but not upper and lower flow limits) for stations affected by factors such as ice.

HYDCOMP uses files of daily-value flows and flow measurement data retrieved from the USGS Automated Data Processing System (ADAPS) (National Water Information System, 1997) database. Descriptions of the software components of HYDCOMP, as well as the installation and maintenance procedures of the program, are contained in the appendixes.

Although HYDCOMP and its documentation were written primarily for quality control and estimates of missing streamflow data, the program can be used for the same purposes for other types of daily value data, such as water-quality data or water-level data for rivers, estuaries, or ground water. Instructions for operating HYDCOMP and analysis of results are included in the data input screens of the program. An attempt was made to simplify the analytical methods as much as possible.

PURPOSE AND SCOPE

This report documents the data input, computational methods, output, and analytical methods for the computer program HYDCOMP. In addition, it documents the software components and details the installation and maintenance requirements of the program.

STATISTICAL METHODS

Regression analysis is a simple technique that mathematically relates one variable to other variables by using equations. The user only needs a rudimentary understanding of the measures of regression accuracy to use HYDCOMP; a detailed understanding of statistics or regression methodologies is not required. The practical use and interpretation of the statistical tools within HYDCOMP are described in detail in the following sections.

A major benefit of estimating flows using regression analysis, compared to hydrograph overlays, is that regression results are consistent from user to user, whereas graphical results may not be. Regression also has the added benefit of mathematically quantifying the accuracy of the computed results.

The ordinary least-square (OLS) stepwise regression method is used in HYDCOMP to select the most significant explanatory variables (-8 to +8 lagged flows at 1 to 2 index stations and/or day-of-week) to be used in a prediction equation for flow at the review station (the response variable). A -8 lagged index station flow is the daily flow at an index station 8 days preceding the daily flow at the review station. Stepwise regression analysis (not the user) selects the most-correlated explanatory variables at the 5 percent acceptance level. The program prints out the regression equation coefficients, which is useful only for determining if the lags selected by the regression seem reasonable compared to lags estimated by the user.

The use of lagged flows in the regression has the effect of moving the regressed hydrograph forward or backward in time, and forcing the regressed hydrograph shape to match the shape of the hydrograph at the review station. Use of lagged flows from two index stations will improve the accuracy of the equation somewhat if the stations are not on the same stream and can improve the accuracy considerably if the two index stations are tributaries to the stream at the review station.

The regression also accounts for statistical differences because of weekend power cycles at regulated stations, as indicated by day-of-week variables W1 through W7. These variables equal either zero or one and are established for each day of the week. For example, if the day is Sunday, W1 is given a value of one and W2 through W7 are given values of zero. The effect is to adjust the intercept of the regression of the equation if one or more days of the week have a significant effect on the regression.

Flows estimated by regression techniques may be larger or smaller than observed flows for long periods of time because of the limited ability of regression analysis to simulate the dynamics of the stream. Therefore, large discontinuities could exist between non-missing flows and regressed flows at the edges of gaps in measured data. To remedy this problem, HYDCOMP uses a Kalman algorithm (Grewal and Andrews, 1997) that statistically adjusts regression flows used for estimating missing data to match non-missing flows at the edge of the missing data while, simultaneously, approaching the regression flows toward the center of the period of missing data. The Kalman algorithm is one of the most widely used techniques for smoothing data, and efficiently combines information about the physical process being modeled and measurement-device dynamics with a statistical description of the process and measurement errors. The relation of the non-missing flows to the regression flows at the edges of the gap in the data may have little to do with estimated flows toward the center of a large gap in data. In the Kalman smoothing method, differences between non-missing flows and regressed flows are diminished with distance from the edges of the gap according to the size of the gap, the coefficient of a serial correlation of the residuals from the OLS regression, and the variance statistic from the serial correlation procedure. Adjustments were not made to maintain the mean or variance of the observed data at the review station for logarithmic regression[1].

Adjusted estimated flows for missing data are automatically tabulated by HYDCOMP. If data are erroneous, but not missing, the user can still request tabulation of estimated flows. Otherwise, the estimated flows are not tabulated on the assumption that visual inspection of the hydrographs suffices to accomplish the graphic quality control.

---

[1]Ordinary least-squares (OLS) regression maintains the mean of the independent variable (flows at the review station), but not the variance of the independent variable. However, if flows are transformed to logarithms before regression to maintain homoscedascity of the data about the regression line (variance), the mean of the antilogs of the predicted logarithms of flow will not equal the mean of the arithmetic flows that are input to the regression. A “log-bias” adjustment could be made to the logarithmic regression equation to maintain the means of the input flow data. The larger the standard error of estimate, the larger the log-bias correction would need to be.

Whether the data are transformed to logarithms or not, the OLS method of regression will not maintain the variance of the input flow data. Hirsch (1982) suggested the Maintenance of Variance Extension (MOVE.1) method for adjusting the regression equation to maintain the variance of the input flow data. William Kirby (U.S. Geological Survey Office of Surface Water, oral comm., September 2001) gives the following explanation for not making adjustments to maintain the mean or variance of the observed flow data at the review station: “These two adjustments are usually made to OLS regression equations used for simulating long periods of record or surrogates for long periods of record for use in statistical analyses. For water-volume studies, such as planning storage capabilities for reservoirs, the log-bias adjustment is made to maintain average volumes of flow. For low-flow or high-flow frequency studies, the MOVE.1 method is made to maintain the variance of the flow. An OLS regression, without these two adjustments, provides the best estimate of flow on any one day or short period of days, because for any unique combination of explanatory variables, half the observed data will be larger than the regression value and half the observed data will be less than the regression value.”

Flows at most USGS gaging stations are usually estimated for fairly short periods of missing data, and not for long periods of record where the log-bias or MOVE.1 adjustments would be applicable. However, 2 to 4 months of flow data may be estimated by regression each year during periods of backwater from ice flow in northern States. For those stations, it would be desirable to maintain both the mean and variance, but it is impossible to maintain both at the same time. The USGS is tasked to estimate the most suitable flow for any one day without adjustments to maintain the mean or variance of input flows (William Kirby, U.S. Geological Survey Office of Surface Water, oral comm., September 2001). Therefore, no adjustments, other than Kalman smoothing, were applied to the regressed flows in HYDCOMP. However, log-bias adjustments are tabulated by the program to show the magnitude of the adjustment factor if very long periods are estimated by HDCOMP.

Residuals are the percentage differences between the regressed flows and the observed flows. In HYDCOMP, if the residuals are positive, the regressed flows are larger than the observed flows. The program plots percent residuals against time, season of the year, and index station flow. These percent residuals should scatter uniformly about the zero percent line on these plots for the regression to be valid. If the scatter is not uniform about the zero percent line, there is nothing the user can do about it; the plots just warn the user that the regression equation is not as good as it should be.

The user must have a basic understanding of the usefulness of regression standard error of estimate and R-square for evaluating the accuracy of the regression analyses as well as a basic understanding of the use of the regression 95 percent confidence limits for making hydrographic comparisons. How these values are used to evaluate HYDCOMP output is important to the user, not how the values are computed.

The regression standard error of estimate is a measure of the central scatter of the regression. Suppose that the standard error of estimate were 15 percent. This means that the hydrologist is about 95% sure that about 2/3 of the observed values fall within 15 percent of the regressed values. The smaller the standard error, the better the regression. Therefore, the standard error of estimate is used to compare the accuracy of the regression equations to the associated index stations, as summarized in table 1. Index stations having smaller standard errors of estimates are more suitable for use in regression than those having larger standard errors. The ranges of standard error and purported "usefulness" in table 1 are entirely arbitrary and subject to individual judgement. Table 1 is presented only as a guide.

Table 1. Guidelines for interpretation of regression standard error of regression

Range of Regression standard error of estimate (percent)	Usefulness for estimating flow	Usefulness for hydrographic comparison
0 - 15	Good	Excellent
15 - 30	Fair	Good
30 - 50	Probably not usable	Fair to poor
>50	Not usable	Not usable

In South Carolina, most best-correlated index stations have standard errors of estimate in the 15-30 percent range. For stations in close proximity on the same stream, the standard error may be less than 10 percent. In the arid areas of the western and central United States and in the swampy, highly regulated areas of South Florida, the standard errors of many stations may be much greater than 40 percent and unusable for hydrographic comparison or estimating missing data. An advantage of estimating missing data using the regression method over the graphical method is that the regression method numerically quantifies the accuracy of the estimated data,

whereas the graphical method does not. The user should not select index stations based on standard error of estimate alone, because the standard error may be small entirely by chance. Therefore, the user should verify that selected index stations are hydrologically similar.

The standard errors of estimate could suggest that stations not previously considered as index stations are usable as well as suggest that stations presently considered usable are not. If two stations are hydrologically similar, the standard error of estimate can help determine the more suitable station. Also, the standard error of estimate numerically quantifies a station's suitability as an index station and, therefore, removes some of the subjectivity of the index station selection process.

The squared correlation coefficient is sometimes referred to as the coefficient of determination, but is referred to as "R-square" in this report. R-square indicates the proportion of the total variation in the response variable (streamflow at the review station) that can be accounted for by changes in the explanatory variable(s) (in this case, flows at an index station and day-of-week). For example, an R-square of 90 percent would indicate that 90 percent of the variation in streamflow at the review station can be explained by the lagged flows of the index station(s) and day-of-week. Regressions having larger R-squares are more accurate than regressions having smaller R-squares; therefore, both R-square and the standard error of estimate can be used to compare regressions. The accuracy of a regression is easier to visualize in terms of standard error of estimate rather than the square, because one can visualize the standard error in percent. However, R-square can warn that a regression could be totally unreliable, such as when R-square is less than about 60 percent.

The 95 percent confidence limits of the regression represent the outer limits of the error associated with the regression, compared to the standard error of estimate, which is a measure of the central error. The 95 percent limit is quantified as flow on the daily value hydrographs by HYDCOMP. If flow at the review station is larger or smaller than the 95 percent confidence limit flow for prolonged periods as shown by the hydrographs produced by HYDCOMP, then something may be wrong with either the flows at the review station or at the index station(s) being used in the regression. As implied by the "95 percent," it is expected that some observed flows could fall outside the 95 percent confidence limits, but still be correct. Therefore, the user can usually ignore very short periods during which observed flows plot outside the confidence limits, particularly during flood events, where the regression does not accurately describe the streamflow dynamics. However, if the 95 percent confidence limits show the possibility of error, the user must investigate the computation of flows at both the review and index stations to determine which station has the possibly erroneous data as described later in this report.

The 95 percent confidence limits can even be used to determine when shifts should be applied, particularly where confidence limits are very narrow. However, the limits are usually so large that hydrographic comparisons are useful mostly for detecting large, catastrophic errors. The user should realize that even though the observed flows at the review station fall within the confidence limits, errors of a magnitude lesser than the confidence limits may still exist in the flow data.

PROGRAM DESCRIPTION

The HYDCOMP menu (fig. 1) summarizes the capabilities and operation of the program. The sequence of operations discussed below in conjunction with the HYDCOMP menu illustrates the suggested overall sequence of program operation.

(1) In option 4, one knowledgeable user (not each successive person using HYDCOMP within a District) correlates flows at each currently operated station in a District database against flows at every other currently operated station in the District database to create a most-correlated index station file and a printed list showing the five most-correlated index stations for each station in the District. This must be done only when initializing the file, when new stations are activated, or when hydrologic conditions change drastically. Currently (August 2002), stations in other States can be incorporated in option 4 only by copying the flow data from the neighboring State into the local database.

The user should understand the difference between the "most-correlated index station" file (file **best5sta.sas7bdat)** produced by option 4 and the "selected-index station" file (file **stadat.sas7bdat**) associated with options 1, 2, and 3. For each review station, both files contain lists of index station numbers and dd numbers. The files also contain the dd numbers of the review stations, and standard error of estimate and R-square associated with each index station. However, the index stations in the most-correlated index station file were selected only by statistics in option 4, whereas the index stations in the selected-index station file are final selections by the user based on both statistics and hydrologic similarity.

The list of stations in the most-correlated index station file can be viewed for any one station using option 5 described below. Option 4 will automatically populate both the selected-index station file (if empty the first time option 4 is run) and the most-correlated index station file. However, option 4 will not overwrite the index stations already stored in the selected-index station file on the assumption that the user may have already made a manual selection of index stations. After an initial population of the selected-index station file by option 4, knowledgeable users should eliminate the stations which are not hydrologically suitable or too inaccurate for use.

(2) After option 4 has been run to populate the most-correlated index station file, other users can use option 5 to list the five most-correlated index stations on the computer screen or printer for a particular review station. If option 4 is rerun for another period of time or with a different definition of seasonality, it will not overwrite the selected-index station file.

```
PROGRAM HYDCOMP - Last modified Aug. 2, 2001 by C. L. Sanders

Note: Program HYDCOMP does hydrographic comparisons and estimates
      daily values data by regression. Tutorial screens describe
      how to enter data and analyze output tables and plots.

      Up to 5 sets of index stations and dd numbers are permanently
      stored. Therefore, it is not necessary to retype this
      information with each run.

RETRIEVE OR ADD ONE STATION ONLY; REGRESS AND PLOT THE STATION USING:
  (Note: you can no longer add multiple stations or delete stations
         using options 1 or 2. Use option 3 instead. Multiple users
         can now use options 1 and 2 with minimal chances of collisions.)

   1 - ONE INDEX STATION

   2 - TWO INDEX STATIONS

 ---------------------------------------------------------------------
   3 - ADD, DELETE, OR MODIFY STATION/INDEX STATION DATA FOR MORE THAN
       ONE STATION AT THE TIME.
        (Note: it is now necessary to use option 3 to delete stations
               or to add more than one station at the time.
               Warning - when using option 3, you will lock everybody
               else out of hydcomp.)

 ---------------------------------------------------------------------

   4 - CREATE A PERMANENT FILE OF THE 5 MOST CORRELATED INDEX STATIONS
       FOR EACH STATION IN THE STATE FOR USE IN OPTION 5 TO HELP IN
       SELECTING INDEX STATIONS.
       (Note: Must have already created an ADAPS group file containing
              all the active stations in the state. Not necessary to run
              to run this option to run options 1-3. If used, only needs
              to be rerun a year or so after new data have been collected
              at new stations or if hydrologic conditions have changed.)

 ---------------------------------------------------------------------

   5 - DISPLAY ON THE SCREEN THE 5 MOST CORRELATED INDEX STATIONS FOR A
       SINGLE STATION AFTER OPTION 4 HAS BEEN RUN ONCE.

 ---------------------------------------------------------------------


NOTICE - NOTICE - NOTICE: IF THE ABOVE MENU SEEMS CHOPPED OFF,
      ENLARGE THE SCREEN VERTICALLY!!!

SELECT ONE OF OPTIONS 1-5: 2


NOTE: When you plot to the printer, the tables will also be printed.
      When you plot to the screen, the tables will not be
      printed, but can be printed, by executing the command:

                        lp -y landscape print.file

 1 - PLOT TO SCREEN

 2 - PLOT TO PRINTER

SELECT: 2

NOTE: Unable to open SASUSER.PROFILE. WORK.PROFILE will be opened instead.
NOTE: All profile changes will be lost at the end of the session.
20020525 113559 Processing control file: rdbcntl.upcase

ENTER NAME OF PRINTER: csa
request id is csa-37427 (1 file(s))
request id is csa-37428 (1 file(s))
request id is csa-37429 (1 file(s))
request id is csa-37430 (1 file(s))
request id is csa-37431 (1 file(s))
FINISHED PROGRAM.
```

Figure 1. The HYDCOMP menu.

The selected-index station file can only be overwritten manually, using options 1, 2, or 3 to enter the index stations selected from the list produced by options 4 or 5 or by hydrologic judgement.

(3) Option 1 is used to select, add, or modify review stations one at a time in the selected-index station file and to produce residual plots, hydrographic comparison plots, and tables of estimated flow data using flows at one index station. Option 1 cannot be used to delete review stations or add more than one review station at a time to the selected-index station file. In the most common usage, the user will select a review station using option 1, select an index station or add an index station associated with the review station, and produce the desired plots and tables. In another usage, the user can add a new review station and its associated index stations, and then produce the desired plots and tables.

(4) Option 2 is the same as option 1, except that it uses lagged flows at two index stations in the regression rather than one index station.

(5) Option 3 is used to delete, add, or modify more than one review station at a time in the selected-index station file. Review stations can be deleted from the file only by using option 3. Option 3 will not do hydrographic comparisons or estimate missing data. In option 3, a user can manually add or modify many review stations with their associated index stations at one time.

PROGRAM EXECUTION

To execute the HYDCOMP program, the user must be a member of the NWIS user's group with access to ADAPS. Type in lower case the command shown below:

hydcomp.sh

As can be seen by the menu in figure 1, the user has the opportunity to select from the options described earlier as well as decide whether the output is sent to the screen or printer.

Once the option number and output location are selected by the user, screen forms will appear which permit the user to enter and change information. To exit the screen and execute the program, the user clicks on "File" at the top of the screen, and then "Close" as described below. When plots come to the screen, the user can proceed to the next plot by clicking on the plot with the left mouse button. Backing up through the plots is not possible. The following seven instructions describe how to utilize the screen form and how to get on-line instructions.

1. **To enter data on the forms for all options**: Arrow or tab to the desired position. Alternatively, click on the desired position using the left mouse button. Type in the desired data. Delete using the backspace key.

2. **To change screens for options 1-3**: Click on "View" at the top of the screen and then either "Next Screen" or "Previous Screen." Several screens for data input and on-line documentation may be provided for each station.

3. The selected-index station database contains many review stations with up to five index stations per review station. **To find a review station for options 1-3**: Click on "Search" at the top of the screen, and then on "Where." A box will appear on the screen. To select a review station (02175000, for example), type the following command in the box, including the single quotation marks:

   sta = '02175000'

   If the review station is in the database, its information will appear in the screen form. If a wrong station was retrieved, or another station is to be selected, the user must first click on "Search" and then click on "Undo Last Where" before another review station can be selected with a "search-where" sequence.

4. **To delete a review station for option 3 only**: Find the review station as described in (3) above. Then, click on "Edit" at the top of the screen, and "Delete Record."

5. **To add a new review station for options 1-3**: Try to find the review station as described in (3) above. If it is found, the review station is already created, and another record for the review station should not be created. (A "record" is a single row of data in the selected-index station file containing the review and index station numbers, dd numbers, station names, dates, and so forth.) If it is not found, click on "Edit" at the top of the screen, and then "Add New Record." A blank screen will appear into which new data can be entered. Entering more than one new review station at a time must be done using option 3.

   Caution: If a user types over an existing review station in the file, the data for the original station is lost. Therefore, it is paramount that users understand how to add new stations as described above.

6. **To exit the screen form and execute the program for all options**: Click on "File" at the top of the screen, and then "Close." The program will produce an output table of estimated data to the screen. Exit this table by clicking on "File" and "Close."

7. **For on-line documentation of the program**: Click on "View" at the top of the screen and then on "Next Screen" while in options 1, 2, or 3, to page down through the screens. Screens will appear that describe how to utilize the screen forms, what data to enter on the forms, and how to analyze the tables and plots produced by the program.

Normally, the default size of SAS plots to the screen are too small for visualization of the plotted data. If so, each user must change the screen size by the following procedure:

1. The user must try to access a file named **.Xdefaults** using an editor in the user's home directory. If the file does not exist, the user must create a file by that name in the home directory, preserving the punctuation and case of **.Xdefaults** as shown.

2. Add the following lines of code to the file, maintaining punctuation and case as shown. The "100" is the percentage size screen desired. This can be modified to whatever size the user desires.

   SAS.windowHeight: 100
   SAS.windowWidth: 100
   SAS.windowUnitType: percentage

3. After adding the above lines to the **.Xdefaults** file, the user must log out of the computer and log back in before the modification will be implemented by SAS.

SELECTION OF THE FIVE MOST-CORRELATED INDEX STATIONS FOR EVERY STATION IN THE STATE USING OPTION 4

Option 4 selects the five most-correlated index stations for every station that will be reviewed in the State and initially populates the most-correlated index station file and selected-index station file with the five most-correlated index stations for each review station in the District. Option 4 will not overwrite index stations that already exist in the selected-index station file. The user can then use options 1 or 2 to select the most-correlated and hydrologically similar index stations for one review station at a time, or use option 3 to select index stations for many review stations at a time. The selection is done by deleting the index stations that are too poorly correlated for use or that are hydrologically dissimilar, and possibly by adding hydrologically similar index stations not selected by option 4 as discussed in a later section of this report.

If option 4 is rerun using a different period record or different seasonal regression period, it will not overwrite the index stations already entered in the selected-index station file by a previous run of option 4 or by hand. Therefore, selections of index stations from the newly determined best-correlated index stations must be entered in the selected-index station file by one of the following methods.

(1) Option 4 produces a file named "**out.file**" in the user's directory and a printed list (figs. 2-4) containing a listing of missing data, regression information, and a listing of the five most-correlated index stations for every station in the State, respectively. The user can prevent **out.file** from subsequently being overwritten by HYDCOMP by renaming it to something like "CORRELATED.STATIONS.FILE." This file and list are made available so that the user can select the most-correlated and hydrologically similar index stations for many stations at a time for entry into the selected-index station file using option 3.

MISSING DATA REPORT
11:59 Saturday, May 25, 2002 1

Obs	STATION NUMBER	NUMBER OF MISSING DAYS FOR RETRIEVAL
1	02110500	0
2	02110704	777
3	02110802	703
4	021108044	420
5	02130561	2
6	02130900	0
7	02130910	13
8	02131000	0
9	02131010	0
10	02131472	0
11	02132000	0
12	02135000	0
13	02135210	365
14	02135300	0
15	02135520	1770
16	02136000	0
17	02136361	0
18	02145642	1113
19	02146000	0
20	02146820	1686
21	02147020	1
22	02147403	1598
23	02147500	0
24	02148000	0
25	02148315	168
26	02153200	360
27	02153551	759
28	02153680	1704
29	02153780	0
30	02153800	1506
31	02154500	0
32	02154790	0
33	02155500	365
34	021556525	0
35	02156050	0
36	021563931	1529
37	02156500	15
38	02157510	1229
39	02158408	1600
40	02160105	0
41	02160200	516
42	02160326	0
43	02160381	0
44	02160390	0
45	02160700	0
46	02161000	7

Figure 2. OUT.FILE file reports produced by Option 4 missing-data report in OUT.FILE file produced by Option 4.

REGRESSION COEFFICIENTS

11:59 Saturday, May 25, 2002 4

sta=02110500

Obs	indxsta	stderr	rsquare	coeff	llag8	llag7	llag6	llag5	llag4	llag3	llag2	llag1	lindxq	lahd1
1	02135000	77.6	74.88	0.040	.	.	0.332	.	.	.	.	.	0.693	.
2	02135300	131.6	41.63	29.596	0.381	.	.	0.277	.	.	.	0.394	.	.
3	02132000	136.4	38.45	0.909	0.322	.	.	.	.	.	0.278	.	0.446	.
4	02175000	138.9	36.75	0.342	.	.	.	.	.	.	.	.	1.055	.
5	02136000	145.5	32.57	44.690	-0.324	.	0.316	.	.	.	.	.	0.608	.

Obs	lahd2	lahd3	lahd4	lahd5	lahd6	lahd7	lahd8	w1	w2	w3	w4	w5	w6	w7
1	.	0.274	.	.	.	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	-0.202	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	-0.113	.	.	.	.	.	.	.

sta=02130561

Obs	indxsta	stderr	rsquare	coeff	llag8	llag7	llag6	llag5	llag4	llag3	llag2	llag1	lindxq	lahd1
6	02131000	23.7	94.31	0.436	0.028	.	.	.	.	-0.117	-0.054	.	-0.510	1.715
7	02131010	25.3	93.55	0.314	0.042	.	.	.	.	-0.156	.	.	-0.667	1.804
8	02148000	57.8	68.90	1.031	.	.	-0.088	.	.	.	0.190	0.354	0.320	0.162
9	02132000	59.0	67.79	15.332	0.132	.	.	.	.	-0.191	.	.	0.391	.
10	02130910	61.5	64.98	7.386	.	.	.	.	.	.	.	.	1.022	.

Obs	lahd2	lahd3	lahd4	lahd5	lahd6	lahd7	lahd8	w1	w2	w3	w4	w5	w6	w7
6	-0.056	.	.	0.063	.	.	.	.	.	.	.	.	.	021
7	.	.	.	0.084	.	.	.	-0.017	.	.	.	.	.	0.017
8	.	.	.	.	.	.	0.109	-0.172	-0.225	-0.041	.	.	.	.
9	.	0.261	.	0.263	.	0.507	-0.464	-0.199	-0.248	-0.052	.	.	.	.
10	0.246	.	.	.	.	.	.	-0.199	-0.247	-0.053	.	.	.	.

Figure 3. OUT.FILE file reports produced by Option 4 report of index station numbers, ADAPS data-descriptor numbers, standard errors of estimate, R-square, and regression coefficients in OUT.FILE file produced by Option 4.

INDEX STATIONS HAVING THE LOWEST STANDARD ERROR OF REGRESSION

11:59 Saturday, May 25, 2002 37

Obs	STATION NUMBER	STATION DD NUMBER	INDEX STATION NUMBER	INDEX STATION DD NUMBER	STANDARD ERROR (PER CENT)	R SQUARED (PER CENT)
106	02156050	5	02160326	11	28.0	82.18
107	02156050	5	02154790	14	28.8	81.13
108	02156050	5	02160390	11	29.6	80.21
109	02156050	5	021630967	11	29.8	79.90
110	02156050	5	02160105	5	32.3	76.45
111	02156500	5	02160105	5	23.9	89.74
112	02156500	5	02154500	14	24.7	89.04
113	02156500	5	021556525	13	26.5	87.41
114	02156500	5	02160700	5	27.2	86.80
115	02156500	5	02165200	5	28.7	85.25
116	02160105	5	02160700	5	21.9	92.76
117	02160105	5	02160390	11	25.3	90.41
118	02160105	5	02156500	5	25.6	90.31
119	02160105	5	02161000	1	27.0	89.24
120	02160105	5	02165200	5	27.5	88.71
121	02160326	11	02160390	11	18.3	92.94
122	02160326	11	02164000	1	22.9	88.99
123	02160326	11	021630967	11	24.0	87.98
124	02160326	11	02160700	5	25.5	86.43
125	02160326	11	02154790	14	25.6	86.38
126	02160381	11	02160390	11	29.6	86.23
127	02160381	11	02165200	5	30.5	85.41
128	02160381	11	021630967	11	30.6	85.25
129	02160381	11	02160700	5	30.8	85.11
130	02160381	11	02160326	11	34.9	81.02
131	02160390	11	02160700	5	16.5	94.76
132	02160390	11	02160326	11	18.2	93.69
133	02160390	11	02165200	5	24.6	88.52
134	02160390	11	02160105	5	25.6	87.64
135	02160390	11	02154790	14	27.3	85.93
136	02160700	5	02160390	11	15.3	95.78
137	02160700	5	02160105	5	20.9	92.16
138	02160700	5	02165200	5	21.6	91.64
139	02160700	5	02160326	11	24.2	89.55
140	02160700	5	02156500	5	28.4	85.80
141	02161000	1	02156500	5	31.2	86.36
142	02161000	1	02169500	1	31.9	85.66
143	02161000	1	02160105	5	32.9	84.77
144	02161000	1	02160700	5	34.5	83.32
145	02161000	1	02165200	5	37.4	80.49

Figure 4. OUT.FILE file reports produced by Option 4 report of station numbers, index station numbers, ADAPS data-descriptor numbers, standard errors of estimate, and R-square in OUT.FILE produced by Option 4.

(2) The most-correlated file created by option 4 can be accessed by option 5 to list the five most-correlated stations for each station individually. Option 5 is provided for users who do not have access to the **out.file** or printed list created by option 4. Typically, a user would use option 5 to get the list of most-correlated stations for a specific station, enter selected index stations into the selected-index station file, and do a hydrographic comparison using options 1 or 2.

The index station selection program (option 4) should be run initially by an experienced person to populate the most-correlated index station file and the selected-index station file. Option 4 needs to be rerun only when new stations are installed in the field or when current hydrologic conditions greatly differ from conditions for which the index station selection program was last run. Most users within a District will not have to run option 4.

An important fact to understand is that the correlation and regression processes implemented in option 4 are used only in the initial selection of index stations. Regression equation coefficients are not stored for later use in options 1 and 2. Instead, the regressions implemented in options 1 and 2 are rerun each time options 1 and 2 are run. R-square and the standard error of estimate also are recomputed and stored each time options 1 and 2 are run.

Methods for Determining the Most-Correlated Index Stations

The index station selection portion of HYDCOMP first creates a list of the 20 highest correlated index stations for each review station in the State using daily index station flows lagged from -3 to +3 days. The program computes correlation coefficients between logarithms of flow at every review station and logarithms of individually lagged flows at every other index station for -3 to +3-day lags. Lags are limited to -3 to +3 lags because of the enormity of the computations involved. The program then performs multiple regression between each review station and the corresponding five index stations having the highest correlation coefficients using -8 to +8 lags of daily flows and day of week as is done in options 1 and 2. In the multiple regression, HYDCOMP selects the -8 to +8 lagged flows and days of the week that are significant at the 5 percent acceptance level. All flows are transformed to base-10 logarithms before regression to achieve homoscedacity of the data about the regression lines. These regressions are done because the correlation analysis evaluates only -3 to +3 lags one at a time, whereas multiple lags of -8 to +8 days and day of week are used for final hydrographic comparisons.

This final list of review stations and associated index stations is stored in the most-correlated index station file and printed in a file named "**out.file**" (fig. 4), which is sorted by station number and standard error of regression. **out.file** also contains a missing-data report (fig. 2) and tabulates regression coefficients (fig. 3).

In addition, if the program selects five index stations for a review station, and it finds that no index stations were entered in the selected-index station file for that station, the program will automatically enter the selected index stations in the file for that station. This is useful for initially populating the selected-index station file when the program is first installed. However, option 4 only enters the index station numbers and dd numbers for each review station, and the index station and the begin/end month and day for seasonal regression into the selected-index station file. The user still has to type into the selected-index station file the index and review station names and begin-end dates for regression and plotting using options 1, 2, or 3.

After the selected-index station file has been initially populated by option 4, it might be necessary to rerun option 4 with a different period of record or regression season. If this is done, the rerun of option 4 will not overwrite the stations already entered in the file. Therefore, in this case, the user must enter other index stations manually, using index stations listed in **out.file,** or obtained one review station at a time using option 5. The selected index stations can be entered and plots produced one review station at a time using options 1 or 2. Alternatively, index stations can be entered for several review stations at a time using option 3.

Data Entry for Option 4

For the index station selection portion of the program (option 4), the user creates an ADAPS group file containing the station numbers and dd numbers of the currently active stations in the District. The screen form is shown in figure 5. Data entry requirements are:

1. **ADAPS GROUP FILE NAME** - The user creates an ADAPS group file containing the station numbers of all the currently active stations in the State, as well as a single dd number for each station identifying published, finalized daily value flow data. The name of this ADAPS group file is entered in this field.

2. **BEGIN MM-DD-YYYY and END MM-DD-YYYY** - The begin/end dates of regression should include a period of published, finalized daily flow data when hydrologic conditions are similar to the current year being analyzed. The period should be at least 1-4 years. Computational requirements are so large that option 4 may not work if periods longer than 1 or 2 years are used on some computers. The user should start with the longest period desired, and successively try shorter periods until the program works, if the program fails using the longer periods.

3. **BEGIN SEASON (MM DD) and END SEASON (MM DD)** - For ice-affected stations, it is necessary to restrict regressions to periods when flows are not affected by ice to eliminate estimated data from the regression. For arid stations, it may be necessary to do seasonal regressions. Therefore, the begin/end dates to limit the regression to seasonal periods are entered in these fields.

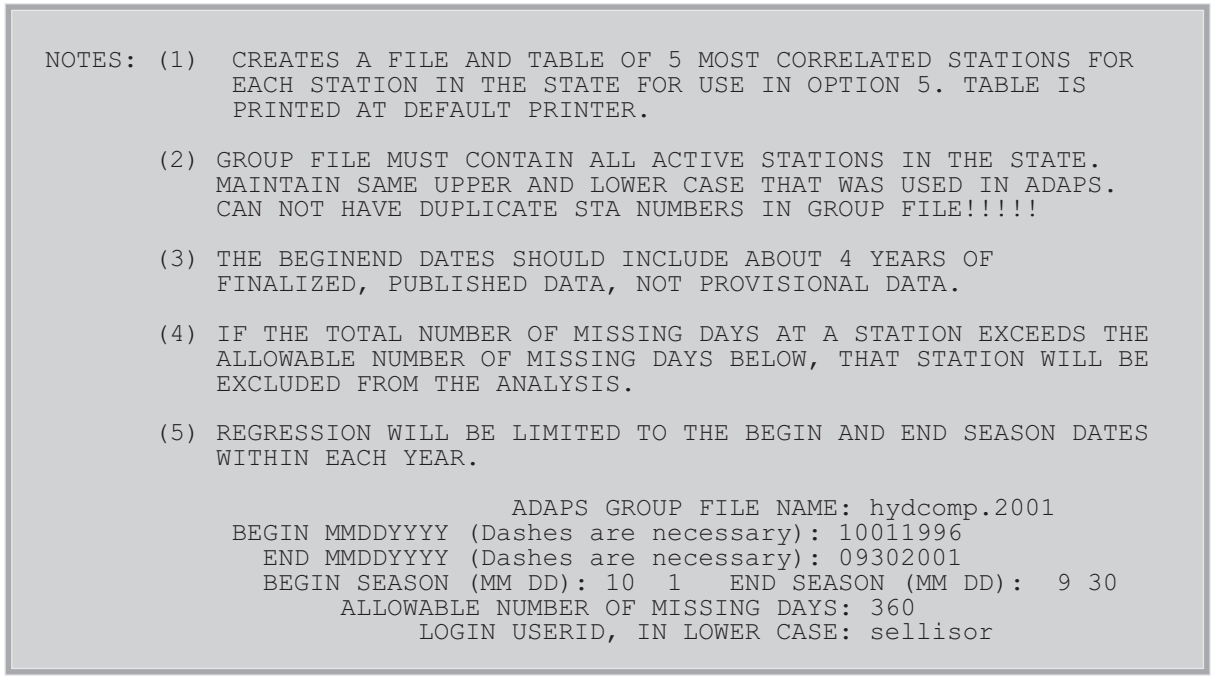


Figure 5. Data input form for Option 4.

4. **ALLOWABLE NUMBER OF MISSING DAYS** - If the number of missing days for a station during the period of regression exceeds the number of days entered in this field, the program will not use that station in the selection process. It was found that if a station had only a few days of record, that station would correlate closely to every other station. For non-arid States, a value of 30 days is reasonable. For arid States, where zero flows may exist for many months of the year, the number of allowable missing days set by the user may have to be quite large. However, the number of allowable missing days for arid States should be reasonable based on the expected number of zero-flow or missing-flow days for the average station in the State.

5. **LOGIN USERID, IN LOWER CASE** - The computer login id, in lower case, of the person who created the group file described earlier should be entered in this field. HYDCOMP uses the user id and the group file name to locate the actual group file so it can retrieve the ADAPS dd numbers associated with each station.

6. To exit the screen and execute the program, click on “File” at the top of the screen, and then “Close.”

DISPLAYING THE FIVE MOST-CORRELATED STATIONS FOR EACH STATION IN THE STATE USING OPTION 5

Once option 4 has been used to populate the most-correlated index station file, option 5 can be used to list the five most-correlated index stations for any individual review station, as shown by the data input screen form in figure 6. The review station number is the only entry that needs to be made. Instructions are shown on the screen form describing how to select index stations from the list. When option 5 is used, the list of selected index stations, associated dd number, and standard error of estimate are presented on the screen, as shown in figure 7, and printed at the user's default printer, as shown in figure 8.

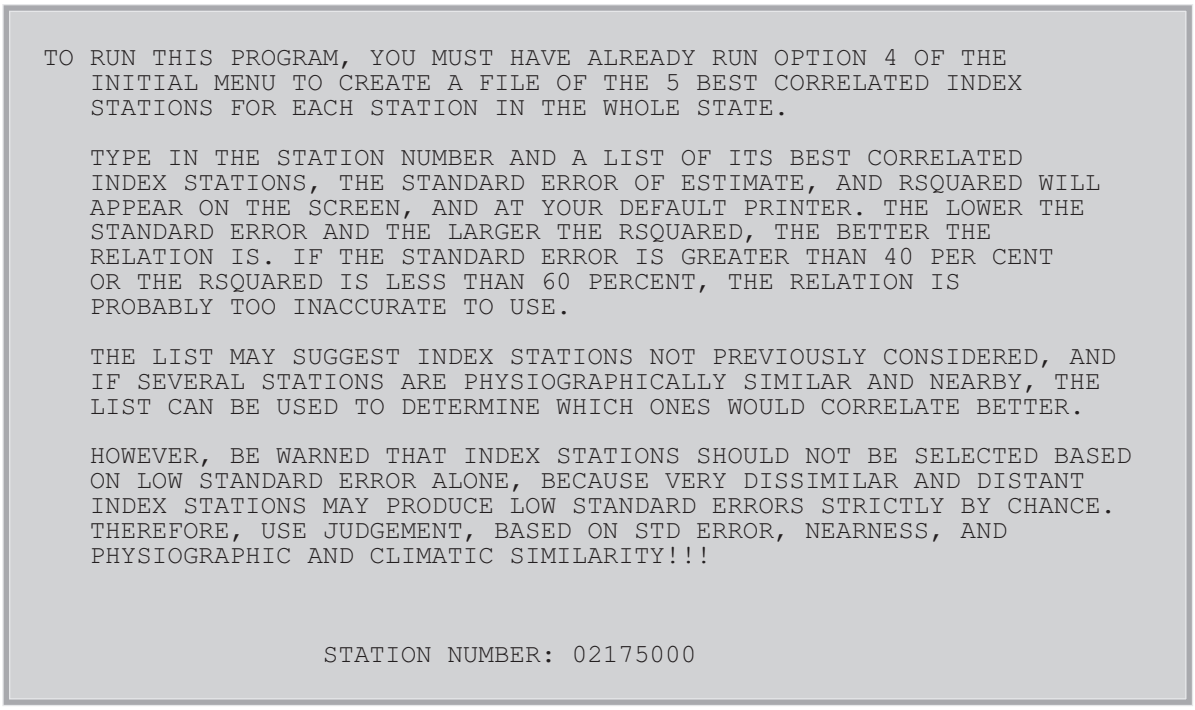


Figure 6. Data input form for Option 5.

Obs	sta	staddno	indxsta	indxddno	stderr
1	02175000	18	02173030	12	25.6
2	02175000	18	02173051	12	26.3
3	02175000	18	02173000	15	28.2
4	02175000	18	02173500	1	29.5
5	02175000	18	02132000	1	35.8

Figure 7. Output by Option 5 to the computer screen, tabulating station numbers, ADAPS DATA descriptor numbers, standard errors of estimate, and R-squares.

INDEX STATIONS HAVING THE LOWEST STANDARD ERROR OF REGRESSION
12:17 Saturday, May 25, 2002 1

STATION NUMBER=02175000

Obs	STATION DD NUMBER	INDEX STATION NUMBER	INDEX STATION DD NUMBER	STANDARD ERROR (PER CENT)	R-SQUARED (PER CENT)
1	18	02173030	12	25.6	90.4
2	18	02173051	12	26.3	89.9
3	18	02173000	15	28.2	88.4
4	18	02173500	1	29.5	87.3
5	18	02132000	1	35.8	81.7

Figure 8. Output by Option 5 to the printer, tabulating station numbers, ADAPS data descriptor numbers, standard errors of estimate, and R-squares.

METHODS FOR SELECTING THE BEST INDEX STATIONS FROM THE MOST-CORRELATED INDEX STATIONS

The option 4 **out.file** and list and the output from option 5 contain the five most-correlated index stations for each review station in the State (fig. 4) and the associated ADAPS dd numbers, standard error of estimate, and R-square for use in the hydrographic comparison and flow estimate part of the program. The list is sorted by station number and then standard error of estimate. From this list (or the output from option 5), the user selects index stations having the lowest standard error of regression and highest R-square that are hydrologically similar to the review station. It should be noted that the standard error may be low and r-square may be high for an index station by coincidence even if the basins are hydrologically dissimilar. Also, the user can select index stations not included in **out.file**.

The standard error of estimate or R-square give the user a statistical method for comparing hydrologically similar index stations. For example, if two potential index stations are judged to be hydrologically similar, the one having the lower standard error of estimate or higher R-square should be selected. Also, a high standard error of regression or low R-square could indicate that a basin may have been erroneously judged to be hydrologically similar. After the selected-index station database has been initially populated by option 4, it is imperative that unsuitable index stations be eliminated from the selected-index station file.

The option 4 **out.file** and list contain a missing data report (fig. 2), which shows the number of missing days for the period of record retrieved. The list may point out index stations that were not selected for a station because there were more missing days than allowed by the user-specified limit. **out.file** also includes a list of standard errors of regression, R-square, and regression coefficients (fig. 3). This list can be used to determine how the most significant lags of flows compare with lags determined by the user.

Another use of option 4 is to allow an outside reviewer to quickly obtain a list of index stations. The outside reviewer can do hydrographic comparisons for any review station in a District without having to search through the station files for previously selected index stations or having to do a complete hydrologic analysis to determine suitable index stations. As a matter of fact, the outside reviewer can conduct hydrographic comparisons without even being in the District being reviewed.

DATA ENTRY, HYDROGRAPHIC COMPARISON, AND ESTIMATES OF MISSING DATA BY REGRESSION USING OPTIONS 1, 2, AND 3

The use of HYDCOMP for data entry of single review stations and hydrographic comparison (options 1 and 2) and multiple-review station data entry (option 3) are best described by a discussion of the data input on the input screen forms (figs. 9-11) and a discussion of the output. The input form is the same for all three options. No plots will be produced using option 3, because it is used only for entering multiple sets of data and for deleting review stations and their index station numbers from the selected-index station file.

In general, a much better regression than using option 1 and one index station can be obtained using option 2 and index stations on two tributaries or main branches of the same stream, if available. If the two index stations are on streams different from the stream at the review station, a somewhat better regression can be obtained by using the two index stations rather than one index station.

```
NOTE: FOR INSTRUCTIONS, CLICK VIEW, NEXT SCREEN, VIEW, NEXT SCREEN

==TO==SELECT==THIS==STATION==YOU==MUST==ENTER=="Y"==HERE : Y
STATION NUMBER 02175000             DD NUMBER 18
STATION NAME   EDISTO RIVER NR GIVHANS,SC

DO NOT, DO NOT,                              NOTE: (1) Regression period is
DO NOT USE THE      BEGIN DATE  END DATE              1 or more water yrs
RIGHT KEY PAD!!    (YYYYMMDD)  (YYYYMMDD)             prior to plot period.
                                                 (2) Do not enter Rsq or
REGRESSION PERIOD:  19951001    19990930             std err; computed by
PLOT PERIOD      :  19990915    20001101             program after each run.

LOG OR ARITH REGRESSION (L/A) L

 INDEX STA NUM   DD STA NAME    INDEX STA NUM  DD STA NAME   R  STD  USE
      #1         #1    #1           #2        #2    #2      SQR ERR (Y/N)
02173030         12 SF ED COPE 02173500         1 NF ED ORAN  93  24   Y
02132000          1 LYNCHES EF _______________ __ ___________ 75  40   N
02173051         12 SF ED BAMB _______________ __ ___________ 92  23   N
02173000         15 SF ED DENM _______________ __ ___________ 91  24   N
02173500          1 NF ED ORAN _______________ __ ___________ 90  28   N
FORMAT OF NWS FILE HOLDING RAINFALL AND MINMAX TEMPERATURE DATA: A
NAME OF NWS FILE: ____________________________________________________
RESTRICT REGRESSIONS TO DV DATA BETWEEN MM DD 10  1 AND MM DD  9 30
RESTRICT REGRESSIONS TO DV DATA BETWEEN 9999999 CFS AND 9999999 CFS

NOTE: FOR MORE DATA SCREENS AND INSTRUCTIONS, CLICK: view, next screen
```

Figure 9. Screen 1 of the data input forms for Options 1, 2, and 3.

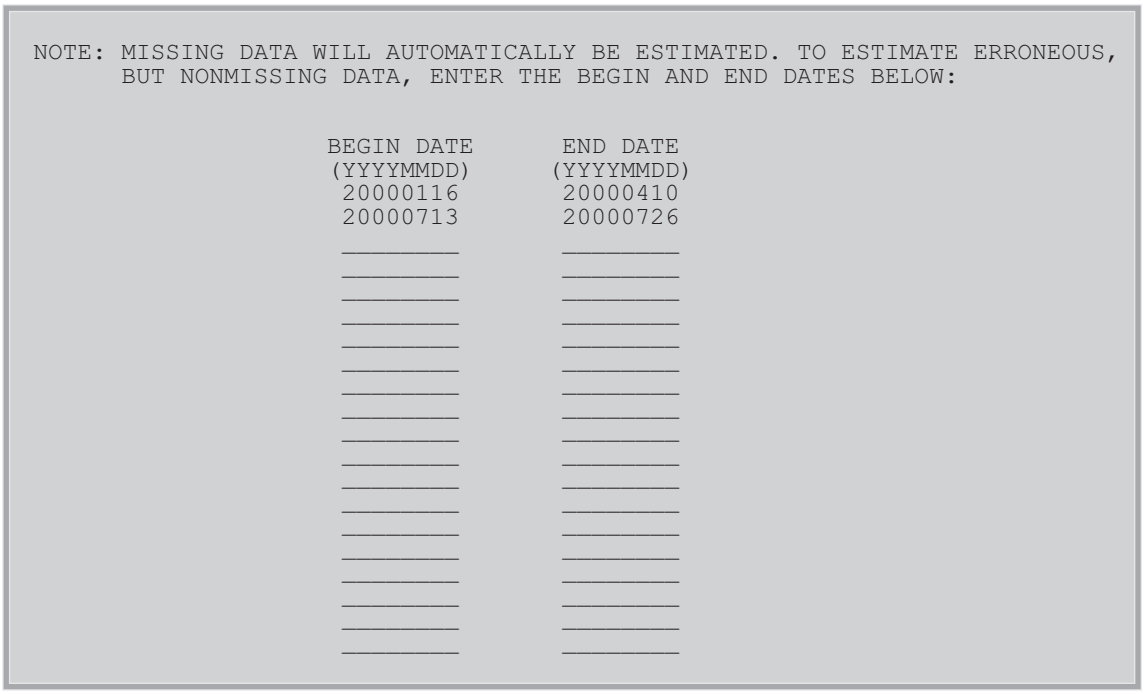


Figure 10. Screen 2 of the data input forms for Options 1-3.

Abbreviated instructions (fig. 11) describing how to use the screen forms, how to enter data on screens 1 and 2, and how to analyze plotted and tabulated results are shown on the on-line screen instructions following screen 2 (fig. 10) for quick reference by the user. The user can page through these instructions by clicking on “View” and “Next Screen” (not “Next Observation)” at the top of the screens. The instructions can be printed by clicking on “File,” “Print Utilities,” “Screen Print,” and “Print.”

How to Enter Data on Screen 1 for Options 1, 2, and 3

Screen 1 (fig. 9) holds the majority of the data needed for hydrographic comparison, estimates of missing data, and data entry by HYDCOMP. Once the selected-index station file is populated, the user will only have to specify in line 1 that this station is to be plotted, specify a regression period and plot period (described below), select the index station(s) to be used, and specify specific periods for which tabular and graphical estimates are desired (in Screen 2, figure 10). After initial population of the selected-index station file, the station names and dd numbers associated with the review stations and the selected index stations do not have to be re-entered with each use of the program.

```
==HOW==TO==USE==THE==SCREEN==FORMS==:        VERY==NECESSARY!!!

  1. TO ENTER DATA ON FORM: arrow or tab to the selected position, or click
        on the position using the left mouse button. Delete with backspace.

  2. TO CHANGE SCREENS, click: view, next screen, or view, previous screen

  3. TO FIND A STATION, click: search, where, and type: sta = '021973000'
         WARNING: the single quotes are NECESSARY!!
         BEFORE FINDING A SECOND STATION, click: search, undo last where

  4. TO DELETE A STATION, find sta as in (3) and click: edit, delete record
     NOTE: Can delete stations only in Option 5, not options 1 and 2.

  5. TO ADD A STATION, first try to find sta as described in (3). Then,

         (a) If found, station is already created, and should not be added
         (b) If not found, click: edit, add record    and type in new data

  6. TO EXIT THE FORM AND RUN THE PROGRAM, click:   file, close

  7. TO STEP THROUGH PLOTS ON SCREEN, click on plot with left mouse button

FOR MORE INSTRUCTIONS, KEEP CLICKING: view, next screen


==DATA==ENTRY==INSTRUCTIONS==FOR==SCREEN 1:

 1. To compute and plot, you MUST type "Y" on the first line of screen 1

 2. "REGRESSION PERIOD" should be 1 or more water yrs prior to plot period

 3. "PLOT PERIOD" is the period to be compared to simulated flows

 4. "LOG OR ARITH"  use log for flows and arith for sw and gw stages

 5. INDEX STATIONS  type in sta nos, dd nos, and short sta name
    If "one index station" was selected from initial menu, only stations
    under the "INDEX STA NUM #1" column will be used in each regression.
    If "two index stations" was selected from initial menu, stations
    under both INDEX STANUM #1 and #2 columns will be used together in
    in each regression.

 6. "RSQR" is the "rsquared" coefficient, which is the percent of
    variation in flow that is explained by the regression. The LARGER
    RSQR is, the more accurate the relation is. If RSQR is less
    than about 60 percent, the relation is probably too inaccuate to
    use.RSQR is automatically entered in this column by the program.

 7. "STD ERR" is the standard error of estimate from the regression, in
    per cent. It is automatically entered in this column by the program,
    on each run. The SMALLER the STD ERR is, the better the relation is.
    Therefore, select closer stations with similar physiography and
    rainfall supply, and with the smaller std errs, where the RSQR is
    greater than 60 percent.

 8. USE (Y/N)  You MUST type 'Y' in this column to specify which of the
    5 index stas or pairs of index stas are to be used in the current
    run of the program.
```

Figure 11. Help information for Options 1-3. (Continues on following pages.)

```
  9. FORMAT OF NWS FILE  The National Weather Service file, obtained
     from the NWIS, contains rainfall and maxmin temperature data
     to be plotted on the hydrograph. See documentation for
     description of format A. Email clsander for other formats.

 10. NAME OF THE NWS FILE  Type in the name of the file.

 11. RESTRICT REGRESSIONS  restrict regressions and plots to the
     begin and end monthday, or to a minmax range of flow, as shown.
     This is to account for seasonality, such as in the west, where
     flow may exist only for certain months of the year.


DATA==ENTRY==INSTRUCTIONS==FOR==SCREEN 2:

  Missing data are automatically estimated and tabulated. The estimated
  flows are automatically adjusted to match the nonmissing data on either
  side of the missing data, as would be done by comparisons on a light
  table. These are called "adjusted simulated" flows.

  Sometimes data are not missing, but are erroneous. Smoothed data can
  be simulated and tabulated by typing the begin and end dates for these
  periods in SCREEN 2.



==HOW==TO==ANALYZE==THE==TABLES==AND==PLOTS:

  1. "REGRESSION COEFFICIENT" table shows the regression equation
     coefficients, and is usually of no use to the analyst, other than to
     see what lagged flows were used. The program uses 8 to +8 lags of the
     index station flows, and dayofweek as explanatory variables in the
     regression. The lagged flows tend to modify the shape and timing
     of the simulated hydrograph, similar to routing flows. The dayofweek
     variables are used in an attempt to quantify the effect of weekly
     power cycles. The largest positive regression coeff shows the best lag.

  2. "ADJUSTED AND UNADJUSTED FLOWS" table contains simulated flows. The
     user should examine the hydrographs to see that the simulated flows are
     reasonable. Normally, the adjusted simulated flows should be used,
     because the Kalman smoothing method smooths simulated flows from the
     nonmissing flows at the edges of the missing data toward the
     unadjusted regression flows at the center of the gap. If the regression
     flows and observed flows cross each other often on the hydrograph, the
     simulated flows will quickly merge with the regression flows.


 3. Residuals and per cent residuals are the differences between the
    simulated and observed data. They are shown plotted against
    time over the period of regression and plot; against day of the
    year, to see if there is a seasonal variation; and against the
    best index flow, to see if some data points may have skewed the
    relation. In general, the residuals should be uniformly scattered about
    the zero residual lines, and should not be bowed, skewed, or offset.
    If they are not uniformly distributed about these lines, be warned that
    the regression may be in error for certain ranges of flow or times of
    year. Uniform oscillations about the zero residual line may be ok.
```

Figure 11. Help information for Options 1-3. (Continued)

```
 4. HYDROGRAPHIC COMPARISON PLOT:

    a. Look at the std err first. Std err means that we're 95% sure that
       2/3 of the observed data are within the std err of the estimated
       data, so this is a measure of the central portion of the scatter.
       If this is less than 10%, you have a really, really good relation.
       However, the rsquare should be greater than 60 percent.
           RANGE OF STD ERR     FOR ESTIMATING DATA  FOR HYD COMPARISON
             0% TO 15%               GOOD                 GOOD
            15% TO 30%               FAIR                 GOOD
            30% TO 50%         PROBABLY NOT USEABLE      FAIRPOOR
            OVER 50%               NOT USEABLE       LOOK AT IT ANYWAY

      b. The 95% confidence limits represent the outer limit of the
         error associated with the regression. Therefore, if the flow
         data are outside these limits for long periods, these periods
         should be investigated in detail for error. Short excursions
         and excursions verified by plotted flow measurements may be ok.

      c. The range of flows used in the regression may not cover the range
         of simulated flows. Flows simulated far outside the range of the
         regression flows may not be reliable. These flows are those
         above or below the dashdot minmax flow lines on the hydrographs.
         The analyst should select earlier begin dates for the regression
         period to try to include a wider range of regression flows.

      d. Adjusted simulated flows for missing data are shown by squares on
         the plots.

      e. As always, check for flat places or sudden changes or spikes
         in both the observed and simulated data for obvious problems
         such as a float on the sand, hungup float, or a bad data point.

      f. KEEP IN MIND THAT THE INDEX STATION MAY BE IN ERROR, RATHER THAN
         THE STATION IN QUESTION. THEREFORE, CHECK FLOWS AT BOTH STATIONS.
         Do two separate hydrographs using two separate index stations to
         determine which of the 3 stations is in error.

CONCLUSION OF ALL INSTRUCTIONS. RETURN BY CLICKING: view, previous screen
```

Figure 11. Help information for Options 1-3. (Continued)

A detailed description of each data item is provided below, assuming that the review station to be plotted is selected as described above. Note the warning "**DO NOT, DO NOT, DO NOT USE THE RIGHT KEY PAD**." Use the number-keys at the top of the keyboard instead. *Using the right key pad for numeric input could lock up the program,* in which case, the system administrator must stop HYDCOMP, and delete the work directory with the prefix "SAS" stored in the **/tmp** directory. The screen-form descriptor of the data field or note is indicated in bold capital letters below.

1. **==TO==SELECT==THIS==STATION==YOU==MUST==ENTER==”Y”==HERE**: The user **must** enter "Y" or "y" to get a plot for the review station; otherwise, the program will produce no results.

2. **STATION NUMBER** is the station number of the review station.

3. **DD NUMBER** is the ADAPS data descriptor number specifying flow for the period of record, including the current year, for the review station.

4. **STATION NAME** is the station name of the review station. The name should include an official, unabbreviated station name and location so that anyone can recognize the station.

5. **REGRESSION PERIOD** is the period, in YYYYMMDD format, to be used by HYDCOMP in developing the regression equation. This period can be any length the user desires. A period of at least 4 years should probably be used and should include only published, finalized flow data prior to the current year for which the comparison is to be made. Thereby, the regression equation will not be influenced by possibly erroneous data from the current year. However, all available data for a new station can be used in the regression. In addition, a period of record should be used with about the same range of flow data as the current year, if possible. For example, if streams in the current year are in extreme drought, it might be necessary to select a prior period where similar low flows were experienced. HYDCOMP hydrograph plots show the range of the input data to the regression as discussed below.

6. **PLOT PERIOD** is the current period for which hydrographic comparison and possible estimates of missing data are to be done. Because the regression uses +8 lagged flows, this period should extend at least 10 days past the end of the period to be compared. The regression also uses -8 lagged flows, but if the regression period adjoins the plot period, the plot period does not have to extend 10 days earlier than the period to be compared.

7. **LOG OR ARITHMETIC REGRESSION (L/A)** - "L" or "l" is entered to transform input values to logarithms before regression. "A" or "a" is entered to use arithmetic input values in the regression. Generally, flows should be converted to logarithms before regression, unless zero or negative flows are observed. For regression of water-quality parameters, the user should do preliminary plots of the dependent data against any potential explanatory variables to see if a log transformation is necessary. In addition, HYDCOMP can be used to compare stages in rivers or ground-water stages, in which case, input values should not be converted to logarithms.

   The purpose of transforming flows or any other parameter to logarithms is to maintain the uniformity of scatter about the regression line, as required by the OLS regression method. For instance, a plot of stage against flow on arithmetic paper would result in a shotgun scatter increasing with increasing stage. However, a plot of the same data on logarithmic paper would result in uniform scatter about a regression line through the data.

8. **INDEX STA NUM #1, DD #1, and STA NAME #1** - These are the station numbers, dd numbers and abbreviated station names of the index stations to be used in option 1 or of the first index station to be used in option 2, where two index stations are to be used in the regression. Up to five index stations can be listed to create up to five separate hydrographic comparison plots. Only the final selected index station information should be included in these fields.

9. **INDEX STA NUM #2, DD #2, and STA NAME #2** - These are the station numbers, dd numbers and abbreviated station names of the second index station required for option 2, where two index stations are to be used in the regression. If option 1 is selected, the information in these fields is ignored. If option 2 is selected, and no information is entered in these fields for the second index station, the program will fail.

10. **R-SQR** is the R-square of the regression, which is the percentage variation in flow that is explained by the regression. The larger the R-square, the more accurate the regression is. If the R-square is less than 60 percent, the regression is likely too inaccurate to be used. The user does not enter this value, because HYDCOMP enters the R-square computed by the last run of option 1 or 2. The primary use of R-square is to determine which regressions are completely unusable.

11. **STD ERR** is the standard error of estimate of the regression, which is a measure of the central scatter about the regression line. The user does not enter this value, because HYDCOMP enters the standard error of estimate computed by the last run of option 1 or 2.

    The standard error of estimate can be used, like R-square, to determine which index stations produce the most accurate regressions. The smaller the standard error of estimate, the more accurate the regression is. The standard error is more useful than R-square, because the user can visualize the percentage scatter of the data points about the regression line (about two-thirds of the input data points will be within the percentage standard error of estimate of the regression line).

12. **USE (Y/N)** - Type a “Y” or “y” in this column beside the index station(s) to be used for hydrographic comparison and flow estimates. A separate set of comparisons will be produced for each set of index station(s) selected in this column in the current run of the program.

13. **FORMAT OF NWS FILE HOLDING RAINFALL AND MIN-MAX TEMPERATURE DATA** - For ice-affected stations, it is useful to overlay rainfall and daily minimum-maximum temperature data over the hydrograph of daily flows, to help determine if stage changes occur because of surface runoff or backwater from ice.

Several weather-data formats may exist, so the format has to be hard-coded in the HYDCOMP program. When the format is hard-coded, it is assigned a character identity (A, for example), and that character is entered in this field.

14. **NAME OF NWS FILE** - This is the name of the file holding the weather data.

15. **RESTRICT REGRESSIONS TO DV DATA BETWEEN MM DD AND MM DD** - For ice-affected stations, it is sometimes customary to estimate ice-affected flow data using regressions based on periods of the year when flows are not ice-affected, thereby eliminating the use of estimated data from the regression. For arid regions, it may be desirable to limit regressions to seasons where flow is most likely to exist. Therefore, HYDCOMP can restrict the regression to the month-day segments of the year as entered in these fields. The default is the entire water year. If a seasonal period is specified in option 4, that period will be loaded into the selected-index station file used in options 1, 2, and 3. However, the user is free to change the seasonal period for each station as necessary.

16. **RESTRICT REGRESSIONS TO DV DATA BETWEEN ____ CFS AND ____ CFS** - For arid regions, it may be desirable to limit regressions to some range of flow that can be entered in these fields.

17. To proceed to screen 2 in order to request tabulation of estimated data for periods of non-missing but erroneous data, click on “View” at the top of the current screen, and then click on “Next Screen.”

18. If it is not necessary to go to screen 2, check to see that a “Y” is coded at the top of the screen to request the regression and plot, and click on “File” at the top of the screen, and then “Close” to exit the screen and execute the program.

How to Enter Data on Screen 2 for Options 1 and 2

The following instructions appear on Screen 2 (fig. 10):

NOTE: MISSING DATA WILL AUTOMATICALLY BE ESTIMATED. TO ESTIMATE ERRONEOUS, BUT NON-MISSING DATA, ENTER THE BEGIN AND END DATES BELOW.

HYDCOMP will not tabulate estimated data, where the daily value data is not missing for the review station, to reduce the amount of tabled data. A common occurrence is that the daily value data is erroneous, rather than missing, because of a hung float, float on the mud, recorder malfunction, and so forth. To have HYDCOMP tabulate estimated data for these periods, enter the begin/end dates in YYYYMMDD format in the fields of Screen 2.

Data can be entered on Screen 2 from option 3. However, when entering many stations at one time using option 3, the user usually does not know which periods of erroneous, but non-missing data need to be estimated and, therefore, does not have data to enter on Screen 2.

To execute the program from Screen 2, if a "Y" is already coded at the top of Screen 1, click on "File" at the top of the screen and then "Close." Otherwise, return to Screen 1, type "Y" in the field of the first line, and click on "File" and then "Close."

How to Analyze Tabular and Graphic Output From Options 1 and 2

As stated earlier, an extensive knowledge of statistics is not necessary to conduct hydrographic comparisons and estimate missing data using HYDCOMP. A discussion of the following example output from HYDCOMP demonstrates the simplicity and straight-forwardness of the analytical methods used by HYDCOMP, beginning with the tabular outputs.

The tabular output in figure 12 summarizes the data controlling the current run of the program. The table lists the date and time of the run, the station number, dd number, station name, the regression periods, the plot periods, seasons, requested ranges of flow to be used in the regression, and whether the regression was logarithmic or arithmetic.

The tabular output (fig. 13) summarizes the index stations used in the regression, the standard error of estimate, the R-square, and regression coefficients. This information and the information in figure 12 should be retained to describe the input and output from the

IDENTIFICATION OF COMPUTER RUN

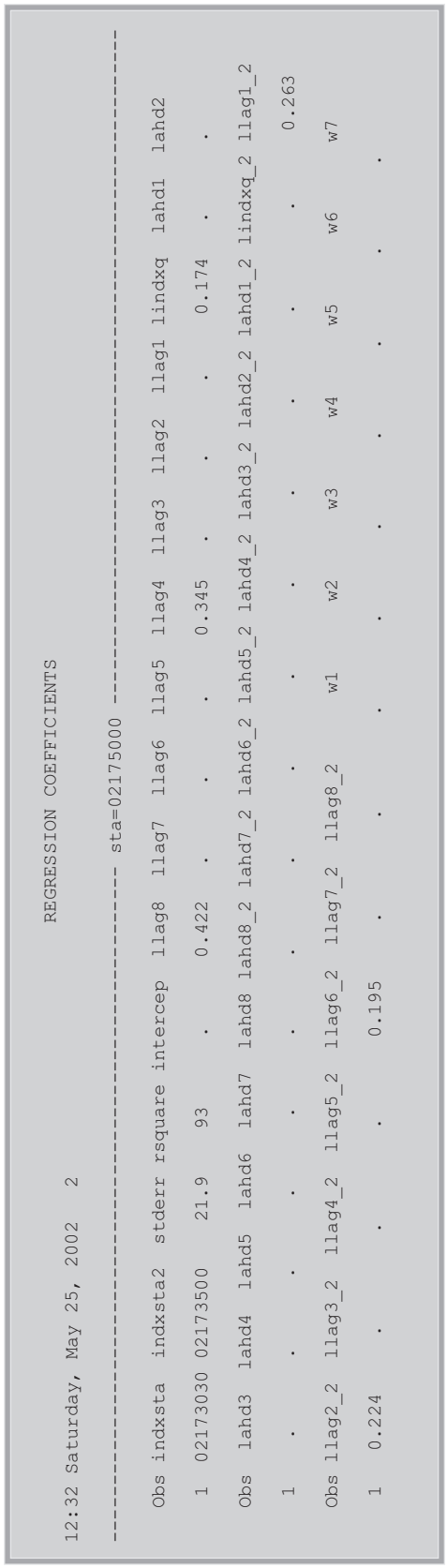
12:32 Saturday, May 25, 2002 1

Note: Keep this as a cover sheet to describe the computer run.

Obs	DATE-TIME OF RUN	STATION NUMBER	DD-NUMBER	STATION NAME	BEGIN MONTH-DAY OF SEASON	END MONTH-DAY OF SEASON	LOW SEASONAL FLOW	HIGH SEASONAL FLOW	NUMBER OF DAYS PER PLOT	L=LOGARITHMIC A=ARITHMETIC	BEGIN REGRESSION
1	25MAY02:12:34	02175000	18	EDISTO RIVER NR GIVHANS,SC							19951001

Obs	END REGRESSION	BEGIN PLOT	END PLOT	BEGIN MONTH-DAY OF SEASON	END MONTH-DAY OF SEASON	LOW SEASONAL FLOW	HIGH SEASONAL FLOW	NUMBER OF DAYS PER PLOT	L=LOGARITHMIC A=ARITHMETIC
1	19990930	19990915	20001101	1001	0930	-9999999	9999999	100	L

Figure 12. Report listing date and time of run, station and ADAPS data descriptor numbers, station name, begin and end dates of regressions and plot periods, seasonal limits of regression, regression limits in terms of flow, and if flows were transformed to logarithms for Options 1 and 2.



```
                                REGRESSION COEFFICIENTS
12:32 Saturday, May 25, 2002   2

------------------------------------------ sta=02175000 ------------------------------------------------------

 Obs indxsta  indxsta2  stderr rsquare intercep  llag8  llag7  llag6  llag5  llag4  llag3  llag2  llag1 lindxq  lahd1  lahd2

   1  02173030 02173500   21.9    93      .     0.422   .      .      .      0.345  .      .      .     0.174   .      .

 Obs  lahd3  lahd4  lahd5  lahd6  lahd7  lahd8 lahd8_2 lahd7_2 lahd6_2 lahd5_2 lahd4_2 lahd3_2 lahd2_2 lahd1_2 lindxq_2 llag1_2

   1    .      .      .      .      .      .      .      .      .      .      .      .      .      .      .      0.263

 Obs llag2_2 llag3_2 llag4_2 llag5_2 llag6_2 llag7_2 llag8_2    w1     w2     w3     w4     w5     w6     w7

   1  0.224    .      .      .      0.195   .      .      .      .      .      .      .      .      .      .
```

Figure 13. Report listing selected index station numbers, standard error of estimate, R-square, and regression coefficients for the regressions done by Options 1 and 2.

regression. The output table (fig. 13) can be used to determine if the lag time of the most significant lagged flow (which had the highest regression coefficient) seems to be a reasonable time difference between flows at the two stations. In the example, llag1_2 is a regression coefficient of 0.263 for flows at index station 2, lagged behind by 2 days. The user might want to check if Saturdays (w7) and Sundays (w1) have regression coefficients for regulated stations, signifying that weekend flows differ from weekday flows in the regression equation. In figure 13, the regression coefficients for w1-w7 are missing, as signified by periods, indicating that they were not important explanatory variables.

The dates and adjusted and unadjusted simulated flows are printed in the tabular output (fig. 14). The table also lists the adjustment factor produced by the Kalman algorithm for smoothing regression flows into the non-missing data at the edge of the gaps in data. The unadjusted simulated flows are also presented in the table, for use in manually smoothing the regressed data with the non-missing data, if necessary. However, the adjusted simulated flows should be used, unless there is evidence that the adjusted flows are not reasonable. The regression log-bias correction factor is also printed in the table, as well as the index station numbers, standard error of estimate, and R-square. This table is printed whether or not the user sends plots to the screen or printer.

When the user requests that plots and data are directed to the screen, the index station numbers, dates, log-bias, adjustment factor, unadjusted simulated flows (regressq), and adjusted simulated flows (final_q) are presented in a table on the computer screen as shown in figure 15. As stated above, the adjusted simulated flows should usually be used.

ADJUSTED AND UNADJUSTED REGRESSION FLOWS
12:32 Saturday, May 25, 2002 3

lgstacnt=1 sta=02175000 indxsta=02173030 indxsta2=02173500 stderr=21.9
rsquare=93

date	UNADJUSTED SIMULATED LOG BIAS (PER CENT)	ADJUSTED SIMULATED ADJUSTMENT FACTOR	FLOW (CFS)	FLOW (CFS)
20000116	0.2	1.36	1474.1	2010.7
20000117	0.3	1.35	1531.6	2068.3
20000118	0.5	1.34	1627.0	2176.0
20000119	0.6	1.33	1703.6	2257.5
20000120	0.7	1.31	1668.7	2191.8
20000121	0.8	1.30	1631.0	2124.2
20000122	0.9	1.29	1674.9	2163.6
20000123	1.0	1.28	1687.4	2162.9
20000124	1.1	1.27	1720.3	2188.6
20000125	1.2	1.26	1847.8	2334.1
20000126	1.2	1.25	2147.4	2694.1
20000127	1.3	1.25	2538.5	3164.1
20000128	1.4	1.24	2878.8	3565.9
20000129	1.4	1.23	3118.5	3840.0
20000130	1.5	1.22	3319.6	4064.5
20000131	1.5	1.22	3805.7	4634.7
20000201	1.6	1.21	4279.1	5184.5
20000202	1.6	1.21	4661.5	5620.4
20000203	1.7	1.20	4804.4	5766.0
20000204	1.7	1.19	4922.9	5882.4
20000205	1.8	1.19	4627.0	5506.0
20000206	1.8	1.19	4260.4	5050.0
20000207	1.8	1.18	4175.7	4931.5
20000208	1.8	1.18	4028.4	4741.1
20000209	1.9	1.17	3577.6	4197.0
20000210	1.9	1.17	3130.9	3661.9
20000211	1.9	1.17	2784.3	3247.4
20000212	1.9	1.16	2496.9	2904.6
20000213	2.0	1.16	2283.1	2649.5
20000214	2.0	1.16	2134.4	2471.6
20000215	2.0	1.16	2079.4	2403.0
20000216	2.0	1.15	2080.9	2400.5
20000217	2.0	1.15	2048.3	2359.0
20000218	2.0	1.15	2012.1	2314.0
20000219	2.0	1.15	1973.3	2266.6
20000220	2.0	1.15	1973.3	2264.2
20000221	2.1	1.15	1966.5	2254.5
20000222	2.1	1.15	1988.6	2278.2
20000223	2.1	1.15	1992.3	2281.3
20000224	2.1	1.14	1952.9	2235.5

(Note: remainder of table is not shown.)

Figure 14. Report listing date, log-bias, Kalman algorithm adjustment factor, unadjusted simulated flow, and adjusted simulated flow output by Options 1 and 2 to the printer.

Obs	indxsta	indxsta2	date	log_bias	adjust	regressq	final_q
134	02173030	02173500	20000116	0.2	1.36	1474.1	2010.7
135	02173030	02173500	20000117	0.3	1.35	1531.6	2068.3
136	02173030	02173500	20000118	0.5	1.34	1627.0	2176.0
137	02173030	02173500	20000119	0.6	1.33	1703.6	2257.5
138	02173030	02173500	20000120	0.7	1.31	1668.7	2191.8
139	02173030	02173500	20000121	0.8	1.30	1631.0	2124.2
140	02173030	02173500	20000122	0.9	1.29	1674.9	2163.6
141	02173030	02173500	20000123	1.0	1.28	1687.4	2162.9
142	02173030	02173500	20000124	1.1	1.27	1720.3	2188.6
143	02173030	02173500	20000125	1.2	1.26	1847.8	2334.1
144	02173030	02173500	20000126	1.2	1.25	2147.4	2694.1
145	02173030	02173500	20000127	1.3	1.25	2538.5	3164.1
146	02173030	02173500	20000128	1.4	1.24	2878.8	3565.9
147	02173030	02173500	20000129	1.4	1.23	3118.5	3840.0
148	02173030	02173500	20000130	1.5	1.22	3319.6	4064.5
149	02173030	02173500	20000131	1.5	1.22	3805.7	4634.7
150	02173030	02173500	20000201	1.6	1.21	4279.1	5184.5
151	02173030	02173500	20000202	1.6	1.21	4661.5	5620.4
152	02173030	02173500	20000203	1.7	1.20	4804.4	5766.0
153	02173030	02173500	20000204	1.7	1.19	4922.9	5882.4
154	02173030	02173500	20000205	1.8	1.19	4627.0	5506.0
155	02173030	02173500	20000206	1.8	1.19	4260.4	5050.0
156	02173030	02173500	20000207	1.8	1.18	4175.7	4931.5
157	02173030	02173500	20000208	1.8	1.18	4028.4	4741.1
158	02173030	02173500	20000209	1.9	1.17	3577.6	4197.0
159	02173030	02173500	20000210	1.9	1.17	3130.9	3661.9
160	02173030	02173500	20000211	1.9	1.17	2784.3	3247.4
161	02173030	02173500	20000212	1.9	1.16	2496.9	2904.6
162	02173030	02173500	20000213	2.0	1.16	2283.1	2649.5
163	02173030	02173500	20000214	2.0	1.16	2134.4	2471.6
164	02173030	02173500	20000215	2.0	1.16	2079.4	2403.0
165	02173030	02173500	20000216	2.0	1.15	2080.9	2400.5
166	02173030	02173500	20000217	2.0	1.15	2048.3	2359.0
167	02173030	02173500	20000218	2.0	1.15	2012.1	2314.0
168	02173030	02173500	20000219	2.0	1.15	1973.3	2266.6
169	02173030	02173500	20000220	2.0	1.15	1973.3	2264.2
170	02173030	02173500	20000221	2.1	1.15	1966.5	2254.5
171	02173030	02173500	20000222	2.1	1.15	1988.6	2278.2
172	02173030	02173500	20000223	2.1	1.15	1992.3	2281.3
173	02173030	02173500	20000224	2.1	1.14	1952.9	2235.5
174	02173030	02173500	20000225	2.1	1.14	1878.6	2150.1
175	02173030	02173500	20000226	2.1	1.14	1811.1	2073.0
176	02173030	02173500	20000227	2.1	1.14	1740.9	1993.0
179	02173030	02173500	20000228	2.1	1.15	1669.7	1912.3
180	02173030	02173500	20000229	2.1	1.15	1623.5	1860.5
181	02173030	02173500	20000301	2.1	1.15	1581.4	1813.7

Figure 15. Report listing date, index station numbers, log-bias, Kalman algorithm adjustment factor (adjust), unadjusted simulated flow (regressq), and adjusted simulated flow (final_q) output by Options 1 and 2 to the terminal screen.

Plots of residuals against several variables are shown in figures 16-18. Residuals are the percentage difference between flows from the regression equation and the measured flows. A positive percent residual indicates that the regression is overcomputing flows. The user should inspect the plots as described below.

1. The user should note the magnitude of the scatter about the zero percent residual line which gives a visualization of the accuracy of the regression equation, in percent.

2. Primarily, the user should determine if the percent residuals are uniformly scattered about the zero percent residual lines. In other words, the general pattern should not form a sinusoidal curve, parabola, angled line, and so forth with respect to the zero percent residual line. It may be somewhat acceptable if the residuals oscillate back and forth around the zero percent line. Residual plots warn the user of flow zones and conditions where the regression equation may produce biased results; adjustments to remove seasonal bias are discussed under figure 20 below.

Residuals are plotted against date (fig. 16) to show regression error with time. It can be seen that residuals during the 1996-1999 regression period seem to oscillate back and forth across the zero percent residual line by as much as 30 percent, so the regression is not completely satisfactory with respect to time. Also, the residuals are consistently negative for the 2000 water year in question. This could indicate a systematic error in the computed flows for the 2000 water year being reviewed. This plot can warn the user that some of the published flows in the regression period are in error, as indicated by unusually large residuals on the plot. If so, the user should investigate computations for those periods to see if an error was made. Options 1 or 2 can be run for the period of record at a station to detect possible errors in the historic record.

Residuals also are plotted against month and day of year (fig. 17) to show possible seasonal biases in the regression equation. If 4 years of flow data were specified by the regression period, then the percent residuals for January 1 for all 4 years would be plotted at January 1 on the graph. Residuals are uniformly scattered about the zero percent line appropriately in figure 17. If seasonal differences are pronounced, such as when the differences form a sinusoidal curve about the zero percent line, the code of the program can be modified to adjust for them, if several years of data are available. The adjustment could involve making a sinusoidal adjustment using day of the year.

Residuals are plotted against the “best” lagged index station flow (fig. 18) to show the possible regression bias with flow magnitude. For example, if the residuals are not uniformly scattered about the zero percent line at low flows, the regression may over or underestimate flows in that range. The residuals in figure 18 are not uniformly distributed about the zero percent residual line. The regression may use several lagged flows, so the lagged flow with the largest regression coefficient is used for this plot.

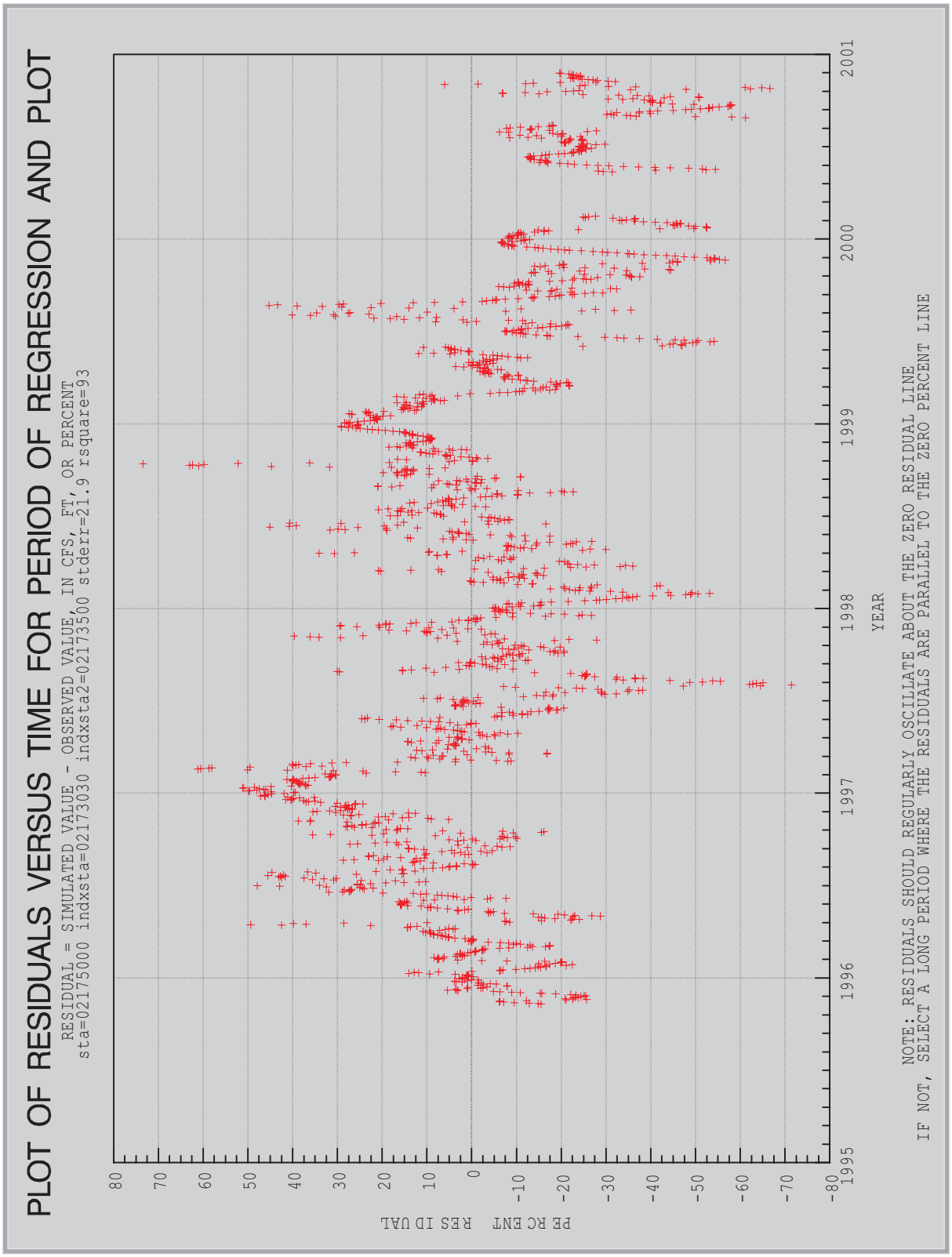


Figure 16. Plots of residuals output by Options 1 and 2 against time period.

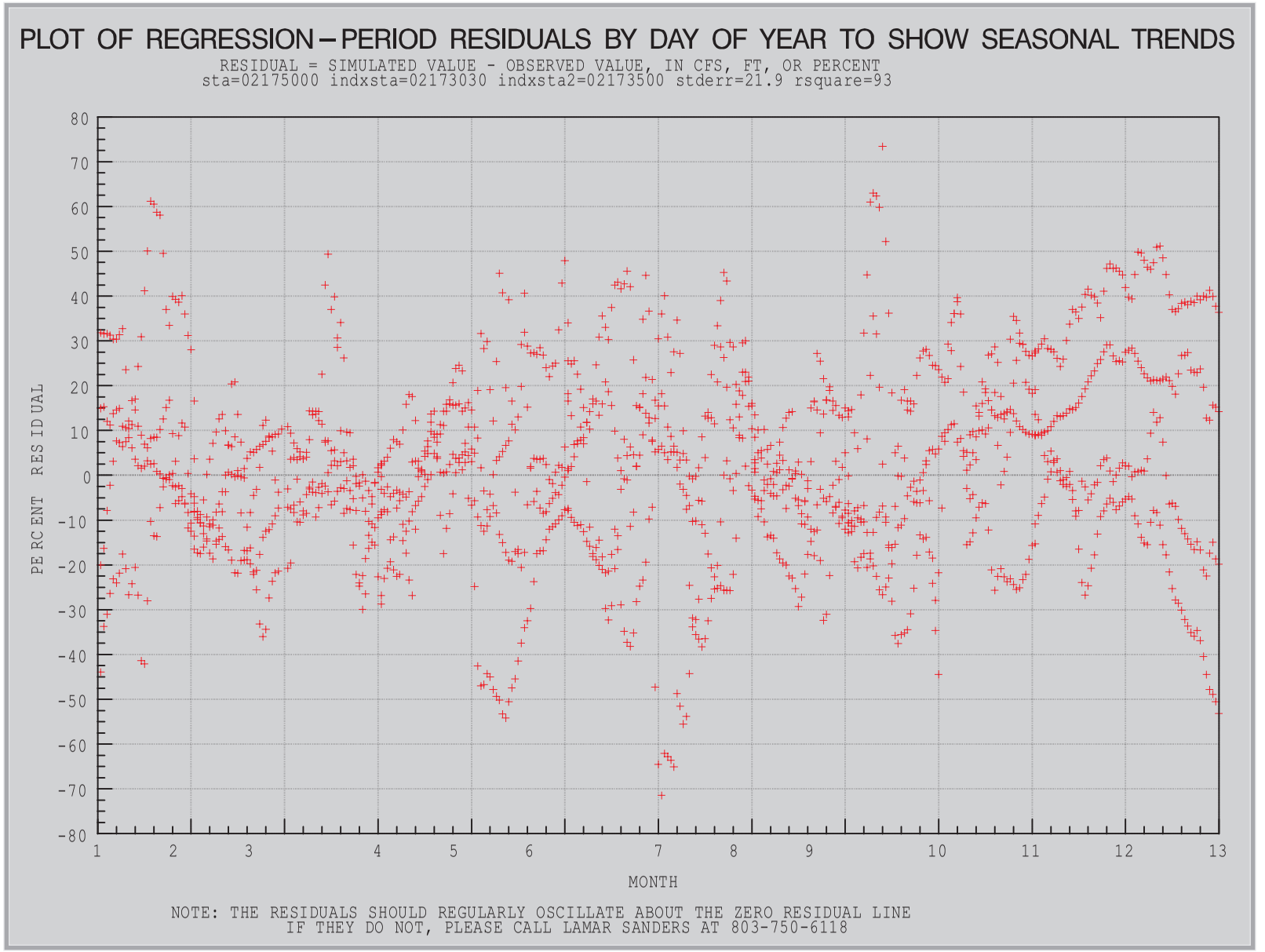


Figure 17. Plots of residuals output by Options 1 and 2 against day of year.

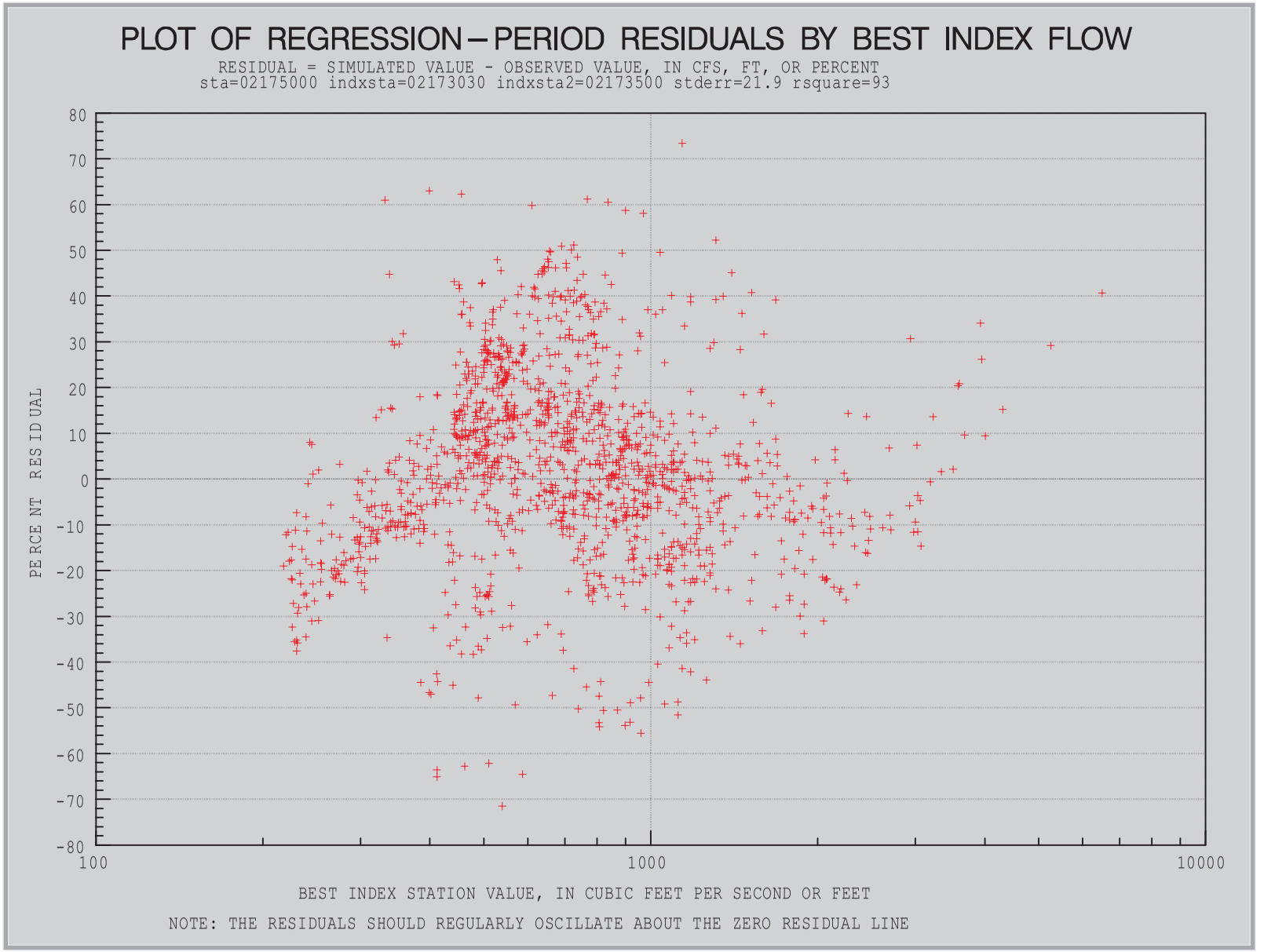


Figure 18. Plots of residuals output by Options 1 and 2 against best index station flow.

Figures 19-25 illustrate the HYDCOMP method for hydrographic comparison and estimates of flow data. The daily value hydrographs have horizontal scales of 2 months per plot. Using figure 19, the simple steps for hydrographic comparison are:

1. The standard error, tabulated at the top of the page, is 21.9 percent, which is about normal for regressions for streams in South Carolina, and indicates a usable regression equation, as suggested in table 1 of this report.
2. The R-square is 93 percent, which shows that there is no major problem with the regression.
3. The regression used index stations 02173030 and 02173500. The user should verify that these stations are hydrologically similar. The two index stations used in this example are the two main tributaries to the stream in question and are, therefore, hydrologically suitable as index stations. Also, the user should verify that flows at the index stations were reviewed and finalized before being used in the regression.
4. The horizontal long-dash short-dash line in the lower and upper parts of the hydrograph are the minimum and maximum flows at the review station that were used in the regression. If the observed flows and confidence-limit flows are very far outside this range of flows, then the regression has been extended too far outside the data used to develop the regression equations. If so, the user should use a regression period where data are available in this range of flows and rerun the regression. The regression and plot periods can be completely separate periods, if necessary. Figures 23-25 indicate that, for this example, a period having lower flows should have been used in the regression.
5. The user must check to see if the observed-flow hydrograph for the review station, indicated by the solid line with triangles in the graph has a reasonable shape, without unusual flat places or sharp discontinuities. On August 22-23, 2000 (fig. 24), the shape of the observed-flow hydrograph has an unusual shape consistent with a clogged intake that suddenly became unclogged, a problem that had been known to exist at this station.
6. The user must check to see if the observed-flow line is not outside the 95 percent confidence limit lines (the dashed lines) for prolonged periods, perhaps 5 days or longer. The “95 percent” indicates that some data may fall outside these limits and still be correct. In particular, the observed-flow data may fall outside the confidence limits around the peaks of flood events, because of timing problems with the regression. However, data on figures 19, 20, and 24 in particular show that flows are outside the confidence limits for several days at a time. In figure 19, the regression flows (shown by the dotted line and X’s) seems flatter than the observed flows. One might suspect that there could be intake trouble at one of the index stations.
7. The user then must check to see if the plotted discharge measurements verify the observed-flow line on the hydrographs. For this example, all the plotted discharge measurements match the observed flow line. However, for streams with small drainage areas and sharply changing flows with time, the discharge measurement may or may not plot close to the observed-flow line. If an observed-flow line falls outside the 95 percent confidence limits, but is verified by a flow measurement, then the regression equation or the flow data at the index station may be in error.

8. If the observed flows are outside the confidence limits for an extended time, the user should be aware that flows at the index station may be wrong, rather than flows at the review station. In addition to closely examining computations at both the review station and index stations, the user can do separate hydrographic comparisons using at least two different index stations. For example, if hydrographs are generated separately using index stations A and B, and flows for the review station (Station C) plot outside the confidence limits for both the station A plot and the station B plot, then the review station (Station C) may indeed be wrong. If flows for the review station (Station C) plot outside the confidence limits for index station A, but not outside the confidence limits for index station B, then there may be an error with index station A. In general, it is a good practice to conduct hydrograhic comparisons, using two or three index stations separately.

9. If the user finds periods of record where the flows are not missing, but obviously wrong for certain reasons (for example, the float is on mud or hung up, or flows are affected by backwater), the user must make a note of the periods, then rerun HYDCOMP, entering those periods in the fields of Screen 2 to request estimated flows for the erroneous data.

Flows were not missing at this station for the entire year. However, for illustration purposes, HYDCOMP was requested to estimate flows for the periods depicted by squares in figures 21, 22, and 24. The period of missing data in figure 24 is fairly short, and it can be seen that the estimated data are smoothed into the non-missing data at the edges of the gap. The estimated data retains the shape of the regression data depicted by the dashed line with X's. One can see that if the regression flows were used without adjustment, the resultant hydrograph would be unrealistically discontinuous at the edges of the gap in data.

The longer period of estimated data in figures 21 and 22 illustrates the results of the Kalman smoothing algorithm. Suppose a completely unrealistic gap in the data existed, such as 300 days. The estimated data should match the non-missing data at the edges of the gap. But, this adjustment should not be applied toward the middle of the gap, where the averaged results of using the unadjusted regression flows would be much more accurate. If the regression flow hydrograph line varies sharply from one side to the other of the observed data line on the hydrograph, the Kalman algorithm will smooth directly to the regression line rather quickly at both ends of the gap in data. However, for this station, the regression line falls above the observed-flow line for about 2.5 months (figs. 22-24). Therefore, the fact that the regression does not match at the edges of the gap may have some long-lasting effect on the simulation of flows. Observe the simulated flows in figures 21 and 22 and the adjustment factors in the table on figure 14. It can be seen that the Kalman algorithm smooths from correction factors of 1.36 and 1.43 at the edges of the gap to a constant correction factor of about 1.15 toward the middle of the gap over a period of about 27 missing days. This slowly changing and constant correction factor seems reasonable, given the long period in figures 22-24 where the regression line did not cross the observed flow line for about 2.5 months. Therefore, this example shows that the Kalman algorithm seems to be a reasonable alternative to manual smoothing.

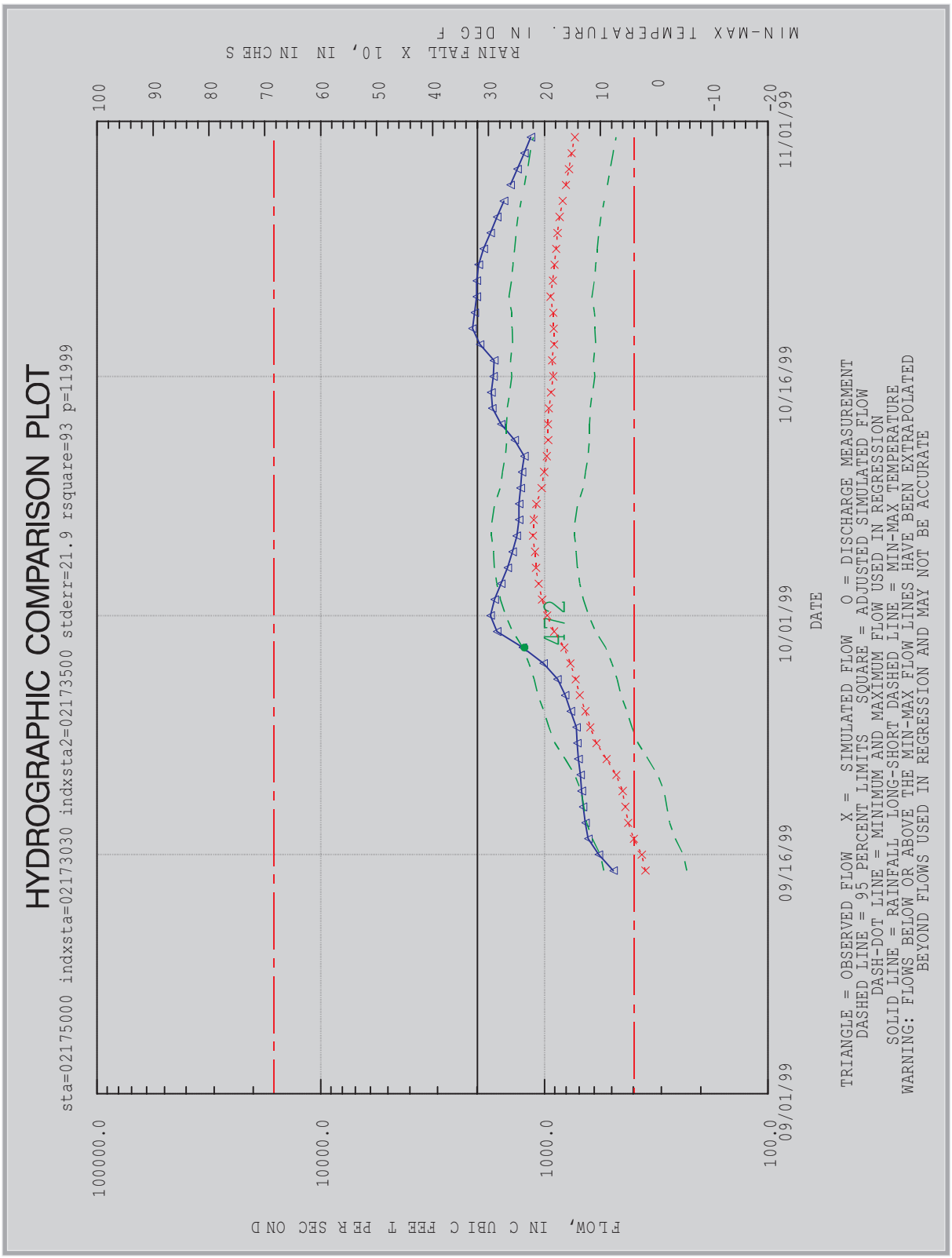


Figure 19. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for September-October, 1999.

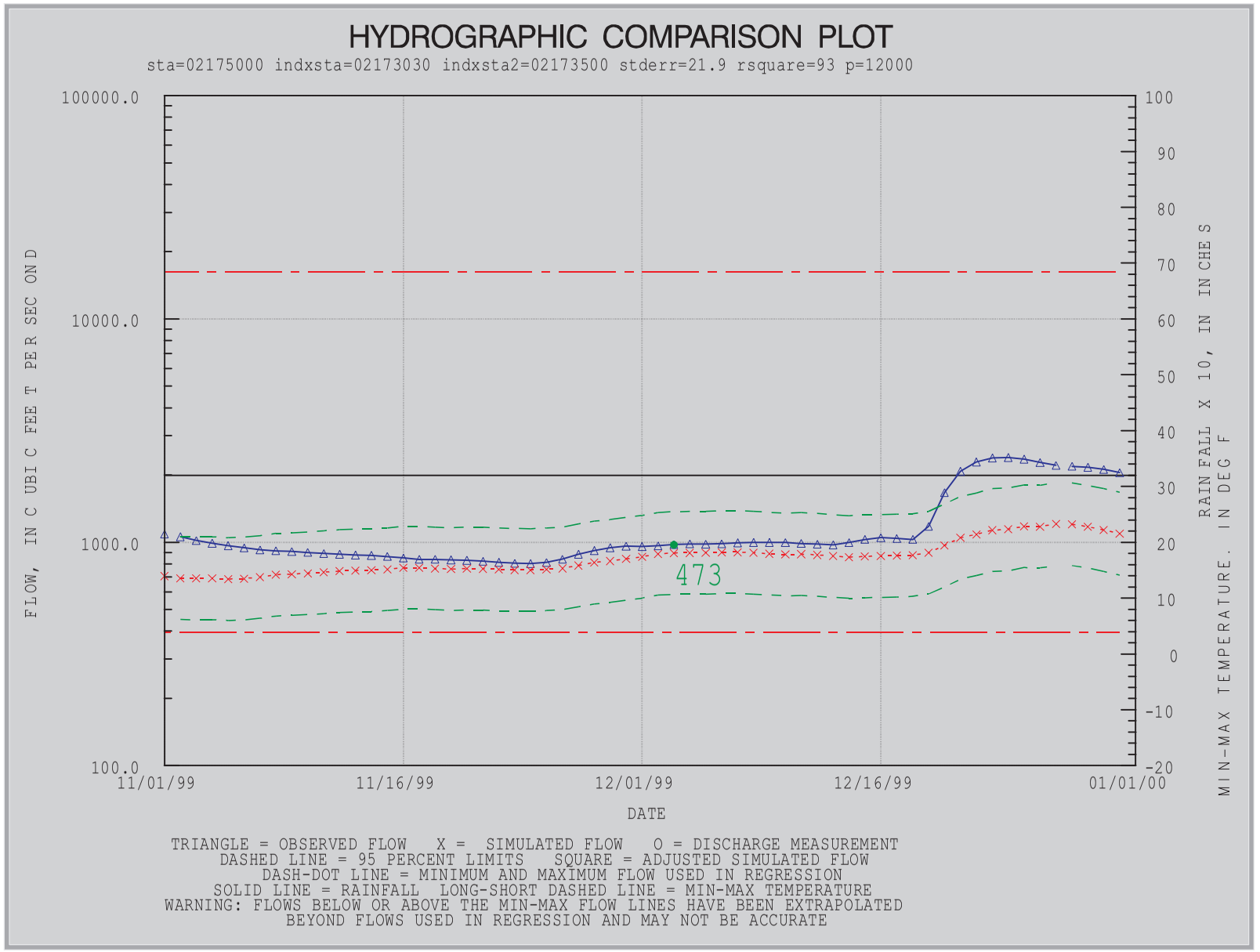


Figure 20. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for November-December, 1999.

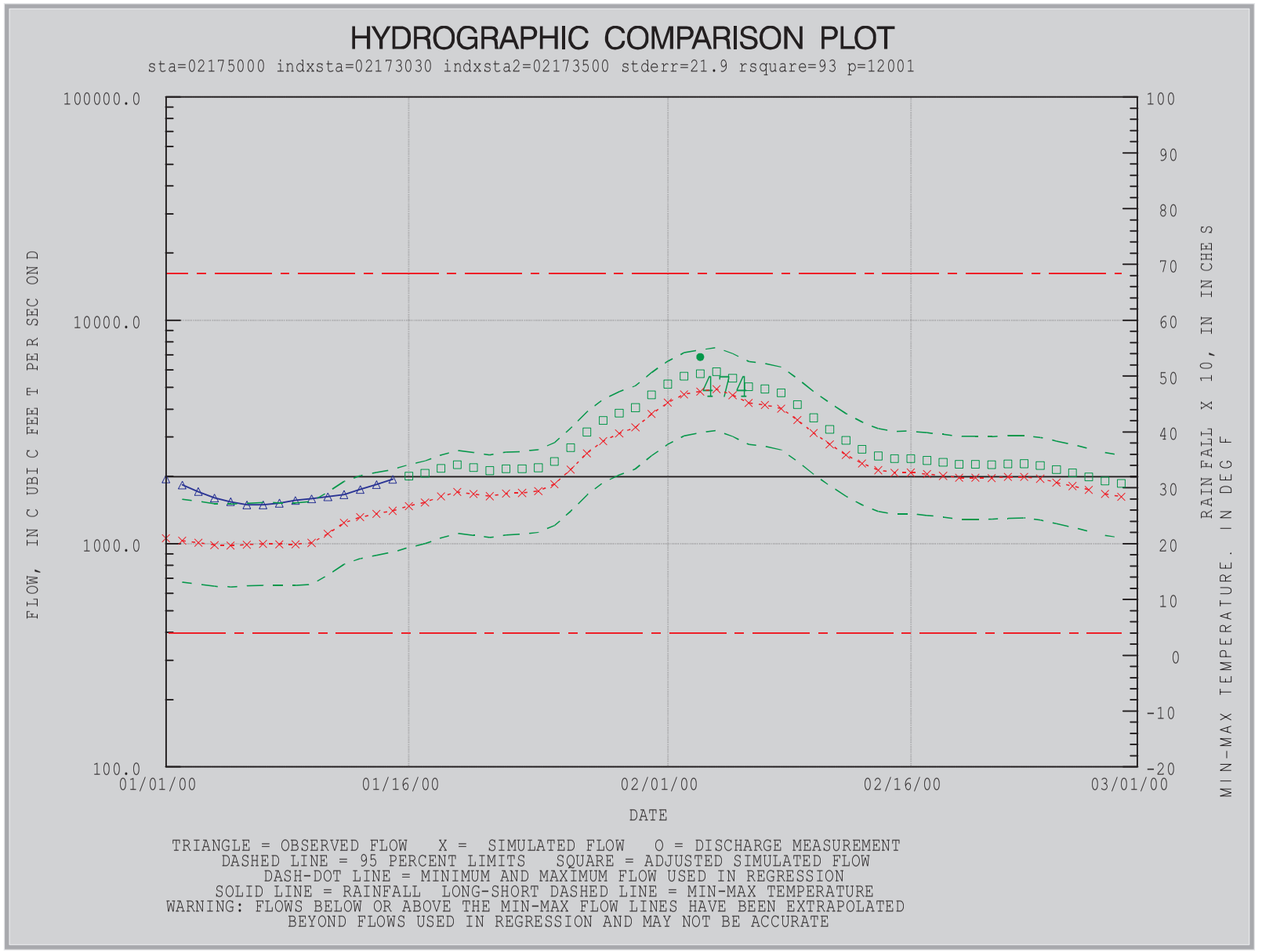


Figure 21. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for January-February, 2000.

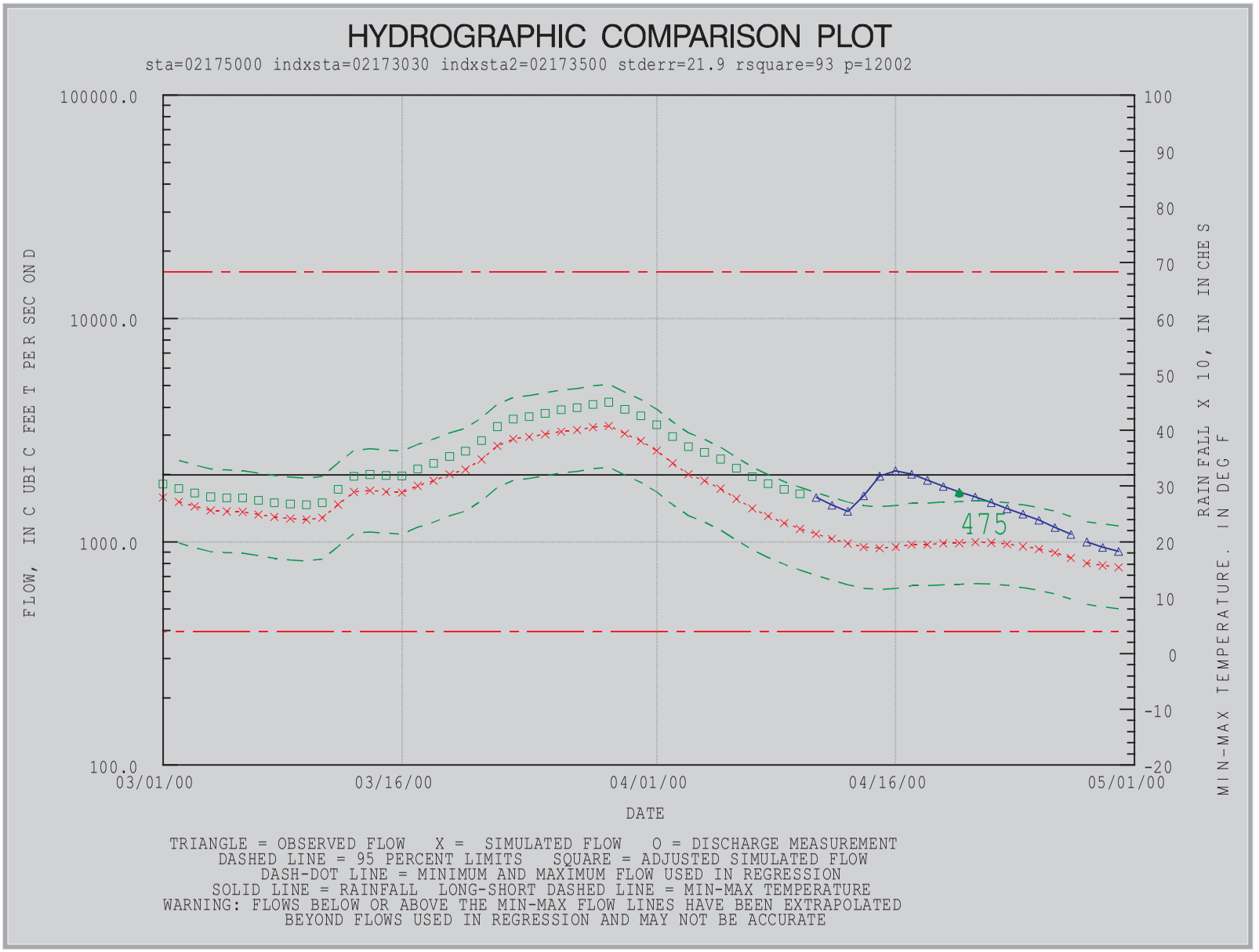


Figure 22. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for March-April, 2000.

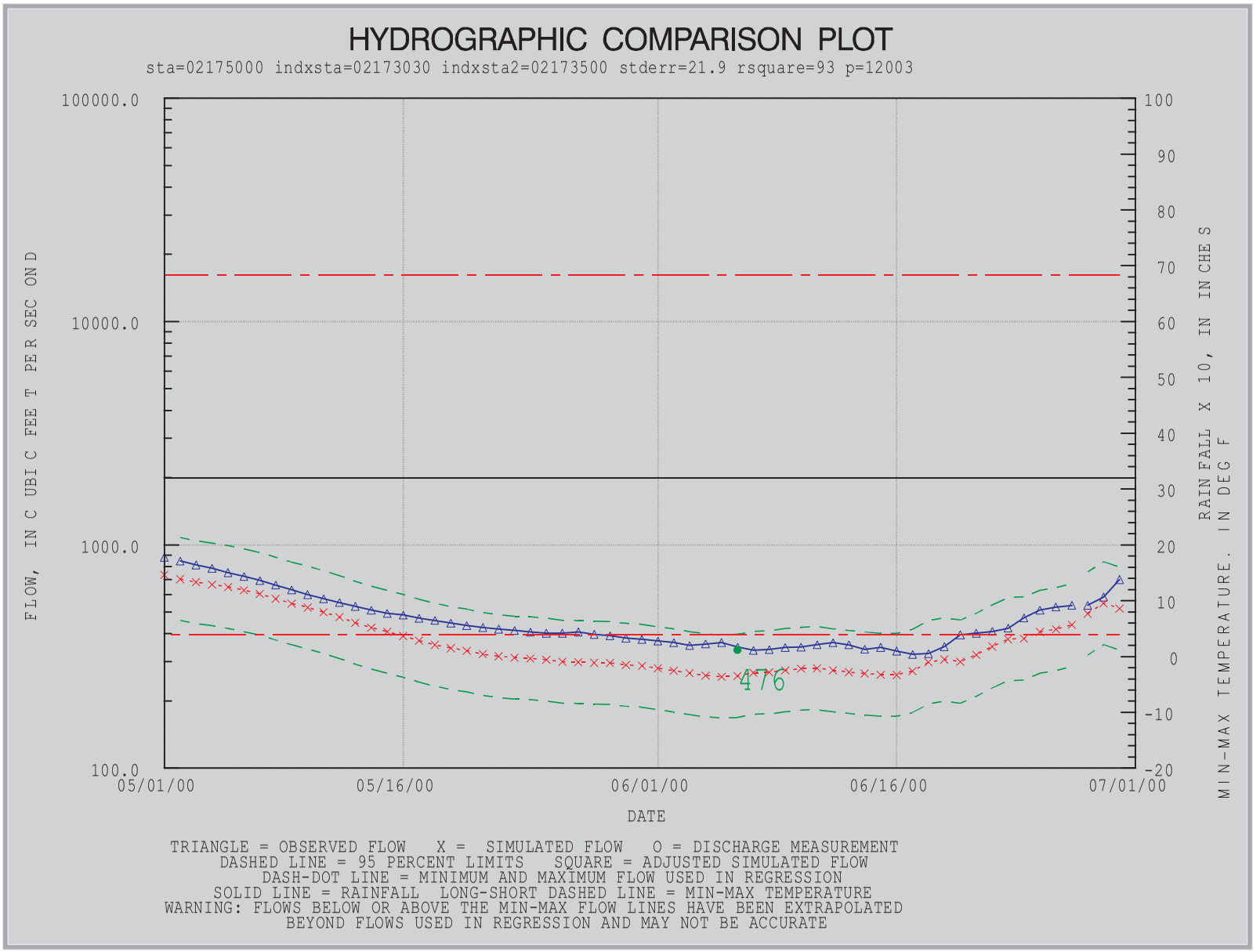


Figure 23. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for May-June, 2000.

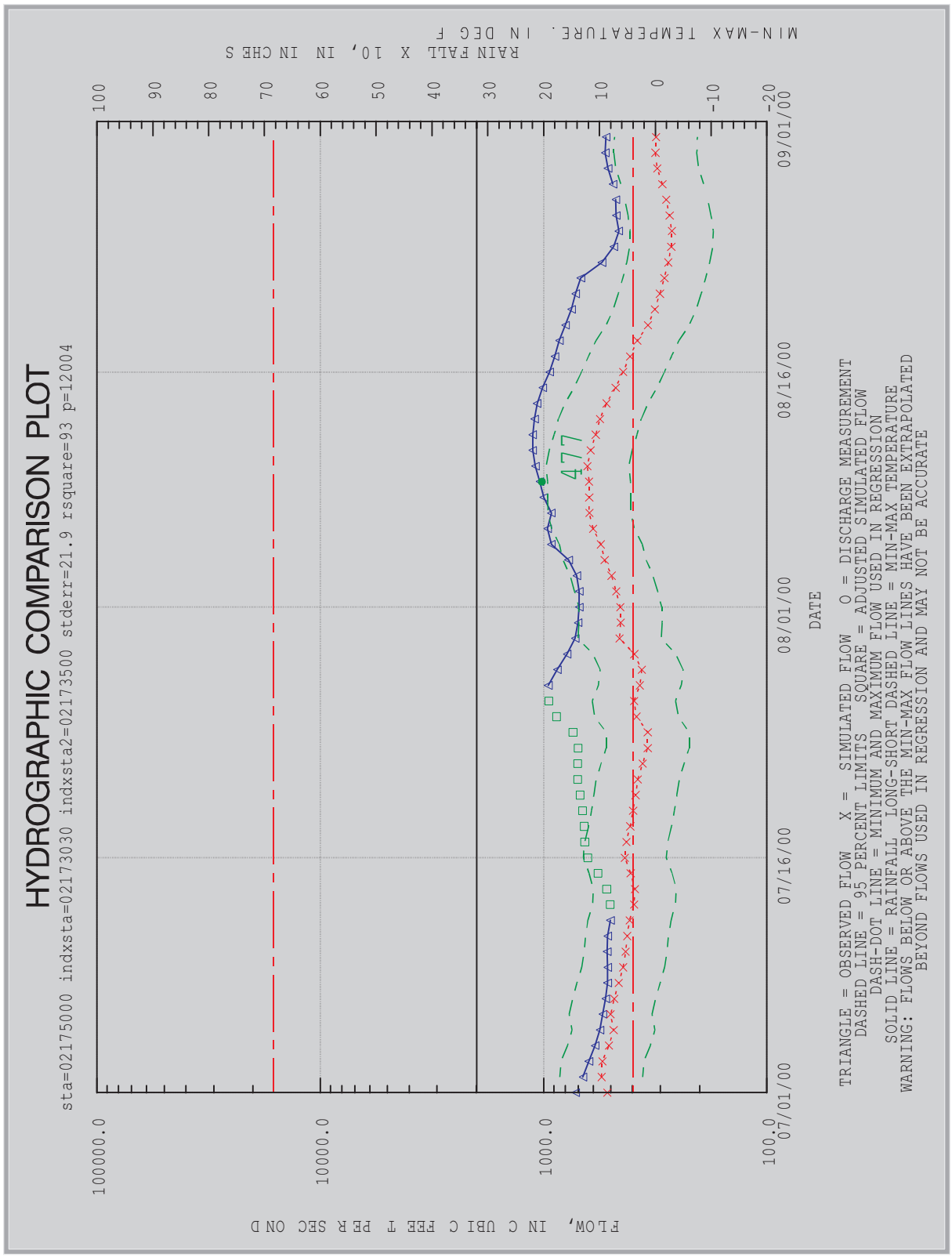


Figure 24. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for July-August, 2000.

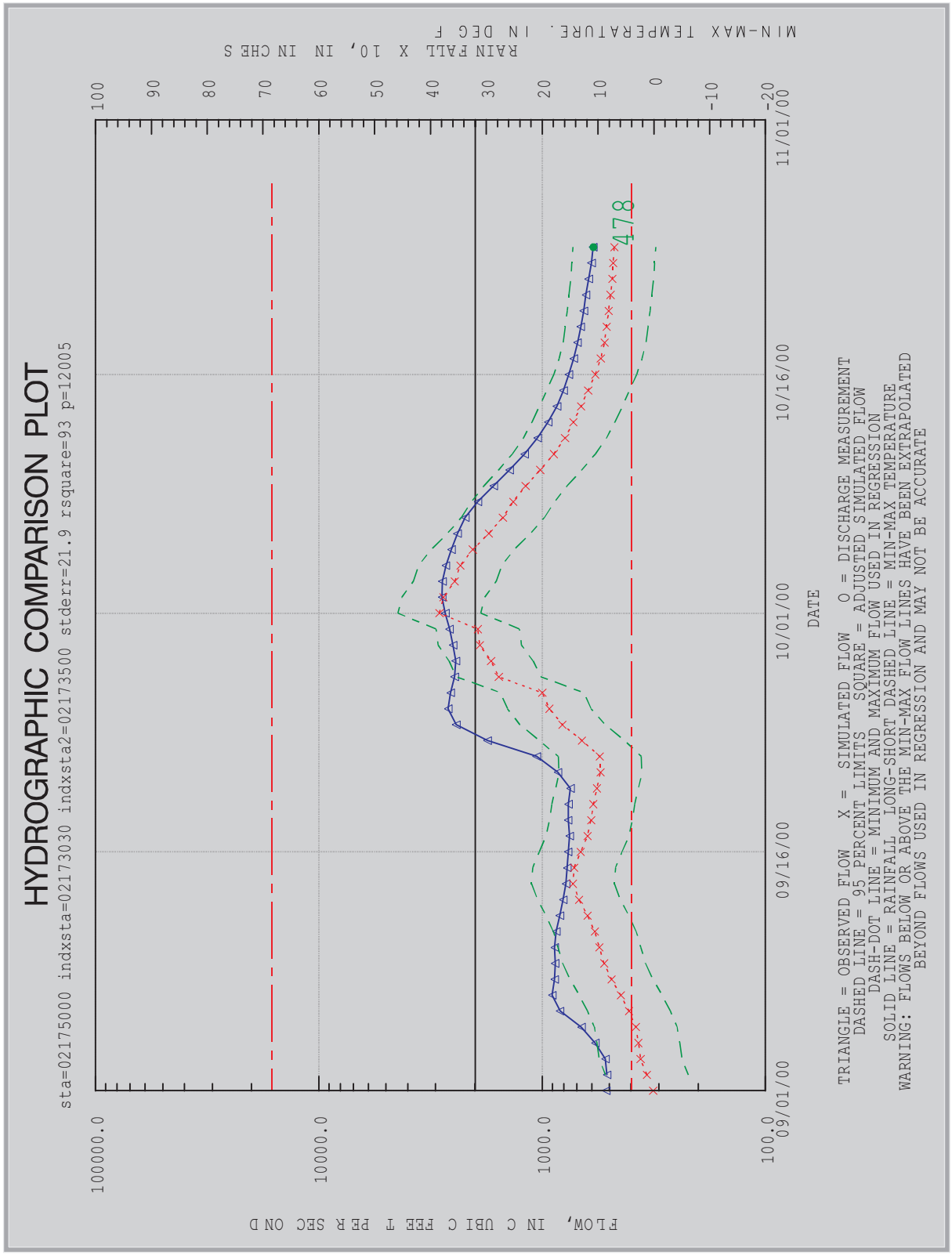


Figure 25. Hydrographs of observed flows, simulated flows, 95 percent confidence limits, and measured flows output by Options 1 and 2 for September-October, 2000.

SUMMARY

HYDCOMP is written in the SAS software language and can be used to (1) automatically select the five most-correlated index stations for each review station in a State by correlation and regression, (2) perform quality control checks by hydrographic comparisons of daily value data using regression, and (3) estimate missing daily value data by regression.

To select the most-correlated index stations, an initial list of 20 index stations with the highest correlation coefficients is produced. The standard error of estimates from regression using -8 to +8 lags of daily flows from the previous water year of published, finalized data, and day of week as explanatory variables are computed for 5 index stations from a list of 20 index stations having the highest correlation coefficients. HYDCOMP automatically loads the selected index stations into the database used for hydrographic comparison for review stations not already entered into the database. The list is also stored in a file and printed. The user can interactively access the list of index stations one review station at a time. If the database has already been populated, rerunning option 4 using different retrieval criteria will not overwrite the loaded database. In this case, the user can select index stations from this list of five index stations having the lowest standard errors of estimate that are hydrologically similar to the review station. The selected index stations can be entered into the database for the hydrographic comparison and data estimate part of the program. Once the database has been loaded, other users have only to access the hydrographic comparison and data estimate part of the program, select a review station, select from the pre-selected index stations, and enter dates for plots and data estimates.

The hydrographic comparison and estimates of missing data part of the program performs regressions using -8 to +8 lags of daily flows from the previous water year of published, finalized data, and day-of-week as explanatory variables. The resulting regression equations are used to estimate flows and 95 percent confidence limits for the current, non-finalized year of data. For hydrographic comparison, the user evaluates residual plots, and checks to see if the non-missing data at the review station are outside the 95 percent confidence limits for several consecutive days on the daily value hydrograph. Additionally, the user checks to see if the non-missing data are verified by plotted flow measurements on the hydrograph, and examines the non-missing hydrograph for shapes indicating erroneous data.

Missing data are estimated by adjusting the regression flows to match non-missing flows at the edges of the missing data using the Kalman smoothing algorithm. In addition, adjusted flows can be produced for periods where data are erroneous, but not missing.

REFERENCES

Grewal, M.S., and Andrews, A.P., 1997, Kalman filtering theory and practice (4th ed.): Englewood Cliffs, N.J., Prentice Hall, 381 p.

Hirsch, R.M., 1982, A comparison of four record extension techniques: Water Resources Research, v. 18, no. 4, p 1081-1088.

National Water Information System, 1997, Automated data processing system, ADAPS user's manual: U.S. Geological Survey, approximately 400 p.

SAS Institute, Inc., 1993, SAS language: reference, version 6, (1st ed.): Cary, N.C., 1042 p.

APPENDIXES

APPENDIX 1: PROGRAM SOFTWARE COMPONENTS

HYDCOMP utilizes several shell programs and SAS macros, programs, and data files, as described in the overview below. This documentation is provided for use by personnel responsible for maintaining the program.

hydcomp.sh is the shell program that presents options for selection by the user, and executes HYDCOMP.

group.dv.sh is a shell program that retrieves daily value data from the ADAPS database for the **mac.corr** macro that correlates every review station in the State to every other index station in the State in option 4. This shell program executes the ADAPS program and answers interactive ADAPS queries; therefore, **group.dv.sh** is subject to updating whenever the interactive ADAPS queries change. Its arguments are:

1. Name of group file for daily value data retrieval.

2. Begin date, in MM-DD-YYYY format.

3. End date, in MM-DD-YYYY format.

run.group.dv.sh is a shell program written and executed from within the **mac.corr** macro which correlates every review station in the State with every other index station in the State in option 4. It executes the **group.dv.sh** shell program. Its arguments are the same as for **group.dv.sh**.

mac.hyd1 is a SAS macro that does the regressions, graphic hydrographic comparisons, and estimates of missing data for options 1 or 2. A SAS macro is a programming subroutine made up of SAS programming statements. Information is passed in and out of the macro by means of arguments. Its arguments are:

1. ***outpt*** directs plotted outputs to the screen (*output* = sc), or to the printer (***outpt*** = pr).

2. ***nindx*** indicates the number of index stations (*nindx* = 1 or *nindx* =2) to be used in the regression.

run.hyd.1.1, **run.hyd.1.2**, r**un.hyd.2.1**, and **run.hyd.2.2** are SAS programs that call **mac.hyd1** from within the **hydcomp.sh** shell program, depending on the options selected within the hydcomp.sh shell program. The arguments are the same as those for **mac.hyd1**.

mac.gopt is a SAS macro included in the **mac.hyd1** file that directs plotted output to a postscript file if the ***outpt*** argument of **mac.hyd1** equals “pr” or the terminal screen if the ***outpt*** argument is “sc.” Its only argument is ***filen***, which is the name of a post script file to which the plot will be written for subsequent spooling.

mac.corr is a SAS macro that produces the five most-correlated index stations for every review station in the State in option 4. There are no arguments.

mac.sta1 is a SAS program (not macro) that gives a list of the five most-correlated index stations for one review station in option 5. (A SAS program is composed of SAS program statements and calls to SAS procedures (PROCS), DATA steps, and SAS macros.) There are no arguments.

s.load.lib is a SAS program that identifies the directory holding the SAS files below. There are no arguments.

s.fsedit.2 is a SAS program that presents the screen forms for data entry for option 3. There are no arguments.

rdbsas.sas is a SAS macro that converts a relational database (rdb) format file into a SAS work file using program **rdb_get**.

rdb_get is a "C" programming language binary-executable program utilized by the **rdbsas.sas** macro to convert the rdb-format data to a SAS work file.

nwts2rdb is a NWIS program that produces Relational Database (RDB) files of unit-value, daily value, and measurement data from the NWIS database.

sasrdb.sas is a SAS macro that converts a SAS file to a rdb-format file using program **rdb_put**.

rdb_put is a "C" programming language executable-binary program utilized by the **sasrdb.sas** macro to convert the SAS file to a rdb-format file.

best5sta.sas7bdat is a SAS file produced by **mac.corr** and option 4 containing the most-correlated index stations for every review station in the State, associated ADAPS data-descriptor (dd) numbers, standard error of estimate, and R-square computed. This file is called the "most-correlated index station" file herein. An ADAPS dd number identifies a category of daily value data stored in the ADAPS data files. It should be kept in mind that these are the "most-correlated" index stations, and that the final, selected index stations are stored in the file described below.

stadat.sas7bdat is a SAS file utilized by **mac.hyd1** and options 1, 2, and 3. It contains the final, selected index stations for every review station in the State, and all the information shown on the screen forms for options 1, 2, and 3. This file is called the "selected-index station" file herein.

cordat.sas7bdat is a SAS file utilized by **mac.corr** and option 4. It contains the ADAPS group file name, begin-end dates, month-day definition of seasonal periods, allowable number of missing days, and computer identity of the owner of the group file.

sta1.sas7bdat is a SAS file for **mac.sta1** and option 5 containing the number of the last review station entered into option 5.

scstadat.sas7bcat is a SAS file containing the information necessary to format the screen forms for **mac.hyd1** and options 1, 2, and 3.

sccordat.sas7bcat is a SAS file containing the information necessary to format the screen forms for **mac.corr** and option 4.

scsta1.sasbcat is a SAS file containing the information necessary to format the screen forms for **mac.sta1** and option 5.

APPENDIX 2: PROGRAM INSTALLATION

The installation shell script **hydcomp_install.sh** is available for downloading over the Internet from the public anonymous File Transfer Protocol (FTP) server **sun1dsccmb.er.usgs.gov (internet address 144.47.8.248)** in the directory **/pub/hydcomp**. When the system administrator executes the installation script file **hydcomp_install.sh** with root access from any directory on the SUN computer, it will transfer, install, and create the necessary programs, directories, and links. The HYDCOMP software should be installed on the same SUN computer where the USGS ADAPS software and database reside. The HYDCOMP program utilizes the commercial SAS software package marketed by the SAS Institute, Inc. (1993). Therefore, a copy of the SAS software needs to be installed on the same SUN computer where HYDCOMP and the ADAPS software and database reside.

The site administrator must also make the following changes after executing **hydcomp_install.sh**:

1. In the uppermost SAS directory (such as /usr/opt/sas), a modification has to be made to the **sasv8.cfg** file. The site administrator must change the directory designated by the **-work** parameter to **/tmp**. The **/tmp** directory should have enough space for work computations being done by HYDCOMP.
2. The site administrator may have to make two modifications to the **hydcomp.sh** shell program in the **/usr/opt/wrdapp/locapp/hydcomp** directory where the HYDCOMP programs and files are stored.
   a. The location and number of versions of the SAS software may vary from District to District. Therefore, it is necessary to tell HYDCOMP which version to use by the ***SASIT*** variable imbedded in the hydcomp.sh code. Note that this is the full pathname to the binary executable SAS program.
   b. The default access rights on most USGS computers are set such that any person writing to an existing file seizes ownership of the file and excludes anyone else from writing to the file thereafter. Therefore, the site administrator must determine what the default access rights on the UNIX system are for the District, and modify them using the ***UMASK*** command embedded in the **hydcomp.sh** code. The access code after modification by the ***UMASK*** command should be 666 so that multiple users can read and write to the files containing the character string "**sas7**" imbedded in the file names.

      If difficulties in establishing access rights are experienced, a user may have seized access rights to the SAS files identified by the included character string "SAS7" in the **/usr/opt/wrdapp/locapp/hydcomp** directory. If so, the site administrator must reset these access codes to 666 before adjusting by the *UMASK* command.

3. The SAS program will be sending graphics to the user's terminal screen. Therefore, if special commands, such as the "setenv" command, are necessary to enable graphics to come to the user's terminal, the site administrator must inform the user what these commands or procedures are.

APPENDIX 3: PROGRAM MAINTENANCE

Maintenance on the HYDCOMP program may be required occasionally as described below. Most changes will be have to be made because of updated versions of the NWIS and SAS software. These changes could include the following task:

1. Changes might have to be made to HYDCOMP because of changes to the NWIS **nwts2rdb** program that produces RDB files from the NWIS database. Prior notice of changes to the **nwts2rdb** program is usually not issued by NWIS to users of the program, so maintenance personnel should be prepared to inquire about the changes if HYDCOMP fails to work after a new release of NWIS software. In addition, because the ADAPS menus sometimes change, the person maintaining HYDCOMP may need to revise the **group.dv.sh** shell file used by **mac.corr** in option 4.

2. Generally, changes should not be necessary with new releases of SAS, but the following changes may have to be made:

   a. The SAS data files (files having "sas7" embedded in the file names) may have to be converted to new SAS file formats. Usually, this is not necessary, but conversion has been necessary in the past. The conversions are done by the easily used SAS "CPORT" and "CIMPORT" procedures (PROCS). PROC CPORT is executed in the original version of SAS to create transport files for conversion by PROC CIMPORT to the new SAS data format. PROC CIMPORT is run using the newer version of SAS.

   b. It is unlikely, but minor changes may be made in some of the other SAS procedures used in HYDCOMP.

District site administrators may have to deal with the following problems.

1. In UNIX, as discussed above, default access rights are such that if person A allows person B to read and write to one of person A's files, then person B seizes ownership to person A's file, who is then denied write-access to his/her own file. An attempt is made to avert this by the UMASK command in the **hydcomp.sh** shell file, as described above. If access problems arise, the site administrator may have to modify the UMASK command specifications, and reset access rights for the SAS datasets and the HYDCOMP shell file programs.

2. If HYDCOMP is aborted for some reason, temporary work files are left in a directory with the prefix "SAS" in the **/tmp** directory. These files could be quite large, so the site administrator should look for them in the **/tmp** directory now and then and if found, delete them using the "rm -r SAS*" command.