USGS
science for a changing world

Regionalization of Surface-Water Statistics Using Multiple Linear Regression

Chapter 12 of
Section A, Statistical Analysis, of
Book 4, Hydrologic Analysis and Interpretation

Techniques and Methods 4–A12
Version 1.1, February 2021

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

Regionalization of Surface-Water Statistics Using Multiple Linear Regression

By William H. Farmer, Julie E. Kiang, Toby D. Feaster, and Ken Eng

Chapter 12 of
Section A, Statistical Analysis of
Book 4, Hydrologic Analysis and Interpretation

Techniques and Methods 4–A12
Version 1.1, February 2021

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

U.S. Geological Survey, Reston, Virginia
First release: 2019
Revised: February 2021 (ver. 1.1)

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov.

Suggested citation:
Farmer, W.H., Kiang, J.E., Feaster, T.D., and Eng, K., 2019, Regionalization of surface-water statistics using multiple linear regression (ver. 1.1, February 2021): U.S. Geological Survey Techniques and Methods, book 4, chap. A12, 40 p., https://doi.org/10.3133/tm4A12.

ISSN 2328-7055 (online)

Contents

Figures

Table

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
	Length	
inch (in.)	2.54	centimeter (cm)
	Area	
square mile (mi^2)	2.590	square kilometer (km^2)
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

International System of Units to U.S. customary units

Multiply	By	To obtain
	Length	
centimeter (cm)	0.3937	inch (in.)
	Area	
square kilometer (km^2)	0.3861	square mile (mi^2)

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as $°F = (1.8 \times °C) + 32$.

Abbreviations

B.L.U.E.	Best Linear Unbiased Estimator
GLS	generalized least squares
MSE	mean squared error
OLS	ordinary least squares
USGS	U.S. Geological Survey
WLS	weighted least squares
WREG	weighted-multiple-linear regression

Regionalization of Surface-Water Statistics Using Multiple Linear Regression

By William H. Farmer, Julie E. Kiang, Toby D. Feaster, and Ken Eng

Abstract

This report serves as a reference document in support of the regionalization of surface-water statistics using multiple linear regression. Streamflow statistics are quantitative characterizations of hydrology and are often derived from observed streamflow records. In the absence of observed streamflow records, as at unmonitored or ungaged locations, other techniques are required. Multiple linear regression is one tool that is widely used to regionalize or transfer information from gaged to ungaged locations. This report provides the background to support regression-based regionalization of streamflow statistics. This background includes tools for data assembly, exploratory data analysis, model estimation in a least-squares framework, and model evaluation.

Introduction

Typically derived from streamflow records, streamflow statistics are quantitative characterizations of hydrologic phenomena at point locations along stream networks and in contributing areas. Engineers, planners, and regulators commonly use streamflow statistics to inform a wide array of projects, including design of water resource systems (for example, water treatment facilities), design of flood control projects, design of infrastructure (for example, bridges, culverts) and risk assessments in many different hydrologic contexts.

Streamflow statistics can be computed directly from observed, at-site streamflow records when sufficiently long records are available for streamflow-gaging stations (commonly abbreviated as streamgages). However, as shown by Kiang and others (2013), such data are not always available. For example, the historical record of streamflow may be short, resulting in poor estimates, or locations may be completely ungaged. However, it is often possible to leverage regional hydrologic information to supplement or improve limited at-site information.

Multiple linear regression is one technique commonly used by the U.S. Geological Survey (USGS) to estimate streamflow statistics using regional information. Hydrologic characterizations using multiple linear regression for regional interpolation are often referred to as "regionalization studies" because they use regional information to provide approximate values for hydrologic variables at ungaged locations. This report describes multiple linear regression as used by the USGS to estimate streamflow statistics. Although other techniques for regionalization of surface-water statistics exist, they are not as commonly used within the USGS and are only mentioned in this report.

What is Regression?

Regression is a statistical method for describing and quantifying the observed relationships and dependent variability between two or more variables. Regression can be used to predict the value of one variable, referred to as the "response variable," based on the value of one or more other variables, referred to as the "explanatory variables." The terms "response" and "explanatory" describe the variable relationships; variations of these names are described in the "Streamflow Statistics as Response Variables" section of this report. (Useful definitions of common terms can be found in appendix 1.) The estimates derived from regression analysis are approximations of the conditional mean of the response variable given the fixed quantities of the explanatory variables. For example, a regression of the annual maximum streamflow with an annual exceedance probability of 1 percent predicts the mean value of the annual maximum streamflow for a given set of explanatory variables. Regression analysis can also be used to describe the partial effects of explanatory variables on the response variable. More commonly, especially for surface-water problems, a regression model is typically developed so that the response variable, which may otherwise be hard to obtain, can be approximated from more readily available information. In this way, regional hydrologic information can be used to characterize intermittently gaged or ungaged watersheds.

Multiple linear regression is used to describe the linear dependence of a response variable on a set of explanatory variables. The process is termed "multiple" because more than one explanatory variable is used and "linear" because it is assumed that the relationships between the response and explanatory variables can be approximated by a straight line. However, as will be discussed in the "Model Estimation" section of this report, linearity can be achieved through variable transformation.

There are numerous mathematical methods to fit a straight line to a set of observations, each observation being a coupled realization of the response variable and the explanatory variables. These methods can be used, for example, to minimize the greatest absolute deviation or minimize the sum of the deviations of observations from the fitted line. One of the most frequently used methods for linear fitting, largely because of its strong theoretical underpinnings, is least-squares regression. This approach seeks to minimize the sum of the squared deviations from the fitted line. Least-squares regression is the method used most widely in the USGS and may be the most widely used in general.

There are several variations of least-squares regression that differ in how each observation is weighted in the analysis. The USGS commonly uses three approaches for surface-water applications: ordinary least-squares (OLS) regression, weighted least-squares (WLS) regression, and generalized least-squares (GLS) regression. The simplest form of least squares, OLS, weights all observations equally. WLS goes beyond OLS by assigning larger or smaller weights to different observations. One such approach is to assign weights based on the length of the streamflow records used to compute each streamflow statistic; this approach assumes that longer periods of record tend to produce more accurate estimates of the response variable in question. Finally, GLS also assigns weights based on the mutual dependence of the observations. As discussed in the "Variable Selection" section of this report, codependence in streamflow characteristics can arise from spatial proximity of streamgages (for example, streamgages upstream or downstream from other ones).

OLS, WLS, and GLS are all standard regression techniques that are used by the USGS for regionalization studies. Researchers have developed specialized methods of assigning weights that work well for the specific task of estimating streamflow statistics using regression. The remainder of this report describes the basic components of a regression analysis, as well as the specialized methods for surface-water datasets. Although the general approach to regression analysis described in this report can be applied to other types of data (for example, indicators of water quality or sediment concentrations), techniques for assigning weights that have been tailored for estimation of streamflow statistics may not be applicable to other problems without modification.

There are many computer programs that can be used to develop regression models. This report does not provide information on the use of any particular program. Rather, the report provides information on how to set up a regression-based regionalization analysis, choose an appropriate model, and interpret the results. The USGS has developed computer programs that can assist in regression analysis. The weighted-multiple-linear regression (WREG) program was tailor-made for use with hydrologic variables and can be used to set up and evaluate OLS, WLS, and GLS regressions (Eng and others, 2009). In addition to WREG, a basic computer program for statistical analysis will be useful to complete the initial steps of a regression-based regionalization study.

Framework for a Regression-Based Regionalization Study

The iterative process of a regression-based regionalization study is represented in figure 1. The process can be divided into three major tasks: assembling data; model development, including conducting exploratory data analysis, model estimation, and model evaluation; and model application and documentation. Assembling input data (described in the "Data Assembly" section) may require using geographic information systems, as well as methods for estimating streamflow statistics, and may take a considerable amount of time and resources. Model development involves the selection and prioritization of explanatory variables, estimating coefficients, and evaluating predictions. Exploratory data analysis (described in the "Exploratory Data Analysis" section) is an important part of a regression study because familiarity with the data can guide decisions when developing and interpreting the model. The principles of model estimation and model evaluation (the most iterative step) are described in the "Model Estimation" and "Model Evaluation" sections. Model evaluation sometimes instigates further data analysis and estimation. The first attempt to

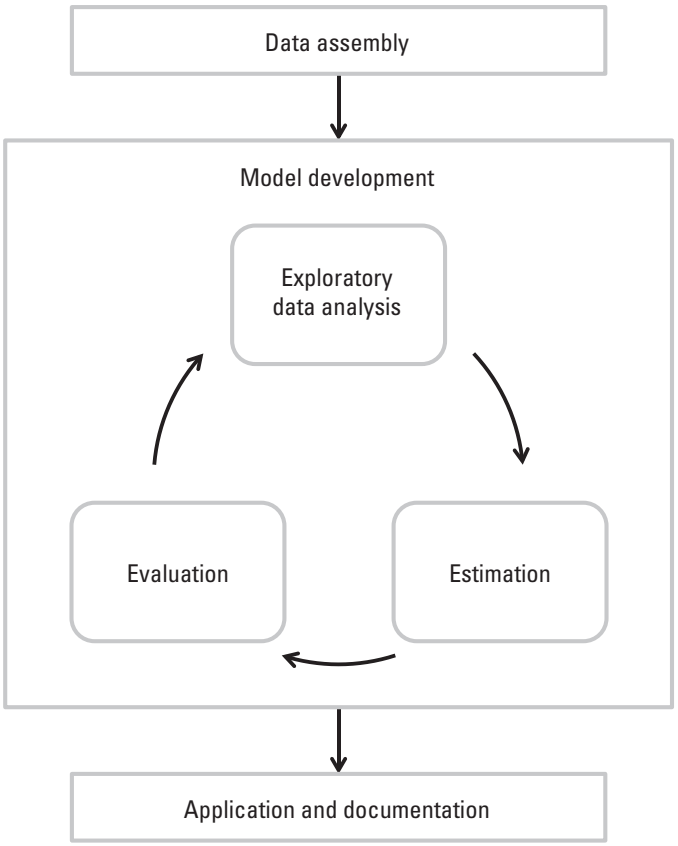

Figure 1. Diagram showing a generalized work flow for a regression-based regionalization study.

develop a regression model rarely results in the best or final model. Careful analysis of the model results and subsequent refinement of the model can result in improved predictive power of the regression model. Finally, model application and documentation are described in the "Model Application and Documentation" section. The guidance provided in this report is intended to ensure that the process used to develop the final regression model is fully documented with sufficient detail to be easily reproduced.

Description of an Example Dataset

The examples in this report are drawn from observed streamflow data and the GAGES-II dataset of geospatial attributes for evaluating streamflow (Falcone, 2011). Daily streamflow information was drawn from 300 sites across the continental United States with 60 years of complete daily records from October 10, 1956, to September 30, 2016. For each site, the annual maximum daily streamflow was selected. From this record of maximums, the 90th percentile was taken as representative of the 10-year event and used as the response variable. For regression, the GAGES-II dataset (Falcone, 2011) provides several possible explanatory variables: drainage area (DRAIN_SQKM), basin-averaged precipitation (PPTAVG_BASIN), basin-averaged temperature (T_AVG_BASIN), at-site temperature (T_AVG_SITE), basin-averaged relative humidity (RH_BASIN), median relief ratio (RRMEDIAN), and average March precipitation (MAR_PPT7100_CM). The example regression is based on a logarithmic transformation of the response variable and a logarithmic transformation of drainage area. Data for this example can be obtained from Farmer (2019).

Data Assembly

The assembly of input data is the first step of a regression-based regionalization study. This process includes the identification of suitable study areas, sites, and regions, as well as the collection and processing of requisite data. Related to this selection process, and often done in tandem, is the identification of the streamflow statistics that will be estimated as response variables in the region. Additionally, suitable descriptors of the watersheds and the climate in the region are needed for use as explanatory variables. At times, both explanatory and response variables are already computed. However, there are times when the response variable needs to be derived by methods beyond the scope of this report (for example, the computation of flood-frequency statistics is described in Bulletin 17C (England and others, 2018) or considerable effort is required to determine potential explanatory variables. In addition to detailing the terminology used to describe these processes, the following subsections outline the appropriate concerns for the assembly of input data for a regression-based regionalization study.

Region Definition and Site Selection

Regression models are defined for a specified domain: spatial, temporal, physiographic, or some combination thereof. When considering the spatial domain, it is ideal to start with the largest reasonable region. However, it is always possible to consider subregionalization to further refine the domain. In addition to spatial domains, regression analyses often are tied to specific temporal domains (for example, the period of record over which observations were made). A conventional goal is to define regions to minimize the degree of unexplained variability in the response variable or to minimize or eliminate bias in regression estimates. After a region is defined, it is not advisable to extrapolate beyond the region by using regression models outside the target spatial, temporal, and physiographic bounds. Extrapolating the models outside the defined bounds will require additional analysis.

After a region is defined, the sites and streamgages from which to compile data on the response and explanatory variables are selected. The purpose of the regression model should be kept in mind during the site and streamgage selection process. For example, regression models are most commonly used to estimate natural streamflow conditions. In such cases, there is an implicit assumption that the data used to develop the regression equation are representative of natural systems. Alterations to natural streamflow resulting from human modifications, such as reservoirs, withdrawals, discharges, or other nonnatural alterations, can invalidate this assumption. Consequently, regressions used to estimate natural streamflow conditions should rely only on streamgages that are not substantially affected by human modifications in the watershed. Although the definition of "substantial effects" may be somewhat subjective, every effort should be made to present a justifiable defense.

Conducting a regression-based regionalization study that includes altered watersheds is not appropriate without attempting to account for those alterations in the regression. For example, if watersheds with substantial urbanization are included, information on the degree of urbanization may be captured in land-use information, population-density information, or other variables. Although the streamflow statistics predicted from such models are not derived from purely natural processes, the explanatory variables provide a means of describing and accounting for the degree to which changes to the landscape affect hydrological processes in the basin. However, it is not entirely accurate to identify streamgages as categorically altered or unaltered. Within the realm of altered watersheds there may exist a wide range of types or degrees of alteration, producing variability that may not be accounted for by, or appropriate in, a single regression model.

One of the primary concerns in the process of site selection for regression-based regionalization studies is the length of record. Streamflow statistics can be interpreted as representative only of the period of record used to compute the statistic. In this case, use of a common base period for all sites within a region may be appropriate, such as the mean flow for

the period 1951–2000. More commonly, computed streamflow statistics are intended to be indicative of long-term conditions, and sites selected for use in regionalization studies have different record lengths.

Natural variability in streamflow, often driven by multidecadal wet and dry shifts in climate (for example, Mauget, 2003; McCabe and Wolock, 2014), using different periods of record can result in time-sampling errors. As such, the length of record used to compute streamflow statistics can have a direct effect on the accuracy of those statistics. Closely related to the length of record, time-sampling errors tend to be larger for streamflow statistics describing rare or unusual conditions than for streamflow statistics describing commonplace conditions. That is, although streamflow statistics describing rare events may differ considerably when different periods and lengths of record are used, streamflow statistics describing commonplace events generally will be less variable. As an example, the long-term mean annual streamflow can be more reliably estimated from a shorter record length than the instantaneous-peak streamflow that has a 1-percent annual probability of exceedance. Whatever the streamflow statistic, a long period of record is desirable as long-term persistence, as well as episodic conditions, may significantly affect the ultimate statistic.

The USGS generally advises that at least 10 years of data be available to fit a frequency curve and estimate recurrence-interval statistics (Riggs, 1972; U.S. Interagency Advisory Committee on Water Data, 1982). Other authors have suggested a minimum length of 15 years (Kennard and others, 2010), though the appropriate length largely depends on the statistic of interest. In general, 10–15 years of data should be considered a minimum record length for calculation of streamflow statistics used in regression-based regionalization studies in which the goal is to estimate streamflow statistics representative of long-term conditions. For some statistics (for example, regional skews), a period of 25 years may be more appropriate. If data are available at enough streamgages, a longer minimum record length can be specified for a particular study. In any case, a defensible justification and documentation of the minimum record length should be included in any regression-based regionalization study.

Commonly, streamflow characteristics are calculated from either (1) the longest available period of homogeneous record for each streamgage, or (2) a fixed range in time for all streamgages. When using the longest homogeneous period of record at each streamgage, the maximum information available at each streamgage is used, which minimizes the sampling error of the streamflow estimate at each streamgage. However, calculating statistics using different time periods across sites and different lengths of record can introduce additional noise into a regression model that is difficult to address appropriately. For example, consider changing climates or land uses over the period of record. Errors in the fit of the regression line may result from these climatic shifts or land-use changes and not from differences in the explanatory variables, as presumed by the regression model. Use of a uniform period of record may help to minimize some of these issues, but time-sampling errors will be larger than necessary for streamgages where records are short. Furthermore, using the same period of record for all stations may artificially increase the cross correlations among the streamgages, effectively limiting information content.

Finally, the length and period of record also may dictate the basin characteristics and climate attributes used as explanatory variables. Explanatory variables used in the regression analysis should be representative of basin conditions during the period of record used to derive the streamflow characteristics used as response variables. Failure to consider the period of record may lead to use of less than optimal explanatory variables and substantially weaken model performance or inadvertently misguide inferential analysis.

Record Extension and Augmentation

Although every effort should be made to identify sites with a sufficient length of record, specialized techniques have been developed for short-term and partial-record streamgages. Although definitions differ, short-term, continuous-record stations are those with less than approximately 10–15 years of record. Because statistics calculated from shorter records generally are not as accurate as those calculated at stations with longer records, time-sampling errors associated with short-term streamflow records can result in substantial biases. Additional statistics can sometimes be estimated for sites with short-term records by “augmenting” the record with information from a long-term, continuous-record site. Examples of techniques that have been used to augment streamflow records include the maintenance-of-variance extension (Hirsch, 1982) and base-flow correlation techniques (Stedinger and Thomas, 1985).

Noncontinuous streamflow records are available at crest-stage streamgages and at low-streamflow, partial-record streamgages. A crest-stage streamgage is designed to record stages of all peaks above a set base elevation at the streamgaging location, including annual peaks. Provided the crest-stage streamgage is adequately maintained, it is appropriate to treat peak-streamflow records from a crest-stage streamgage similarly to a continuous-record streamgage for regional regression. However, at partial-record streamgages at sites with low streamflow, measurements are made somewhat sporadically during low-flow periods, and these measurements sometimes do not correspond with the timing of annual minimum flows. Similar to short-term, continuous-record stations, information from a long-term, continuous-record station can be used to augment streamflow records at partial-record stations (Tasker, 1975).

Streamflow statistics for short-term, continuous-record streamgages or partial-record streamgages can be augmented with estimates derived from record augmentation techniques. These estimates can be used in a regression study if it can be demonstrated that they extend the range of the regression model by extending or filling in gaps in the range of response and explanatory variables. Estimates from augmented records should not be used simply because they exist, as estimates

can degrade the accuracy of a regression model if they are not properly weighted to account for cross correlations among streamgage statistics. Current methods for assigning weights and estimating cross correlations are imperfect. Vogel and Kroll (1991) provide a discussion of the advantages and disadvantages of record augmentation with respect to cross correlations. In the end, this decision is highly subjective and extreme caution is advised.

Nested Basins

Another complication for site selection is the prevalence of nested basins. Streamgages on the same stream or river system, where a smaller drainage is completely contained, or nested, within a larger drainage, can contain redundant information. This redundancy can negatively affect the regression analysis, and so can be considered when selecting streamgages for inclusion in the regression analysis. If two streamgages are immediately upstream or downstream from one another, drain comparable basin areas, and include a similar period of record, they should not both be included in the development of a regression model. The farther apart the streamgages are, the more dissimilar the basin attributes and streamflow are likely to become. A headwater streamgage may be quite different from a main-stem streamgage in terms of both streamflow and basin attributes, so both could be included in development of a regression model. Simple metrics, such as used by Parrett and others (2011), can be used to help screen for possibly redundant sites.

Some judgment is required in identifying closely related, redundant basins. One commonly used criterion is to include streamgages on the same river system only if their drainage areas differ by at least a factor of two (Sauer, 1974). That is, the size of the smaller basin would be no more than 50 percent of size of the larger basin. More stringent criteria are also reasonable, but less stringent criteria are not suggested (Sauer, 1974; McCuen and Levy, 2000). The simplest method for deciding which streamgage to retain in the analysis is to choose the one with the longest period of record based on the assumption that the data from the station with the shorter record is mostly concurrent with the station with the longer record. In addition, it may be useful to consider which streamgage best extends the range of basin attributes used in the analysis. If the records are not concurrent, the determination to use all available records has been made, and there are no other concerns, both stations should be included in the analysis. However, the errors in the resulting regression should always be evaluated for correlation.

Streamflow Statistics as Response Variables

Regression models are mathematical descriptions of the relationships between a response variable and one or more explanatory variables. The term “response variable” is used because the variable responds to changes in other variables but other names are also used. The response variable is also known as a dependent variable because its magnitude depends on the magnitude of other variates. Because it is often displayed on the vertical or Y axis of a bivariate plot, the response variable is also known as the Y variable. Finally, the response variable, because it is the predicted quantity, is also known as the predictand. Although all terms are valid, the variable is known as the response variable in this report to maintain consistency.

In a regression-based, surface-water regionalization study, the response variable of interest is typically a streamflow statistic. Examples of statistics that a hydrologist may be interested in estimating from a regression model include the mean annual streamflow, the flood that has a 1-percent annual exceedance probability, and an annual low streamflow of specified frequency of occurrence. Streamflow-duration statistics or percentiles are other commonly estimated streamflow statistics. The desired statistics will be calculated for the available streamgages in the region of interest prior to building the regression model. Specific methods for calculating streamflow statistics are not discussed in this report but the selected methods are an important element of a regression study, as the quality of the input data will affect the quality of the regression estimates. Consequently, every effort should be made to ensure that streamflow statistics are calculated in an accurate and consistent manner and “suspect data” should be reviewed and removed if warranted. When data are removed from an analysis, the reason for the removal should be documented in accordance with standard policies of quality assurance and control.

The USGS has been conducting regionalization studies using regression techniques for many years. Benson and Carter (1973) summarize studies prior to publication of their report. They state that the desired accuracy of a regression model is the degree of uncertainty obtained when calculating that statistic from a 10-year record. In general, this goal was most often met in the eastern United States for mean annual streamflow, mean monthly streamflow, and 50-year flood. Regression estimates of the mean annual streamflow had the lowest uncertainty. Benson and Carter (1973) found that low streamflows are most difficult to estimate, and none of the studies available at the time met the criteria of producing estimates with uncertainties equivalent to those obtained using 10 years of observed streamflow record. They attributed the difficulty to an inability to adequately describe aquifer characteristics critical to low streamflows with regionally available data.

Basin Attributes as Explanatory Variables

Similar to response variables, explanatory variables are known by several different names. They are explanatory because, in the context of the regression model being developed, they explain the variability in a response variable. In the context of regression, explanatory variables do not depend on the variability of a response variable, so they are called independent variables. Because they are customarily plotted on the horizontal or X axis of bivariate plots, explanatory variables can be denoted as X variables. Because they

produce a prediction of the response variable, they can be called predictors. Finally, because they are regressed upon to produce estimates of the response variable, they can be called regressors. In this report, for consistency, they are known as explanatory variables.

The purpose of a regression-based, surface-water regionalization study is to predict the value of a streamflow statistic from readily available data that describe the watershed. In this way, these watershed characteristics, or basin attributes, act as explanatory variables. These may include variables such as drainage area, mean annual precipitation, or a variable that characterizes the underlying soils or geology. Although regressions are purely statistical, every effort should be made to consider explanatory variables that have a plausible hydrologic linkage to the response variable. Rarely are hydrologic processes governed by linear mechanistic processes, but the inclusion of explanatory variables with plausible hydrologic linkages aids model interpretation.

Calculating the values of explanatory variables typically relies heavily on the use of geographic information systems and previously published data for the region. Specific methods for calculating the values of explanatory variables are not described in this report. For use in a regression model, explanatory variables should be relatively easy to calculate for an ungaged location. In addition, it is preferable to use explanatory variables that can be calculated with relatively small errors. That is, explanatory variables that are difficult to measure or have high uncertainty are less desirable.

The least-squares-regression methods described in this report assume that the explanatory variables are measured without error. Errors can affect the accuracy of the regression model and estimates of its uncertainty. Issues can include biased estimates of model coefficients (Allison, 1999, p. 55). Draper and Smith (1981, p. 7 and 124) suggest that, although the assumption that there is no error in explanatory variables is rarely met in practice, adequate models can be developed if the errors are small compared to the range of observed values of the explanatory variable. If this criterion cannot be met, the variable should not be used in a least-squares-regression model. A discussion of ways to manage uncertainty in explanatory variables is beyond the scope of this report.

Many basin attributes have the potential to influence streamflow and can be considered as explanatory variables. For surface-water streamflow statistics, explanatory variables can generally be classified into a few broad categories: (1) basin geometry, including drainage area, slope, and elevation; (2) climatological, including precipitation, temperature and potential evapotranspiration; (3) land-cover descriptors, including land use, soils, and geology; and (4) available storage, including quantifications of the coverage and presence of lakes, ponds, and other open water. Many permutations of these variables can be developed, but it is generally not necessary to test every possible variable. Selecting one or two variables associated with a hydrologic process should be sufficient to determine if that class of variable is worth further consideration. Report authors should consider providing theorized descriptions of how their selected explanatory variables are expected to be related to variations in the regionalized streamflow statistics.

Other descriptors of streamflow, such as the variability in streamflow or the average rate of recession, are sometimes used. Even though these can be powerful indicators and may be easily calculated at streamgages, they are difficult to estimate at ungaged locations, so their applicability is limited. As such, using streamflow-derived variates as explanatory variables is discouraged. If such variables are considered, the effects of uncertainty in the estimated value of the variable at ungaged locations should also be evaluated, quantified, and incorporated using advanced techniques.

Benson and Carter (1973) compiled information from previous USGS studies on basin characteristics most commonly used in regression-based regionalization studies. They found that drainage area was uniformly the most widely used variable, regardless of whether mean annual streamflows or flood streamflows were being predicted. Measures of precipitation, land cover, and elevation or slope were also widely used. In northern areas of the United States, a measure of snowfall was an important explanatory variable. The region of interest, underlying processes, and selected response variable are important considerations in selecting appropriate explanatory variables.

Kiang and others (2013) summarized basin characteristics that were used in models that have been entered in the USGS National Streamflow Statistics Program (http://water.usgs.gov/software/NSS/). Models were summarized for peak streamflows (1-percent and 10-percent exceedance probability flows), low streamflows (7Q10—7-day minimum flow that happens on average only once every 10 years), and average streamflows (mean and median annual streamflows). Although peak-streamflow models were available for nearly all states in the United States, low streamflow and average streamflow models were available for only about one-third of the states. Consistent with Benson and Carter (1973), Kiang and others (2013) found that drainage area was nearly always included as an explanatory variable. Measures of precipitation, elevation, and slope were also commonly included as significant variables for all types of streamflow. Less commonly used, but also appearing in many models, were variables describing the soils in the basin, land cover, and the amount of water storage available in the basin. Other variables included measures of the basin shape, indicators of geology, and streamflow indices. When considering which variables to use for regression models, it may be useful to explore previously developed equations found in the National Streamflow Statistics Program and related publications.

Exploratory Data Analysis

After assembling the relevant data, candidate variables can be qualitatively evaluated for use as explanatory variables in a regression model. This process, exploratory data analysis, helps to develop a basic understanding of how the variables relate to one another and to the response variable. Multiple linear

regression techniques require that the relationship between response and explanatory variables be linear. Sometimes a transformation of the response variable, explanatory variables, or both will help to linearize the relationships; the need for such a transformation will be made apparent through the intuition-developing process of exploratory data analysis. Further, exploratory data analysis will help identify dependencies within the explanatory variables. Dependencies, which should be avoided, are discussed in the "Variable Selection" section.

Before starting the exploratory data analysis, it is important to consider possible errors in the candidate explanatory variables. Hopefully, quality assurance in the data assembly phase addressed any major concerns, but additional checks may be required. Viewing data graphically can help to identify errors in the data. For example, a plot of the annual time series used to calculate streamflow statistics may reveal an unusually high or low value, often called an outlier, or a sudden or gradual change in the magnitude of the streamflow over time. Any data that do not conform to expectations should be checked to be sure there is not an error. However, data should not be removed without strong evidence of inaccuracy. Complete documentation of such conditions is required.

Exploratory data analysis begins with an assessment of the response variable dependence on each explanatory variable and the dependence among explanatory variables. Unlike least-squares regression, this exercise is typically a visual and qualitative assessment. One is seeking to better understand the variability in the datasets and the potential usefulness thereof. Quantitative assessment follows.

First, it is useful to consider the possible dependencies among explanatory variables. As demonstrated quantitatively in the "Variable Selection" section, the precision of multiple linear regression is sensitive to redundancies and strong dependencies in explanatory variables. A starting point of analysis is to plot each explanatory variable against each other explanatory variable and observe the relationships. Figure 2 displays the relationships among several explanatory variables for the example dataset. Ideally, candidate explanatory variables will not be correlated; instead, plots should show a nearly horizontal, smoothed relationship. In figure 2, there are several variable pairs that do not show the ideal relationship. Of most concern is the strong linear dependence between average temperature in the drainage basin and average temperature at the streamgage. Because of this strong dependence, the variables provide highly redundant information. Consequently, it is inappropriate to include both variables in the same regression analysis. In these types of situations, there is no standard guidance on which of the variables should be retained. Arguments of physical plausibility or ease of collection may prove useful, or it may be necessary to consider each in turn.

In addition to identifying troublesome dependencies, visualizations such as those in figure 2 begin to document the distributional behavior of each explanatory variable. Histograms of each explanatory variable, though not shown here, are also useful. The goal is to ensure that the explanatory variables present a distribution with reasonable spread or variability. Not using data that exhibit distinct clustering around specific values will increase the physiographic range of the resulting regression model, thereby improving the specification of regression coefficients when a least-squares method is applied.

After controlling for redundancy and dependency in explanatory variables, it is useful to visually and qualitatively consider the dependence of the response variable on each explanatory variable in turn. Multiple linear regression seeks to quantify the linear partial dependencies of the response variable on each explanatory variable. Consequently, one seeks roughly linear correspondence between the response variable and each candidate explanatory variable. Figure 3 shows example scatterplots of the response variable and several explanatory variables for the example dataset. Panel *A* shows a strong linear correspondence. Although transformation seems necessary, the drainage area may be a strong candidate for an explanatory variable in multiple linear regression. Panels *B* and *C* show less linear correspondence, suggesting that average March precipitation and the median relief ratio might be poor explanatory variables. In the example dataset, the median relief ratio had the least correlation with the selected variable of interest. Finally, panel *D* shows what might be a linear dependence. Closer inspection reveals substantial clumping; the perceived linearity is a result of visual attenuation produced by the points in the farthest right-hand part of the plot. For this reason, such an explanatory variable may not be a strong explanatory variable. Used only for illustrative purposes, panel *D* shows the same data as panel *C*, but artificial data points have been appended to the right side of the graph.

Using visualization to identify the potential linear dependencies of explanatory variables on the response variable can be difficult. Dependency of an exploratory variable on one or more of the other exploratory variables may obscure the dependency of that exploratory variable on the response variable. For this reason, some combinations of explanatory variables, if physically justifiable, may prove useful. Because drainage area is often selected as an explanatory variable for regressions dealing with streamflow statistics, it may be helpful to standardize the response variable as a function of the upstream drainage area, and then consider the correspondence with explanatory variables. After the effects of drainage area are removed, other relationships between the response and candidate explanatory variables may be more apparent. Figure 4 shows one such example with the example dataset. Panel *A* shows a weaker relationship when the response variable is not standardized. However, when the response variable is divided by the confounding explanatory variable not displayed (drainage area), a stronger correspondence is indicated (panel *B*).

If there is not a linear relationship between the response variable and an explanatory variable, mathematical transformations of either variable may improve linearity. For surface-water statistics of perennial streams, streamflow is a commonly used response variable that is often paired with upstream drainage area as an explanatory variable; these variables are typically log-transformed to achieve linearity. For intermittent or ephemeral streams, this practice is

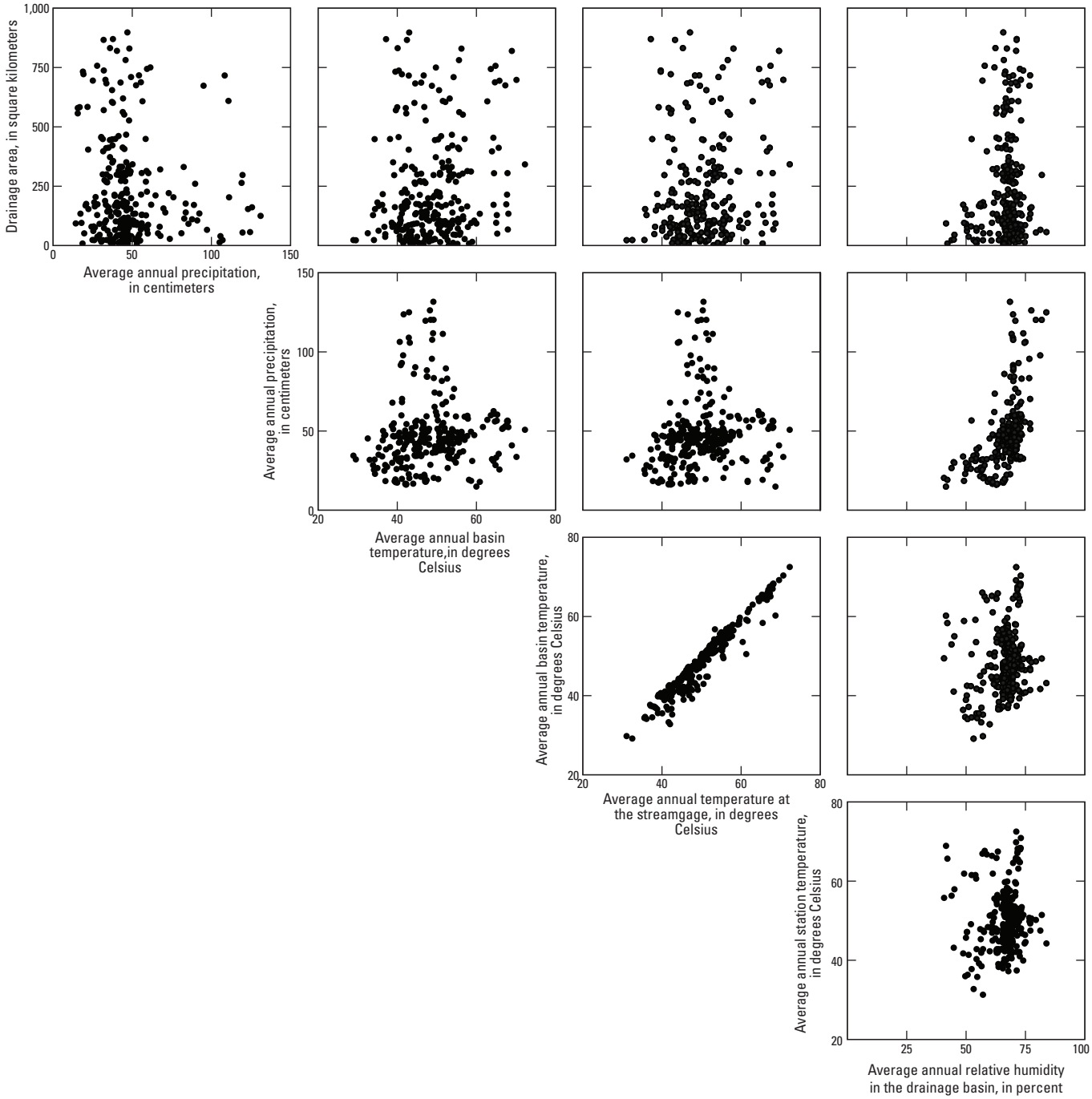

Figure 2. Correlation matrix plots comparing selected explanatory variables.

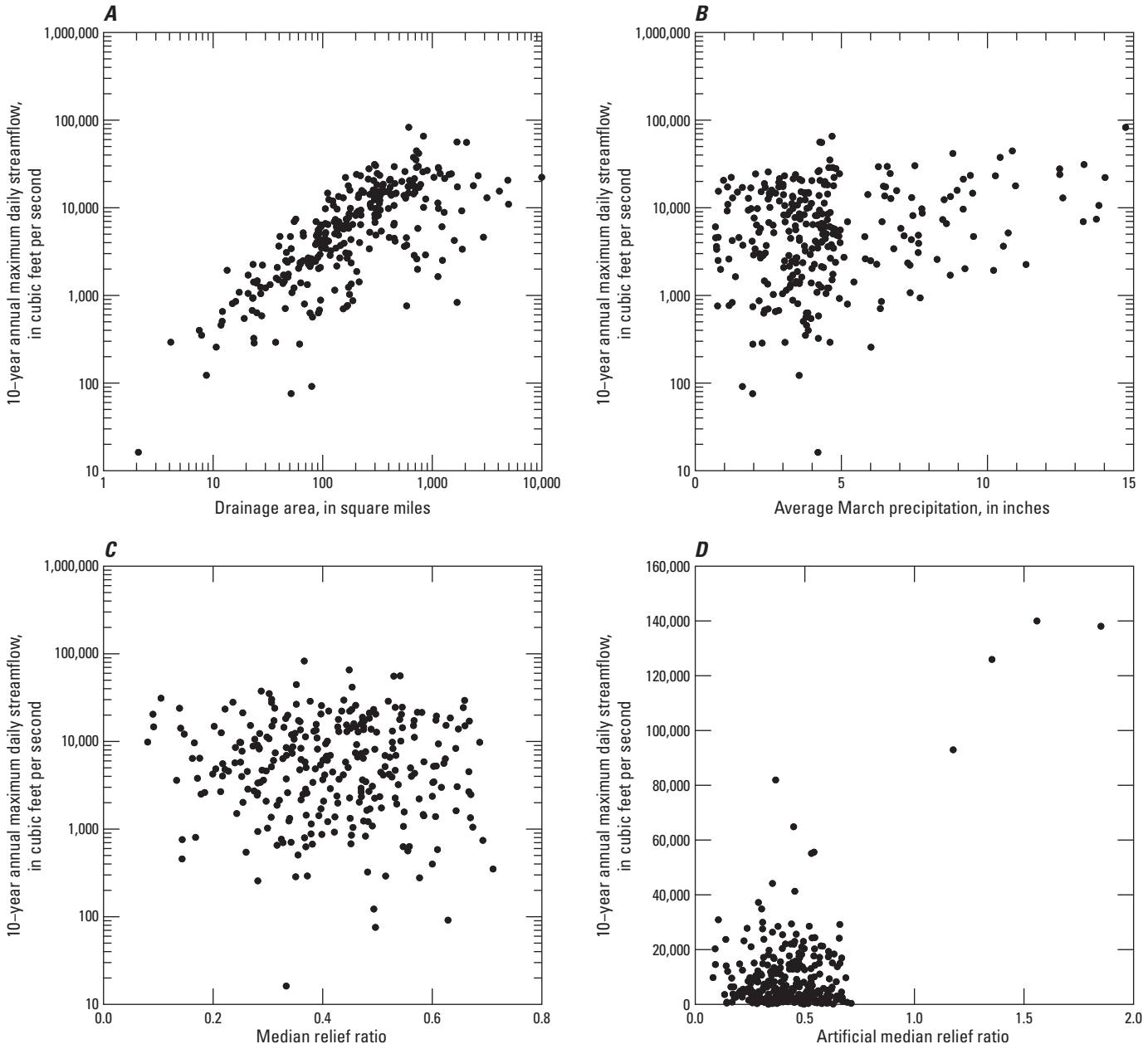

Figure 3. Scatterplots showing relationships between the response variable (10-year annual maximum daily streamflow) and selected explanatory variables. *A*, Scatterplot showing a strong linear relationship between the response variable and drainage area. *B*, Scatterplot showing a weaker linear relationship between the response variable and average March precipitation. *C*, Scatterplot showing the least correlation between the response variable and the median relief ratio. *D*, Scatterplot showing the same data as panel *C*, but with artificial data points appended to the right side of the graph.

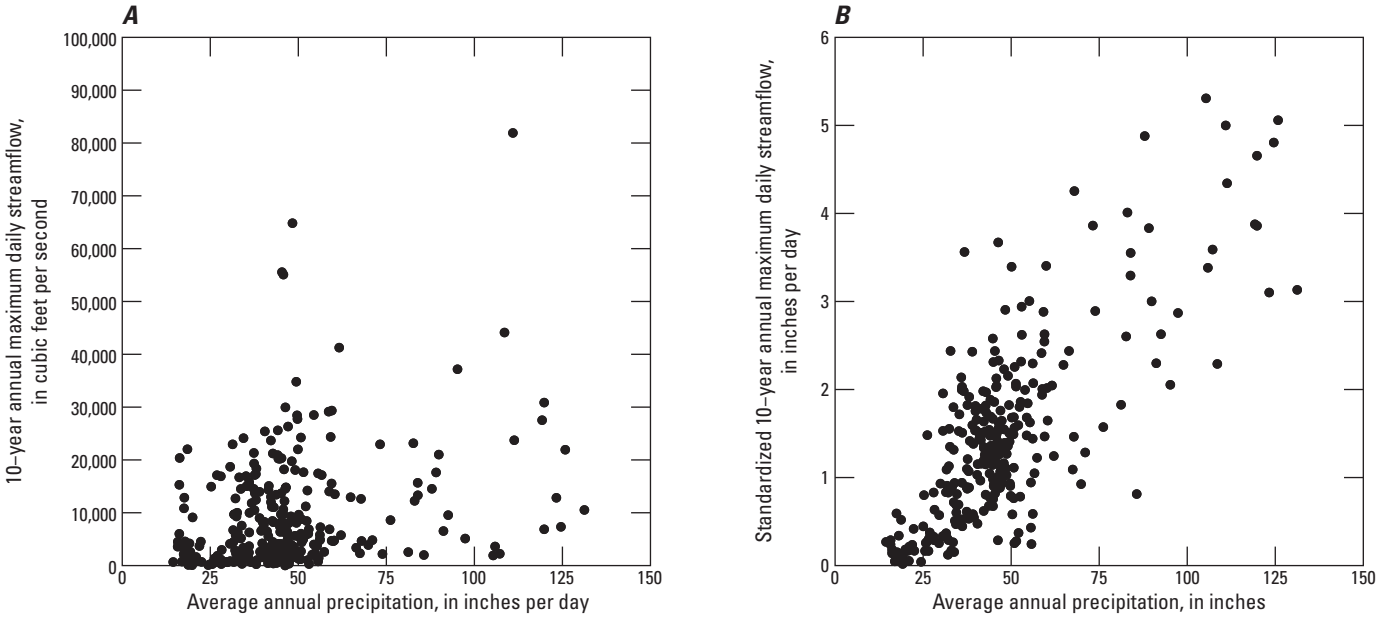

Figure 4. Scatterplots showing the relationships between the response variable (10-year annual maximum daily streamflow) and average annual precipitation. The plots show how the relationships can be improved by standardizing the response variable by another explanatory variable. In this case, the effects of drainage area have been removed. *A*, Scatterplot showing a weak relationship when the response variable is not standardized. *B*, Scatterplot showing the relationship after the response variable is standardized. The relationship is tighter in panel *B* than in panel *A*.

not suggested because the response variables will often equal zero. As an example of such a relationship, panel *A* of figure 5 shows the response variable (10-year annual maximum daily streamflow) plotted against an explanatory variable (drainage area). Although some correspondence is evident, the relationship between them is distinctly nonlinear. In panel *B*, both the explanatory and response variables are plotted on a logarithmic scale, which produces a more linear relationship. (Taking the logarithm of a variable achieves the same transformation as plotting on a logarithmic scale.) In addition to linearizing a relationship, a transformation can be useful in standardizing the spread of data points around the linear relationship.

Logarithmic transformation is only one of several transformation methods. Common tools for identifying potential variable transformations are Mosteller's bulging rule (Mosteller and Tukey, 1977, p. 84), Tukey's ladder of powers (Tukey, 1977, p. 89), and Box-Cox transformations (Box and Cox, 1964). These tools describe certain power transformations that can be used to straighten different degrees and angles of curvature. In addition to power transformations, there are several types of algebraic transformations.

When a response variable is transformed, the regression model predicts the transformed response variable rather than the untransformed response variable. To obtain an estimate of the untransformed response variable, the user will back transform the prediction or the regression model itself. This process of back transformation can result in substantial bias, although methods for handling this bias for common transformations have been documented. Transformations of the response variable should only be done when the transformed response variable is linearly related to all the explanatory variables in the model. Explanatory variables, however, can be transformed independently or not at all. In hydrology, logarithmic transformation is arguably the most common. The properties of logarithms, particularly as they pertain to variables whose domain includes zero or negative values, are considerations in any decision to perform a logarithmic transformation.

Multiple linear regression describes a regression plane or hyperplane (Montgomery and others, 2006), meaning that regression analysis quantifies the isolated influence of each explanatory variable in turn by holding constant all other variables. Consequently, a pairwise comparison of response and explanatory variables may fail to identify a strong linear dependency that may later prove influential through formal regression analysis. Although considering a few standard variable combinations and transformations is useful, it is far from exhaustive. For this reason, such exploratory data analysis can be viewed as a tool by which to increase familiarity with, and insight about, the dataset at hand. This increased familiarity aids in the iterative interpretation of candidate regression models. However, although exploratory data analysis can determine the adequacy of candidate explanatory variables, it cannot determine if a variable can be categorically excluded from subsequent regression analysis.

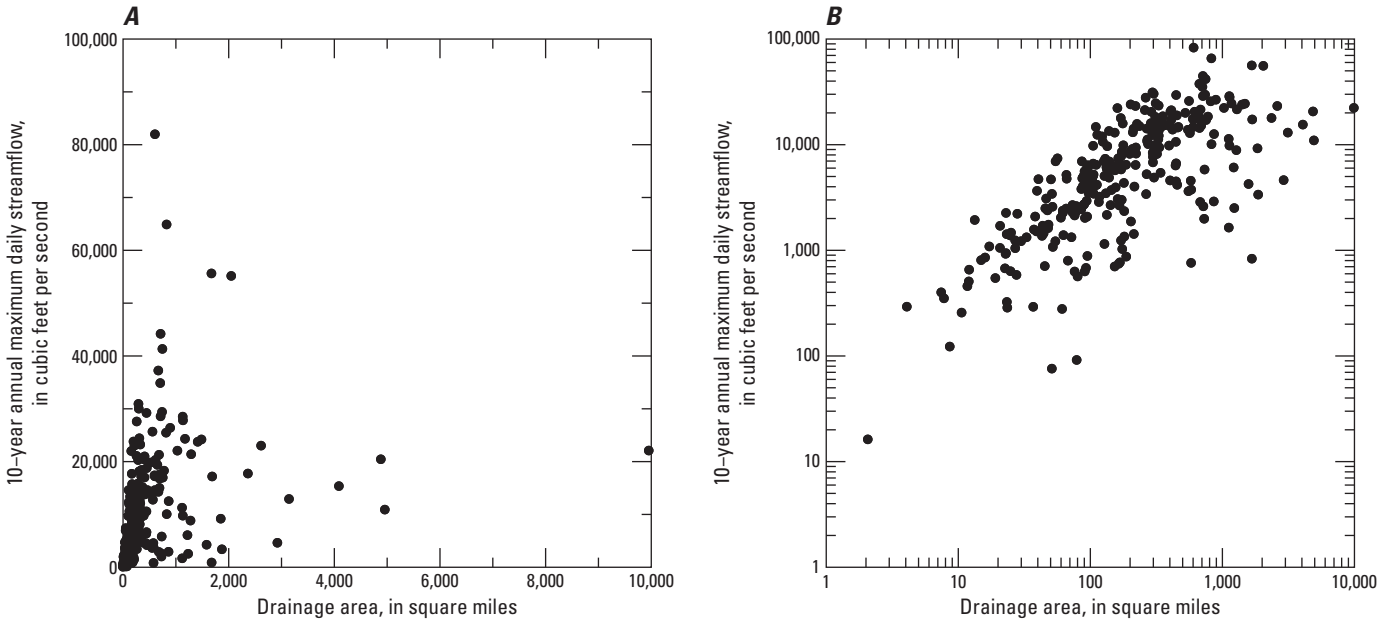

Figure 5. Scatterplots showing the effect of transforming response and explanatory variables. *A*, Scatterplot of the nonlinear relationship between untransformed explanatory and response variables. *B*, Scatterplot showing the improved linear relationship after logarithmic transformation of both variables.

Model Estimation

Formal regression analysis is required to quantify the relationships between the selected response variable and the explanatory variables. The following sections discuss the basic principles of regression, the principles of least-squares regression, and variations of least-squares regression. A glossary of symbols can be found in appendix 2.

Principles of Linear Regression

In linear regression, it is assumed that the response variable can be represented as a linear combination of the explanatory variables. This assumption is commonly expressed by representing the response variable Y and the explanatory variables as Xs such that

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_{[,1} + \beta_2 \boldsymbol{X}_{[,2} + \ldots + \beta_M \boldsymbol{X}_{[,\boldsymbol{M}} , \quad (1)$$

where

M is an arbitrary number of explanatory variables; and

β are unobserved linear coefficients of the underlying, true model.

If equation 1 is a valid underlying model, then the regression coefficients can be estimated by several different methods. Regardless of the calculation method, approximations are commonly represented by a hat to the relevant character (for example, β_0 is estimated as $\hat{\beta}_0$). Furthermore, with the approximation of linear coefficients, the estimated linear model deviates from the true underlying model, resulting in some residual error, ε. The reformulation of the linear model, after regression estimation, is therefore

$$\boldsymbol{Y} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{X}_{[,1} + \hat{\beta}_2 \boldsymbol{X}_{[,2} + \ldots + \hat{\beta}_M \boldsymbol{X}_{[,\boldsymbol{M}} + \boldsymbol{\varepsilon} . \quad (2)$$

In the application of equation 2, the model residual is unknown and the resulting prediction of response variable, $\hat{Y}$, is given as

$$\hat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{X}_{[,1} + \hat{\beta}_2 \boldsymbol{X}_{[,2} + \ldots + \hat{\beta}_M \boldsymbol{X}_{[,\boldsymbol{M}} . \quad (3)$$

A graphical representation of the observed response, the fitted model, and model residuals for a bivariate case of the 10-year annual maximum streamflow and drainage area are shown in figure 6.

An analysis of residuals is the focus of most tools for regression analysis. In model development, the residuals are defined as the difference between the observed and estimated values of the Y variable:

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}. \quad (4)$$

Because streamflow statistics are, themselves, estimates of unobserved values (for example, Y is an estimate of the unknown true value, $\tilde{Y}$), they are subject to some degree of error. Unless a statistic is defined as being for the specific time period that was used to compute the statistic, it is impossible to know the true value of a streamflow statistic without an

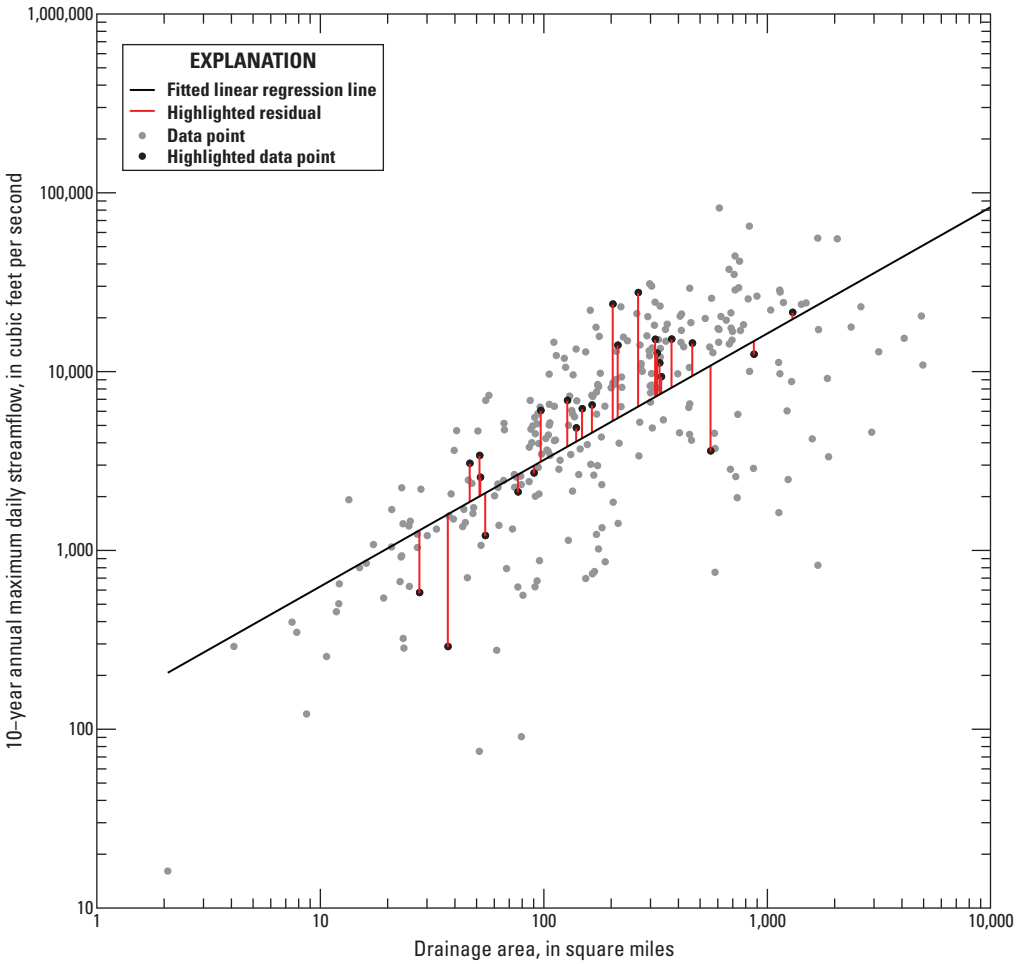

Figure 6. Graph showing the response variable (10-year annual maximum daily streamflow) and explanatory variable (drainage area), whose observations are represented in the scatterplot, linearly regressed against each other to produce the dashed line in this bivariate example. The red line segments represent residuals of the regression.

infinitely long streamflow record. Variability in the streamflow record introduces time-sampling error into estimates calculated using a finite record length. Therefore, because finite samples are required for model development, the residual error in a regression model is composed of two types of errors: a model error, δ , and a sampling error of the streamflow statistic, η. The model error is the difference between the unknown, true value and the regression-estimated value:

$$\boldsymbol{\delta} = \tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}, \tag{5}$$

but the sampling error, η, is the difference between the observed or computed value and the true value:

$$\boldsymbol{\eta} = \boldsymbol{Y} - \tilde{\boldsymbol{Y}}. \tag{6}$$

Summing equations 5 and 6, in conjunction with equation 4, yields

$$\boldsymbol{\varepsilon} = \boldsymbol{\delta} + \boldsymbol{\eta}. \tag{7}$$

Further analysis of this residual error is possible after assumptions about the method of model fitting have been made.

Least-Squares Regression

Least-squares regression is one of the most widely used techniques for estimating the regression model presented in equation 2. Other methods can be used (for example, minimizing mean deviation, minimizing the maximum absolute deviation), but least-squares regression provides a closed-form solution to minimize the sum of the squared residuals, SS_{ε}, across N observations, as

$$SS_{\varepsilon} = \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{N}\left(\varepsilon_i\right)^2. \tag{8}$$

When working with only one explanatory variable, it may be easy to draw a line by eye to fit the data. However, two analysts may draw lines with different slopes and intercepts. The least-squares regression method ensures consistency by defining and estimating an optimal fit. Note that the residual measures only the differences in the response variable and not in the explanatory variables.

Consider the bivariate linear modeling, having one response and one explanatory variable,

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_{[,1}\,, \tag{9}$$

with the least-squares estimate of

$$\hat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{X}_{[,1}\,. \tag{10}$$

The least-squares estimators of β_0 and β_1, can be obtained by substituting equation 10 into equation 8 and taking the partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. Setting the resulting two equations to zero, as if determining the location of extrema, and solving the system of equations gives the following estimators:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{N} Y_i - \hat{\beta}_1 \sum_{i=1}^{N} X_{i,1}}{N}\,, \tag{11}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(X_{i,1} - \frac{\sum_{i=1}^{N} X_{i,1}}{N} \right) \left(Y_i - \frac{\sum_{i=1}^{N} Y_i}{N} \right)}{\sum_{i=1}^{N} \left(X_{i,1} - \frac{\sum_{i=1}^{N} X_{i,1}}{N} \right)^2}\,. \tag{12}$$

The derivation of these equations can be carried further to a proof that the least-squares regression model presents a Best Linear Unbiased Estimator (B.L.U.E.). Beyond the fact that an unbiased estimate is produced, the discussion of B.L.U.E. is beyond the scope of this report.

The development of equations 11 and 12 rely on the assumptions that the linear model is the correct underlying model and that the mean of the residuals is zero. However, to draw additional inference from the development of the regression model and estimated coefficients, additional assumptions are needed. Commonly, the residuals are assumed to be independent and identically distributed variates that follow a normal distribution with a constant variance; the latter condition is called homoscedasticity. The result is that any estimate derived from the least-squares regression model can be considered to be a conditional mean of the response variable, given the fixed values of the explanatory variable. Further, the conditional mean is at the center of a conditionally normal distribution, as depicted in figure 7. These assumptions aid in model interpretation and evaluation.

Least-squares regression can be extended to multiple linear regression, which can evaluate multiple explanatory variables without loss of inferential fidelity. For multiple linear regression, it is convenient to consider matrix notation. By considering equation 2 to be representative of each observation of the response variable across a range of observations, a system of N equations can be developed. This system can be summarized by (1) $\boldsymbol{Y}$, an $N \times 1$ matrix of the observations of the response variable; (2) $\boldsymbol{X}$, an $N \times (M + 1)$ matrix of the observations of the explanatory variables with the first column consisting of unit values; (3) $\hat{\boldsymbol{\beta}}$, an $(M + 1) \times 1$ matrix of the linear coefficients of the response variable; and (4) $\boldsymbol{\varepsilon}$, an $N \times 1$ of the residuals, such that

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}\,. \tag{13}$$

In this formulation, the least-squares regression coefficients can be estimated as

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{Y}\,, \tag{14}$$

where

$\boldsymbol{\Lambda}$ is a square $N \times N$ matrix of weights on the observations of Y.

The differences in the weighting matrix lead to the different modes of regression: OLS, WLS, and GLS.

For hydrologic applications, if the response variable, a streamflow statistic, and the residual error therein contain both modeling and sampling errors, the variance of the residual errors, σ_ε^2, is a summation of sampling variability and modeling variability:

$$\sigma_\varepsilon^2 = \sigma_\eta^2 + \sigma_\delta^2\,, \tag{15}$$

where

σ_η^2 is the sampling error variance that reflects that part of the residual error variance that can be attributed to imprecise estimates of the observed response variable, primarily due to finite record length; and

σ_δ^2 is the modeling error variance that reflects that part of the residual error variance that results from an imperfect model that does not adequately explain all the variability seen in the observations.

In OLS regression, the sampling error variance cannot be separated from the modeling error variance. The modeling error reflects the variability in the difference between the regression estimate and the unknown, true value. Specific applications of WLS and GLS regression developed by Tasker (1980) and Tasker and Stedinger (1989) allow the sampling and modeling error variances to be estimated. Another advantage of the applications is that the availability of alternate weighting matrices account for different record lengths and sampling uncertainty.

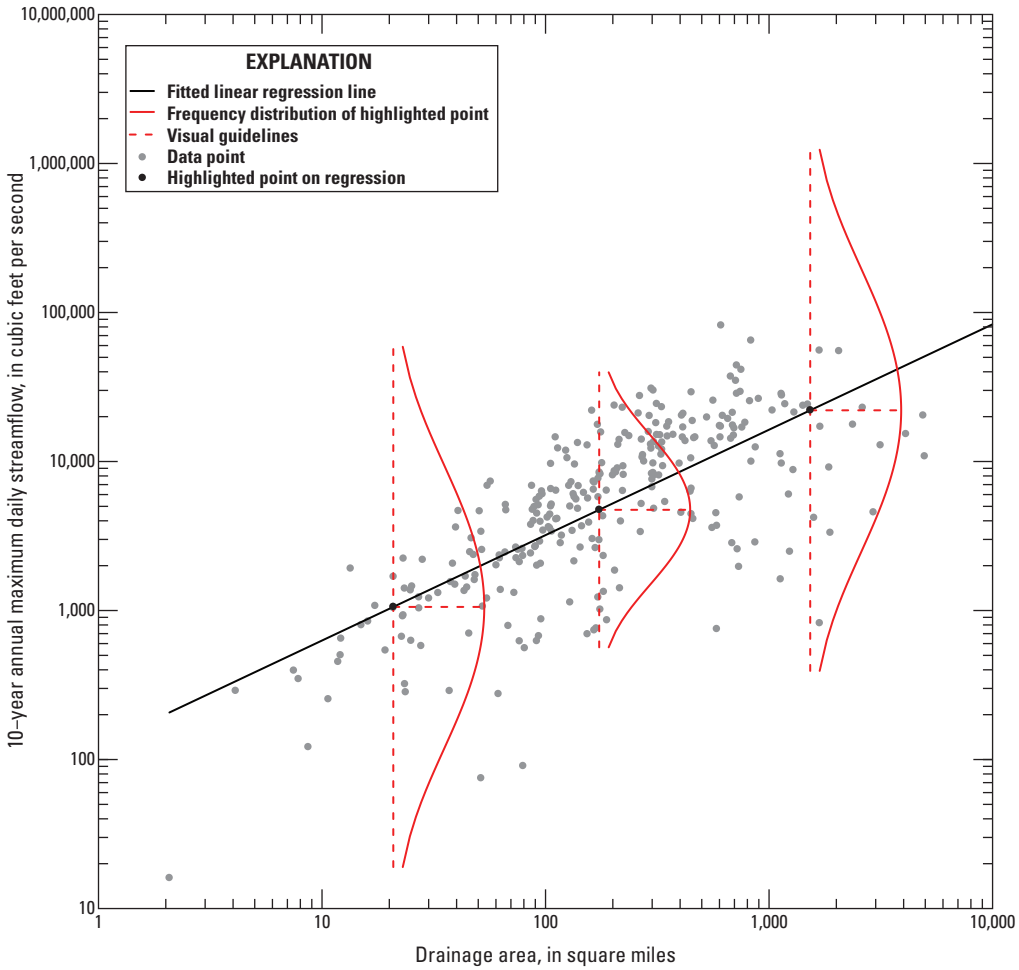

Figure 7. Graph showing the relationship between the response variable (10-year annual maximum daily streamflow) and explanatory variable (drainage area). The linear regression, as represented by the heavy black line, represents the mean of a normal distribution conditional on the observed explanatory variables. The bell-shaped normal curves represent the distribution of the observations about each black dot. Such a distribution can be developed for any point along the regression.

Ordinary Least-Squares Regression

OLS regression, the most traditional form of least-squares regression, assumes that the uncertainty associated with each of the observations is approximately equal. Accordingly, in the estimation of regression coefficients, all observations are weighted equally. In this case, the weighing matrix for OLS regression, Λ_{OLS}, is the identity matrix:

$$\Lambda_{OLS,i,j} = \begin{cases} 1 & for \quad i = j \\ 0 & for \quad i \neq j \end{cases} \tag{16}$$

The OLS approach is suitable for estimating regression parameters when there is little variation in the precision of observed response variables and the residuals are independent of each other.

Weighted Least-Squares Regression

WLS regression is a modification of OLS regression that accounts for heterogeneous uncertainties in the observed response variables. The uncertainty in streamflow statistics, which are commonly used as response variables, is often due to the length of the streamflow record used to compute the statistics. Record augmentation or extension also can introduce uncertainties into streamflow statistics. WLS regression accounts for these uncertainties by assigning weights to each variable based on the perceived uncertainty or variance in that observation. A variable with a small variance would be assigned a greater weight because there is less uncertainty in the estimated streamflow statistic. Conversely, a variable with a higher variance would be assigned a low weight because there is more uncertainty with respect to the true value of that streamflow statistic. Therefore, for WLS regression, the

elements of the weighting matrix, **Λ**, are a product of the type and source of the response variable.

There are several different approaches that can be used to estimate the elements of the weighting matrix. In all of the methods, the main diagonal of the matrix will have nonzero values, but the off-diagonal elements are all zero. The values in the main diagonal will sum to the number of observations used in the analysis. Otherwise, use of the weighting matrix will result in incorrect estimates of error. The optimal method for assigning weights is to divide the inverse of the variance of the estimates at each station by the mean of the inverses of the variances for all stations. Dividing by the mean, which is called centering, ensures that the total of the individual weights is equal to the number of observations. However, it is not always possible to quantify the exact variance. In these cases, there are several other approaches that can be used, such as defining weights based on the length of the record used to estimate the streamflow statistics, the variance of the time series used to calculate the streamflow characteristic, or some combination of these measures. These alternative approaches can only be used when the variances of the streamflow estimates are not available. When using WLS regression, the method used to assign weights for the different variables will be documented in the resulting report.

Tasker (1980) developed a method for estimating the values in the weighting matrix specifically for use with frequency statistics calculated using a log-Pearson Type III analysis (for example, as advised by Bulletin 17C (England and others, 2018). When the response variable is such a frequency statistic, the weighting matrix can be calculated as

$$\hat{\Lambda}_{WLS,i,j} = \begin{cases} \sigma_\delta^2 + \dfrac{c_1}{m_i} & for \quad i = j \\ 0 & for \quad i \neq j \end{cases}. \tag{17}$$

The modeling error variance as mentioned in equation 16, σ_δ^2, is estimated as

$$\sigma_\delta^2 = max\left[0 \quad , \quad \sigma_{\varepsilon|OLS}^2 - c_1\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\right)\right], \tag{18}$$

where

$\sigma_{\varepsilon|OLS}^2$ is the mean squared error (MSE) of estimates derived from an OLS regression on the same explanatory variables,

m_i is the record length used to compute the *i*th observation of the response variable, and

N is the total number of observations.

The coefficient, c_1, is given as

$$c_1 = max\left(0 \quad , \quad \bar{\sigma}^2\left(1 + \frac{\bar{K}^2}{2}\left(1 + 0.75\overline{g_w^2}\right) + \bar{K}\overline{g_w}\right)\right), \tag{19}$$

where

$\bar{\sigma}$ is the arithmetic average of the standard deviations of the annual-time series of the streamflow used to compute the streamflow statistics used as the explanatory variable,

$\bar{K}$ is the arithmetic average deviate from the log-Pearson Type III distribution used to estimate the response variable at each streamgage considered in the analysis, and

$\overline{g_w}$ is the arithmetic average skewness from the log-Pearson Type III distribution used to estimate the response variable at each streamgage considered in the analysis.

The log-Pearson Type III deviate is determined as a function of the probability of exceedance and skewness (U.S. Interagency Advisory Committee on Water Data, 1982).

The skewness values used in this development can be either the at-site skewness or the weighted skewness, though the weighted values (g_w) are shown in equation 19. As described in Bulletin 17C (England and others, 2018), the weighted skewness for the *i*th streamgage is given as

$$g_{w,i} = \omega_i g_i + \left(1 + \omega_i\right) g_{Reg,i}, \tag{20}$$

where

$g_{Reg,i}$ is the regional skewness estimate applicable to the *i*th gage.

The weight, ω_i, is a function of the estimated MSE of the skewness value at the gage, MSE_{gi} and the estimated MSE of the regional skewness values, $MSE_{g_{reg}}$, such that

$$\omega_i = \frac{MSE_{g_{reg}}}{MSE_{g_i} + MSE_{g_{reg}}}. \tag{21}$$

A variety of methods are available to determine g_{reg} values (Bulletin 17C; England and others, 2018). As given by Griffis and others (2004) and Griffis and Stedinger (2009), the MSE of the skewness value at the gage, MSE_{gi}, is estimated as

$$MSE_{g_i} = \left[\frac{6}{m_i} + c_2\right]\left[1 + \left(\frac{9}{6} + c_3\right)g_i^2 + \left(\frac{15}{48} + c_4\right)g_i^4\right], \tag{22}$$

where

$$c_2 = -\frac{17.75}{m_i^2} + \frac{50.06}{m_i^3},$$

$$c_3 = \frac{3.93}{m_i^{0.3}} - \frac{30.97}{m_i^{0.6}} + \frac{37.1}{m_i^{0.9}}, \text{ and}$$

$$c_4 = \frac{6.16}{m_i^{0.56}} + \frac{36.83}{m_i^{1.12}} - \frac{66.9}{m_i^{1.68}}. \tag{23}$$

An important advantage of Tasker's (1980) approach to WLS regression is that it provides unique estimates of the modeling error variance and the sampling error variance. An

alternative approach to calculating the modeling error variance, σ_δ^2, is presented by Stedinger and Tasker (1986). Stedinger and Tasker's (1986) estimator has been demonstrated to be more precise than equation 17, but neither the estimator nor equation 17 includes a mix of approaches to compute streamflow characteristics at partial-record streamgages (Funkhouser and others, 2008).

Generalized Least-Squares Regression

There is often a high degree of similarity among streamflow statistics from neighboring streamgages. Therefore, streamflow statistics from different streamgages, and the response variables selected from them, cannot be assumed to be independent of each other. For streamflow statistics calculated from a log-Pearson Type III frequency analysis, Stedinger and Tasker (1985) introduce a GLS approach to parameter estimation that builds on WLS regression by accounting for both correlated streamflows and time-sampling errors. This GLS approach incorporates estimates of the covariances among residual errors at each streamgage into the elements of $\boldsymbol{\Lambda}$.

Tasker and Stedinger (1989) estimate the weighting matrix for GLS, $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{GLS}}$, as

$$\hat{\Lambda}_{GLS,i,j} = \begin{cases} \sigma_\delta^2 + \dfrac{\sigma_i^2}{m_i}\left[1 + K_i g_{w,i} + \dfrac{K_i^2}{2}\left(1 + \dfrac{3g_{w,i}^2}{4}\right)\right] & for \quad i = j \\ \dfrac{\hat{\rho}_{ij}\sigma_i\sigma_j m_{ij}}{m_i m_j}\left[1 + \dfrac{K_i g_{w,i}}{2} + \dfrac{K_j g_{w,j}}{2} + \dfrac{K_i K_j}{2}\left(\hat{\rho}_{ij} + \dfrac{3g_{w,i}g_{w,j}}{4}\right)\right] & for \quad i \neq j \end{cases}, \quad (24)$$

where

subscripts are indices of streamgages in the region of interest,
σ_δ^2 is the modeling error variance,
σ_i is the standard deviation of the streamflow time series used to estimate the streamflow statistic at the subscripted site,
$g_{w,i}$ is the weighted skewness at the subscripted site,
K is the log-Pearson Type III deviate of the subscripted site,
m_{ij} is the concurrent record length for the subscripted streamgages, and
$\hat{\rho}_{ij}$ is the estimated cross correlation among the time series of streamflow used to calculate the streamflow statistic.

The main diagonal elements of $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{GLS}}$ include the model error, $\boldsymbol{\delta}$, and all elements include the effect of the time-sampling error, $\boldsymbol{\eta}$, at the subscripted site.

For the system described in equation 13 to be solvable, it is necessary to estimate cross-site correlations. The sampling uncertainty in observed correlations may produce a singular, or noninvertible weighting matrix. Tasker and Stedinger (1989) suggest approximating the cross correlations as

$$\hat{\rho}_{ij} = \theta_1^{\left[\frac{d_{ij}}{\theta_2 d_{ij}+1}\right]}, \quad (25)$$

where

d_{ij} is the distance between the subscripted sites, and
θ_1 and θ_2 are dimensionless parameters, which are estimated from the observed data.

As described by Tasker and Stedinger (1989), the modeling error variances, σ_δ^2 in $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{GLS}}$, and the estimated coefficients are simultaneously determined by iteratively searching for a non-negative solution to

$$\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{\Lambda}_{GLS}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = N - (M+1), \quad (26)$$

where

$\hat{\beta}$ is determined from equation 14,

N is the number of observations, and

M is the number of explanatory variables used in the regression.

Equation 24 is an improvement compared to WLS regression because it accounts for correlated streamflows and time-sampling errors. However, it does not account for the uncertainty associated with estimating values of skewness. Depending on the actual magnitude of errors in estimation of skewness, this additional error may unduly influence the estimation of regression parameters. A method presented by Griffis and Stedinger (2007) accounts for this uncertainty in the skewness. In this instance, the skewness values must be weighted skewness values rather than at-site skewness values. The revised weighting matrix is

$$\hat{\Lambda}_{GLS-skew,i,j} = \begin{cases} \sigma_\delta^2 + \frac{\sigma_i^2}{m_i}[1 + K_i g_{reg,i} + \frac{K_i^2}{2}(1 + \frac{3g_{reg,i}^2}{4}) + \omega_i K_i \frac{\partial K}{\partial g_{reg,i}}(3g_{reg,i} + \frac{3g_{reg,i}^3}{4}) + \\ \omega_i^2 (\frac{\partial K}{\partial g_{reg,i}})^2 (6 + 9g_{reg,i}^2 + \frac{15g_{reg,i}^4}{8})] + (1-\omega_i)^2 \sigma_i^2 MSE_{g_{reg}} (\frac{\partial K}{\partial g_{reg,i}})^2 \end{cases} for\ i=j \\ \begin{cases} \frac{\hat{\rho}_{ij}\sigma_i\sigma_j m_{ij}}{m_i m_j}[1 + \frac{K_i g_{reg,i}}{2} + \frac{K_j g_{reg,j}}{2} + \frac{K_i K_j}{2}(\hat{\rho}_{ij} + \frac{3g_{reg,i}g_{reg,j}}{4}) + \\ \frac{\omega_i K_j g_{reg,i}}{2}\frac{\partial K}{\partial g_{w,i}}(3\hat{\rho}_{ij} + \frac{3g_{reg,i}g_{reg,j}}{4}) + \frac{\omega_j K_i g_{reg,j}}{2}\frac{\partial K}{\partial g_{reg,j}}(3\hat{\rho}_{ij} + \frac{3g_{reg,i}g_{reg,j}}{4}) + \\ \frac{\omega_i\omega_j\sigma_i\sigma_j m_{ij}}{\sqrt{m_i m_j}}\frac{\partial K}{\partial g_{reg,i}}\frac{\partial K}{\partial g_{reg,j}}\hat{\rho}_{ij}^3\sqrt{\sigma_{g_i}^2\sigma_{g_j}^2}] \end{cases} for\ i\neq j \end{cases} \quad (27)$$

Most of the variables in equation 27 are defined for equation 24 and the weights, ω_i, are defined for equation 20.

The partial derivative of the log-Pearson Type III deviates is calculated from Kite's (1975, 1976) approximations as

$$\frac{\partial K}{\partial g_{reg,i}} = \frac{\left(Z_p^2 - 1\right)}{6} + \frac{\left(Z_p^3 - 6Z_p\right)}{54} - \frac{\left(Z_p^2 - 1\right)g_{reg,i}^2}{72} + \frac{Z_p g_{reg,i}^3}{324} + \frac{5g_{reg,i}^4}{23{,}328}, \quad (28)$$

where

Z_p is the standard normal deviate corresponding to probability p.

Griffis and Stedinger (2009) approximate the variance of the skewness, $\sigma_{g_i}^2$, as

$$\sigma_{g_i}^2 = \left[\frac{6}{m_i} + c_2\right]\left[1 + \left(\frac{9}{6} + c_5\right)g_{reg,i}^2 + \left(\frac{15}{48} + c_6\right)g_{reg,i}^4\right], \quad (29)$$

where

$$c_5 = \frac{3.92}{m_i^{0.3}} - \frac{31.1}{m_i^{0.6}} + \frac{34.86}{m_i^{0.9}}, \text{ and}$$

$$c_6 = \frac{7.31}{m_i^{0.56}} + \frac{36.83}{m_i^{1.18}} - \frac{66.9}{m_i^{1.77}}. \quad (30)$$

Conclusions

Although OLS regression is the simplest approach to multiple linear regression, WLS or GLS regression is generally preferred when applicable. WLS and GLS regression methods better account for differences in the accuracy of streamflow statistics at different streamgages. GLS regression is preferred over WLS regression if the correlation in streamflow statistics among streamgages can be approximated by using a relationship between distance and degree of correlation. If there is no clear relationship between correlation and distance, the covariance matrix estimated using Tasker and Stedinger's (1989) method will not be accurate. In this case, WLS regression would be a better choice. However, note that the methods presented by Tasker (1980) and Tasker and Stedinger (1989) are specific to streamflow frequency statistics and cannot be used for other streamflow statistics or other response variables without modification.

Model Evaluation

The development of a regression model is a highly iterative process. Even after conducting an exploratory data analysis, it is necessary to consider several groups of explanatory variables to identify the model structure that provides the best predictive capacity while satisfying the assumptions of least-squares regression. As discussed in the following sections, model evaluation can be divided into (1) structural assessments to ensure the validity of assumptions and the representativeness of the model structure and (2) performance assessments to determine the goodness-of-fit and predictive capacity of the candidate models.

Structural Diagnostics

Before considering the predictive performance of a regression model, it is important to check that the fitted model validates the assumptions used to develop the model. Namely, one is most interested in the normality of residuals, homoscedasticity of residuals, and independence of residuals with respect to each other and all possible variables. If these assumptions are validated, inferences can be made from the fitted model and its coefficients. The most straightforward approach to assessment is to begin by examining the residuals. From there, it is possible to assess the model structure by considering which explanatory variables are included and their added value.

Residuals

The residuals of a fitted model provide important diagnostics of model adequacy and should be examined for normality, homoscedasticity, and independence. The validity of the fitted model should be questioned if the residuals are found to be nonnormal, heteroscedastic, or dependent. Without well-behaved residuals that are normally distributed, homoscedastic, and independent, the model assumptions are invalid and inferences cannot be drawn without numerous caveats.

A scatterplot showing the relationships among the residuals and the observations or predictions of the response variable can be used to assess the residuals. Figure 8 shows residual errors plotted as a function of the response variable, which is the predicted 10-year annual maximum daily streamflow. As described by Helsel and Hirsch (2002, p. 245), the residuals should be distributed around zero with a constant variability. Although the condition of unbiasedness (that is, the mean being zero), is important, figure 8 is more useful for assessing the constancy of variability in the residuals. Ideally, the residuals will exhibit homoscedasticity, which means that the variability or spread of the residuals around the zero-residual line is not dependent on the observed or predicted response variable. Panel *A* of figure 8 illustrates homoscedastic residuals, but panel *B* illustrates the opposite, heteroscedastic residuals.

In addition to exhibiting homoscedasticity, the residuals should be approximately normally distributed. As shown in figure 9 for the residuals of the example regression, a normal probability plot is a graphical method for evaluating the validity of this assumption. A normal probability plot is built by plotting the ordered residuals against the normal quantile of the ordered residuals, assuming the sample mean and variance of the errors are reasonable approximations of their true values. This process requires the assumption of a probability plotting position. A common approach to assigning a plotting position is to use the Weibull plotting position,

$$p_r = \frac{r}{N+1}, \tag{31}$$

where

r is the rank of the residual errors being considered, in increasing order;
N is the total number of observations; and
p_r is the plotting position or nonexceedance probability of the specified residual error.

Helsel and Hirsch (2002) discuss several alternative plotting positions. The normal quantile of this residual error is then given as

$$\hat{\varepsilon}_r = \overline{\varepsilon} + Z_{p_r}\sigma_{\varepsilon}, \tag{32}$$

where

$\hat{\varepsilon}_r$ is the theoretical rth quantile,
$\overline{\varepsilon}$ is the mean of the residuals,
σ_{ε} is the standard deviation of the residuals, and
Z_{p_r} is the standard normal quantile of the residual error given the plotting position p_r.

Plotting $\hat{\varepsilon}_r$ against the observed residuals, ε_r, gives a normal probability plot. Deviations from a line with a slope of one and an intercept of zero (thereby passing through the origin) demonstrate deviations from normality. Assuring the homoscedasticity and normality of residuals is an attempt to validate the assumption that the marginal distribution of residuals is not a function of the predicted response variables.

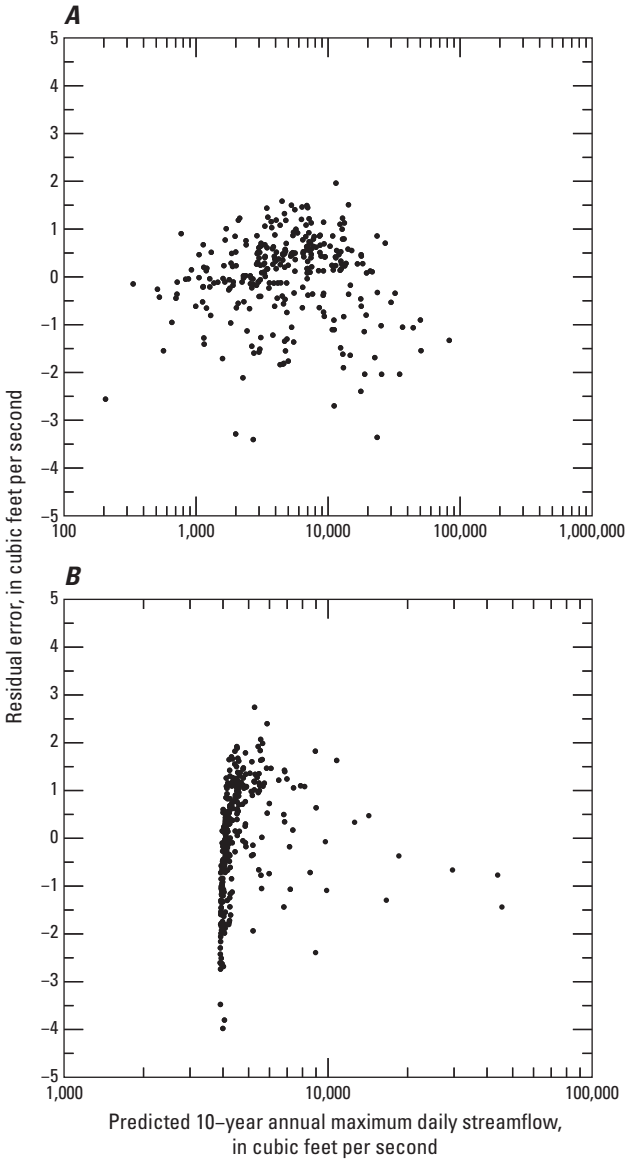

Figure 8. Scatterplots showing residual errors plotted as a function of the response variable (predicted 10-year annual maximum daily streamflow). *A*, Scatterplot showing a homoscedastic relationship between the residuals and the response variable. Homoscedasticity is the ideal situation because the variability or spread of the residuals around the zero-residual line is not dependent on the observed or predicted response variable. *B*, Scatterplot showing a heteroscedastic relationship between the residuals and the response variable. The variability of the residuals in panel *B* noticeably decreases with an increasing magnitude in the response variable.

Leverage and Influence

After assessing the distribution of residuals, it is useful to consider the influence that the specific observations of the explanatory and response variables contained in the dataset may have on regression. This process is conducted by considering two metrics: leverage and influence. Leverage measures the distance of each independent observation from the other observations. Influence measures the sensitivity of regression parameters to any particular observation. Observations with high leverage and substantial influence require further examination.

Leverage measures how far away the value of one observation of the set of explanatory variables is from the centroid of all other observations. Leverage gives an indication of whether the values of the explanatory variables for any particular observation are unusual when compared to other observations. In matrix notation, the leverage of each observation is given on the main diagonal of

$$\boldsymbol{h} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}. \tag{33}$$

Leverage values are considered large if

$$h_{i,i} > h_{limit} = \frac{C_h}{N}\sum_{i=1}^{N} h_{i,i}, \tag{34}$$

where

C_h is a constant.

Although the leverage metric identifies unusual observations, such unusual observations may or may not have any significant influence on the estimated regression coefficients. An influence metric, such as Cook's D (Cook, 1977), indicates whether an observation has a large influence on the estimated regression parameter values. Cook's D is calculated as

$$D_i = \frac{e_i^2 L_{i,i}}{(M+1)\left(\text{›}_{i,i} - L_{i,i}\right)^2}, \tag{35}$$

where

M is the number of explanatory variables considered,

$\Lambda_{i,i}$ is the *i*th main diagonal of the Λ weighting matrix, and

$L_{i,i}$ is the *i*th main diagonal of $\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ (Tasker and Stedinger, 1989).

Commonly (for example, Montgomery and others, 2006), an observation can be considered to have large influence if

$$D_i > D_{limit} = \frac{4}{N}. \tag{36}$$

An observation that follows the relationship indicated by other observations can have high leverage but low influence. Although neither model would be acceptable for interpolation beyond the main cluster of data points, figure 10 shows examples of an observation with high leverage but without high influence (panel *A*) and high leverage with high influence (panel *B*). Figure 10 is based on this report's example regression, with an artificial data point added for illustration.

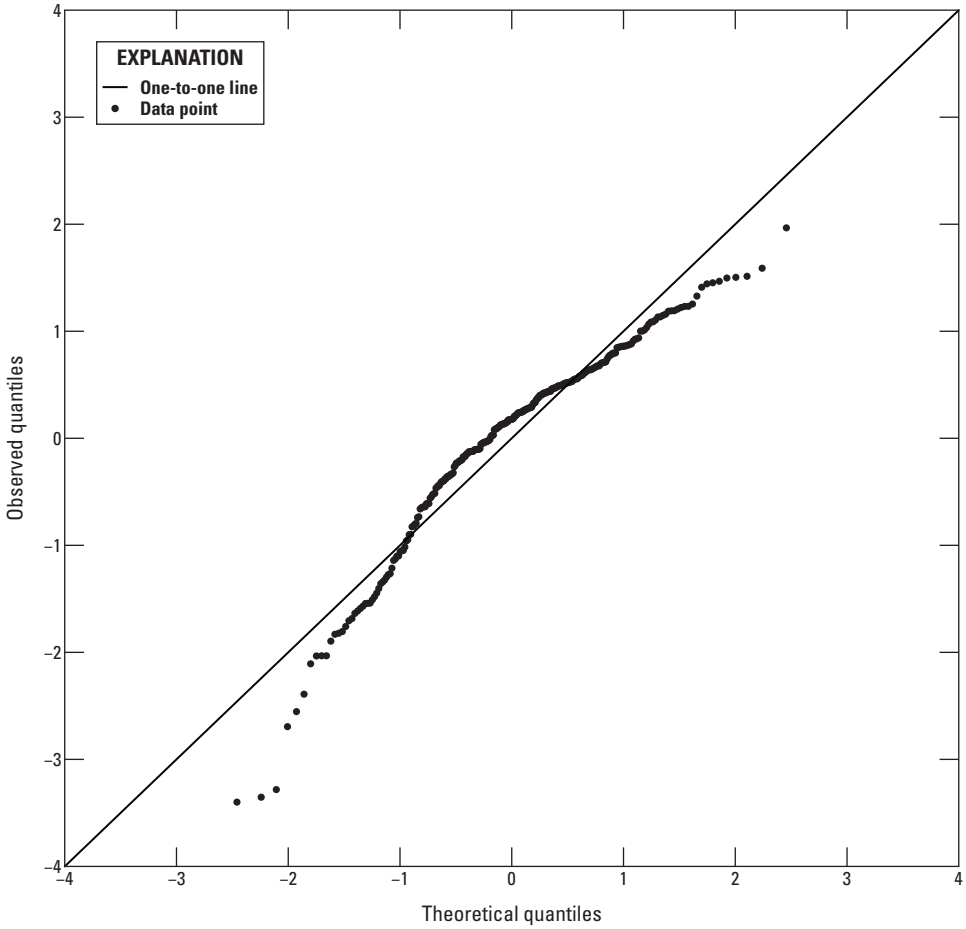

Figure 9. Example of a normal probability plot for checking the normality of residuals. The plot shows the relationship among observed quantiles and theoretical quantiles. If normally distributed, the points should plot near the straight line with a slope of 1 and an intercept of 0 thereby passing through the origin. This example suggests that the data may not be distributed normally.

All high leverage and high influence observations should be checked to ensure that the values of both the response and explanatory variables were correctly calculated. If an error is found and cannot be corrected, the observation should be removed from the analysis. However, removing the observation may lessen the applicable range of the regression models. In general, observations should not be removed from an analysis simply because they exhibit high leverage or high influence. Every effort should be made to identify the differences between the high leverage or high influence observations and the other observations. This understanding may inform the limitations of the regression model. If the applicable range is not constrained, the error metrics calculated using the smaller set of observations will not reflect the true uncertainty in predictions from the regression model.

There are several ways to lessen the influence of a high influence observation without obvious errors. First, efforts should be made to find other sites with similar explanatory variables. Further improvements may be obtained by separating observations into different groups (for example, groundwater-dominated sites, very small basins). By creating a new class of sites, the observation may no longer be an outlier or may have less influence. Using a model with other explanatory variables may also prove useful. The most extreme action is the removal of the observation with high influence, thereby limiting the applicability of the regression.

Variable Selection

The most important elements of a model's structure are the explanatory variables included in the model. However, determining which variables should be included in the model is one of the greatest challenges of model development. In general, a variable should be included if it provides unique, linear information; the associated coefficient is significantly different from zero; and the sign and magnitude of the associated coefficient are reasonable given a physical interpretation of the result. For each explanatory variable, a minimum

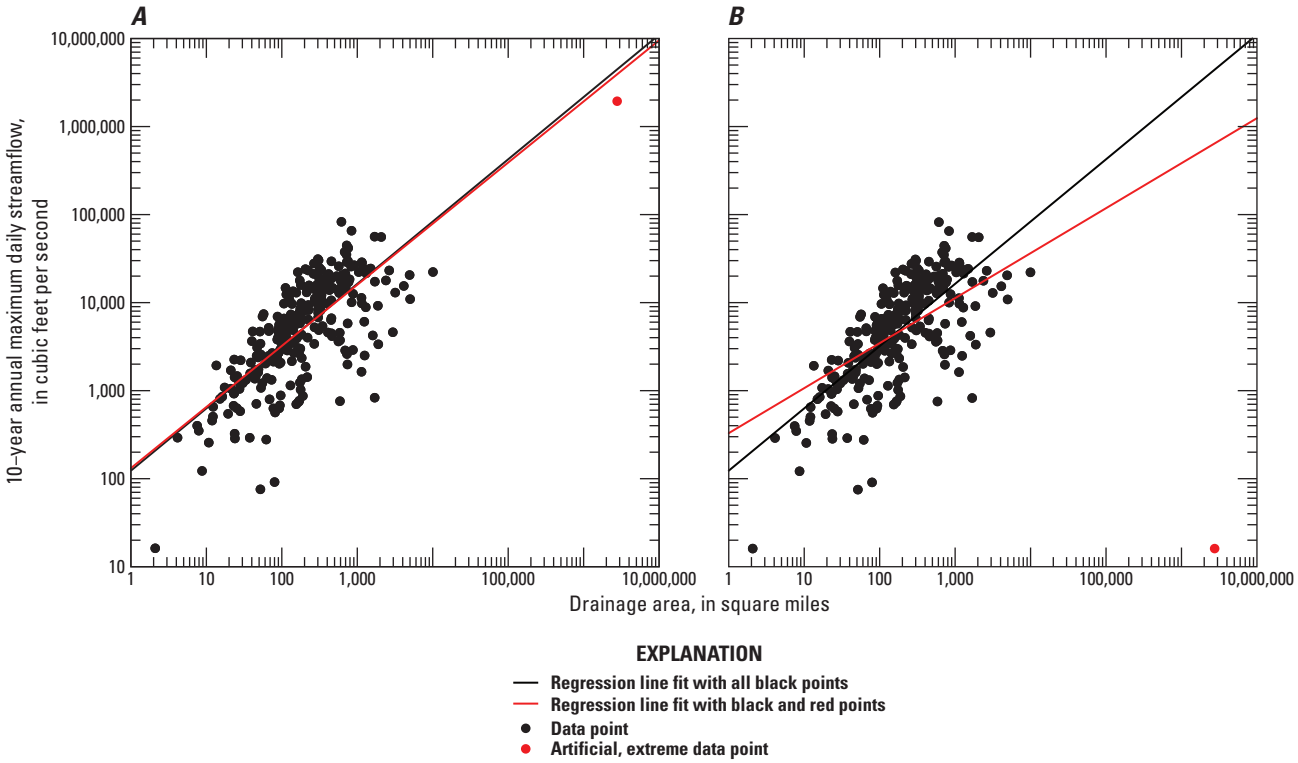

Figure 10. Scatterplots showing examples of leverage with and without high influence. An artificial data point (in red) has been added to illustrate high leverage. *A*, Scatterplot showing high leverage without high influence. The artificial data point produces a linear regression (in red) that is similar to the dotted, black line of regression based on all data excluding the artificial point. *B*, Scatterplot showing high leverage with high influence. The artificial point exhibits the same degree of leverage because the associated explanatory value has not changed, but the influence is substantial, shifting the red regression line downward from the regression line with the value omitted.

number of observations should be used in the final regression. A common minimum number is 10 observations for each variable, however, this requirement has not been codified.

Each explanatory variable should contain unique information that is linearly related to the response variable. During exploratory data analysis, the linearity among explanatory variables and the linearity with the response variable were considered. After a regression has been fitted, it is possible to provide a more robust assessment. The uniqueness of information can be assessed by considering the phenomenon of multicollinearity. Partial-regression plots are useful for determining the linearity of the relationship with each explanatory variable.

Partial-regression plots for each explanatory variable document the linearity of the relationships between the explanatory variable and the response variable. Partial-regression plots show the partial residual against each adjusted explanatory variable. Partial residuals are unique to the explanatory variable being assessed. They are calculated as a residual error from a regression built with all explanatory variables except the one being assessed. The adjusted explanatory variable is the residual error between the observed explanatory variable and the predicted explanatory variable. The predicted explanatory variable is treated as a response variable and is predicted from a regression built on all other explanatory variables. The partial-residual plots in figure 11 show an assessment of the partial residual from regressing the logarithm of the 10-year annual maximum streamflow against the logarithm of drainage area and the untransformed value of average annual precipitation. Partial-residual plots should appear linear. Curvature in a partial-residual plot is the best indicator that variable transformation is needed. The partial-residual plots should also demonstrate homoscedasticity and normality when assessed against the response variable.

In addition to a linear relationship, each explanatory variable should provide unique information. Multicollinearity is a condition in which two explanatory variables have strong linear dependencies and are in essence moving with each other (Myers, 1990). Such a condition inflates the variance of the regression coefficients, affecting the precision of the resultant estimates. Consequently, variables that are highly correlated should not be included in the same regression model because the redundant information contained in highly correlated explanatory variables affects the fit of the regression coefficients. For example, main-channel length and drainage area are likely to be highly correlated because as the drainage area gets larger, the main-channel length tends to get longer.

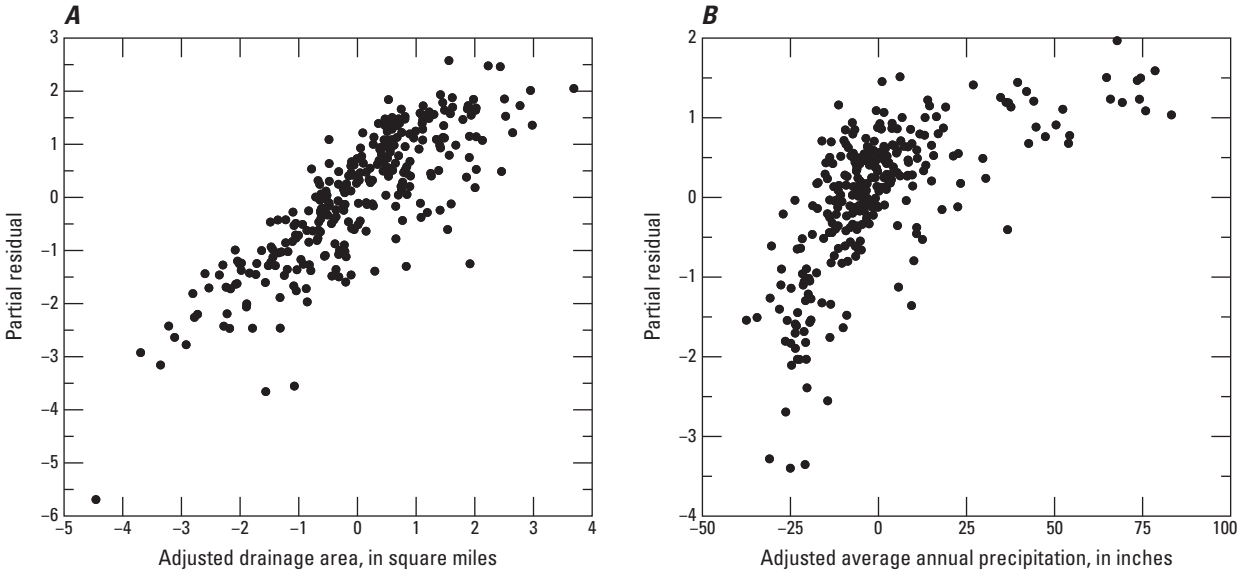

Figure 11. Partial-residual plots for the example regression with two explanatory variables. The regression is of the logarithmically transformed 10-year annual maximum streamflow against *A*, the logarithmically -transformed drainage area and *B*, the untransformed basin precipitation. The curvature in panel *B* suggests that variable transformation may be appropriate for precipitation.

Having both variables in the same equation is not adding new information, and in fact, can cause the equation to produce erroneous results. Indications of multicollinearity include coefficients of explanatory variables that change noticeably when adding or removing an explanatory variable, or signs of coefficients that are not what would be expected.

A quantitative check for multicollinearity is the variance inflation factor. The variance inflation factor, *VIF*, is calculated as

$$VIF_k = \frac{1}{1 - R_k^2}, \tag{37}$$

where

R_k^2 is the coefficient of determination from a regression designed to predict the *k*th explanatory variable as a function of all other explanatory variables.

Equation 37 is the same regression used in the development of adjusted explanatory variable for partial-residual analysis. A variance inflation factor greater than 5 or 10 (corresponding to R_k^2 greater than 0.8 to 0.9) indicates highly correlated predictors that warrant further investigation. This rule is a commonly used decision rule, but there may be cases where highly correlated variables are worth retaining. If highly correlated variables are retained, the variance of the fitted coefficients will be magnified by the variance inflation factor, which increases the uncertainty about the correct value of a coefficient. This uncertainty carries through from regression fitting to prediction and characterization of the confidence in estimates.

Some degree of multicollinearity can be mitigated through sample design, such as ensuring that all observations capture a wide range of the possible combinations of explanatory variables. Gaps in the representativeness of data can result in multicollinearity. For example, consider a regression of streamflow with two explanatory variables: drainage area and percentage of basin urbanized. If the selected monitoring network included large drainage areas with low urbanization and small drainage areas with high urbanization, and these observations were not representative of the true underlying distribution, a possibly misleading negative correlation between the two explanatory variables would result. When a correlation exists between two variables but does not result from any causal relationship between them, it may be possible to collect observations that deviate from this sampling bias, for example, large drainage areas with high degrees of urbanization. However, this approach is not always possible due to data unavailability or nonexistence and the strategy will not work when two variables are correlated.

Another possible remedy for multicollinearity is to eliminate one or more of the explanatory variables. As mentioned in the "Exploratory Data Analysis" section, trying each of the correlated explanatory variables in turn, while keeping other explanatory variables constant, can help determine which of the correlated explanatory variables is most useful for the final regression model. Kroll and Song (2013) note that multicollinearity is most effective when trying to draw inferences from regression coefficients. However, if interest is only in prediction, then multicollinearity is not a major concern. Kroll and Song (2013) provide a deeper discussion of multicollinearity and possible approaches to correction.

If an explanatory variable is to be included in the regression model, it is appropriate to determine if the fitted

coefficients are significant. In addition to estimating the values of the model coefficients, least-squares regression provides methods for estimating the variance of these estimates as

$$\hat{\sigma}^2_{\beta_k} = \left\{ \begin{array}{cc} \frac{SS_\varepsilon}{N-M-1}\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1} & for\,OLS \\ \left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1} & else \end{array} \right\}_{k,k}. \quad (38)$$

The need to append the leading term, which is equivalent to the MSE, for OLS regression arises because the weighting matrix $\boldsymbol{\Lambda}$ contains the MSE for the specialized forms of regression discussed in this report except for OLS regression. Based on the assumptions of least-squares regression, each coefficient approximately follows a Student's t-distribution with $(N-M-1)$ degrees of freedom. Accordingly, the t-value for each coefficient can be estimated as

$$t_{\beta_k} = \frac{\hat{\beta}_k}{\hat{\sigma}_{\beta_k}}. \quad (39)$$

However, these variances and the distributive representation are only valid if all the assumptions of least-squares regression are valid. Any deviation restricts the use of these approximations and the inference derived thereof.

Because the coefficients follow a Student's t-distribution, it is possible to develop confidence intervals on the coefficients or apply a hypothesis test to determine the significance of each coefficient. Such tests, producing *p*-values, are widely used in regression analysis. However, caution should be used in interpreting significance. The American Statistical Association recently reminded practitioners that *p*-values are closely linked with development of a null hypothesis and are not a measure of the accuracy of any alternative hypothesis (Wasserstein and Lazar, 2016). Science and regression should not be based only on the significance of *p*-values (Wasserstein and Lazar, 2016).

The idea of a significance test is to evaluate the null hypothesis that a coefficient is zero against an alternative hypothesis that the coefficient is not zero. The tests on each variable are treated as independent. The *p*-value of a given coefficient is

$$p_{\beta_k} = 2Prob\left(t > \left|t_{\beta_k}\right|\right). \quad (40)$$

The *p*-value, p_{β_k}, gives the probability of a similarly extreme value of $\hat{\beta}_k$ based on the assumption that the null hypothesis is true. A small p_{β_k} suggests that similarly extreme observations are unlikely under the null hypothesis and is commonly considered evidence that the null hypothesis is invalid. This test is often simplified by specifying a level of significance, commonly 0.05. A *p*-value of less than this level of significance is evidence that the null hypothesis should be rejected and the coefficient is considered significant. If there is evidence to reject the null hypothesis, it is wise to retain the explanatory variable in question. Finally, each significance test is conducted independently. Therefore, a standard F-test should be used to test for dependence or joint significance.

Variables that are determined to represent unique, linear information and have a regression parameter significantly different from zero are used for the regression. For these variables, it is important to consider the estimated coefficients with respect to expectations based on the physical system. As with exploratory data analysis, the users' intuition and system understanding are pivotal. The coefficients and fitted model should yield reasonable predictions when reasonable values of the explanatory variables are used. It is important to consider the estimated intercept carefully and ask if the quantified relationships behave as would be expected given the physical constraints of the system. In such an analysis, more attention is paid to the slopes than the intercept.

Each slope coefficient should reasonably represent the isolated effect of the associated explanatory variable. Both the sign and magnitude of the coefficient should make sense. A positive coefficient indicates that the larger the value of the explanatory variable, the larger the response variable. For example, with a streamflow statistic, a positive coefficient makes sense for explanatory variables such as drainage area or precipitation. Similarly, although special attention should be paid to units, the magnitude of the coefficient should be checked for reasonableness. Signs or coefficient magnitudes determined by the model that are not reasonable indicate that there is a problem with the model or inputs.

Before leaving the subject of variable selection, it is important to address the field of automated variable selection. To date, nothing can replace the intuitive understanding of the physical system supplied by a knowledgeable user. However, there are several tools for automated variable selection; such tools are often included in statistical software packages. These methods can allow quick screening of variables but the applicability of the suggested variables should be carefully reviewed. Exploratory data analysis should give the analyst a feel for what variables are likely to be significant, even before automated procedures are used.

Automated algorithms for variable selection can be classified into four groups: forward selection, backward elimination, stepwise selection, and all-possible-subsets selection (Helsel and Hirsch, 2002). Forward selection schemes first select the most significant explanatory variable from the candidate variables supplied by the user. Then, one by one, the scheme adds the next most significant variable until a criterion is reached. With backward elimination, the program starts with all explanatory variables and eliminates them one at a time, based on which of the remaining variables is least significant. Stepwise selection is similar to forward selection, but, at each iteration, all variables are also checked to see if one should be removed. The final method, all-possible-subsets selection, is computationally intensive because every possible combination of the candidate explanatory variables is tested. The software for this method typically allows the user to select a maximum number of explanatory variables to use. The all-possible-subsets selection method is the suggested method because it

allows for extensive exploration without omitting possible combinations. With any of these methods, inclusion of highly correlated explanatory variables should be avoided. If it is desired to test highly correlated variables, only one of them should be included in a single model-selection run at a time. In addition to variable selection techniques, there are procedures that can evaluate models on other metrics, such as the Mallows Cp. Whichever approach is used, the model identified should undergo rigorous review of model adequacy following the procedures described in this report.

Performance Diagnostics

After validating underlying model assumptions and evaluating the structure of the candidate model, it is appropriate to quantify the performance and predictive capacity of the regression model. There are a wide array of tools to assess model performance and some of the most common tools are discussed in this section. These include the coefficient of determination, the MSE, confidence intervals, and prediction intervals. Although other metrics may be useful for specific applications, the methods discussed in this report are considered essential evaluations of model performance.

Coefficient of Determination and Mean Squared Error

The coefficient of determination, commonly referred to as R^2, is probably the most common metric of regression performance. The coefficient of determination measures the proportion of the variation in the response variable that is explained by the linear combination of the explanatory variables represented in the regression model. Montgomery and others (2006) and numerous other sources provide the basic calculation as

$$R^2 = 1 - \frac{SS_\varepsilon}{SS_A}, \tag{41}$$

where

SS_ε is the residual sum of squares, as given in equation 8; and

SS_A is the total sum of squares, given as

$$SS_A = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2, \tag{42}$$

where

$\bar{Y}$ is the mean of the observed response variable.

The total sum of squares represents the total variability in the response variable, but the residual sum of squares represents the variability in the response variable that remains after the regression is applied. The coefficient of determination ranges from 0 to 1, with higher values indicating that the regression explains more of the variability in the response variable.

Adding more explanatory variables will almost always reduce the residual sum of squares. Therefore, adding additional variables, whether appropriate or not, will almost always improve the coefficient of determination. This improvement is typically artificial. To address this phenomenon when comparing candidate regression models with different numbers of explanatory variables, the coefficient of determination can be adjusted for the number of explanatory variables such that

$$R^2_{adj} = 1 - \frac{SS_\varepsilon / \left(N - M - 1\right)}{SS_A / \left(N - 1\right)}. \tag{43}$$

Consequently, adding explanatory variables to a regression that do not reduce the residual sum of squares will decrease R^2_{adj}, decreasing the value of increases in model performance.

For WLS and GLS regressions, Griffis and Stedinger (2007) suggest a performance metric based on the modeling error variance. Because this metric relies on the modeling error variance, it cannot be used for OLS regressions because the modeling error variance is not separated from sampling error. Regardless, this pseudo coefficient of determination, R^2_{pseudo}, is given as

$$R^2_{pseudo} = 1 - \frac{\sigma^2_{\delta|M}}{\sigma^2_{\delta|0}}, \tag{44}$$

where

$\sigma^2_{\delta|M}$ is the modeling error variance from a WLS or GLS regression with M explanatory variables, and

$\sigma^2_{\delta|0}$ is the modeling error variance from a WLS or GLS regression with no explanatory variables.

R^2_{pseudo} is based on the variability in the response variable explained by the regression after the effect of the time-sampling error is removed.

In addition to the coefficient of determination, the MSE is a common performance metric of regression models. In a manner, the MSE is the variance of the residuals, as discussed with respect to equation 15. The residuals are the deviations of observations from the regression line; therefore, the variability describes the spread of the observations around the regression line. A smaller spread, as represented by a smaller MSE, indicates a better model fit. The MSE is calculated by dividing the residual sum of squares by the degrees of freedom in the model such that

$$MSE = \frac{SS_\varepsilon}{N - M - 1}. \tag{45}$$

As with all variances, the MSE takes on squared units of the response variable. It is therefore often convenient to take the square root of the MSE, typically called the root MSE or standard error. The root MSE is in the same units as the response variable and represents the standard deviation of the residuals. In cases where the response variable is a logarithmic transform using the common logarithm, the root MSE can be expressed as a percentage as

$$\sqrt{MSE}_{\%} = 100\sqrt{e^{(ln(10))^2 MSE} - 1}. \tag{46}$$

Aitchison and Brown (1957) provide the appropriate conversion for other logarithmic transformations.

The variance of the residuals, that is, the MSE, can be divided into the sampling error variance and the modeling error variance. Although it is not always possible to make this delineation in practice, the methods presented in this report for WLS and GLS regression of frequency statistics allow for this separation. Because WLS and GLS regression weight observations uniquely and MSE regression does not, it is strongly advisable to consider the separate sampling and modeling error variances when possible. Although the sampling error cannot be controlled by regression, consideration of both sampling and modeling error allows an analyst to determine the isolated modeling error variance.

Confidence Intervals

A confidence interval is a range that purports to contain the true, conditional estimate with some degree of confidence. The variability is the result of uncertainty in the regression parameters. The regression line represents a conditional mean of the response variable, given the observed explanatory variables. Furthermore, as shown in figure 7, the conditional mean is the center of a conditionally normal distribution of the response variables for the given observations of the explanatory variables. Given the assumptions of least-squares regression, it is possible to describe the conditional variance of each estimate and thereby develop confidence intervals around the estimate.

The conditional variance of the estimated response variable describes uncertainty in the conditional mean, for example, the estimated response variable, and describes the placement of the regression line. The conditional variance is calculated as

$$\sigma^2_{\hat{Y}_i|\boldsymbol{X}_i} = \begin{cases} MSE\left[\boldsymbol{X}_i^T\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i\right] & for\,OLS \\ \boldsymbol{X}_i^T\left(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_i & else \end{cases}, \quad (47)$$

where

X_i is the $(M + 1) \times 1$ vector of the explanatory variables associated with observation *i*.

OLS regression requires multiplication by the MSE because the MSE is not embedded in the weighting matrix. For WLS and GLS regression, as discussed in this report for streamflow frequency statistics, the MSE is already embedded in the weighting matrix. This conditional variance represents the variability resulting from uncertainty in the regression coefficients. The conditional variance is often referred to as the "sampling error variance of the regression." However, the conditional variance is different from the sample error variance of the response variable described with respect to equation 15.

Because estimated response variables are the means of a conditional normal distribution, the mean follows a Student's t-distribution with *N*–*M*–1 degrees of freedom. This fact is used to construct the confidence intervals surrounding the estimated response variables. Specific confidence intervals are parameterized to a specific level of confidence. By tradition, confidence intervals are presented symmetrically around the estimate. The level of confidence is defined as 100(1–α)% where α is a value between 0 and 1. Such a confidence implies that observations can be expected to fall outside of the confidence interval 100α% of the time, with each end equally probable. The bounds are the interval are given as

$$\left[\hat{Y}_{i,lower|\boldsymbol{X}_i} \quad , \quad \hat{Y}_{i,upper|\boldsymbol{X}_i}\right] = \hat{Y}_i \pm t_{(\alpha/2),(N-M-1)}\sigma_{\hat{Y}_i|\boldsymbol{X}_i}, \quad (48)$$

where

$t_{(\alpha/2),(N-M-1)}$ is the t-value that is exceeded 100(α/2)% of the time from a Student's t-distribution with *N*–*M*–1 degrees of freedom. This value can be determined from a standard statistical table or most statistical software.

Confidence intervals can be constructed for any level of confidence, but, by convention, the 90- or 95-percent confidence interval is typically reported.

Confidence intervals describe the uncertainty in an estimated response variable given observations of the explanatory variable. They arise from uncertainty in the regression coefficients. However, the interval also quantifies the uncertainty in the placement of the regression line. Figure 12 shows a 95-percent confidence interval for the regression of logarithmically transformed 10-year annual maximum daily streamflow and the logarithm of drainage area. The example demonstrates that the results of several different regressions can be included within the confidence interval. Each blue line is generated from an identically sized resample, with replacement, of the original data.

Prediction Intervals

The confidence interval quantifies the uncertainty in the placement of the regression line or the uncertainty of the estimate that results from the uncertainty in regression coefficients. However, a confidence interval does not account for the uncertainty associated with a limitation of the model, namely the failure to explain all the residuals. For this problem, prediction intervals, which combine the uncertainty in the placement of the regression line with the uncertainty associated with the residuals, are needed. Prediction intervals give the expected range of estimated response variables when the regression is used in practice.

Similar to confidence intervals, prediction intervals are based on the distribution assumed for least-squares regression and an associated variance. In this case, the appropriate variance is the variance of prediction, $\sigma^2_{\hat{Y}_i,pred|\boldsymbol{X}_i}$, which is given as

$$\sigma^2_{\hat{Y}_i,pred|\boldsymbol{X}_i} = \begin{cases} MSE + \sigma^2_{\hat{Y}_i|\boldsymbol{X}_i} & for\,OLS \\ \sigma^2_\delta + \sigma^2_{\hat{Y}_i|\boldsymbol{X}_i} & else \end{cases}. \quad (49)$$

Because it is not possible to estimate the modeling error variance for OLS regression, the MSE is used instead. The square

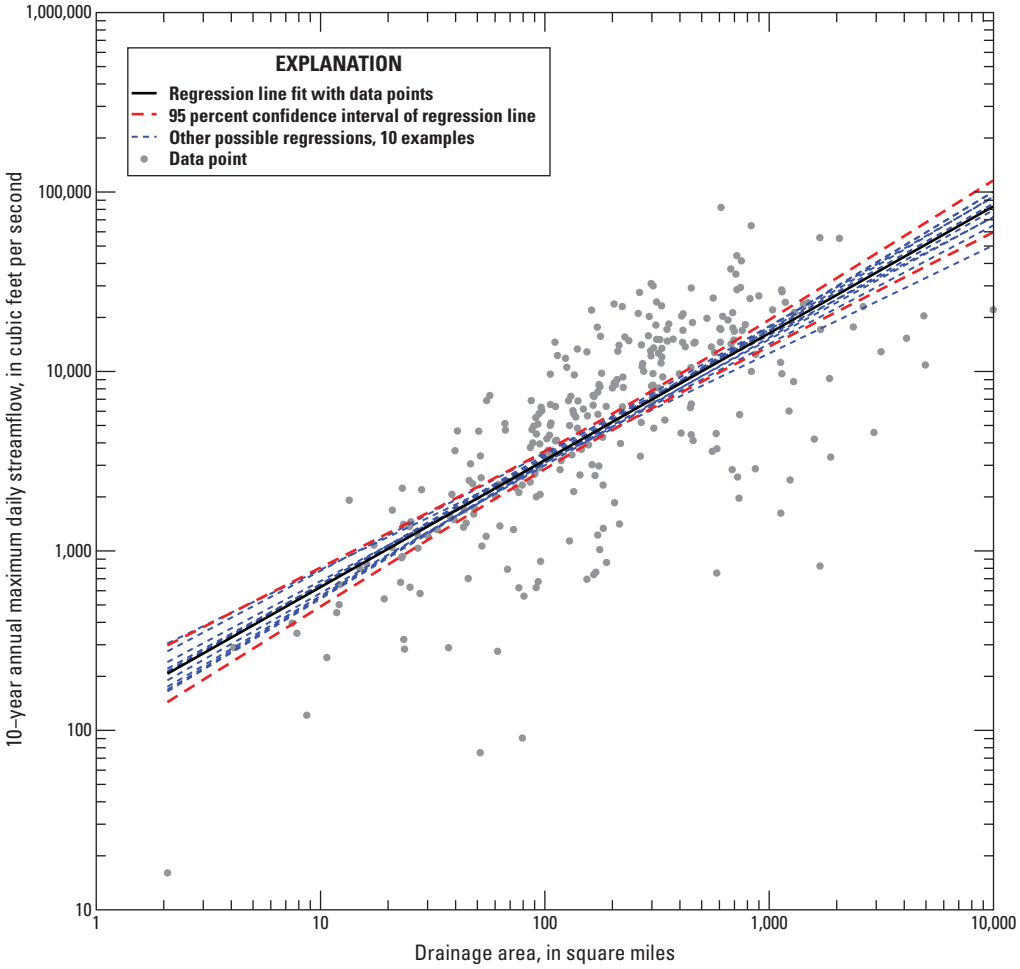

Figure 12. Graph showing a 95-percent confidence interval, in red, for the linear regression line, in black, based on data points, in gray, of the regression of logarithmically transformed 10-year annual maximum daily streamflow and the logarithm of drainage area. The dashed blue lines indicate the uncertainty in the regression coefficients and, therefore, the placement of the regression line. The confidence interval contains 95 percent of the possible regression lines, but not 95 percent of the observed data.

root of the variance of prediction is often called the standard error of prediction. If the response variable is a logarithmic transform, it can be presented as a percentage using the approach discussed with respect to the MSE. Regardless, the prediction interval is then estimated as

$$\left[\hat{Y}_{i,pred,lower|\boldsymbol{X}_i} \quad , \quad \hat{Y}_{i,pred,upper|\boldsymbol{X}_i}\right] = \hat{Y}_i \pm t_{(\alpha/2),(N-M-1)}\sigma_{\hat{Y}_i,pred|\boldsymbol{X}_i}. \quad (50)$$

As suggested by the additive term, prediction intervals are always wider than confidence intervals because of the additional uncertainty in the residuals and the regression itself. Figure 13 presents a prediction interval of the regression of logarithmically transformed 10-year annual maximum daily streamflow and the logarithm of drainage area.

The average variance of prediction across all observations used to develop the regression is a common summary performance metric that is computed by taking the arithmetic average of variances of prediction. If the observations used in the analysis are representative of the population of possible observations, the average variance of prediction is a good measure of, as the name implies, the uncertainty inherent in using this regression for prediction. However, these summary metrics are not a substitute for a full cross validation of model performance.

Cross Validation

Coefficients of determination, MSEs, confidence intervals, and prediction intervals are extremely valuable tools for understanding the performance, appropriateness, and relative power of a regression model. However, they are limited in scope. Evaluating the accuracy of the response variable

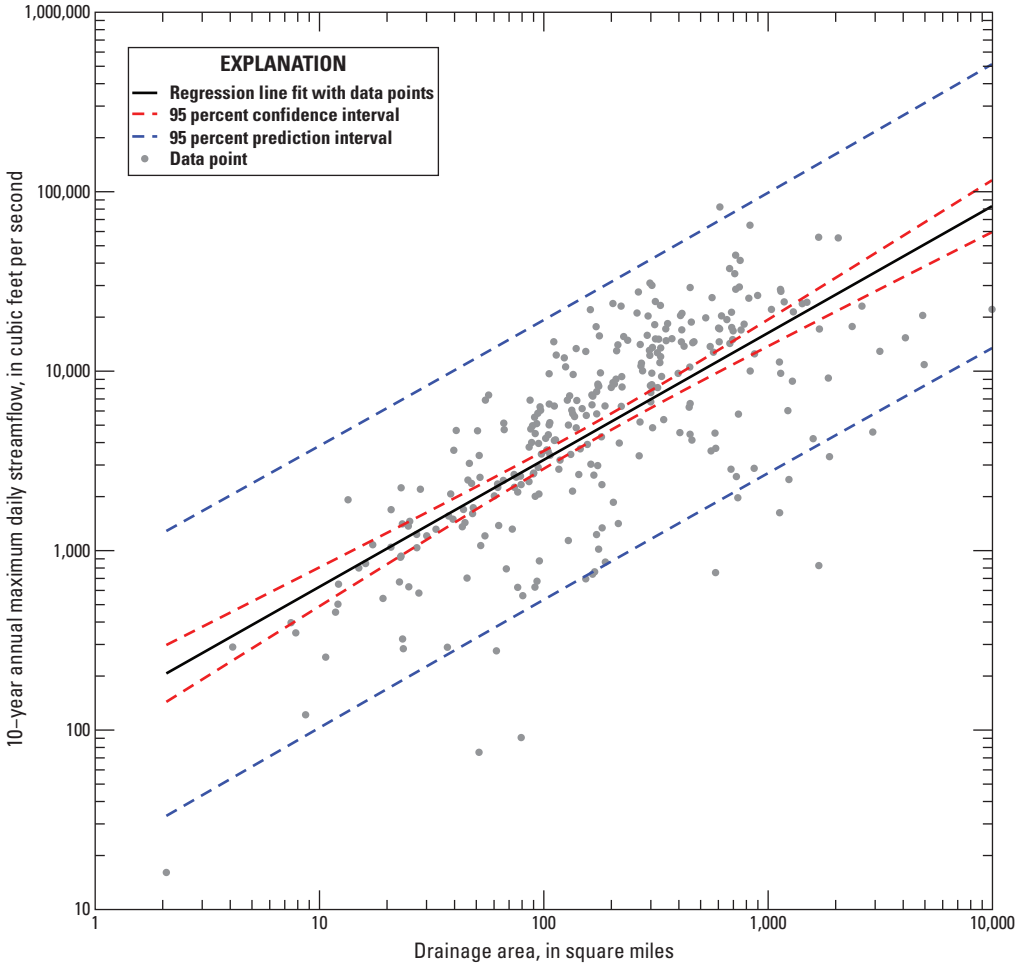

Figure 13. Graph showing a 95-percent prediction interval, in red, for the linear regression line, in black, based on data points, in gray, of the regression of logarithmically transformed 10-year annual maximum daily streamflow and the logarithm of drainage area. The dashed blue lines show the 95-percent prediction interval, which contains approximately 95 percent of the data points.

estimated from the regression by using the same observations that were used to develop the regression can be misleading. What needs to be known is how well the model is likely to perform when used to predict the response variable at sites not used for model development. Although tools such as the standard error of prediction try to answer this question theoretically based on the existing data, cross-validation procedures allow analysts to begin to answer this question more directly.

Cross validation of a model refers to the process where a portion of the data is set aside, and the remaining observations are used to estimate the parameters of the model. The developed model is then used to predict the response variables in the set-aside observations. Cross validation treats the set-aside observations as separate from the regression development, loosely simulating the process of true unobserved prediction.

Cross validation requires splitting the dataset into groups. For each split, there is one group that is set aside and one group that is used for model calibration. The size of each of these groups can be any portion of the data. For example, the data could be randomly split into halves for a two-fold validation. Similarly, the data could be split into thirds, using two-thirds of the data to calibrate the model and predicting the remaining one-third. This three-fold validation would be dense because the calibration set is larger than the validation set. When conducting a data split, it is important to consider random samples to avoid sampling bias in the calibration or validation set.

Another cross-validation scheme is the leave-one-out validation, also known as the remove-one or *N*-fold validation. Because this approach uses a calibration set that consists of all but one observation, the calibration set most closely represents

the full sample. By considering each observation in turn, it is possible to assemble a complete set of truly predicted values, $\check{Y}_i$. These predictions can then be used to compute a prediction residual error sum of squares,

$$SS_P = \sum_{i=1}^{N}\left(Y_i - \check{Y}_i\right)^2, \tag{51}$$

which is also commonly abbreviated as the PRESS statistic. By analogy, the prediction error sum of squares can be combined with the total sum of squares to produce a prediction coefficient of determination, R^2_{pred}, or a cross-validated coefficient of determination such that

$$R^2_{pred} = 1 - \frac{SS_P}{SS_A}\,. \tag{52}$$

Although other metrics are required for proper assessment, the prediction error sum of squares and the prediction coefficient of determination are some of the best measures of predictive performance.

Model Refinements and Other Issues

The development of a multiple linear regression model can be a highly iterative process. Only by considering several sets of explanatory variables and proceeding through some degree of performance evaluation is it possible to develop an optimal model. After evaluating a specific model, there are countless modifications that can be made to improve model parameterization. Some methods are discussed in this section, including additional transformations and subregionalization. This section also discusses the steps to take when model development reveals problematic conditions, such as the presence of trends or censored values. Because least-squares regression may not be the best method to use, the section also discusses several available alternatives.

Variable Transformation and Retransformation Bias

Variable transformation was discussed at length in the "Exploratory Data Analysis" section. However, after fitting and evaluating a regression model, new tools are available for improved variable transformation. Partial-regression plots, as discussed in the "Variable Selection" section, are immensely useful for discovering advantageous variable transformations. Caution should be used when considering transformations of the response variable because, after estimation, retransformation of the response variable can introduce significant bias. Beauchamp and Olson (1973) provide a seminal discussion of transformation bias for logarithmic transformations. The bias happens when values above and below the predicted response variable are not equally probable. The need for a bias correction is largely a function of intended use and should be considered based on the acceptability of an asymmetrical distribution around the prediction.

Subregionalization

Multiple linear regression proceeds by accounting for the linear heterogeneity of the observations represented by the explanatory variables. However, unless the underlying physical processes are purely linear, it often is not possible to separately consider all sources of variation. In such cases, further partitioning may control for the heterogeneity in the underlying data or processes. When considering spatially distributed observations, such as streamgages, this partitioning is called subregionalization. However, partitioning does not need to be geographically based and several advanced tools exist for identifying subregions on other bases.

Subregionalization can improve the predictive accuracy and precision of regression models by grouping hydrologically similar observations together. If a regression model is developed for an entire set of observations, the model's results can be used to evaluate if

subregionalization may be helpful. Although not the only approach, an analysis of the residuals from the domain-wide regression model can suggest where subregional models may prove advantageous. In such cases, data visualization and maps can be useful. However, it should be noted that subregionalization introduces additional parameters into the model.

After defining groups of observations, indicator regression can be used to simultaneously fit regression models to all groups. Indicator regression relies on a set of binary variables that indicate the membership of each group. Consider a bivariate relationship between two arbitrary variables Y and X with a set of observations that can be divided into three groups by underlying soil types, producing regions A, B, and C. Indicator regression requires defining two binary variables, $I_{A,i}$, which takes the value 1 if observation i is contained in region A and a value of 0 if the observation is contained in another region, and $I_{B,i}$, which takes the value 1 if observation B is contained in region B and a value of 0 if the observation is in another region. In this example, a third indicator variable is not necessary because the regions are mutually exclusive. The regressions for each region can be estimated simultaneously as

$$\boldsymbol{Y} = \beta_{0,C} + \nu_{0,A}\boldsymbol{I_A} + \nu_{0,B}\boldsymbol{I_B} + \beta_{1,C}\boldsymbol{X} + \nu_{1,A}\boldsymbol{I_A}\boldsymbol{X} + \nu_{1,B}\boldsymbol{I_B}\boldsymbol{X} \; . \tag{53}$$

The values of β and ν can be estimated by standard regression techniques. Careful inspection reveals that, if the observation is contained in region A, the regression devolves to

$$\boldsymbol{Y} = \beta_{0,C} + \nu_{0,A} + \beta_{1,C}\boldsymbol{X} + \nu_{1,A}\boldsymbol{X} = \left(\beta_{0,C} + \nu_{0,A}\right) + \left(\beta_{1,C} + \nu_{1,A}\right)\boldsymbol{X} = \beta_{0,A} + \beta_{1,A}\mathbf{X} \, . \tag{54}$$

If the observation exists within region B, the regression devolves to

$$\boldsymbol{Y} = \beta_{0,C} + \nu_{0,B} + \beta_{1,C}\boldsymbol{X} + \nu_{1,B}\boldsymbol{X} = \left(\beta_{0,C} + \nu_{0,B}\right) + \left(\beta_{1,C} + \nu_{1,B}\right)\boldsymbol{X} = \beta_{0,B} + \beta_{1,B}\mathbf{X} \, . \tag{55}$$

Finally, if the observation is categorized in region C, the regression devolves to

$$Y = \beta_{0,C} + \beta_{1,C}\mathrm{X} \, . \tag{56}$$

Indicator regression allows for significance testing, so it can be used to determine if subregionalization is appropriate for a specific situation. In practice, it is often assumed that indicator variables only affect the intercept. However, interaction terms, those binary variables that are multiplied with other explanatory variables to produce alternate slopes, should also be considered (see Helsel and Hirsch, 2002) even though the interaction terms may or may not be significant. The example shows that subregionalization effectively introduces $(M + 1)(N_R - 1)$ new variables, where N_R represents the number of regions. Recall that it was suggested that a minimum of 10 observations be available to support the addition of any explanatory variable or requisite coefficient. Clearly, extensive use of indicator regression can quickly exhaust the degrees of freedom in the dataset.

In lieu of indicator regression or subregionalization, a region-of-influence regression may prove useful. Region-of-influence regression defines, based on geographic, physiographic, or some other measure of proximity, observation-specific regions for each observation. This observation-specific region is then used to develop an observation-specific regression. In hydrology, these are typically site-specific, proximate regions. Further examples are provided by Burn (1990), Tasker and others (1996), and Eng and others (2005, 2007).

Trends

Multiple linear regression produces conditional estimates that are specific to a particular timeframe unless a representation of time is considered explicitly and tools for temporal regression are applied. For example, records of average rainfall from 1971 to 2000 may not provide accurate predictions of average streamflows in 2020. Because temporal structure is not addressed in standard regression models, it is extremely important to detect and account for any temporal trends or nonstationarity in the underlying dataset. In the regression of hydrologic frequency statistics, temporal structure is typically represented by trends in the streamflow time

series used to compute frequency statistics. Although beyond the scope of this report, the presence of trends can be tested for using several tests; Helsel and Hirsch (2002) suggest using either a Kendall's Tau test or a Wilcoxon ranked-sum test to determine if trends are present.

Observations with trends may need to be treated differently than observations without trends depending on the purpose of the analysis and the suspected cause of the trends. Natural variability in meteorological conditions can cause streamflow records to exhibit trends for short periods of time. An observation with a trend that is related to variability in precipitation should remain in the analysis because the trend is a result of natural variability. However, an observation with a trend that is the result of anthropogenic or nonnatural factors should be removed because the observation may confound the regression analysis. It may be possible to apply the statistical process of detrending to the data. However, if the aim of the analysis is not to detect purely natural conditions, neither trend may be cause for concern. Regardless, trends should be documented and discussed in all reports of regression-based analyses.

Model Consistency

In streamflow frequency analyses, regression models are often developed for several different streamflow statistics, such as events with a 1-, 5-, and 50-percent exceedance probability. If each event is considered independently, the resulting regression may produce nonintuitive results, such as a 5-percent event that is greater than a 1-percent event when the latter should be greater. Such nonintuitive results are more likely to happen when one or more explanatory variable is near the limit of the values for the sites used in the regression analysis. Testing should be done using hypothetical values for the explanatory variables to determine if nonintuitive results are obtained. One way to prevent the production of nonintuitive results is to have all the models contain the same explanatory variables. Although constraining the models will not guarantee consistent results, it is less likely that problematic estimates will be produced. This approach may cause some loss of predictive capability or parameter significance, which should be assessed in conjunction with the overall predictability across all events. Constraining the models is a common approach, but the argument can be made that different events are controlled by different processes. Therefore, the analyst needs to decide how to proceed but all justifications should be documented with appropriate evidential support.

Zero-Valued Response Variables

In hydrology, streamflow statistics sometimes equal zero. In many traditional regression-based analyses, using streamflow statistics that equal zero is problematic because the statistics are often logarithmically transformed and the logarithm of zero is undefined. Some simple, but not ideal, approaches to dealing with zero flows have included avoiding logarithmic transformations, omitting zero-valued observations, and augmenting the observations by some small, positive value. However, each of these approaches have problems.

In most cases, failing to consider a particular transformation is not possible because the relationship between the response variable and the explanatory variable may require a particular transformation for linearity. Failing to treat this nonlinearity precludes the use of the linear regression techniques discussed in this report. Omitting zero-valued observations has been used in some cases, but this approach biases the regression because the omission misrepresents the lower tail of the distribution of the response variable and limits the range of applicability of the equations. Adding a small, positive correction to all values allows for the use of the logarithmic transformation, but the selection of that constant is problematic and can substantially influence results.

One viable approach to the handling zero-valued observations is to use censored, or Tobit, regression (Helsel and Hirsch, 2002). Censored regression argues that zero-valued observations are near-zero-valued observations that represent a measurement below some censoring limit rather than true zero. Censored regression develops a linear regression on values above (for right-censored data) or below (for left-censored data) the user-provided threshold. Right-censoring refers to the situation where values above a threshold are censored and left-censoring refers to the situation where values less than a specified threshold are censored. When the exact value of very small observations is effectively unimportant, left-censored regressions can be utilized for regression-based regionalization studies. Because zero-valued observations are below the specified threshold, they do not need to be logarithmically transformed and the regression analysis can proceed. However, rather than relying on least-squares regression, censored regression uses a method of maximum-likelihood estimator.

In a hydrologic application, Kroll and Stedinger (1999) tested regression models that left out stations with zero-valued statistics, added a constant to all flows, and used a censored regression. The results suggest that censored regression is preferable to the other options discussed by Kroll and Stedinger (1999). Generally, censored regression is most appropriate when the proportion of values below the censored threshold is not large. Helsel and Hirsch (2002, p. 375) suggest that censored regression is appropriate for small to moderate amounts of censoring. Hirsch and others (1993, p. 17.50) suggest that censored regression is appropriate if censoring does not exceed 50 percent of the observations.

Another viable approach to handling zero-valued observations is to develop a two-step regression. In the first step, a logistic regression is applied to quantify the probability that an observation is equal to zero. If the logistic regression indicates the likelihood of a nonzero observation, then a multiple linear regression model is used to estimate the observation. This linear regression is set up using only nonzero-valued observations. Logistic regression, as the name implies, first attempts to fit observations as binary data, either nonzero-valued or zero-valued, to the logistic function rather than a straight line.

Logistic regression is most appropriate when there are many zero-valued observations. If there are not many zero-valued observations, it is not practical to fit the logistic function to the data. Helsel and Hirsch (2002, p. 375) suggest that logistic regression is appropriate for moderate to large frequencies of censoring. Hirsch and others (1993, p. 17.50) suggest that logistic regression is appropriate when there is at least 20 percent censoring of observations. Both Hirsch and others (1993) and Helsel and Hirsch (2002) discuss censoring.

Alternatives to Least-Squares Regression and Linear Fitting

Least-squares regression is not the only tool for modeling relationships, linear or otherwise. Although beyond the scope of this report, it is important to note that alternatives to the methods presented in this report are available and constantly evolving and new and novel models are constantly being proposed. With respect to linear relationships, there are several method-of-moments and method-of-maximum-likelihood estimators. Another method, called robust regression, is less sensitive to individual outliers, and in that sense the solution is more "robust." Helsel and Hirsch (2002) suggest that the Kendall Theil Robust Line is an alternative to censored regression for regressions involving low censoring. In Europe, a common approach to regionalization is to scale events by some index value, which in hydrology is known as the index flood method. Beyond the realm of linear fitting, artificial neural networks, regression trees, random forests, factor analysis, and principal components analysis have all proven useful for regionalization studies in hydrology. Given the wide range of analytical methods, this report makes no effort to assess the appropriateness of alternatives to least-squares regression.

Model Application and Documentation

An appropriately developed regression model can be used to estimate unobserved response variables. In hydrology, model application typically involves the estimation of streamflow statistics at ungaged points along a stream network or predicting natural streamflow characteristics at gaged watersheds that have undergone human modifications. However, such applications rely on the validity of the assumptions used in model development and the appropriate characterization of regression uncertainty.

When applying a regression model for prediction, it is important to keep in mind that the predictions are only as good as the underlying data and assumptions. Extrapolation beyond the range of the explanatory variables used for regression development is strongly discouraged. Outside of this range, estimates and their estimated uncertainties are unreliable. If sample observations used in the model are not representative of conditions at the target sites at which statistics are going to be predicted, the model should not be applied. Furthermore, if the underlying assumptions of the distributions of residuals or other diagnostics are invalid, the predictions are also unreliable.

In some hydrologic applications, an improved estimate of a streamflow statistic may be obtained by weighting at-site estimates, regression estimates, or scaled estimates. Bulletin 17C (England and others, 2018) describes a weighting procedure for blending regional estimates and at-site estimates. The manual of the National Streamflow Statistics Program (Ries, 2007) describes tools for scaling proximate estimates by upstream drainage area that are particularly useful for ungaged sites immediately upstream or downstream from a gage.

The culmination of any regression-based regionalization study is complete documentation of data collection and exploration and model development, evaluation, and application. Complete documentation is essential for reproducible science. A report on the model development allows users of the regression model to assess the suitability of the model for their purposes and provides a useful record of the analysis for future updates. Although it is not necessary to exhaustively document every aspect of data and model exploration, it is important to document all computation and analysis methods and justify all conclusions. All data used to develop the models should be published and archived, including geographic information system data used to compute the basin characteristics. In addition, guidance and examples should be provided on how users can apply the models to obtain estimates at ungaged locations.

Conclusions

Multiple linear regression is a powerful tool that can be used to quantify the relationships among the response variable and the explanatory variables. In practice, this tool aids in the regionalization of surface-water statistics including streamflow frequency statistics. Some regression-based regionalization studies may require an alternative approach (for example, in the presence of highly nonlinear processes), but the essential steps of a standard regression-based regionalization study are

- *Project outline*.—Determine the questions that are to be assessed and the variablthanes of interest.
- *Data collection and quality assurance*.—Gather data on the selected variables of interest, taking care to note, address, and document any concerns with the quality of the underlying data.
- *Exploratory data analysis*.—Develop an insight of the data by evaluating the summary statistics, distributions, and relationships among the different variables.
- *Candidate model development*.—Develop several candidate regression models based on the information developed through exploratory data analysis. Based on the project goals and the characteristics of the data, use appropriate tools to estimate model parameters and evaluate model performance.

- *Finalized model development*.—Based on evidence derived from exploratory data analysis and the development of candidate models, arrive at an appropriate and defensible model of the system in question.
- *Model application*.—If necessary, implement the final model to meet the project goals.
- *Documentation*.—Document all aspects of the project, from goal setting through candidate model evaluation to model application. Although exhaustive documentation of all intermediate steps is not necessary, the final documentation should present a defensible and logical narrative supporting the acceptance of the final model and a cknowledge any limitations encountered along the way.
- *Archive*.—All data and model formulations included in the documentation should be archived in a stable format so that the study can be reproduced in the future.

References Cited

Aitchison, J., and Brown, J.A.C., 1957, The lognormal distribution: Cambridge, England, Cambridge University Press, 176 p.

Allison, P.D., 1999, Multiple regression—A primer: Thousand Oaks, Calif., Pine Forge Press, 202 p.

Beauchamp, J.J., and Olson, J.S., 1973, Corrections for bias in regression estimates after logarithmic transformation: Ecology, v. 54, no. 6, p. 1403–1407, accessed June 6, 2019, at https://doi.org/10.2307/1934208.

Benson, M.A., and Carter, R.W., 1973, A national study of the streamflow data-collection program: U.S. Geological Survey Water-Supply Paper 2028, 44 pp. [Also available at https://doi.org/10.3133/wsp2028.]

Box, G.E.P., and Cox, D.R., 1964, An analysis of transformations: Journal of the Royal Statistical Society, Series B, v. 26, no. 2, p. 211–252, accessed June 6, 2019, at http://www.jstor.org/stable/2984418.

Burn, D.H., 1990, Evaluation of regional flood frequency analysis with a region of influence approach: Water Resources Research, v. 26, no. 10, p. 2257–2265, accessed June 6, 2019, at https://doi.org/10.1029/WR026i010p02257.

Cook, R.D., 1977, Detection of influential observation in linear regression: Technometrics, v. 19, no. 1, p. 15–18, accessed June 6, 2019, at https://doi.org/10.2307/1268249.

Draper, N.R., and Smith, H., 1981, Applied regression analysis, second edition: Hoboken, N.J., John Wiley and Sons, 709 p.

Eng, K., Chen, Y., and Kiang, J.E., 2009, User's guide to the weighted-multiple-linear-regression program (WREG version 1.0): U.S. Geological Survey Techniques and Methods, book 4, chap. A8, 21 p. [Also available at http://pubs.usgs.gov/tm/tm4a8.]

Eng, K., Milly, P.C.D., and Tasker, G.D., 2007, Flood regionalization—A hybrid geographic and predictor-variable region-of-influence regression method: Journal of Hydrologic Engineering, v. 12, no. 6, p. 585–591, accessed June 6, 2019, at https://doi.org/10.1061/(ASCE)1084-0699(2007)12:6(585).

Eng, K., Tasker, G.D., and Milly, P.C.D., 2005, An analysis of region-of-influence methods for flood regionalization in the Gulf-Atlantic Rolling Plains: Journal of the American Water Resources Association, v. 41, no. 1, p. 135–143, accessed June 6, 2019, at https://doi.org/10.1111/j.1752-1688.2005.tb03723.x.

England, J.F., Jr., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Jr., Veilleux, A.G., Kiang, J.E., and Mason, R.R., Jr., 2018, Guidelines for determining flood flow frequency—Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p., accessed June 6, 2019, at https://doi.org/10.3133/tm4B5.

Falcone, J.A., comp., 2011, GAGES-II—Geospatial attributes of gages for evaluating streamflow [digital spatial dataset], U.S. Geological Survey, accessed June 6, 2019, at http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml.

Farmer, W.H., 2019, An example dataset for exploration of multiple linear regression: U.S. Geological Survey data release, https://doi.org/10.5066/P9T5ZEXV.

Funkhouser, J.E., Eng, K., and Moix, M.W., 2008, Low-flow characteristics and regionalization of low-flow characteristics for selected streams in Arkansas: U.S. Geological Survey Scientific Investigations Report 2008–5065, 161 p. [Also available at http://pubs.usgs.gov/sir/2008/5065.]

Griffis, V.W., and Stedinger, J.R., 2007, The use of GLS regression in regional hydrologic analyses: Journal of Hydrology, v. 344, p. 82–95, accessed June 6, 2019, at https://doi.org/10.1016/j.jhydrol.2007.06.023.

Griffis, V.W., and Stedinger, J.R., 2009, Log-Pearson type 3 distribution and its application in flood frequency analysis. III—Sample skew and weighted skew estimators: Journal of Hydrologic Engineering, v. 14, no. 2, p. 121–130, accessed June 6, 2019, at https://doi.org/10.1061/(ASCE)1084-0699(2009)14:2(121).

Griffis, V.W., Stedinger, J.R., and Cohn, T.A., 2004, Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments: Water Resources Research, v. 40, W07503, 17 p., accessed June 6, 2019, at https://doi.org/10.1029/2003WR002697.

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water Resources Investigations, Book 4, chapter A3, 523 pages. [Also available at https://doi.org/10.3133/twri04A3.]

Hirsch, R.M., 1982, A comparison of four streamflow record extension techniques: Water Resources Research, v. 18, no. 4, p. 1081–1088, accessed June 6, 2019, at https://doi.org/10.1029/WR018i004p01081.

Hirsch, R.M., Helsel, D.R., Cohn, T.A., and Gilroy, E.J., 1993, Statistical treatment of hydrologic data, *in* Maidment, D.R., ed., Handbook of hydrology: New York, McGraw-Hill Inc., p. 17.1–17.55.

Kennard, M.J., Mackay, S.J., Pusey, B.J., Olden, J.D., and Marsh, N., 2010, Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies: River Research and Applications, v. 26, no. 2, p. 137–156, accessed June 6, 2019, at https://doi.org/10.1002/rra.1249.

Kiang, J.E., Stewart, D.W., Archfield, S.A., Osborne, E.B., and Eng, K., 2013, A national streamflow network gap analysis: U.S. Geological Survey Scientific Investigations Report 2013–5013, 79 p. plus one appendix as a separate file. [Also available at http://pubs.usgs.gov/sir/2013/5013.]

Kite, G.W., 1975, Confidence limits for design events: Water Resources Research, v. 11, no. 1, p. 48–53, accessed June 6, 2019, at https://doi.org/10.1029/WR011i001p00048.

Kite, G.W., 1976, Reply [to "Comment on 'Confidence limits for design events' by G.W. Kite"]: Water Resources Research, v. 12, no. 4, p. 826, , accessed June 6, 2019, at https://doi.org/10.1029/WR012i004p00826.

Kroll, C.N., and Song, P., 2013, Impact of multicollinearity on small sample hydrologic regression models: Water Resources Research, v. 49, no. 6, p. 3756–3769, accessed June 6, 2019, at https://doi.org/10.1002/wrcr.20315.

Kroll, C.N., and Stedinger, J.R., 1999, Development of regional regression relationships with censored data: Water Resources Research, v. 35, no. 3, p. 775–784, accessed June 6, 2019, at https://doi.org/10.1029/98WR02743.

Mauget, S.A., 2003, Multidecadal regime shifts in U.S. streamflow, precipitation, and temperature at the end of the twentieth century: Journal of Climate, v.16, p. 3905–3916, accessed June 6, 2019, at https://doi.org/10.1175/1520-0442(2003)016%3C3905:MRSIUS%3E2.0.CO;2.

McCabe, G.J., and Wolock, D.M., 2014, Spatial and temporal patterns in conterminous United States streamflow characteristics: Geophysical Research Letters, v. 41, no. 19, p. 6889–6897, accessed June 6, 2019, at https://doi.org/10.1002/2014GL061980.

McCuen, R.H., and Levy, B.S., 2000, Evaluation of peak discharge transposition: Journal of Hydrologic Engineering, v. 5, no. 3, p. 278–289, accessed June 6, 2019, at https://doi.org/10.1061/(ASCE)1084-0699(2000)5:3(278).

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2006, Introduction to linear regression analysis (4th ed.): Hoboken, N.J., John Wiley and Sons, 640 p.

Mosteller, F., and Tukey, J.W., 1977, Data analysis and regression—A second course in statistics: Reading, Mass., Addison-Wesley, 608 p.

Myers, R.H., 1990, Classical and modern regression with applications (2d ed.): Boston, Mass., PWS-Kent Publishing Company, 488 p.

Parrett, C., Veilleux, A., Stedinger, J.R., Barth, N.A., Knifong, D.L., and Ferris, J.C., 2011, Regional skew for California, and flood frequency for selected sites in the Sacramento–San Joaquin River Basin, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2010–5260, 94 p. [Also available at http://pubs.usgs.gov/sir/2010/5260.]

Ries, K.G., III, comp., 2007, The National Streamflow Statistics Program—A computer program for estimating streamflow statistics for ungaged sites: U.S. Geological Survey Techniques and Methods book 4, chap. A6, 37 p. (Also available at http://pubs.usgs.gov/tm/2006/tm4a6.)

Riggs, H.C., 1972, Low-flow investigations: U.S. Geological Survey Techniques of Water Resources Investigations, book 4, chap. B1, 18 p. [Also available at http://pubs.er.usgs.gov/publication/twri04B1.]

Sauer, V.B., 1974, Flood characteristics of Oklahoma streams—Techniques for calculating magnitude and frequency of floods in Oklahoma, with compilations of flood data through 1971: U.S. Geological Survey Water-Resources Investigations Report 73–52, 306 p. [Also available at https://doi.org/10.3133/wri7352.]

Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis—1. Ordinary, weighted, and generalized least squares compared: Water Resources Research, v. 21, no. 9, p. 1421–1432, accessed June 6, 2019, at https://doi.org/10.1029/WR021i009p01421.

Stedinger, J.R., and Tasker, G.D., 1986, Regional hydrologic analysis—2. Model-error estimators, estimation of sigma and log-Pearson type 3 distributions: Water Resources Research, v. 22, no. 10, p. 1487–1499, accessed June 6, 2019, at https://doi.org/10.1029/WR022i010p01487.

Stedinger, J.R., and Thomas, W.O., Jr., 1985, Low-flow frequency estimation using base-flow measurements: U.S. Geological Survey Open File Report 85–95, 22 p., accessed June 6, 2019, at https://doi.org/10.3133/ofr8595.

Tasker, G.D., 1975, Combining estimates of low-flow characteristics of streams in Massachusetts and Rhode Island: Journal of Research of the U.S. Geological Survey, v. 3, no. 1, p. 107–112.

Tasker, G.D., 1980, Hydrologic regression with weighted least squares: Water Resources Research, v. 16, no. 6, p. 1107–1113, accessed June 6, 2019, at https://doi.org/10.1029/WR016i006p01107.

Tasker, G.D., Hodge, S.A., and Barks, C.S., 1996, Region of influence regression for estimating the 50-year flood at ungaged sites: Journal of the American Water Resources Association, v. 32, no. 1, p. 163–170, accessed June 6, 2019, at https://doi.org/10.1111/j.1752-1688.1996.tb03444.x.

Tasker, G.D., and Stedinger, J.R., 1989, An operational GLS model for hydrologic regression: Journal of Hydrology, v. 111, p. 361–375, accessed June 6, 2019, at https://doi.org/10.1016/0022-1694(89)90268-0.

Tukey, J.W., 1977, Exploratory data analysis: Reading, Mass., Addison-Wesley, 500 p.

U.S. Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency, Bulletin 17-B of the Hydrology Subcommittee: Reston, Va., U.S. Geological Survey, Office of Water Data Coordination, 183 p.

Vogel, R.M., and Kroll, C.N., 1991, The value of streamflow record augmentation procedures in low flow and flood-flow frequency analysis: Journal of Hydrology, v. 125, no. 3-4, p. 259–276, accessed June 6, 2019, at https://doi.org/10.1016/0022-1694(91)90032-D.

Wasserstein, R.L. and Lazar, N.A., 2016, The ASA's statement on *p*-values—Context, process, and purpose: The American Statistician, v. 70, no. 2, p. 129–133, accessed June 6, 2019, at https://doi.org/10.1080/00031305.2016.1154108.

Publishing support provided by the Science Publishing Network, Denver Publishing Service Center

For more information concerning the research in this report, contact the
Director, Integrated Modeling and Prediction Division, Water Mission Area
US Geological Survey Mailstop 415
12201 Sunrise Valley Drive
Reston, VA 20192

Appendix 1. Glossary of Terms

Bias The difference between the expected value of an estimator and the true value; bias is often understood as the difference between the observed mean and the mean of estimated values. More generally, the tendency to overestimate or underestimate the value of a population parameter.

Correlation A measure of the degree to which two variables change together.

Covariance Similar to correlation, a measure, although in the multiplicative units of the two variables, of the degree to which two variables change together.

Heteroscedasticity The state in which the variability of a variable is not constant across the range of values of a second variable used to predict it. The opposite of homoscedasticity.

Homoscedasticity The state in which the variability of a variable is constant across the range of values of a second variable used to predict it. The opposite of heteroscedasticity.

Population Often unmeasurable, the entire, exhaustive set of observations from which a sample is drawn.

Residual error The difference between an observed value of the response variable and its estimate.

Sample A finite set of observations drawn from the larger population.

Spurious correlation A false correlation between two variables that can be attributed to coincidence or to another factor that is not apparent at the time of examination.

Statistical independence Two observations or variables are considered to be statistically independent if the value of one has no effect on the probability distribution of the other. That is, a specific realization of an observation or variable does not make the second observation or variable more or less likely to take on any specific value.

Statistical significance A result is considered statistically significant if it can be demonstrated that there exists only a small probability that the given result could have happened by random chance. Significance is only meaningful in the context of a statistical test of null and alternative hypotheses.

Variance The second central moment of a dataset, a measure of variability, taken as the expectation of the squared deviation of a random variable from its mean.

Appendix 2. Glossary of Symbols

Table 2.1. Glossary of symbols.

[Shading indicates not applicable. MSE, mean squared error]

General symbol	Specific realizations	Description	Equation number(s)
Variables			
Ch		A coefficient used to determine the critical value of leverage.	34
D		Cook's D, a measure of influence	35, 36
	D_{limit}	The critical value of Cook's D, above which observations are considered to be particularly influential.	36
GLS		Generalized least squares	24
$GLS-skew$		Generalized least squares with an adjustment for uncertain skewness	27
$\boldsymbol{I_A}$ and $\boldsymbol{I_B}$		Binary indicator variables that take values of 1 only if the observation is in the subregion defined by the subscript.	53
$\boldsymbol{K}$		The deviate of the log-Pearson Type III distribution that is used to estimate the response variable for each observation.	24, 27, 28
	$\overline{K}$	The average log-Pearson Type III deviate used across all observations	19
L		A matrix used for the calculation of Cook's D	35
M		The number of explanatory variables used in a model	1, 2, 3, 26, 35, 38, 43, 45, 48, 50
MSE		MSE—When not indexed or indexed with i, this symbol refers to the mean square residual from the regression.	45, 46, 47, 49
	$\sqrt{MSE}_{\%}$	The root MSE expressed as a percentage	46
	MSE_{g_i}	The estimated MSE of the at-gage estimate of skewness. The estimate is gage specific, so it is indexed with i.	21, 22
	$MSE_{g_{reg}}$	The estimated MSE of the regional estimate of skewness	21
N		The number of observations available for model fitting	8, 11, 12, 18, 26, 31, 34, 36, 38, 42, 43, 45, 48, 50, 51
	N_R	The number of regions used for subregionalization	
OLS		Ordinary least squares	16, 38, 47
R^2		The coefficient of determination	41
	R^2_{adj}	The coefficient of determination adjusted for the number of explanatory variables used in regression	43
	R^2_k	The coefficient of determination from a regression of explanatory variable k as predicted by all other variables	37

Table 2.1. Glossary of symbols.—Continued

[Shading indicates not applicable. MSE, mean squared error]

General symbol	Specific realizations	Description	Equation number(s)
	R^2_{pred}	An analog to the coefficient of determination based on the prediction residual sum of squares	52
	R^2_{pseudo}	An analog to the coefficient of determination based on the relative improvement in the model error variances	44
SS		Sum of squared errors	
	SS_A	Total or sum of all squares	41, 42, 43, 52
	SS_P	Prediction residual sum of squares (PRESS)	51, 52
	SS_ε	Residual sum of squares	8, 38, 41, 43, 45
VIF		Variance inflation factor, typically indexed with *k*	37
WLS		Weighted least squares	17
$\boldsymbol{X}$		The matrix explanatory variables with the number of rows (*N*) rows and *M* columns. Rows represent unique observations (for example, sites) and columns represent individual explanatory variables. The sites are typically indexed with *i* of *j*, but the variables are typically indexed with *k*.	1, 2, 3, 9, 10, 11, 12, 13, 26, 33, 38, 47, 48, 49, 53, 54, 55, 56
$\boldsymbol{Y}$		The response variable, which is typically represented as a vector of observations, and indexed with *i*.	1, 2, 4, 6, 8, 9, 11, 12, 13, 14, 26, 42, 51, 53, 54, 55, 56
	$\hat{Y}_{i,pred,upper\|\boldsymbol{X}_i}$ $\hat{Y}_{i,pred,lower\|\boldsymbol{X}_i}$	The upper and lower limits of a prediction interval around the regression estimate of *Y*	50
	$\hat{Y}_{i,upper\|\boldsymbol{X}_i}$ $\hat{Y}_{i,lower\|\boldsymbol{X}_i}$	The upper and lower limits of a confidence interval around the regression estimate of *Y*	48
	$\check{Y}_i$	The estimate of Y_i from a regression built on all observations except Y_i	51
	$\overline{Y}$	The mean of all observations of the response variable	42
	$\hat{Y}$	A vector of estimates of the response variable, typically indexed with *k*	3, 4, 5, 8, 10, 47, 48, 49, 50
	$\tilde{Y}$	A vector of the unobservable true value of the response variable, typically indexed with *k*	5, 6
Z_p or Z_{pr}		The standard normal quantile with the subscripted nonexceedance probability	28, 32
c		This symbol is used to represent a set of generic coefficients used in numerical computation.	
	c_1		17, 18, 19
	c_2		22, 23, 29
	c_3		22, 23

Table 2.1. Glossary of symbols.—Continued

[Shading indicates not applicable. MSE, mean squared error]

General symbol	Specific realizations	Description	Equation number(s)
	c_4		22, 23
	c_5		29, 30
	c_6		29, 30
d_{ij}		The separation distance between the two subscripted sites	25
m		This symbol is used to indicate the length of the data record.	
	$m_{i,j}$	The length of the overlapping period of record between the subscripted sites	24, 27
	m_i	The record length at the subscripted site	17, 18, 22, 23, 24, 27, 29, 30
p		This symbol is typically reserved for probabilities	28
	p_{β_k}	The probability that the estimate of the subscripted coefficient would be observed if the true value of the coefficient were zero	40
	p_r	Nonexceedance probability of the subscripted rank	31, 32
r		The rank of a residual error	31, 32
t		A Student's t statistic	40
	$t_{(\alpha/2),(N-M-1)}$	A Student's t statistic with an exceedance probability of $\frac{\alpha}{2}$ and $(N-M-1)$ degrees of freedom	48, 50
	t_{β_k}	The Students t statistic associated with the estimated value of the subscripted regression coefficient. The degrees of freedom are taken to be $(N-M-1)$	39, 40
$\boldsymbol{g}$		The at-gage skewness of the log-Pearson Type III distribution used to estimate the response variable for each observation; typically indexed with i	20, 22
	$\overline{g_w}$	The average log-Pearson Type III weighted skewness used across all observations	19
	$\boldsymbol{g_{Reg}}$	The regional log-Pearson Type-III skewness used at each site	20, 27, 28, 29
	$\boldsymbol{g_w}$	The weighted skewness of the log-Pearson Type III distribution used to estimate the response variable for each observation; typically indexed with i	20, 24
$\boldsymbol{h}$		A square matrix whose main diagonal elements represent the leverage of each observation	33, 34
	h_{limit}	The critical value of leverage above which observations are considered to have high leverage	34
$\boldsymbol{\Lambda}$		An N-by-N matrix of weights on the observations of Y, typically indexed by i and j. This symbol is also commonly called a covariance matrix.	14, 33, 35, 38, 47

Table 2.1. Glossary of symbols.—Continued

[Shading indicates not applicable. MSE, mean squared error]

General symbol	Specific realizations	Description	Equation number(s)
	$\hat{\Lambda}$	An estimate of the weighting matrix	17, 24, 26, 27
α		A probability used for assessment of statistical results, often known as the significance level	48, 50
$\boldsymbol{\beta}$		Coefficients describing the linear relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, typically indexed with k Without any modification, this symbol usually represents the true values that cannot be observed.	1, 38, 39, 40
	$\beta_{0,A}$ and $\beta_{1,A}$ $\beta_{0,B}$ and $\beta_{1,B}$ $\beta_{0,C}$ and $\beta_{1,C}$	Regression coefficients derived for subregionalization using indicator regression	53, 54, 55, 56
	β_0 or $\hat{\beta}_0$	A coefficient with the reserved subscript to define the constant additive term of the model	1, 2, 3, 9, 10, 11
	$\hat{\boldsymbol{\beta}}$	An estimation of the linear coefficients, typically indexed with k	2, 3, 11, 12, 13, 14, 26, 39
$\boldsymbol{\delta}$		A vector of model errors from a fitted regression, typically indexed with i	5, 7, 15, 17, 18, 24, 27, 49
$\boldsymbol{\varepsilon}$		A vector of residual errors from a fitted regression, typically indexed with i	2, 4, 7, 8, 13, 15, 18, 35
	$\hat{\varepsilon}$	The theoretical residual based on the assumptions of normality	32
	$\overline{\varepsilon}$	The average residual	32
$\boldsymbol{\eta}$		A vector of sampling errors for a particular statistic, typically indexed with i	6, 7, 15
θ_1 and θ_2		Fitting coefficients used to approximate inter-site correlation based on separation distance	25
$\hat{\rho}$		Approximated cross-correlation among the time series of streamflow used to calculate the streamflow statistic, typically indexed with i and j	24, 25, 27
σ^2		This symbol is used to indicate some sort of variance, although the square root is used to indicate a standard deviation.	
	$\hat{\sigma}^2_{\beta_k}$	Variance of the estimated value of the subscripted regression coefficient	38, 39
	$\overline{\sigma}^2$	The squared arithmetic average of the standard deviation of the annual time series of streamflow used to compute the response variable at each site	19
	$\sigma^2_{\hat{Y}_i, pred\|\boldsymbol{X}_i}$	The conditional prediction variance on $\hat{Y}$ given a particular observation of X	49, 50
	$\sigma^2_{\hat{Y}_i\|\boldsymbol{X}_i}$	The conditional variance on $\hat{Y}$ given a particular observation of X	47, 48, 49
	$\sigma^2_{g_i}$	The approximated variance of the at-gage skewness	27, 29
	σ^2_i	The standard deviation of the annual time series of streamflow used to compute the response variable at site i	24, 27

Table 2.1. Glossary of symbols.—Continued

[Shading indicates not applicable. MSE, mean squared error]

General symbol	Specific realizations	Description	Equation number(s)
	$\sigma^2_{\delta\|0}$	The modeling error variance for a model with no explanatory variables, using only a regression constant	44
	$\sigma^2_{\delta\|M}$	The modeling error variance for a model with M variables	44
	σ^2_{δ}	The modeling error variance	15, 17, 18, 24, 27, 49
	$\sigma^2_{\varepsilon\|OLS}$	The modeling error variance derived from a model fit with ordinary least-squares regression	18
	σ^2_{ε}	The variance of residual errors	15, 32
	σ^2_{η}	The sampling error variance	15
$v_{0,A}$ and $v_{1,A}$ $v_{0,B}$ and $v_{1,B}$		Regression adjustments fitted through indicator regression for subregionalization	53, 54, 55
$\boldsymbol{\omega}$		The weights defined by MSE to compute weighted skewness values, typically indexed by i	20, 21, 27
Functions			
$\|\cdots\|$		The absolute value of the argument	40
$\frac{\partial \ldots}{\partial \ldots}$		The partial derivative of the numerator with respect to the denominator	27, 28
$\ldots\|\ldots$		Typically read as "given," implying that the preceding argument is realized based on the condition of the following argument	18, 44, 47, 48, 49
$Prob(\ldots)$		The probability of the argument	40
$ln(\ldots)$		The natural logarithm of the argument	46
$max(\ldots)$		The maximum value of the arguments	18
T		The transpose of the item being superscripted	14, 26, 33, 38, 47
e		A mathematical constant	46
Indices			
i		This index is reserved to point to different sites; as such, it runs from 1 to N.	8, 11, 12, 16, 17, 18, 24, 27, 47, 48, 49, 50
j		This index is reserved to point to different sites; as such, it runs from 1 to N.	16, 17, 24, 27
k		This index is reserved to point to different explanatory variables; as such, it runs from 1 to N.	37, 38, 39, 40
[		This subscript is used to indicate that all values in the row or column are being described.	1, 2, 3, 9, 10

ISSN 2328-7055 (online)
https://doi.org/10.3133/tm4A12