Hydrologic Regression and Network Analysis Using Program GLSNET

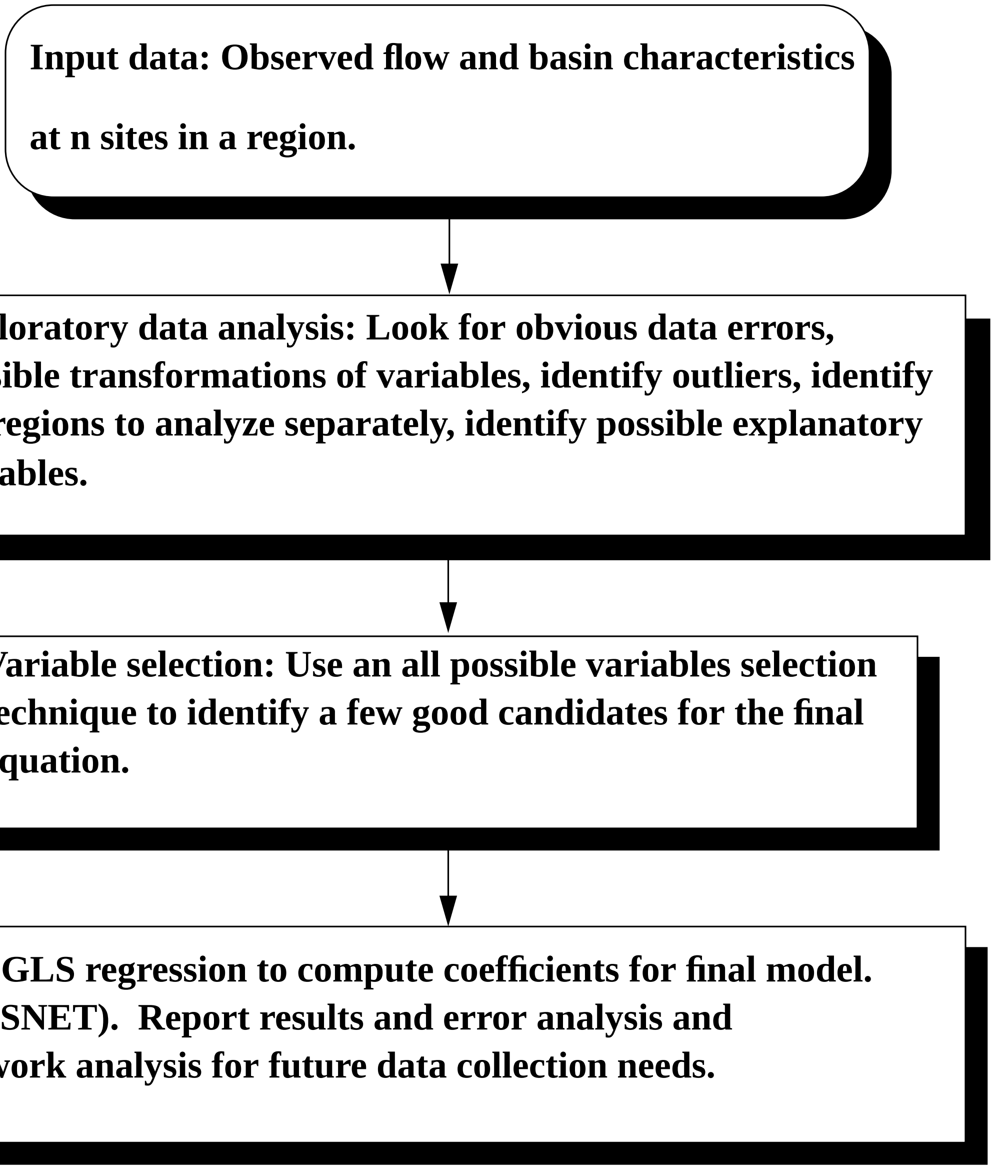
By G.D. Tasker, K.M. Flynn, A.M. Lumb, and W.O. Thomas, Jr.

U.S. GEOLOGICAL SURVEY
Water-Resources Investigations Report 95-_____

Reston, Virginia
1995

INTRODUCTION

In regional hydrologic regression we want to relate a flow characteristic, such as the 50-year peak or 7-day 10-year low flow, to basin characteristics that can be readily determined at an ungaged site. The purpose of the regression is usually to predict the flow characteristic at an ungaged site. The chart below gives a general idea of the steps needed to arrive at a final model.



Input data: Observed flow and basin characteristics at n sites in a region.

Exploratory data analysis: Look for obvious data errors, possible transformations of variables, identify outliers, identify subregions to analyze separately, identify possible explanatory variables.

Variable selection: Use an all possible variables selection technique to identify a few good candidates for the final equation.

Use GLS regression to compute coefficients for final model. (GLSNET). Report results and error analysis and network analysis for future data collection needs.

Purpose and Scope

The purpose of this document is to provide guidelines and examples of how to create a Watershed Data Management (WMD) file containing the necessary data to do a regional regression of a flow characteristic with basin characteristics and how to use GLSNET to perform a regional regression using the generalized-least squares method described in Tasker and Stedinger (1989). The next section describes how to create a Watershed Data Management (WDM) file for GLSNET. It is followed by a section that describes the GLS regression procedure. Finally, the procedure for analyzing the network for future data collection is described. Thus this document deals with the first and last box of the above chart. The second and third boxes of the chart can be approached using other software such as STATIT, SAS, or MINITAB.

DATA MANAGEMENT FOR GENERALIZED LEAST SQUARES (GLS)

WDM Files

A WMD file is used to store and manage the data required for the GLS analysis. The program IOWDM (Input and/or Output for a WDM file) is used to store basin characteristics, n-day high- and low-flow annual time series, and time series or tables of annual peak flows. The program ANNIE is used to add or modify basin characteristics, add or modify time-series data, and modify table data. The ANNIE program is also used to examine the contents of a WDM file. The ANNEX option in the GLSNET program is used to compute low flows at partial-record sites.

A WDM file is a binary, unformatted, direct-access file. It cannot be examined using a text editor. If the file is opened by an editor, it will be corrupted if it is saved or filed. Some editors will corrupt the file even when the file is not saved or filed. It is recommended that the suffix wdm be used in naming WDM files to make them easy to identify.

Data in a WDM file are arranged in data sets. A WDM file may contain a single data set or as many as 32,000 data sets. A data set contains a collection of data values, such as the annual time series of 7-day low flows at a station and selected attributes describing basin characteristics of the station and time series. There are over 300 attributes that can be used to describe a data set. See Appendix B, table B.1 for a detailed description of the available attributes and table B.2 for the relationship between WATSTORE basin and streamflow characteristics and WDM attributes.

Table 1 contains a summary of the steps for identifying and preparing the data for a GLSNET analysis. The first step is to identify the data needs. What are the boundaries of the study area? Which stations are required and which can be ignored? Which basin characteristics are required, and which basin characteristics may be important?

Table 1. Steps in identifying and preparing data for GLSNET

1. Identify data needs
   - area or region of interest
   - annual peaks
   - annual high and/or low flows
   - basin characteristics and attributes
2. Acquire data
   - use a consistent naming convention
   - WATSTORE
   - NWIS
   - other
3. Use IOWDM
   - build WDM file
   - input data
4. Use ANNIE
   - add additional attributes
   - list and plot attributes to examine and verify
   - list and plot time series to examine and verify
   - modify data as required

Identifying and Preparing Data for GLSNET

Step 2 is to acquire the data that has been identified. Using a consistent and descriptive naming convention will simplify the task of file management. Appendix C contains descriptions of the data formats that will be processed by IOWDM. Appendix D contains example Job Control Language (JCL) for retrieving data from the WATSTORE data base.

Input to a WDM File Using IOWDM

Step 3 is to use IOWDM to build the WDM file and enter the data. All of the stations to be analyzed by GLSNET will be stored in a single WDM file. Figure 1 contains an example of building a WDM file and adding basin characteristics and 7-day low flows to the file.

Adding Attributes Using ANNIE

Step 4 in identifying and preparing data for GLSNET is to use ANNIE to add any additional attributes that may be important to the data sets. Figure 2 contains an example of adding the characteristic percent of basin underline by Devonian to the data sets as attribute UBC024.

Using ANNEX to Compute Low Flows at Partial-record Sites

Figure 1. Example of building a WDM file and adding basin characteristics and 7-day low flows using IOWDM.

```
 #  screen
--  --------------------------------------------------------------------------
 1  Opening screen (File)
 2      File (Build)
 3          Build (va.wdm)
 4*     File (Return)
 5* Opening screen (Input)
 6      Input (Basin)
 7          Basin (Source)
 8              Source (va.bcd)
 9*         Basin (Options)
10              Options:  Confirm processing:   No data sets
                                              X First data set
                                                All data sets
                             Data-set status: X New
                                                Find
                                                Existing
                               Data-set type: X Time
                                                Table
                                   First dsn: 1
                                   Increment: 1
11*         Basin (Process)
12              Process (Location)
13                  Location:  TSTYPE = blank
                               ISTAID STAID STANMA STFIPS DSCODE AGENCY
14                  Location:  TSTYPE = L007
15              Process (Continue)
16*             Return
17*     Input (n-day)
18          N-Day (Source)
19              Source (va.ndy)
20*         N-Day (Options)
21              Options:  Confirm processing:    No data sets
                                               X First data set
                                                 Each station
                                                 All data sets
                              Data-set status: X New
                                                 Find
                                                 Enter
                                    First dsn: 36
                              Minor increment: 1
                              Major increment: 1
21.a            Options:  cursor moved to find
21.b            Options:  Find turned on
22*          N-day (Process)
23              Process (Location)
24                  Location:  TSTYPE:  L007
                               SEASBG:  4 - beginning of season (April)
                               SEASND:  3 - end of season (March)
25              Process (Continue)
26*         N-day (Return)
27      Input (Return)
28* Opening screen (File)
29      File (Summarize)
30          Summary
31*     File (Return)
32  Opening screen (Return)

Keystrokes:

F           File
B           Build
```

```
va.wdm      name for new wdm file
<F2>        accept screen
R           Return to Opening screen from File menu
I           Input
B           Basin characteristics data for processing
S           Source of basin characteristics
va.bcd      name of file containing basin characteristics data
<F2>        accept
O           Options for basin characteristics processing
<F2>        accept as is
P           Process the basin characteristics file
L           Location description for first station in input file
<dn arrow>  move cursor down to TSTYPE field
L007        set TSTYPE to L007 for 7-day low flows
<F2>        accept
C           Continue processing basin characteristics file
R           Return from Basin characteristics format to Input screen
N           N-day data for processing
S           Source of n-day data
va.ndy      name of file containing n-day data
<F2>        accept
O           Options for n-day processing
<rt arrow>  Move cursor to data-set status column
<dn arrow>  Move cursor to Find option under data-set status
<X>         select Find option to add n-day data to basin characteristics
<F2>        accept
P           Process the n-day file
L           Location
<F2>        accept
C           Continue processing n-day file
R           Return from N-day format to Input screen
R           Return from Input to Opening Screen
F           File to get a summary of contents of wdm file
S           Summary of wdm file
<F2>        Accept
R           Return from File to Opening Screen
R           Return to operating system
```

Figure 2. Example of adding additional attributes to data sets using ANNIE.

```
 #  screen
--  ------------------------------------------------------------------------------
1   Opening screen (File)
2       File (Open)
3           Open (va.wdm)
4*       File (Return)
5*   Opening screen (Data)
6       Data sets (Attributes)
7           Attributes (Select)
8               Select (Find)
9                   Find (Execute)
10                      Execute - search criteria
11                      Execute - 36 data sets checked match & added
12*                 Find (Return)
13*             Select (List)
14*                 list of dsn
15*             Select (Return)
16          Attribute (Modify)
17              Modify:  which attribute, blank
18                    :  ubc024
19              Modify:  for dsn 1, ubc024 none
20                    :  80
 .
 .
 .
21              Modify:  for dsn 36, ubc024 none
22                    :  0.0
23              Modify:  which attribute, blank
24                    :  done
25*         Attribute (Return)
26*     Data sets (Return)
27* Opening screen (Return)

Keystrokes:

F           File
O           Open the wdm file
va.wdm      name of the wdm file
<F2>        accept
R           Return to Opening Screen from File menu
D           Data sets
A           Attributes
S           Select the data sets to be worked with
F           use Find method of Selecting data sets
E           Execute with no search criteria will find all data sets
<F2>        summary of search criteria, Accept to continue
<F2>        found 36 data sets, Accept to continue
R           Return to the Select menu from Find
L           List the dsn of the selected data sets
<F2>        data sets 1-36 were selected, Accept to continue
R           Return to the Attribute menu from Select
M           Modify or Add attribute(s) in the selected data sets
ubc024      Modify or Add attribute ubc024 in/to the selected data sets
80          there is no data value for ubc024 in dsn 1, set it to 80
        .
        .
        .
0.0         there is no data value for ubc024 in dsn 36, set it to 0.0
done        ubc024 modified/added for all selected data set, DONE here
R           Return to Data set menu from Attribute
R           Return to Opening screen from Data sets
R           Return to operating system
```

LOW-FLOW FREQUENCY ESTIMATION AT UNGAGED SITES USING BASE-FLOW MEASUREMENTS

Introduction

Estimates of low-flow statistics (such as the 7-day 10-year low flow) are needed at ungaged sites for water-quality management. Experience has indicated that these low-flow characteristics currently cannot be accurately estimated by regression on drainage-basin characteristics. An alternative is the use of base-flow measurements at the ungaged site and concurrent daily flows at a nearby gaged site to establish relation between low flows at the two locations. Traditionally the 7-day 10-year low flow (or other low flows of interest) at the ungaged site is estimated by using the computed 7-day 10-year low flow at the nearby gaged site and the established regression relation. This technique is shown to be biased. An alternative estimator is proposed that utilizes the same regression relationship to estimate the mean and standard deviation of the annual events at the ungaged site in order to estimate the 7-day 10-year low flow. When applied to an actual data set, the new estimator appears to be unbiased in log space and to have the minimum mean square error among the five estimators considered.

Water-quality management often requires estimation of low-flow streamflow characteristics at sites without long or perhaps any daily flow records. In particular, the annual minimum 7-day consecutive low flow, which on average will be exceeded in 9 of 10 years, or in 19 of 20 years, is often employed as a design flow. Thomas and Benson (1970) found that such 7-day 10-year or 20-year low-flow values cannot be accurately estimated as a function of basin characteristics such as drainage area, stream channel length, or the percentage of the drainage area in forest or lakes. More recent reports, such as Arihood and Glatfelter (1986) and Bingham (1982), that utilize basin characteristics indicative of geology have achieved greater accuracy in estimating low-flow statistics. However, an alternative to the basin characteristics approach is still needed to improve the accuracy of low-flow estimation. One alternative traditionally used was suggested by Riggs (1965, 1972). In this approach, base-flow measurements (instantaneous values) are obtained at the site in question and correlated with concurrent daily flows (a 24-hour average) at a nearby gaged site for which a long flow record is available. Ideally the watershed for the nearby gaged site should be of similar drainage area size and geologic characteristics and have similar base-flow recession characteristics. The base-flow measurements and the concurrent daily flows at the gaged site can be used to establish a relationship between the flows at the two sites. That relationship, and the long-term flow record at the gaged site, can then be used to estimate the low-flow frequency relationship at the ungaged site. Riggs (1965, 1972) focused on graphical procedures. Hardison and Moss (1972) and Gilroy (1972) substituted analytical regression procedures for establishing a linear relationship between the logarithms of the flows and for estimating the accuracy of the d-day T-year low-flow estimate for the ungaged sites. In this report, deficiencies with their approach are discussed. An improved estimator is developed and a first-order estimate of its variance provided. The performance of this new estimator and four other estimators is examined using data from several stations. The new estimator is extended to allow use of concurrent daily flow values and one or more gages.

The Basic Problem

The analysis here is based on a linear model describing the relationship between the logarithms of the annual minimum d-day low flows, y_t, at the ungaged site and those, x_t, at a nearby gaged site:

$$y_t = \alpha + \beta x_t + e_t \qquad e_t \sim N(0, \sigma_e^2) \ . \tag{1}$$

With this model, the ε_t are independent residual errors that are assumed to be uncorrelated with the x_t. Letting μ_x, μ_y, σ_x^2, σ_y^2, and ρ_{xy} denote means, variances, and correlations of y and x, (1) implies that

$$\mu_y = \alpha + \beta\mu_x \tag{2}$$

and

$$\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_e^2 . \tag{3}$$

Equation (2) can also be written

$$\alpha = \mu_y - \beta\mu_x. \tag{4}$$

Multiplying both sides of (1) by x and taking expectations yields the additional relationship

$$\rho_{xy}\sigma_y\sigma_x = \beta\sigma_x^2$$

or

$$\beta = \rho_{xy}\sigma_y/\sigma_x. \tag{5}$$

In order to use the model in (1) and annual d-day minima at the gaged site to estimate the distribution of d-day low flows at the ungaged site, estimators of the parameters α, β, and σ_e^2 of the model in (1) are required. However, no record of d-day low flows at the ungaged site is available for this purpose. To overcome this difficulty, the logarithms of concurrent base-flow measurements $\tilde{y}_t$ and daily flows $\tilde{x}_t$ are used to estimate those parameters. Such observations should be separated by significant storm events so as to represent reasonably independent observations of the low-flow processes. Thus, one would base their analysis on the assumption or approximation that the relationship between $\tilde{y}_t$ and $\tilde{x}_t$ can be described by

$$\tilde{y}_t = \alpha + \beta \tilde{x}_t + e_t \qquad e_t \sim N(0, \sigma_e^2) \tag{6}$$

where α, β, and σ_e^2 have the same values as the model in (1).

In a subsequent section, this approximation is evaluated by comparing values of α and β based on base-flow measurements (equation 6) and annual 7-day minima (equation 1) at several sites. Although the α and β values vary significantly for given pairs of stations, on the average the assumption of similar α and β values was reasonable. The assumption that the relationship between instantaneous base flows is the same as the relationship between the minimum 7-day annual flows at the two sites is a necessary one if the proposed method is to be employed; it allows concurrent base-flow measurements to be used to construct a model that also relates annual minimum d-day low flows at the two sites. While this approximation appears reasonable for 7-day means, it may not be satisfactory for durations significantly longer than 7 days.

The derivations to follow use the definitions:

$$m_x = \frac{1}{n} \sum_{t=1}^{n} x_t$$
sample mean of the logarithms of annual d-day flows at the gaged site

$$m_{\tilde{y}} = \frac{1}{L} \sum_{t=1}^{L} \tilde{y}_t$$
sample mean of the logarithms of base-flow measurements at the ungaged site

$$m_{\tilde{x}} = \frac{1}{L} \sum_{t=1}^{L} \tilde{x}_t$$
sample mean of the logarithms of concurrent daily flows at the gaged site

$$s_x^2 = \frac{1}{(n-1)} \sum_{t=1}^{n} (x_t - m_x)^2$$
sample variance of the logarithms of annual d-day low flows at the gaged site

$s_{\tilde{y}}^2 = \frac{1}{(L-1)} \sum_{t=1}^{L} (\tilde{y}_t - m_{\tilde{y}})^2$ sample variance of the logarithms of the base-flow measurements at the ungaged site

$s_{\tilde{x}}^2 = \frac{1}{(L-1)} \sum_{t=1}^{L} (\tilde{x}_t - m_{\tilde{x}})^2$ sample variance of the logarithms of concurrent daily flows at the gaged site (7)

n = number of years record at the gaged site

L = number of base-flow measurements and concurrent daily flows

and also

$$b = \sum_{t=1}^{L} [(\tilde{y}_t - m_{\tilde{y}})(\tilde{x}_t - m_{\tilde{x}})] / (s_{\tilde{x}}^2 (L-1))$$

$$a = m_{\tilde{y}} - bm_{\tilde{x}}$$

$$s_e^2 = \frac{1}{L-2} \sum_{t=1}^{L} (\tilde{y}_t - a - b\tilde{x}_t)^2 \qquad (8)$$

Here a, b, and s_e^2 are the ordinary least squares estimators of the α, β, and σ_e^2 in (6). Furthermore, assume that the e_i in (6), corresponding to the base-flow measurements, are independent.

The issue is how to estimate the logarithm of the d-day T-year low flow

$$Y_T = \mu_y + K_y\sigma_y \qquad (9)$$

at the y-site (ungaged site) given the logarithm of the d-day T-year low flow

$$X_T = \mu_x + K_x\sigma_x \qquad (10)$$

at the x-site (gaged site). Here K_y and K_x are the appropriate frequency factors for the two sites for the computed skew values at the T-year recurrence interval. If the logarithms of the d-day low flows at both sites are assumed to have the same standardized distribution, then $K_y = K_x$.

A tempting estimator of Y_T suggested by Riggs (1965, 1972) and Hardison and Moss (1972) is

$$\hat{Y}_T^{(R)} = a + b\hat{X}_T . \qquad (11)$$

It was assumed that $\hat{Y}_T^{(R)}$ would be unbiased. However, if $\hat{X}_T = X_T$, and with the assumptions and approximation employed here

$$\underset{a,\,b}{E}\left[\hat{Y}_T^{(R)}\right] = \underset{a,\,b}{E}\,[a + bX_T]$$

$$= \alpha + \beta\,[\mu_X + K_X\,\sigma_X]$$

$$= (\mu_y - \beta\mu_X) + \rho_{xy}(\sigma_y/\sigma_X)\,[\mu_X + K_X\sigma_X]$$

$$= \mu_y + \rho_{xy}K_X\sigma y. \tag{12}$$

$\hat{Y}_T^{(R)}$ will be an unbiased and consistent estimator of Y_T only if

$$K_y = \rho_{xy}\,K_X.$$

This is unlikely. If K_y and K_X are approximately equal (implying the skew coefficients are approximately equal), then $\hat{Y}_T^{(R)}$ is only unbiased if $\rho_{xy} = 1$, given the other assumptions. In a subsequent section of this paper, it is shown that K_y and K_X are approximately equal for watersheds in similar hydrologic environments.

A reasonable, consistent, and simple estimator of Y_T can be obtained by using the base flows to calculate the estimators a and b of α and β. These values can be used with m_X and s_x^2 to estimate μ_y and σ_y^2 via equations (2) and (3). Our moment estimators are

$$\hat{\mu}_y = a + bm_x \tag{13a}$$

$$\hat{\sigma}_y^2 = b^2 s_x^2 + s_e^2\left[1 - \frac{s_x^2}{(L-1)\,s_{\tilde{x}}^2}\right]. \tag{13b}$$

The extra factor in brackets in (13b) is employed to obtain an unbiased estimator of σ_y^2 as shown below. Clearly, for independent base-flow observations and annual d-day low-flow measurements at the x-site

$$E\,[\mu_y] = (\hat{\mu}_y - \beta\mu_x) + \beta\,(\mu_x) = \mu_y\,. \tag{14a}$$

For fixed $\{\tilde{x}_1,\ldots,\tilde{x}_m\}$

$$E\left[\hat{\sigma}_y^2\right] = [\beta^2 + \mathrm{Var}\,(b)\,]\,\sigma_x^2 + \sigma_e^2\left[1 - \frac{\sigma_x^2}{(L-1)\,s_{\tilde{x}}^2}\right]$$

$$= \beta^2\sigma_x^2 + \sigma_e^2 + \sigma_x^2\left[\mathrm{Var}\,(b) - \frac{\sigma_e^2}{(L-1)\,s_{\tilde{x}}^2}\right] = \sigma_y^2\,. \tag{14b}$$

Thus, $\hat{\sigma}_y^2$ is also unbiased given that for every set $\{\tilde{x}_1,...,\tilde{x}_m\}$,

$$\mathrm{Var}\,(b) \;=\; \frac{\sigma_e^2}{(L-1)\,s_{\tilde{x}}^2} \tag{15}$$

provided the residuals in (6) are independent.

Finally, our moment estimator of Y_T is

$$\hat{Y}_T^{(M)} \;=\; \hat{\mu}_y + K_y\hat{\sigma}_y \tag{16}$$

where K_y can be estimated by K_x.

Precision of $Y_T^{(M)}$

Use of (13a, b) to estimate the mean μ_y and variance σ_y^2 of the annual minimum d-day flows at the y-site to facilitate estimation of $Y_T = \mu_y + K_y\sigma_y$ is theoretically the most attractive alternative considered. A first-order estimate of the variance of that estimator can be derived assuming that the residuals in (6) are normally distributed. To first order

$$\mathrm{Var}\left[\hat{Y}_T^{(M)}\right] \;=\; \mathrm{Var}\,[\hat{\mu}_y] + \frac{K_y^2}{4\sigma_y^2}\mathrm{Var}\left[\hat{\sigma}_y^2\right] + \frac{K_y}{\sigma_y}\mathrm{Cov}\left[\hat{\mu}_y,\, \hat{\sigma}_y^2\right] \tag{27}$$

where

$$\mathrm{Var}\,[\hat{\mu}_y] \;=\; \sigma_e^2\left[\frac{1}{L} + \frac{(\mu_x - m_{\tilde{x}})^2}{(L-1)\,s_{\tilde{x}}^2}\right] + \beta^2\left(\frac{\sigma_x^2}{n}\right) \tag{28}$$

which neglects the second order term Var(b)•Var(m_x). To first order in 1/L and 1/n

$$\left[1 - \sigma_x^2 / \left(\left(L-1\right) s_{\tilde{x}}^2\right)\right]\; \mathrm{Var}\,(s_e^2) \cong \mathrm{Var}\,(s_e^2) \tag{29}$$

where terms such as Var(b^2)•Var (s_x^2) can be neglected. Also, to first order (b^2-β^2) = (b+β)(b-β) ≅ 2β(b-β) so that E[(b^2-β^2)^2] ≅ 4β^2 Var(b) yielding

$$\mathrm{Var}\left[\hat{\sigma}_y^2\right] \cong 4\sigma_x^4\beta^2\mathrm{Var}\,(b) + \beta^4\mathrm{Var}\,(s_x^2) + \mathrm{Var}\,(s_e^2)\,. \tag{30}$$

Finally,

$$\mathrm{Cov}\,(\hat{\mu}_y,\, \hat{\sigma}_y^2) \cong 2\beta\sigma_x^2\,[\mathrm{Cov}\,(a,b) + \mu_x\mathrm{Var}\,(b)\,] \;=\; 2\beta\sigma_x^2\mathrm{Var}\,(b)\;(\mu_x - m_{\tilde{x}})\;. \tag{31}$$

Combining these results and also assuming that the x_t are themselves independent and normally distributed yields

$$\begin{aligned}
\mathrm{Var}\left[\hat{Y}_T^{(M)}\right] &\cong \frac{\sigma_e^2}{L} + \frac{(\mu_x - m_{\tilde{x}})^2\sigma_e^2}{(L-1)\,s_{\tilde{x}}^2} + \frac{\beta^2\sigma_x^2}{n} \\
&\quad + \frac{K_y^{\,2}}{4\sigma_y^2}\left\{\frac{4\beta^2\sigma_x^4\sigma_e^2}{L s_{\tilde{x}}^2} + \frac{2\beta^4\sigma_x^4}{n} + \frac{2\sigma_e^4}{L}\right\} + \frac{2\beta\sigma_x^2(\mu_x - m_{\tilde{x}}K_y\sigma_e^2)}{L\sigma_y s_{\tilde{x}}^2} \\
&\cong \frac{\sigma_e^2}{(L-1)}\left\{1 + \frac{(\mu_x - m_{\tilde{x}})^2}{s_{\tilde{x}}^2} + \frac{K_y^{\,2}}{2\sigma_y^2}\left[\sigma_e^2 + \frac{2\beta^2\sigma_x^4}{s_{\tilde{x}}^2}\right] + \frac{2\beta K_y(\mu_x - m_{\tilde{x}})\,\sigma_x^2}{\sigma_y s_{\tilde{x}}^2}\right\} \\
&\quad + \frac{\beta^2\sigma_x^2}{(n-1)}\left\{1 + \frac{\beta^2 K_y^2\sigma_x^2}{2\sigma_y^2}\right\} \qquad (32)
\end{aligned}$$

While (32) should be quite adequate for assessing the relative precision or sampling variability of $\hat{Y}_T^{(M)}$, it is only a first-order (in 1/n and 1/L) estimate derived assuming the residuals in (6) as well as the x_t are independent and normally distributed. Moreover, it does not incorporate the error introduced into the analysis by the assumption that the models in (1) and (6) have the same parameter values. However, equation (32) does allow the analyst to directly estimate the variance of d-day T-year low flows.

Possible Accuracy Improvements

Two particular variables are sometimes subject to a hydrologist's control; these are ρ_{xy}, the cross correlation of the flows, and L, the number of concurrent measurements upon which the estimates of α, β, and σ_e^2 are based. By selecting a gage site whose low flows are highly correlated with the flows at the ungaged site of interest, the hydrologist can hope to obtain a pair of stations with a high ρ_{xy}. L is clearly an indication of the effort invested to obtain concurrent measurements. Figure 1 illustrates, using the first-order approximation in (32), a likely relationship between the standard error of estimate (in percent)

$$\mathrm{SE} = 100\left[\exp\{(2.3)^2\mathrm{Var}\left[\hat{Y}_T^{(M)}\right]\} - 1\right]^{1/2} \qquad (35)$$

and values of ρ_{xy} and L. In this example, representative values of n = 25 or 50, K_y = -1.3, $\mu_x \cong m_{\tilde{x}}$, $\sigma_y = \sigma_x$ = 0.35, and $\sigma_{\tilde{x}}$ = 0.25 were used. Note that $\sigma_e^2 = (1 - \rho_{xy}^2)\, \sigma_y^2$. The choice of K_y = -1.3 implies that the standard errors shown in figure 1 are comparable to those for the 10-year event.

FIGURE 1 HERE

Figure 1 shows that for small L, the standard error decreases rapidly as L increases. As L becomes larger, the accuracy of $\hat{Y}_T$ is ultimately determined by the precision of m_x and s_x^2, the estimators of the moments of the flows at the gaged site. The standard error of the gaged site estimator of Y_T is 22 percent for n = 25, and 16 percent for n = 50. These numbers provide a standard with which to compare the values in figure 1. In this particular example, the precision of $\hat{Y}_T^{(M)}$ increases slowly beyond L = 20.

One can also see in the figure that for small L, the accuracy of $\hat{Y}_T$ is highly sensitive to ρ_{xy}: higher correlations yield more accurate estimators. This occurs because for fixed σ_y^2, large ρ_{xy} yields relatively small σ_e^2 meaning that α, β, and σ_e^2 are relatively more accurate than they would be if ρ_{xy} were smaller.

In general, Var$\left[\hat{Y}_T^{(M)}\right]$ decreases with increasing ρ_{xy}. However, as can be seen in figure 1a corresponding to n = 25, the variance of $\hat{Y}_T$ for ρ = 0.50 actually becomes slightly less than the variance for ρ = 0.70 or 0.90 when L is large. This makes sense in that our estimator of Y_T, for small ρ_{xy}, depends as much or more on the parameters of the regression model and the estimated residual variance as it does on m_x and s_x^2, the sample moments of x_t; see equations (13), (2), or (3). This explains mathematically why, with large L, it can happen that

$$\text{Var}\left[\hat{Y}_T^{(M)} \middle| \rho = 0.5\right] < \text{Var}\left[\hat{Y}_T^{(M)} \middle| \rho = 0.7\right].$$

Figure 1a here

Such reversals of precision are probably an illusion because they occur in instances when the basic approximation upon which the analysis is based is probably not satisfactory. The theory leading to our best estimator $\hat{Y}_T^{(M)}$ was based on the approximation that the parameters α, β, and σ_e^2 of the models in (1) and (6) were essentially the same. This assumption is probably true when ρ_{xy} = 1, but becomes an increasingly less precise description of reality as ρ_{xy} decreases. For ρ_{xy} = β = 0, the models in (1) and (6) become

$$y_t = \alpha + e_t \qquad E[e_t] = E[\tilde{e}_t] = 0$$

$$\tilde{y}_t = \alpha + \tilde{\varepsilon}_t \qquad \mathrm{Var}\ [e_t] = \mathrm{Var}\ [\tilde{e}_t] = \sigma_e^2$$

implying that both y_t and $\tilde{y}_t$ have the same mean α and variance σ_e^2. We have argued above that it should be the case that

$$E[y_t] < E[\tilde{y}_t]$$

because y_t are annual minima whereas $\tilde{y}_t$ are only small values, the majority of which will exceed the minima for their year.

Another way of viewing the origin of this problem is by noting that when ρ_{xy} approaches unity, the model in (6) allows low flows $\tilde{x}_t$ at the x-gage to be mapped fairly precisely into the corresponding flows $\tilde{y}_t$ at the y-gage. It is then a reasonable approximation to assume that the annual low flow y_t at the y-gage occurred concurrently with the annual flow x_t at the x-gage and that

$$y_t = \alpha + \beta\ x_t + e_t$$

where α, β, and σ_e^2 can be estimated using low base flows and concurrent daily flows. However, when ρ_{xy} assumes small or even modest values, then it will frequently occur that y_t does not occur concurrently with x_t. Then concurrent base-flow and daily-flow measurements at the two sites do not provide a reliable means of estimating the relationship between the annual minima. In retrospect, it would be our recommendation that these regression procedures not be employed to estimate the distribution of annual minima at ungaged sites unless ρ_{xy} exceeds about 0.70. For half of our station pairs, the sample estimates of the cross correlations failed to meet this criterion.

A Multivariate Estimator

Description of Technique

Because the explanatory power of the relationship, (1) or (6), relating flows at the two sites is so important to the validity of the analysis, it may be possible and worthwhile to use multivariate models

$$y_t = \alpha + \sum_{j=1}^{k} \beta_j x_{jt} + e_t \tag{36}$$

where $x_i, \ldots, x_k$ are annual minimum d-day low flows. Again assume that $\beta = (\alpha, \beta_i, \ldots, \beta_k)^T$ can be estimated by analysis of concurrent base-flow measurements; thus, corresponding to (6) one has

$$\tilde{y}_t = \alpha + \sum_{j=1}^{k} \beta_j \tilde{x}_{jt} + e_t \tag{37}$$

where in both cases

$$e_t \sim N\,(0, \sigma_e^2)\ .$$

By including more than one station in the regression model to explain the value of $\tilde{y}_t$ or y_t, it may be possible to substantially reduce σ_e^2 meaning that the explanatory variables $\{\tilde{x}_{it}, \ldots., \tilde{x}_{kt}\}$ explain more of the variation of $\tilde{y}_t$, and the effective ρ_{xy} is increased.

For convenience, let the multivariate model be written

$$\underline{y}_t = \underline{x}_t^T \underline{\beta} + e_t \tag{38a}$$

where $x_t = (1, x_{it}, \ldots, x_{kt})^T$. Let

$$\underline{b} = (a, b_1, \ldots, b_k)^T \tag{38b}$$

be the least squares estimator of $\underline{\beta}$. Finally, let

$$\underline{m}_x = (1, m_{1x}, \ldots, m_{kx})^T$$

$$\underline{m}_{\tilde{x}} = (1, m_{1\tilde{x}}, \ldots, m_{k\tilde{x}})^T \tag{38c}$$

be the row vector of sample means for the d-day low flows and base flows while

$$V_{xx} = \frac{1}{n-1} \sum_{t=1}^{n} (\underline{x}_t - \underline{m}_x)\ (\underline{x}_t - \underline{m}_x)^T$$

$$S_{\tilde{x}\tilde{x}} = \sum_{t=1}^{L} (\underline{\tilde{x}}_t - \underline{m}_{\tilde{x}})\ (\underline{\tilde{x}}_t - \underline{m}_{\tilde{x}})^T$$

$$S_{\tilde{x}\tilde{y}} = \sum_{t=1}^{L} (\underline{\tilde{x}}_t - \underline{m}_{\tilde{x}})\ (\underline{\tilde{y}}_t - \underline{m}_{\tilde{y}})\ \ .$$

With this notation

$$\underline{b} = [S_{\tilde{x}\tilde{x}}]^{-1} S_{\tilde{x}\tilde{y}} \qquad (39a)$$

and the residual mean square error is

$$s_e^2 = \left\{ \sum_{t=1}^{L} (\tilde{y}_t - m_{\tilde{y}})^2 - \underline{b}^T S_{\tilde{x}\tilde{y}} \right\} / (L-2) \qquad (39b)$$

Using the multivariate regression model, the estimators of y_t's mean and variance in (13) become

$$\hat{\mu}_y = \underline{m}_x^T \underline{b} \qquad (40a)$$

and

$$\hat{\sigma}_y^2 = \underline{b}^T V_{xx} \underline{b} + s_e^2 \{1 - \mathrm{tr}\,[V_{xx} (S_{xx})^{-1}]\} \qquad (40b)$$

where tr[•] corresponds to the trace of the indicated matrix and should correspond to a relatively small correction.

As before,

$$E\,[\hat{\mu}_y] = \mu_y$$

and

$$E\left[\hat{\sigma}_y^2\right] = \sigma_y^2$$

due to the correction employed in the σ_y^2 term in (40b); for modest L, the correction tr $\left[V_{xx} (S_{\tilde{x}\tilde{x}})^{-1}\right]$, which is of the order $(L\text{-}1)^{-1}$, may be omitted.

With these estimators, $\hat{Y}_T^{(M)}$ would still be given by (16) and its variance by (27) with the appropriate values of Var $[\hat{\mu}_y]$, Var $\left[\hat{\sigma}_y^2\right]$, and Cov $\left[\hat{\mu}_y, \hat{\sigma}_y^2\right]$. An analysis similar to that leading to (29), (30), and (31) yields first-order approximations for those quantities assuming the pertinent random variables are normally distributed.

$$\mathrm{Var}\,[\hat{\mu}_y] \cong \sigma_e^2 \underline{\mu}_x^T [S_{\tilde{x}\tilde{x}}]^{-1} \underline{\mu}_x + \underline{\beta}^T [V_{xx}/n]\, \underline{\beta}$$

$$\mathrm{Cov}\left[\hat{\mu}_y, \hat{\sigma}_y^2\right] \cong 2\sigma_e^2 \beta^T [S_{\tilde{x}\tilde{x}}]^{-1} V_{xx} \underline{\mu}_x$$

$$\mathrm{Var}\left[\hat{\sigma}_y^2\right] \cong \sigma_e^4 \{2/(L-k) + 4\underline{\beta}^T V_{xx}^T [S_{\tilde{x}\tilde{x}}]^{-1} V_{xx}\underline{\beta}\} + \underline{\beta}^T A \underline{\beta}/(n-1) \qquad (41)$$

where

$$A_{ij} = \sum_{h=1}^{n} \sum_{k=1}^{n} \beta_h \beta_k (V_{ih} V_{kj} + V_{ij} V_{hk})$$

in which V_{ij} are the elements of V_{XX} (see Zellner, 1971, p. 389). Here k is the number of explanatory variables in the regression model; here $S_{\tilde{x}\tilde{x}}/(L-1)$ and V_{XX} have been substituted for the respective population moments. Note the first row and column of V_{XX} are zero because of the definition in equation 38a. As before, these expressions neglect the error introduced by using $\tilde{y}_t$ and $\tilde{x}_t$ to estimate the relationship between y_t and x_t.

GLS REGIONAL REGRESSION USING GLSNET

Model Description and Assumptions

Consider a region in which we have data for n gaging stations as follows:

At each gaged site we estimate a streamflow characteristic, such as the logarithm of the 50-year peak flow,

$$y_i = \Psi_i + \eta_i, \qquad (1)$$

in which ψ_i is the true (but unknown) log of the 50-year peak, and η_i is a random error. If y_i is an unbiased estimate of ψ_i, then η_i (sometimes called time sampling error) has mean zero and variance that is a function of how many years of data are available at the site and the standard deviation of annual peaks. In addition, we have k basin characteristics, such as log of drainage area, that are measured with negligible error.

If we are willing to assume that (within the region defined by the basin characteristics at the n stations) ψ is approximately linearly related to the basin characteristics (x’s), then the model formulation can be written

$$\Psi_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots + \beta_k x_{ki} + \varepsilon_i \quad \text{(i=1,2,...,n; n>k)}, \qquad (2)$$

in which ε_i is a model error assumed uncorrelated from observation to observation, with mean zero and constant variance, γ^2. Substituting into equation 1,

$$y_i \;=\; \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \eta_i + \varepsilon_i . \tag{3}$$

In matrix notation

$$\mathbf{Y} \;=\; \mathbf{X}\beta + \upsilon \,, \tag{4}$$

in which

$$\mathbf{Y} \;=\; \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \qquad \beta \;=\; \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix} \qquad \upsilon \;=\; \begin{bmatrix} \varepsilon_1 + \eta_1 \\ \varepsilon_2 + \eta_2 \\ \ldots \\ \varepsilon_n + \eta_n \end{bmatrix} \quad , \tag{5}$$

in which $E[\upsilon]=\mathbf{0}$, and $E[\upsilon\upsilon^T]=\Lambda$. Now the GLS estimator of β is

$$\mathbf{b} \;=\; (\mathbf{X}^T\Lambda^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Lambda^{-1}\mathbf{Y} \;. \tag{6}$$

The problem with this estimator is we do not know Λ and must estimate it from the data. In OLS, Λ is estimated as $\sigma^2\mathbf{I}$, which would not be bad if all stations in that region had approximately the same lengths of record, or if the variance of η_i is small relative to the variance of ε_i at every station in the region.

In most studies this assumption may be hard to justify, so we try to make a better estimate of Λ. We will denote this estimated covariance matrix $\hat{\Lambda}$, and the GLS estimator, **b**, will be referred to as an Estimated Generalized Least Squares (EGLS) estimator.

EGLS Regression

To illustrate how $\hat{\Lambda}$ is estimated, we use an example. Suppose that y_i is the log of the 50-year peak estimated from m_i years of record and that the annual peaks follow a Log-Pearson Type III distribution at all sites. Further, to minimize notation, assume that the skew coefficient at all sites is zero. The elements of $\hat{\Lambda}$ would be given by:

$$\lambda_{ij} \;=\; \begin{cases} \gamma^2 + \dfrac{\sigma^2{}_i\,(1+0.5K^2)}{m_i} & \text{(for i=j)} \\ \qquad\qquad\qquad\qquad \text{or} & \\ \dfrac{\rho_{ij}\sigma_i\sigma_j m_{ij}\,(1+0.5K^2)}{m_i m_j} \;(& \text{for } i \neq j) \end{cases} \quad . \tag{7}$$

In this equation we know K (Log-Pearson Type III standard deviate for zero skewness and 50-year recurrence interval), m_i (record length at station i), and m_{ij} (concurrent record length for stations i and j), but we must estimate σ_i (standard deviation of annual peaks at station i), ρ_{ij} (cross correlation of annual peaks at stations i and j), and γ^2 (variance of model error) from the data. Furthermore, we cannot use s_i (the sample estimate of σ_i) as an estimate of σ_i without introducing bias, and the use of r_{ij} (sample cross correlations) for ρ_{ij} often causes numerical problems. Therefore, we estimate σ_i and ρ_{ij} as follows:

The standard deviation of annual peaks, σ_i, is estimated from a regional regression of the form

$$\ln(s_i) = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}. \tag{8}$$

The cross correlation coefficient, ρ_{ij}, is estimated by developing an empirical relationship between sample cross correlations, r_{ij}, and distance between stations of the form

$$r_{ij} = \Theta^{\left[\frac{d_{ij}}{\alpha d_{ij} + 1}\right]}. \tag{9}$$

Now the only parameters left to find in the EGLS model is the model error variance γ^2. This value is determined by a numerical search method so that

$$(Y\text{-}Xb)^T\Lambda^{-1}(Y\text{-}Xb) = n\text{-}k. \tag{10}$$

Finally, in order to estimate all the quantities we need, we have to run two different regressions. First, as explained, a regional regression of ln(s) is performed to get $\hat{s}_i$, and then the final regression of Y to find b and γ.

Leverage in GLS Regression

Recall that in OLS regression the leverage of site i is the ith diagonal element of

$$H=X(X'X)^{-1}X'. \tag{11}$$

The analogous statistic in GLS regression is the ith diagonal element of

$$H^*=X(X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}. \tag{12}$$

The sum of h^*_{ii}=p' and the average h^*_{ii}=(p'/n), and a high leverage site would be one in which h^*_{ii}>(2p/n) as a rule of thumb.

Cook’s D Statistic in GLS Regression

A GLS version of Cook’s D is

$$D'_i = \frac{e_i^2 h'_{ii}}{p'\,(\gamma_i - h'_{ii})^2}\ , \qquad (13)$$

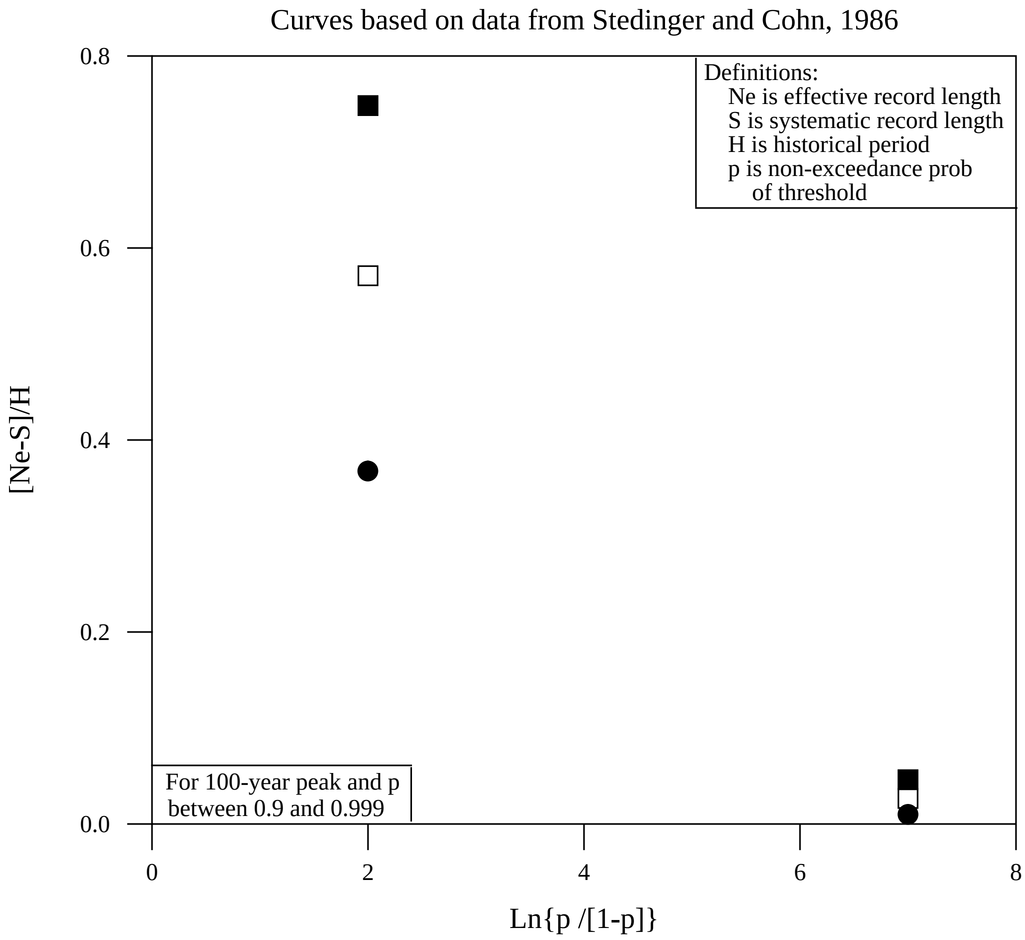
where h'_{ii} are diagonal elements of

$$H' = X(X'\Lambda^{-1}X)^{-1}X'. \qquad (14)$$

D'_i is large if it exceeds about (4/n).

Adjustment for Historical Information

The formulas above assume we have just systematically recorded data at each site. In reality, we often have additional information about unusually large floods that might have occurred outside the record. This historical information can be used to improve the estimate of T-year floods at gaged sites. To incorporate this additional information into the EGLS regression method, we compute an effective record length at the site with historical information to reflect the additional accuracy introduced by the historical information (see graph below). The GLSNET program in ANNIE makes this adjustment automatically if the number of historical peaks and the historical period are stored in the WDM file as attributes.



Average Mean Squared Error of Prediction

One measure of how good the regression model is for prediction is the average mean squared prediction error, where the average is taken over prediction sites with X variables identical to the observed data. This is a good measure when the observed data have been collected at a representative set of sites in the region. It is computed as

$$\mathrm{AMSEP} = \left(\sum_{i=1}^{n}\left(\hat{\gamma}^2_i + x_i\,(X'\hat{\Lambda}^{-1}X)^{-1}\,x'_i\right)\right)n^{-1}. \qquad (15)$$

Prediction Interval

Users of the regression model are probably more interested in a measure of error in a particular prediction rather than an average prediction. A good measure of the error of a particular prediction is the confidence interval of a prediction, or prediction interval. Let x_0 represent the usual row vector of basin characteristics at a prediction site. As usual, x_0 is augmented by a 1 as the first element. The predicted value is $\hat{y}_0 = x_0 b$. A 100(1-α) prediction interval would be

$$\hat{y}_0 - T \le y_0 \le \hat{y}_0 + T, \tag{16}$$

where

$$T = t_{\frac{\alpha}{2}, n-p'} \sqrt{\left(\hat{\gamma}^2_0 + x_0 (X'\hat{\Lambda}^{-1} X)^{-1} x'_0\right)}, \tag{17}$$

where $t_{\alpha/2, n-p'}$ is the critical value from a t distribution for n-p' degrees of freedom.

If a log transform had been made so that $y_0 = \log_{10}(q_0)$, then the prediction interval would be

$$10^{\hat{y}_0 - T} \le q_0 \le 10^{\hat{y}_0 + T}. \tag{18}$$

Extrapolation Beyond Sample Data

In making predictions from a regression model, one should be very careful about extrapolating beyond the region of the original sample data. As noted earlier, it is possible in multiple regressions to be within the range of every X variable without being within the region of the sample data. All the original data are contained in an ellipsoid defined by

$$h'_{max} \ge x (X'\hat{\Lambda}^{-1} X)^{-1} x' , \tag{19}$$

where h'_{max} is the maximum of the diagonal elements of H' (see 2.15). For a prediction at a new site x_0, if h'_{00} computed by $x_0(X'\Lambda^{-1}X)^{-1}x'_0$ is greater than h'_{max}, then the prediction point is outside the ellipsoid of the original sample data and is an extrapolation.

The calculations needed to compute prediction intervals and to test to see if a prediction is an extrapolation are tedious and time consuming. However, they can be easily made with a small computer program that could be given to the user as part of the report.

Computer Output

The program prints summary regression results for three separate regressions: (1) the regression of log of standard deviation of flows on basin characteristics, (2) the regression of mean of flows on basin characteristics, and (3) the regression of the T-year quantile on basin characteristics. These regression summaries include estimates of the regression coefficients along with their standard error, T statistic, and P value for testing the hypothesis that the coefficient is zero.

Following the final regression summary, a table is printed to help one analyze the results. This table includes the following information for each observation.

Column

1 Station number.

2 Observed value of the flow statistic.

This value serves as the observed response, y_i, and is calculated as

$$y_i = \hat{\mu}_i + k_i s_i$$

where $\hat{\mu}_i$ is the sample mean (usually in log units), k_i is the Log-Pearson Type III standard deviate, and s_i is the sample standard deviation (also usually in log units). The value of k_i is a function of the skew coefficient at site i and the recurrence interval of interest. It is computed internally in the program.

3 Predicted value of y_i, computed as $\underline{x}_i\underline{b}$ where $\underline{x}_i$ is a row vector of basin characteristics and $\underline{b}$ is a column vector of estimated regression coefficients.

4 Variance of the predicted value of y_i, computed as

$$\hat{\gamma}_i^2 + \underline{x}_i (\underline{X}^T \Lambda^{-1} \underline{X})^{-1} \underline{x}_i^T$$

where $\hat{y}_i^2$ is the estimated model error variance at site i (column 8).

5 Residual, r_i (column 2 - column 3)

$$r_i = \tilde{y}_i - \underline{x}_i\underline{b}$$

6 Weighted average of predicted and observed values of y at site i

$$(\text{weighted } y_i) = \frac{n_i y_i + en_i \underline{x}_i \underline{b}}{n_i + en_i}$$

where n_i is the actual record length and en_i is the equivalent record length of the regression estimate (column 10).

7 Standardized residual, rs_i, is the residual r_i divided by the square root of its variance. It is calculated as

$$rs_i = \frac{r_i}{\left[\lambda_i - \underline{x}_i\,(\underline{X}^T\underline{\Lambda}^{-1}\underline{X})^{-1}\,\underline{x}_i^T\right]^{1/2}}$$

where λ_i is the ith diagonal of Λ.

If the residuals are approximately normal, then about one-sixth of the rs_i's will fall above 1 and about one-sixth of the rs_i's will fall below -1, and about 95 percent of the rs_i's will fall between -2 and +2.

8 Model error variance, γ_i^2, is a measure of the error inherent in the model that cannot be changed by collecting more data.

9 Estimated sampling error variance, $\hat{\Sigma}_i$, calculated as

$$\hat{\Sigma}_i = \underline{x}_i\,(\underline{X}^T\hat{\underline{\Lambda}}^{-1}\underline{X})^{-1}\,\underline{x}_i^T.$$

Sampling error is the error in predicting y_i due to estimating the true regression parameters, $\underline{\beta}$, with $\underline{b}$. The variance of a prediction (column 4) is the sum of model error variance (column 8) and sampling error variance (column 9).

10 Equivalent years of record, en_i (Hardison, 1971), expresses the accuracy of prediction in terms of the number of years of record required to achieve results of equal accuracy. It is calculated as

$$\frac{\hat{s}_i^2\left[1 + k_i g_i + \frac{k_i^2}{2}\,(1 + 0.75 g_i^2)\right]}{\hat{\gamma}_i^2 + \hat{\Sigma}_i}$$

where g_i is the coefficient of skewness at site i.

11 The hat diagonals, h_{ii}, are related to the “distance” between the row vector, $\underline{x}_i$, of basin characteristics and the row vector, $\bar{x}$, of basic characteristic means. They are the diagonal elements of

$$H = \underline{X}\,(\underline{X}^T\underline{\Lambda}^{-1}\underline{X})^{-1}\,\underline{X}^T\hat{\underline{\Lambda}}^{-1}.$$

The sum of all n values of h_{ii} is p, and values greater than 2p/n are considered “large.”

12 This statistic is a GLS version of "Cook's D." It summarizes the influence of each observation on the final regression result. It depends on both the "leverage," h_{ii}, and the residual, r_i, and is calculated by

$$D_i = \frac{h_{ii}^* r_i^2}{p\,(\lambda_i - h_{ii}^*)^2}$$

where h_{ii}^* is the ith diagonal element of $\underline{H}^* = \underline{X}\,(X^T \underline{\hat{\Lambda}}^{-1} \underline{X})^{-1} \underline{X}^T$.

Values of D_i greater than 4/n are often considered to be particularly influential and their validity should be examined. A large value of D_i does not mean that the ith observation should be deleted. It does indicate that deletion of this observation will have a greater effect on the regression result than observations with smaller values of D_i.

Finally, the averages for sampling error variance (column 9), model error variance (column 8), and equivalent years of record (column 10) are printed. These average values summarize the strength of the regression model. The square root of the average model error variance is comparable to the standard error of estimate; and, if the dependent variable is in log units, it could be converted to a percent error. An overall measure of predictive ability is the average equivalent years of record. Another measure of overall predictive ability is the average variance of prediction (sum of average sampling error variance and average model error variance).

In the example output shown on the next two pages, the dependent variable was the log (base 10) of the 100-year peak (Q_{100}). The final equation could be written as

$$\log Q_{100} = 1.77291 + 0.70472 \log(\text{area}) + 0.33609 \log(\text{slope}) + 1.54420 \log (I_{24\text{-}2} - 1.0)$$

or

$$Q_{100} = 59.3(\text{area})^{0.705}(\text{slope})^{0.336}(I_{24\text{-}2}-1.0)^{1.54}.$$

The average equivalent years of record would be 4.1. The standard error of the model would be $(.2356)^{1/2}$ or .189. This standard error could be expressed as a percent by the formula

$$SE = 100\,[\exp(0.0356*5.3019)-1]^{1/2} = 46\%.$$

The average standard error of prediction would be

$$\overline{SE}_p = (0.0356 + 0.0070)^{1/2} = 0.206 \text{ log units.}$$

In percent, $\overline{SE}_p = 100\,[\exp(.226)-1]^{1/2} = 50\%$.

Note: The constant 5.3019 in the above formula is $(\ln 10)^2$. It converts the units from squared common logs to squared natural logs. If the dependent variable was in natural (base e) units, then this factor would be 1.

NETWORK ANALYSIS USING GLSNET

Identifying Stream Gages to Operate for Regional Information

The problem of identifying at which sites to collect future streamflow data is formulated as a mathematical program to optimize regional information subject to a budget constraint. An approximate solution is obtained using a step-backward technique that identifies gaging station sites, either existing or new, to discontinue data collection, or not start data collection, respectively, if the budget is exceeded. The method allows a network manager to design a nearly optimal streamflow data network for collecting regional information. The method is illustrated by a network of stream gages in Illinois.

Formulation of the Network Analysis Problem

1. Decision variables:

$$u_i = \begin{cases} 1, & \text{if gage i is operated} \\ 0, & \text{if gage i is not operated} \end{cases} \qquad (20)$$

2. Objective function:

$$\text{minimize } Z = f\{u_i\} \qquad (21)$$

2. Constraints:

$$\sum \text{cost}_i u_i \le \text{BUDGET} \qquad (22)$$

Objective: Minimize the sum of squared prediction errors of a regional regression model over a representative set of sites.

The mean square error of a prediction at a given site, x_k, is

$$MSE_k = \gamma_k^2 + x_k\{X'\Lambda(u)^{-1}X\}^{-1}x_k \qquad (23)$$

where the inverse covariance matrix Λ is written as $\Lambda(u)$ to show that it is a function of the decision variables u.

Let the set R denote a representative set of sites in the region. The objective function is

$$\min Z = \sum_{k \subset R} MSE_k \tag{24}$$

An approximate solution for the mathematical program is obtained by the following algorithm:

1. Set u_i=1 for all stations.
2. For all stations in which u_i=1, find station j such that $(Z_{(j)}-Z_j)/C_j$ is smallest and set u_j=0.
3. Check constraint, if $\Sigma C_i u_i$>Budget, go to 2, otherwise stop.

Computer Output

The NETWORK option in GLSNET produces an output file and, optionally, a TELAGRAF plot file. The output file (see the example that follows) shows at each step the expected sampling mean square error at the end of the specified planning horizon. At each step the calculations are made assuming that all the stations are operated during the entire planning horizon except the ones indicated in the column "discontinued stations" for all steps down to the current step. The plot file is a plot of the "average sampling error" column against the "cost" column.

Note that if all "costs" are 1, as in the example, the "cost" is actually the number of stations operated during a planning horizon. The units of "average sampling error" often are squared log units that are difficult to interpret in a practical sense. However, one may convert these units to more meaningful units. For example, one could convert “average sampling error” at end of planning horizon to the expected percent increase in equivalent years of record, Δn_e, by the formula

$$\Delta n_e = 100 \left[\frac{s_o^2 - s_f^2}{s_m^2 + s_f^2} \right] \tag{25}$$

where s_m^2 is the average mean square model error, s_o^2 is the current average mean square sampling error, and s_f^2 is the future average mean square sampling error. For example, in step 24 of the NETWORK program, cost is 17 and s_f^2 is 0.00479. From the GLS output, s_m^2 is 0.0356 and s_o^2 is 0.0070. Therefore, for a cost of 17, Δn_e is given by

$$\Delta n_e = 100 \left[\frac{0.0070 - 0.0048}{0.0356 + 0.0048} \right] = 5.4\% . \tag{26}$$

REFERENCES

Alley, W.M., and Burns, A.W., 1983, Mixed-station extension of monthly streamflow records: American Society of Civil Engineers Journal of Hydraulic Engineering, v. 109, no. 10, p. 1272-1284.

Arihood, L.D., and Glatfelter, D.R., 1986, Method for estimating low-flow characteristics for ungaged streams in Indiana: U.S. Geological Survey Open-File Report 86-323.

Benson, M.A., and Matalas, N.C., 1967, Synthetic hydrology based on regional statistical parameters: Water Resources Research, v. 3, no. 4, p. 931-935.

Bingham, R.H., 1986, Low-flow characteristics of Alabama streams: U.S. Geological Survey Water-Supply Paper 2083, 27 p.

Dalrymple, T., 1960, Flood-frequency analyses: U.S. Geological Survey Water-Supply Paper 1543-A.

Fiering, M.B., 1962, On the use of correlation to augment data: Journal of the American Statistical Association, v. 57, p. 20-32.

______ 1963, Use of correlation to improve estimates of the mean and variance: U.S. Geological Survey Professional Paper 434-C.

Gilroy, E.J., 1972, Outline of derivations: in U.S. Geological Survey Water-Supply Paper 1542-B, p. 48-55.

Hardison, 1971

Hardison, C.H., and Moss, M.E., 1972, Accuracy of low-flow characteristics estimated by correlation of base-flow measurements: U.S. Geological Survey Water-Supply Paper 1542-B.

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

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

Moran, M.A., 1974, On estimators obtained from a sample augmented by multiple regression: Water Resources Research, v. 10, no. 1, p. 81-85.

Riggs, H.C., 1965, Estimating probability distributions of drought flows: Water and Sewage Works, v. 112, no. 5, p. 153-157.

_____ 1972, Low-flow investigations: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chapter B1.

Stedinger and Cohn, 1986, graph

Tasker and Stedinger, 1989, p. 1

Thomas, D.M., and Benson, M.A., 1970, Generalization of streamflow characteristics from drainage-basin characteristics: U.S. Geological Survey Water-Supply Paper 1975, 55 p.

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

Zellner, A., 1971, An Introduction to Bayesian Inference in Econometrics: New York, Wiley and Sons, p. 389.

APPENDIX II - NOTATION

The following symbols are used in this paper:

a = sample estimator of α

b = sample estimator of β

$BIAS_j$ = average bias of j^{th} estimator over the data set

k_x = frequency factor for gaged x site

K_y = frequency factor for ungaged y site

L = number of concurrent base-flow and daily flow measurements

m_x = sample mean of logarithms of gaged annual low flows

$m_{\tilde{x}}$ = sample mean of logarithms of concurrent daily flows observed at gaged site

m_y = sample mean of logarithms of concurrent annual low flows at y-site

$m_{\tilde{y}}$ = sample mean of logarithms of concurrent base flows observed at gaged site

n = number of years of record at gaged site

r = sample estimator of ρ_{xy}

$RMSE_j$ = average root mean square error of j^{th} estimator over the data set

s_e^2 = sample estimator of σ_ε^2

s_x^2 = sample variance of logarithms of gaged annual low flows

$s_{\tilde{x}}^2$ = sample variance of logarithms of concurrent daily flows observed at gaged site

s_y^2 = sample variance of logarithms of annual flows at y-site

$s_{\tilde{y}}^2$ = sample variance of logarithms of concurrent base flows observed at ungaged site

SE = standard error of estimate (in percent) as defined in equation (35)

x_t = logarithm of annual low flows at gaged site

$\tilde{x}_t$ = logarithm of concurrent daily flows at gaged site

X_T = value of logarithm of T-year d-day low flow at gaged site

y_t = logarithm of annual low flows at ungaged site

$\tilde{y}_t$ = logarithm of concurrent base flows at ungaged site

Y_T = value of logarithm of T-year d-day low flow at ungaged site

$\hat{Y}_T^{(G)}$ = estimator of logarithm of T-year d-day low flow based upon Gilroy's variance estimator

$\hat{Y}_T^{(H)}$ = estimator of logarithm of T-day d-day low flow based upon Hirsch's MOVE.1 technique applied to X_T

$\hat{Y}_T^{(M)}$ = recommended moment estimator of logarithm of T-year d-day low flow

$\hat{Y}_T^{(R)}$ = regression estimator of logarithm of T-year d-day low flow

$\hat{Y}_T^{(S)}$ = mean-scaling estimator of logarithm of T-year d-day low flow

$Y_{7,T}$ = logarithmic value of 7-day T-year low flow estimates at ungaged sites

α = constant in regression model

β = slope parameter in regression model

ε_t = residual error for observation t in regression model

μ_x = mean of logarithms of annual low flows at gaged site

μ_y = mean of logarithms of annual low flows at ungaged site

$\hat{\mu}_y$ = estimated mean of logarithms of annual low flows at ungaged site

ρ or ρ_{xy} = cross-correlation between logarithms of annual low flows at gaged x and ungaged y sites

σ_e^2 = variance of residual errors e_t in regression model

σ_x^2 = variance of logarithms of annual low flows at gaged site

$\sigma_{\tilde{x}}^2$ = variance of logarithms of concurrent daily flows at gaged site

σ_y^2 = variance of logarithms of annual low flows at ungaged site

$\sigma_{\tilde{y}}^2$ = variance of logarithms of base flows at ungaged site

$\hat{\sigma}_y^2$ = estimated variance of logarithms of annual low flows at ungaged site