HYDROLOGICAL PROCESSES
Hydrol. Process. **15**, 1133–1155 (2001)
DOI: 10.1002/hyp.207

Optimal estimation of suspended-sediment concentrations in streams

David J. Holtschlag*
US Geological Survey, Water Resources Division, 6520 Mercantile Way, Suite 5, Lansing, MI 48911, USA

Abstract:

Optimal estimators are developed for computation of suspended-sediment concentrations in streams. The estimators are a function of parameters, computed by use of generalized least squares, which simultaneously account for effects of streamflow, seasonal variations in average sediment concentrations, a dynamic error component, and the uncertainty in concentration measurements. The parameters are used in a Kalman filter for on-line estimation and an associated smoother for off-line estimation of suspended-sediment concentrations. The accuracies of the optimal estimators are compared with alternative time-averaging interpolators and flow-weighting regression estimators by use of long-term daily-mean suspended-sediment concentration and streamflow data from 10 sites within the United States. For sampling intervals from 3 to 48 days, the standard errors of on-line and off-line optimal estimators ranged from 52·7 to 107%, and from 39·5 to 93·0%, respectively. The corresponding standard errors of linear and cubic-spline interpolators ranged from 48·8 to 158%, and from 50·6 to 176%, respectively. The standard errors of simple and multiple regression estimators, which did not vary with the sampling interval, were 124 and 105%, respectively. Thus, the optimal off-line estimator (Kalman smoother) had the lowest error characteristics of those evaluated. Because suspended-sediment concentrations are typically measured at less than 3-day intervals, use of optimal estimators will likely result in significant improvements in the accuracy of continuous suspended-sediment concentration records. Additional research on the integration of direct suspended-sediment concentration measurements and optimal estimators applied at hourly or shorter intervals is needed.

KEY WORDS suspended sediments; optimal estimation; Kalman filtering; computation; flux; loads

INTRODUCTION

Since 1995, the National Stream Quality Accounting Network (NASQAN) of the US Geological Survey (USGS) has monitored water quality at 40 stations on four of the nation's largest basins, including the Columbia, the Colorado, the Mississippi, and the Rio Grande. Monitored constituents include sediment concentrations, major ions, trace elements, nutrients, pesticides, carbon, and support variables including streamflow, dissolved oxygen, temperature, pH, and conductivity. To ensure that these data are used as effectively as possible, the NASQAN programme, through the USGS Office of Water Quality, has supported the development of optimal estimators of suspended-sediment concentrations. The general form of these estimators may also be applicable to other constituents.

Sediment in streams results from erosion in the basin and transport by flowing water (Guy, 1970). Either process can limit the occurrence of sediment in streams. Total sediment discharge includes a bed-load component and a suspended-sediment component. Bed load refers to sediment particles rolling, sliding, or tumbling along the streambed. Suspended sediment represents that component of sediment that stays in suspension for an appreciable length of time and represents a dynamic equilibrium between the upward forces of turbulence holding particles in suspension against the downward force of gravity. Turbulent forces are

* Correspondence to: D. J. Holtschlag, US Geological Survey, Water Resources Division, 6520 Mercantile Way, Suite 5, Lansing, MI 48911, USA. E-mail: dholtschlag@usgs.gov

Received 15 November 1999
Accepted 15 May 2000

directly related to streamflow rate, and the effectiveness of gravitational forces is related to particle sizes and densities. In most natural rivers, sediments are transported mainly as suspended sediment (Yang, 1996).

Estimation of average sediment concentrations and flux rates requires the integration of continuous data on streamflow with discrete measurements of sediment concentration. This integration is commonly carried out by interpolating discrete measurements by use of time-averaging or flow-weighting methods. Phillips *et al*., (1999) compared 20 existing and two proposed methods for computing loads and found that a time-averaging method produced the most precise estimates for two stations analysed. The precision of this method decreases significantly, however, as the sampling interval increases (Phillips *et al*., 1999). Furthermore, Bukaveckas *et al*. (1998) conclude that time-averaging methods may produce biased estimates of flux during periods of variable discharge.

A sediment-rating curve approach (Helsel and Hirsch, 1992) is a common method of flow weighting. This type of rating curve generally describes the relation between the logs [where log refers to the common (base 10) logarithm function] of suspended-sediment concentration and the logs of streamflow. This rating-curve approach, however, has been found to underestimate river loads (Ferguson, 1986). In addition, Bukaveckas *et al*. (1998) indicate that flow-weighting methods may produce biased estimates if the concentration–streamflow relation is affected by antecedent conditions or has seasonal variability. Seasonal rating curves are used to reduce the scatter and to eliminate this bias at some sites (Yang, 1996). Although beyond the scope of this study, Richards and Holloway (1987) evaluate the accuracy and precision of tributary load estimates as affected by both sampling frequency and pattern.

The concentrations of suspended sediments are measured at stream-gauging stations throughout the United States because of the environmental and economic significance of the effect of sediment on receiving waters. The USGS operates many of these stations, which are funded through several cooperative and federal programmes. Daily-mean concentrations of suspended sediment are determined for sites where sufficient direct measurements of suspended-sediment concentration and continuous streamflow data are available (Randy Parker, US Geological Survey, written communication, 1999). Because of the uniform data-collection methods (Guy and Norman, 1973) and computational procedures (Porterfield, 1972; Glysson, 1987) used by the USGS, these data provide the best available information on suspended-sediment flux in the nation's rivers. In addition, the USGS maintains a nationwide database of suspended-sediment concentration, streamflow, and ancillary data (Randy Parker, written communication, 1999). The database contains daily values for 1593 stations in the United States that have an average period of record of 5·3 years. This database was used for the analysis reported in this paper and is accessible from the Internet at http://webserver.cr.usgs.gov/sediment/

Purpose and scope

This paper develops optimal on-line and off-line estimators of suspended-sediment concentrations for streams on the basis of daily values of computed suspended-sediment concentration and streamflow information. Data from 10 sites are used to compare the accuracy of the optimal estimators with interpolation and regression estimators (Koltun *et al*., 1994) that are commonly used to compute suspended-sediment concentration records. The estimators were restricted to those that could be readily implemented with data that are generally available at gauging stations. The choice of daily-value rather than unit-value (hourly or less) computational intervals, however, was based on the greater accessibility of daily-value data. The optimal estimators developed in this paper are intended for eventual application at unit-value intervals.

Site selection

Ten USGS gauging stations (Figure 1, Table I) were selected to develop the estimators and assess their accuracy. The sites represent a broad range of basin sizes, suspended-sediment concentration characteristics, and streamflow characteristics. Basin drainage areas range from 1610 to 116 000 km^2. Median suspended-sediment concentrations (Figure 2) range from 8 to 3040 mg/l. Median streamflow (Figure 3) ranged from

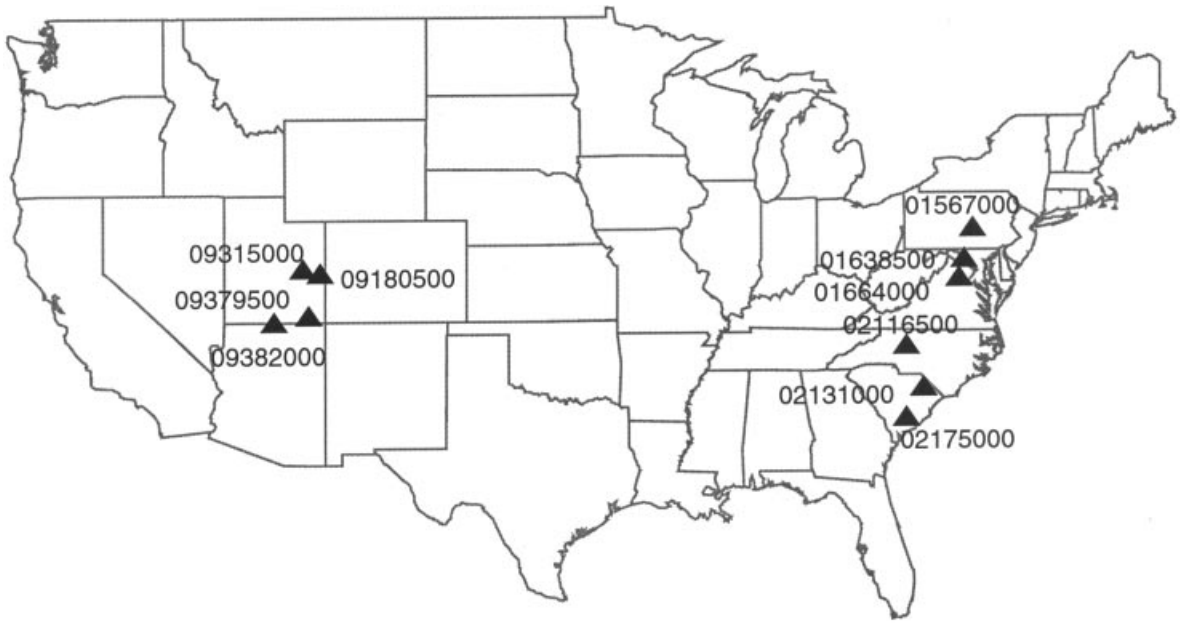

Figure 1. Locations of selected US Geological Survey sediment gauging stations in the United States

0·34 to 241 m^3/s. Available suspended-sediment particle-size distribution data indicates that the percentage of suspended sediment finer than 0·125 mm ranged from 70 to 95% (Figure 4) among sites.

ESTIMATORS OF SUSPENDED-SEDIMENT CONCENTRATIONS

Previous estimators for computing records of suspended-sediment concentrations (Koltun *et al.*, 1994) include both time-averaging and flow-weighting procedures. In this paper, time-averaging procedures refer to linear or cubic-spline interpolators between direct suspended-sediment concentration measurements, typically obtained at unequal time intervals. These estimators have no rigorous mechanism for quantifying the uncertainty of the values produced, but provide estimates that are consistent with data at the times of direct measurements. Flow-weighting procedures include linear regressions, which condition estimates of suspended-sediment concentrations on streamflow (and sometimes other explanatory variables). The regression estimators account for a primary source of variability between measurements of sediment concentration and provide a measure of uncertainty. Regression estimators, however, do not converge appropriately to measured values at times of direct measurement. Optimal estimators combine the benefits of both time-averaging and flow-weighting procedures in a formal mathematical model that also describes the statistical uncertainty in the estimates.

Both suspended-sediment concentrations and streamflow tend to be highly skewed to the right. Logarithmic (log) transformation, however, creates a more symmetrical distribution (Figure 5). The linearity between percent frequency, on a normal probability scale, and both suspended-sediment concentrations (Figure 2) and streamflow (Figure 3), on the log scale, supports the common assumption that suspended-sediment concentrations and streamflow are approximately log-normally distributed. The log transformation improves linearity between the sediment concentrations and streamflow (Figure 6) and reduces heteroscedasticity (nonconstant variance that may be a function of magnitude) in the model errors. Thus, interpolation and regression methods generally apply a log transformation prior to method application. In addition, some estimators provide an adjustment for the seasonal variations (Figure 7) in average sediment concentrations.

Table I. Identification and location of selected sediment gauging stations

USGS station number	Station name	Basin area (km^2)	Latitude	Longitude	Date of the beginning of the period of record used in analysis	Date of the ending of the period of record used in analysis	Process variance, $\overline{Q}_{\Delta k}$ for a 1-day time step computed with a measurement error variance of $\overline{R} = 0{\cdot}04$ and $P_0{}^+ = 1$
01567000	Juniata River at Newport, PA	8687	40°28′42″	77°07′46″	July 29, 1952	Sept. 30, 1989	0·1357
01638500	Potomac River at Point of Rocks, MD	25 000	39°16′25″	77°32′35″	July 12, 1966	Sept. 30, 1990	0·1015
01664000	Rappahannock River at Remington, VA	1610	38°31′50″	77°48′50″	July 9, 1965	Sept. 30, 1993	0·2893
02116500	Yadkin River at Yadkin College, NC	5910	35°51′24″	80°23′10″	Jan. 3, 1951	Sept. 30, 1989	0·1421
02131000	Pee Dee River at Peedee, SC	22 900	34°12′15″	79°32′55″	Sept. 25, 1968	Sept. 30, 1972	0·0799
02175000	Edisto River near Givhans, SC	7070	33°01′40″	80°23′30″	Mar. 26, 1967	Sept. 30, 1972	0·1959
09180500	Colorado River near Cisco, UT	16 400	38°48′38″	109°17′34″	May 1, 1968	Sept. 30, 1984	0·1665
09315000	Green River at Green River, UT	116 000	38°59′10″	110°09′02″	Aug. 19, 1968	Sept. 30, 1984	0·1025
09379500	San Juan River near Bluff, UT	59 600	37°08′49″	109°51′51″	Dec. 17, 1951	Dec. 3, 1958	0·1127
09382000	Paria River at Lees Ferry, AZ	3650	36°52′20″	111°35′38″	Oct. 1, 1948	Sept. 30, 1976	0·5496

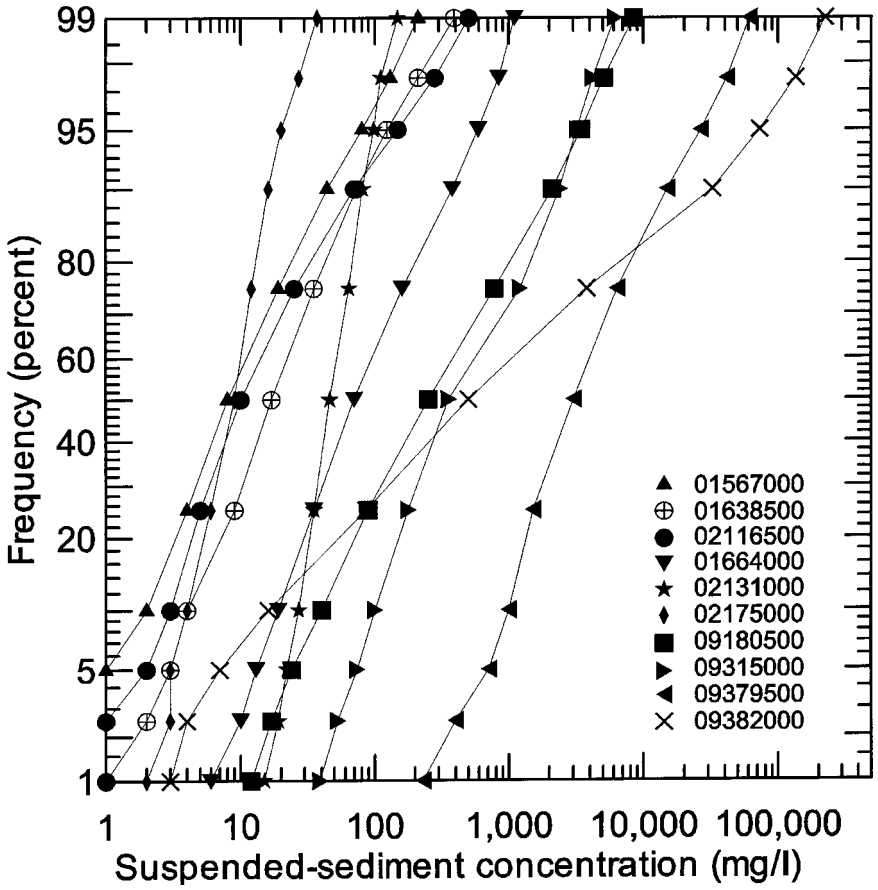

Figure 2. Distribution of daily mean concentrations of suspended sediment at selected gauging stations

Interpolators

Interpolation provides a mechanism for time-averaging direct measurements. Both linear and nonlinear interpolation is used to compute daily suspended-sediment concentrations from unequally-spaced measurements (Koltun *et al*., 1994). Interpolation provides estimates that match direct measurements of concentration exactly at the time of measurement. Linear interpolation approximates concentrations between times of direct measurements as straight-line segments connecting log-transformed concentrations on linear time scales. Nonlinear interpolation is based on a cubic-spline function between log-transformed values on a linear time scale. This interpolation produces a continuously differentiable arc that approximates a manually drawn curve. Estimates of concentrations from interpolations are obtained by inverse log transformation (exponentiation).

Regression estimators

Regression models can be used to flow-weight estimates of suspended-sediment concentration. These models describe a static statistical relation between suspended-sediment concentrations and the corresponding set of explanatory variables. Explanatory variables are selected based on their correlation with suspended-sediment concentrations and their general availability. Once developed, regression equations are used to estimate suspended-sediment concentrations during periods when direct measurements are unavailable.

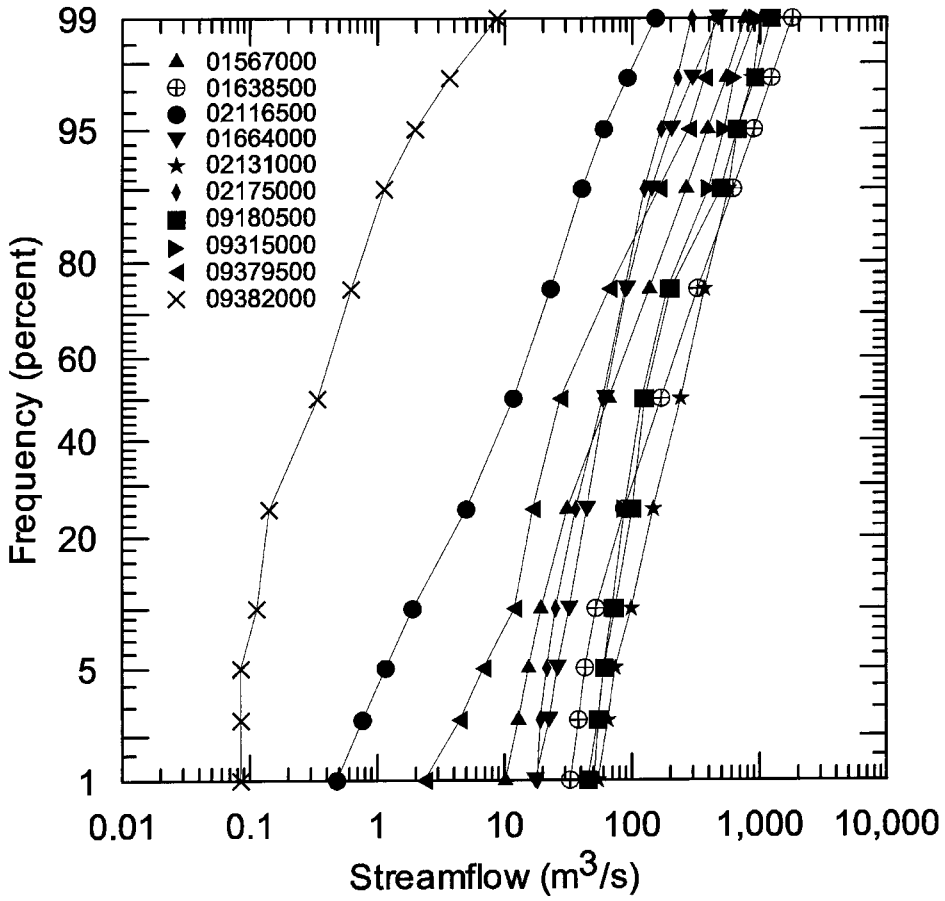

Figure 3. Distribution of daily mean streamflow rates at selected sediment gauging stations

The general form of a regression equation is:

$$y_k = \mathbf{u}_k \cdot \boldsymbol{\beta} + \varepsilon_k \tag{1}$$

where y_k is the log of the suspended-sediment concentration at time k. Suspended sediment concentrations, Y_k, are commonly measured in milligrams per litre; $\mathbf{u}_k$ is a $(p+1)$-dimensional row vector of explanatory variables at time k, where p is the number of explanatory variables, with 1 added to provide an intercept term; $\boldsymbol{\beta}$ is a $(p+1)$-dimensional column vector of parameters; and ε_k is the regression residual at time k. The set of residuals from regression equations is generally assumed to be independent and normally distributed with a mean of zero and a variance of σ^2, commonly written as $\varepsilon \sim \mathrm{NI}(0, \sigma^2)$.

In ordinary least-squares regression, the estimate of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta}_{\mathrm{ols}}$, is computed as:

$$\boldsymbol{\beta}_{\mathrm{ols}} = (\mathbf{U}^{\mathrm{T}} \cdot \mathbf{U})^{-1} \mathbf{U}^{\mathrm{T}} y \tag{2}$$

where the $n \times (p+1)$ matrix $\mathbf{U}$ is formed by horizontally concatenating n rows (observations) of the $\mathbf{u}_k$ vectors. The superscript T indicates a matrix transpose and the superscript -1 indicates a matrix inversion. Finally, the regression estimate of suspended-sediment concentration for time indexed by k is computed as:

$$y_{[\mathrm{ols}]k} = \mathbf{u}_k \cdot \boldsymbol{\beta}_{\mathrm{ols}} \tag{3}$$

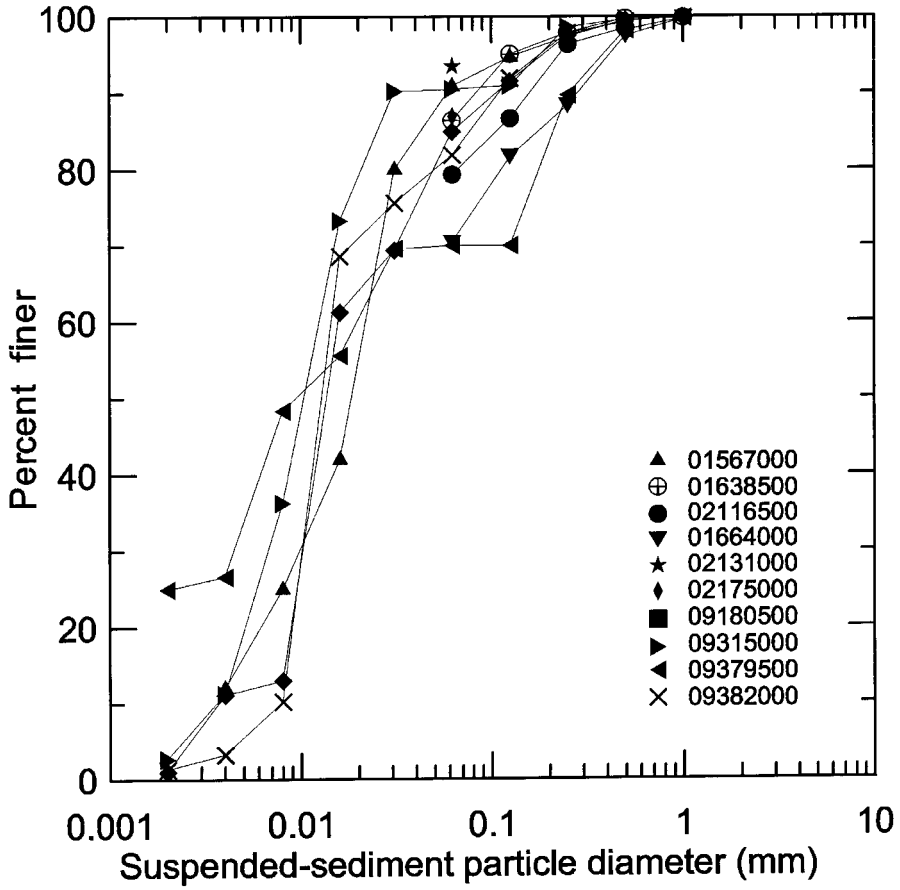

Figure 4. Distribution of suspended-sediment particle diameters at selected sediment gauging stations

The development of simple and multiple linear regression equations for estimating suspended-sediment concentrations is described in the following sections.

Simple linear regression

In this paper, a simple linear regression [slr] estimator refers to an equation for computing the log of suspended-sediment concentration on the basis of its variation with the log of streamflow. The estimator has the form:

$$y_{[\mathrm{slr}]k} = [\, u_{k,0} \quad u_{k,1} \,] \cdot \boldsymbol{\beta}_{[\mathrm{ols}]} \tag{4}$$

where $y_{[\mathrm{slr}]k}$ is the simple linear regression estimate at time indexed by k; $u_{k,0} = 1$ for all k; $u_{k,1}$ = log[streamflow] at time k; and $\boldsymbol{\beta}_{[\mathrm{ols}]}$ is a column vector containing the elements $\beta_{[\mathrm{ols}]_0}$ and $\beta_{[\mathrm{ols}]_1}$ of corresponding least-squares estimates.

Simple linear regression equations for computing the log of suspended-sediment concentrations for the 10 selected sites are summarized in Table II. Results indicate that the logs of suspended-sediment concentrations consistently were related positively to the logs of streamflow. The accuracy of the regression equations, however, varied widely among sites. Interpreting the coefficient of determination, r^2, as the fraction of variability explained by the regression, the slr estimator described a minimum of 1·1% of the variability in logs of suspended-sediment concentrations at Edisto River near Givhans, SC, and a maximum of 51·4% of the

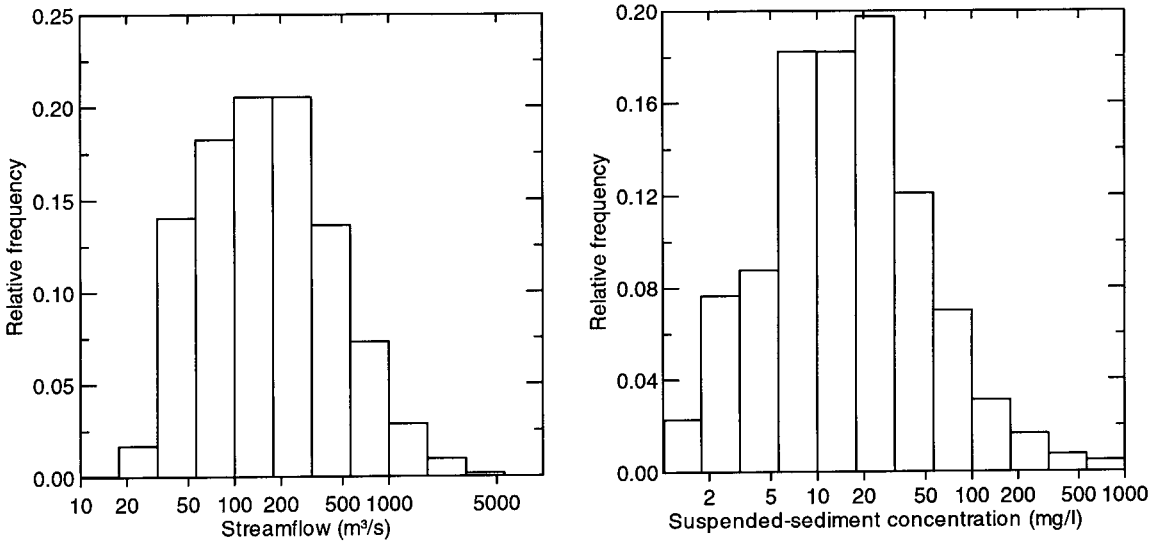

Figure 5. Distribution of suspended-sediment concentrations and streamflow at Potomac River at Point of Rocks, MD

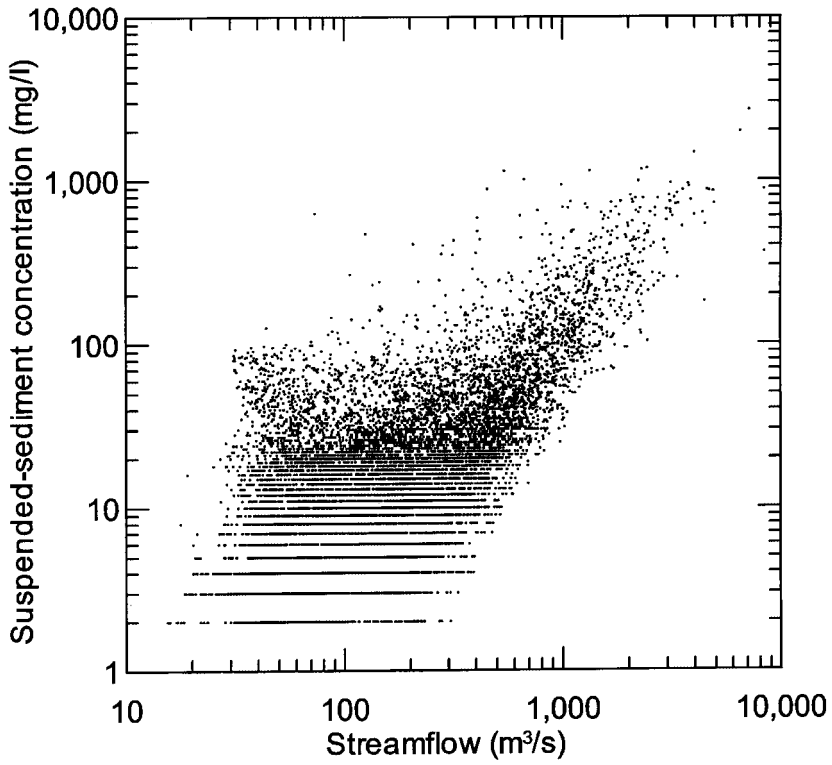

Figure 6. Relation between streamflow and suspended-sediment concentration at Juniata River at Newport, PA

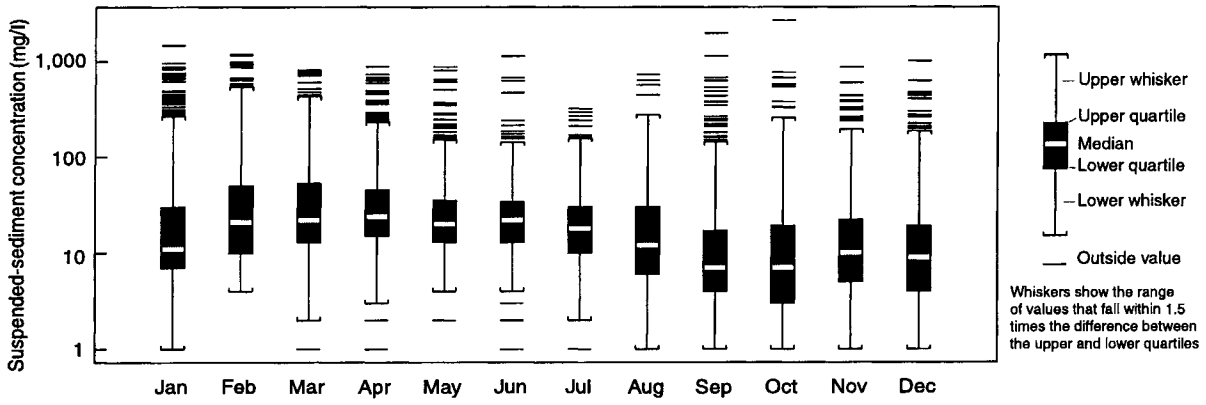

Figure 7. Monthly variation in suspended-sediment concentrations at Potomac River at Point of Rocks, MD

variability in logs of suspended-sediment concentrations at Paria River at Lees Ferry, AZ. Residuals of all slr equations were highly autocorrelated as indicated by Durbin–Watson d-statistic values less than 2 (Table II), thus violating the assumption of independent residuals associated with ordinary least-squares regression.

Multiple linear regression

Relative to the simple linear regressions, multiple linear regression [mlr] equations increase the number of explanatory variables to improve the estimation accuracy. The form of this equation is:

$$y_{[\mathrm{mlr}]k} = [\, u_{k,0} \quad u_{k,1} \quad u_{k,2} \quad u_{k,3} \quad u_{k,4} \,] \cdot \boldsymbol{\beta}_{[\mathrm{ols}]} \tag{5}$$

where $u_{k,0}$ and $u_{k,1}$ are as defined previously; $u_{k,2} = \sin[2\pi \cdot j_k/366]$, where j_k is the day number of the year such that, in a leap year, $j_1 = 1$ corresponds to January 1 and $j_{366} = 366$ corresponds to December 31; $u_{k,3} = \cos[2\pi \cdot j_k/366]$; and $u_{k,4} = (u_{k,1} - u_{k-1,1})/(t_k - t_{k-1})$, where t_k is the time indexed by k. Together the numerator and denominator provide an Euler approximation to the change in streamflow at time k. Again $\boldsymbol{\beta}_{[\mathrm{ols}]}$ is a column vector of corresponding ordinary least-squares parameter estimates.

The parameter estimates $\beta_{[\mathrm{ols}]_2}$ and $\beta_{[\mathrm{ols}]_3}$ describe a sinusoid of amplitude A and phase φ to approximate the seasonal variability of log-transformed suspended-sediment concentrations. In particular, the amplitude and phase parameters are computed as:

$$A = \sqrt{\beta^2_{[\mathrm{ols}]_2} + \beta^2_{[\mathrm{ols}]_3}} \tag{6}$$

$$\varphi = \tan^{-1}\left[\frac{\beta_{[\mathrm{ols}]_3}}{\beta_{[\mathrm{ols}]_2}}\right] \tag{7}$$

and the resulting seasonal component can be written:

$$S_k = A \sin\left[\frac{2\pi \cdot j_k}{366} + \varphi\right] \tag{8}$$

In the mlr equations (Table II), logs of streamflow were again positively associated with logs of suspended-sediment concentrations. In addition, with the exception of Paria River at Lees Ferry, AZ, positive changes in daily streamflow were associated with increasing suspended-sediment concentrations. Finally, a seasonal component in logs of sediment concentrations was consistently detected at all sites. Conditioned on streamflow, the day of lowest average suspended-sediment concentrations was February 15 and the corresponding day of highest average concentrations was August 17. The amplitude of the seasonal component varied from

Table II. Model parameters and summary statistics for selected stations (slr indicates simple linear regression, mlr indicates multiple linear regression, gls indicates generalized least squares)

Station name	Equation	Number	Model parameter estimates (standard errors)						Coefficient of determination: total (partial)	Root mean square error	Dubin–Watson *d*-statistic
			β_0	β_1	β_2	β_3	β_4	ϕ			
Juniata River at Newport, PA	slr	13 576	−1·4181	0·8506	—	—	—	—	0·4814	0·8857	0·2225
			(0·0329)	(0·0076)							
	mlr		−1·8063	0·9419	−0·1855	−0·3678	0·8000	—	0·5531	0·8222	0·2484
			(0·0394)	(0·0091)	(0·0128)	(0·0101)	(0·0301)				
	gls		−3·2337	1·2787	−0·4795	−0·4176	0·2405	−0·9000	(0·4563)	0·3782	2·0160
			(0·0762)	(0·0163)	(0·0474)	(0·0454)	(0·0149)	(0·0037)	0·9055		
Potomac River at Point of Rocks, MD	slr	8845	−0·6054	0·6757	—	—	—	—	0·3115	0·9375	0·1657
			(0·0562)	(0·0107)							
	mlr		−2·0158	0·9467	−0·4504	−0·6860	0·6511	—	0·5389	0·7673	0·2216
			(0·0574)	(0·0110)	(0·0141)	(0·0118)	(0·0376)				
	gls		−3·8526	1·3011	−0·7132	−0·7656	0·1427	−0·9103	(0·4498)	0·3371	1·9024
			(0·1056)	(0·0189)	(0·0572)	(0·0558)	(0·0188)	(0·0044)	0·9110		
Rappahannock River at Remington, VA	slr	10 309	0·8394	0·7193	—	—	—	—	0·4516	0·9564	0·4121
			(0·0205)	(0·0078)							
	mlr		0·5849	0·8266	−0·2893	−0·3593	0·5557	—	0·5335	0·8823	0·4274
			(0·0226)	(0·0089)	(0·0145)	(0·0128)	(0·0224)				
	gls		0·0950	1·0358	−0·4733	−0·4386	0·2585	−0·8098	(0·4755)	0·5331	2·1599
			(0·0473)	(0·0164)	(0·0414)	(0·0395)	(0·0135)	(0·0058)	0·8297		
Yadkin River at Yadkin College, NC	slr	14 149	−1·1009	1·2976	—	—	—	—	0·5113	0·7980	0·2655
			(0·0451)	(0·0107)							
	mlr		−1·8505	1·4760	−0·2676	−0·6131	0·1280	—	0·6743	0·6516	0·3907
			(0·0419)	(0·0099)	(0·0086)	(0·0078)	(0·0203)				
	gls		−1·6176	1·4203	−0·2469	−0·6083	0·1012	−0·8059	(0·5373)	0·3863	2·0556
			(0·0642)	(0·0148)	(0·0242)	(0·0236)	(0·0124)	(0·0050)	0·8855		

Pee Dee River at Peedee, SC	slr	1465	2·3975	0·2657	—	—	—	—	0·1490	0·4186	0·7699
			(0·0912)	(0·0166)							
	mlr		2·0884	0·3221	−0·1076	−0·1825	0·6187	—	0·3300	0·3718	0·8438
			(0·0938)	(0·0171)	(0·0154)	(0·0141)	(0·0473)				
	gls		1·5565	0·4194	−0·1476	−0·1977	0·4770	−0·5977	(0·2739)	0·3020	2·2011
			(0·1662)	(0·0302)	(0·0303)	(0·0282)	(0·0404)	(0·0213)	0·5584		
Edisto River near Givhans, SC	slr	2014	1·7393	0·0947	—	—	—	—	0·0112	0·5530	0·7885
			(0·0814)	(0·0198)							
	mlr		1·1650	0·2328	−0·1821	−0·2372	0·4535	—	0·1334	0·5181	0·8966
			(0·0863)	(0·0210)	(0·0180)	(0·0169)	(0·1579)				
	gls		1·1594	0·2344	−0·1824	−0·2375	0·4760	−0·5515	(0·0458)	0·4323	2·3326
			(0·1571)	(0·0382)	(0·0333)	(0·0313)	(0·2024)	(0·0186)	0·3970		
Colorado River near Cisco, UT	slr	5995	−0·2568	1·1631	—	—	—	—	0·3059	1·2554	0·1318
			(0·1151)	(0·0226)							
	mlr		2·1337	0·6849	0·1625	−0·8998	2·8417	—	0·4607	1·1069	0·1967
			(0·1204)	(0·0238)	(0·0213)	(0·0232)	(0·1688)				
	gls		−5·2931	2·1628	−0·2003†	−0·1957†	0·2827	−0·9525	0·1892	0·4360	1·9313
			(0·4063)	(0·0773)	(0·1586)	(0·1620)	(0·0703)	(0·0040)	0·9164		
Green River at Green River, UT	slr	5885	0·3445	1·1624	—	—	—	—	0·3857	0·9593	0·1343
			(0·09523)	(0·0191)							
	mlr		2·5918	0·7069	0·2248	−0·6401	0·8802	—	0·4828	0·8805	0·1691
			(0·1129)	(0·0228)	(0·0178)	(0·0197)	(0·1180)				
	gls		0·2314	1·1856	0·0693	−0·4079	0·0179†	−0·9261	(0·1315)	0·3451	1·7784
			(0·2720)	(0·0537)	(0·0853)	(0·0882)	(0·0480)	(0·0049)	0·9206		
San Juan River near Bluff, UT	slr	2542	6·2933	0·5177	—	—	—	—	0·2540	0·9757	0·1623
			(0·0649)	(0·0176)							
	mlr		6·1320	0·5636	−0·2243	0·0075	0·6477	—	0·2849	0·9558	0·1777
			(0·0796)	(0·0220)	(0·0297)	(0·0316)	(0·0913)				
	gls		3·6322	1·2743	−0·6377†	0·5436†	−0·1473	−0·9483	(0·3165)	0·3607	1·9641
			(0·2051)	(0·0434)	(0·1871)	(0·1875)	(0·0355)	(0·0063)	0·8982		
Paria River at Lees Ferry, AZ	slr	10225	8·4937	1·9950	—	—	—	—	0·5144	1·9206	0·2317
			(0·0278)	(0·0192)							
	mlr		9·0441	2·5180	−0·3541	−1·2712	−1·0168	—	0·6315	1·6733	0·2935
			(0·0261)	(0·0191)	(0·0234)	(0·0258)	(0·0327)				
	gls		7·8562	1·3864	−0·2789	−0·6351	−0·2662	−0·9191	(0·3736)	0·7660	1·8892
			(0·0961)	(0·0209)	(0·1293)	(0·1304)	(0·0148)	(0·0039)	0·9228		

† Parameter not statistically significant at the 0·05 level.

a maximum of 1·0463 (in log of milligram per litre units) at Potomac River at Point of Rocks, MD, to a minimum of 0·2467 at Pee Dee River at Peedee, SC. Similar to the slr results, significant autocorrelation was present in the residuals of mlr estimates.

State-space estimators

State-space models provide a basis for optimal estimation, that is minimizing the error of the estimate of the state, by utilizing knowledge of system and measurement dynamics, assumed statistics of system noises and measurement errors, and initial condition information (Gelb, 1974). State-space models disaggregate a dynamic system such as suspended-sediment concentrations into process and measurement components. The process component describes the evolution of the system dynamics and their associated uncertainty. The measurement component describes the static effects of explanatory variables and the uncertainty in the measurement process. Together, these components form an estimator that continually accounts for the effects of known inputs (such as streamflow) and optimally adjusts model estimates for periodic direct measurements that contain some uncertainty.

The process of adjusting for direct measurements is described as predicting, filtering, or smoothing, depending on the set of direct measurements used in estimation. Predicting uses only measurements prior to the time of estimation; filtering uses only measurements up to and including the time of estimation; and smoothing uses measurements before and after the time of estimation. Predicted and filtered estimates provide on-line data, that is, information that can be continuously updated to the present (real time). Smoothed estimates provide off-line data only, that is, estimates are delayed until subsequent direct measurements become available. The accuracy of off-line estimates, however, is generally greater than that of corresponding on-line values. The length of the delay is determined by measurement frequency. Both on-line and off-lines estimators will be developed and analysed in this paper, although the off-line estimators are of primary interest for publication of suspended-sediment concentrations. In this paper, the state-space models are developed in discrete, rather than continuous, time based on average values that are reported (sampled) at a time step of 1 day. Thus, the state-space models are in the form of difference rather than differential equations. This implies that there is no information available about the distribution of sediment concentrations at shorter (unit) time intervals. Although this reporting interval limits the dynamic characterization of sediment concentrations, the characterization based on the 1-day sampling interval can be applied to shorter time intervals if appropriate adjustments are applied. Applications to daily values are discussed in the following sections and modifications needed for estimation at shorter time intervals are discussed in a subsequent section.

A generalized least-squares regression equation is used as a transition between regression models and state-space models. In this paper, the generalized least-squares [gls] model is of the form:

$$\begin{aligned} y_k &= [\, u_{k,0} \quad u_{k,1} \quad \ldots \quad u_{k,p} \,] \cdot \boldsymbol{\beta} + \chi_k \\ \chi_k &= \phi \chi_{k-1} + \varepsilon_k \end{aligned} \tag{9}$$

which is similar to the multiple-regression equation developed previously, with the addition of a second equation containing the parameter ϕ describing the error component as a first-order autoregressive process. The parameters of this model, $[\boldsymbol{\beta}_{[\mathrm{gls}]} \quad \phi_{[\mathrm{gls}]}]$, were estimated iteratively by specification of the maximum likelihood option in the SAS/ETS® AUTOREG procedure (SAS Institute, 1988).[1] Although only series without missing data were used in this paper, the maximum likelihood estimation technique is applicable to series with missing values.

Results of estimation for selected sites (Table II) indicate that the significance of $\boldsymbol{\beta}_{[\mathrm{gls}]}$ is maintained when the $\phi_{[\mathrm{gls}]}$ parameter is added. The root mean square error (RMSE) of residuals from predicting logs of

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

suspended-sediment concentration 1 day in advance of the current time step is significantly lower than for the residuals from the corresponding multiple-regression equations. The total r^2 of the full model, including both static and dynamic components, improved significantly, while the autocorrelation of the residuals diminished greatly. The r^2 value associated with the static component, $\mathbf{u}_k \cdot \boldsymbol{\beta}_{[\mathrm{gls}]}$, however, averaged slightly less than the comparable multiple-regression component, $\mathbf{u}_k \cdot \boldsymbol{\beta}_{[\mathrm{ols}]}$ (Table II).

To complete the transition from the gls regression model to the state-space model, the error component is disaggregated into a process noise component, w_k, and a measurement noise component, v_k, as:

$$\varepsilon_k = w_k + v_k \tag{10}$$

where the w_k and v_k are normally distributed sequences with expected values of zero, $E[w_k] = E[v_k] = 0$; the covariance of the process error is uncorrelated and has magnitude Q_k, $E[w_k w_j^{\mathrm{T}}] = Q_k \delta_{kj}$; [2] measurement errors are uncorrelated with covariance R_k, $E[v_k v_j^{\mathrm{T}}] = R_k \delta_{kj}$, and process and measurement errors are assumed to be uncorrelated at all times, $E[w_k v_j^{\mathrm{T}}] = 0$ for j and k. Because of this last assumption, a delay can be introduced in process error without consequence, so that the model can be written in standard state-space form as:

$$\begin{aligned} \chi_k &= \phi_{[\mathrm{gls}]} \cdot \chi_{k-1} + w_{k-1} \\ y_k &= \chi_k + \mathbf{u}_k \cdot \boldsymbol{\beta}_{[\mathrm{gls}]} + v_k \end{aligned} \tag{11}$$

where the upper equation is referred to as the state equation and the corresponding autoregressive error component is considered to be the state variable. The lower equation is referred to as the measurement equation. For this analysis, the measurement equation was simplified by subtracting the static component $\mathbf{u}_k \cdot \boldsymbol{\beta}_{[\mathrm{gls}]}$ as:

$$\begin{aligned} \chi_k &= \phi_{[\mathrm{gls}]} \chi_{k-1} + w_{k-1} \\ \tilde{y}_k &= \chi_k + v_k \end{aligned} \tag{12}$$

On-line estimation

Given initial estimates of the state χ_0 (the autoregressive error), and an associated state error covariance P_0, the state-space model can be revised continuously to estimate the magnitude and uncertainty of the state by use of the Kalman filtering algorithm. In this algorithm, the revision takes place in two parts: a temporal projection, determined by the system dynamics having an error covariance Q_k, and a measurement update, determined by the accuracy of measurements having an error covariance R_k. In this paper, the estimate $\chi^-_{[\mathrm{F}]k}$ refers to the temporal projection of the state at time k, based on information available at $k-1$. In contrast, $\chi^+_{[\mathrm{F}]k}$ refers to the estimate of the state at time step k, formed by updating the temporal projection with measurement information available at time k. Similar notation is used to designate corresponding state error covariances $P^-_{[\mathrm{F}]k}$ and $P^+_{[\mathrm{F}]k}$. The subscript [F] denotes a filtered or on-line estimate of the state.

Details of the algebra of the Kalman filtering algorithm are described in the following. A projection of the state error covariance at time k, just prior to a measurement, is computed as:

$$P^-_{[\mathrm{F}]k} = \phi_{[\mathrm{gls}]} \cdot P^+_{[\mathrm{F}]k-1} \cdot \phi^{\mathrm{T}}_{[\mathrm{gls}]} + Q_{k-1} \tag{13}$$

Then, the Kalman gain, which optimally weights the reliability of the model estimate with the reliability of measurement data at time k, is computed as:

$$K_k = P^-_{[\mathrm{F}]k} [P^-_{[\mathrm{F}]k} + R_k]^{-1} \tag{14}$$

[2] The Kronecker delta function is defined as $\delta_{kj} = 1$ for $k = j$ and 0 otherwise.

If a measurement is available at time k, a measurement update is computed for the state and the state error covariance at time k, just after the measurement, as:

$$\chi^{+}_{[F]k} = \chi^{-}_{[F]k} + K_k[\tilde{y}_k - \chi^{-}_{[F]k}] \tag{15}$$

$$P^{+}_{[F]k} = [1 - K_k]P^{-}_{[F]k} \tag{16}$$

If no measurement is available, the updated state and state error covariances are the same as the temporal estimates. Finally, the index for k is incremented and a state prediction is computed as:

$$\chi^{-}_{[F]k} = \phi \cdot \chi^{+}_{[F]k-1} \tag{17}$$

The process is repeated as additional data become available.

In practice, highly reliable estimates of Q_k, R_k, $\chi^{-}_{[F]0}$, and $P^{+}_{[F]0}$ are seldom available. Useful estimates, however, can be obtained readily. Specifically, both Q_k and R_k are variances and therefore take on positive values. An estimate of R_k can be computed for each direct measurement by use of methods described by Burkham (1985), if sufficient information is available. Alternatively, temporal estimates have a low sensitivity to R_k so an average value, $\overline{R}$, may be adequate. If Q_k is assumed constant for a given time step of size k, $\overline{Q}_{\Delta k}$, then an initial estimate can be computed as $\sigma^2 - \overline{R}_k$, where σ^2 is estimated as the mean square error of the generalized least-squares equation. The final estimate of $\overline{Q}_{\Delta k}$ can be selected as the value that generates $P^{-}_{[F]}$, such that:

$$1 - \alpha = \Pr\left[\chi^{-}_{[F]} - z_{\alpha/2}\sqrt{P^{-}_{[F]}} < \tilde{y} < x^{-}_{[F]} + z_{\alpha/2}\sqrt{P^{-}_{[F]}}\right] \tag{18}$$

is satisfied (Siouris, 1996), where $z_{\alpha/2}$ is the standard normal quantile corresponding to a probability of $1 - \alpha/2$, and Pr indicates the probability. Finally, $\chi^{-}_{[F]0}$ is typically set equal to its expected value of zero and $P^{+}_{[F]0}$ is set somewhat larger than Q plus R. In this paper, suspended-sediment concentration data were available as daily averages rather than direct measurements. Estimates of $\overline{Q}_{\Delta k}$ (Table 1) were determined to satisfy Equation (18) based on $\overline{R} = 0{\cdot}04$, corresponding to measurement error of about 20%. A value of $P^{+}_{[F]0} = 1$ was used for all selected stations.

To complete the estimation process, the temporal estimate of the state is added to the static component to predict the log of suspended-sediment concentration at time k as:

$$y^{-}_{[F]k} = \mathbf{u}_k \cdot \boldsymbol{\beta}_{[gls]} + \chi^{-}_{[F]k} \tag{19}$$

The corresponding $100(1 - \alpha)\%$ confidence interval for y_k based on the prediction information is:

$$y^{-}_{[F]k} \pm z_{\alpha/2}\sqrt{P^{-}_{[F]k}} \tag{20}$$

Similarly, the measurement update of the state is added to the static component to compute a filtered estimate of the log of suspended-sediment concentration at time k as:

$$y^{+}_{[F]k} = \mathbf{u}_k \cdot \boldsymbol{\beta}_{[gls]} + \chi^{+}_{[F]k} \tag{21}$$

The $100(1 - \alpha)\%$ confidence interval for y_k based on the filtered information is:

$$y^{+}_{[F]k} \pm z_{\alpha/2}\sqrt{P^{+}_{[F]k}} \tag{22}$$

Off-line estimation

In this paper, a smoother is a mathematical procedure that combines a forward running filter estimate with a backward running filter estimate. Thus, all data before and after the time for which the estimate is computed

are used to determine an optimal value. Smoothers are based on more data than forward running filters and are generally more accurate. Smoothing is considered an off-line estimation procedure because estimates are delayed until measurements at the end of the estimation intervals become available.

There are three types of smoothers in general use. First, a fixed-interval smoother is an estimator that provides optimal values of all states within an estimation interval defined by beginning and ending measurements. Second, a fixed-point smoother is an estimator that provides an optimal value for a fixed point in time (such as an initial condition). And third, a fixed-lag smoother is an estimator that provides an optimal value of the state at a fixed time interval after the most recent measurement. For this application, a fixed-interval smoother algorithm referred to as the Rauch–Tung–Striebel (RTS) smoother (Gelb, 1974; Grewal and Andrews, 1993) is used.

To implement the smoother, the Kalman filter is run up to the measurement ending the estimation interval. All state and covariance elements computed by the filter are utilized. The initial condition for the smoother, denoted by a subscript [S], is the filter estimate formed by the measurement update at the end of the estimation interval as:

$$\chi_{[S]N} = \chi^{+}_{[F]N} \tag{23}$$

Then, starting at the end of the period, the smoother gain is computed as:

$$A_k = P^{+}_{[F]k} \cdot \phi^{T}_{[gls]} [P^{-}_{[F]k+1}]^{-1} \tag{24}$$

and moving backward in time through the estimation interval, the smoother estimate is updated as:

$$\chi_{[S]k} = \chi_{[F]k} + A_k(\chi_{[S]k+1} - \chi^{-}_{[F]k+1}) \tag{25}$$

The state error covariance of the smoother is:

$$P_{[S]k} = P^{+}_{[F]k} + A_k[P_{[S]k+1} - P^{-}_{[F]k+1}]A^{T}_k \tag{26}$$

The time index is decremented and the above procedure is repeated until the beginning of the estimation interval to complete the computation of the smoothed estimate of the state and state error covariance.

Again, to complete the estimation process, the smoothed state estimate is added to the static component to estimate the log of suspended-sediment concentration at time k as:

$$y_{[S]k} = \mathbf{u}_k \cdot \boldsymbol{\beta}_{[gls]} + \chi_{[S]k} \tag{27}$$

Finally, a $100(1-\alpha)\%$ confidence interval for y_k based on the smoothed information is:

$$y_{[S]k} \pm z_{\alpha/2} \sqrt{P_{[S]k}} \tag{28}$$

COMPARISON OF ESTIMATORS

Techniques for estimation of suspended-sediment concentrations described in this paper include interpolation, regression, and optimal estimators. Figure 8 provides a basis for graphical comparison of some of these estimators for the hypothetical situation in which only one of 12 daily values is available to estimate the complete record. Results for the interpolation techniques, which are represented here by cubic-spline interpolation, indicate that interpolation fits the selected (direct) measurements, but fails to account for streamflow influences and often results in a poor match between the estimates and suspended-sediment concentrations not used in the estimation. Similarly, simple linear regression may result in poor estimates because it fails to adequately account for the concentration measurements used in estimation, even though the estimates are all conditioned on streamflow. Optimal estimators, represented by the off-line [Smooth] estimator, effectively account for streamflow (and seasonal) influences, and for information provided by (direct)

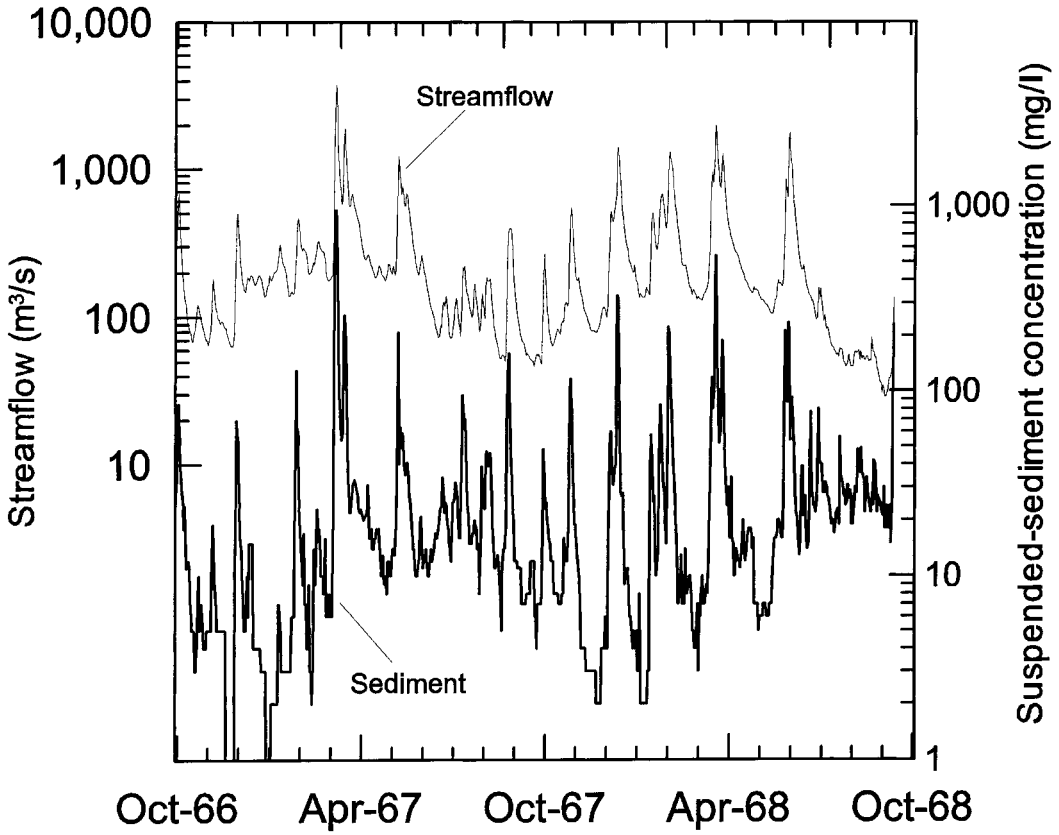

Figure 8. Suspended-sediment concentration and streamflow at Juniata River at Newport, PA (US Geological Survey gauging station 01567000)

measurements used in the estimation. In addition, the optimal estimators provide a measure of uncertainty of the estimated record (Figure 9).

In addition to graphical methods, summary statistics were computed to facilitate numerical comparison of the accuracies of the alternative estimators. Specifically, the accuracies of simulated sampling intervals of 3, 6, 12, 24, and 48 days, separated by corresponding estimation intervals of 2, 5, 11, 23, and 47 consecutive unsampled days, were investigated. The interpolators, filters, and smoothers were updated using data from the sampled days, and the accuracy of the estimators was assessed on the basis of RMSE of log concentration estimates on unsampled days (thus the effective sample size increased with the length of the sampling interval). Results indicate that the off-line [S] estimator was the most accurate (Table III, Figure 10). Although the accuracy of the interpolators was high at shorter sampling intervals, this accuracy decreased rapidly with increasing sampling interval. The accuracy of regression estimators did not improve locally in response to direct measurements of suspended-sediment concentrations.

The RMSE can be expressed either in log units (as above) or as a standard error in percent. The RMSE in percent reflects the coefficient of variation of the estimator as:

$$\mathrm{RMSE}_{\mathrm{percent}} = 100\left|\frac{\sigma_y}{\mu_y}\right| = 100\left|\sqrt{e^{\sigma_y^2} - 1}\right| \tag{29}$$

Results from this analysis indicate that the average standard error for the slr estimator is 124% and that the average standard error for the mlr estimator is 105%. The average standard error of the linear interpolator

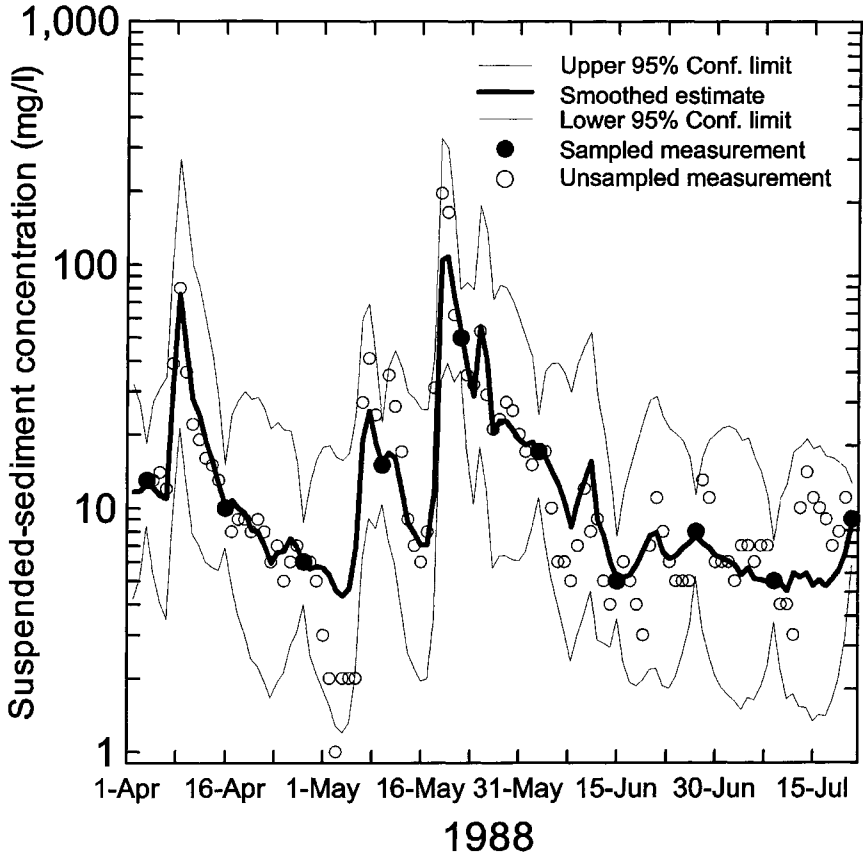

Figure 9. Estimates and uncertainties of suspended-sediment concentrations at Juniata River at Newport, PA (US Geological Survey gauging station 01567000)

ranged from 48·4% for 3-day sampling intervals to 159% for 48-day sampling intervals. The average standard error of the cubic-spline interpolator ranged from 50·6% for 3-day sampling intervals to 176% for 48-day sampling intervals. The average standard error for the on-line [F] estimator ranged from 52·7% for a 3-day sampling interval to 107% for a 48-day sampling interval. The average standard error for the off-line [S] estimator ranged from 39·9% for a 3-day sampling interval to 93·0% for a 48-day sampling interval. Thus, the off-line [S] estimator has the lowest standard error, especially at the shorter sampling intervals that are needed to compute continuous records of suspended-sediment concentrations.

Possible sources of systematic variation in estimation errors among sites were investigated. In particular, off-line root mean square errors showed a slight tendency to decrease with increasing median discharges. Although the number of sites analysed is thought to be too small to provide conclusive results, the finding is consistent with results by Phillips *et al.* (1999), who indicated that the accuracy and precision of the estimators that they evaluated declined with a reduction in drainage area. No relation between model errors and sediment characteristics was detected.

APPLICATION TO SEDIMENT COMPUTATION

Bias transformation adjustment

Exponentiation provides a simple transformation back to the original metric for estimates computed using a log transformation. The mean of the exponentiated values, however, estimates the median of the original

Table III. Estimation error characteristics at selected sites for specified sampling intervals (slr indicates simple linear regression, mlr indicates multiple linear regression, LinInt indicates linear interpolation, and CubSpl indicates cubio-spline interpolation). The average root mean square error (mg/l) is given for days without measurements

Station name	Estimator	Simulated measurement interval (days)				
		3	6	12	24	48
Juniata River at Newport, PA	slr	0·8868	0·8875	0·8877	0·8880	0·8875
	mlr	0·8231	0·8229	0·8240	0·8242	0·8235
	LinInt	0·4372	0·6430	0·8287	0·9930	1·1791
	CubSpl	0·4488	0·6762	0·8768	1·0455	1·2450
	On-line	0·4495	0·5555	0·6600	0·7483	0·8063
	Off-line	0·3339	0·4339	0·5272	0·6297	0·7297
	Sample size	9050	11 310	12 441	12 995	13 254
Potomac River at Point of Rocks, MD	slr	0·9388	0·9365	0·9380	0·9387	0·9387
	mlr	0·7686	0·7653	0·7664	0·7674	0·7675
	LinInt	0·3878	0·5780	0·7669	0·9514	1·0680
	CubSpl	0·3893	0·6068	0·8124	1·0193	1·1509
	On-line	0·4049	0·5044	0·5945	0·6859	0·7532
	Off-line	0·3054	0·3840	0·4746	0·5723	0·6869
	Sample size	5896	7370	8107	8464	8648
Rappahannock River at Remington, VA	slr	0·9598	0·9567	0·9596	0·9584	0·9590
	mlr	0·8812	0·8806	0·8834	0·8834	0·8845
	LinInt	0·6725	0·8823	1·0172	1·1663	1·2800
	CubSpl	0·6989	0·9368	1·0705	1·2338	1·3480
	On-line	0·6018	0·6878	0·7793	0·8449	0·8795
	Off-line	0·4888	0·5791	0·6698	0·7756	0·8458
	Sample size	6872	8590	9449	9867	10 058
Yadkin River at Yadkin College, NC	slr	0·7984	0·8007	0·8004	0·7983	0·7985
	mlr	0·6541	0·6553	0·6540	0·6526	0·6525
	LinInt	0·4908	0·6929	0·8979	1·0140	1·1103
	CubSpl	0·5086	0·7407	0·9604	1·0810	1·1814
	On-line	0·4406	0·5164	0·5766	0·6150	0·6360
	Off-line	0·3529	0·4353	0·5141	0·5712	0·6136
	Sample size	9432	11 790	12 969	13 547	13 818
Pee Dee River at Peedee, SC	slr	0·4162	0·4174	0·4167	0·4167	0·4197
	mlr	0·3678	0·3706	0·3694	0·3705	0·3715
	LinInt	0·3400	0·4133	0·4440	0·4830	0·5024
	CubSpl	0·3583	0·4388	0·4688	0·5098	0·5544
	On-line	0·3176	0·3434	0·3627	0·3706	0·3721
	Off-line	0·2834	0·3189	0·3489	0·3628	0·3673
	Sample size	976	1220	1342	1403	1410
Edisto River near Givhans, SC	slr	0·5580	0·5557	0·5555	0·5536	0·5529
	mlr	0·5235	0·5220	0·5212	0·5186	0·5191
	LinInt	0·4272	0·4526	0·4622	0·4918	0·5215
	CubSpl	0·4604	0·4794	0·4893	0·5277	0·5407
	On-line	0·4543	0·4822	0·5038	0·5109	0·5160
	Off-line	0·4129	0·4481	0·4849	0·5009	0·5124
	Sample size	1342	1675	1837	1909	1927
Colorado River near Cisco, UT	slr	1·2530	1·2580	1·2573	1·2584	1·2560
	mlr	1·1054	1·1097	1·1083	1·1097	1·1068
	LinInt	0·4056	0·5936	0·7888	0·9177	1·0585
	CubSpl	0·4226	0·6222	0·8369	0·9609	1·1271
	On-line	0·5147	0·6913	0·8617	1·0281	1·1851
	Off-line	0·3704	0·5062	0·6617	0·7788	0·9226
	Sample size	3996	4995	5489	5727	5828

(*continued on next page*)

Table III. (*continued*)

Station name	Estimator	Simulated measurement interval (days)				
		3	6	12	24	48
San Juan River near Bluff, UT	slr	0·9616	0·9570	0·9597	0·9588	0·9590
	mlr	0·8823	0·8776	0·8802	0·8797	0·8790
	LinInt	0·3148	0·4306	0·5876	0·7411	0·8683
	CubSpl	0·3183	0·4506	0·6245	0·7836	0·9230
	On-line	0·4383	0·5384	0·6557	0·7403	0·8449
	Off-line	0·3079	0·4020	0·5340	0·6460	0·7550
	Sample size	3922	4900	5390	5635	5734
Green River at Green River, UT	slr	0·9730	0·9757	0·9770	0·9775	0·9775
	mlr	0·9536	0·9563	0·9552	0·9562	0·9562
	LinInt	0·3500	0·5145	0·7679	0·9256	1·1197
	CubSpl	0·3619	0·5357	0·8225	0·9588	1·1775
	On-line	0·4164	0·5627	0·6966	0·8291	0·9546
	Off-line	0·3041	0·4018	0·5354	0·6558	0·8054
	Sample size	1694	2115	2321	2415	2444
Paria River at Lees Ferry, AZ	slr	1·9171	1·9195	1·9209	1·9191	1·9187
	mlr	1·666	1·6712	1·6731	1·6729	1·6722
	LinInt	0·7965	1·1408	1·5457	2·0766	2·5059
	CubSpl	0·8090	1·1846	1·6546	2·2199	2·6274
	On-line	0·9152	1·2240	1·4594	1·6589	1·8124
	Off-line	0·6450	0·8960	1·1448	1·4237	1·6553
	Sample size	6816	8520	9372	9798	10 011

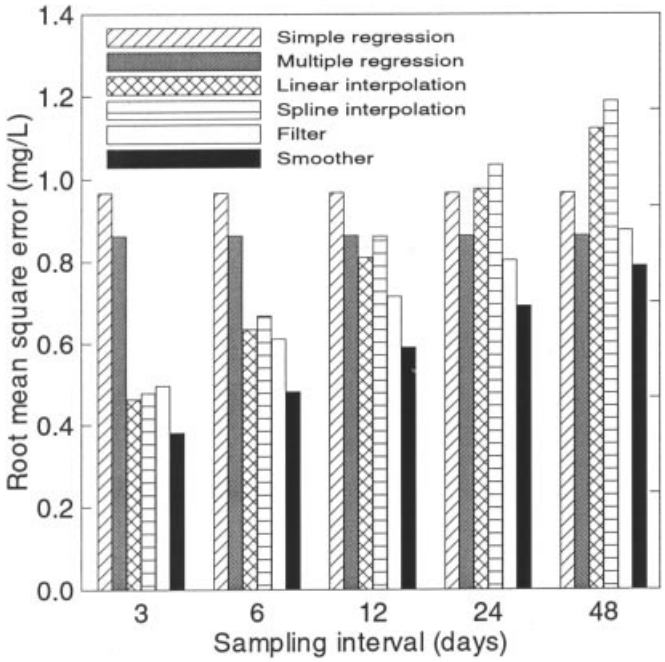

Figure 10. Average root mean square error of selected estimators for specified sampling intervals

population rather than the mean. For log-normally distributed populations, the median is less than the mean. Thus, the mean of the exponentiated estimates generally is thought to underestimate the mean suspended-sediment concentration. In contradiction, Walling and Webb (1988) show that exponentiated estimates that are adjusted for this possible bias are significantly less accurate and less precise than estimates that are not adjusted, based on load calculations derived from rating curves. If adjustment is desired, however, the 'smearing' approach (Helsel and Hirsch, 1992) provides a nonparametric bias correction factor (BCF) applicable to optimal estimates of suspended-sediment concentrations. For off-line [S] estimates, the BCF is computed as the mean of the exponentiated residuals:

$$\mathrm{BCF} = \sum_{i=1}^{n} \frac{\exp[y_i - y_{[S]i}]}{n} \tag{30}$$

The residuals, which may follow any distribution, are only assumed to be independent and homoscedastic. For application, the BCF is multiplied by the exponentiation smoothed estimates to compute the mean suspended-sediment concentrations.

Unit value computation

Because of the nonlinear relation (in arithmetic space rather than log space) between streamflow and suspended-sediment concentrations and occasionally large daily ranges in streamflow, daily average suspended-sediment concentrations are typically computed based on unit (less than daily) values rather than daily average streamflow values. It is anticipated that future applications of the estimators will be based on a development from unit values and direct measurement data. If the unit value data are difficult to obtain, however, the dynamics determined from analysis of daily values data can be applied to shorter intervals, with adjustments described in the following section.

The state-space model contains static components, associated with the effects of streamflow and seasonality, and dynamic components, associated with the error characteristics. The explanatory variable associated with the static effect of changes in streamflow rate, $u_{k,4}$, on sediment concentrations must be recomputed for the size of the unit time step. The parameters associated with the static effects, however, may still be applicable. In contrast, values of ϕ and Q_k characterize the dynamic error in sediment concentrations at a time step of 1 day. The following section describes how to adjust daily estimates of ϕ and Q_k for computation at unit time intervals.

The discrete time (difference) equation for the state equation was developed with a time index k having a constant size Δk equal to 1 day and was expressed as:

$$\chi_k = \phi \cdot \chi_{k-1} + w_{k-1} \tag{31}$$

To convert this equation for use with an alternate size of time step, Equation (31) is first converted to a continuous time (differential) equation of the form:

$$\dot{\chi}_t = F\chi_t + w_t \tag{32}$$

which indicates that the rate of change in the state variable is proportional to its present state. This proportionality factor is simply computed as:

$$F = \log[\phi] \tag{33}$$

For application at a new sampling interval Δk_1, equal to say 1/24 day, and indexed by k_1, ϕ and the state equation are revised as:

$$\begin{aligned} \phi_1 &= \mathrm{e}^{F \cdot \Delta k_1} \\ \chi_{k_1} &= \phi_1 \cdot \chi_{k_1 - 1} + w_{k_1 - 1} \end{aligned} \tag{34}$$

with a result that $\phi_1 > \phi$ for unit time intervals shorter than 1 day. Converting Q_k to an alternate time step is an extension of this procedure. First, the discrete process variance determined at $\Delta k = 1$ is related to the continuous process variance Q_τ as:

$$Q_{\Delta k} = \mathrm{e}^{F\Delta k}\left\{\int_0^{\Delta k} \mathrm{e}^{-F\tau} Q_\tau \,\mathrm{e}^{-F_\tau^{\mathrm{T}}}\,\mathrm{d}\tau\right\}\mathrm{e}^{F_{\Delta k}^{\mathrm{T}}} \tag{35}$$

(Grewal and Andrews, 1993). Solving Equation (35) for Q_τ results in:

$$Q_\tau = \frac{2F \cdot Q_{\Delta k}}{-1 + \mathrm{e}^{2F\cdot\Delta k}} \tag{36}$$

The corresponding discrete time covariance for discrete time interval Δk_1 can then be computed as:

$$Q_{\Delta k_1} = Q_\tau \,\mathrm{e}^{2F\Delta k_1}\left[\frac{-1}{2\,\mathrm{e}^{2F\cdot\Delta k_1}F} + \frac{1}{2F}\right] \tag{37}$$

Finally, the magnitude of the measurement covariance, R_k, may be decreased because the direct (instantaneous) measurements of suspended-sediment concentrations contain smaller time-sampling errors than those present in daily average values. Methods described by Burkham (1985) may be helpful in revising the estimate of R_k. Once the values of ϕ, Q_k, R_k, and $u_{k,4}$ are revised, the previously developed Kalman filter and smoother equations may be applied.

Implication for sampling design

The analyses described in this paper were based on long-term suspended-sediment concentration records, which had an average length of 6901 daily observations without missing values. Although application to shorter record length with missing data was not assessed, the following observations may provide a useful preliminary guide to these types of situations. The SAS AUTOREG procedure was used to estimate model parameters. The documentation for this procedure suggests that the maximum likelihood estimation method be used if missing values are 'plentiful' (SAS Institute, 1988, p. 177). Although 'plentiful' is not explicitly defined, it is assumed to be related to both the length of record and the percentage of missing values. Should data characteristics limit the number of parameters that can be estimated or found statistically significant by use of this method, a simpler model structure might be used initially, perhaps eliminating terms for the streamflow derivative or seasonal component. Additionally, systematic sampling at a fixed interval limits information available on the covariance structure, which is used in filtering and smoothing operations, to multiples of the fixed interval. Then, covariance information at shorter intervals must be projected from multiples of the sampling interval. For example, if systematic sampling occurs at 15-day intervals, the covariance structure is empirically defined only for 15, 30, 45-day intervals and so on, and must be extrapolated to 1-day or unit intervals for record computation. This type of extrapolation introduces additional uncertainty in the estimation process, and might be avoided by varying the sampling interval while maintaining the same total number of samples. Additional analyses are needed to more fully investigate the implications for sampling design.

SUMMARY

This paper develops optimal estimators for on-line and off-line computation of suspended-sediment concentrations in streams and compares the accuracies of the optimal estimators with results produced by time-averaging interpolators and flow-weighting regression estimators. The analysis uses long-term daily-mean suspended-sediment concentration and streamflow data from 10 sites within the United States to compare accuracies of the estimators. A log transformation was applied to both suspended-sediment concentration and

streamflow values prior to development of the estimates. Standard techniques are described for removing the possible bias in estimates of the mean computed form exponentiated values of sediment concentration computed in log units and for approximating instantaneous or unit dynamics on the basis of daily samples.

The optimal estimators are based on a Kalman filter and an associated smoother to produce the on-line and off-line estimates, respectively. The optimal estimators included site-specific parameters, which were estimated by generalized least squares, to account for influences associated with ancillary variables, including streamflow and annual seasonality, on suspended-sediment concentrations. In addition, the optimal estimators account for autoregressive error components and uncertainties in the accuracy of direct measurements in computing continuous records of suspended-sediment concentrations. Results were compared with estimates produced by both linear and cubic-spline interpolators, which do not account for ancillary variables, and with simple and multiple-regression estimators, which do not locally account for direct measurements of suspended-sediment concentration.

The average standard error of simple and multiple regression estimates was 124 and 105%, respectively. The accuracies of interpolators, and on-line and off-line estimators are related to measurement frequency, and were compared at simulated measurement intervals of 3, 6, 12, 24, and 48 days. The average standard error of the linear interpolator ranged from 48·4% for 3-day sampling intervals to 159% for 48-day sampling intervals. The average standard error of the cubic-spline interpolator ranged from 50·6% for 3-day sampling intervals to 176% for 48-day sampling intervals. The average standard error for the on-line estimator ranged from 52·7% for a 3-day sampling interval to 107% for a 48-day sampling interval. The average standard error for the off-line estimator ranged from 39·9% for a 3-day sampling interval to 93·0% for a 48-day sampling interval. Thus, the off-line estimator had the lowest standard error, especially at the shorter sampling intervals that are needed to compute continuous records of suspended-sediment concentrations.

The use of the optimal estimators rather than interpolators or regression estimators will improve the accuracy and quantify the uncertainty of records computed on the basis of suspended-sediment concentrations measured at intervals less than 48 days. Although in this paper, parameters for the estimators were developed on the basis of daily values data, it is anticipated that in typical applications the estimators will be developed on the basis of unit-value data and direct measurement information.

REFERENCES

Bukaveckas PA, Likens GE, Winter TC, Buso DC. 1998. A comparison of methods for deriving solute flux rates using long-term data from streams in the Mirror Lake watershed. *Water, Air, and Soil Pollution* **105**: 277–293.

Burkham DE. 1985. An approach for appraising the accuracy of suspended-sediment data. *US Geological Survey Professional Paper* **1333**: 18.

Ferguson RI. 1986. River loads underestimated by rating curves. *Water Resources Research* **22**(1): 74–76.

Gelb A (ed.). 1974. *Applied Optimal Estimation*. MIT Press: Cambridge, MA; 374.

Glysson GD. 1987. Sediment-transport curves. *US Geological Survey Open-File Report* **87–218**: 47.

Grewal MS, Andrews AP. 1993. *Kalman Filtering—Theory and Practice*. Prentice Hall Information and System Science Series: Englewood Cliffs, NJ; 38.

Guy HP. 1970. Fluvial sediment concepts. US Government Printing Office: Washington, DC; *US Geological Survey Techniques of Water-Resources Investigations*, Vol. 3, Chap. C1. 55.

Guy HP, Norman VW. 1973. Field methods for measurement of fluvial sediment. US Government Printing Office: Washington, DC; *US Geological Survey Techniques of Water-Resource Investigations*, Vol. 3, Chap. C2. 59.

Helsel DR, Hirsch RM. 1992. *Statistical Methods in Water Resources*. Studies in Environmental Science Vol. 49. Elsevier: New York; 522.

Koltun GF, Gray JR, McElhone TJ. 1994. User's manual for SEDCALC, a computer program for computation of suspended-sediment discharge. *US Geological Survey Open-File Report* **94–459**: 46.

Phillips JM, Webb BW, Walling DE, Leeks GJL. 1999. Estimating the suspended sediment loads of rivers in the LOIS study area using infrequent samples. *Hydrological Processes* **13**(7): 1035–1050.

Porterfield G. 1972. Computation of fluvial-sediment discharge. US Government Printing Office: Washington, DC; *US Geological Survey Techniques of Water-Resources Investigations*, Vol. 3, Chap. C3. 66.

Richards RR, Holloway J. 1987. Monte Carlo studies of sampling strategies for estimating tributary loads. *Water Resources Research* **23**(10): 1939–1948.

SAS Institute. 1988. *SAS/ETS User's Guide*, Vers. 6, 1st ed. SAS Institute: Cary NC; 560 pp.

Siouris GM. 1996. *Optimal Control and Estimation Theory*. Wiley: New York; 407.
Walling DE, Webb BW. 1988. The reliability of rating curve estimates of suspended sediment yield; some further comments. In *Sediment Budgets*, Bordas MP, Walling DE (eds). IAHS Publication No. 174; 350.
Yang CT. 1996. *Sediment Transport—Theory and Practice*. McGraw-Hill: New York; 396.