Computer Programs for Obtaining and Analyzing Daily Mean Streamflow Data from the U.S. Geological Survey National Water Information System Web Site

Appendix 3. Streamflow Record Extension Facilitator (SREF Version 1.0)—A program for extending and augmenting available streamflow data using long-term streamflow records from hydrologically similar sites

By Gregory E. Granato

Open-File Report 2008–1362

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

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

Contents

Figures

Tables

Conversion Factors and Abbreviations

Multiply	By	To obtain
Area		
square mile (mi^2)	259.0	hectare (ha)
square mile (mi^2)	2.590	square kilometer (km^2)
Volume		
cubic foot (ft^3)	0.02832	cubic meter (m^3)
Flow rate		
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
cubic foot per second per square mile [(ft^3/s)/mi^2]	0.01093	cubic meter per second per square kilometer [(m^3/s)/km^2]

Abbreviations

1B3	1-day 3-year biological low flow
4B3	4-day 3-year biological low flow
7Q10	7-day 10-year low flow
ADAPS	USGS Automated Data Processing System
BCF	bias correction factor
BMP(s)	best management practice(s)
CD-ROM	computer disk-read only memory
dv	daily value
dd	data descriptor
EMC	event mean concentration
FHWA	Federal Highway Administration
GNWISQ	Get National Water Information System Streamflow (Q)
KTRLine	Kendall-Theil robust line
log10	common logarithm
MkDF	Make U.S. Environmental Protection Agency DFLOW3 batch input Files
MkPP	Make plotting position file
MOVE	maintenance of variance
NWIS	National Water Information System
NWISWeb	National Water Information System Web
PC	Personal Computer
Q	Streamflow
QSTATS	Streamflow (Q) Statistics
RDB	relational database
ROS	regression-on-order statistics
SELDM	stochastic empirical loading and dilution model
SQL	structured query language
SREF	Streamflow Record Extension Facilitator
TMDL(s)	Total Maximum Daily Load(s)
USGS	U.S. Geological Survey

Appendix 3.   Streamflow Record Extension Facilitator (SREF Version 1.0)—A program for extending and augmenting available streamflow data using long-term streamflow records from hydrologically similar sites

By Gregory E. Granato

Abstract

The Streamflow Record Extension Facilitator (SREF Version 1.0) program was developed by the U.S. Geological Survey (USGS) in cooperation with the Federal Highway Administration to provide an estimated long-term record of daily mean streamflows (record extension) or long-term estimates of streamflow statistics (record augmentation) at sites with limited data. SREF can be used to extend or augment limited data from a partial-record site or short-term streamflow-gaging station by using data from a representative long-term continuous-record streamflow-gaging station. Record extension and augmentation are based on the assumption that long-term streamflow records from hydrologically similar sites can be used to estimate values at a site of interest with limited data. The SREF program implements the maintenance of variance extension type 1 (MOVE.1) method and the maintenance of variance extension type 3 (MOVE.3) method. SREF also facilitates record extension with a single or multisegment regression model developed by use of the nonparametric Kendall-Theil Robust Line (KTRLine) program. SREF provides graphical displays of the data, regression lines, and residual statistics to facilitate the analytical process. The program outputs a file of paired data values and a file with regression statistics. The user may output files containing an estimated record of daily mean streamflow data and files containing estimates of selected long-term statistics for the site of interest. This computer program has a graphical user interface that follows standard Microsoft Windows interface conventions.

Introduction

Stream-discharge data, defined herein as published daily mean stream-discharge values, are necessary to define statistics for potential streamflows at gaged sites and to estimate streamflows at ungaged sites. The U.S. Geological Survey (USGS) streamflow data collection program is designed to provide streamflow data at gaged sites and to provide information that can be used to estimate streamflows at almost any point along any stream in the United States (Benson and Carter, 1973; Wahl and others, 1995; National Research Council, 2004). Streamflow statistics from such data are important for many planning and design activities including water-supply analysis, habitat protection, bridge and culvert design, calibration of surface- and ground-water models, and water-quality assessments. Streamflow information is especially critical for water-quality assessments (Warn and Brew, 1980; Di Toro, 1984; Driscoll and others 1989; Driscoll and others, 1990a, b; U.S. Geological Survey, 2006). Calculation of streamflow statistics for receiving waters is necessary to estimate the potential effects of point sources such as wastewater-treatment plants and nonpoint sources such as highway and urban-runoff discharges on receiving water. Streamflow statistics indicate the amount of flow that may be available for dilution and transport of contaminants (U.S. Environmental Protection Agency, 1986; Driscoll and others, 1990a, b). Streamflow statistics also may be used to indicate receiving-water quality because concentrations of water-quality constituents commonly vary naturally with streamflow. For example, concentrations of suspended sediment and sediment-associated constituents (such as nutrients, trace elements, and many organic compounds) commonly increase with increasing flows, and concentrations of many dissolved constituents commonly decrease with increasing flows in streams and rivers (O'Connor, 1976; Glysson, 1987; Vogel and others, 2003, 2005).

Reliable, efficient, and repeatable methods are needed to access and process streamflow information and data. For example, the Nation's highway infrastructure includes an innumerable number of stream crossings and stormwater-outfall points for which estimates of stream-discharge statistics may be needed. The USGS maintains the National Water Information System (NWIS), a distributed network of computers and

file servers used to store and retrieve hydrologic data (Mathey, 1998; U.S. Geological Survey, 2008). NWISWeb is an on line version of this database that includes water data from more than 24,000 streamflow-gaging stations throughout the United States (U.S. Geological Survey, 2002, 2008). NWISWeb also has USGS data for than 46,000 sites with concurrent measurements of stage and flow and more than 24,000 surface-water-quality monitoring sites with one or more measurements of stage and flow in the conterminous United States that may provide short-term data at or near a potential site of interest. Information from NWISWeb is commonly used to characterize streamflows at gaged sites and to help predict streamflows at ungaged sites.

Long records (years to decades) are commonly needed to characterize variations in streamflow through time. For example, at least one decade—preferably two to four decades—are considered necessary to develop a reliable estimate of the 7-day 10-year (7Q10) hydrologic low flow (U.S. Environmental Protection Agency, 1986), and at least 20 years of data are recommended for analysis of the 1-day 3-year biological low flow (1B3) and the 4-day 3-year biological low flow (4B3), which are used for setting water-quality criteria and total maximum daily load (TMDL) waste-load allocations (Rossman, 1990; Saunders and others, 2004).

Streamflow Record Extension and Augmentation

The Streamflow Record Extension Facilitator (SREF) uses long-term continuous-record streamflow data with relatively short-term continuous-record streamflow data or partial-record streamflow measurements to augment or extend existing data to provide estimates of long-term streamflow characteristics for sites with limited data. A long-term continuous-record streamflow-gaging station is defined herein as a streamflow-measurement site for which stage-discharge relations have been developed and used with automated stage recording devices to produce a record of interpreted daily discharge (parameter code P00060) for a period greater than or equal to twenty years. A partial-record site is defined herein as a streamflow-measurement site without a continuous stage-recording device for which a limited number of instantaneous stream-discharge measurements (parameter code P00061) are available. A short-term continuous-record streamflow-gaging station is defined herein as a streamflow-measurement station for which stage-discharge relations have been developed and used with automated stage recording devices to produce a record of interpreted daily discharge (parameter code P00060) for a period of 1 to 20 years. However, the distinction between a long-term and short-term streamflow-gaging station may differ for different applications, depending on data availability and information needs. In this manual, the term "site of interest" will generally be used to indicate a partial-record site or a short-term streamflow-gaging station for which streamflow estimates will be made, unless this distinction is necessary to frame the discussion.

The SREF program facilitates streamflow-record extension and augmentation. Record extension commonly is defined as a method for estimating individual discharge values for a site of interest from comparable values collected at one or more longer-record site(s), or data from other sites with data that cover a gap in the record at the site of interest (Hirsch, 1979, 1982; Hirsch and Gilroy, 1984; Vogel and Stedinger, 1985). Streamflow-record augmentation commonly is defined as a method for generating long-term estimates of statistics such as the mean, variance, 1B3, 4B3, 7Q10 and other streamflow statistics for a site of interest (Matalas and Jacobs, 1964; Vogel and Stedinger, 1985). Hirsch (1982) demonstrated that the regression method commonly referred to as the line of organic correlation is suitable for record extension or augmentation with maintenance of variance, commonly described as the MOVE method. Hirsch and Gilroy (1984) and Helsel and Hirsch (2002) discuss the relative merits of regression equations developed with ordinary least squares, the line of organic correlation, and least normal squares methods, indicating that the line of organic correlation is the best method of the three for record extension and record augmentation. Vogel and Stedinger (1985) provide methods to adjust the record-extension equations for differences in the mean and variance of streamflows in the short-term concurrent period and the long-term streamflow record. Vogel and Stedinger (1985) designated the adjusted method as MOVE.3. Similar methods designated as MOVE.2 (Hirsch, 1982) and MOVE.4 (Vogel and Stedinger, 1985) have been developed but are more complex than MOVE.1 and MOVE.3, respectively, and do not offer significant improvements in predictive capability.

Record extension or augmentation at a site of interest must be done with an understanding of the potential effects of outliers on the process. If partial-record stations are used, potential differences between instantaneous measurements and the daily mean streamflow available from historical records must be examined. Record extension or augmentation of partial-record data is commonly used to estimate low-flow statistics, so partial-record data is commonly collected only during times of stable-streamflow recession (Riggs, 1972; Ries and Friesz, 2000). Measurement of streamflows over a wide range of values is necessary for extending a record of representative daily mean streamflows at the site of interest. Collection of such a record, however, may be difficult without multiple instantaneous discharge measurements on days with high, rapidly changing streamflows. For a given site record extension using short-term continuous records is more practicable and accurate than extension of a partial record of instantaneous measurements because continuous records of daily mean streamflows are more comparable to index station values than are the instantaneous partial-record measurements. Also, the process of developing a stage-discharge relation with multiple measurements over the range of streamflows for a continuous streamflow-gaging station reduces the potential for individual outliers among measurements. This is because each measurement can be compared to the stage-discharge rating curve as the measurement is collected to determine if it is a

potential outlier. Thus, a measurement can be evaluated in the field and repeated to verify a potential change in the stage-discharge relation (Rantz and others, 1982).

If the user is interested in augmenting a short continuous record or an extensive partial record, results may be improved by selecting values from the record for the site of interest that best characterize the statistic of interest. The means and standard deviations of the sample of concurrent data are used in the MOVE.1 method to develop the regression equation for the line of organic correlation. Streamflow values that are representative of the statistic of interest may improve predictions of flows in the range of interest. For example, a user interested in low-flow statistics may extract low-flow values from the record at the site of interest to develop a regression line. Conversely, the user may extract high-flow values from the record at the site of interest to develop a regression line if the user is interested in high-flow statistics. The user should use MOVE.1 rather than MOVE.3 for this type of analysis because the MOVE.1 regression equation is not adjusted for the difference between the concurrent record and the mean and standard deviation of the complete long-term record. Furthermore, regression equations developed with a subset of data should not be used for record extension if the selected data and therefore the regression equation are not representative of the full range of streamflows.

Purpose and Scope

This manual describes the implementation, use, and interpretation of results from the SREF (Version 1.0) program. SREF is a program for extension or augmentation of available streamflow data at a site of interest by using long-term streamflow records at hydrologically similar sites. The SREF program implements the maintenance of variance extension type 1 (MOVE.1) method described by Hirsch (1982) and the maintenance of variance extension type 3 (MOVE.3) method described by Vogel and Stedinger (1985). The SREF program also facilitates record extension with a single or multisegment regression model developed by use of the nonparametric Kendall-Theil Robust Line (KTRLine) method (Helsel and Hirsch, 2002; Granato, 2006). The SREF program was developed by the USGS in cooperation with the Federal Highway Administration (FHWA) and the Rhode Island Water Resources Board (RIWRB) for use in the analysis of local, regional, and national hydrologic data sets. SREF uses streamflow files retrieved and processed from NWISWeb using the GNWISQ program (appendix 1). Methods and governing equations for estimating streamflow statistics in the SREF program are described. The formats of input data and output statistics are described. Step-by-step use of the program's graphical user interface is illustrated. The program code written in Microsoft Visual Basic 6.0 is documented as individual files in a Visual Basic project directory on the computer disk containing this manual.

Selecting Index Stations

The theory supporting record extension and augmentation is based on the idea that drainage basins upstream of the short- and long-term-record sites are hydrologically similar and that the record of streamflow from an index (or predictor) station may be used as a surrogate for streamflow information at the site of interest. For example, in comparing three record-extension techniques, Hirsch (1979) demonstrated that even simple techniques such as drainage-area-ratio techniques (which do not require streamflow data from the site of interest) could produce fairly accurate estimates (with bias within about 15 percent) if basins were in close proximity, of similar size, and of similar hydrogeology. Hirsch (1979) also indicated that regression-based methods would be more appropriate than the drainage-area-ratio technique if differences in basin characteristics substantially affect the flows per unit area.

The requirements for selecting one or more index stations to be used for extending or augmenting available streamflow data for a site of interest depend on the objectives of the analysis. Hydrogeologic similarity, relative basin size, and the proportion of consumptive water use between the index stations and the site of interest are important considerations for both record extension and augmentation because some statistical characteristics (such as higher moments of the index-streamflow population) are transferred by record extension and augmentation techniques (Timothy Cohn, U.S. Geological Survey, Office of Surface Water, written commun., 2006). In general, basins of index stations should be in close proximity to the site of interest, have similar physiographic and hydrogeologic characteristics, and have similar drainage areas. For both extension and augmentation, the analyst must consider regression statistics (from the line of organic correlation) and visually examine the relation between concurrent measurements to ensure that the relation is linear and that residuals are approximately normal and homoscedastic.

If the objective is to adjust short-term statistics at one site so they represent long-term statistics (record augmentation), then the selection of index stations may be more heavily weighted on regression statistics than proximity to the site of interest. This is because the best regression model will provide the best possible estimates of selected streamflow statistics. In this case, the proximity and relative sizes of drainage areas of the site of interest and the index stations are not the most critical considerations.

If the objective is to generate a representative sequence of daily, weekly, monthly, or annual streamflows (record extension), then it is important to consider basin characteristics and regression statistics. The tendency for high flows to follow high flows and low flows to follow low flows in a streamflow data set, commonly called serial correlation, is considered an important feature to preserve in streamflow-extension techniques (Hirsch, 1982). The proximity of the index stations to the site of interest is important for maintaining the serial correlation of streamflows in the extended daily mean streamflow

records. Records from nearby streamflow-gaging stations are more likely than records from more distant stations to reflect the effects of individual precipitation events (or lack thereof) in both basins in the extended record. Basin characteristics such as topography, geology, and basin area also are important considerations for providing an estimated time-series of streamflow data with the correct flow statistics and serial correlation of streamflows. Drainage-basin area is important to represent the watershed response to a given precipitation event. For example, a commonly used rule-of-thumb for hydrograph separation is that the number of days required for streamflow to return to base flow after a storm peak is the drainage area (in square miles) raised to the power 0.2 (Linsley and others, 1975; Sloto and Crouse, 1996). Therefore, one may expect stormflows to persist (after the peak) for about 1 day for a 1-mi^2 basin, 2 days for a 32-mi^2 basin, and 3 days for a 243-mi^2 basin. Similarly, the time to the hydrograph peak, commonly expressed as a function of stream length and distance to the upstream divide, is related to basin size (Chow and others, 1988). Basin size may be less critical for record augmentation than for extension because the regression-equation coefficients should account for differences in between the means and variances of streamflows of different basins, and the MOVE equation is not being used to estimate a streamflow for a particular day.

Hydrologists commonly want to use two or more index stations for record extension or augmentation because different streamflow sites rarely are similar in all physiographic, hydrogeologic, land-cover, land-use, and water-use features. Use of multiple stations is done in an attempt to improve individual estimates and compensate for perceived limitations in hydrologic similarity or the relative proximity of the index station with the best regression relation. For example, Ries and Friesz (2000) use MOVE.1 regression statistics to derive a weighted-average estimated streamflow value at the site of interest from individual MOVE.1 prediction values for different stations. Multiple index stations, with good regression statistics, commonly are used for record augmentation in an attempt to generate estimates that are representative of streamflow statistics in adjacent areas. Estimates with multiple stations are commonly weighted by the strength of regression statistics. Although the effects of random differences in spatial and temporal patterns in precipitation on the relative timing and magnitude of streamflow variation from day to day will tend to be balanced over long periods of time (for record augmentation), it is thought that use of multiple stations for estimating individual streamflow values may better reflect the occurrence of individual storm events (for record extension). The magnitude of local storm events, however, may be reduced by including streamflow estimates from neighboring index stations that did not receive precipitation during an individual event.

Governing Equations and Statistical Methods

The SREF program is designed to help the user develop regression relations and implement them for streamflow-record extension or augmentation. This program facilitates the development of log-log regression relations for streamflow data with the MOVE.1 and MOVE.3 implementations of the line of organic correlation. The program facilitates this process by finding paired values of streamflow from one or more index stations and the site of interest and calculating regression relations and related statistics. The program automatically implements either MOVE.1 or MOVE.3 for streamflow-record extension or augmentation from a one-segment regression line. The program also provides an interface to enter regression coefficients for a one-segment or multisegment regression line developed outside the program. For example, the program provides an output file to facilitate development of regression relations with the Kendall-Theil Robust Line program (Granato, 2006). If the user selects record extension, the program outputs the estimated and measured values for the site of interest in the USGS NWISWeb daily value relational database (RDB) format. If the user selects record augmentation, the program generates estimates of selected statistics for the site of interest from one or more input files with the desired streamflow statistics for the index stations. The program outputs paired measurements and regression statistics in a tab-delimited ASCII text file.

Implementation of the Line of Organic Correlation

The MOVE.1 and MOVE.3 methods are implementations of the regression method known as the line of organic correlation (Hirsch, 1982; Hirsch and Gilroy, 1984; Vogel and Stedinger, 1985). The underlying assumption for using regression to extend or augment the streamflow record for the site of interest using record for the long-term streamflow-gaging station is that the population of streamflows at the two sites have similar properties in terms of the shapes of the distributions of flows, the serial correlations of flows, and seasonal variations (Hirsch, 1982; Vogel and Stedinger, 1985). It is also assumed that the proportions and magnitudes of any anthropogenic effects on streamflows at the stations covary. Although these record-extension techniques adjust the statistics for location and scale (for example, the mean and variance, respectively) of the long-term record to the short-term record, variations in the daily record for the long-term site are projected onto the estimated daily record for the short-term site. Although most of the concurrent record may indicate a high degree of

correlation, individual values may be affected by local temporal variations in precipitation, infiltration, and runoff. The quality of individual streamflow estimates will be a function of the fit of the regression model, the proximity of the sites, and the hydrological similarity of the basins upstream of each site. This is because day-to-day variations are more critical for record extension than for record augmentation.

The fundamental regression equation is

$$Y_i = mX_i + b + e_i, \qquad (1)$$

where

i	is the index value for a given measurement,
Y_i	is streamflow value for the ith measurement at the short-term-record site,
X_i	is the streamflow value for the ith measurement at the long-term-record site,
m	is the regression-line slope,
b	is the regression-line intercept, and
e_i	is the difference between the regression-line estimate and the measured Y value for the ith measurement (Helsel and Hirsch, 2002).

The slope (m) of the line of organic correlation (MOVE.1) is calculated as the product of the sign of the correlation coefficient (r) and the ratio of the standard deviation of the Y values (S_y) to the standard deviation of the concurrent X values (S_x), and is expressed mathematically as

$$m = sign[r]\left(\frac{S_y}{S_x}\right). \qquad (2)$$

The program calculates several statistics (Helsel and Hirsch, 2002) to calculate the MOVE.1 and MOVE.3 estimates, including the sum-of-squares for Y as:

$$SS_y = \sum_{i=1}^{N}(Y_i - \overline{Y})^2, \qquad (3)$$

where

SS_y	is the sum-of-squares Y,
$\overline{Y}$	is the average Y value, and
N	is the number of paired-data values.

The variance of Y values is

$$S_y^2 = \frac{SS_y}{(N-1)}. \qquad (4)$$

The standard deviation of Y values (S_y) is the square root of the variance. Similarly, the sum-of-squares of X is

$$SS_x = \sum_{i=1}^{N}(X_i - \overline{X})^2, \qquad (5)$$

where

$\overline{X}$ is the average X value.

The variance of X values is

$$S_x^2 = \frac{SS_x}{(N-1)}. \qquad (6)$$

The standard deviation of X values (S_x) is the square root of the variance. The sum of the cross products (S_{XY}) is

$$S_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}). \qquad (7)$$

The correlation coefficient (r) may be expressed as

$$r = \frac{S_{XY}}{\sqrt{SS_X S_{XY}}} \qquad (8)$$

as in (Helsel and Hirsch, 2002), or by use of, an intermediate beta term (β)

$$\beta = \frac{S_{XY}}{S_X} \qquad (9)$$

to calculate r as

$$r = \beta \frac{S_X}{S_Y} \qquad (10)$$

as in Vogel and Stedinger (1985).

Unlike ordinary least squares regression, the MOVE.1 preserves the variance of the predicted Y population because the sign of the correlation coefficient is used rather than the magnitude of the correlation coefficient (Hirsch, 1982; Hirsch and Gilroy, 1984; Hirsch and others, 1993; Helsel and Hirsch, 2002). Matalas and Jacobs (1964) proposed use of the ordinary least squares estimates of slope and intercept with a stochastic realization of the random-error term to recover the lost variance, but this was deemed unsatisfactory for record extension because results would depend on individual stochastic realizations of the extended record, and the random error would artificially reduce the serial correlation in subsequent streamflow values (Hirsch, 1982; Vogel and Stedinger, 1985; Hirsch and others, 1993).

The intercept (b) of the line of organic correlation (MOVE.1) is calculated so that the line with slope m passes through a point with XY coordinates that equal the mean of the X and Y populations, respectively. This is expressed mathematically as

$$b = \overline{Y} - m\overline{X}. \qquad (11)$$

Equation (1) shows that the residual error or uncertainty in the predicted Y value for each XY pair is expressed mathematically as

$$e_i = Y_i - (mX_i + b) . \tag{12}$$

(Hirsch, 1982; Hirsch and Gilroy, 1984; Vogel and Stedinger, 1985).

Vogel and Stedinger (1985) noted that MOVE.1 estimates produce populations of Y values based on statistics of the X and Y samples during the concurrent period of record, and this may not fully utilize information available in the total population of X values from the long-term record station. Therefore, Vogel and Stedinger (1985) developed the MOVE.3 method to adjust the coefficients of the regression equation to reflect differences in statistics between the concurrent record and the full long-term record at the index station. To do this, Vogel and Stedinger (1985) used estimates of the mean and standard deviation from the concurrent period (N_1) from both stations, and estimates of the mean and standard deviation from the nonconcurrent period (N_2) of the index station, to estimate statistics for the entire period (N_1+N_2). The estimate of the long-term average value ($\overline{Y_{lt}}$) for the short-term-record site is

$$\overline{Y_{lt}} = \overline{Y_1} + \frac{N_2}{(N_1 + N_2)} \beta(\overline{X_2} - \overline{X_1}) . \tag{13}$$

The estimate of the long-term variance at the short-term-record site is

$$S_{ylt}^2 = \frac{1}{(N_1 + N_2 - 1)} \times$$
$$\left[(N_1 - 1)S_{y1}^2 + (N_2 - 1)\beta^2 S_{x2}^2 + \frac{N_1 N_2}{(N_1 + N_2)} \beta^2 (\overline{X_2} - \overline{X_1})^2 + (N_2 - 1)\alpha^2 (1 - r^2) S_{y1}^2 \right], \tag{14}$$

in which the alpha coefficient (α) is

$$\alpha = \left[\frac{N_2(N_1 - 4)(N_1 - 1)}{(N_2 - 1)(N_1 - 3)(N_1 - 2)} \right]^{\frac{1}{2}} . \tag{15}$$

The MOVE.3 equation is expressed as

$$Y_i = a' + m'(X_i - \overline{X_2}) , \tag{16}$$

in which the modified intercept is

$$a' = \frac{(N_1 + N_2)\overline{Y_{lt}} - N_1\overline{Y_1}}{N_2} , \tag{17}$$

and the modified slope is

$$m' = \left[\frac{\left[(N_1 + N_2 - 1)S_{ylt}^2 - (N_1 - 1)S_{y1}^2 - N_1(\overline{Y_1} - \overline{Y_{lt}})^2 - N_2(a' - \overline{Y_{lt}})^2 \right]}{\left[(N_2 - 1)S_{x2}^2 \right]} \right]. \tag{18}$$

Rearranging the terms in equation 16 yields

$$Y_i = (a' - m'\overline{X_2}) + m'X_i , \tag{19}$$

which, when values for a', m' and the mean of the noncuncurrent values of X are used, is in the standard format of a slope and intercept for a linear equation.

Regression Statistics

The SREF program calculates several regression statistics for both the MOVE.1 and MOVE.3 regression lines to help the user evaluate both the data used in the regression and the regression results. These statistics include the residual for each point, the leverage of each point in the X and Y direction, the mean square error, the root mean square error, and a modified Nash-Sutcliff coefficient of efficiency. The residual for each point (e_i) is defined by equation (12) on the basis of the concurrent data and either the MOVE.1 or MOVE.3 estimates of the slope or intercept as appropriate. The mean square error (MSE) is calculated as

$$MSE = \frac{\sum_{i=1}^{N} e_i^2}{(N-2)} , \tag{20}$$

and the root mean square error (RMSE) is the square root of the mean square error (Helsel and Hirsch, 2002). The RMSE indicates the magnitude of the population of residual errors. The leverage in the X direction (h_i) is calculated as

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SS_x} . \tag{21}$$

The leveraged y-residual $e_{(i)}$ is calculated as

$$e_{(i)} = \frac{e_i}{(1 - h_i)} . \tag{22}$$

The SREF program also calculates a modified Nash-Sutcliff coefficient of efficiency (Legates and McCabe, 1999). The coefficient of efficiency for a regression model ranges from one, indicating that the predicted values Y_P perfectly characterize the observed data Y_i, to minus infinity, indicating that the regression model has no predictive ability. A coefficient of efficiency of zero indicates that the average value of the observed data is as efficient for predicting observed values as the regression model. Legates and McCabe (1999) indicate that the squared terms used in the Nash-Sutcliff coefficient of efficiency tend to put too much weight on outliers. Legates and McCabe (1999) also indicate that a similar statistic, based on absolute values rather than squared terms, provides a similar measure of efficiency without the undue influence of outliers. This efficiency statistic is calculated as

$$E = 1.0 - \frac{\sum_{i=1}^{n} |Y_i - Y_p|}{\sum_{i=1}^{n} |Y_i - \overline{Y}|} , \tag{23}$$

where

E	is the efficiency ratio,
i	is the index value for a given measurement,
n	is the number of measurements,
Y_i	is the streamflow value for the ith measurement at the short-term-record site,
Y_p	is the predicted streamflow value for the ith measurement at the short-term-record site, and
$\overline{Y}$	is the average of the streamflow values at the short-term-record site.

The efficiency ratio may be examined with the residual plots, correlation coefficients, and other regression statistics to compare the predictive capability of regression equations developed by using data from different index stations.

Treatment of Zero Flows in the Predictor Record Set

The SREF program was designed for the possibility of zero flows in the input data set. Zero-flow values may occur at the index site, the site of interest, or both. The existence of zero flows in either input data set may affect regression-coefficient estimates and regression statistics. The SREF program, however, uses the logarithms of concurrent nonzero streamflow values to develop a regression model. Streamflows that are reported as zero may represent actual dry stream conditions or may be a small flow that is below the limit of detection for stream-stage or discharge measurements. For example, water may seep under or through the streamflow control that is used as the point of zero flow.

Statistical treatment of censored data (unknown values below a measurement-detection level) has received extensive study for water-quality values (Helsel and Hirsch, 2002; Helsel, 2004). Different methods for estimating sample statistics such as the means and standard deviations of populations with varying proportions of censored values have been developed and reported in the literature (Helsel, 2004). Although such methods commonly use estimates of individual censored values to estimate sample statistics, use of these estimates as substitutes for individual values is not recommended (Helsel, 2004). Future enhancements of the SREF program may include the use of statistical methods for estimating sample statistics for censored data sets. However, statistical methods such as regression-on-order statistics (ROS), survival analysis, and other methods described by Helsel (2004) are not implemented within SREF for record extension. This is because these statistical methods are not designed to predict individual values, and streamflow data values are serially correlated rather than being independent values.

If the number of zero flows in the input data sets is substantial, the user may estimate the slope and the intercept

of the regression line by use of the Kendall-Theil Robust line program (Granato, 2006). Because the Kendall-Theil Robust line method is based on the median of pairwise slopes and the medians of the X and Y data sets, regression coefficients are not sensitive to assumptions made about censored values as long as the biased values represent less than about 29 percent of the data (Wilcox, 2001; Helsel and Hirsch, 2002; Granato, 2006). The user also may calculate MOVE.1 or MOVE.3 estimates by first calculating the means and standard deviations of the X and Y values using censored-data methods (Helsel, 2004) and then enter the resulting estimates of the slope and intercept for this adjusted line of organic correlation in the Kendall-Theil Robust Line equation interface of the SREF program for streamflow-data extension and augmentation.

The SREF program has four options for handling zero flows in each record-generation interface because it is necessary to address the existence of zero flows in the index station data set to generate a streamflow record at the site of interest. The four choices are (1) to use the zero-flow values for the index site, (2) substitute a value of 0.1 ft^3/s, (3) substitute a user-defined streamflow value, or (4) substitute a streamflow-recession constant (ranging from greater than 0.0 to 1) to be applied from the last nonzero streamflow value. Selection of an index value of zero will result in a predicted value of zero for a regression equation formulated from the logarithms of data. Selection of an index value of 0.1 ft^3/s or a user-defined substitute value will provide a constant estimate based on the regression equations for all zero flows at the index station. Selection of a streamflow-recession constant for the index station will create a series of flow values that decrease with each successive day of zero flows. The streamflow-recession constant sets the rate of decrease in the streamflow values. For example, if the recession constant is 0.5, the index stream-flow value on the first day of zero flow will be one-half of the previous nonzero value. The index streamflow on the next day will be one-quarter of the last nonzero streamflow value. Index streamflow values will approach zero asymptotically at the user-selected rate. Each time a streamflow value of zero appears in the index-station record, estimated streamflow values in the output file are denoted by a comment code "e0" to warn the user that these estimated values are based on the user-selected substitutions for censored index-station flow values.

Bias Correction

The SREF program calculates regression statistics in common (base 10) logarithmic space because logarithmic transformation commonly helps linearize relations between streamflows collected at different stations and improves the structure of residuals (Ries and Friesz, 2000). For example, Hirsch (1979) indicated that streamflow-record extension by log-log regression may produce estimates with less bias and smaller root mean square errors than regression estimates in

linear space. In theory, each individual streamflow estimate from the regression-line equation is the most probable value of the logarithm of streamflow at the site of interest, and the retransformed value of this estimate is also the most probable value of streamflow (Timothy Cohn, U.S. Geological Survey, Office of Surface Water, written commun., 2006). Because the MOVE.1 and MOVE.3 equations are developed with the logarithms of streamflow, it is the means and variances of the logarithms of streamflow that are preserved. Bias occurs in the regression of transformed variables because the intercept of the regression line is fit through the geometric mean values of the X and Y variables, and the geometric mean (the retransformed average of the transformed data) does not equal the arithmetic average of the untransformed data (Koch and Smillie, 1986; Glysson, 1987; Gilroy and others, 1990; Crawford, 1991; Cohn and others, 1992; Hirsch and others, 1993; Helsel and Hirsch, 2002). Therefore, the streamflow estimate is the geometric rather than the arithmetic mean of streamflow data at the site of interest for a given value at the index site, and normal distribution of the residuals in logarithmic space is skewed when retransformed.

The arithmetic average has special meaning in hydrologic studies designed to determine total streamflows or constituent loads because the arithmetic averages of the streamflow or constituent load are, in theory, the total masses of water and solutes, respectively, that flow past the measurement point divided by the total time of the measurement interval. Even a small bias in individual predictions may produce a substantial error in the calculated total. Ignoring bias correction can result in the underprediction of total flows, but the use of bias correction can result in the overprediction of total flows especially if high outliers are included in the measurements (Koch and Smillie, 1986). Several bias-correction methods have been proposed in the literature (Duan, 1983; Koch and Smillie, 1986; Cohn and others, 1992; Helsel and Hirsch, 2002). The SREF program calculates a bias correction factor (BCF) as one of the regression statistics and allows the user to implement this BCF. Use of the BCF, however, is not the default option because of the limited applicability for record extension with MOVE.1 or MOVE.3 (Timothy Cohn, U.S. Geological Survey, Office of Surface Water, written commun., 2006).

The nonparametric smearing estimator (the average of the retransformed log-regression residuals) proposed by Duan (1983) was selected for implementation as one of the regression statistics provided with the SREF program because it performs reasonably well and is not sensitive to statistical assumptions regarding residual-population characteristics (Gilroy and others, 1990; Crawford, 1991; Hirsch and others, 1993; Helsel and Hirsch, 2002). The Duan (1983) BCF is called the smearing estimator because the method applies or "smears" the average retransformed error over all measurements. This approach was developed for log-normal transformations with the natural log. With the appropriate retransformation method, however, the smearing estimator should be applicable for other transformations. A generalized expression

of the smearing estimator, applicable for any log-based transformation, is

$$BCF = \frac{\sum_{i=1}^{n} G(e_i)}{N},$$ (24)

where

BCF	is the bias correction factor,
G	is the retransformation function;
e_i	is the residual error or uncertainty in the predicted Y value for each data point (i), and
N	is the number of XY data pairs.

In this case, G is

$$G(e_i) = 10^{e_i}.$$ (25)

Use of Multiple Index Stations

Multiple index stations for record extension or augmentation are commonly used in an attempt to provide improved estimates that are representative of streamflow characteristics in adjacent areas. In theory, the best predictive equation will produce the lowest residual errors and this equation will indicate the index station that is most representative of the site of interest. In practice, the selection of the equation with the lowest residual errors may depend as much on the data that is available as the representativeness of the index basin. For example, similarity in available low-flow data may be related to similarity in water withdrawals or in aquifer materials and extent. In contrast, similarity in high-flow data may be related to similarity in basin slope and imperviousness. Although there may be such artifacts in the individual record-extension predictions, the correlation coefficient and the residual-regression statistics provide a quantitative measure of the predictive value of data from each index station.

The SREF program calculates a weighted average of predicted streamflow values from two or more representative index basins by use of the mean square error. This weighting factor is calculated as

$$Weight_j = \frac{\left(\sum_{j=1}^{NS} MSE_j\right) - MSE_j}{\sum_{j=1}^{NS} MSE_j}$$ (26)

where

$Weight_j$	is the weighting factor for station j,
NS	is the number of index stations used, and
MSE_j	is the mean square error for the predictive

equation using station j as defined in equation 20.

Equation 26 effectively assigns the highest weights to the stations with the lowest mean square errors and ensures that the sum of the weighting factors for all index stations is one.

Use of the Kendall-Theil Robust Line Program

The SREF program provides an interface that allows the user to enter a one-segment or multisegment regession model generated by the Kendall-Theil Robust Line Program (KTRLine; Granato, 2006). The SREF program interface for KTRLine regression equations also may be used for regression equations developed with other methods. Wilcox (2001) demonstrates that parametric-regression techniques are not robust to outliers because of the effect of such values on the mean and variance statistics that form the basis of these techniques. The KTRLine interface is provided to enable data extension and augmentation with nonparametric-regression methods. Wilcox (2001) indicated that even one outlier may grossly distort parametric-regression relations. He also indicates that non-normality and heteroscedasticity can increase the standard error by one or more orders of magnitude. In contrast, Wilcox (2001) indicates that more than 29 percent of data points must be outliers for the Kendall-Theil regression method to provide a distorted estimate of the relation for the bulk of the points. Streamflow measurements from partial-record data sets may contain outliers because the hydrographer does not have an established stage-discharge rating to evaluate each measurement. Streamflow measurements from continuous-record data sets may contain outliers for a number of reasons, including equipment error, intermittent accumulation of debris or ice in the control section and other factors affecting the precision of measurements at extremely high and extremely low streamflows (Rantz and others, 1982).

The KTRLine program is a Visual Basic program that may be used with the Microsoft Windows operating system to calculate parameters for robust, nonparametric estimates of linear-regression coefficients between two continuous variables. The KTRLine program was developed because this robust nonparametric method is resistant to the effects of outliers and non-normality in residuals that commonly characterize hydrologic data sets. The slope of the line is calculated as the median of all possible pairwise slopes between points. The intercept is calculated so that the line will run through the median of input data. A single-line model or a multisegment model may be specified. The number of segments that may be specified is one if there are fewer than 20 points, two if there are 20–29 points, and increases by one up to a maximum of five with each additional decade of available data. The number of segments is limited to provide stability in slopes for small data sets and resultant convergence points in multisegment models. The user, however, may specify one or more segments

that have fewer than 10 data points as long as the total number of points in the data set exceeds 20. The KTRLine program is a graphical tool that facilitates development of regression models by generating graphs of the regression line with data, the regression residuals (with X or Y), percentile plots of the cumulative frequency of the X variable, the cumulative frequency of the Y variable, and the cumulative frequency of the regression residuals. The user may individually transform the independent and dependent variables to reduce heteroscedasticity and to linearize data. The KTRLine program also prints model specifications and regression statistics to the screen.

The user may save and print the regression results. The KTRLine program can accept data sets that contain up to about 15,000 XY data points. A daily mean streamflow data set with 15,000 values would represent a concurrent data set that is about 41 years long. The program may be perceptibly slow with data sets that contain more than about 1,000 points because the program must sort the array of all pairwise slopes and a data set of this size results in the calculation and sorting of about 500,000 pairwise slopes.

The SREF program provides leverage statistics for each data pair in the output file so the user may identify potential outliers. Use of the KTRLine program (Granato, 2006), however, may indicate if these data are substantially changing the regression relation between streamflow at the index station(s) and the site of interest. Outliers may have a substantial influence on the regression coefficients from the line of organic correlation if these coefficients substantially depart from the regression coefficients from the Kendall-Theil robust line. Helsel and Hirsch (2002, p. 11) indicated that "Outliers may be the most important points in the data set, and should be investigated further." Such high-leverage outliers, usually at high or low values of the predictor variable, indicate that different physical processes may affect relations in different flow regimes.

Wilcox (2001) also indicates that parametric methods may provide unrealistic estimates if a data set includes values that characterize two or more processes. Use of multisegment regression models may be justified if differences in hydrologic characteristics between the index station and the site of interest differ in different discharge ranges. Differences in water withdrawals or in aquifer materials and extent may affect the base-flow recession. For example, streamflow depletions caused by withdrawals that are one percent of the average flow may be a much higher percentage of a low-flow statistic. In contrast, different relations between stations in high-flow data may be caused by factors such as differences in basin slope, wetland storage, and connected impervious areas. Use of a multisegment regression model may provide better estimates for the site of interest if such factors affect correlations in different flow ranges.

Ries and Friesz (2000) demonstrated the use of graphical methods (Searcy, 1959; Riggs, 1972) for estimating streamflows at a site of interest if relations between the index station and the site of interest were not well characterized by a one-segment regression model. Such a graphical fit can be subjective if done manually, but methods such as locally weighted scatterplot smoothing (LOWESS) can be used to indicate a relation (Helsel and Hirsch, 2002). Neither the graphical method nor the LOWESS method, however, provides an equation to facilitate record extension for thousands of data points. Furthermore, these methods do not provide regression statistics to aid the user in selecting the best index station. The KTRLine program (Granato, 2006) does provide equations for multisegment relations for data sets with more than 19 points and regression statistics for comparing the predictive values of alternate regression relations based on data from different index stations.

Use of the SREF Program

SREF is a Visual Basic program that provides regression statistics for the MOVE.1 and MOVE.3 regression methods and produces an estimated streamflow record for a site with limited data from long-term continuous streamflow records from one or more streamflow-gaging stations. The user interface for the program consists of an interactive specification form, a graphical-output selection form, a graphical-output form, and three forms for generating long-term streamflow record and estimates of selected streamflow statistics for the site of interest. The SREF program, if installed properly, should be compatible with commonly used Microsoft Windows operating systems. A screen resolution of at least 1024 by 768 pixels is needed to display the entire graphical-output form. Streamflow data from selected stations in Massachusetts and Rhode Island are available on the computer disk containing this manual. These data are used in the examples in the following discussion on the installation and use of the program.

Installation and Removal

The SREF program depends on a number of software drivers and dynamic-link libraries that may not be installed and available on the user's computer. This program must be installed by someone who has administrative rights on the program user's computer. The administrator must test the installation with the user's profile to ensure that all permissions are set properly. A readme.txt file with installation instructions and three installation files are in the folder SREFInstall on the CD-ROM containing this manual. Three files—setup.exe, setup.LST, and SREFv1.CAB—are needed to install the program. The file setup.exe is the installation program. The setup.LST file is a text file that provides the necessary installation specifications. The file SREFv1.CAB is the file containing the program and support files. The setup program is a standard Microsoft installation wizard that should be familiar to the user or system administrator. These three files must be located in the same directory on the CD-ROM or in a directory on the user's computer. The person installing the program should follow all the standard choices for installation. The installation

program creates the directory C:\Program Files\FHWA\HEP\ SREF\ and includes the SREF program in the computer's registry. If desired, a shortcut to the program can be added to the desktop manually after installation. Sample files are saved on the CD-ROM containing this manual and should be copied to a directory in which the user has read, write, and execute rights. The user may uninstall the SREF program and its support files by use of the standard Microsoft Windows Add or Remove Programs wizard on the control panel of the user's computer.

Input-File Formats

The SREF program uses several input-file formats to retrieve the information and data necessary to develop regression relations between the data from one or more long-term streamflow-gaging stations and data from the site of interest (table 1). Input-data files include an initialization file, a station-list file, an input file with data from the site of interest, one or more long-term-record data files, and one or more long-term-record statistics files. The program uses an initialization file SREF.ini to provide some flexibility in the format of the other input files. The program uses the station-list file to indicate the working directory, which is the location of the station-list file and the input-file names, which are based on the station numbers. The program reads streamflow data from the site of interest in five possible file formats. Long-term-record data files are used for input of daily mean streamflow values to develop the regression relation and for generation of an extended daily mean streamflow record for the site of interest (record extension). A flexible-format input-statistics file is used for generation of estimate the selected statistical estimates for the site of interest (record augmentation). Examples of each type of file are provided in the example-data directory for the SREF program on the CD-ROM containing this manual.

Stream-discharge data for the site of interest may come from various sources, so there are five possible input-file formats for this data. The first format is a two- or three-column tab-delimited format that can be used for partial-record or short-term data. The second format is the NWISWeb RDB daily value format intended for input of short-term continuous-record data. The third format is the NWISWeb RDB instantaneous-discharge-measurement format intended for input of partial-record data. The fourth and fifth formats are designed to read FORTRAN-style input files. The fourth input format is configured to read individual discharge measurements obtained from the USGS Automated Data Processing System (ADAPS) (U.S. Geological Survey, 2003). The first option (FORTRAN01) is the ADAPS short (9-207) format. The second option (FORTRAN02) is the ADAPS full format.

The format for each partial-record file may include more than one instantaneous discharge measurement on any given date. Multiple measurements made at the partial-record site on a given day are paired with repeated values of the daily mean streamflow for that day from each long-term continuous-record station for the purpose of regression-model development. The resulting regression-model equation may be used for record extension and augmentation. Inclusion of multiple measurements for a given day will, however, preclude use of the option to use measured values from the partial-record site for generating the record of estimated daily mean streamflows at the site of interest. Depending on the objective of the analysis and the nature of the available partial-record data, the user may wish to average the available measurements for a given day before proceeding with the analysis.

Initialization File Format

The initialization file (SREF.ini) is a tab-delimited ASCII text file that consists of information about the format of each input file. The initialization file (fig. 3-1) must include sixteen lines of information. The first two lines, which start with a pound symbol (#), indicate the file type and purpose. The next four lines, which start with the alphanumeric designations 1A through 1D, specify the columns of a tab-delimited NWISWeb daily value input-file format with column specifications for the station number, the date, the daily mean streamflow, and associated column codes. The next four lines, which start with the alphanumeric designations 2A through 2D, specify the columns of a tab-delimited NWISWeb stream-measurement format with specifications for the station number, the measurement number, the date, and the measured discharge (flow) value. The next eight lines with the remaining two input-file-specification definitions start with the alphanumeric

```
# SREF Initialzation file
# Data Column Specification
1A. NWIS Web Station No.    2
1B. NWIS Web Date           3
1C. NWIS Web Mean Daily Flow        4
1D. NWIS Web Mean Daily Flow Code           5
2A. NWIS Web Measurement Station No.        2
2B. NWIS Web Measurement No.        3
2C. NWIS Web Measurement Date       4
2D. NWIS Web Measurement Flow       11
3A. ADAPS 9207 Measurements No.    1        6
3B. ADAPS 9207 Measurements Date   9        10
3C. ADAPS 9207 Measurements Flow   62       8
4A. ADAPS Full Measurements No.    1        6
4B. ADAPS Full Measurements Date   9        10
4C. ADAPS Full Measurements Flow   70       9
```

Figure 3-1. Initialization file for the Streamflow Record Extension Facilitator.

Table 3-1. Summary of input and output files for the Streamflow Record Extension Facilitator (SREF Version 1.0) program.

[GNWISQ, Get NWIS streamflow (Q) software version 1; ADAPS, Automated Data Processing System; MOVE, maintenance of variance; NWIS, National Water Information System; RDB, relational database]

File type	Example files on the CD-ROM directory: CDROM:\Applications\SREFEx\	Comments
Input files		
Initialization input file	SREF.ini	Provides flexibility in the input files
Station-list input file	IndexStations.txt	Lists long-term index-station numbers
Column-Format Input File	01118000PR.txt	Tab-delimited input file created by user
NWISWeb daily value RDB streamflow data file	01118000Q.txt	The 2005 daily value format NWISWeb RDB file, formatted by GNWISQ
RDB Measurement Input File	01118000NWM.txt	NWIS Web RDB surface-water measurement file
FORTRAN01 Input File	01117465FORT01.txt	ADAPS short (9-207) discharge-measurement format
FORTRAN02 Input File	01117465FORT02.txt	ADAPS long discharge-measurement format
Streamflow-Statistics Input File	01117500STATS.txt	Tab-delimited input file with long-term index-station statistics
Output files		
Paired-Record Output File	01118000DAT0.txt	Tab-delimited ouput file with concurrent measurements
Period-of-Record Date Output File	01118000-PeriodOfRecordDates.txt	Tab-delimited ouput file indicating available data for the period of record
Maintenance-of-Variance Regression-Results Output File	01118000MOV0.txt	MOVE.1 and MOVE.3 regression equations and regression statistics
Kendall-Theil Robust-Line Input File	01118000-01117500KT.txt	Tab-delimited output with the paired streamflow measurements and the date
Extended-Record Output File	01118000by01117500MV1Q.txt	Estimated daily mean values in NWISWeb RDB daily value format
Streamflow-Statistics Output File	01118000by01117500STATSMV1Q.txt	Estimated streamflow statistics for the site of interest
Error output file	SREFErrFile0.txt	File containing errors or warnings generated during the analysis process

designations 3A through 3C and 4A through 4C. These lines include the information necessary to define input formats that may be expected from a FORTRAN program. In FORTRAN formats, each variable is assigned a set number of spaces on a line and a preset length. Lines 3A through 3C and 4A through 4C each have a line definition, a tab, the position of the variable, another tab, and the length of the input variable. In each case the SREF initialization file instructs the program to read a six-character measurement number, a ten-character measurement date, and a discharge value. These initialization variables must be specified in the proper order, but the position and length may be used to read files with different FORTRAN formats as long as each measurement number, date, and discharge value are on the same input line.

Station-List Input-File Format

The program reads a station-list file to identify potential long-term streamflow-gaging stations for use as index stations for record extension or augmentation. This list file does not contain any header lines. The list file consists of 8- to 12-character streamflow-gaging-station numbers on separate lines without leading or following spaces. The station-list file should not include any text except the station numbers of interest and should not end in a blank line. The station numbers in this file are used with a "Q.txt" suffix to identify NWISWeb data files for each long-term index station.

Column-Format Input File

The column-format data file is designed to provide a 2- or 3-column tab-delimited partial-record input-data-file option. This type of input file can be easily generated from many computer applications, including plain-text editing programs, databases, spreadsheets, and by text-editing the USGS ADAPS RDB measurement format (U.S. Geological Survey, 2003). The column-format data file consists of tab-delimited stream-discharge data that are concurrent with the period of record of the long-term record station. The first line of the file must include an 8- to 12-digit station-identification number, a tab, and a descriptor (such as text describing the station name and location). Subsequent lines include the date, a tab, and stream-flow measurements from the site of interest. An additional tab-delimited column of metadata may be included in the column-format input file for informational purposes, but the SREF program will not process this information. Visual Basic will recognize a number of date formats, but use of a standard year-month-day format (for example, 1990-12-19) will ensure that data are read properly. The column-format input file should not end in a blank line.

NWISWeb Daily Value RDB Input-File Format

The NWISWeb daily value RDB input file is used to read a short-term-record data file for the site of interest and a data file for each index stations. The NWISWeb daily value RDB format (Manis and others, 1988) input file consists of tab-delimited daily mean streamflow data. The daily discharge files consist of comment lines denoted by a pound symbol (#) in the first space, a tab-delimited column-name line, a tab-delimited RDB column-format line, and multiple lines of tab-delimited data. The SREF program is written to read a line with the fields

agency_cd[TAB]site_no[TAB]dv_dt[TAB]dv_va[TAB]dv_cd

in which [TAB] is the invisible tab character (ASCII character 9) and the data fields are the agency code, the site number, the date, the daily mean streamflow value, and the associated qualification code, respectively. SREF will also read the newer format with the fields

agency_cd[TAB]site_no[TAB]datetime[TAB]02_00060_00003[TAB]02_00060_00003_cd

in which [TAB] is the invisible tab character and the data fields are the agency code, the site number, the date, the daily mean streamflow value, and the associated qualification code, respectively. The column-name line is followed by a column-format line that indicates the maximum character length for each variable and the type of variable (d for date, n for number, and s for string). All remaining lines in the RDB file contain the data. NWISWeb daily value RDB input files should not end in a blank line. In the current (2006) format, values are denoted as approved (A) or provisional (P) in the qualification-code column. Approved values have been examined, and, if necessary, adjusted for variations in equipment performance or stream conditions; estimated values are denoted as (Ae) (Rantz and others, 1982).

SREF is designed to extend or augment available stream-flow records. Records from short-term stations may include gaps, but SREF can be used to estimate the record during such gaps. Index-station record files also may have gaps in record. If there are gaps in the Index-station record, regression lines will be estimated by using concurrent measurements from the short-term stations and index stations. SREF cannot be used to estimate record at the site of interest during gaps in the index-station record. If multiple stations are selected, however, SREF will properly weight available index values on each day to estimate a complete record for the short-term site of interest. Data from index stations may exist for such gaps, however, because some data were missed in the conversion from NWIS to NWISWeb. The user should examine available records and, if necessary, contact the appropriate state water-science center to inquire about any gaps in long-term streamflow-gaging station records. If there are gaps in the index-station record the SREF program will output information about the gaps in the long-term record for each station in the error file to facilitate analysis of the data, and the program will launch a message box to inform the user that there are such gaps in the record.

SREF has some flexibility for reading NWISWeb formats, but for the most consistent results it may be best to obtain the older daily value format files. The initialization

file may be used to specify the proper input columns, but the SREF program will verify that the proper column is specified by looking for the proper field names. Unfortunately, the new (2006) format files for daily mean streamflow records may include daily minimum and/or daily maximum flows for different stations if the records were coded for such values at any point in time. Also the NWISWeb group reserves the right to make unannounced changes in output formats (which may include mandatory inclusion of extra columns not requested in the query to NWISWeb) and descriptors at any time (James Slack, U.S. Geological Survey Office of Water Information, written commun. 2006). For these records, short-term-record data files and long-term index-station files should be obtained and reformatted to the older, more consistent, daily value (dv) format by use of the GNWISQ program (appendix 1). The GNWISQ program also has the ability to read a program specific initialization file GNWISQ.ini, which can be configured to adapt to future changes in NWISWeb.

RDB Surface-Water-Measurement Input-File Format

The RDB surface-water-measurement input-file format specification matches the current (2007) NWISWeb surface-water-measurement format. Currently, measurements are available for continuous-record stations but not for existing partial-record measurement sites. However, the ability to read a RDB-measurement input-file format is provided because inclusion of partial-record measurements on NWISWeb has been designated as an objective for further NWISWeb development. As an RDB format, the NWISWeb surface-water-measurement input files start with one or more comment lines with a pound symbol as the first character of each line, one column name and one column-format line, and one or more lines of data. The SREF program reads the station number, measurement number, date, and measured instantaneous stream-discharge value. The initialization file may be used to specify the proper input columns. The default (current format) columns are column 2 (station number), column 3 (measurement number), column 4 (date), and column 11 (measured instantaneous stream-discharge value).

FORTRAN01 and FORTRAN02 Input-File Formats

Both the FORTRAN01 and FORTRAN02 measurement input-file format specifications match current (2003) ADAPS-format output. The FORTRAN01 format matches the ADAPS short (9-207) format. The FORTRAN02 format matches the ADAPS full format. Each of these formats is completely defined in the ADAPS Manual (U.S. Geological Survey, 2003). These ADAPS formats are organized, but complex. Each measurement and associated columns may be displayed on more than one line within the file, there are numerous blank lines, and the block of header text may be repeated multiple times within each file. These files are labeled as FORTRAN formats because each variable is defined with a fixed starting position and a fixed length within a line of data (rather than having a flexible location and length delimited by a special character such as a tab or comma). The SREF program examines each line of these FORTRAN-format files to find lines that have a valid measurement number, a valid date, and a valid streamflow measurement in one line. If the user selects the FORTRAN01 option, the SREF program looks for a valid measurement number that starts with the first character and is 6 characters long, a valid date value that starts with the 9th character and is 10 characters long, and a valid stream-discharge value that starts with the 62nd character and is 8 characters long. If the user selects the FORTRAN02 option, the SREF program looks for a valid measurement number that starts with the first character and is 6 characters long, a valid date value that starts with the 9th character and is 10 characters long, and a valid stream-discharge value that starts with the 70th character and is 9 characters long. The location and length of each variable may be edited in the initialization file to read different FORTRAN format files. The SREF program creates a tab-delimited column-format input file from these FORTRAN files and then reads the tab-delimited column-format input file.

Streamflow-Statistics Input-File Format

The streamflow-statistics input-file format is a flexible two-column tab-delimited text-file format. The streamflow-statistics input file is read during the record augmentation phase of the analysis so that the regression equation may be used to estimate streamflow statistics at the site of interest from comparable statistics at the index station. The streamflow-statistics input-file format may have one or more comment lines at the head of the file or within the file as long as there are no tabs in the comment lines. There can be one or more input statistics in the file. The format of each input line includes the name of the statistic, a tab character, and the value of the statistic. The name can be almost any combination of letters, numbers, symbols, and spaces (for example, "99.5_%"). A zero value for a statistic at the long-term index station will result in a prediction of zero for that statistic at the site of interest. Streamflow statistics may be entered in any order. It should be noted, however, that the SREF program prints the name of each statistic listed in the input file to the output-statistics file but does not attempt to match the names. If the user is doing a multistation analysis, the streamflow-statistics input file for each index station must have the same number of statistics in the same reporting order to obtain valid results. Otherwise, SREF will combine estimates assumed to represent the same statistic from each index station in the order that they are read from each streamflow-statistics input file.

Output-File Formats

The output-file formats include a paired-record file, an optional period-of-record date file, a maintenance-of-variance regression-results file, an optional KTRLine input file, one or more extended-record files, one or more streamflow-statistics files, and an error file (table 1). The default filename for the paired-data file is the 8-digit short-term station number with the suffix "DAT#.txt" (for example, 01116800DAT0.txt), in which the "#" is an automatically generated sequence number if files with lower sequence numbers exist in the target directory. The default filename for the maintenance-of-variance results file is the short-term-record station number with the suffix "MOV#.txt," which also has a sequence number (for example, 01116800MOV0.txt). The default filename for each KTRLine input file is the 8-digit short-term-record station number, a hyphen, the 8-digit index station number, and the suffix "KT.txt" (for example, 01116800-01117000KT.txt). The default filename for each extended-record file is the short-term-station number with the suffix "MVQ.txt." The user has the option to create a period-of-record date file indicating what data are available for each date of the user-specified long-term daily record. The default filename for the period-of-record date file is the short-term-record station number and the suffix "-PeriodOfRecordDates.txt" (for example, 01116800-PeriodOfRecordDates.txt). An error file is written if there are missing dates in the index-station records or if there is a data-processing error. The error filename is SREFErrFile#.txt, in which the "#" symbol indicates a sequence number (for example, SREFErrFile3.txt).

Paired-Record Output File Format

The paired-record file provides the user with the date of each partial/short-term-record measurement and the corresponding measurements from selected long-term record index stations. This file may be used with other statistical software, spreadsheets, or graphing packages to analyze the data and relations between measurements. Finding and formatting matching measurements from multiple records was one of the most time-consuming aspects of the manual record-extension process. The paired-record file includes a one-line header row with the date, the partial/short-term-record station number, and each long-term-record index station number. Subsequent data rows include the date, the partial/short-term-record measurement and the corresponding streamflow measurement from each index station. It is assumed that all records from the long-term-record index stations will be concurrent with short-term record.

Period-of-Record Date Output File Format

The period-of-record date file allows the user to examine the availability of streamflow data on each day between the user-specified beginning and ending dates of the long-term daily record. The period-of-record date file consists of one tab-delimited line of data for each day in the record. The first line is a heading line with the title "date," the station number for the site of interest, and each long-term index station. The following lines include the date (in YYYY-MM-DD format) and either a one or a zero to indicate whether each station has data (1) or does not have data (0) for each date in the user-specified period of record.

Maintenance-of-Variance Regression-Results Output File Format

The maintenance-of-variance regression-results file provides the user with the MOVE.1 and MOVE.3 regression equations and the regression statistics discussed previously in the Governing Equations and Statistical Methods section. Regression statistics for all stations in the station file are provided for the user. The regression-results file includes the partial/short-term-record station number, the date and time at which the file was created (to help with version control), and some explanatory text. The results for each long-term-record index station are listed sequentially in order that they appear in the station-list input file. For each station, the MOVE.1 statistics are listed first, and then the MOVE.3 statistics are listed. Regression coefficients are listed for each method in several formats. The correlation coefficient and the coefficient of determination (R-squared) are listed next. The maximum and minimum concurrent values (and the common logarithms of these values) are provided to allow evaluation of the potential range of the regression relation. The date, X (long-term-record index, daily mean discharge), Y (partial/short-term-record discharge measurement), the predicted Y value, the residual (or error), the X-leverage value, and the Y-leverage value are provided for each data pair. This information allows the user to do a preliminary analysis, check leverage statistics, and recompute the regression statistics without selected outliers to check the effects of outliers on the regression equations. The mean square error, the root mean square error, the Duan (1983) bias-correction factor, and the modified Nash-Sutcliff coefficient of efficiency (Legates and McCabe, 1999) are provided after the paired-data statistics to help the user evaluate regression results. All of this information is repeated for the MOVE.3 regression results.

Kendall-Theil Robust-Line Input File (Output from SREF) Format

The KTRLine input file is a three-column tab-delimited file written by SREF for use as an input file for the KTRLine program (Granato, 2006). The KTRLine input file provides the paired data necessary to check the parametric regression (MOVE.1 or MOVE.3) results with nonparametric-regression results. These files also allow the user to develop

a multisegment regression model with KTRLine if a one-segment model is insufficient to characterize the relations between stations. A KTRLine input file will be generated for each long-term record station selected if the "Make Kendall-Theil Robust Line Input File" check box is clicked on the SREF input form before the "5. Get Streamflows" command button is clicked. The first line of the file is a column-heading line identifying the selected long-term-record station number (the X variable), the partial/short-term-record station number (the Y variable), and "Date" (because the date is the metadata that helps identify the pair of values).

Extended-Record Output File Format

An extended-record output file will be created if the user does record extension. These files are created by left-clicking the "Generate" command button on one of the record generation forms in the SREF graphical user interface. The extended-record output file consists of the estimated values for the site of interest from data collected at one or more long-term index stations. The extended-record output-file format mimics the (2005) NWISWeb daily value RDB input-file format. The headings of these output files, however, indicate that the streamflow values are estimated with MOVE.1, MOVE.3 or KTRLine equations. The heading lines also indentify the index station or stations used to make the estimates. The extended-record output file also includes comment codes in the daily-value qualification-code column. If a value is estimated, the comment code begins with an "e." The comment code also includes any codes from the long-term index-station files. These codes indicate if the long-term index-station data are approved (A), provisional (P), estimated (e), or have some other qualifier. The SREF program also adds a plus sign "+" if the index-station value is greater than the maximum concurrent value and a minus sign "-" if the index-station value is less than the minimum concurrent value used to develop the MOVE.1 or MOVE.3 equation. These symbols indicate whether the estimated values are extrapolated beyond the concurrent data.

Streamflow-Statistics Output File Format

A streamflow-statistics output file will be created if the user chooses to augment the record for the site of interest. These files are created by left-clicking the "Estimate" command button on one of the record generation forms in the graphical user interface. The file consists of several header lines and statistical streamflow estimates in a tab-delimited RDB format. The streamflow-statistics output file includes an initial header line indicating that the file documents statistical estimates for the station of interest. If the user is doing a one-station analysis, an example of this line would be "Statistical estimates for station: 01117600 from station: 01117500." If the user is doing a multistation analysis an example of this

line would be "Statistical estimates for station: 01117600 from Multiple stations:" This initial line is followed by one (for a one-station analysis) or more (for a multistation analysis) lines that indicate the equation type (MOVE.1, MOVE.3, or KTRLine) and the equation of the line. The next line includes the column-variable names, which are followed by a line of RDB column-format codes. The following lines include the name of the statistic, comment codes, and the estimated statistics for the site of interest. If the user is doing a one-station analysis, each line will include the name, the value of the statistical estimate, and any applicable comment codes. If the user is doing a multistation analysis, each line will include the name, a composite comment code, the value of the composite statistical estimate, and the values of the individual statistical estimates for each station used in the analysis.

The comment codes generated by the SREF program and listed in the streamflow-statistics output file are important for interpreting the results of the analysis. All statistics in this output file will be given a comment code of "e" to distinguish them as estimated values. If an input statistic is above the maximum concurrent value for the associated index station, the comment code will include a plus sign "+" to indicate that the estimated statistic is beyond the range of the regression relation developed with input data. Similarly, if an input statistic is below the minimum concurrent value for the associated index station the comment code will include a minus sign "-." If for example, data from a low-flow partial-record station is being used to estimate flood statistics the existence of "e+" should alert the user that the estimates are beyond the range of the regression relation. If data from multiple stations are being used, there will be one e value and a space, a plus, or a minus in the same order as the stations in the list. These codes are printed so the user can determine which statistics are beyond the range of available data for which stations. For example, a code of "e- -" would indicate that the statistical estimate is below concurrent data for the first and third stations listed in a three-station analysis.

Error Output File Format

If one or more errors are detected during a SREF data-analysis session, error codes will be printed to a sequentially numbered error file. If errors were recorded, the program will prompt the user to examine the error file when the user closes the program. The error-file format consists of an explanatory text line for each nonfatal error encountered in the data-specification and analysis process. Each error line includes, in order, the path and filename of the problem file, the type of problem, and problem details. A common error pertains to date specification at the beginnings or ends of available records. An example of an error message for a specified date beyond the end of available records is: "C:\ExampleData\01121000Q.txt Missing Days between: 9/30/2004 and 9/30/2005."

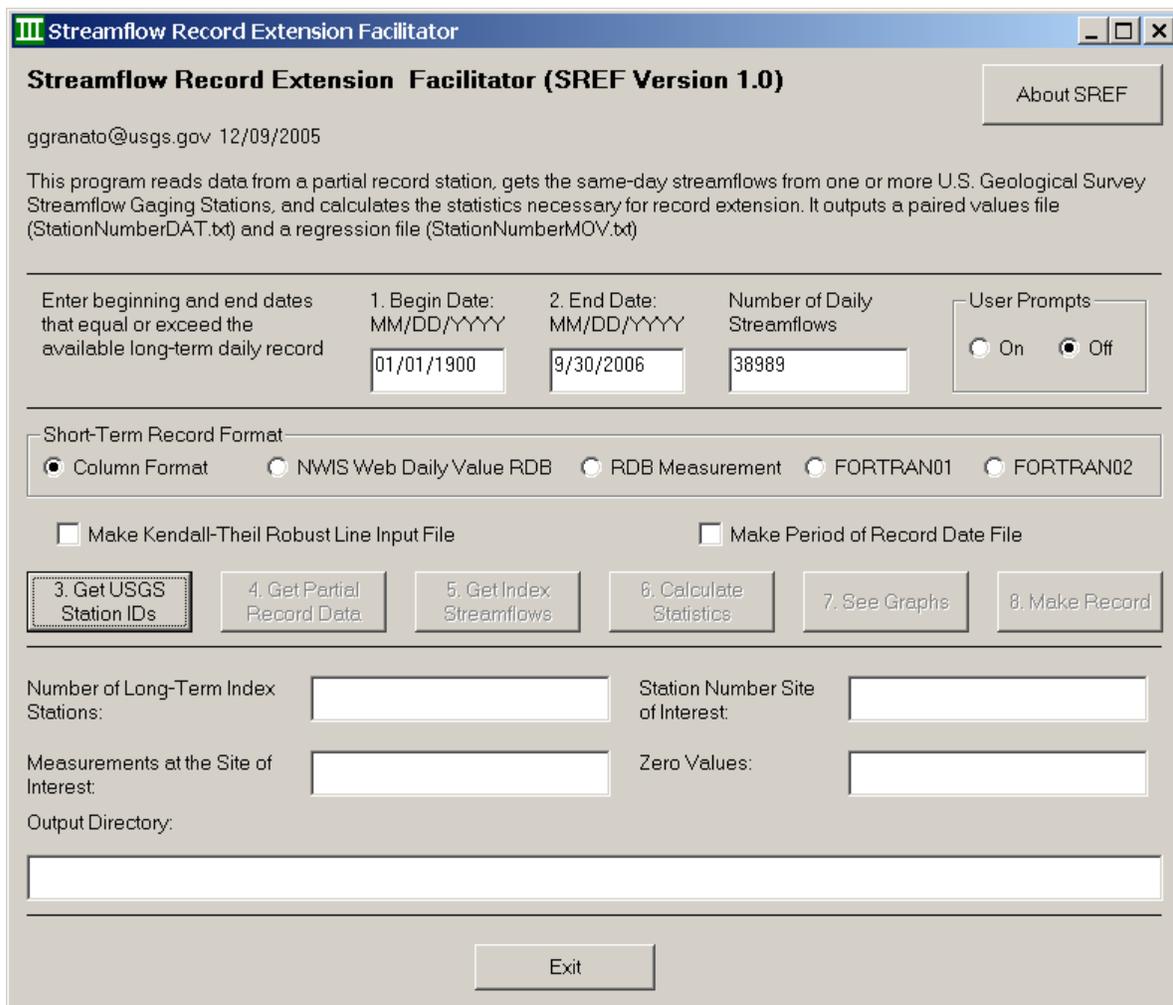**Figure 3-2.** Splash screen for the Streamflow Record Extension Facilitator.

Figure 3-3. Initial view of the Streamflow Record Extension Facilitator input-specification form showing the date fields and the active "3. Get USGS Station IDs" command button.

Graphical User Interface

The graphical user interface consists of six user-input forms and a number of file-specification forms, input boxes, and message boxes to facilitate data input, analysis, and output. The program is activated by left-clicking the icon for the executable program file SREFv1.exe or a Microsoft Windows shortcut to this file on the user's hard drive. When the program is activated, an initial splash screen with information about the program and directions on the user-prompt option is activated (fig. 3-2). The user must left-click the "OK" command button to proceed to the interactive specification form. The interactive specification form (fig. 3-3) is the initial program interface. This form provides information about the program and guides the user through the data-specification process, which is designed to be a linear progression through the steps necessary to define the input-file characteristics, to specify and initialize the output filenames, and to begin the regression analysis. As such, most of the controls on the interactive specification form are initially disabled when the form is initialized and are activated sequentially as they are required in the input-data process. Once a step is complete, each control is disabled so that the user may follow the proper input process. If a mistake is made in the process, the user may close the program at any point and restart the process with only a small investment in time. The user-prompt option near the upper right edge of the interactive specification form (fig. 3-3) provides detailed information about each control on the graphical user interface. By default, the user-prompt option is turned off because it is usually required only during the first use of the program. However, other user prompts may appear to provide information or direction for settings that affect program performance regardless of the user-prompt option.

Input Specification

The interactive specification form (fig. 3-3) provides information and controls the input and analysis process. The interactive specification form provides a graphical user interface to set the beginning and end dates, select input stations, read partial or short-term-record formats for the site of interest, read daily mean streamflow data from index stations, launch the graphing interface, and launch the record-generation interface.

Enter beginning and ending dates: The user must examine the input files and select appropriate dates for the analysis. By default, the beginning date is set to the beginning of the 1900s and the end date is set to the current computer date. The user can specify any beginning or end date as long as the end date is after the beginning date. The maximum number of daily mean streamflows is recalculated automatically as the dates are changed. The beginning and end dates should be selected to match the period of interest during which long-term records are available. The dates may be selected so that provisional streamflow data are used, but provisional data are subject to change until the record is reviewed and approved.

Get USGS Station IDs: After entering the beginning and end dates, the third step in the data-specification process is to read the USGS station identification numbers for the long-term index stations from a station-list file. The standard Windows file-specification common-dialog box is launched when the command button labeled "3. Get USGS Station IDs" is left-clicked. The file selected must be a list of station numbers in the station-list-file format. Selecting the wrong file will cause a nonfatal error message that echoes the content of the first few lines of the selected files. If such an error occurs, the user must reselect this command button and try again. If the index-station specification is correct, the number of stations in the file is displayed in the text box below the command button "3. Get USGS Station IDs," which is disabled, and the next command button "Get Partial Record Data" is enabled so that the user may select the input file for the site of interest (fig. 3-4).

Get Partial-Record or Short-Term Data: The fourth step in the data-specification process is to specify the partial or short-term-record data file. The user must first select one of the five input-format option buttons in the "Short-Term Record Format" option frame. (The input formats are described in the "Input-File Formats section of this manual.) The input formats may be modified by the user by editing the initialization file described in the input-file formats section of this manual. The command button will read "4. Get Partial Record Data" unless the "NWISWeb Daily Value RDB" option is selected, in which case it will read "4. Get Short Record Data." Left-click this command button to proceed. Once this command button is left-clicked, a message box (fig. 3-5) will appear to prompt the user to ensure that the correct file-format option is selected. Once the user confirms the selection by left-clicking "OK" on the message box, the program reads the data file for the site of interest and updates text boxes on the form to show the station number, the number of available values, and the number of zero values in the partial-record/short-term-record file for the site of interest. If the user wishes to test a nonparametric regression analysis, the user must left-click the "Make Kendall-Theil Robust Line Input File" checkbox before clicking the fifth command button because the checkbox will be locked when the fifth command button is clicked.

Get Index Streamflows: The fifth step in the data-specification process is to read the long-term data for the index stations from the station-list file by left-clicking the "5. Get Index Streamflows" command button (fig. 3-6). The SREF program uses the station numbers from the station-list file to locate these files and read the data in the station-list-file directory. Each filename is the station number with a "Q.txt" suffix (for example, 01117500Q.txt). If the data in the long-term-record station do not meet the beginning or end-date criteria or if there are missing data within the record, a nonfatal error will be reported (fig. 3-7). The SREF program will use only paired data to develop the regression equations and only available data to generate estimated streamflow data for the site of interest. If data are generated from one station, missing data in the long-term record will result in missing data in the estimated
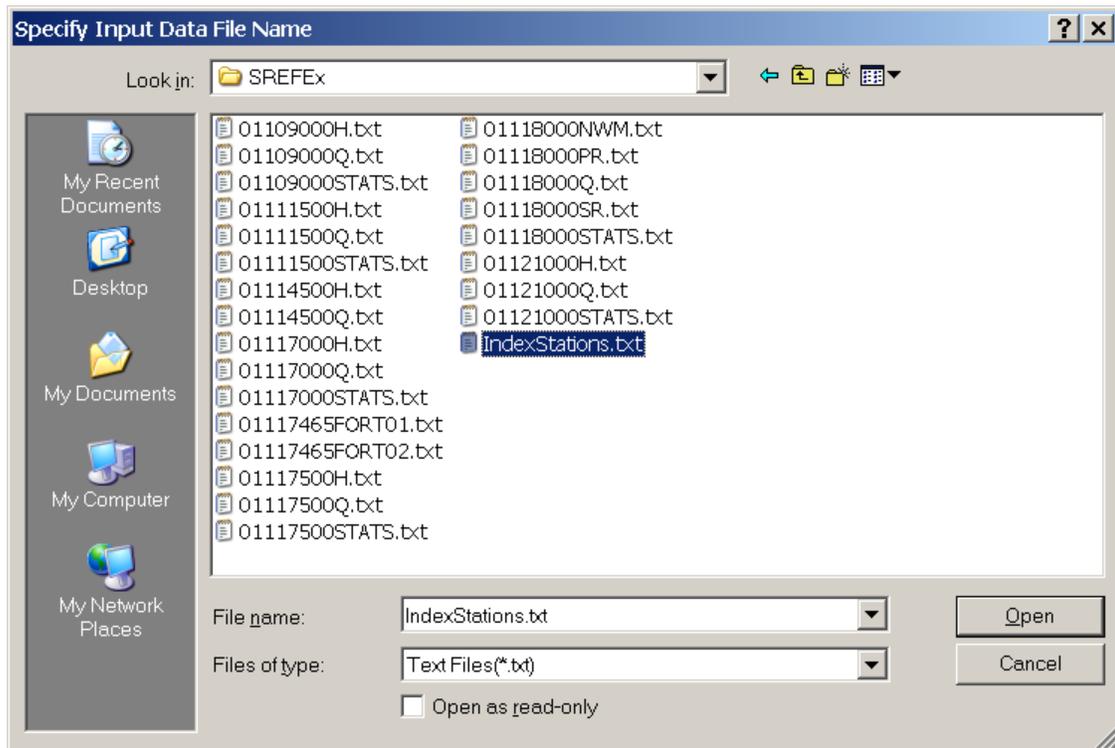
Figure 3-4.   View of an input-file common-dialog form.

Figure 3-5. View of the Streamflow Record Extension Facilitator input-specification form showing the focus on the active "4. Get Partial Record Data" command button, the number of long-term-record index stations, and the output working directory.

Figure 3-6. View of the Streamflow Record Extension Facilitator input-specification form showing the active "5. Get Index Streamflows" command button, the number of measurements, and the number of zero values in the record for the site of interest.
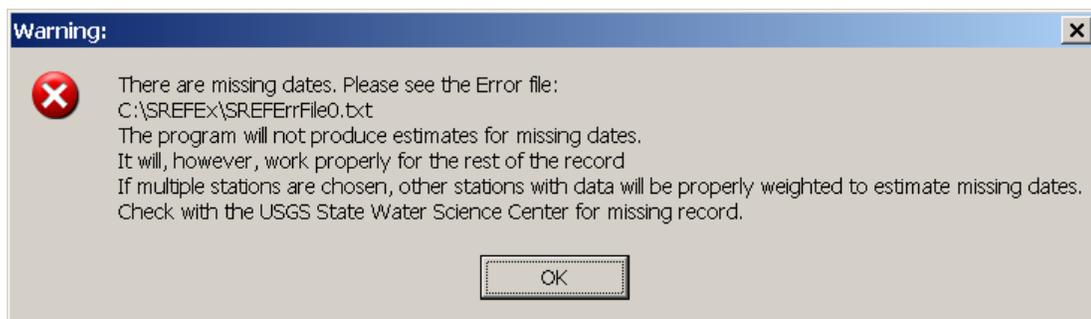
Figure 3-7. View of the nonfatal date-warning notification screen. The nonfatal date warning occurs if one or more long-term-record site measurements in the specified date interval are missing. In this case, the missing dates are the nonapproved record after water year 2004.

record and the dates of missing record will not appear in the output file. If data are generated from multiple stations, however, an estimate will be made for any date within the specified range that is covered by any of the selected stations. In this case, estimates will be reweighted among the stations used for a given day.

Calculate Statistics: The sixth step in the data-specification process is to calculate the regression statistics by left-clicking the "Calculate Statistics" command button. The SREF program automatically calculates regression coefficients and regression statistics in log-log space for each long-term-record index station and prints the results to the maintenance-of-variance regression-results file. Once the calculations are complete, the SREF program enables the "See Graphs" and "Make Record" command buttons (fig. 3-8). These command buttons allow the user to open the graphing interface and record generation interface (for record extension and augmentation), respectively.

Graph Selection and Viewing

Graphs of the data, regression lines, and regression residuals provide information about the data and statistical relations that are difficult to get by use of numerical statistics (Helsel and Hirsch, 2002). For example, relations between data from a site of interest and an index station may have high correlation coefficients, but the relation between measurements may be curved rather than linear. Similarly, relations between measurements may be best described by using a multiline regression model, especially if different physical processes dominate within different flow ranges. The SREF program provides six options for graphical analysis of data, regression lines, and residuals.

"See Graphs" and the Graph Selection Form: The seventh step in the analysis process on the interactive specification form is to view the graphs of data and regression lines and the associated residual plots by left-clicking the

Figure 3-8.   View of the Streamflow Record Extension Facilitator input-specification form showing the active "7. See Graphs" command button.

"7. See Graphs" command button (fig. 3-8). This command button launches the Graph Selection Form (fig. 3-9). The Graph Selection Form allows the user to select a long-term index station in the combobox at the top of the form. As the combobox is activated, a list of station numbers and the associated correlation coefficients appears. The user may select a station by right-clicking on the station number in the combobox-popup list. The station-identification number and the correlation coefficient are provided so the user may select the most promising sites for further analysis. This form allows the user to select four options for viewing the paired data and the regression lines. These options are the paired data, the paired data and the MOVE.1 line, the paired data and the MOVE.3 line, or the paired data and both regression lines (ALL). The form also allows the user to select four options for viewing residual plots. Residuals for the MOVE.1 line or the MOVE.3 may be plotted with respect to the predictor variable (X_i) or with the predicted (Y_i) values. The user may close the Graph Selection Form and relaunch it from the input-specification form with different options multiple times. These choices allow the user to graph data, regression lines, and residuals from another index site.

Paired Data and Regression Lines: If one of the data and regression-line options is selected, left-clicking the "Make Graph" command button on the Graph Selection Form (fig. 3-9) launches the Graphing Form with paired data or the paired data with the regression lines for the selected index station (fig. 3-10). The base 10 (common) logarithms of data are plotted on the graph. The daily mean streamflow data from the long-term index station are plotted with respect to the horizontal X axis as predictor variables, and the paired values from the site of interest are plotted with respect to the Y axis as the predicted variables. The lines may be plotted with data depending on the selection option. If the MOVE.1 and MOVE.3 regression lines are selected, the equation for each line is printed along the bottom of the graph. Figure 3-10A shows the graph of results for a 55-point set of randomly selected data from March 1941 through September 2004 for the Wood River at Hope Valley (USGS station 01118000) plotted against paired data from the Pawcatuck River at Wood River Junction (USGS station 01117500). Figure 3-10B shows the graph of results for a 1,725-point data set from January 2000 through September 2004 for the same stations.

Regression Residuals: If one of the residual-plot options is selected, left-clicking the "Make Graph" command button on the Graph Selection Form (fig. 3-9) launches the Graphing Form with the selected regression-residual plot. A regression residual is the difference between the predicted and measured Y value for a given X value (equation 12). Residuals for the MOVE.1 or MOVE.3 predictions may be plotted against the predictor (X_i) (fig. 3-10C) or predicted (Y_i) values (fig. 3-10D). The predicted data for a particular station may have a good correlation coefficient but the regression line may not be suitable if the residuals do not appear random with constant variance for the full range of X_i or Y_i. (The USGS techniques of water-resources investigations report by Helsel and Hirsch (2002) is available on the internet and describes the graphical interpretation of residual plots.) The user may select a
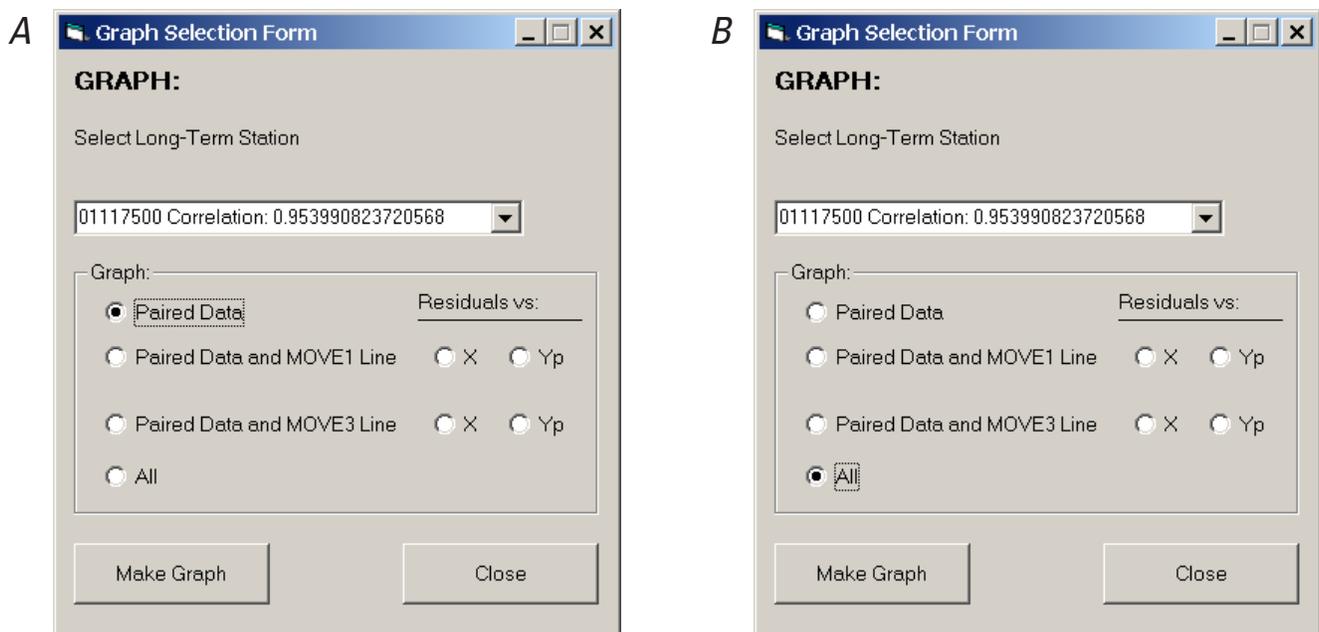

Figure 3-9.   View of the Graph-selection form showing a long-term-record selection in the combobox and (A) the "Paired Data" graph option and (B) the "All" (paired data, MOVE.1 line, and MOVE.3 line) graph option selected.
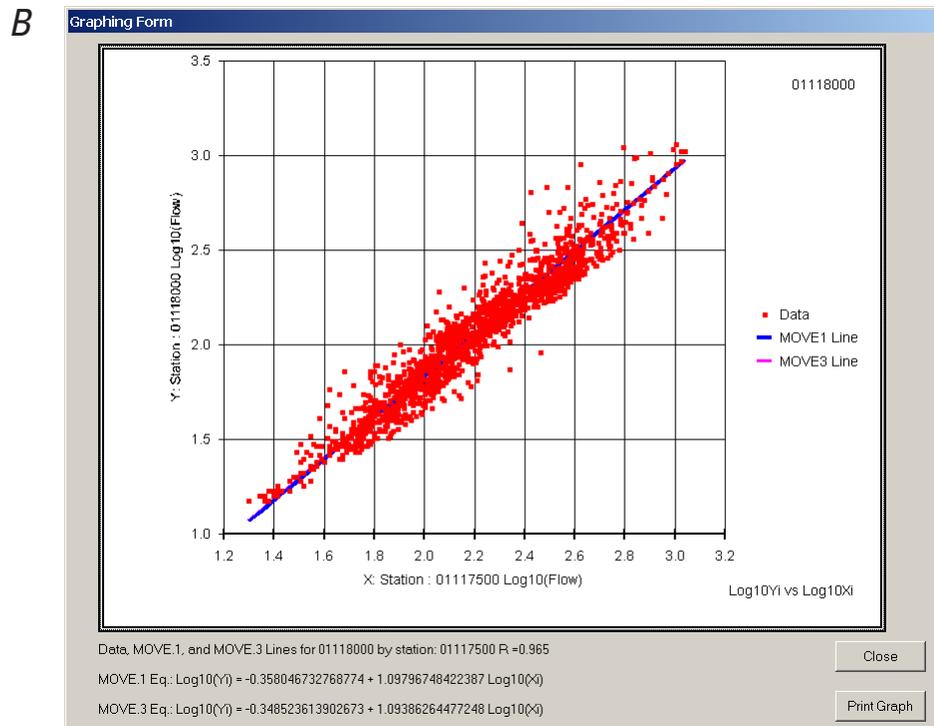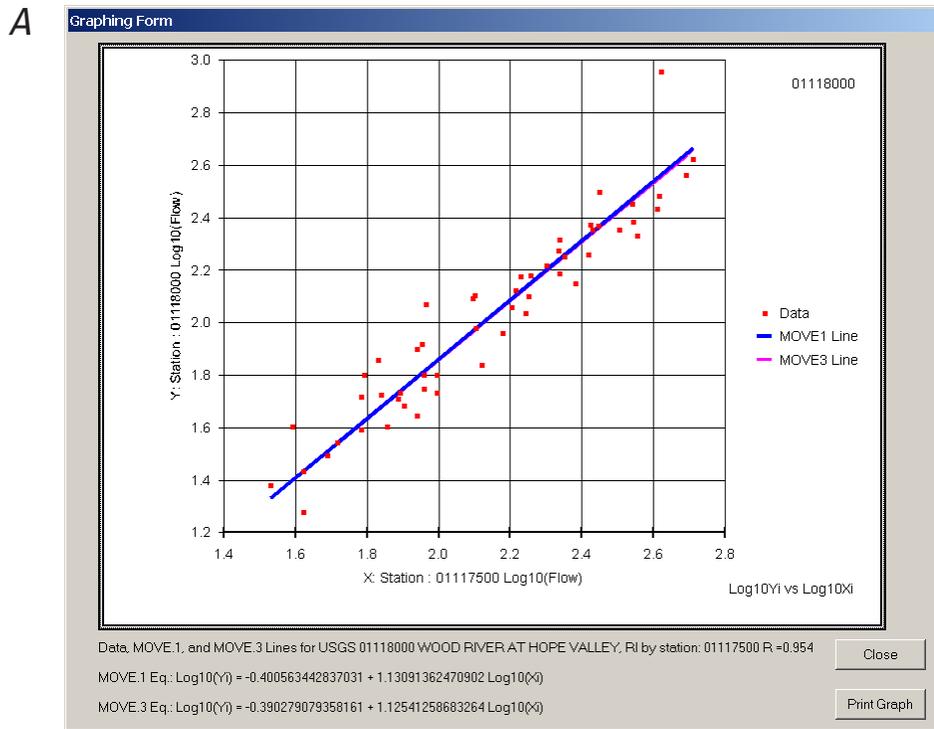
A

B

Figure 3-10. View of the graphing form showing (A) MOVE.1 and MOVE.3 lines for a partial record (B) MOVE.1 and MOVE.3 lines for a short-term record, (C) the residuals plotted with respect to X data, and (D) the residuals plotted with respect to Y data.
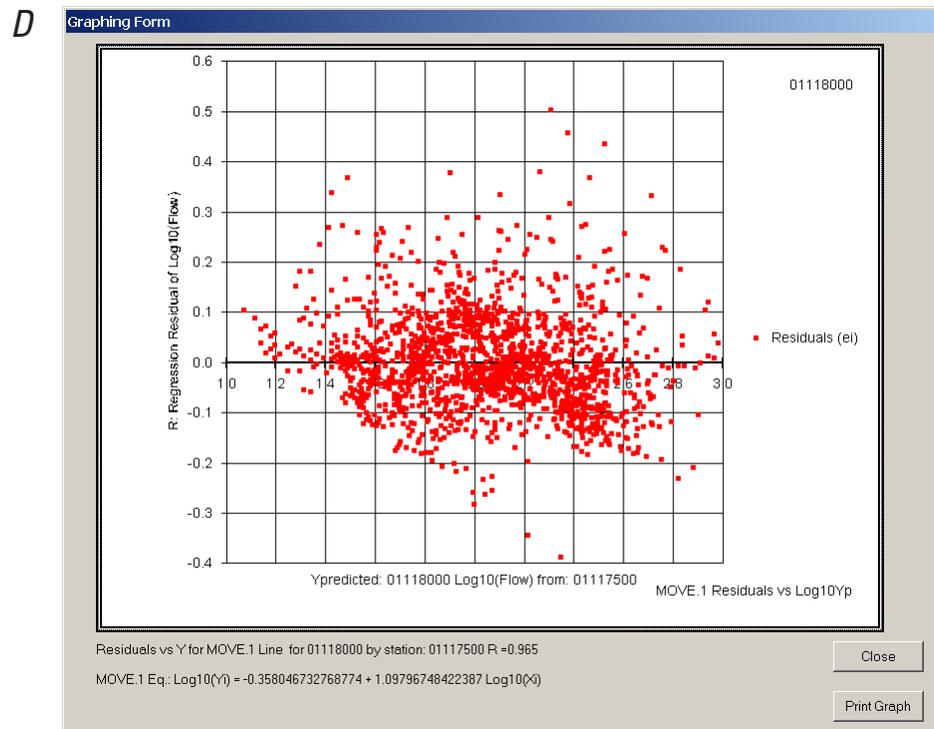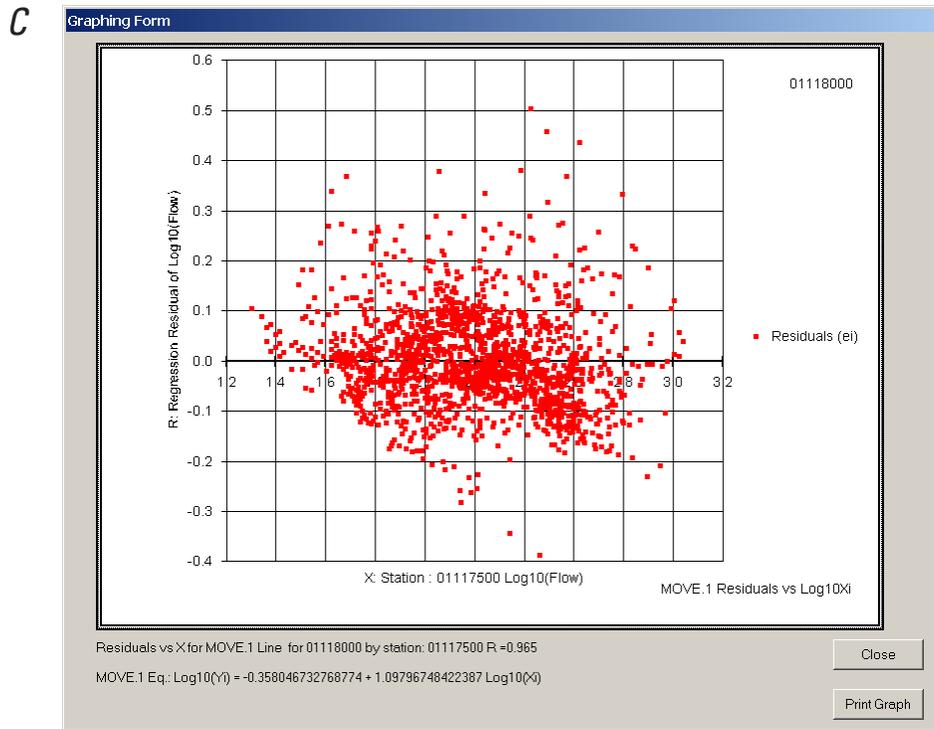
C

D

Figure 3-10. View of the graphing form showing (A) MOVE.1 and MOVE.3 lines for a partial record (B) MOVE.1 and MOVE.3 lines for a short-term record, (C) the residuals plotted with respect to X data, and (D) the residuals plotted with respect to Y data.—Continued

different station or consider the possibility of developing a multisegment model to characterize different relations for different processes. The residual plots may help the user to identify high-leverage outliers that may affect the regression estimates.

Data-Point Identification: The graphing form has a data-identification utility that allows the user to see information about a selected data point by left-clicking the point on the graph. When the user left-clicks a point on the graph, a message box appears (fig. 3-11) providing information about the point selected. For most choices, the message box identifies the long-term and short-term station numbers, the measurement date, streamflow values, the logarithms of the streamflow values, the streamflow value predicted with the MOVE.1 or MOVE.3 equation, and the associated residual value (fig. 3-11). If the graph shows both the MOVE.1 and MOVE.3 equations, then the message box also shows both predicted value Y_i and associated residual values (fig. 3-11). The user may append this information to the end of the maintenance-of-variance regression-results file for subsequent examination of potential outliers by left-clicking the "Yes" command button. The user can return to the graphing form without saving the data-point-identification information by left-clicking the "No" command button.

Printing and Saving the Graphing Form: Once the user has examined the data and regression lines they may print the graphing form to document the analysis or close the form to return to the Graph Selection Form. The user may print the Graphing Form by left-clicking the "Print Graph" command button and saving an image of the Graphing Form. The user may print to a file if they have a print-to-image utility (such as an Adobe PDF writer). On Windows-based personal computers, the user can do a screen capture of the active form. To do a screen capture, the user must left-click the title bar of the form to ensure that the form has the focus, hold down the "Alt" button and press the "Print Scrn" button. The user must paste the captured image into software that will accept images, such as Microsoft Paint, Word, or PowerPoint.

Streamflow-Record Extension and Augmentation

The SREF program was written to facilitate record extension (generation of individual daily mean streamflow values at the site of interest from index-station values) and record augmentation (estimation of values of selected streamflow statistics at the site of interest from index-station values). The streamflow-record extension and augmentation forms provide an interface to complete these tasks. The user can extend or augment the record by using the MOVE.1 technique, the MOVE.3 technique, and by entering regression-line statistics from KTRLine program or another regression tool.

The Select Record Generation Method Form: The eighth step in the process for record extension or augmentation is record generation. The user must left-click the "8. Make Record" command button (fig. 3-8) to launch the Select Record Generation Method form (fig. 3-12). This form allows

the user to select a record-generation method. The first option allows the user to extend or augment the record with either the MOVE.1 or MOVE.3 equations by selecting one long-term index station. The second option allows the user to do these tasks by selecting one or more long-term index stations. The third option allows the user to extend or augment the record with a single-segment or multisegment equation developed with the KTRLine program (Granato, 2006) or another regression tool. To initiate the process the user must left-click the desired option and then left-click the "Generate Record" command button.

Use MOVE.1 or MOVE.3 with One Station: If this option is selected, right-clicking the "Generate Record" command button on the Select Record Generation Method form (fig. 3-12) will launch the Generate Long-Term Record form (fig. 3-13A). The Select Record Generation Method form allows the user to select a long-term station of interest in the combobox at the top of the form. As the combobox is activated, a list of station numbers and the associated correlation coefficients appears (fig. 3-13B). The user may select a station by right-clicking on the station number in the combobox popup list. The user may select different options and generate a daily mean streamflow record or estimate selected streamflow statistics for the site of interest. The user also has the option to apply the Duan (1983) BCF to the record that is generated. The user may generate the record or estimate statistics with either MOVE.1 or MOVE.3 by selecting the associated option. For record generation, the user may select either predicted or measured values for days with concurrent measurements. Measured values of daily mean streamflow for a short-term-record site are preferable to predicted values, but hydrographs of the common period may indicate the representativeness of the estimated record. The user must decide if instantaneous flow values for partial-record sites are better than estimated daily mean streamflow records for subsequent use of the generated record. Comparisons may be done outside the SREF interface after records are generated

If the record of the selected long-term index station includes one or more zero flows, a warning message box indicating the presence of zero flows (fig. 3-14) will appear when the station is selected, and the zero-flow option frame will be enabled (fig. 3-15). Similarly, if the record of the default long-term index station (the first station in the index station list) includes zero flows, the user will see a warning message box (fig. 3-14) indicating the presence of zero flows before the record-generation form appears, and the zero-flow option frame will be enabled when the form appears (fig. 3-15). The user has four choices for dealing with zero-flow values in the long-term record, including the use of zero flows for the index site, the use of a substitute value of 0.01 ft^3/s, the use of a specified substitute value (fig. 3-16), or the use of a recession constant (fig. 3-17). Zero-flow estimates will be annotated with the comment code "e0" in the output file. The zero-flow options were discussed in this program manual in the section "Governing Equations and Statistical Methods."

Print to output data file:C:\SREFEx\01118000MOV0.txt

Point selected:
Long-Term Gaging Station (Xi): 01117500
Partial Record Gaging Station (Yi): 01118000
Measurement Date: 6/18/2001
Data:
Xi: 419   Log10(Xi): 2.6222
Yi: 906   Log10(Yi): 2.9571
MOVE.1 Log10(Yi): 2.5649 Residual for Log10(Yi): 0.39219
MOVE.3 Log10(Yi): 2.5608 Residual for Log10(Yi): 0.39633

Yes   No

Figure 3-11.   View of the point-identification message box allowing the user to view and save information about paired data points on the graph that are selected by left-clicking the point of interest.

Select Record Generation Method

Select Generation Method

◉ Use MOVE.1 or MOVE.3 with one station

○ Use MOVE.1 or MOVE.3 with multiple stations

○ Use Kendall-Theil Robust Line equation(s)

Generate Record   Close This Form
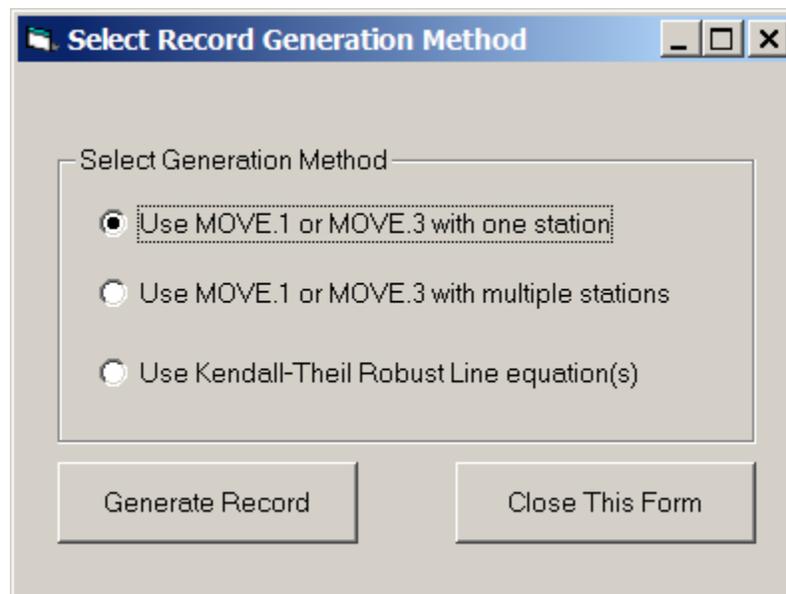
Figure 3-12.   View of the Select Record Generation Method form.

A

B

Figure 3-13.   View of the one-station MOVE.1, MOVE.3 record-generation form showing (A) the initial view and (B) a view with the combobox activated for the selection of index stations.

Figure 3-14. View of the zero-flow warning-message box indicating that the record of the selected index station includes one or more zero flow values.
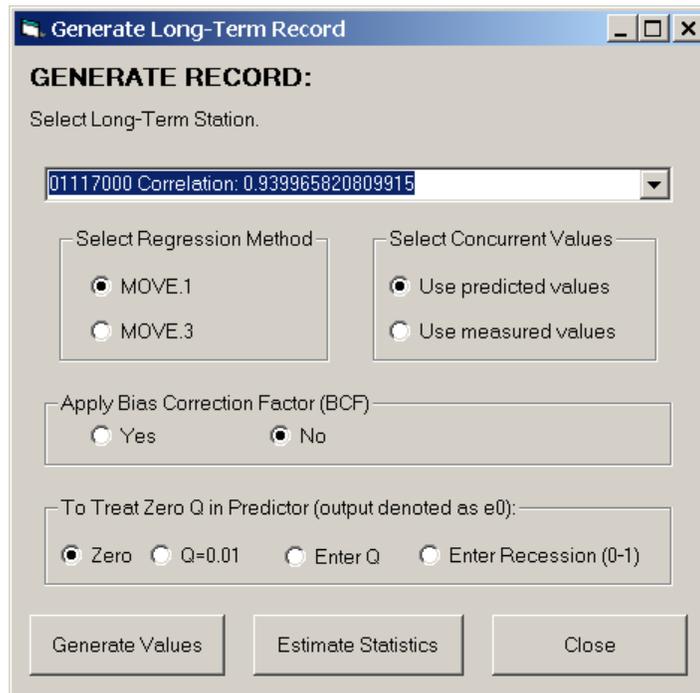
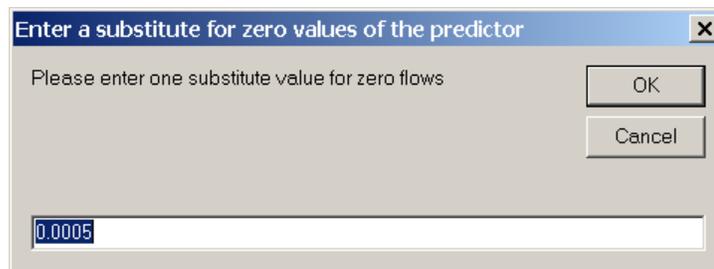**Figure 3-15.** View of the one-station MOVE.1, MOVE.3 record-generation form showing the zero-flow substitution options.

Figure 3-16. View of the zero-flow substitution input box for defining a single user-defined replacement value for all zero flows in the index-station record.

The user may select the "Zero" option on the record-generation form (fig. 3-15) to specify the use of zero-flow values for the index station that will result in zero-flow values for the site of interest. This option may be a good choice if both the index and short-term-record stations are believed to be intermittent or ephemeral streams, and it can be assumed that zero flows at the index site are a good predictor for zero flows at the site of interest. The user may wish to examine this assumption with a logistic-regression analysis.

The user may select the "Q=0.01" option to specify a minimum value of 0.01 ft^3/s, which is a common detection limit for measuring streamflow in a natural channel. In this case, the 0.01 ft^3/s value will be applied in the regression equation as a predictor (X) to estimate a value (Y) at the short-term/partial-record station of interest. This option may be used if it is believed that a zero flow in the index record actually represents a small nonzero flow, or that consumptive water use in the index basin may cause a small number of zero flows that are not representative of conditions in the basin of interest.

The user may choose "Enter Q" to enter a single index-station streamflow value to be substituted for every zero flow in the record (for example, the minimum nonzero flow in the index-station record). If this option is clicked, an input box will appear (fig. 3-16) with a default value of 0.0005; however, the user can enter an estimated substitute value. If an invalid entry is made (for example, a non-numeric or negative value), the user will be given the choice to retry or cancel. If the user selects "Cancel," the program will reset the default to zero. If the user decides to try again, the input box will display the default value or last valid choice entered. Similarly, if the user selects a valid choice and then clicks another option from the zero-flow option frame, the last valid choice entered in the input box will be preserved as the default value if the user reselects the "Enter Q" option. This option also may be used if it is believed that a zero flow in the index record actually represents a small nonzero flow or that consumptive water use in the index basin may cause a small number of zero flows that are not representative of conditions in the basin of interest.

The final option for replacing zero values in the index station record is the "Enter Recession (0-1)" option. When this option is selected, an input box appears prompting the user to enter a recession constant for the index station (fig. 3-17). The user must enter a recession constant that is greater than zero and less than or equal to one. The recession constant is applied to the index streamflows, not the logarithm of index streamflows. If this option is selected, the program will apply the recession constant to the last nonzero streamflow value to estimate a nonzero flow as a replacement for the first zero-flow value in a series. If the flow on the next day is zero, the program will apply the recession constant to the estimate of the index-station flow for the previous day, and so forth. For example, if the last measured nonzero flow is 1.0 ft^3/s followed by four zero flows, selection of a recession constant of 0.8 would result in the series 0.8, 0.64, 0.51, and 0.41 ft^3/s; selection of recession constant of 0.5 would result in the series

0.5, 0.25, 0.125, and 0.0625 ft^3/s to replace the zeros as the predictor value (X) in the regression equation. This zero-flow option is most consistent with base-flow-recession theory (Chow and others, 1988). The user may analyze the index-station record to estimate a suitable base-flow-recession constant. This method may produce streamflow estimates that are effectively zero; such extreme low-flow values are probably extrapolations beyond the data used to develop the regression equations.

The user can generate a daily mean streamflow record for the site of interest in the NWISWeb RDB daily value format by left-clicking the "Generate" command button once the appropriate options for record extension have been selected (fig. 3-13). When the user initiates the process, a Microsoft Windows common-dialog box appears so the user can save the generated record to the target directory or to a user-specified subdirectory (fig. 3-18). The default filename for these records is the short-term/partial-record-site number, the text "by," the long-term index-site number, and the equation suffix "MV1Q.txt" or "MV3Q.txt." The user may modify the default filename, especially to explore the effect of different specification options. The RDB header comments in the output file indicate the index station, the regression method (MOVE.1 or MOVE.3), and the regression equation used.

The user can estimate streamflow statistics for the site of interest in the streamflow-statistics output-file format by left-clicking the "Estimate" command button once the appropriate options for record augmentation have been selected (fig. 3-13). The streamflow augmentation function is dependent on the existence of a properly formatted streamflow-statistics input file designated with the proper filename. When the user initiates the augmentation process, a Microsoft Windows common-dialog box appears so the user can save the estimated statistics to the target directory or to a user-specified subdirectory (fig. 3-18). The default filename for the streamflow-statistics output-file is the short-term/partial-record-site number, the text "by," the long-term index-site number, and the equation suffix "STATSMV1Q.txt" or "STATSMV3Q.txt." The user may modify the default filename, especially to explore the effect of different specification options. The streamflow-statistics output-file format is described in the section "Output-File Formats" in this manual.

Use MOVE.1 or MOVE.3 with Multiple Stations: If this option is selected, right-clicking the "Generate Record" command button on the Select Record Generation Method form (fig. 3-12) will cause the program to launch the Multiple Station Record Generation form (fig. 3-19). This form provides the same options and methods as the Generate Long-Term Record form (fig. 3-13). The user is directed to that subsection to read the description of each option for choosing the regression method, concurrent values, BCF, and treatment of zero flows in the index-station record. The difference between the forms is that all the index-station numbers appear with their respective correlation coefficients in a list box that allows multiple selections rather than in a combobox, which
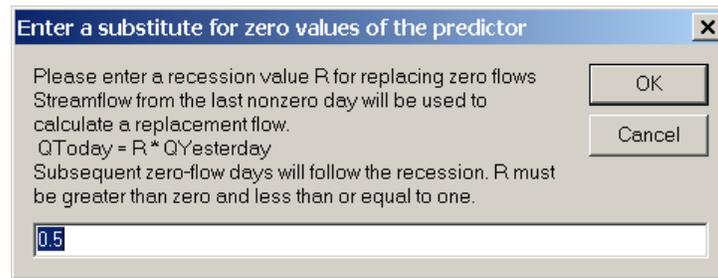
Figure 3-17. View of the zero-flow substitution input box for defining a user-defined recession constant for calculating replacement values for all zero flows in the index-station record.
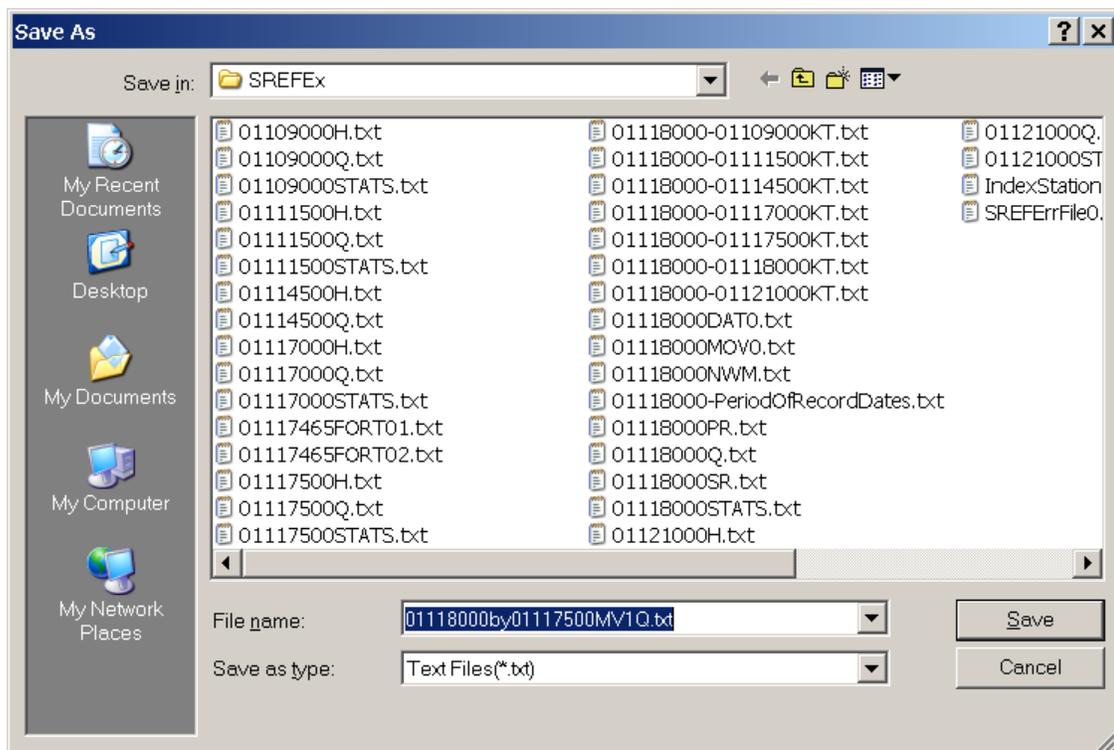

Figure 3-18. View of the common-dialog box for saving the estimated-record values. The default filename indicates the partial-record station (01118000), the index station (01117500), and the extension method (MV1).

limits the user to one selection. If more than 11 index stations are specified in the input file, a scroll bar will appear on the right side of the list box to allow the user to scroll down and select additional stations.

To select two or more sequential stations in the list, the user may hold the shift key and left-click at the beginning and end of the sequential-station list. To select nonsequential stations or nonsequential stations in addition to a sequential-station list, the user must hold the control key and left-click any additional stations. Left-clicking an individual station will clear a list. Holding the control key and left-clicking will allow the user to deselect individual stations. Once the user selects the index stations, left clicking the "Finalize Multiple Index-Station Selection" command button (fig. 3-19) will lock the list and enable the "Generate Values" and "Estimate Statistics" command buttons. The default filenames for multistation-record generation include the short-term/partial-record-station number and the appropriate method suffix. Because the multistation estimates are weighted by the mean-square error (discussed in the "Record Extension With Multiple Index Stations" subsection in the "Governing Equations and Statistical Methods" section), the user is advised to examine individual equation regression statistics in the regression-results output file.

Use Kendall-Theil Robust Line Equation(s) Option: If this option is selected, right-clicking the "Generate Record" command button (fig. 3-12) will cause the program to launch the KTRLine Record Generation form. KTRLine has a number of options for developing single-segment or multisegment regression lines, including data transformations for the X and Y variables (Granato, 2006). Figure 3-20 shows an example of an implementation of the KTRLine program with the same data that is displayed in figure 3-10B. The output file from the KTRLine program includes the regression coefficients and statistics necessary for generating a long-term record or selected streamflow statistics for the site of interest. The KTRLine Record Generation form in the SREF program (fig. 3-21) is closely matched to the output of the KTRLine program. The first step in using this form is to select the long-term index station from the combobox at the top of the form. As in the other record-generation forms, the default selection is the first station in the index-station list. Station selection is critical in the KTRLine Record Generation form because the record-generation equation is entered manually rather than being automatically paired with the station selection. The next selection indicates the number of segments in the regression model. This selection will enable or disable the regression-equation input text boxes lower on the form (fig. 3-21).

The KTRLine Record Generation form (fig. 3-21) allows the user to specify the transformation for the predictor (X) and predicted (Y) variables independently. The user must enter the intercept, slope, the BCF, and maximum X value for each segment. The maximum X value for each segment is used as the breakpoint between segments for a multisegment model. It is important to note that all of these regression parameters must be entered with respect to the selected transformation as they appear in the KTRLine program output. If, for example, the regression parameters are developed in log-log space the regression coefficients would be for the linear equation in log-log space.

The KTRLine Record Generation form (fig. 3-21) includes the same options as the one-station record-generation form (fig. 3-13A). The common options are the use of concurrent values, application of the BCF, and treatment of zero flows in the predictor record. The user is directed to the subsection "Use MOVE.1 or MOVE.3 with One Station" option on to read the description of each common option on the forms.

Summary

The Streamflow Record Extension Facilitator (SREF Version 1.0) program was developed by the U.S. Geological Survey (USGS) in cooperation with the Federal Highway Administration to estimate long-term streamflow characteristics at sites with limited data. The SREF program allows the user to extend (estimate daily mean streamflow values) or augment (estimate statistics) a record for a site of interest. Record extension and augmentation are based on the assumption that long-term streamflow records from hydrologically similar sites can be used to estimate values at a site of interest with limited data. SREF implements the maintenance of variance extension type 1 (MOVE.1) method, and the maintenance of variance extension type 3 (MOVE.3) method. SREF also facilitates record extension and augmentation with a single or multisegment regression model developed by use of the nonparametric Kendall-Theil Robust Line (KTRLine) program.

This manual describes the implementation, use, and interpretation of results from the SREF program. SREF was written in Microsoft Visual Basic 6.0 and has a graphical user interface that follows standard Microsoft Windows interface conventions. An executable version of the program, example files, and the Visual Basic source code are documented in the SREF directory on the computer disk containing this manual.

Acknowledgments

Figure 3-19. View of the Multiple Station Record Generation form showing two stations selected and the zero-flow substitution options active.
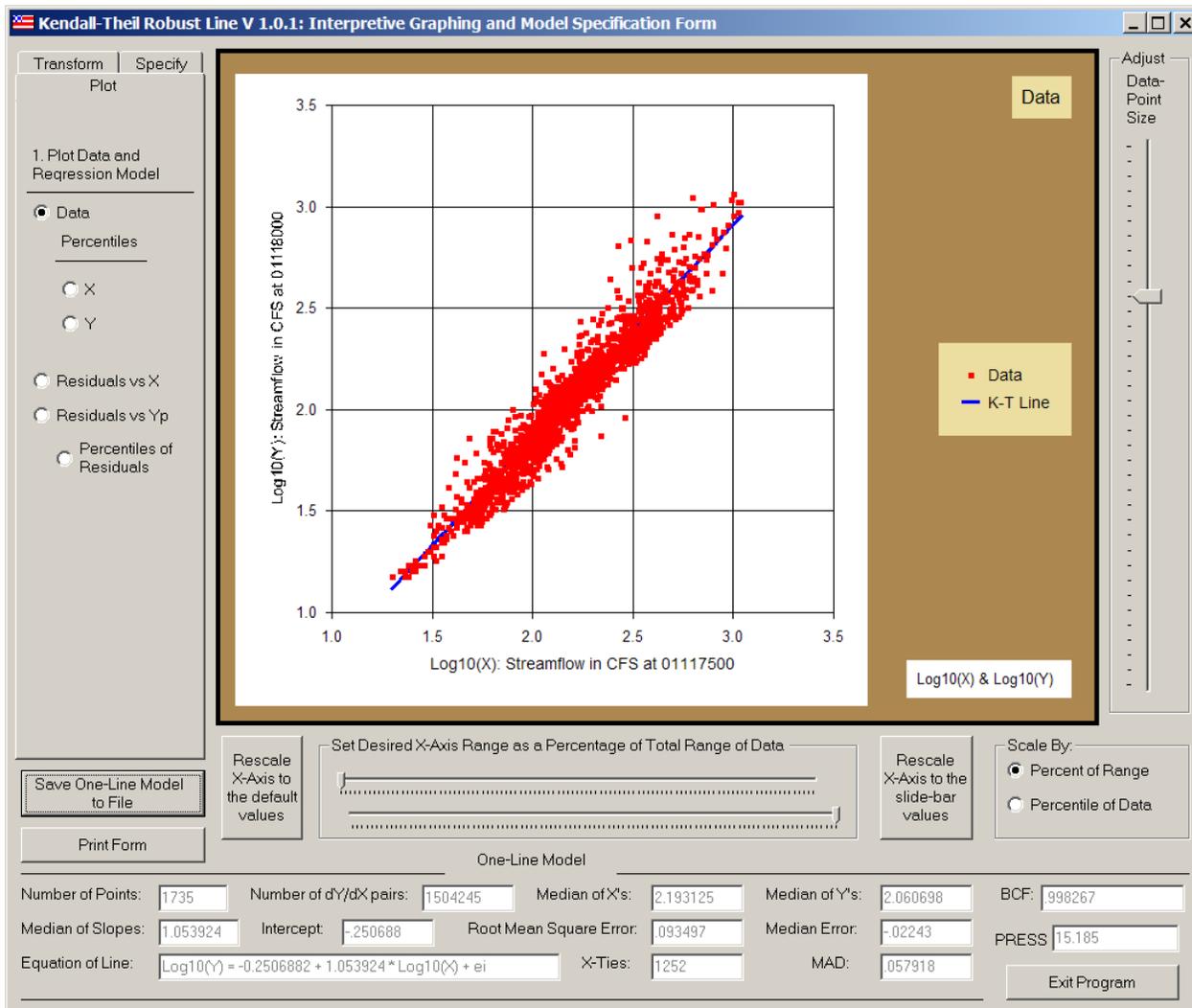
Figure 3-20.   View of the Kendall-Theil Robust Line (Granato, 2006) Interpretive Graphing and Model-Specification form showing results of a nonparametric log-log regression analysis based on the data displayed in figure 3-10B.

Figure 3-21. View of the Kendall-Theil Robust Line Record-Generation form showing input for a one-segment log-log model.

References Cited

Benson, M.A., and Carter, R.W., 1973, A national study of the streamflow data-collection program: United States Geological Survey Water-Supply Paper 2028, 44 p.

Chow, V.T., Maidment, D.R., and Mays, L.W., 1988, Applied Hydrology: New York, McGraw-Hill Inc., 572 p.

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., and Summers, R.M., 1992, The validity of a simple log-linear model for estimating fluvial constituent loads—An empirical study involving nutrient loads entering Chesapeake Bay: Water Resources Research, v. 25, no. 9, p. 2353–2363.

Crawford, C.G., 1991, Estimation of suspended-sediment rating curves and mean suspended-sediment loads: Journal of Hydrology, v. 129, p. 331–348.

DiToro, D.M., 1984, Probability model of stream quality due to runoff: Journal of Environmental Engineering, v. 110, no. 3, p. 607–628.

Driscoll, E.D., Shelley, P.E., Gaboury, D.R., and Salhotra, Atul, 1989, A probabilistic methodology for analyzing water quality effects of urban runoff on rivers and streams: U.S. Environmental Protection Agency, Office of Water, EPA 841-R89-101, 128 p.

Driscoll, E.D., Shelley, P.E., and Strecker, E.W., 1990a, Pollutant loadings and impacts from highway stormwater runoff, volume I—Design procedure: Federal Highway Administration Final Report FHWA-RD-88-006, 67 p.

Driscoll, E.D., Shelley, P.E., and Strecker, E.W., 1990b, Pollutant loadings and impacts from highway stormwater runoff, volume II—User's guide for interactive computer implementation of design procedure: Federal Highway Administration Final Report FHWA-RD-88-007, 23 p.

Duan, Naihua, 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610.

Gilroy, E.J., Hirsch, R.M., and Cohn, T.A., 1990, Mean-square error of regression-based constituent transport estimates: Water Resources Research, v. 26, no. 9, p. 2069–2077.

Glysson, G.D., 1987, Sediment-transport curves: U.S. Geological Survey Open-File Report 87-218, 47 p.

Granato, G.E., 2006, Kendall-Theil Robust Line (KTRLine—version 1.0)—A Visual Basic program for calculating and graphing robust-nonparametric estimates of linear-regression coefficients between two continuous-variables: Techniques and Methods of the U.S. Geological Survey, chap. A7, book 4, 31 p.

Helsel, D.R., 2004, Nondetects and data analysis—Statistics for censored environmental data: Hoboken, N.J., John Wiley and Sons Publishing Inc., 268 p.

Helsel D.R., and Hirsch, R.M., 2002, Statistical methods in water resources—Hydrologic analysis and interpretation: Techniques of Water-Resources Investigations of the U.S. Geological Survey, chap. A3, book 4, 510 p.

Hirsch, R.M., 1979, An evaluation of some record extension techniques: Water Resources Research, v. 15, no. 6, p. 1781–1790.

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

Hirsch, R.M., and Gilroy, E.J., 1984, Methods of fitting a straight line to data—examples in water resources: Water Resources Bulletin, v. 20, no. 5, p. 705–711.

Hirsch, R.M., Helsel D.R., Gilroy, E.J., and Cohn, T.A., 1993, Chapter 17—Statistical analysis of hydrologic data, *in* Maidment D.R., ed., Handbook of Hydrology: New York, McGraw-Hill Book Company, p. 17.1–17.55.

Koch, R.W., and Smillie, G.M., 1986, Bias in hydrologic prediction using log-transformed regression models: Water Resources Bulletin, v. 22, no. 5, p. 717–723.

Legates, D.R., McCabe G.J., 1999, Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation: Water Resources Research, v. 35, no.1, p. 233–241.

Linsley, R.K., Jr., Kohler, M.A., and Paulhus, J.L.H., 1975, Hydrology for engineers, 2d ed., New York, McGraw-Hill Book Company, 482 p.

Manis, Rod, Schaffer, Evan, and Jorgensen, Robert, 1988, Unix relational database management—application and development in the Unix environment: Englewood Cliffs NJ, Prentice Hall, 476 p.

Matalas, N.C., and Jacobs, Barbara, 1964, A correlation procedure for augmenting hydrologic data: U.S. Geological Survey Professional Paper 434-E, 7 p.

Mathey, S.B., ed., 1998, National Water Information System: U.S. Geological Survey Fact Sheet FS-027-98, 2 p.

National Research Council, 2004, Assessing the National Streamflow Information Program: Washington, DC, The National Academies Press, 164 p.

O'Connor, D.J., 1976, The concentration of dissolved solids and river flow: Water Resources Research, v. 12, no. 2, p. 279–294.

Rantz, S.E., and others, 1982, Measurement and computation of streamflow—Volume 1 and 2, Computation of discharge: U.S. Geological Survey Water-Supply Paper 2175, 631 p.

Ries, K.G., III, and Friesz, P.J., 2000, Methods for estimating low-flow statistics for Massachusetts streams: U.S. Geological Survey Water-Resources Investigations Report 2000-4135, 81 p.

Riggs, H.C., 1972, Low-flow investigations: U.S. Geological Survey Techniques of Water-Resources Investigations, Book 4, chap. B1, 18 p.

Rossman, L.A., 1990, DFLOW User's Manual: Cincinnati, OH, U.S. Environmental Protection Agency, Office of Research and Development, Risk Reduction Engineering Laboratory, 26 p.

Saunders, J.F., III, Murphy, Marylee, Clark, Martyn, and Lewis, W.M., 2004, The influence of climate variation on the estimation of low flows used to protect water quality—a nationwide assessment: Journal of the American Water Resources Association, v. 40, no. 5, p. 1339–1349.

Searcy, J.K., 1959, Flow-duration curves, in Manual of Hydrology—Part 2. Low-flow techniques: U.S. Geological Survey Water-Supply Paper 1542-A, p. 1–33.

Sloto, R.A., and Crouse, M.Y., 1996, HYSEP: a computer program for streamflow hydrograph separation and analysis: U.S. Geological Survey Water-Resources Investigations Report 96-4040, 46 p.

U.S. Environmental Protection Agency, 1986, Technical guidance manual for performing wasteload allocation—Book 6, Design conditions, Chapter I Stream design flow for steady-state modeling: Washington, DC, U.S. Environmental Protection Agency Report EPA 440/4-86-014, 64 p.

U.S. Geological Survey, 2002, NWISWeb: New site for the Nation's water data: U.S. Geological Survey Fact Sheet FS-128-02, 2 p.

U.S. Geological Survey, 2003, User's manual for the National Water Information System of the U.S. Geological Survey—Automated Data Processing System (ADAPS): U.S. Geological Survey Open-File Report 2003-123, 413 p.

U.S. Geological Survey, 2008, NWISWeb data for the Nation, accessed February 12, 2008, at *http://waterdata.usgs.gov/nwis/*

Vogel, R.M., Rudolph, B.E., and Hooper, R.P., 2005, Probabilistic behavior of water-quality loads: Journal of Environmental Engineering, v. 131, no. 7, p. 1081–1089.

Vogel, R.M., and Stedinger, J.R., 1985, Minimum variance streamflow record augmentation procedures: Water Resources Research, v. 21, no. 5, p. 715–723.

Vogel, R.M., Stedinger, J.R., and Hooper, R.P., 2003, Discharge indices for water-quality loads: Water Resources Research, v. 39, no. 10, p. 1273–1281.

Wahl, K.L., Thomas, W.O., Jr., and Hirsch, R.M., 1995, Stream-gaging program of the U.S. Geological Survey: U.S. Geological Survey Circular 1123, 22 p.

Warn, A.E., and Brew, J.S., 1980, Mass balance: Water Research, v. 14, p. 1427–1434.

Wilcox, R.R., 2001, Modern insights about Pearson's correlation and least-squares regression: International Journal of Selection and Assessment, v. 9, no. 1-2, p. 195–205.

This page intentionally left blank.