Identifying Stream Gauges to Operate for Regional Information

GARY D. TASKER

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 gauging station sites, either existing or new, and whether to discontinue or not start data collection 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 gauges in Illinois.

How to determine which gauging stations (old or new) to operate, given a limited budget, to minimize the errors, or uncertainty, in estimated design flows at ungauged sites is addressed. The benefit of estimating design flows with less uncertainty is that it decreases the probability of overestimating or underestimating the size of bridge or culvert openings. In this manner, optimizing the stream gauging network for regional information can reduce costs of highway design and maintenance. The approach to solving the problem is to define the objective, formulate the problem as a mathematical program, and find an approximate solution.

The technique is based on the regional regression method of estimating design flows, such as the 50-year peak flow, from physiographic characteristics, such as drainage area, channel slope, soil type, and land use; and meteorological variables, such as mean annual precipitation and 2-year, 24-hr rainfall intensity. The regional regression model is a multivariable regression model that can be written, after suitable transformations, in the linear form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

- $\mathbf{Y}$ = $(n \times 1)$ vector of flow characteristic at n sites,
- $\mathbf{X}$ = $(n \times p)$ matrix of $(p - 1)$ basin characteristics at the n sites augmented by a column of 1s,
- $\boldsymbol{\beta}$ = $(p \times 1)$ vector of regression parameters to be estimated, and
- $\mathbf{e}$ = $(n \times 1)$ vector of random errors.

The dependent variable is a flow characteristic, such as the logarithm of the 50-year flood (commonly used in the design of bridges and culverts), that is derived from a sample of observed flows, such as the logarithms of observed annual peak discharges at each site.

U.S. Geological Survey, 430 National Center, Reston, Va. 22092.

Hydrologists have used ordinary least squares (OLS) methods to estimate $\boldsymbol{\beta}$. The OLS estimates are appropriate and statistically efficient when the at-site flow estimates are equally reliable, the natural variability is the same at each site, and observed concurrent flows at every pair of sites are independent. In practice, hydrologic data sets are not so homogeneous.

A more appropriate estimator of $\boldsymbol{\beta}$ is the generalized least squares (GLS) estimator,

$$\boldsymbol{\beta} = (\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{Y}$$

in which it is assumed that the errors have zero mean $E[\mathbf{e}] = 0$, and covariance matrix $E[\mathbf{ee}^T] = \boldsymbol{\Lambda}$. The matrix $\boldsymbol{\Lambda}$ is a weighting matrix that accounts for unequal record lengths used for computing flow characteristics and for cross correlations between flow characteristics at different sites.

The operational difficulty with the GLS estimator of $\boldsymbol{\beta}$ is that $\boldsymbol{\Lambda}$ is unknown and must be estimated from the data. Stedinger and Tasker (*1*) proposed that $\boldsymbol{\Lambda}$ be estimated as

$$\boldsymbol{\Lambda} = \gamma^2\mathbf{I} + \mathbf{W}$$

where γ^2 is an estimate of the model error variance and $\mathbf{W}$ is an $(n \times n)$ matrix of sampling covariance. The model error variance (γ^2) is the portion of total error variance that is because of an imperfect model; it is not a function of the sample data. Model error occurs because many of the factors that explain variations in $\mathbf{Y}$ are unknown or imperfectly specified. Estimation of γ^2 is made by an interactive search technique outlined in Tasker and Stedinger (*2*). The sampling error variance $(\mathbf{W})$ is the error variance caused by estimating the regression coefficients $(\boldsymbol{\beta})$ with limited data in both the temporal and spatial dimensions.

The critical feature of this GLS technique is that it allows the covariance matrix $\boldsymbol{\Lambda}$ to be expressed in terms of record length at each gauge in the network. For the interested reader, the elements of the $\boldsymbol{\Lambda}$ matrix are given in Tasker and Stedinger (*2*, p. 363). For the problem at hand, it is enough to note that the values of the diagonal elements of the $\boldsymbol{\Lambda}$ matrix change as the record length associated with any gauging station change, and that the value of the off-diagonal element associated with any pair of stations is a function of the relative location of the sites. These changes reflect the decrease in variance of the response variable $(\mathbf{Y})$ as the record length increases, and the redundant information collected from cross-correlated annual peaks. Record length may refer to current record length or record length at the end of a planning horizon. Thus $\boldsymbol{\Lambda}$ may be calculated for different sets of record-length combinations that reflect different future data collection strategies.

DEFINITION OF OBJECTIVE

The first step in solving the problem at hand is to define the objective to be minimized. Let $\theta(\mathbf{x})$ represent the true streamflow statistic at a site with basin characteristics described by the column vector $\mathbf{x}$. The precision with which the regression estimate at the site $(\mathbf{x}\hat{\beta})$ approximates $\theta(\mathbf{x})$ can be described by the mean square error of prediction (MSE) at an ungauged site with basin characteristics $\mathbf{x}$

$$\mathrm{MSE}(\mathbf{x}) = \gamma^2 + \mathbf{x}^T(\mathbf{X}^T\Lambda^{-1}\mathbf{X})^{-1}\mathbf{x}$$

The regional statistical information contained in the regression model for the site is proportional to the reciprocal of $\mathrm{MSE}(\mathbf{x})$ (*3*).

Generally, interest lies in the average MSE over a representative set of basin characteristics rather than the MSE for a particular $\mathbf{x}$ as in the previous equation. Let that representative set be described by a set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Then the average MSE criterion would be

$$Z = \gamma^2 + m^{-1} \sum_{i=1}^{m} \mathrm{MSE}(\mathbf{x}_i)$$

Because γ^2 is not a function of how long gauges are operated, the first term on the right side of this equation may be ignored in the objective function Z.

Consider how to identify a feasible gauging strategy that minimizes Z. This analysis assumes that a gauging network already exists and is providing some data on which to base the design. Let $j = 1, \ldots, m_e$ index the existing stations of which the available record lengths are n_j. To this basic set can be added additional new gauging stations indexed by $j = m_e + 1, \ldots, m_e + m_a$, budgets permitting. The goal will be to identify some set, indexed by k, of these $m_e + m_a$ stations to continue to operate over the planning horizon (some 5 to 20 years). If the planning horizon is too short, it would seldom be advantageous to add a new gauge because it would not have an opportunity to accumulate a useful record. Thus, too short a time horizon will lead to myopic strategies that are suboptimal in the long run.

In designing an optimal data collection strategy, it must be recognized that some gauging stations will continue to be operated because of specific data needs, such as managing outflows from a reservoir or monitoring long-term trends. Other sites, perhaps, must be eliminated, such as a site that will be flooded by a new dam and can no longer be operated as a gauging station. Both of these types of gauging stations are termed "nondiscretionary" because the decision to continue or discontinue operation is made outside of this analysis. Of the $m_e + m_a$ gauging sites, let F equal the set of discretionary sites that one can choose to operate or not. The problem then is to find the feasible subset (S_k) of F that minimizes Z, subject to a budget limit (B) on discretionary gauging. Recall that $\mathbf{X}$ is the set of basin characteristics corresponding to both the existing gauging station and any additional gauging station sites specified by S_k, and Λ is the GLS weighting matrix pertaining to this set of gauging stations and record lengths that will pertain at end of the planning period. Fundamentally, the optimization of Z poses a large nonlinear integer programming problem. Because the number of subsets of F that are within the gauging budget can be large, a direct attack on the problem is not attractive. However, an approximate solution may be obtained using a step-backward approach (*4*). The step-backward approach starts by considering all possible gauges as being part of the network and incrementally dropping the least valuable stations until the remaining set is within the discretionary data collection budget constraint.

APPLICATION OF NETWORK ANALYSIS

In order to illustrate the network analysis procedure, a network of 85 gauges in Illinois was selected for analysis. The flow and basin characteristics for these sites were taken in a regional analysis of flood characteristics conducted by the U.S. Geological Survey in cooperation with the Illinois Department of Transportation (*4*). At some of these stations, much more data than annual peak flows are collected, but for this analysis only the annual peak flow regional information is considered. Thirty-six of the stations were selected as stations that must be continued, and two stations were selected as stations that cannot be continued. These 38 stations are classified as nondiscretionary (stations for which the choice of continuing or discontinuing operation was made before the analysis). The remaining 47 stations, along with 15 potentially new stations, are considered discretionary. The approximate locations of all these sites are shown in Figure 1, and flow and basin characteristics are presented in Table 1. The selection and classification of these gauges as discretionary or nondiscretionary were made for illustrative purposes and do not necessarily reflect the true classification of the gauges.

The first task in analysis of this network was to run a GLS regression of the 50-year peak discharge on basin characteristics using the methods described by Tasker and Stedinger (*2*). The final regression was

$$\begin{aligned}\log \mathrm{Q50} = {} & 2.1763 + 0.7538(\log \mathrm{AREA}) \\ & + 0.4767(\log \mathrm{SLOPE}) \\ & + 0.6791[\log (\mathrm{I24{,}2} - 2.5)]\end{aligned}$$

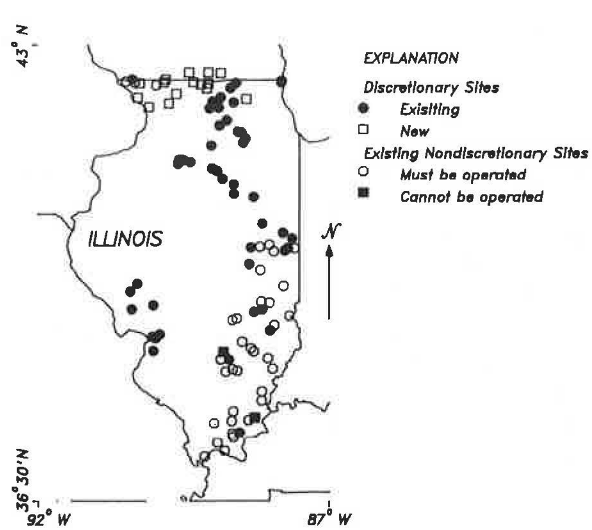

FIGURE 1 Location of gauging stations used in study.

TABLE 1 DATA USED IN REGRESSION

Station Number	AREA	SLOPE	I24,2	50-Yr Peak	Station Number	AREA	SLOPE	I24,2	50-Yr Peak
3336100	1.0	21.0	2.9	310.	5438300	0.8	87.3	2.8	254.
3336500	35.0	6.9	3.0	5100.	5438390	88.1	8.3	2.8	4140.
3336900	134.0	5.5	3.0	6610.	5438500	538.0	4.6	2.8	13400.
3337500	68.0	2.6	3.0	3510.	5438850	1.7	28.7	2.9	422.
3338000	340.0	3.0	3.0	9430.	5439000	77.7	2.7	2.9	2840.
3338100	2.2	15.8	3.0	758.	5439500	387.0	2.3	2.8	9720.
3338500	958.0	3.1	3.0	30500.	5439550	1.7	53.8	2.8	570.
3338800	1.3	33.2	3.0	1300.	5440000	1099.0	4.1	2.8	21600.
3339000	1290.0	3.2	3.0	36800.	5440500	117.0	6.3	2.9	8000.
3341700	1.1	44.3	3.1	588.	5551700	70.2	5.6	2.8	2030.
3341900	0.0	52.8	3.2	62.	5551800	0.4	87.1	2.9	541.
3343400	186.0	3.0	3.1	7110.	5551900	32.6	8.8	2.8	1640.
3344000	919.0	1.5	3.1	24200.	5551930	21.1	10.0	2.8	735.
3344250	0.1	10.5	3.2	75.	5554000	186.0	5.4	2.9	5940.
3344425	0.1	97.1	3.2	137.	5554500	579.0	1.1	3.0	14400.
3344500	7.6	15.7	3.2	3820.	5554600	0.2	60.7	3.0	186.
3345500	1516.0	1.6	3.2	46600.	5555000	1084.0	1.3	3.0	21900.
3346000	318.0	4.3	3.2	26100.	5555300	1251.0	1.4	3.0	36100.
3378000	228.0	2.8	3.3	6400.	5555400	0.1	50.4	3.0	292.
3378635	240.0	5.3	3.2	11000.	5555775	0.4	24.5	2.9	130.
3378650	1.6	19.6	3.3	806.	5556500	196.0	6.1	3.0	13000.
3378900	745.0	2.7	3.3	29600.	5557000	86.7	9.0	3.0	11100.
3378980	0.4	73.6	3.2	502.	5557100	0.3	97.1	3.0	396.
3379500	1131.0	2.0	3.3	49100.	5557500	99.0	12.7	3.0	8200.
3379650	1.6	36.1	3.3	1350.	5586350	1.8	53.9	3.4	1750.
3380300	0.1	98.7	3.4	119.	5586500	2.3	24.3	3.4	982.
3380350	208.0	2.8	3.3	19100.	5586850	0.0	63.4	3.4	48.
3380400	1.1	36.1	3.4	728.	5587000	868.0	2.3	3.4	36600.
3380450	0.4	87.6	3.4	352.	5587850	0.4	42.5	3.4	675.
3380475	97.2	4.1	3.4	9360.	5587900	212.0	5.1	3.4	9660.
3380500	464.0	1.9	3.4	29700.	5588000	36.7	7.9	3.5	8410.
3381500	3102.0	1.2	3.3	39100.	5589500	22.6	11.1	3.5	7280.
3381600	0.2	89.8	3.3	352.	5590000	12.4	17.2	3.0	1480.
3382025	0.5	75.5	3.5	516.	5590400	109.0	2.5	3.1	3720.
3382100	147.0	4.3	3.5	5860.					
3382170	13.3	12.2	3.4	2500.	--POSSIBLE NEW STATIONS--				
3382510	8.5	25.5	3.4	704.	990001	0.9	157.9	3.0	
3382520	1.1	28.3	3.4	772.	990002	247.0	10.9	3.0	
3384450	42.9	16.2	3.5	15800.	990003	230.0	6.6	3.0	
3385000	19.1	21.4	3.5	7780.	990004	3340.0	0.7	2.8	
3385500	1.0	145.2	3.5	1690.	990005	202.0	2.7	2.8	
3612000	244.0	2.7	3.5	11700.	990006	1034.0	2.3	2.9	
3612200	0.3	141.0	3.5	460.	990007	1.3	40.9	3.0	
3614000	2.0	23.9	3.6	887.	990008	1326.0	1.6	3.0	
4087300	1.5	34.3	2.7	357.	990009	2.0	29.4	3.0	
4087400	5.0	21.7	2.6	939.	990010	523.0	3.2	2.9	
5414820	39.6	18.9	3.0	14500.	990011	0.5	97.1	2.9	
5415000	125.0	11.3	3.0	17100.	990012	2550.0	0.9	2.9	
5415500	20.1	37.3	3.0	12400.	990013	6363.0	0.8	2.7	
5418750	1.9	35.2	3.0	700.	990014	2.2	40.3	2.8	
5438250	85.1	5.7	2.8	4360.	990015	14.4	7.4	2.7	

where Q50 is the 50-year peak discharge, AREA is the basin's drainage area, SLOPE is the average basin slope, and I24,2 is the 2-year, 24-hr rainfall intensity.

Next, the step-backward search procedure was run for two cases. In the first case, only 47 existing stations were considered as discretionary gauges. Recall that 38 of the 85 stations were not discretionary because they were selected as stations that must be continued or could not be continued. In the second case, 15 new stations were added to the set of discretionary stations. The two stations identified as stations that could not be continued are forced to be the first stations to be dropped by the step-backward search by giving them a relatively large cost. All stations were given a relative cost of 1 except for these two stations, which were given a cost of 1,000.

The results of the analysis for Cases 1 and 2 are presented in Tables 2 and 3, respectively. Figure 2 shows that, in this example, operating new stations provides a greater marginal decrease in the objective function (increase in regional information) than does operating only existing stations. The most effective stations to operate for the planning horizon can be determined by reading from the bottom to the top of Tables 2 and 3. That is, the last station listed for a given case is the best to operate, the next to last is the next best to operate, and so on.

CONCLUSIONS

The step-backward search method was presented as a procedure for the problem of deciding which gauging stations to operate in the future (including possible new stations) to minimize the sampling error in a regional hydrologic regression. The step-backward search method can be used to identify a feasible network design that is probably near the optimal; the method also gives the network manager a management tool that can (a) identify a nearly optimal gauging plan or network design, (b) provide insight into how much regional information is lost or gained by decisions to reduce or increase the operating budget, and (c) evaluate any proposed network design by comparing it with a nearly efficient design at the same budget level.

TABLE 2 COMPUTER RESULTS FOR CASE 1 IN WHICH NO NEW STATIONS ARE CONSIDERED AND FOR WHICH THE PLANNING HORIZON IS 20 YEARS

Step	Discontinued stations	Average sampling error	Relative Cost	Number of stations available for regression analysis
0	NONE	0.00223	2083.	85
1	3382520	0.00223	1083.	85
2	3380300	0.00223	83.	85
3	3345500	0.00223	82.	85
4	5555300	0.00223	81.	85
5	5590000	0.00223	80.	85
6	5587000	0.00223	79.	85
7	5554500	0.00223	78.	85
8	5557000	0.00223	77.	85
9	5556500	0.00223	76.	85
10	5557500	0.00223	75.	85
11	3338500	0.00223	74.	85
12	5415000	0.00223	73.	85
13	5555000	0.00223	72.	85
14	5554000	0.00223	71.	85
15	3380350	0.00223	70.	85
16	3336500	0.00223	69.	85
17	5590400	0.00223	68.	85
18	5440500	0.00223	67.	85
19	5588000	0.00223	66.	85
20	5589500	0.00223	65.	85
21	5439000	0.00223	64.	85
22	5439500	0.00223	63.	85
23	3338800	0.00223	62.	85
24	5438500	0.00223	61.	85
25	5440000	0.00223	60.	85
26	5586500	0.00223	59.	85
27	5438850	0.00223	58.	85
28	5587900	0.00223	57.	85
29	3338100	0.00224	56.	85
30	5555400	0.00224	55.	85
31	5551800	0.00224	54.	85
32	3336100	0.00225	53.	85
33	5557100	0.00225	52.	85
34	5551700	0.00226	51.	85
35	3384450	0.00226	50.	85
36	3344425	0.00226	49.	85
37	5438250	0.00227	48.	85
38	5551900	0.00227	47.	85
39	5587850	0.00228	46.	85
40	5554600	0.00229	45.	85
41	5438390	0.00230	44.	85
42	5586350	0.00232	43.	85
43	5551930	0.00233	42.	85
44	5586850	0.00234	41.	85
45	5555775	0.00236	40.	85
46	3344250	0.00238	39.	85
47	5438300	0.00239	38.	85
48	5439550	0.00241	37.	85
49	4087400	0.00247	36.	85

The following stations are not eligible to be discontinued:

3336900	3337500	3338000	3339000	3341700	3341900	3343400	3344000
3378000	3378635	3378650	3378900	3378980	3379500	3379650	3380400
3380500	3381500	3381600	3382025	3382100	3382170	3382510	3385000
3612200	3614000	4087300	5414820	5415500	5418750	3344500	3346000
3380450	3380475	3385500	3612000				

TABLE 3 COMPUTER RESULTS FOR CASE 2 IN WHICH 15 NEW STATIONS (NUMBERS BEGINNING WITH 9900) ARE CONSIDERED AS DISCRETIONARY SITES AND FOR WHICH THE PLANNING HORIZON IS 20 YEARS

Step	Discontinued stations	Average sampling error	Relative Cost	Number of stations available for regression analysis
0	NONE	0.00188	2098.	100
1	3380300	0.00188	1098.	100
2	3382520	0.00188	98.	100
3	5555300	0.00188	97.	100
4	3345500	0.00188	96.	100
5	5590000	0.00188	95.	100
6	5554500	0.00188	94.	100
7	990011	0.00190	93.	99
8	5557000	0.00190	92.	99
9	3338500	0.00190	91.	99
10	5556500	0.00190	90.	99
11	5557500	0.00190	89.	99
12	5554000	0.00190	88.	99
13	5555000	0.00190	87.	99
14	5587000	0.00189	86.	99
15	5415000	0.00189	85.	99
16	3336500	0.00189	84.	99
17	5440500	0.00189	83.	99
18	5439500	0.00189	82.	99
19	5439000	0.00189	81.	99
20	5590400	0.00189	80.	99
21	3380350	0.00189	79.	99
22	5438500	0.00189	78.	99
23	5440000	0.00189	77.	99
24	3338800	0.00190	76.	99
25	5588000	0.00189	75.	99
26	5589500	0.00189	74.	99
27	990008	0.00190	73.	98
28	5438850	0.00190	72.	98
29	5551700	0.00190	71.	98
30	5551900	0.00190	70.	98
31	5586500	0.00191	69.	98
32	990006	0.00191	68.	97
33	990003	0.00192	67.	96
34	3338100	0.00193	66.	96
35	5438250	0.00193	65.	96
36	3336100	0.00193	64.	96
37	5557100	0.00194	63.	96
38	5587900	0.00194	62.	96
39	5551800	0.00194	61.	96
40	990010	0.00195	60.	95
41	990014	0.00197	59.	94
42	990009	0.00199	58.	93
43	990007	0.00201	57.	92
44	5438390	0.00201	56.	92
45	3384450	0.00201	55.	92
46	5551930	0.00202	54.	92
47	5555400	0.00202	53.	92
48	990002	0.00204	52.	91
49	3344425	0.00205	51.	91
50	990012	0.00206	50.	90
51	990005	0.00208	49.	89
52	5587850	0.00209	48.	89
53	5554600	0.00210	47.	89
54	5586350	0.00211	46.	89
55	5555775	0.00212	45.	89
56	990004	0.00215	44.	88
57	5439550	0.00216	43.	88
58	5438300	0.00217	42.	88
59	5586850	0.00219	41.	88
60	990001	0.00224	40.	87
61	3344250	0.00226	39.	87
62	990013	0.00233	38.	86
63	990015	0.00241	37.	85
64	4087400	0.00247	36.	85

The following stations are not eligible to be discontinued:

3336900	3337500	3338000	3339000	3341700	3341900	3343400	3344000
3378000	3378635	3378650	3378900	3378980	3379500	3379650	3380400
3380500	3381500	3381600	3382025	3382100	3382170	3382510	3385000
3612200	3614000	4087300	5414820	5415500	5418750	3344500	3346000
3380450	3380475	3385500	3612000				

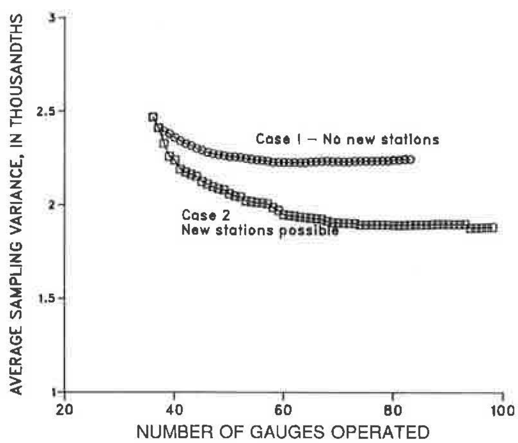

FIGURE 2 Results of network analysis.

REFERENCES

1. J. R. Stedinger and G. D. Tasker. Regional Hydrologic Analysis I. Ordinary, Weighted, and Generalized Least Squares Compared. *Water Resources Research*, Vol. 21, No. 9 1985, pp. I421–I432.
2. G. D. Tasker and J. R. Stedinger. An Operational GLS Model for Hydrologic Regression: *Journal of Hydrology*, Vol. III, 1989, pp. 361–375.
3. R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, London, 1949.
4. G. W. Curtis. Technique for Estimating Flood-Peak Discharges and Frequencies on Rural Streams: Water Resources Investigations Report 87-4207, U.S. Geological Survey, 1987, 79 pp.

Publication of this paper sponsored by Committee on Hydrology, Hydraulics, and Water Quality.