STATISTICAL MODELS FOR ESTIMATING DAILY STREAMFLOW IN MICHIGAN

By D.J. Holtschlag, U.S. Geological Survey, and

Habib Salehi, Michigan State University

U.S. GEOLOGICAL SURVEY

Water–Resources Investigations Report 91– 4194

Prepared in cooperation with the

MICHIGAN DEPARTMENT OF NATURAL RESOURCES

Lansing, Michigan

1992

DEPARTMENT OF THE INTERIOR

MANUEL LUJAN, JR., Secretary

U.S. GEOLOGICAL SURVEY

Dallas L. Peck, Director

For additional information
write to:

District Chief
U.S. Geological Survey
6520 Mercantile Way, Suite 5
Lansing, MI 48911

Copies of this report can
be purchased from:

U.S. Geological Survey
Books and Open-file Reports Section
Federal Center, Building 810
Box 25425
Denver, CO 80225–042

CONTENTS

ILLUSTRATIONS

ILLUSTRATIONS—Continued

TABLES

CONVERSION FACTORS

Multiply	By	To obtain
	Length	
foot (ft)	0.3048	meter
mile (mi)	1.609	kilometer
	Area	
square foot (ft^2)	0.09294	square meter
square mile (mi^2)	2.590	square kilometer
	Volume	
gallon (gal)	3.785	liter
gallon (gal)	0.003785	cubic meter
cubic foot (ft^3)	0.02832	cubic meter
	Flow	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second
cubic foot per second per square mile [$(ft^3/s)/mi^2$]	0.01093	cubic meter per second per square kilometer

STATISTICAL MODELS FOR ESTIMATING DAILY STREAMFLOW IN MICHIGAN

by D.J. Holtschlag and Habib Salehi

ABSTRACT

Statistical models for estimating daily streamflow were analyzed for 25 pairs of streamflow-gaging stations in Michigan. Stations were paired by randomly choosing a station operated in 1989 at which 10 or more years of continuous flow data had been collected and at which flow is virtually unregulated; a nearby station was chosen where flow characteristics are similar. Streamflow data from the 25 randomly selected stations were used as the response variables; streamflow data at the nearby stations were used to generate a set of explanatory variables.

Ordinary-least–squares regression (OLSR) equations, autoregressive integrated moving-average (ARIMA) equations, and transfer–function-noise (TFN) equations were developed to estimate the log transform of flow for the 25 randomly selected stations. The precision of each type of equation was evaluated on the basis of the standard deviation of the estimation errors. OLSR equations produce one set of estimation errors; ARIMA and TFN models each produce l sets of estimation errors corresponding to the forecast lead. The lead-l forecast is the estimate of flow l days ahead of the most recent streamflow used as a response variable in the estimation. In this analysis, the standard deviation of lead–l ARIMA and TFN forecast errors were generally lower than the standard deviation of OLSR errors for $l \leq 2$ days and $l \leq 9$ days, respectively.

Composite estimates were computed as a weighted average of forecasts based on TFN equations and backcasts (forecasts of the reverse–ordered series) based on ARIMA equations. The standard deviation of composite errors varied throughout the length of the estimation interval and generally was at maximum near the center of the interval. For comparison with OLSR errors, the mean standard deviation of composite errors were computed for intervals of length 1 to 40 days. The mean standard deviation of length–l composite errors were generally less than the standard deviation of the OLSR errors for $l \leq 32$ days. In addition, the composite estimates ensure a gradual transition between periods of estimated and measured flows.

Model performance among stations of differing model error magnitudes were compared by computing ratios of the mean standard deviation of the length–l composite errors to the standard deviation of OLSR errors. The mean error ratio for the set of 25 selected stations was less than 1 for intervals $l \leq 32$ days. Considering the frequency characteristics of the length of intervals of estimated record in Michigan, the effective mean error ratio for intervals ≤ 30 days was 0.52. Thus, for intervals of estimation of 1 month or less, the error of the composite estimate is substantially lower than error of the OLSR estimate.

INTRODUCTION

The U.S. Geological Survey (USGS) operates hydrologic data–collection stations nationwide to provide all levels of government, the private sector, and the general public with water–resources information. Daily mean streamflow is a major component of this data-collection program. In the United States in 1989, streamflow records were published for 7,239 continuous–record gaging stations, including records for 140 stations in Michigan (Condes de la Torre, 1989).

Daily mean streamflow is computed on the basis of hourly or more frequent measurements of water-surface elevation (stage) and a rating curve that defines the relation between stage and discharge for a particular stream. The rating curve is developed from periodic measurements of stage and discharge obtained when flow is not affected by variable backwater. The backwater effect is the reduction in flow expected at a specific stage that is attributable to an obstruction in the channel. Periodic measurements are used as a basis for adjusting the rating curve to account for minor changes in the hydraulic characteristics of a stream channel.

At many gaging stations, parts of the flow record are estimated each year because of lost stage record or variable backwater. Record is lost generally because of malfunction of sensing or recording equipment. In Michigan, channel ice is a major cause of variable backwater. The accuracy of the estimates affects the utility of the flow records and the operation of streamflow–gaging stations.

Estimates of daily flow can be computed by use of ordinary least–squares regression (OLSR) equations or by use of hydrologic flow–routing models (Scott and Moss, 1986, p. 298). OLSR equations are developed from recorded streamflows at a station where estimates are needed; these flows are referred to as the dependent or response variable. Flow data from a nearby station are used as a source of independent or explanatory variables. OLSR equations commonly include two or more lagged series of flows from the nearby station as explanatory variables. Lagged series are created by shifting flows ahead or behind their measured time of occurrence by one or more days.

OLSR models are based on the assumption of independent observations; however, consecutive daily streamflows are generally autocorrelated. This discrepancy between the OLSR model and data characteristics increases the difficulty of identifying the explanatory streamflow lags essential for inclusion in the regression equation. In addition, the transition between periods of regression estimates and measured values is seldom smooth. Results of streamflow estimation in Michigan (Holtschlag, 1985, p. 14) indicate that large errors associated with OLSR estimates generally limit potential use of the equations.

Errors in hydrologic flow–routing models are similar to errors found in OLSR models (Scott and Moss, 1986, p. 304).

Alternatively, stochastic models describe the relation between input and output in dynamic systems. For a univariate time series, the input, which can be described by a white–noise series, can be related to the output, or the observed time series, by an autoregressive integrated moving-average (ARIMA) equation. Bivariate input systems, consisting of a white-noise series and an explanatory series, can be related to a single output series by a transfer–function-noise (TFN) equation.

Because stochastic-model assumptions are more consistent with streamflow-data characteristics than OLSR-model assumptions, stochastic equations are more likely to result in adequate estimates than are OLSR equations. In the past, difficulties in the identification and estimation of stochastic equations have limited their use. Recently the AUTOBOX[1] program (Automatic Forecasting Systems, 1988), a computer–based expert system, has been developed to automate the development of ARIMA and TFN equations. Because of its ease of use and its potential for widespread application, the AUTOBOX program was used in this evaluation.

This study was conducted in cooperation with the Michigan Department of Natural Resources to help maintain the accuracy of streamflow records in a cost–effective manner.

Purpose and Scope

The purpose of this report is to discuss the feasibility of using stochastic models for estimating daily mean streamflow in Michigan. Feasibility was evaluated on the basis of a comparison of the relative accuracy of estimates obtained from OLSR equations and estimates determined from stochastic equations for 25 randomly selected streamflow-gaging stations in Michigan. The effect of data transformations and the effect of inclusion of explicit trend and seasonal components in the equations also was assessed.

[1]Use of trade names in this report is for identification only and does not constitute endorsement by the U.S. Geological Survey.

Streamflow data

Twenty–five streamflow-gaging stations (table 1, figs. 1 and 2) were randomly selected from U.S. Geological Survey stations operated in Michigan; each had been in continuous operation for 10 or more years through water year 1989. The stations were restricted to streams in which flow was virtually unregulated. In this report, flow at the randomly selected station is referred to as the response variable; for brevity, the randomly selected gaging station is referred to as the response station. For each response station, a second station in close proximity to the first was selected to aid in the estimation of the response variable. These nearby (explanatory) stations were selected from stations typically used by USGS personnel in Michigan for comparison in estimating flow. Flow data from the nearby stations was used to generate a set of explanatory variables.

Daily streamflow data for 5 years (water years 1985–89) were used in the analysis. Data from the first 4 years were used for model calibration; data from the fifth year were used for model verification. Periods of estimated record, described in the annual water data report (U.S. Geological Survey, 1986–1990) were documented. These data describe 1,012 periods of estimated record with a mean duration of 14 days per period. Thus, about 15 percent of the daily streamflow was estimated.

A high percentage of the periods of estimated record were of short duration (fig. 3); about 65 percent of the periods of estimated record were 1 week or shorter. Further analysis of estimated record characteristics indicated that the distribution of estimated record varied seasonally and peaked in January and February (fig. 4). The peak indicates that the probability of variable backwater and equipment failures is greatest in winter.

Table 1.—Selected U.S Geological Survey streamflow–gaging stations in Michigan

[Map number refers to figs. 1 and 2. Flow estimates at the response station can be based, at least in part, on flow from the corresponding explanatory station]

Response station			Explanatory station		
Map number	Station number	Stream name	Map number	Station number	Stream name
1–R	04040500	Sturgeon River near Sidnaw	1–E	04043050	Trap Rock River near Lake Linden
2–R	04056500	Manistique River near Manistique	2–E	04045500	Tahquamenon River near Tahquamenon Paradise
3–R	04059500	Ford River near Hyde	3–E	04059000	Escanaba River at Cornell
4–R	04061500	Paint River at Crystal Falls	4–E	04033000	Middle Branch Ontonagon River near Paulding
5–R	04096400	St. Joseph River near Burlington	5–E	04096600	Coldwater River near Hodunk
6–R	04096515	South Branch Hog Creek near Allen	6–E	04096900	Nottawa Creek near Athens
7–R	04102500	Paw Paw River at Riverside	7–E	04101500	St. Joseph River at Niles
8–R	04105000	Battle Creek at Battle Creek	8–E	04105500	Kalamazoo River near Battle Creek
9–R	04111500	Deer Creek near Dansville	9–E	04111379	Red Cedar River near Williamston
10–R	04113000	Grand River at Lansing	10–E	04109000	Grand River at Jackson
11–R	04114500	Looking Glass River near Eagle	11–E	04116000	Grand River at Ionia
12–R	04115000	Maple River at Maple Rapids	12–E	04117500	Thornapple River near Hastings
13–R	04122100	Bear Creek near Muskegon	13–E	04121900	Little Muskegon River near Morley
14–R	04122200	White River near Whitehall	14–E	04121500	Muskegon River at Evart
15–R	04122500	Pere Marquette River at Scottville	15–E	04121300	Clam River at Vogel Center
16–R	04127918	Pine River near Rudyard	16–E	04057510	Sturgeon River near Nahma Junction
17–R	04128000	Sturgeon River near Wolverine	17–E	04127800	Jordan River near East Jordan
18–R	04135500	Au Sable River at Grayling	18–E	04135700	South Branch Au Sable River near Luzerne
19–R	04146000	Farmers Creek near Lapeer	19–E	04146063	South Branch Flint River near Columbiaville
20–R	04160570	North Branch Belle River at Imlay City	20–E	04160600	Belle River at Memphis

Table 1.—Selected U.S. Geological Survey streamflow-gaging stations in Michigan—Continued

Response station			Explanatory station		
Map number	Station number	Stream name	Map number	Station number	Stream name
21–R	04163400	Plum Brook at Utica	21–E	04161100	Galloway Creek near Auburn Heights
22–R	04164000	Clinton River near Fraser	22–E	04161540	Paint Creek at Rochester
23–R	04164100	East Pond Creek at Romeo	23–E	04161580	Stony Creek near Romeo
24–R	04164500	North Branch Clinton River near Mount Clemens	24–E	04168000	Lower River Rouge at Inkster
25–R	04166000	River Rouge at Birmingham	25–E	04166100	River Rouge at Southfield

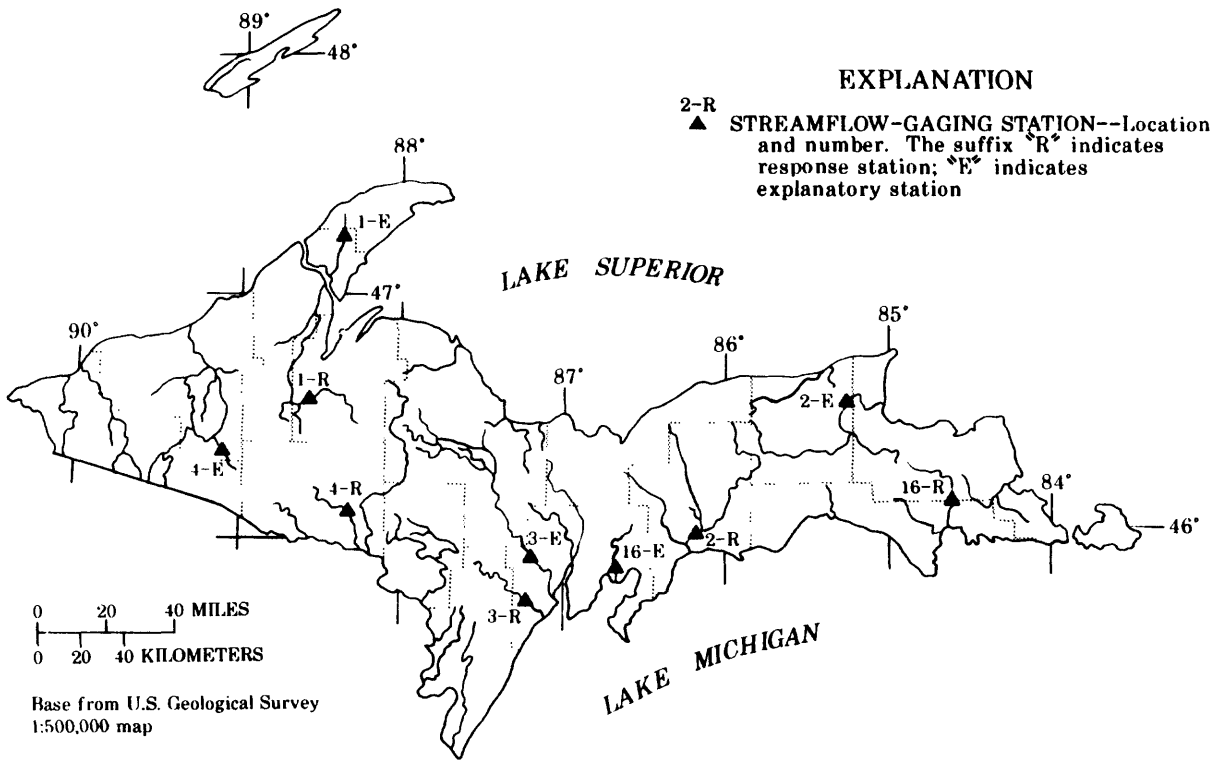

Figure 1.—Locations of selected streamflow-gaging stations in the Upper Peninsula of Michigan. (See Table 1 for U.S. Geological Survey station names and numbers.)

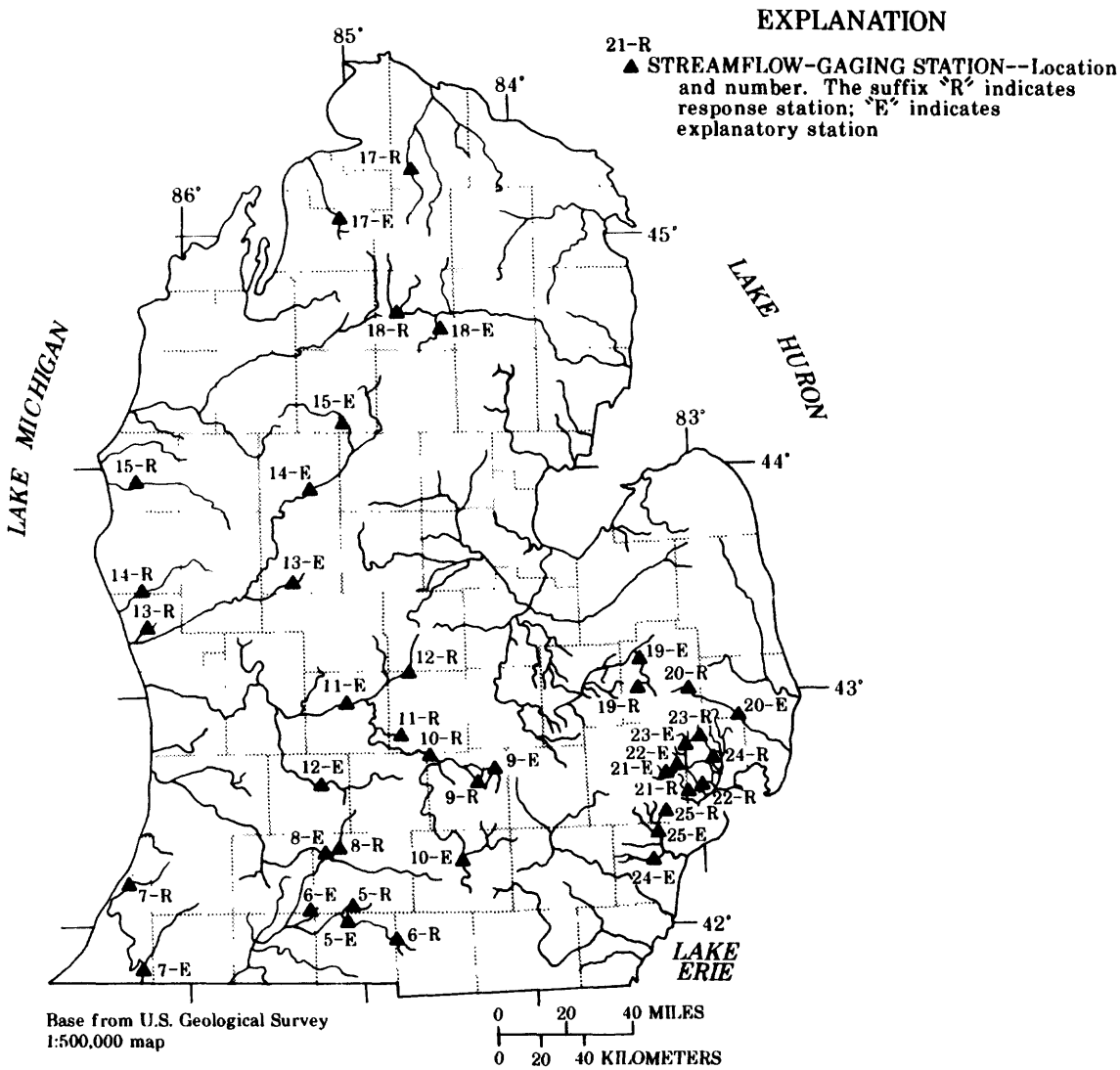

Figure 2.—Locations of selected streamflow-gaging stations in the Lower Peninsula of Michigan. (See Table 1 for U.S. Geological Survey station names and numbers.)

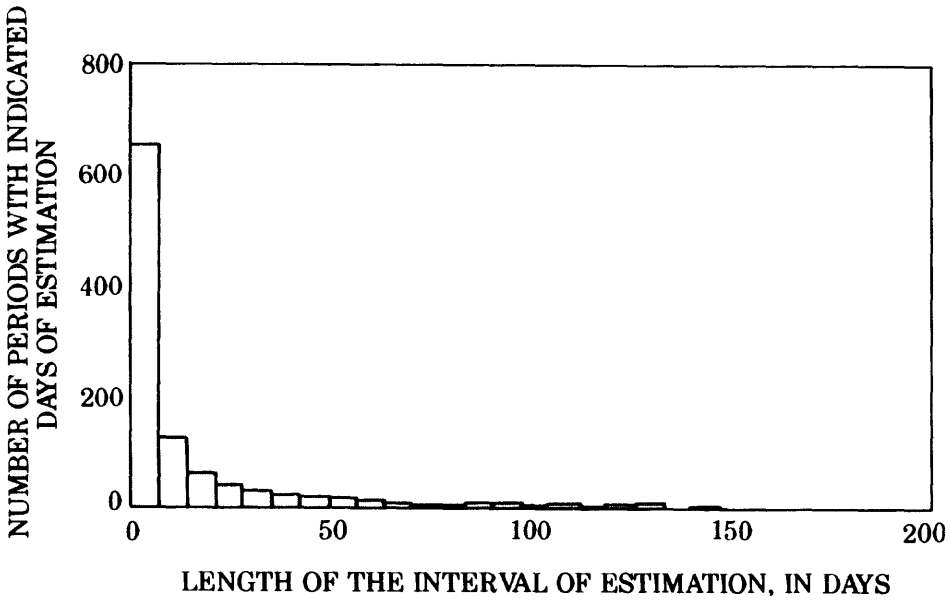

Figure 3.—Histogram of the length of intervals of estimated streamflow record.

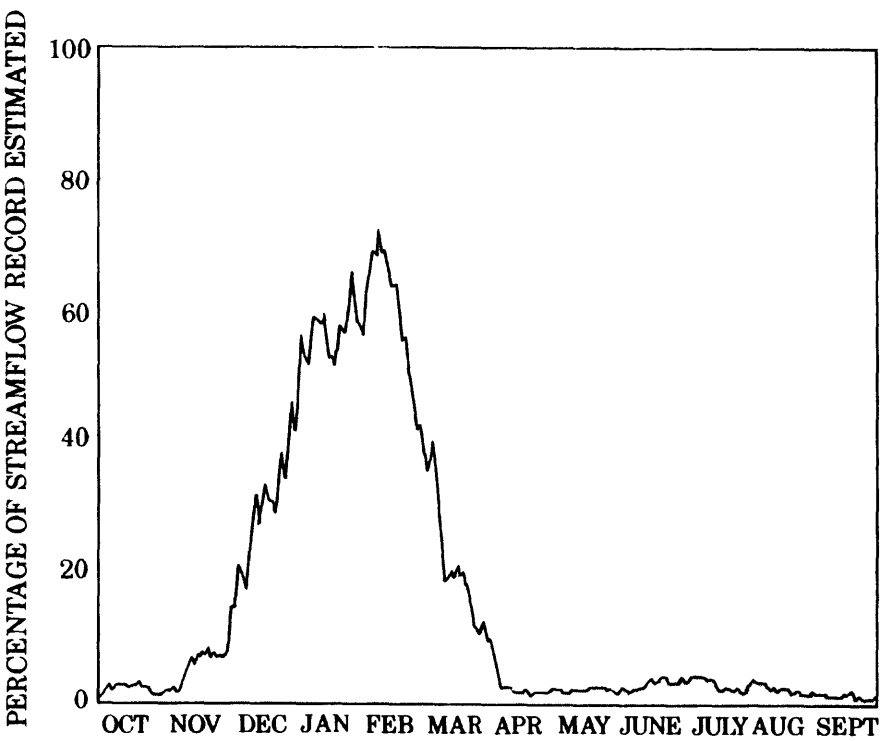

Figure 4.—Seasonal distribution of estimated streamflow record.

STATISTICAL MODELS FOR ESTIMATING DAILY STREAMFLOW

The OLSR models and the stochastic models discussed in this report are linear statistical models with additive error components. Both classes of models are based on the assumption that the variability of flow values can be disaggregated as

$$y_t = M_t + S_t + z_t \beta + e_t , \tag{1}$$

where

y_t is the daily mean streamflow at time t (in this analysis, upper-case letters X_t and Y_t indicate streamflows in cubic feet per second, at the explanatory and response sites; lower-case letters x_t and y_t indicate a transformed metric);

M_t is the deterministic trend component of y_t;

S_t is the seasonal component of y_t;

z_t is a p-dimensional row vector that includes a value of 1 as the first element and transformed streamflows as additional elements (z_t may include values of x_{t-l} for small integers l and values of y_{t-l} for small positive whole numbers l);

β is a p-dimensional column vector of coefficients relating streamflows in z_t to y_t; and

e_t is the model-error component generally assumed to be normally distributed and independent with mean zero and variance σ_e^2, $e_t \sim N(0,\sigma_e^2)$.

OLSR models and stochastic models differ fundamentally in the variates that can potentially be included in the vector z, the method for identifying appropriate variates for inclusion, and the technique for estimating the coefficient vector, β. The form of equation 1 is not typically used for the development of stochastic equations, but it can be derived by algebraic manipulation from a more parsimonious form to facilitate estimation.

Premodeling Considerations

Statistical models are developed to describe natural phenomena in a simple and useful form. Simplicity is commonly measured by the number of parameters in the estimating equation; usefulness, by the accuracy of the estimates. The nonlinear data transformations and techniques for trend and seasonal estimation discussed in the following sections aid in the development of statistical models.

Data Transformations

Nonlinear data transformations are often applied so that the equation relating response and explanatory variables is parsimonious (Box and Draper, 1987, p. 288) and so that model variance is homoscedastic. The effect of such transformations is generally to expand the data scale in one part of the range and to contract it in another. The choice of data transformation is related to the distribution of the response and explanatory variates and to the relation between the two variates.

The frequency distribution of daily mean streamflow data (referred to as "daily streamflow data" hereafter) is typically skewed to the right (fig. 5). A natural logarithm (log) transformation, which contracts the high end of the range, is commonly applied to facilitate model development (Riggs, 1968, p. 10); however, the log transformation corresponds to only one value in a family of power transformations, $x_t^{\lambda_x}$, $y_t^{\lambda_y}$. Maximum likelihood methods for choosing the power-transformation parameters λ_x and λ_y are described by Box and Draper (1987, p. 289).

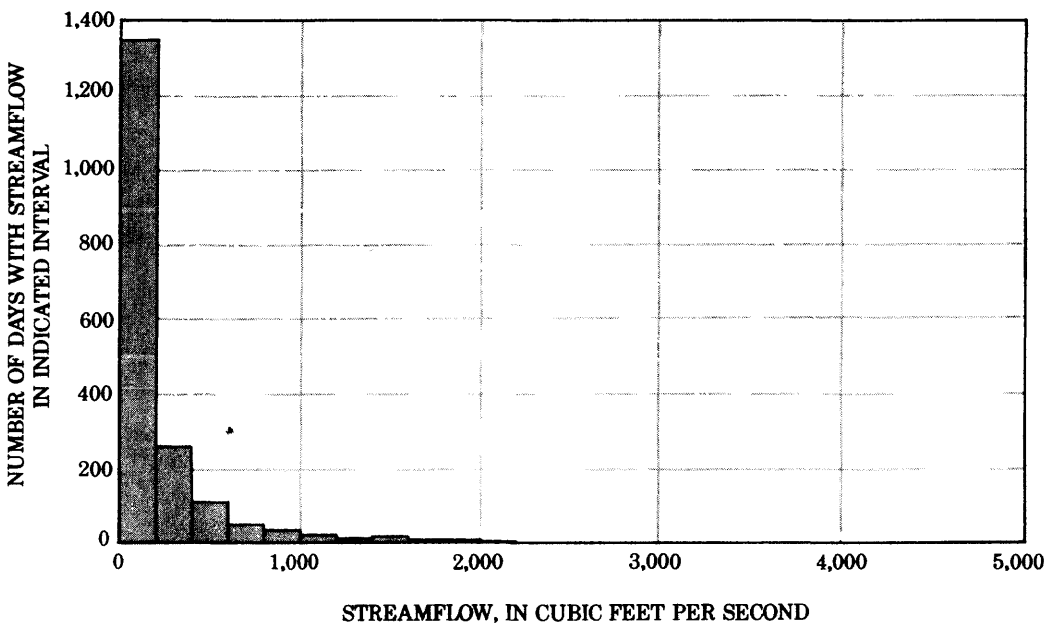

Figure 5.—Histogram of daily streamflows of Sturgeon River near Sidnaw, water years 1985-89.

The avas transformation (Statistical Sciences, 1990, chap. 5, p. 14) is a generalization of the power transformation. The avas transformation is composed of a pair of nonlinear–monotonic functions, f and g, chosen so that the model $f(Y_t) = g(X_t)$ results in an additive, homoscedastic model–error component. Although the avas transformations varied among station pairs, log- and avas-transformed streamflows tended to differ primarily by a constant (fig. 6). The adequacy of the log transformation was assessed by comparing the accuracy of the OLSR equations based on log transformation with the accuracy of equations based on avas transformation.

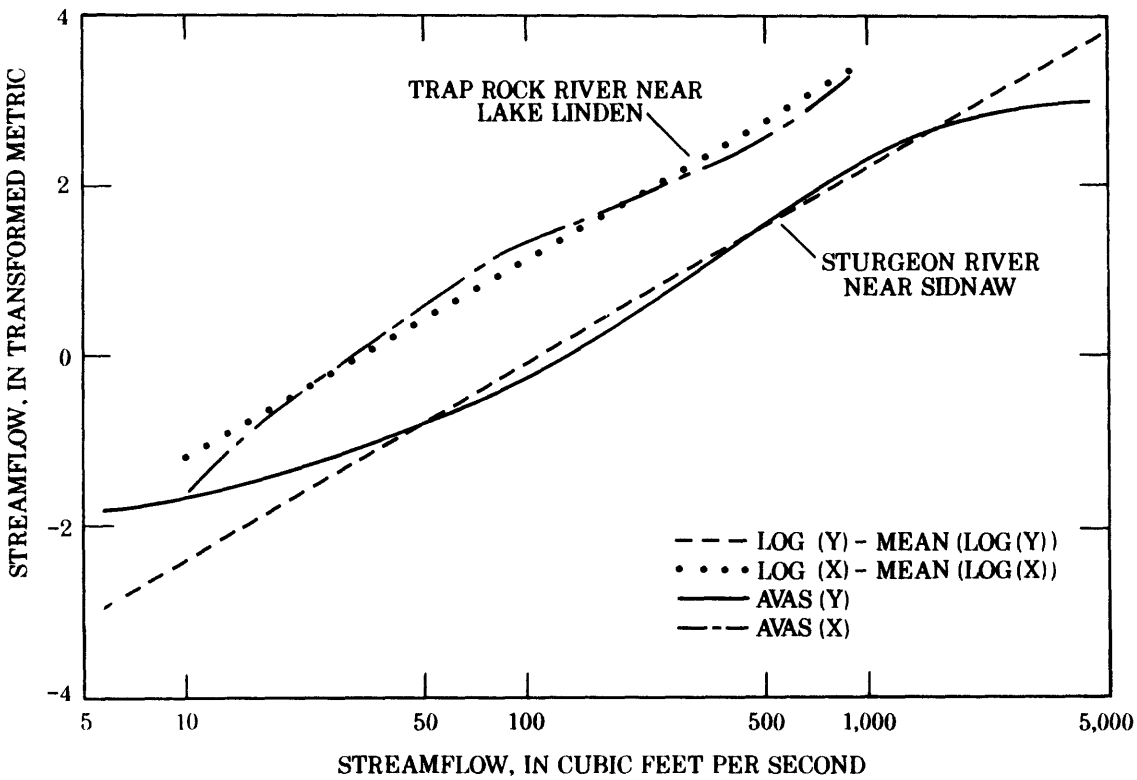

Figure 6.—Relation between streamflows and the log and avas transformations of streamflows, Sturgeon River near Sidnaw and Trap Rock River near Lake Linden.

Trend and Seasonal Components

In time-series modeling, the trend and seasonal components in equation 1 are commonly approximated by use of a deterministic function of time and are subtracted from the streamflows before the dependence on the explanatory series is modeled. The trend is typically approximated by a low–order polynomial equation, whereas the seasonal component is approximated by a Fourier series (Box and Jenkins, 1976, p. 301). Because of the dynamic characteristics of streamflow data, however, the true significance level of parameters describing the trend or seasonality are generally unknown.

In this study, trend components were generally not evident by inspection of the streamflow series. For completeness, the possibility of a linear-trend component was included in the development of the stochastic equations. Residuals from all model were inspected for trends.

A prominent seasonal component was apparent in all of the streamflow series (fig. 7). Because use of seasonal models containing parameters with uncertain statistical significance is undesirable, seasonal components were described by means of a variable span moving–average (VSMA) vector (Statistical Sciences, 1990, chap. 5, p. 45). The

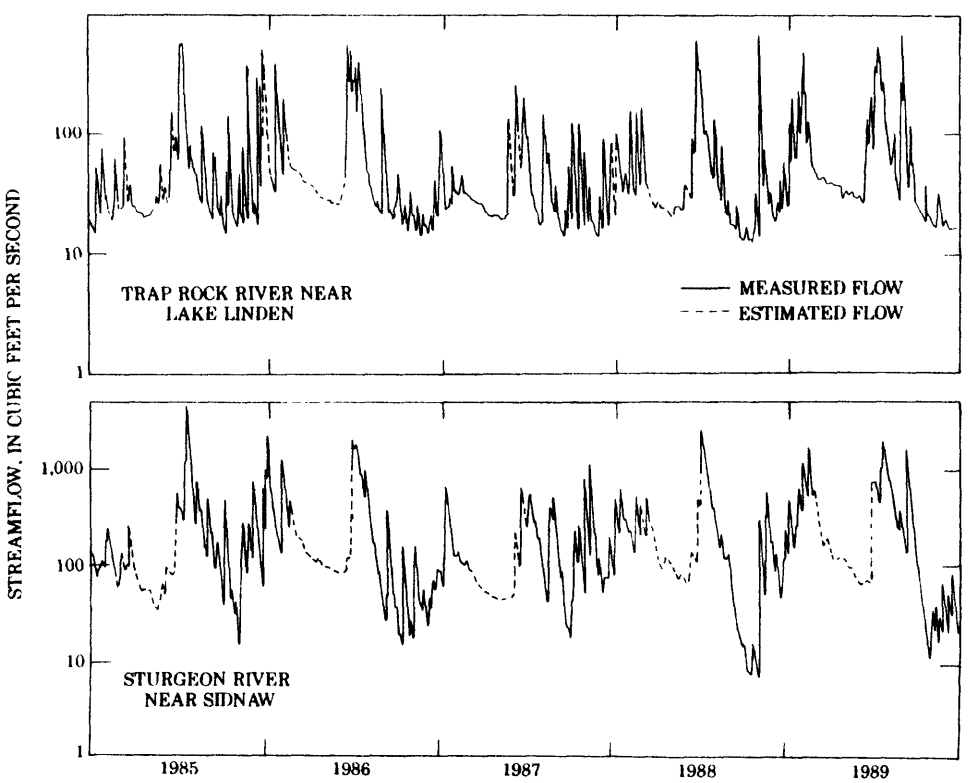

Figure 7.—Streamflow of Sturgeon River near Sidnaw and Trap Rock River near Lake Linden, water years 1985-89.

VSMA vector is computed from estimates of the log of daily streamflow based on the average of four log-transformed flow values for each day of the year during the calibration period. A moving average of adjacent daily averages was used to estimate a VSMA vector through the 365 daily values. The span of the moving average changed with the local curvature and the variability of residuals about the VSMA vector. Smoothing parameters, used to control the width of the span, were based on local cross–validation analysis. VSMA vectors were constrained to maintain a 365–day period.

Scatter plots (fig. 8) showed that the VSMA vectors were consistent with the major features of the apparent seasonal variability of daily averages and that VSMA vectors for paired stations were similar in shape. OLSR and stochastic models were developed with and without seasonal components estimated by the vectors to assess the need for an explicit seasonal component.

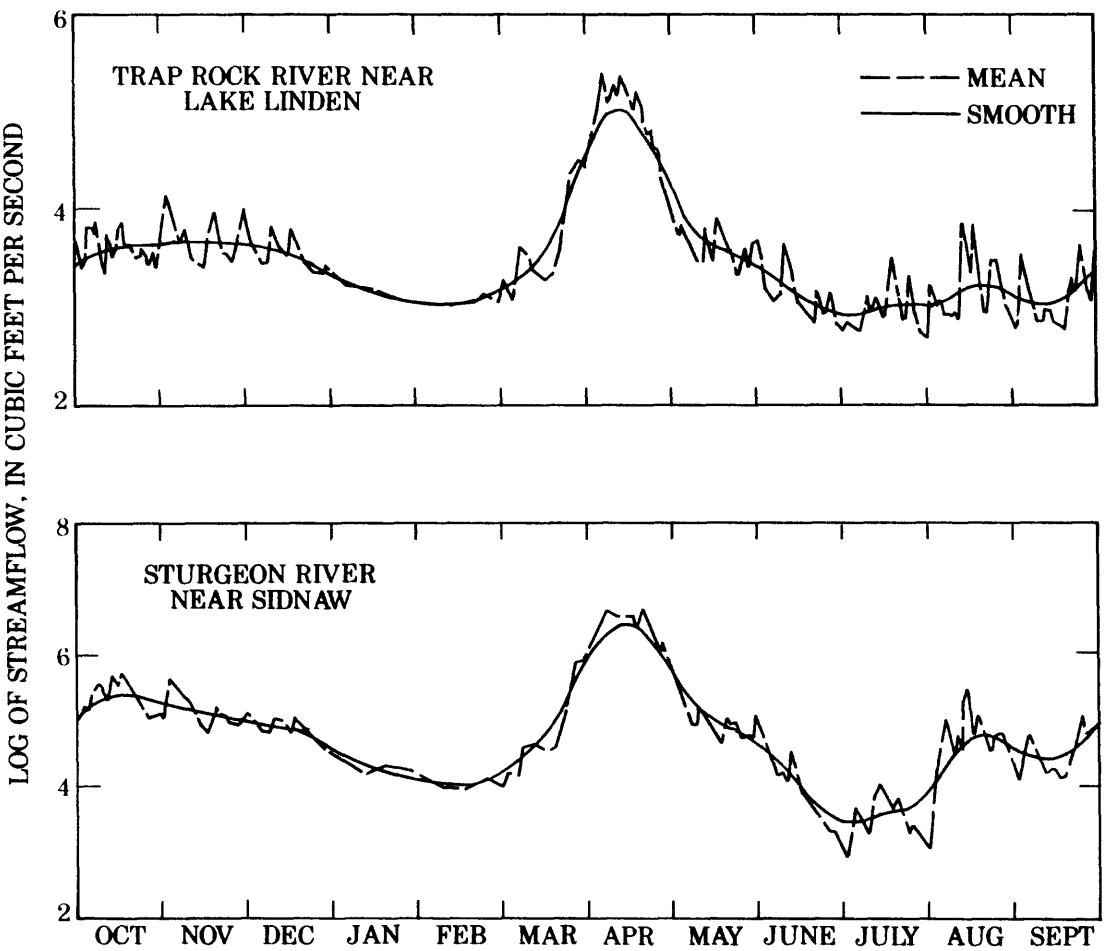

Figure 8.—Estimated seasonal components of streamflow, Sturgeon River near Sidnaw and Trap Rock River near Lake Linden.

Ordinary–Least–Squares Regression Model

OLSR models have been used extensively in hydrology because they are objective and easy to apply, they produce indices of accuracy, and they are generally accepted as a tool for estimation. In a nationwide investigation of the cost–effectiveness of the U.S. Geological Survey streamflow–gaging program (Fontaine and others, 1984, p. 13), OLSR models were evaluated as an alternative to field data collection for determining daily streamflow; however, because of the inaccuracies of the resulting streamflow estimates, few gaging-station records were reproduced with sufficient accuracy to be considered viable alternatives to measured streamflow data (Scott and Moss, 1986, p. 303). The formulation and assumptions associated with the OLSR model are discussed in the following section.

Formulation

The OLSR model can be written without trend or seasonal components as

$$y_t = \boldsymbol{x}_t \boldsymbol{\beta} + e_t , \tag{2}$$

where

y_t is the response variable at time t;

$\boldsymbol{x}_t$ is the value of a p–dimensional row vector at time t (The vector contains 1 as its first element and explanatory streamflows from an interval of width 12 days and includes $x_{t+4}, x_{t+3}, \cdots, x_{t-7}$ for each y_t. This interval was chosen to include all significant cross–correlations between $\frac{\Phi_x(B)}{\Theta_x(B)} y_t$ and $\frac{\Phi_x(B)}{\Theta_x(B)} x_t$ for the 25 paired stations. The rational polynomial $\frac{\Phi_x(B)}{\Theta_x(B)}$ is the estimated linear filter that transforms x_t to a white-noise series α_t (Box and Jenkins, 1976, p. 380). Here, B is the backshift operator such that $B^m x_t = x_{t-m}$. Values of x_t are assumed to be measured without error in the OLSR model.);

$\boldsymbol{\beta}$ is a column vector of coefficients, of the same length as $\boldsymbol{x}_t$, relating streamflow at a response site to streamflow at the explanatory site (an intercept term is included as the first element), and

e_t is the OLSR error at time t where $e_t \sim \mathcal{NI}(0, \sigma_e^2)$. (This form of the error component implies that the error vector $\boldsymbol{e} = [e_1 \; e_2 \; e_3 \; \cdots \; e_n]'$ has the property that $\mathcal{E}(\boldsymbol{e}\boldsymbol{e}') = \sigma_e^2 \mathcal{I}_n$ where $\mathcal{I}_n$ is an $n \times n$–dimensional identity matrix and $\mathcal{E}$ is the expected value operator; the prime symbol associated with a matrix indicates the transposition operation.

OLSR estimates of the coefficient vector β are computed as

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{3}$$

where X is an $n \times p$ matrix of transformed explanatory streamflows and y is an n–dimensional column vector of transformed response streamflows (Beck and Arnold, 1977, p. 235). The number n corresponds to the number of days of streamflow values used in the estimation. The covariance of the estimated coefficient vector $\hat{\beta}$ is

$$\mathrm{Cov}(\hat{\beta}) = (X'X)^{-1}X'\mathcal{E}(ee')X(X'X)^{-1}. \tag{4}$$

For uncorrelated errors, such that $\mathcal{E}(ee') = \sigma_e^2 \mathcal{I}_n$, $\hat{\beta}$ is the minimum variance estimator of β, and the covariance of $\hat{\beta}$ is

$$\mathrm{Cov}(\hat{\beta}) = \sigma_e^2(X'X)^{-1}; \tag{5}$$

however, if the residuals are correlated, then the OLSR estimate of $\hat{\beta}$ is inefficient (Beck and Arnold, 1977, p. 239), and the $\mathrm{Cov}(\hat{\beta})$ is difficult to assess.

The OLSR estimate of transformed streamflow for time t is computed as $\hat{y}_t = x_t\hat{\beta}$. In addition to the point estimate, probability limits can be used to describe a random interval that has a probability of $1-a$ of containing y_t, where a is a specified significance level. For large numbers of measurements used in estimation, the probability interval for a new observation based on the OLSR equation and the assumption of independent, constant-variance errors can be computed as

$$\hat{y}_t \pm \mathcal{N}_{a/2}\sqrt{\hat{\sigma}_e^2\left[1 + x_t(X'X)^{-1}x_t'\right]}, \tag{6}$$

where $\mathcal{N}_{a/2}$ is the upper $100(a/2)^{\text{th}}$ percentile of a normal distribution (Johnson and Wickern, 1982, p. 311). A sample estimate of σ_e is referred to as the root mean square of the OLSR estimation errors (RMSE). In general, the width of the probability interval increases as the RMSE and the distance between $\mathbf{x}_t$ and the mean, $\bar{\mathbf{x}}$, ($\|x_t - \bar{x}\|$), increase; however, as the number of measurements used in the estimation increases, the interval width tends to become nearly constant over a wide range of streamflow magnitudes.

Implementation

OLSR equations were developed by selecting an appropriate subset of explanatory streamflow variates from a full equation. The full equation included an intercept term and 12 positive and negative lags of streamflow from the nearby station $(x_{t+4}, x_{t+3}, \ldots, x_{t-7})$ as explanatory variates for each value of the response variable (y_t) at time t. The RMSE of the full equation was used to provide an unbiased estimate of σ_e for the true equation, σ_{ef}. The true equation is defined here as the regression equation containing the appropriate subset of explanatory variates.

The full equation tended to include explanatory variables that had little computed statistical significance. More parsimonious equations, which included subsets of the explanatory series from the full equation, were identified by means of the C_p statistic (Daniel and Wood, 1980, p. 86),

$$C_p = \frac{RSS_p}{\hat{\sigma}^2_{ef}} - (n - 2p) , \qquad (7)$$

where RSS_p is the residual sum of squares for an OLSR model containing an intercept term and $p-1$ explanatory variables. Results for equations with small bias tend to cluster about the line $C_p = p$.

As a further means of excluding unnecessary parameters, final selection criteria for the OLSR equations included the requirement that all parameters, except the intercept, have a individual significance level (P–value) of 5 percent or less. In addition, a minimum increase of 0.001 in the coefficient of determination (R^2) also was required to increase the number of parameters in an estimation equation.

OLSR equations developed from log-transformed streamflows (table 2) contain an average of 4.3 explanatory variates. The most commonly included lags of the explanatory series were at $t+4$ (60 percent), t (100 percent), $t-1$ (56 percent), and $t-7$ (64 percent) relative to the streamflow response at time t. Selection of explanatory series at times t and $t-1$ indicate that the explanatory and response series were virtually contemporaneous. The extreme lags were likely included because they were the least redundant among available series after inclusion of the explanatory series near time t. Inclusion of even more extreme explanatory lags in the full equation is unlikely to improve estimating equations substantially because of the general decrease in cross–correlation with increasing lag.

Table 2.—Ordinary–least–squares regression equations for estimating daily streamflow

[x_t and y_t are log transformed daily streamflows at the explanatory and response sites; B is the backshift operator, $B^m z_t = z_{t-m}$; F is the forward shift operator, $F^m z_t = z_{t+m}$; e_t is a residual series; $\hat{\sigma}_e$ is the estimated standard error of e_t; and $\mathcal{R}^2$ is the coefficient of determination]

Response station	Explanatory station	Equation $\hat{\sigma}_e$	Equation $\mathcal{R}^2$	Ordinary–least–squares regression equation
04040500	04043050	0.5241	0.749	$y_t = 0.1458 + (0.1154F^3 + 0.5067 + 0.3362B + 0.1912B^3 + 0.1621B^5)\, x_t + e_t$
04056500	04045500	.2312	.850	$y_t = 1.5443 + (0.2010F^4 + 0.5241 + 0.1099B^7)\, x_t + e_t$
04059500	04059000	.3216	.895	$y_t = -3.6561 + (0.2132F^4 + 0.5652 + 0.3744B + 0.2490B^3)\, x_t + e_t$
04061500	04033000	.1919	.890	$y_t = -0.4188 + (0.0924F^4 + 0.6206 + 0.4793B + 0.1095B^5)\, x_t + e_t$
04096400	04096600	.2148	.926	$y_t = 0.7019 + (0.1175F^4 + 0.6117 + 0.0841B^4)\, x_t + e_t$
04096515	04096900	.4609	.845	$y_t = -5.3031 + (0.2453F^4 + 1.5523 - 0.4429B + 0.3647B^7)\, x_t + e_t$
04102500	04101500	.1958	.808	$y_t = 0.2341 + (0.1362F^4 + 0.2532F + 0.2693 + 0.2157B + 0.1593B^3 - 0.3079B^7)\, x_t + e_t$
04105000	04105500	.1550	.961	$y_t = -4.6954 + (0.2060F + 0.5298 + 0.3938B + 0.3326B^2 + 0.0643B^7)\, x_t + e_t$
04111500	04111379	.4303	.906	$y_t = -3.8078 + (0.1522F^4 + 0.5731F + 1.3134 - 0.7365B)\, x_t + e_t$
04113000	04109000	.2673	.903	$y_t = 0.0088 + (0.1973F^4 + 0.1250F^3 + 0.4377 + 0.2417B + 0.3007B^3 + 0.0932B^7)\, x_t + e_t$
04114500	04116000	.2905	.899	$y_t = -2.9855 + (0.4924F + 0.3337 + 0.2408B^7)\, x_t + e_t$
04115000	04117500	.6622	.735	$y_t = -3.3127 + (0.1763F^3 + 0.9580 + 0.3650B^7)\, x_t + e_t$
04122100	04121900	.4469	.707	$y_t = -3.6039 + (0.0914F^4 + 0.3312F + 0.8952 - 0.2251B + 0.2083B^7)\, x_t + e_t$
04122200	04121500	.2113	.761	$y_t = 2.0762 + (0.0818F^4 + 0.3147 + 0.4666B - 0.1588B^3 + -0.1145B^7)\, x_t + e_t$
04122500	04121300	.1894	.769	$y_t = 3.4023 + (0.1092F^4 + 0.3631 + 0.2841B^2 - 0.0766B^7)\, x_t + e_t$
04127918	04057510	.4166	.701	$y_t = 0.4439 + (0.2097F^4 + 0.8972 - 0.1651B^7)\, x_t + e_t$

Table 2.—Ordinary–least–squares regression equations for estimating daily streamflow—Continued

Response station	Explanatory station	Equation $\hat{\sigma}_e$	Equation R^2	Ordinary–least–squares regression equation
04128000	04127800	.0916	.853	$y_t = -2.6376 + (0.6954 + 0.4321B + 0.0733B^2 + 0.0737B^3 + 0.0732B^4 + 0.0891B^5 + 0.1042B^7)\, x_t + e_t$
04135500	04135700	.1099	.792	$y_t = 1.7594 + (0.5916 + 0.1490B - 0.1224B^4 - 0.1111B^7)\, x_t + e_t$
04146000	04146063	.3825	.900	$y_t = -3.2123 + (0.4482F + 0.5259 + 0.3202B^2)\, x_t + e_t$
04160570	04160600	.3653	.881	$y_t = -1.2114 + (0.4153F + 0.4284)\, x_t + e_t$
04163400	04161100	.5315	.803	$y_t = -0.0994 + (0.0557F^4 + 0.9656 - 0.0466B^3)\, x_t + e_t$
04164000	04161540	.2721	.822	$y_t = 2.9550 + (1.008 + 0.1863B - 0.1366B^2 - 0.1278B^3 - 0.1100B^5 - 0.0590B^7)\, x_t + e_t$
04164100	04161580	0.2911	0.874	$y_t = 0.4804 + (0.1124F^4 + 0.7017 + 0.0568B^7)\, x_t + e_t$
04164500	04168000	.7759	.744	$y_t = 0.3683 + (0.0844F^4 + 0.2107 + 0.3056B + 0.1939B^2 + 0.1052B^3 + 0.1155B^5 + 0.1645B^7)\, x_t + e_t$
04166000	04166100	.1924	.932	$y_t = -0.8647 + (-0.0340F^2 + 0.1290F + 0.7906 + 0.0303B^2 + 0.0341B^5)\, x_t + e_t$

Application

Estimation of daily streamflow by means of the OLSR equation is illustrated with log-transformed streamflow from station 04040500 as the response variable and log-transformed flow from station 04043050 as the explanatory variable. On the basis of the C_p plot (fig. 9) a six-parameter and a seven-parameter model were evaluated. Individual $\mathcal{P}$–values of all explanatory series in the six-parameter model were less than 0.0001; $\mathcal{P}$–values of all explanatory series in the seven-parameter model were less than 0.002. The increase in $\mathcal{R}^2$ from the six-parameter model (0.7489) to the seven-parameter model (0.7497), however, was less than the specified 0.001. Therefore, the six-parameter model was selected.

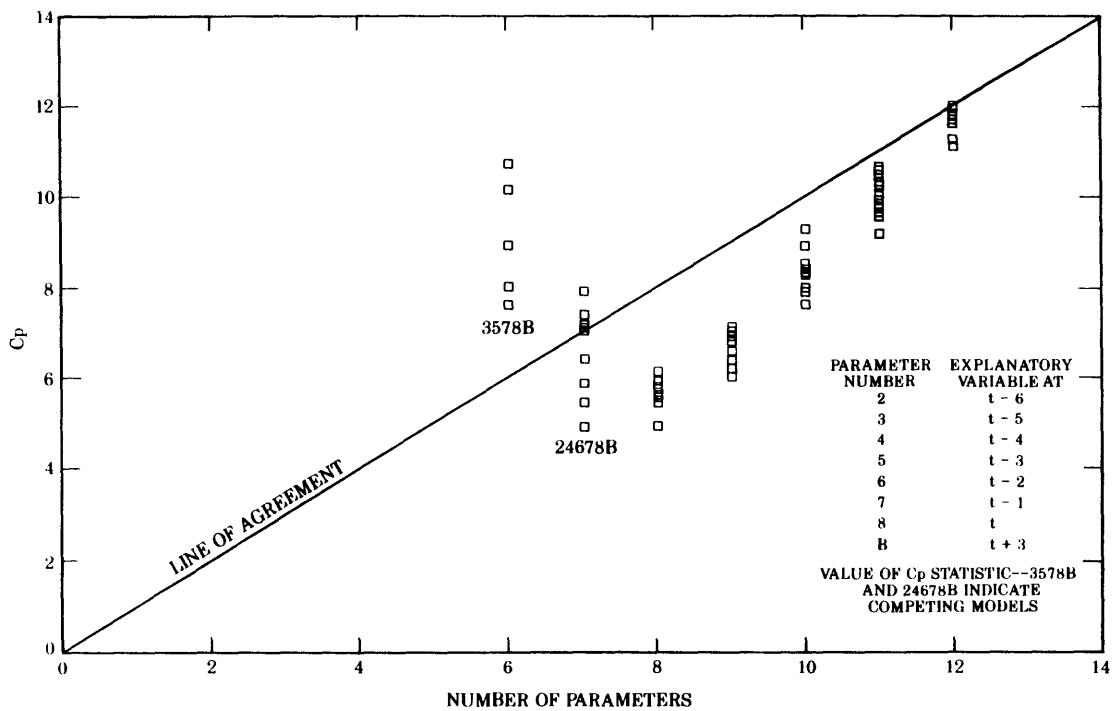

Figure 9.—C_p for alternative ordinary–least–squares regression equations for estimating streamflow at Sturgeon River near Sidnaw on the basis of streamflow at Trap Rock River near Lake Linden.

The OLSR estimate of the log of streamflow for October 1, 1988, based on the selected equation and log-transformed streamflow at station 04043050 from September 26 through October 4, 1988, is

$$\hat{y} = \boldsymbol{x}\hat{\boldsymbol{\beta}} = [1 \ \log(35) \ \log(21) \ \log(22) \ \log(25) \ \log(21)] \begin{bmatrix} 0.14576 \\ 0.11542 \\ 0.50674 \\ 0.33624 \\ 0.19120 \\ 0.16210 \end{bmatrix} = 4.2472.$$

The exponentiated estimate, $\exp(\hat{y})$, is 69.9 ft^3/s.

The 95–percent probability interval obtained from the calibration data is

$$4.2472 \pm 1.96\sqrt{0.52412^2(1 + 0.0016716)} = [3.2191, 5.2753],$$

which corresponds to an exponentiated interval of [25.0 ft^3/s, 195 ft^3/s]. Although the measured flow on October 1, 1988 of 131 ft^3/s is contained within this interval, the interval is wide, and the correlation of residuals in time is highly positive (fig. 10).

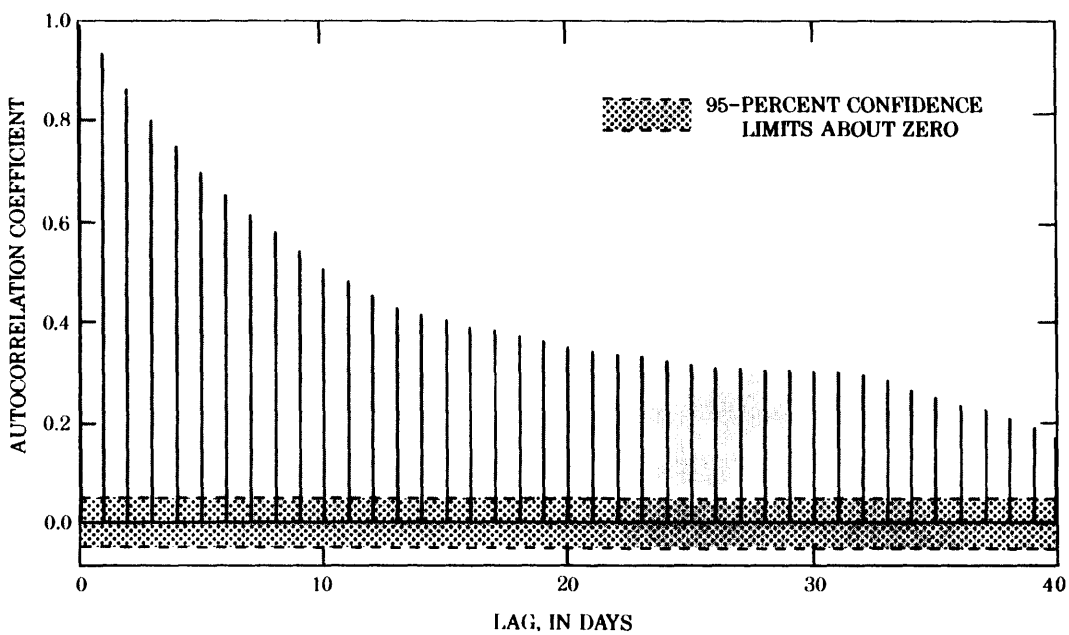

Figure 10.—Correlogram of residuals of the selected ordinary–least–squares regression equation for estimating streamflow at Sturgeon River near Sidnaw.

Stochastic Models

Two types of stochastic equations were developed by use of AUTOBOX as part of this analysis. Univariate processes were described by ARIMA equations, and bivariate - input-univariate output processes were described by TFN equations. Data from water years 1985–88 were used in identification and estimation stochastic models. Sample autocorrelation and partial autocorrelation functions were computed up to lag 12 to aid in identification. All estimated parameters were statistically significant at the 5–percent level. Diagnostic checks for model sufficiency, parameter necessity, and invertibility were computed automatically for each tentatively identified model. Because of the large period of the seasonality of daily values, seasonality components were not included within the ARIMA or TFN models.

Autoregressive Integrated Moving-Average Models

Formulation

The general form of an ARIMA model used to describe the univariate processes is

$$\nabla_z (z_t - \mu_z) = M_t + \frac{\Theta_z(B)}{\Phi_z(B)} \alpha_t , \tag{8}$$

where

z_t refers to either the explanatory or the response series;

μ_z is the mean of the univariate series;

∇_z is the backward difference operator on the univariate series (This operator causes lagged values of the univariate series to be subtracted from itself a specified number of times. The backward difference operator, ∇, equals $(1-B^o)^d$ where the order, o, corresponds to the lag used in the differencing and the degree, d, specifies the number of consecutive times that the differencing operator is applied. Differencing operators were applied as needed to create stationarity in the univariate series.);

M_t is a deterministic trend;

$\Phi_z(B)$ is an autoregressive operator of order p associated with variate z, where

$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$;

$\Theta_z(B)$ is a moving-average operator of order q associated with variate z, where $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$; and

α_t is a white-noise process associated with the ARIMA model of variate z. (This process is generally assumed to be distributed as a $NI(0,\sigma_\alpha^2)$. Values of α_t are computed as $z_t - \breve{z}_{t-1}(1)$, where $\breve{z}_{t-1}(1)$ is the lead–1 forecast of z_t at time origin t–1 based on the ARIMA equation.)

The lead-l forecast is the estimate of streamflow l-days ahead of the most recent streamflow used as a response variable in the estimation. Forecasts, expressed as a weighted sum of of previous streamflows, can be computed readily from the model. The lead–l ARIMA forecasts, $\breve{z}_t(l)$, can be computed by letting

$$[z_{t+l}] = \begin{cases} z_{t+l} & l \leq 0 \\ \breve{z}_t(l) & l > 0 \end{cases} \tag{9}$$

(simply replacing unknown z's by forecasts) and replacing the unknown α_t by zero (Box and Jenkins, 1976, p. 307).

For a series with μ_t=0, moving-average operator, $\Theta_z(B)$; autoregressive operator $\Phi_z(B)$; and differencing operator, ∇_z; equation 8 can be written

$$\Theta_z^{-1}(B)\, \Phi_z(B)\, z_t \nabla_z = \Pi(B)\, z_t = \alpha_t \tag{10}$$

for $\Pi(B) = 1 + \sum_{j=1}^{\infty} \pi_j B^j$. The π_j weights are computed by equating coefficients in equation 10. The ARIMA lead–l forecast is then computed recursively as

$$\breve{z}_t(l) = \sum_{j=1}^{\infty} \pi_j [z_{t+l-j}] \,. \tag{11}$$

The estimated variance of the ARIMA lead–l forecast error is

$$\breve{\sigma}_e^2(l) = \breve{\sigma}_\alpha^2 \sum_{j=0}^{l-1} \psi_j^2 \tag{12}$$

for $\Psi(B) = \Pi^{-1}(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$.

Probability limits associated with the ARIMA model forecast are a function of the forecast lead-l and are computed as

$$\breve{z}_t(l) \pm \mu_{\alpha/2}\, \breve{\sigma}_e(l) \,. \tag{13}$$

Implementation

ARIMA equations were developed by use of AUTOBOX and log-transformed streamflows obtained from the 50 selected gaging stations (table 3). First-order first–degree differencing operators were used in 72 percent of the equations. No other differencing operators were included. In equations without differencing operators, the mean was estimated and removed.

Integrated moving-average (IMA) equations were identified for 44 percent of the stations; autoregressive integrated (ARI) equations were developed for 28 percent of the stations; autoregressive (AR) equations accounted for 20 percent of the stations; the remaining 8 percent of the stations were described by autoregressive moving-average (ARMA) equations. The coefficient of determination associated with equations containing differencing operators was higher than that for equations lacking differencing operators (fig. 11). No significant deterministic trend components were identified. On the basis of criteria specified within AUTOBOX, the assumption of independence and normality of equation residuals (α_t) cannot be rejected.

Application

The univariate model developed from log-transformed streamflow at station 04040500 includes one differencing operator and one moving-average operator. The differencing operator is of first order and first degree. The moving-average operator consists of one parameter with a value of –0.3897 associated with a backorder power of 1.

Computation of forecasts can be facilitated by converting the rational polynomial equation (by algebraic manipulation) to the regression form of equation 11 as

$$\breve{y}_t(l) \approx 1.390[y_{t+l-1}] - 0.542[y_{t+l-2}] + 0.211[y_{t+l-3}] - 0.0822[y_{t+l-4}] + 0.0320[y_{t+l-5}].$$

The lead–1 estimate of the log of streamflow for October 1, 1988, based on the IMA model is 4.950. The corresponding exponentiated estimate, $\exp(\breve{y}_t(1))$, is 141.1 ft^3/s.

A 95–percent probability interval is obtained by computing coefficients $\Psi(B) = 1 + 1.3897B + 1.3897B^2 + \cdots$ and applying equation (12). The lead–1 95–percent probability interval equals $4.949 \pm 1.96 \times 1 \times 0.216 = [4.527, 5.372]$, which corresponds to an exponentiated interval of [94.6 ft^3/s, 215 ft^3/s]. This interval excludes the OLSR estimate, and it is 59 percent narrower than the corresponding regression interval; however, the estimated standard deviation of the forecast error and associated probability interval width increases sharply with the forecast lead. In this case, the standard deviation of the lead–4 forecast error exceeds the RMSE of the OLSR equation (fig. 12). TFN models, discussed subsequently, combine some of the desirable properties of ARIMA and OLSR estimators.

Table 3.—Autoregressive integrated moving–average equations for estimating daily streamflow

[z_t is log transformed daily streamflow; B is the backshift operator, $B^m z_t = z_{t-m}$; ∇ is the backward difference operator, $z_t\nabla = z_t - z_{t-1}$; α_t is a white-noise series; $\breve{\sigma}_e(1)$ is the estimated standard deviation of the lead–1 forecast error; and $\mathcal{R}^2$ is the coefficient of determination for the lead–1 forecast]

Station number	Equation $\breve{\sigma}_e(1)$	Equation $\mathcal{R}^2$	Autoregressive integrated moving–average equation
04033000	0.1032	0.943	$(1 - 0.4296B + 0.2626B^2)z_t\nabla = \alpha_t$
04040500	.2157	.957	$z_t\nabla = (1 + 0.3897B)\alpha_t$
04043050	.2667	.872	$(1 - 1.1273B + 0.3895B^2 - 0.1948B^3)(z_t - 3.4564) = \alpha_t$
04045500	.0490	.995	$(1 - 0.9526B + 0.3291B^2 - 0.0995B^3)z_t\nabla = \alpha_t$
04056500	.0372	.996	$(1 - 1.0290B + 0.3927B^2 - 0.0797B^3)z_t\nabla = \alpha_t$
04057510	.1175	.969	$(1 - 0.4970B + 0.2430B^2)z_t\nabla = \alpha_t$
04059000	.1210	.969	$z_t\nabla = (1 + 0.4906B + 0.1646B^2)\alpha_t$
04059500	.1110	.987	$z_t\nabla = (1 + 0.7317B + 0.3098B^2 + 0.0984B^3)\alpha_t$
04061500	.1524	.930	$z_t\nabla = (1 + 0.1051B - 0.0399B^2)\alpha_t$
04096400	.0903	.987	$z_t\nabla = (1 + 0.5620B + 0.1527B^2)\alpha_t$
04096515	.1779	.977	$(1 - 0.4046B + 0.2144B^2)z_t\nabla = \alpha_t$
04096600	.1178	.985	$z_t\nabla = (1 + 0.5727B + 0.0948B^2)\alpha_t$
04096900	.0835	.984	$(1 - 0.8544B + 0.3526B^2)z_t\nabla = \alpha_t$
04101500	.1045	.961	$z_t\nabla = (1 + 0.0552B - 0.0916B^2)\alpha_t$
04102500	.0620	.981	$z_t\nabla = (1 + 0.6493B)\alpha_t$
04105000	.1007	.983	$z_t\nabla = (1 + 0.6480B + 0.2849B^2 + 0.1153B^3)\alpha_t$
04105500	.1094	.955	$(1 - 0.2112B)z_t\nabla = \alpha_t$
04109000	.1816	.914	$z_t\nabla = (1 - 0.1056B^2 - 0.0930B^3)\alpha_t$
04111379	.1549	.977	$z_t\nabla = (1 + 0.5074B + 0.0790B^2)\alpha_t$
04111500	.3488	.938	$z_t\nabla = (1 + 0.2134B - 0.2246B^2 - 0.09462B^3)\alpha_t$
04113000	.2251	.931	$(1 - 0.7904B - 0.2256B^2 + 0.04871B^4)(z_t - 6.4776) = \alpha_t$
04114500	.1699	.966	$(1 - 0.2458B + 0.1718B^2 - 0.0754B^3)z_t\nabla = \alpha_t$
04115000	.1513	.986	$z_t\nabla = (1 + 0.6752B + 0.2248B^2 + 0.0813B^3)\alpha_t$
04116000	.1888	.951	$(1 - 1.0310B + 0.0598B^3)(z_t - 7.4194) = \alpha_t$
04117500	.0754	.991	$(1 - 2.0117B + 1.4203B^2 - 0.3953B^3)(z_t - 5.6118) = \alpha_t$

Table 3.—Autoregressive integrated moving–average equations for estimating daily streamflow—Continued

Station number	Equation $\breve{\sigma}_e(1)$	Equation $\mathcal{R}^2$	Autoregressive integrated moving–average equation
04121300	0.0856	0.973	$z_t\nabla = (1 + 0.6504B - 0.1819B^3)\alpha_t$
04121500	.0776	.984	$z_t\nabla = (1 + 0.5994B + 0.1441B^2)\alpha_t$
04121900	.1471	.931	$(1 - 0.4106B + 0.2876B^2 + 0.0803B^4)z_t\nabla = \alpha_t$
04122100	.2356	.918	$z_t\nabla = (1 + 0.0957B - 0.2780B^2 - 0.0982B^3)\alpha_t$
04122200	.0777	.967	$z_t\nabla = (1 + 0.6533B + 0.2002B^2 - 0.0582B^3)\alpha_t$
04122500	.0483	.985	$z_t\nabla = (1 + 0.8130B + 0.4997B^2 + 0.1906B^3)\alpha_t$
04127800	.1193	.522	$(1 - 0.6536B - 0.1295B^4)(z_t - 5.2349) = \alpha_t$
04127918	.1958	.934	$(1 - 0.3029B + 0.1704B^2)z_t\nabla = \alpha_t$
04128000	.0990	.827	$(1 - 1.0924B + 0.4254B^2 - 0.2459B^3)(z_t - 5.4289) = \alpha_t$
04135500	.0491	.958	$z_t\nabla = (1 + 0.6732B - 0.1450B^3)\alpha_t$
04135700	.0568	.979	$z_t\nabla = (1 + 0.4871B + 0.08545B^2)\alpha_t$
04146000	.1798	.978	$z_t\nabla = (1 + 0.4468B + 0.1654B^2 + 0.0871B^3)\alpha_t$
04146063	.1633	.967	$z_t\nabla = (1 + 0.4393B - 0.0463B^3)\alpha_t$
04160570	.3156	.912	$z_t\nabla = (1 + 0.1475B - 0.1798B^2 - 0.1351B^3)\alpha_t$
04160600	.2774	.945	$(1 - 0.4373B + 0.2845B^2)z_t\nabla = \alpha_t$
04161100	.4901	.805	$(1 - 1.0085B + 0.2756B^2 - 0.1680B^3)(z_t - 3.2088) = \alpha_t$
04161540	.2396	.882	$(1 - 0.9160B - 0.0528B^3)(z_t - 3.8090) = (1 - 0.2681B^2)\alpha_t$
04161580	.2449	.928	$(1 - 0.1294B + 0.1490B^2)z_t\nabla = \alpha_t$
04163400	.5796	.765	$(1 - 0.9576B + 0.2666B^2 - 0.1941B^3)(z_t - 5.8942) = \alpha_t$
04164000	.3491	.706	$(1 - 0.7991B - 0.1143B^4)(z_t - 5.8512) = (1 - 0.1965B^2)\alpha_t$
04164100	.2385	.915	$(1 - 0.1009B + 0.1827B^2)z_t\nabla = \alpha_t$
04164500	.2964	.962	$(1 - 0.5921B + 0.3598B^2)z_t\nabla = \alpha_t$
04166000	.4023	.703	$(1 - 0.8436B - 0.0685B^4)(z_t - 2.9150) = (1 - 0.2407B^2)\alpha_t$
04166100	.4609	.651	$(1 - 0.8007B - 0.0878B^4)(z_t - 3.9794) = (1 - 0.2029B^2)\alpha_t$
04168000	.6266	.772	$(1 - 0.9940B + 0.3339B^2 - 0.2285B^3)(z_t - 4.2960) = \alpha_t$

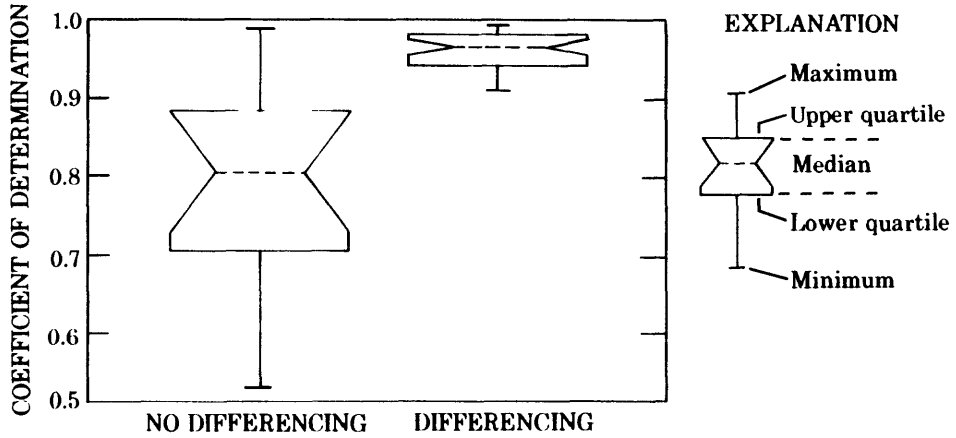

Figure 11.—Box plots of coefficients of determination for autoregressive integrated moving–average equations.

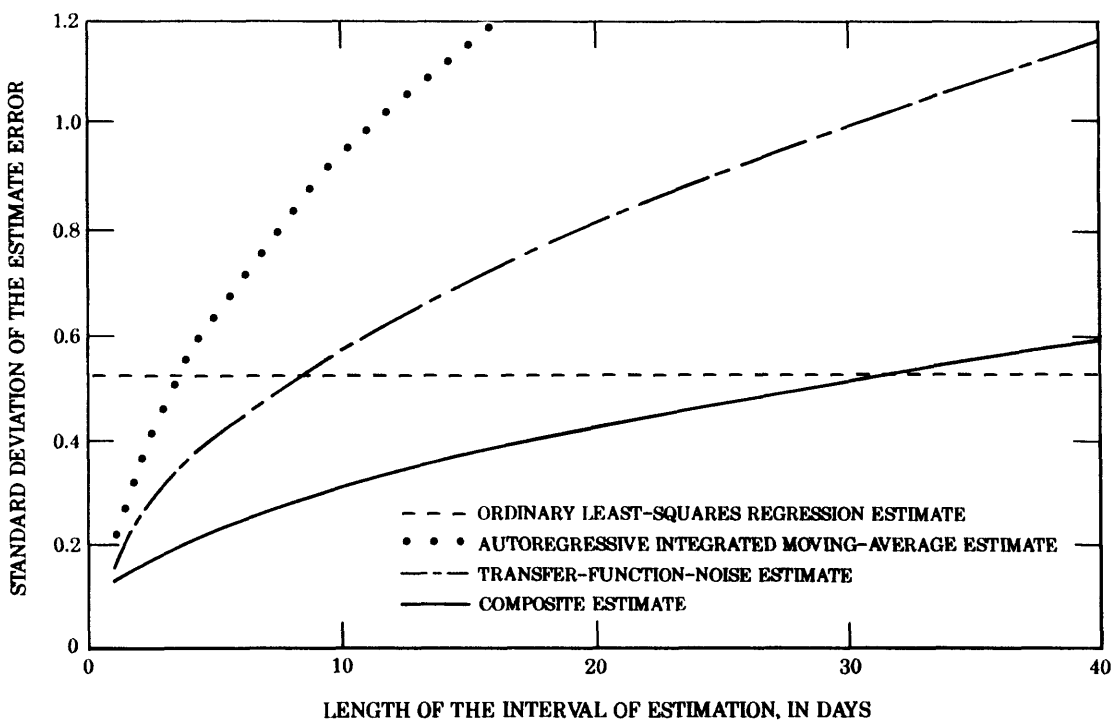

Figure 12.—Relation between errors of ordinary–least–squares regression, autoregressive moving–average, transfer–function–noise, and composite estimates and the length of the interval of estimation for log transformed streamflows from Sturgeon River near Sidnaw.

Transfer–Function-Noise Model

Formulation

The general form of a TFN model can be written as a rational polynomial in B as

$$\nabla_y (y_t - \mu_y) = M_t + \frac{\Theta(B)}{\Phi(B)} a_t + \frac{\Omega(B)}{\delta(B)} \nabla_x (x_t - \mu_x) , \qquad (14)$$

where

y_t is the value of the response series at time t;

μ_y is the mean of the explanatory series;

∇_y is the differencing operator on the explanatory series;

M_t is the deterministic trend;

$\Phi(B)$ is the autoregressive operator of order p on the white-noise series;

$\Theta(B)$ is a moving-average operator of order q on the white-noise series;

a_t is the value of the white-noise series at time t (Values of a_t are computed as $y_t - \tilde{y}_{t-1}(1)$, where $\tilde{y}_{t-1}(1)$ is the lead–1 forecast of y_t at time origin t–1 based on the TFN equation; $a_t \sim NI(0, \sigma_a^2)$.);

δ(B) is the denominator operator of order r on the explanatory series;

$\Omega(B)$ is the numerator operator of the explanatory series ($\Omega(B)$ is the product of the operator $\omega(B)$ of order s and pure delay factor B^b. The pure delay describes the time interval in days between the explanatory input and the streamflow response.);

x_t is the value of the explanatory series at time t;

μ_x is the mean of the explanatory series; and

∇_x is the differencing operator on the explanatory series.

As in the ARIMA model with no trend component, the lead–l forecast from time origin t can be computed recursively (Box and Jenkins, 1976, p. 407) as

$$\tilde{y}_t(l) = \sum_{j=1}^{\infty} P_j [y_{t+l-j}] + \sum_{j=1}^{\infty} Q_j [x_{t+l-j}] , \qquad (15)$$

where the weights P and Q can be obtained by equating coefficients in the expressions

$$\Theta(B) \left[1 - \sum_{j=1}^{\infty} P_j B^j \right] = \phi(B) , \qquad (16)$$

where $\phi(B) = \Phi(B)\nabla_y$,

$$\Theta(B)\delta(B)\sum_{j=1}^{\infty} Q_j B^j = \phi(B)\omega(B)B^b \tag{17}$$

and, because x_t is assumed to be known throughout the period of estimated record at the response station,

$$[x_{t+l-j}] = x_{t+l-j}. \tag{18}$$

The variance of the lead–l forecast error (Box and Jenkins, 1976, p. 405) is

$$\tilde{\sigma}_e^2(l) = \sigma_\alpha^2 \sum_{j=b}^{l-1} v_j^2 + \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2, \tag{19}$$

where σ_α^2 is the variance of the white-noise component of the explanatory series filtered by an ARIMA model and $\Phi_x(B)x_t = \Theta_x(B)\alpha_t$, σ_a^2 is the variance of the white-noise series from the TFN model. The v weights and Ψ weights can be obtained explicitly by equating coefficients in

$$\delta(B)\phi_x(B)v(B) = \omega(B)\Theta_x(B)B^b \tag{20}$$

and

$$\phi(B)\Psi(B) = \Theta(B) \tag{21}$$

(Box and Jenkins, 1976, p. 405). Because x_t is assumed to be known, equation 19 can be written simply as

$$\tilde{\sigma}_e^2(l) = \tilde{\sigma}_a^2 \sum_{j=0}^{l-1} \psi_j^2. \tag{22}$$

Probability limits associated with the TFN model forecast are a function of the forecast lead-l and are computed as

$$\tilde{y}_t(l) \pm N_{\alpha/2}\, \tilde{\sigma}_e(l). \tag{23}$$

Implementation

TFN models were identified and estimated from streamflows from the selected 25 station pairs. Forward shifting of the explanatory series was simulated by specifying the first observation of the explanatory series 4 days ahead of the first observation of the response series. The forward shift was needed in some cases to make full use of the cross–correlation between the explanatory and response series in cases where the two series were contemporaneous or where the explanatory series lagged behind the response series.

The TFN models developed for the log transformed streamflow data (table 4) included, at most, one autoregressive operator (68 percent), one numerator operator (100 percent), one denominator operator (36 percent), and one or more moving-average operators (68 percent). The maximum number of moving-average operators in a single equation was five. In 84 percent of the equations, the response series received first–order first–degree differencing. The explanatory series was differenced in all the equations in which the response series was differenced and in three additional equations. The explanatory series also received first–order first–degree differencing except for one station that received first-order second–degree differencing. The mean was estimated and removed from all series not containing a differencing operator. The pure delay for the forward-shifted explanatory series varies from 0 to 4 days. No deterministic trend components were required. On the basis of criteria specified within AUTOBOX, the assumption of independence and normality of equation residuals (a_t) was not rejected.

Application

The TFN model developed to estimate the log of streamflow at station 04040500 based on the log of streamflow at station 04043050 includes a first–order and first–degree differencing operator on the response series. The noise series includes one autoregressive operator with two parameters. Parameters are associated with backorder powers of 1 and 2 and are estimated as 0.1943 and –0.1213, respectively. The explanatory series, which is assumed to be is stationary about an explicitly included mean of 3.4566, includes one numerator operator. The numerator operator contains five parameters 0.4808, 0.2267, 0.1448, 0.5884, and 0.0530—associated with backorder powers of 0, 1, 2, 3, and 7, respectively. The pure delay of the shifted explanatory series, b, is 4. On the basis of a one–step–ahead forecast, $\tilde{y}_t(1)$, the $\mathcal{R}^2$ value of the equation is 0.975, and the estimated standard deviation of lead–1 forecast errors, $\tilde{\sigma}_e(1)$, is 0.166.

The TFN forecast for October 1, 1988, is obtained by use of equation 16,

$$\sum_{j=1}^{\infty} P_j B^j = 1 - (1 - 0.1943B + 0.1213B^2)(1-B)$$

to compute the P_j coefficients ($P_1 = 1.194$, $P_2 = -0.3156$, and $P_3 = 0.1213$, and by use of equation 17,

$$\sum_{j=0}^{\infty} Q_j B^j = (1 - 0.194B + 0.121B^2)(1-B)(0.481 - 0.227B - 0.145B^2 - 0.0588B^3 - 0.0530B^7)B^4$$

Table 4.—Transfer–function–noise equations for estimating daily streamflow

[x_t and y_t are log transformed daily streamflows at the explanatory and reponse sites; B is the backshift operator, $B^m z_t = z_{t-m}$; ∇ is the backward difference operator, $z_t\nabla = z_t - z_{t-1}$; a_t is a white–noise series; $\tilde{\sigma}_e(1)$ is the estimated standard deviation of the lead–1 standard error; and $\mathcal{R}^2$ is the coefficient of determination for the lead–1 forecast]

Response station	Explanatory station	Equation $\tilde{\sigma}_e(1)$	Equation $\mathcal{R}^2$	Transfer–function–noise equation
04040500	04043050	0.1664	0.975	$y_t\nabla = (1 - 0.1943B + 0.1213B^2)^{-1} a_t + (0.4808 - 0.2267B - 0.1448B^2 - 0.0588B^3 - 0.0530B^7)(x_t - 3.4566)B^4$
04056500	04045500	.0320	.997	$y_t\nabla = (1 - 0.7899B + 0.2409B^2)^{-1} a_t + (0.1191 + 0.2556B + 0.2542B^2)\, x_t\nabla\, B^3$
04059500	04059000	.0879	.992	$y_t\nabla = (1 - 0.3956B + 0.2078B^2)^{-1} (1 - 0.0972B^7)(1 - 0.0891B^{10})\, a_t + (1 - 0.4824B)^{-1} (0.1096 + 0.4032B + 0.1095B^2)\, x_t\nabla\, B^3$
04061500	04033000	.1106	.963	$y_t\nabla = (1 + 0.2860B + 0.1649B^2)^{-1} (1 - 0.1286B^4)(1 - 0.1668B^5)(1 - 0.1711B^6)(1 + 0.1349B^7)(1 - 0.1807B^9)\, a_t + (0.1145 + 0.5398B + 0.4794B^2)\, x_t\nabla\, B^3$
04096400	04096600	.0691	.992	$y_t\nabla = (1 + 0.3266B)(1 - 0.0974B^5)\, a_t + (0.1404 + 0.4607B)\, x_t\nabla\, B^3$
04096515	04096900	.1288	.988	$y_t\nabla = (1 - 0.0974B + 0.1569B^2)^{-1} a_t + (0.2450 + 1.2348B - 0.3893B^2)\, x_t\nabla\, B^3$
04102500	04101500	.0536	.986	$y_t\nabla = (1 + 0.4472B - 0.1893B^2)(1 - 0.1528B^5)(1 - 0.1219B^{10})(1 - 0.1098B^{11})\, a_t + (1 - 0.6118B)^{-1} (0.0860 + 0.1488B + 0.1495B^2 - 0.0632B^9)\, x_t\nabla\, B^2$
04105000	04105500	.0733	.991	$y_t\nabla = (1 - 0.1817B + 0.2776B^2)^{-1} (1 - 0.0946B^5)\, a_t + (1 - 0.4669B)^{-1} (0.2113 + 0.4406B + 0.1630B^2)\, x_t\nabla\, B^3$
04111500	04111379	.2386	.971	$y_t\nabla = (1 - 0.2246B - 0.3795B^2 - 0.08583B^3)\, a_t + (1 - 0.3272B)^{-1} (0.6679 + 0.9353B - 0.8396B^2)\, x_t\nabla\, B^3$
04113000	04109000	.2148	.937	$y_t\nabla = (1 - 0.1912B)\, a_t + (-0.1015 - 0.3165B + 0.1069B^5)\, x_t\nabla\nabla\, B^2$

Table 4.—Transfer–function–noise equations for estimating daily streamflow—Continued

Response station	Explanatory station	Equation $\tilde{\sigma}_e(1)$	Equation χ^2	Transfer–function–noise equation
04114500	04116000	0.1284	0.980	$y_t\nabla = (1 + 0.2263B^2)^{-1}(1 - 0.0946B)(1 - 0.1170B^4)\, a_t + (0.0807 + 0.3852B + 0.1008B^2 - 0.3757B^3 + 0.0840B^7 - 0.0682B^9)(x_t - 7.4013)\, B^2$
04115000	04117500	.1290	.990	$y_t\nabla = (1 + 0.5298B)\, a_t + (1.0365 - 0.2283B + 0.1394B^3)\, x_t\nabla\, B^4$
04122100	04121900	.1723	.956	$y_t\nabla = (1 - 0.0965B - 0.1965B^2 - 0.0826B^3)\, a_t + (1 + 0.3506B)^{-1}(0.3321 + 1.0320B - 0.1134B^3 + 0.1009B^6)\, x_t\nabla\, B^3$
04122200	04121500	.0629	.979	$y_t\nabla = (1 - 0.4861B + 0.2587B^2)^{-1}(1 - 0.0974B^6)\, a_t + (1 - 0.8577B)^{-1}(0.1778 + 0.3682B^2 - 0.4839B^3)\, x_t\nabla\, B^3$
04122500	04121300	.0425	.988	$y_t\nabla = (1 - 0.6145B + 0.1220B^2 + 0.1381B^3)^{-1}\, a_t + (0.1034 + 0.1586B + 0.1941B^2 + 0.1204B^3 + 0.0786B^4)\, x_t\nabla\, B^3$
04127918	04057510	.1694	.950	$y_t\nabla = (1 - 0.1942B + 0.1430B^2)^{-1}\, a_t + 0.8204\, x_t\nabla\, B^4$
04128000	04127800	.0525	.952	$(y_t - 5.4304) = (1 - 0.7911B - 0.0962B^4)^{-1}(1 - .1257B^2)\, a_t + (0.0381 + 0.6757B^2 + 0.4303B^3 + 0.0547B^4 + 0.0658B^5 + 0.0651B^6 + 0.0486B^7 + 0.0431B^8)(x_t - 5.2360)\, B^2$
04135500	04135700	.0364	.977	$y_t\nabla = (1 - 0.4591B + 0.3607B^2 + 0.1350B^4)^{-1}\, a_t + (1 - 0.9272B + 0.3580B^2)^{-1}(0.5327 - 0.2692B)\, x_t\nabla\, B^4$
04146000	04146063	.1345	.988	$y_t\nabla = (1 - 0.1469B + 0.1340B^3)^{-1}(1 - 0.1043B^2)\, a_t + (0.3541 + 0.5615B + 0.2446B^3)\, x_t\nabla\, B^3$
04160570	04160600	.2247	.955	$y_t\nabla = (1 - 0.2520B - 0.3025B^2 - 0.1339B^3)\, a_t + (1 + 0.2430B)^{-1}(0.3632 + 0.6211B)\, x_t\nabla\, B^3$

Table 4.—Transfer–function–noise equations for estimating daily streamflow—Continued

Response station	Explanatory station	Equation $\tilde{\sigma}_e(1)$	Equation $\mathcal{R}^2$	Transfer–function–noise equation
04163400	04161100	0.3232	0.927	$(y_t - 1.8997) = (1 - 0.7541B - 0.0756B^4)^{-1} a_t +$ $(0.9620 + 0.0546B - 0.1025B^2)(x_t - 2.0493) B^4$
04164000	04161540	.2163	.887	$(y_t - 5.8527) = (1 - 0.4948B - 0.1429B^3)^{-1} (1 + 0.1178B^6)(1 + 0.0862B^{11}) a_t$ $+ (1 - .6966B)^{-1} (1.0358 - 0.5463B - 0.2575B^2) (x_t - 3.8101) B^4$
04164100	04161580	.1698	.957	$y_t\nabla = (1 - 0.3036B^2 - 0.1171B^3) a_t + (0.0957 + 0.6743B^4) x_t\nabla$
04164500	04168000	.2449	.974	$y_t\nabla = (1 - 0.4196B + 0.2720B^2)^{-1} a_t + (0.1594 + 0.1328B +$ $-0.1232B^2 - 0.1112B^3 - 0.0469B^4)(x_t - 3.1982) B^4$
04166000	04166100	.1433	.962	$(y_t - 2.9172) = (1 - 0.4390B - 0.1198B^2 - 0.2985B^3)^{-1} (1 - 0.1890B^3) a_t +$ $(0.0406 + 0.1270B + 0.7895B^2) (x_t - 3.9819) B^2$

to compute the Q_j coefficients ($Q_4 = 0.481$, $Q_5 = -0.320$, $Q_6 = -0.042$, $Q_7 = -0.058$, $Q_8 = -0.006$, $Q_9 = -0.007$, $Q_{11} = -0.053$, $Q_{12} = 0.010$, $Q_{13} = -0.006$, and $Q_j = 0$ otherwise). Then, from equation 15, the estimate of the log of streamflow on October 1, 1988, is 4.9547. The exponentiated estimate is 142 ft^3/s.

The estimated variance of the lead–l forecast is computed from the TFN equation developed for station 04040500 by equating the ψ_j in equation 21 as

$$\sum_{j=0}^{\infty} \psi_j B^j = [(1 - 0.1943B + 0.1213B^2)(1-B)]^{-1}$$

such that $\psi_0 = 1$, $\psi_1 = 1.194$, $\psi_2 = 1.111$, $\psi_3 = 1.071$, $\psi_4 = 1.073$, $\psi_j \approx 1.079$ for $j \in \{5,6,...\}$. Then, from equation 23, the 95–percent probability interval about $\tilde{y}_t(1)$ equals [4.628, 5.281]; this corresponds to an exponentiated interval of [102 ft^3/s, 197 ft^3/s]. This interval excludes the OLSR estimate but includes the lead–1 ARIMA forecast. The width of the TFN lead–1 interval is 68 percent smaller than the corresponding width of the OLSR estimate and 23 percent smaller than the interval width based on the lead–1 ARIMA forecast; however, the estimated standard deviation of the forecast error and the associated probability-interval width increase as forecast lead increases. In this case, the standard deviation of the lead–9 forecast error exceeds the RMSE of the OLSR equation (fig. 12). A composite model, discussed in the following sections, combines ARIMA and TFN models to produce an improved estimator.

Composite Model

Two stochastic models, an ARIMA model of the response series and a bivariate–input univariate–output TFN model, were developed to estimate streamflow at time $t+l$. In the TFN model, the standard deviation of forecast errors is reduced by use of an explanatory series in addition to measured values of the response before the beginning of period of estimated record; however, the TFN estimate can be improved by use of streamflow measured at the response site after the period of estimated record.

An ARIMA model of the reverse–ordered series serves as a second model to forecast streamflow during the period of estimated record. Forecasting the reverse-ordered time-series values will be referred to as backcasting. If the response series is reversible, in the sense that the distribution of $(y_1, y_2, ..., y_n)'$ is the same as that of $(y_n, y_{n-1}, ..., y_1)'$, then no additional stochastic modeling is required because the ARIMA model developed during TFN model development of the time–ordered series is applicable to the reverse–ordered streamflow data.

An unbiased composite estimate, based on a weighted sum of the TFN forecast and the ARIMA backcast, ensures a gradual transition of streamflows between periods of measured and estimated flow. In addition, the errors associated with the composite estimate are likely to be smaller than those in either of the individual models if the weights are chosen appropriately. A method for choosing the weights to minimize the variance of the composite estimate is discussed in the following section.

Formulation

The composite estimate of streamflow at $t+l$, $\bar{y}_t(l)$—based on a TFN forecast at time t, $\tilde{y}_t(l)$ and an ARIMA backcast to $t+l$ from $t+k+1$, $\breve{y}_{t+k+1}(l-k-1)$, where k is the length of the interval of estimated record and $1 \leq l \leq k$—can be written as

$$\bar{y}_t(l) = w_{\mathcal{T}}(l)\, \tilde{y}_t(l) + w_{\mathcal{A}}(l-k-1)\, \breve{y}_{t+k+1}(l-k-1)\,. \tag{24}$$

Bias in the composite estimates was avoided by constraining the sum of the weights $w_{\mathcal{T}}(l)$ and $w_{\mathcal{A}}(l-k-1)$ to equal 1. In addition, $w_{\mathcal{T}}(l)$ and $w_{\mathcal{A}}(l-k-1)$ were assumed to be nonnegative. Because $w_{\mathcal{T}}(l) + w_{\mathcal{A}}(l-k-1) = 1$, equation 24 can be written

$$(y_t - \bar{y}_t(l)) = w_{\mathcal{T}}(l)\,(y_t - \tilde{y}_t(l)) + w_{\mathcal{A}}(l-k-1)\,(y_t - \breve{y}_{t+k+1}(l-k-1))\,. \tag{25}$$

On the basis of the variance of equation 25, the variance of the length–l composite error, $\bar{\sigma}^2_e(l)$, can be written as

$$\mathrm{Var}(y_t - \bar{y}_t(l)) = w^2_{\mathcal{T}}(l)\,\mathrm{Var}(y_t - \tilde{y}_t(l)) + w^2_{\mathcal{A}}(l-k-1)\,\mathrm{Var}(y_t - \breve{y}_{t+k+1}(l-k-1)) \tag{26}$$

under the assumption that the covariance between the model errors,

$\mathrm{Cov}\left[((y_t - \breve{y}_{t+k+1}(l-k-1)), (y_t - \tilde{y}_t(l))\right]$, equals zero.

The prediction variance in equation 26 is minimized with respect to $w_{\mathcal{T}}(l)$ and $w_{\mathcal{A}}(l-k-1)$, subject to the above constraints, when

$$w_{\mathcal{T}}(l) = \frac{\breve{\sigma}^2_e(l-k-1)}{\tilde{\sigma}^2_e(l) + \breve{\sigma}^2_e(l-k-1)} \tag{27}$$

and

$$w_{\mathcal{A}}(l-k-1) = \frac{\tilde{\sigma}^2_t(l)}{\tilde{\sigma}^2_e(l) + \breve{\sigma}^2_e(l-k-1)}\,. \tag{28}$$

If variances are replaced with sample estimates and terms are rearranged, equation 24 can be expressed as

$$\bar{y}_t(l) = \frac{\frac{1}{\tilde{\sigma}_e^2(l)}\,\tilde{y}_t(l) + \frac{1}{\breve{\sigma}_e^2(l-k-1)}\,\breve{y}_{t+k+1}(l-k-1)}{\frac{1}{\tilde{\sigma}_e^2(l)} + \frac{1}{\breve{\sigma}_e^2(l-k-1)}}, \qquad (29)$$

where it is apparent that the composite estimate, $\bar{y}_t(l)$, is obtained by weighting the TFN forecast and ARIMA backcast inversely proportional to their respective error variances. Similarly, equation 26 can be expressed as

$$\bar{\sigma}_e^2(l) = \left(\frac{\breve{\sigma}_e^2(l-k-1)}{\breve{\sigma}_e^2(l-k-1)+\tilde{\sigma}_e^2(l)}\right)^2 \tilde{\sigma}_e^2(l) + \left(\frac{\tilde{\sigma}_e^2(l)}{\breve{\sigma}_e^2(l-k-1)+\tilde{\sigma}_e^2(l)}\right)^2 \breve{\sigma}_e^2(l-k-1)\,. \qquad (30)$$

Implementation

ARIMA equations, developed from the reverse–ordered streamflows, were virtually identical to the equations developed from the time–ordered streamflows. Length–1 composite estimates were computed from the 4 years of calibration data for the 25 selected station pairs. Empirical results indicate that the composite estimates are unbiased, $\Sigma(y_{ti} - \bar{y}_{ti}(1))/\,(4 \times 365) \simeq 0$, for $i = 1$ to 25 of the selected stations. In addition, the empirical standard deviations of length–1 estimation errors, $\bar{\sigma}_{e_i}(1) = \sqrt{\Sigma(y_{ti}-\bar{y}_{ti}(1)^2)/(4\times365)}$, average 24.4 percent lower than standard deviations of the lead–1 TFN forecast errors, $\tilde{\sigma}_e^2(1)$.

Use of $\bar{\sigma}_e^2(l)$ as an estimator of the true error variance is based on the assumption of independence between TFN forecast errors and ARIMA backcast errors. This assumption was investigated by examining the linear correlation between model errors. Results based on the 25 selected station pairs indicate that, although the distribution of correlation coefficients was roughly symmetrical about zero (fig. 13), most of the individual sample correlation coefficients were significantly different from zero at the 5–percent level.

The effect of this correlation on the estimated standard deviation of the composite error is shown in figure 14. If the model errors were negatively correlated, the empirical standard deviation, $\hat{\bar{\sigma}}_e(1)$, was less than the estimate based on independence, $\bar{\sigma}_e(1)$. In contrast, if the correlation was positive, $\bar{\sigma}_e(1)$ was less than $\hat{\bar{\sigma}}_e(1)$. Thus, although

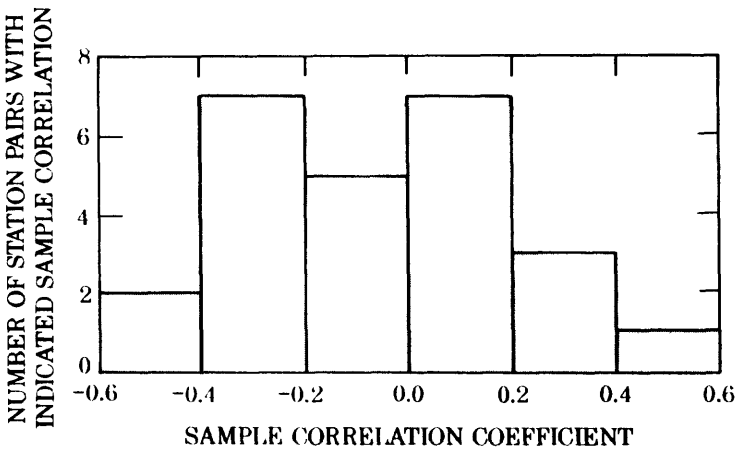

Figure 13.—Histogram of sample correlation coefficients between lead–1 forecast and lead–1 backcast errors computed by use of transfer–function–noise and autoregressive integrated moving–average equations.

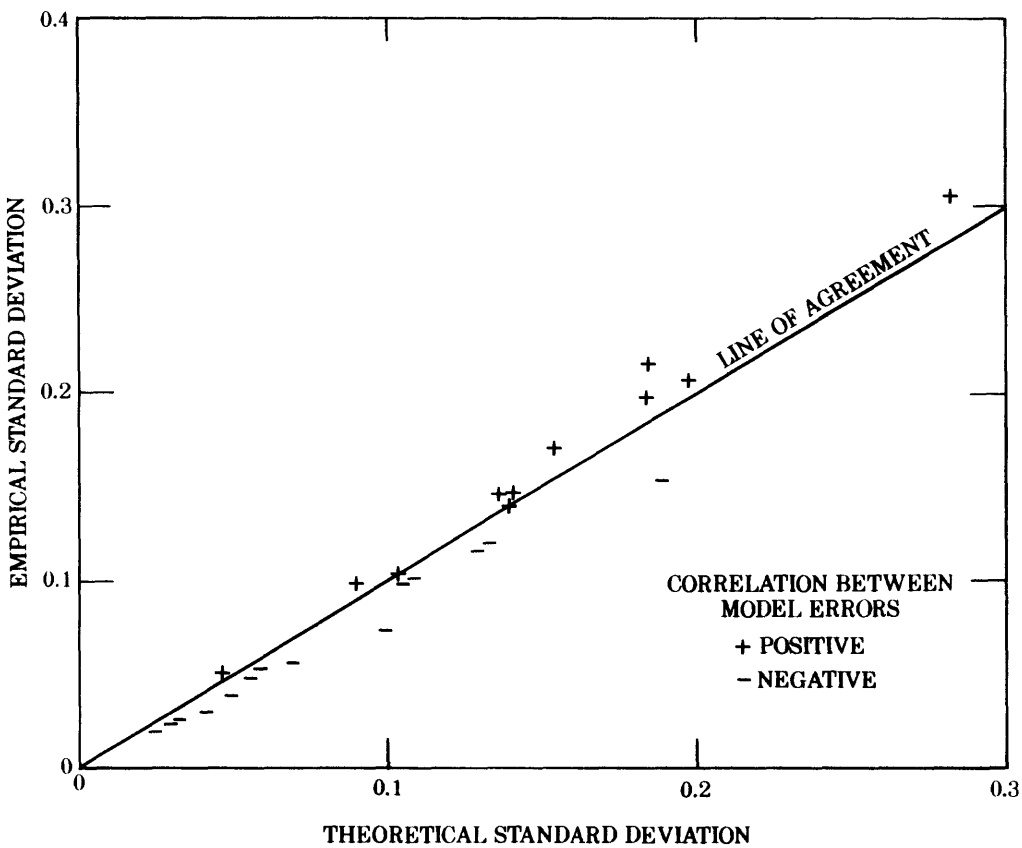

Figure 14.—Relation between estimates of the standard deviation of length–1 composite errors based on empirical analysis and estimates based on the assumption of independence between transfer–function–noise-forecast and autoregressive integrated moving–average-backcast errors.

$\bar{\sigma}_e(1)$ is apparently an unbiased estimator of the true standard error[3] $\bar{\sigma}_e(1)$ is likely to differ somewhat from the true standard error. Detailed analysis of the model-error correlation structure is needed if precise estimates of standard deviation of composite errors are required.

Application

The potential improvement associated with the composite estimator is illustrated by means of a hypothetical 21–day interval of estimated record beginning on October 1, 1988, at station 04040500 (fig. 15). The explanatory series, based on streamflow records from station 04043050, includes two small peaks, one on October 6 and a second on October 19. The streamflow response associated with these peaks seems inconsistent. During the October 6 peak, the relative increase in streamflow at the response site is greater than the relative increase at the explanatory site; during the October 19 peak, the relative increase in streamflow at the explanatory site is large but the relative increase at the response site is small. This inconsistency may be attributed to the differences in the distribution of rainfall over the two basins.

The exponentiated TFN forecasts, which alternately underestimate and overestimate the measured streamflow response for these two peaks, indicate that the response was alternately greater and smaller than expected on the basis of average response during the calibration period. Although the TFN forecasts result in a smooth transition between measured flow on September 30 and estimated flow October 1, the transition from estimated to measured flow between October 21 and October 22 is abrupt. In contrast, the exponentiated ARIMA backcasts are consistent with flow on October 22 but not with measured flow on September 30. The exponentiated composite estimate matches the TFN forecasts near the beginning of the interval and approaches the ARIMA backcasts near the end of the estimated interval. In contrast, the exponentiated OLSR estimates differ sharply from measured streamflow anytime within the first 16 days of the interval. Within this 21–day interval, the exponentiated composite estimates appear unbiased, whereas the exponentiated OLSR estimates tend to underestimate measured streamflow.

[3]Results of a Wilcoxon rank–sum test (Conover, 1980, p. 280) failed to reject H_o: $\hat{\bar{\sigma}}_e(1) - \bar{\sigma}_e(1) = 0$ at the 5–percent level of significance)

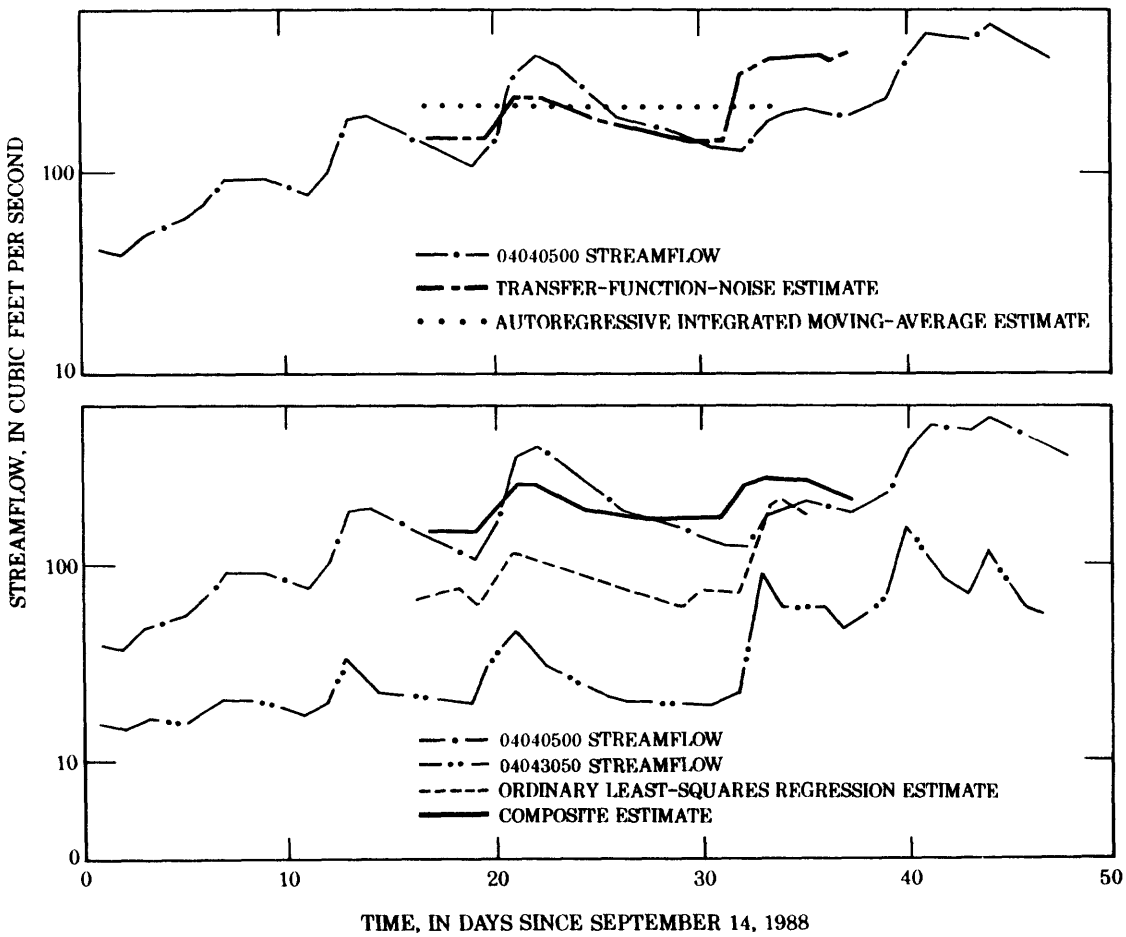

Figure 15.—Measured and estimated streamflow of Sturgeon River near Sidnaw from September 15 through October 31, 1988.

COMPARISON OF MODEL BUILDING ALTERNATIVES

The following sections discuss the variation in accuracy of streamflow estimation (1) between alternative data-transformations, (2) between models developed with and without removal of seasonal components, and (3) among alternative statistical models.

Data Transformations

Two sets of OLSR equations were developed from log- and avas-transformed streamflow data. The two sets contained similar numbers of parameters; the equations based on log-transformed streamflow data contained an average of 5.3 parameters, whereas equations based on the avas-transformed streamflow data contained an average of 4.9 parameters. The average coefficient of determination, $\mathcal{R}^2$, of the equations based on the log transformation (0.84) differed slightly from the average $\mathcal{R}^2$ of the equations based on the avas transformation (0.85); however, the residuals of the equations based on the avas transformation were more likely to have a constant variance than were residuals from equations based on the log transformation (fig. 16).

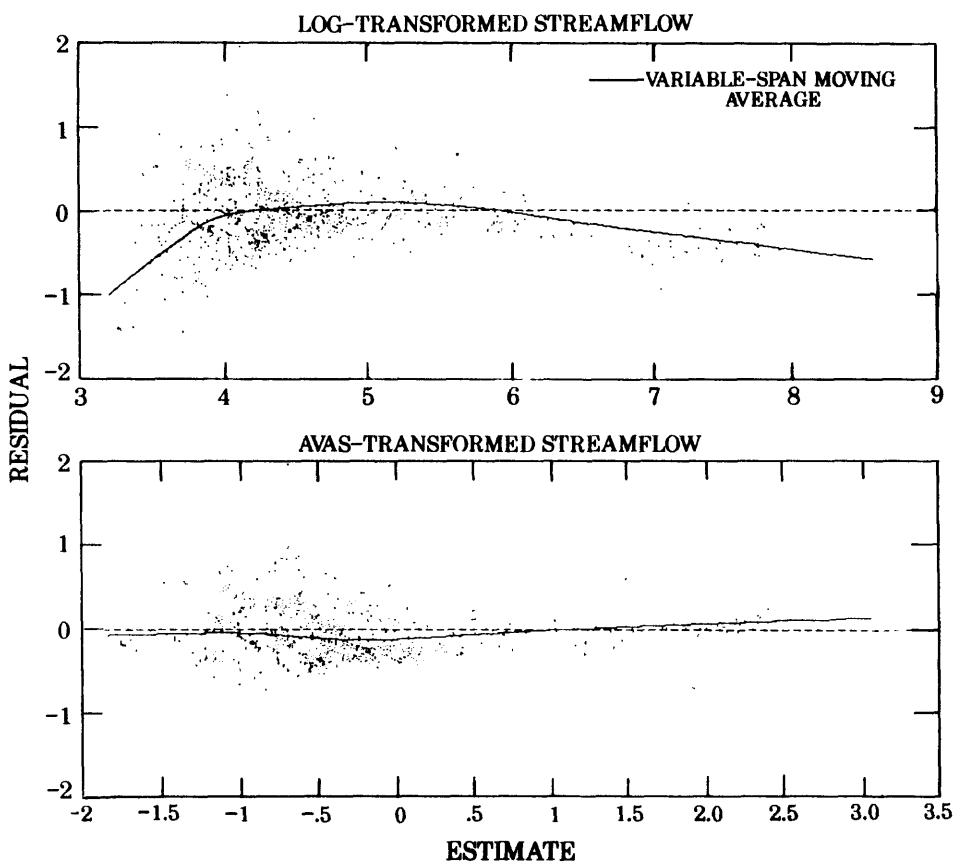

Figure 16.—Variation of residuals with estimates based on ordinary–least–squares regression equations developed from log- and avas-transformed streamflow values from Sturgeon River near Sidnaw.

The bias and the precision of the estimates also were compared to assess the adequacy of the log transformation commonly applied in the analysis of streamflow data. A common metric for comparison for the log and the avas estimates was provided by inversely transforming both estimates to cubic feet per second and subtracting them from measured streamflows included within the verification period as

$$\epsilon_{lt} = y_t - \exp(\hat{y}_{lt}) \tag{31}$$

$$\epsilon_{at} = y_t - f^{-1}(\hat{y}_{at}) , \tag{32}$$

where ϵ_{lt} are the residuals based on equations developed from log-transformed data and ϵ_{at} are the residuals based on avas-transformed data. Inverse transformations of the log and avas functions were obtained, respectively, by exponentiation and by linear interpolation between tabled values based on the calibration data.

Analysis of the residual statistics showed no significant difference at the 5–percent level between the mean or standard deviation of residuals derived from the log-transformation-based OLSR equations and those derived from the avas-transformation-based OLSR equations. Therefore, although the avas transformation improved the error distribution somewhat, the accuracy of estimation differed little between equations sets. For simplicity of analysis and for consistency with other investigations, the log transformation was not rejected in favor of the avas transformation.

Trend and Seasonal Components

The potential for improving the accuracy of the estimating equations by including explicit trend and seasonal components was investigated. No significant linear-trend components were identified during ARIMA or TFN model development. No other trend components were evident from inspection of time–series plots of residuals.

Seasonal components were estimated by use of the VSMA vectors and the log-transformed streamflow data within the calibration period. The VSMA vectors were subtracted from the log-transformed calibration data to form deseasonalized series. OLSR models and TFN models were developed from the deseasonalized data. Seasonalized estimates for the verification period were formed by adding the deseasonalized estimates from the estimating equations and the seasonal component estimated by use of the VSMA vector.

OLSR equations based on deseasonalized flow data (OLSR′ equations) contained an average of 5.0 parameters, whereas OLSR equations based on data without deseasonalization contained an average of 5.3 parameters; however, the $\bar{R}^2$ of the OLSR′

equations averaged only 0.74, about 0.10 units below the average for OLSR equations. The r^2 of lead-1 forecasts with flow based on TFN′ equations (deseasonalized flow data) was 0.94, about 0.03 units below the r^2 for the TFN equations. The bias and precision of estimates that were reseasonalized and transformed back to cubic feet per second were not significantly different (at the 5–percent level) than bias and precision of estimates computed from deseasonalized data. The results indicate that inclusion of seasonal components is unlikely to substantially improve either the OLSR estimates or lag-1 forecasts based on stochastic models; however, deseasonalization results in time series that satisfy the stationarity assumption associated with the statistical models and may improve forecasts at longer lead times than those investigated in this analysis.

Statistical Models

OLSR models and stochastic models were used to describe daily streamflows for 25 gaging-station pairs in Michigan. Important differences were found between the consistency of model assumptions with data characteristics, the accuracy of estimation, and the ease of model development and use.

Both models are based on the assumption that the errors are uncorrelated. This assumption was satisfied in the stochastic models. In contrast, the OLSR residuals were positively autocorrelated for all stations. The autocorrelation indicates that the variance of the OLSR errors should be represented as an n–dimensional matrix rather than as a scalar. This violation of a basic assumption of OLSR models creates uncertainty as to the applicability of the OLSR equations to estimation of daily streamflows. The limitations of the OLSR equations were most apparent in the estimation of short intervals of daily streamflow.

The standard deviation of errors of the stochastic models increased monotonically with increased length of the interval of estimation (fig. 12). In contrast, the standard deviation of the OLSR errors varies with the distance between $x_{t+l} - \bar{x}$; however, because of the large number of measurements used in the estimation, the standard deviation of the OLSR errors was approximated as a constant equal to its root mean square error ($\hat{\sigma}_e$). On the basis of this approximation, 60 percent of the $\breve{\sigma}_e(2)$ in the ARIMA equations were less than or equal to $\hat{\sigma}_e$; however, only 20 percent of $\breve{\sigma}_e(4)$ were less than or equal to $\hat{\sigma}_e$. In comparison, 60 percent of the $\tilde{\sigma}_e(9)$ in the TFN equations were less than or equal to $\hat{\sigma}_e$ and 12 percent of the $\tilde{\sigma}_e(14)$ were less than or equal to $\hat{\sigma}_e$.

The mean standard deviation of the length–l estimation errors is a basis for comparing the accuracy of the OLSR and the composite estimators. The estimated mean standard deviation of the length–l composite error is computed as

$$\bar{\bar{\sigma}}_e(l) = l^{-1} \sum_{l} \bar{\sigma}_e(l) . \tag{33}$$

On this basis, the stochastic equations provide the more accurate estimates of daily streamflow for short durations of estimated record. Specifically, $\bar{\bar{\sigma}}_e(l) \leq \hat{\sigma}_e$ at all stations for $l \leq 10$ days, at 92 percent of the stations for $l \leq 21$ days, and at 52 percent of the stations for $l \leq 32$ days (fig. 12).

The mean error ratio is a basis for assessing the potential reduction in streamflow estimation error by use of the composite estimate rather than the OLSR estimate. The mean error ratio was computed as

$$\bar{\hat{r}}_\sigma(l) = n_s^{-1} \sum_{i} \frac{\bar{\bar{\sigma}}_{e_i}(l)}{\hat{\sigma}_{e_i}} \tag{34}$$

for the $n_s = 25$ selected response stations. The results indicate that $\bar{\hat{r}}_\sigma(l) < 1.0$ for $l \leq$ 32 days (fig. 17). By weighting $\bar{\hat{r}}_e(l)$ by the frequency of length–l intervals of estimated record, $f(l)$, for $l = 1,2,...30$, an estimate of the effective mean error ratio, $\bar{\bar{\hat{r}}}_\sigma$, is obtained. The value of $\bar{\bar{\hat{r}}}_\sigma$ indicates the potential for reducing estimation errors by use of the composite estimate instead of the OLSR estimate for periods of estimated record not exceeding 1 month. The estimate of $\bar{\bar{\hat{r}}}_\sigma = 0.52$ was computed as

$$\bar{\bar{\hat{r}}}_\sigma = \left[\sum_{l} f(l) \right]^{-1} \sum_{l} f(l)\, \bar{\hat{r}}_\sigma(l) . \tag{35}$$

Because the estimated ratio is much less than 1, use of the composite estimate is likely to substantially reduce the errors of streamflow estimation in Michigan compared to the OLSR estimate.

Finally, both OLSR equations and stochastic equations were developed by means of commercially available computer software. Development of OLSR equations required the specification of arbitrary criteria (an increase of 0.001 in R^2 value for the addition of a model parameter) to avoid an excessive number of parameters that may have occurred if model identification had been based on the computed significance level of parameters.

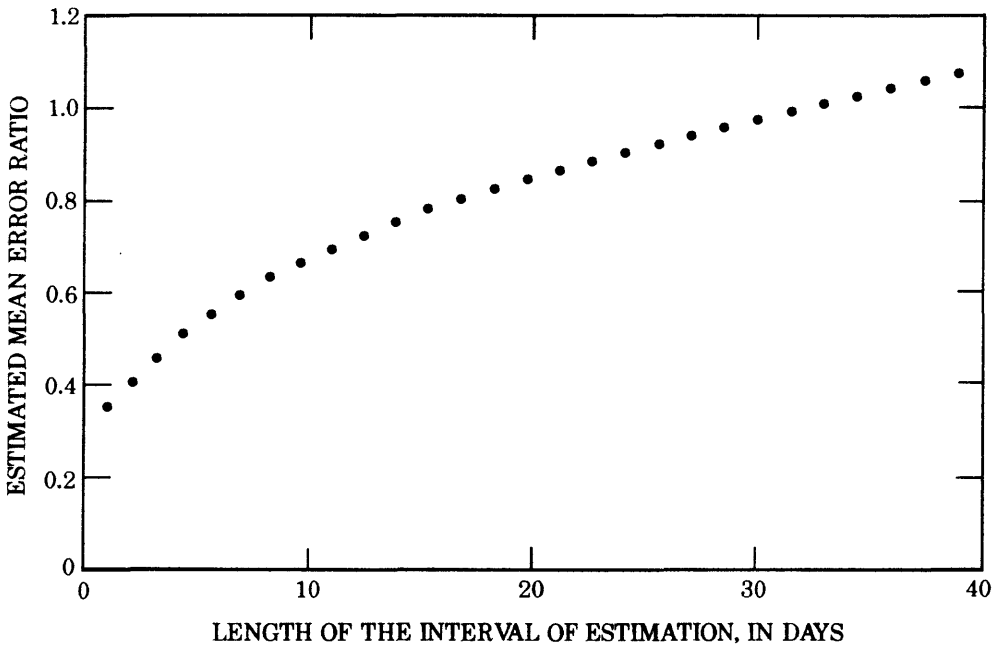

Figure 17.—Relation between the mean error ratio and the length of the interval of estimation.

The computed significance level of OLSR parameters were overestimated because the streamflow residuals were positively correlated. Stochastic model development by means of AUTOBOX was completely automatic once the time series, modeling options, and minimum significance of parameters was specified.

The regression form is convenient for estimating streamflow values. Stochastic equations can be converted to a regression form through algebraic manipulation. An option in AUTOBOX provides this converted output or the forecast values directly. Although the OLSR equations are developed in the regression form, the estimates may not be useable because they tend to be inconsistent with streamflows before and after the period of estimated record.

SUMMARY

Daily streamflows from October 1, 1984 through September 30, 1989 were analyzed for 25 pairs of gaging stations in Michigan. Stations were paired by randomly choosing a station operated in 1989 at which 10 or more years of continuous record were available and at which flow is virtually unregulated and then selecting a nearby station where flow characteristics are similar. Streamflow data from each of the 25 randomly selected stations were used as the dependent or response variable; streamflow data at the nearby stations were used to generate a set of independent or explanatory variables.

About 15 percent of the daily values during this period were estimated because of either missing record (often resulting from equipment failure) or variable backwater (often related to channel ice formation). The durations of periods of estimated record averaged 14 days, although about 65 percent were less than or equal to 7 days. Estimates are computed to complete daily streamflow records obtained by the U.S. Geological Survey at most of the 140 stations operated in Michigan.

Daily streamflow estimates can be obtained by use of statistical models. Because of the asymmetrical distribution of streamflow data, nonlinear data transformations are commonly applied to facilitate model development. In this analysis, two nonlinear data transformations were investigated, the natural logarithm (log) and the additive variance stabilizing (avas) transformations. In addition, the need for explicit trend and seasonal components in the statistical models also was studied.

Ordinary least–squares regression (OLSR) is a commonly used statistical model for estimating a response variable based on one or more explanatory variables. The OLSR model is based on the assumption that the model errors are independent and normally distributed with a mean of zero and a constant variance. OLSR equations were developed for estimating the streamflow response at time t based on streamflow at the explanatory station at time t+4, t+3,...,t–7. Subsets of the explanatory variables were selected on the basis of a statistic (C_p) that reflects parsimony in a selected model. Final OLSR model-selection criteria included a minimum computed significance of 5 percent for all explanatory variables and a minimum increase in the coefficient of determination (R^2) of 0.001 for each added parameter.

Stochastic models were developed to describe dynamic univariate processes (ARIMA models) and bivariate–input univariate-output processes (TFN models). The stochastic models were developed using AUTOBOX, a computer–based expert–system program (Automatic Forecasting Systems, 1988). A composite estimator was developed from the forecast computed by the TFN equation and a backcast (a forecast of the

reverse-ordered response series following the end of the period of estimated record) computed by the ARIMA equation.

OLSR analysis of log- and avas-transformed streamflow data indicate that, although the avas transformation tends to improve the distribution of the transformed residuals, the streamflow estimates are not significantly improved. Therefore, the widely used log transformation was considered adequate to develop the streamflow models. No trends were detected in the streamflow series, and the removal of seasonal components did not significantly improve the accuracy of OLSR- or stochastic-model estimates examined.

Residuals from the OLSR models of daily streamflow were characterized by consistently positive correlations which violates the assumption of independence associated with the OLSR model. OLSR estimates of streamflow were commonly inconsistent with measured values of streamflow at the response station immediately before and after the period of estimated record. Neither the ARIMA nor the TFN model residuals were autocorrelated; furthermore, their distributions were virtually normal. The standard deviations of the forecast errors were generally lower than the OLSR estimates at small forecast leads; however, the forecast errors increased monotonically with the forecast lead and generally exceeded the OLSR estimates by lead–12.

A composite estimate, formed by weighting the TFN forecast with the ARIMA backcast in inverse proportion to their respective error variances, decreased the average standard deviation of the estimate errors and ensured a smooth transition between periods of measured and estimated streamflow. The average standard deviation of the errors of the combined estimate was generally less than the OLSR estimate for intervals less than or equal to 32 days.

The mean error ratio indicates the potential reduction in streamflow estimation error if the composite estimate rather than the OLSR estimate is used. The mean error ratio for the 25 selected response stations was less than 1 for lengths of estimated record less than or equal to 32 days. Weighting the mean error ratio by the frequency of length–l intervals of estimated record provides an estimate of the effective mean error ratio. Because the estimated value of the effective mean error ratio of 0.52 is much less than 1, use of the composite estimate rather than the OLSR estimate could substantially reduce streamflow-estimation errors in Michigan.

SELECTED REFERENCES

Automatic Forecasting Systems, 1988, Autobox Plus User's Guide Version 2.0: Hatboro, Pennsylvania, 143 p.

Beck, J. V., and Arnold, K. J., 1977, Parameter estimation in engineering and science: New York, John Wiley, 501 p.

Box, G. E. P., and Draper, N. R., 1987, Empirical model–building and response surfaces: New York, John Wiley, 669 p.

Box, G. E. P., and Jenkins, G. M., 1976, Time series analysis–forecasting and control: Oakland, California, Holden–Day, 575 p.

Condes de la Torre, Alberto, 1989, Operation of hydrologic data–collection stations by the U.S. Geological Survey in 1989: U.S. Geological Survey Open–File Report 90–171, 52 p.

Conover, W. J., 1980, Practical nonparametric statistics (2d ed.): New York, John Wiley, 493 p.

Daniel, Cuthbert, and Wood, S.F., 1980, Fitting equations to data–computer analysis of multifactor data (2d ed.): New York, John Wiley, 458p

Fontaine, R. A., Moss, M. E., Smath, J. A., and Thomas, Jr., W. O., 1984, Cost effec–tiveness of the stream–gaging program in Maine—a prototype for nationwide implementation: U.S. Geological Survey Water–Supply Paper 2244, 39 p.

Holtschlag, D. J., 1985, Cost effectiveness of stream–gaging program in Michigan: U.S. Geological Survey Water–Resources Investigations Report 85–4293, 72 p.

Johnson, R. A., and Wichern, D. W., 1982, Applied multivariate statistical analysis: Englewood Cliffs, NJ, Prentice-Hall, 594 p.

Riggs, H. C., 1968, Some statistical tools in hydrology, Hydrologic analysis and interpretation: U.S. Geological Survey Techniques of Water–Resources Investigations, Book 4, Chap. A1, 39 p.

Scott, A. G., and Moss, M. E., 1986, Analysis of the cost effectiveness of the U.S. Geological Survey stream–gaging program, *in* Moss, M.E., ed, Integrated design of hydrological networks, International Association of Hydrological Sciences publication 158, 50 p.

Statistical Sciences, 1990, S–PLUS for DOS User's Manual; Seattle, Washington, 206 p.

U.S. Geological Survey, 1986-90, Water resources data for Michigan, water year 1985-89 volume 1; U.S. Geological Survey Water-Data Reports MI-85-1 to MI-89-1 (published annually).

DEFINITIONS OF TERMS

Autocorrelation. The correlation between elements of a series of observations or measurements ordered in time. The collection of autocorrelation coefficients for two or more lags of a series is called the autocorrelation function.

Coefficient of determination. The square of the product–moment correlation coefficient.

Cubic feet per second. A unit expressing rate of discharge. One cubic foot per second is equal to the discharge of a stream 1 foot wide and 1 foot deep flowing at an average velocity of 1 foot per second.

Degrees of freedom. A number based on the difference between the sample size and the number of parameters in the model.

Discharge. The volume of water (or more broadly, volume of fluid plus suspended sediment) that passes a given point within a given period of time.

Expected value. The mean of a random variable.

Fourier series. A set of sine and cosine functions capable of approximating a variety of mathematical functions.

$\mathcal{P}$–value. The probability of obtaining a value of a statistic as unusual as that observed.

Polynomial trend. A trend of the general form

$$y = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \cdots + \alpha_n t^n.$$

Low order implies that the highest exponent of t will be 3 or less.

Prewhitening. A filtering procedure that reduces autocorrelation.

Root–mean–square error. The square root of the second moment of a set of differences between measured values and model estimates. Equal to the square root of the mean-square error.

Stationary. As used in this report, a stochastic process is stationary in the covariance if the covariance function $\mathrm{Cov}(k) = \mathcal{E}[x(t)\, x(t+k)]$ exists and is independent of t for all integer values of k.

Stochastic model. A model in which estimates are based on their probability of occurrence rather than on physical laws. In this report, stochastic models refer to statistical models that are appropriate for describing dynamic systems.

Streamflow. The discharge that occurs in a surface stream channel.

Variance. The variance is the average value of the squared deviation of a variate from its mean.

Water year. In U.S. Geological Survey reports, the water year is the 12–month period from October 1 through September 30. The water year is designated by the calendar year in which the water year ends; thus, the water year ending September 30, 1988, is called "water year 1988."

★U.S. GOVERNMENT PRINTING OFFICE: 1992-644-891/ 52058