WATER RESOURCES RESEARCH, VOL. 36, NO. 7, PAGES 1823–1833, JULY 2000

Multisite disaggregation of monthly to daily streamflow

D. Nagesh Kumar

Department of Civil Engineering, Indian Institute of Technology, Kharagpur, India

Upmanu Lall[1]

Department of Civil and Environmental Engineering and Utah Water Research Laboratory
Utah State University, Logan

Michael R. Petersen

Keller-Bliesner Engineering, Logan, Utah

Abstract. Streamflow disaggregation is used to preserve statistical attributes of time series across multiple sites and timescales. Several algorithms for spatial disaggregation and for disaggregation of annual to monthly flows are available. However, the disaggregation of monthly to daily or weekly to daily flows remains a challenge. A new algorithm is presented for simultaneously disaggregating monthly flows at a number of sites and daily flows at an index site to daily flows at a number of sites on a drainage network. The continuity of flow in time across months at each site as well as the intersite flow pattern are preserved. The disaggregated daily flows at the multiple sites are conditioned on the spatial (across site) pattern of monthly flows at the respective sites. The probability distribution of the vector of disaggregated flows conditional on the multisite monthly flows is approximated nonparametrically using the k nearest neighbors of the monthly spatial flow pattern. A constrained optimization problem is solved to adaptively estimate the disaggregated flows in space and time for each such neighborhood. An application to data from a tributary of the Colorado River is used to illustrate the modeling process.

1. Introduction

Disaggregated streamflow sequences that are statistically similar to observed streamflow records are very useful for analyzing multireservoir operation policies and river basin management. There is renewed interest in disaggregation methods as climate-related issues (regional El Niño Southern Oscillation (ENSO) forecasts or downscaling of climate change scenarios) have come to the fore. The disaggregation models proposed by *Valencia and Schaake* [1972, 1973] have been used to divide annual flows into seasonal flows [*Mejia and Rousselle*, 1976; *Tao and Delleur*, 1976; *Srikanthan*, 1978; *Lane*, 1979; *Salas et al*., 1980] and to divide aggregate basin flows (monthly or annual) into flows at individual sites [*Loucks et al*., 1981; *Lane*, 1979, 1982; *Salas et al*., 1980]. *Mejia and Rousselle* [1976], *Lane* [1979], and *Stedinger and Vogel* [1984] further extended this model to reproduce the correlation between disaggregated flow volumes between subperiods (e.g., months) of different years. Other disaggregation models include models proposed by *Harms and Campbell* [1967], *Stedinger et al*. [1985], *Grygier and Stedinger* [1988], *Santos and Salas* [1992], *Bartolini and Salas* [1993], *Koutsoyiannis* [1992], and *Koutsoyiannis and Manetas* [1996] with various improvements. By and large, these approaches have focused on space or time disaggregation and on annual to seasonal or seasonal to subseasonal flows. Parametric assumptions of the probability distribution of the underlying streamflow are usually invoked. The disaggregated flows (monthly from annual sum or individual sites from index site) are obtained using the correlation structure of the respective time or space flow. Exceptions are the works of *Lall et al*. [1996] and *Tarboton et al*. [1998]. They proposed a nonparametric approach for space or time disaggregation based on kernel density estimation.

A number of factors complicate the development of operational monthly to daily and space-time disaggregation schemes for streamflow. A primary difficulty is the rapid increase in the dimensionality of the parameter space relative to the finite amount of data available. Staged disaggregation procedures (e.g., monthly to weekly to daily) are sometimes advocated to address this problem. However, it can be difficult to maintain continuity of flows across subperiods in such schemes. Another problem is that the flow dynamics can be quite nonlinear at the finer timescales. Thus state-dependent models that allow the disaggregated sequence to adapt to the flow conditions may be needed. The time and/or space correlation structure used for disaggregation by traditional methods may actually change by flow condition (e.g., extreme wet or dry or average conditions) in this setting. The work presented here seeks to overcome these difficulties by putting the space-time disaggregation problem in a rather different context than the statistical estimation procedures used thus far. Disaggregation of monthly flows at upstream sites and daily flows at an index gage to daily flows at upstream sites is sought as the solution of an optimization problem for each month. The "best" values of the space-time components are sought subject to flow continuity across

[1]On sabbatical at Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York.

Copyright 2000 by the American Geophysical Union.

Paper number 2000WR900049.
0043-1397/00/2000WR900049$09.00

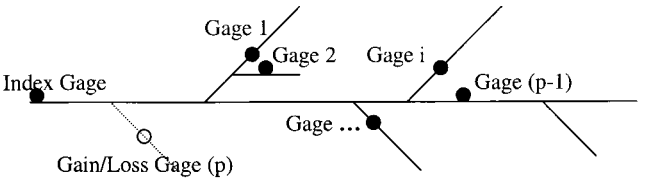

Figure 1. Generic layout of stream gages on a river basin for the disaggregation problem. Monthly flows are available at the $(p - 1)$ upstream gages and for the hypothetical gain/loss gage. Daily and monthly flows are available at the index gage. The disaggregation problem is to solve for the daily flows at each of the p upstream gages for each month. Historical daily and monthly flow data at all p gages and the index gage provide the basis for the disaggregation.

days and sites (i.e., no unreasonable jumps or drops in flow), proper summability in space and time, and other constraints that serve to regularize the solution. The disaggregated daily flows are conditioned on the spatial pattern of monthly flows across sites. For a given month they are selected as the optimal (closest in a weighted L_1 norm to the historical daily flows in the conditioning set) conditional expectation of the daily flow vector for each site, conditional on the monthly spatial flow pattern and the daily flow pattern at the index gage. An empirical, nonparametric estimate of the multivariate conditional density function based on k nearest neighbors in state space is used to determine the subset of historical monthly flow data used for conditioning the estimate.

The precise problem solved is defined in section 2. The solution approach and the implementation of the algorithm are discussed in section 3. An application to data from Colorado that motivated the development of the algorithm is presented in section 4.

2. Problem Statement

The setting for the problem considered is illustrated in Figure 1. An "index" gage is located on the main stem of the river system. Monthly and daily flows are available at this gage. A number $(p - 1)$ of stream gages are located on stream reaches influent into the index gage. A hypothetical gain/loss gage representing changes in total flow from the $(p - 1)$ upstream gages to the index gage is also considered. Historical daily and monthly flow data at all sites are presumed to be available for MK years (the subset of months for which daily and monthly flows are available at all sites is denoted MK). These data are used to estimate the statistical relationships for disaggregation. A second data set of length MD years (the subset of months for which only monthly flows are available at the p sites to be disaggregated and monthly and daily flows are available at the index site is denoted MD) is also available. These data correspond to the period for which disaggregation of monthly to daily flows is needed. Monthly flow data are presumed to be available for each of the p upstream sites and for the downstream index site. In addition, the daily flows for each month in the MD year period are also available at the index site. The disaggregation problem considered in this paper entails the temporal disaggregation of monthly to daily flows at the p sites (including the gain/loss gage) and the spatial disaggregation of the index site daily flows to the upstream sites. The generated daily flows should sum to the monthly flows at each site, and the daily flows across sites for a given day should sum to the daily flow at the downstream index site. The gain/loss site conceptually accounts for timing issues as well as ungaged tributaries, diversions, or return flows in this setting. Also, we need to preserve continuity of flow in time across months and across sites.

Let the monthly flow at the index station in month m be denoted as Q_m and the daily flow on day j in month m as q_{jm}. For the upstream stations to be disaggregated, denote the monthly flow in month m at site i as X_{mi} and the daily flow on day j in month m at site i as x_{jmi}. Daily flow on day j in month m at site i as a proportion of the index site is given by

$$p_{jmi} = x_{jmi}/q_{jm}. \tag{1}$$

We need to generate daily flows x_{jmi} for each site i, for every month m that belongs to the set MD. For each month m, we need to solve for $(nd_m{}^*p)$ daily flows across all the sites where nd_m is the number of days in month m. For 30 days and 10 sites this leads to 300 unknowns to solve for each month. The disaggregation problem is solved one month at a time. Let us denote the set of disaggregated flows $(x_{jmi}, j = 1 \ldots nd_m; i = 1 \ldots p)$ for month m as the vector $\mathbf{x}_m$. The disaggregation problem may then be addressed through the estimation of the conditional probability density function $f(\mathbf{x}_m|\mathbf{y}_m)$ or the conditional expectation $E[\mathbf{x}_m|\mathbf{y}_m]$, where the vector $\mathbf{y}_m$ includes the monthly flows X_{mi}, $i = 1 \ldots p$ at the upstream sites and the daily $(q_{jm}, j = 1 \ldots nd_m)$ flows at the index site. For 30 days and 10 sites this leads to 40 conditioning variables in each month. Clearly, this translates into a formidable stochastic estimation problem if it is approached in the classical framework of multivariate density estimation. The estimation of such a density function from the limited $\mathbf{x}_m$ and $\mathbf{y}_m$ data in the MK years of common record is unlikely to be successful even under strong parametric assumptions. Consequently, an empirical optimization strategy that significantly constrains the estimation problem is used to develop useful estimates of $\mathbf{x}_m$. In a parametric framework the problem could be handled by a condensed, staged disaggregation procedure. The main problem with parametric models is the inability to produce realistic sequences of daily flows. That is, they can be designed to preserve autocorrelation and other basic moments, but they do not look like "real" daily streamflows. This is the main strongpoint of the nonparametric approach.

First, we consider the summability conditions for each month m in the disaggregation period MD. We have p conditions (equation (2)) requiring the daily flows over month m to sum to the recorded flow at each of the p sites and nd_m conditions (equation (3)) requiring the daily flows at all the p sites for any day to sum to the daily flow at the downstream sites. This results in $(p + nd_m)$ equations to solve for (p^*nd_m) unknowns. Usually, the number of unknowns will far exceed the number of equations (e.g., 300 versus 40), and the problem as stated is not well posed since many combinations of values for the x_{jmi} will satisfy these equations. Further, these equations will need to be solved anew for every combination of flow values for any given month:

$$\sum_{j=1}^{nd_m} x_{jmi} = X_{mi} \qquad i = 1, \ldots, p \tag{2}$$

$$\sum_{i=1}^{p} x_{jmi} = q_{jm} \qquad j = 1, \ldots, nd_m. \tag{3}$$

Equation (3) can also be written as

$$\sum_{i=1}^{p} p_{jmi} = 1 \qquad j = 1, \ldots, nd_m. \tag{3'}$$

An optimization algorithm that seeks a solution to these summability equations in the disaggregation period MD and simultaneously yields optimal "prediction" for the $\mathbf{x}_m$ with reference to "similar" months in the period MK is described in Section 3. Here similar is defined in terms of closeness of the $\mathbf{y}_m$ for the month to be disaggregated in terms of the L_1, L_2, Mahalanobis, or other distance metric to values of $\mathbf{y}_m$ for the same calendar month in the MK period.

3. Multisite Disaggregation Algorithm

The algorithm (see Figure 2 for an overview) seeks to generate daily flows for a specific month m^* in the period MD. The following are the key steps:

1. For current month m^*, identify calendar month m_c.
2. Define season window M_c for month m_c (e.g., 1 or 3 months centered about m_c). Only flow vectors from this season window in the past MK years of record are considered as representative of the current conditions. A seasonal window is used to ensure the selection of an appropriate seasonal pattern of monthly flows.
3. Identify spatial monthly flow patterns in the historical MK record that are similar to the current monthly flow pattern $\mathbf{z}_m$ defined as $(Q_{m^*}, X_{m^*1}, i = 1 \ldots p)$. Find the K nearest neighbors $\mathbf{z}_{m^*k}$, $k = 1 \ldots K$ of $\mathbf{z}_{m^*}$ in the vectors $\mathbf{z}_m$, $m \in M_c \in MK$ in the historical period of MK years. The nearest neighbors are identified on the basis of a Euclidean or other distance metric applied to $\mathbf{z}_{m^*}$ and $\mathbf{z}_m$, $m \in M_c \in MK$. *Lall and Sharma* [1996] and *Rajagopalan and Lall* [1999] present time series resampling approaches using multivariate K nearest-neighbor density estimation approaches. They recommend a choice K equal to the $\sqrt{n}$, where n is the sample size, to be effective as a rule of thumb. These K neighbors specify a conditioning slice of the multivariate density of $\mathbf{z}_m$ defined in a neighborhood of $\mathbf{z}_{m^*}$. The K neighbors are associated with monthly indices in the historical (MK) data set. For example, if m_c is July, the seasonal window is 3 months (June, July, and August), the MK period is from 1901 to 1980, and K is 5. We will identify the 5 seasons out of the 80 seasons in the seasonal window with spatial monthly flow patterns that are the five closest in some distance norm to that of the month to be disaggregated. It is assumed that the number of sites is generally smaller than the number of days in a month; hence using the spatial monthly flow pattern to select the best matching historical period will be more effective given a finite data set. After the neighbors are selected, the subsequent computations are performed with the same calendar month as m_c selected from the seasonal window for each of the k neighbors.
4. Define an optimization problem to solve for the daily flow proportions $\mathbf{p}_{m^*}$ to minimize total weighted daily flow prediction error across all p sites for each of the K nearest-neighbor months in the MK year historical period, while satisfying summability and continuity constraints for the current month m^* (which is part of the period MD). This seeks to determine an optimal set of values of the $\mathbf{p}_{m^*}$, conditional on the current monthly flow pattern $\mathbf{z}_{m^*}$. In other words, the suitability of a certain vector of disaggregation proportions $\mathbf{p}_{m^*}$ to be used with data for the current month is evaluated through a predictive exercise over the K most similar months in the period where all daily and monthly flows were available. The weights applied to the error in predicting flow for each day (j) at each site (i) for a neighbor month (k) are based on a measure of similarity of the monthly flows at site i, and the daily flow at the index site for the month m^*, and the historical month corresponding to k.

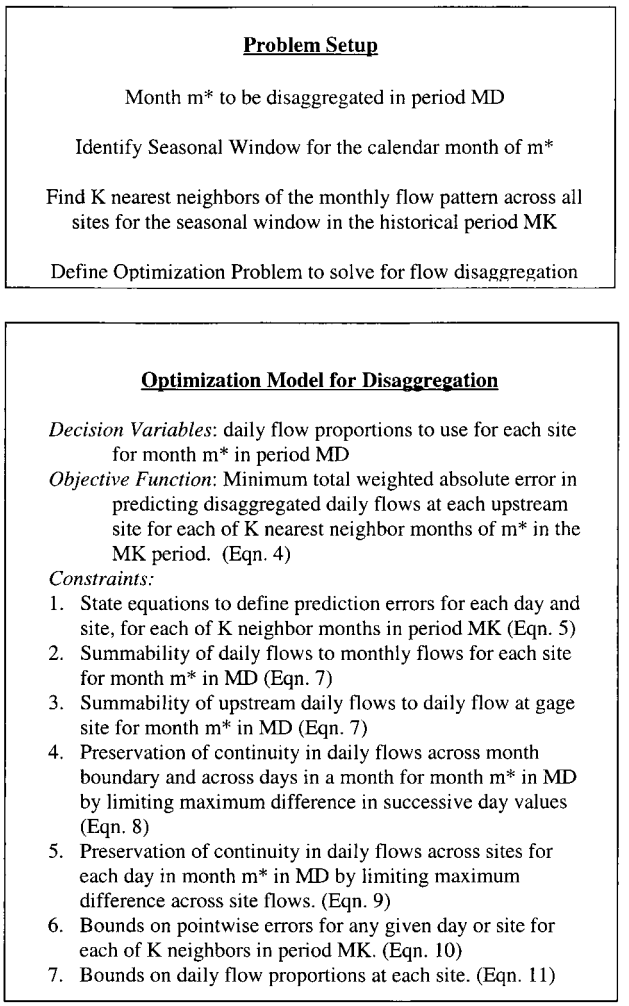

Figure 2. Schematic of disaggregation algorithm.

The linear optimization problem solved for disaggregating the monthly flow for a month is now formally presented. The objective function is defined using a weighted L_1 norm as

$$\min \sum_{k=1}^{K} \sum_{j=1}^{nd_m} \sum_{i=1}^{p} w_{jki} |x_{jki} - p_{jm^*i} q_{jk}| \tag{4}$$

where the weight $w_{jki} = 1/d_{jki}$, with $d_{jki} = [(q_{jm^*} - q_{jk})^2 + (X_{m^*i} - X_{ki})^2]^{1/2}$.

This objective function can be rewritten as

$$\min \sum_{k=1}^{K} \sum_{j=1}^{nd_m} \sum_{i=1}^{p} w_{jki}(u_{jki} + v_{jki}), \tag{4'}$$

where $(u_{jki} - v_{jki})$ is the error in the prediction of the observed daily flow x_{jki} for site i on day j in neighbor month k in

the historical data set MK. Here the error is defined as the difference of two positive variables, u_{jki} and v_{jki}. Consequently, the term $(u_{jki} + v_{jki})$ in the objective function translates into an absolute error in the linear programming framework. The errors are defined in terms of the historical daily flow data at upstream sites (x_{jki}) and index site (q_{jk}) and the candidate value of the daily flow disaggregation proportion p_{jm^*i} through

$$x_{jki} - p_{jm^*i}q_{jk} + u_{jki} - v_{jki} = 0 \tag{5}$$

$$j = 1 \ldots nd_m;\ k = 1 \ldots K;\ i = 1 \ldots p,\ m \in M_c \in MK$$

$$u_{jki} \geq 0; \quad v_{jki} \geq 0, \tag{6}$$

$$j = 1 \ldots nd_m; \quad i = 1 \ldots p; \quad k = 1 \ldots K.$$

Note that the weights for each day, at each site, and each neighbor month k are defined through a distance for the pattern defined by the daily flow at the index gage and the monthly flow for the ith site for the month m^* being disaggregated and the kth historical neighbor of that month. The weights in (4) recognize how similar each day's flow for neighbor month k is to each day's flow in month m^* at the index site and also how similar the monthly flow at site i is in neighbor month k and month m^*. Using this weighting scheme we try to match the current month's spatial flow pattern across sites as well as the daily flow pattern at the index site. Given the likely magnitude differences in the flows across the sites and the daily versus monthly flows, it is a good idea to compute the distance so that each variable has first been scaled by its mean (or otherwise standardized) for the purpose of computing the distance. Other distance metrics can also be used.

The minimization in (4) is done with respect to the proportions p_{jm^*i}. The linear programming problem we solved includes the error variables u_{jki} and v_{jki}. We must recall that the daily flow predictions represented by the terms $p_{jm^*i}q_{jk}$ are being computed for each of the K neighbor months in the historical record MK and are being compared with the actual observed daily flows x_{jki} at each site for each month, as shown in (5). This process is similar to a locally weighted regression or loess [*Cleveland and Devlin*, 1988], where k nearest neighbors of the current prediction point are selected and a minimum weighted least squares solution to a linear or quadratic regression problem is sought in this neighborhood, with each data point in the neighborhood weighted inversely proportional to its distance from the prediction point in the predictor space. Weighted square error or other error norms, instead of the weighted absolute error, could be used in (5). However, a nonlinear optimization scheme would then be necessary.

The algorithm presented here also considers the specification of a number of constraints to regularize the local regression solution. These include (5) and (5′) that apply to the historical period MK. Additional constraints that may be specified are enumerated below.

The summability constraints applied to month m^* flows in period MD across time and space are represented as

$$\sum_{j=1}^{nd_m} p_{jm^*i}q_{jm^*} = X_{m^*i} \qquad i = 1, \ldots, p$$

$$\sum_{i=1}^{p} p_{jm^*i} = 1 \qquad j = 1, \ldots, nd_{m^*}. \tag{7}$$

One can also introduce constraints to ensure "continuity" of flow from one day to the next in month m^* in period MD. This can be done by requiring that the first day's flow be within some range of the previous day's flow at each site. This is stated as

$$xcl_{m^*i} \leq (p_{(j+1)m^*i} - p_{jm^*i})q_{jm^*} \leq xcu_{m^*i} \tag{8}$$

$$i = 1, \ldots, p, \quad j = nd_{(m^*-1)}, 1, \ldots, (nd_{m^*} - 1),$$

where xcl_{m^*i} and xcu_{m^*i} are user specified lower and upper limits for interday flow differences for month m^* at site i. The first index of j in (8) specifies continuity from the last day of the previous month to the 1st day of the current month m^*.

Similar constraints can also be applied to maintain intersite flow continuity. For instance, we could look at pairs of sites (e.g., i and o) and recognize that the difference in the flow proportions for any day in the calendar month m_c between those pairs of sites lies in a certain range historically (for the K neighbors) and they restrict this range as

$$pcl_{m^*i,o} \leq p_{jm^*i} - p_{jm^*o} \leq pcu_{m^*i,o} \tag{9}$$

$$i \neq o,\ i = 1, \ldots, p,\ o = 1, \ldots, p,\ j = 1, \ldots, nd_{m^*}.$$

In addition to a minimization of the global error, we can also require the solution to be "well behaved" in terms of pointwise approximation error by requiring that the percent error in each prediction in the period MK be limited to some number:

$$u_{jki}/x_{jki} \leq E \quad v_{jki}/x_{jki} \leq E \tag{10}$$

$$j = 1 \ldots nd_{m^*} \quad k = 1 \ldots K \quad i = 1 \ldots p,$$

where E is a user specified permissible fractional error in each prediction.

The daily flow proportions for the $(p - 1)$ sites, excluding the gain/loss site, are restricted to lie between 0 and 1. The proportion for the gain/loss site is not restricted:

$$0 \leq p_{jm^*i} \leq 1 \qquad j = 1 \ldots nd_{m^*}, \quad i = 1 \ldots p - 1. \tag{11}$$

The allowable range of the p_{jm^*i} is further restricted by examining the maximum range of these proportions in the period MK for the seasonal window M_c. Suppose that the range of the daily flow for month m_c at site i as a proportion of the index gage ranges lies between pl_{m^*i} and pu_{m^*i} in the historical set MK. It may then be reasonable to replace (11) by the following constraint:

$$pl_{m^*i} \leq p_{jm^*i} \leq pu_{m^*i} \quad j = 1 \ldots nd_{m^*}, \quad i = 1 \ldots p. \tag{12}$$

Summarizing, we find the best performing solution for the disaggregation problem by checking performance in months that were historically the most similar to the current month, while restricting the range of solutions using summability requirements and also a variety of other conditions applied to the month to be disaggregated that make sense operationally. A number of the constraints (equations (8)–(12)) may not be active in a given optimal solution. They largely serve to regularize the solution. Since the size of the problem and the LP computation time increase with the number of constraints, a pragmatic strategy may be to first attempt a solution without some of these constraints and then to add them if the pointwise approximation is indicated to be poor.

The algorithm has been implemented using linear programming in Language for interactive general optimization (LINGO) as a general procedure for the monthly to daily

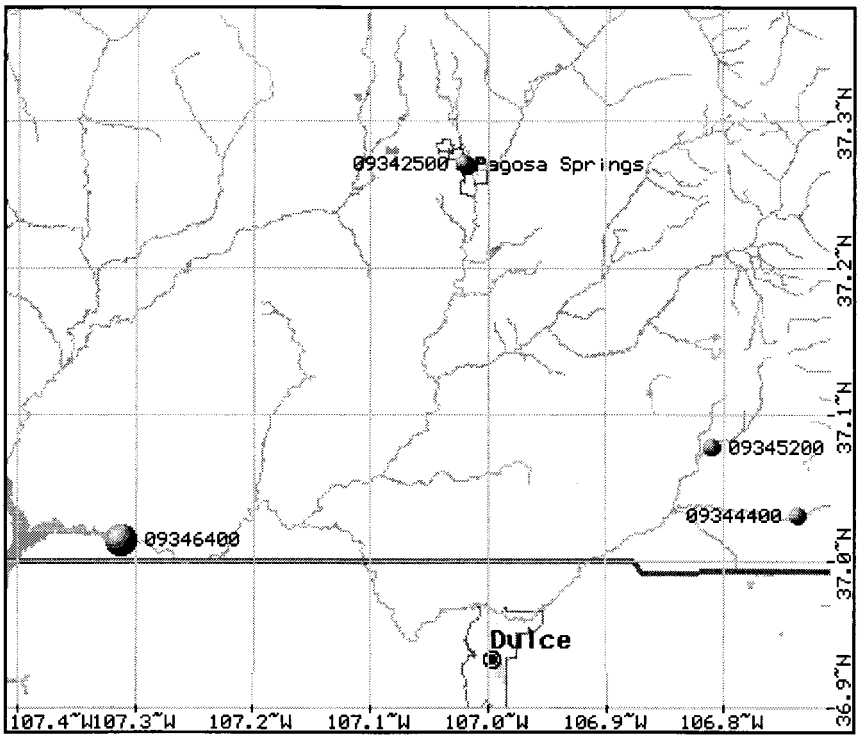

Figure 3. River network for San Juan River, Colorado, application. The index gage is 09346400 at the San Juan River at Caraccas. The three stations on tributaries upstream of this gage are indicated by the shaded circles and the station identification numbers.

disaggregation for multiple sites. A typical optimization problem that has been solved will have $nd_{m*}p(1 + 2K)$ decision variables (including the error terms) and up to $\{nd_{m*}(1 + p + Kp + p(p - 1)/2) + p\}$ constraints (depending on which ones are actually used). Five sites, a 30-day month, and five nearest-neighbor months translate to 1650 decision variables and up to 1235 constraints. Staged spatial disaggregation may be desirable to control problem size where a large number of sites need to be processed. The selection of parameters (e.g., the number of neighbors K) exogenous to the optimization process and some issues in implementation are illustrated through an example application that motivated the development of the algorithm presented here.

4. Model Application

The algorithm presented in section 3 has been tested with a number of synthetic (i.e., generated from known multivariate parametric probability models) and real data sets. Given the nature of the model presented, its attributes are best exhibited through an application to a real data set that motivated our formulation. The San Juan River originates in the San Juan Mountains of southern Colorado. The river flows southwest into New Mexico, through Utah, and ultimately into Lake Powell. In the extreme upper basin the first major tributary to the San Juan River is the Navajo River. Streamflow gaging stations on the main stem of the San Juan River lie above and below the mouth of the Navajo River at Pagosa Springs and near Carracus, Colorado, respectively. The U.S. Bureau of Reclamation required the extension and disaggregation of the natural flow series of the major rivers used in the San Juan River Basin Recovery Implementation Program. The process of developing a recovery program requires long-term daily natural flows. For estimating the flows the study needed that (1) the sequence of monthly natural flows be extended over a longer time period and (2) the monthly natural flows be disaggregated into daily flows using the historic records at nearby gaging stations. A river system operational model then simulated the net flows (natural flows minus agricultural, municipal, recreational demands etc.) required for recovery of the threatened and endangered fish species in the San Juan River. Monthly naturalized streamflow records were available for a number of gages on a tributary basin to the Colorado River. However, daily flows were available at these gages only for a subset of the record. Daily flows at an index site were available for the full record. The interest was in using the data at the index site to develop daily flow records for the full period at all gages upstream of the index site. The network of daily flows was to be used to aid subsequent analyses of daily streamflow variations in different sections of the river as part of an environmental and water resource management project for the river basin. The river network is illustrated in Figure 3.

The streamflow gage on the San Juan River near Carracas, Colorado (station 09346400), was used as index station. Data from three upstream sites (1) station 09342500, San Juan River at Pagosa Springs; (2) station 09344400, Navajo River below the Oso diversion dam near Chromo, Colorado; and (3) station 09345200, Little Navajo River below the Little Oso diversion dam near Chromo, Colorado, were used. A gain/loss site was added as the fourth site. It may be observed from Figure 3 that some of the tributaries are not gaged. In addition, there is an ungaged diversion to another basin. Consequently, the mean flow at the gain/loss site was different from zero. The sum of

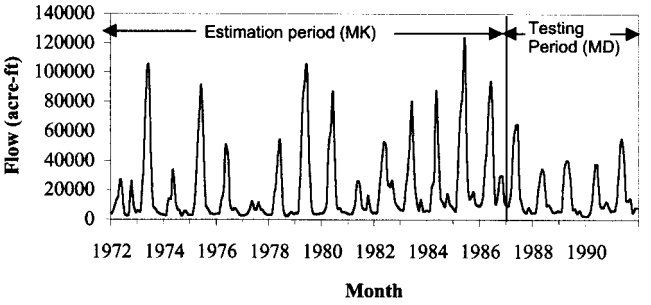

Figure 4. Monthly streamflow time series for the index gage 09346400. The model estimation period (MK) is 1972–1986. The fitted parameters are used for predictions of disaggregated daily flows in the 1987–1991 period (1 acre-foot = 1234 m^3).

flows at these four sites is presumed to sum to the flow at the index site for any day.

Estimated daily natural flows (acre-feet, 1 acre-foot = 1234 m^3) at all these sites were available for a 20-year period from 1972 to 1991. The data for the first 15 years were used for estimation, and the last 5 years of data were used for model testing. A time series plot for the monthly flows at the index site is shown in Figure 4. A 3-month window was considered to choose "good" analogs of the seasonal flow pattern and to help ensure flow continuity across month boundaries. The autocorrelation functions of daily data for a wet season (May–July) were analyzed separately for a wet (1979), an average (1991), and a dry (1977) year to assess state dependence of the serial correlation structure. The time series and the corresponding acre-feet are shown in Figure 5. Interestingly, the daily flows exhibit much more persistence in an average year than in a wet or dry year. Comparable results from an analysis of dry season (January–May) flows are presented in Figure 6. Interestingly, the flow persistence for this season is weakest in the dry year. Similar behavior was noticed for the daily data at other sites. Cross correlation among monthly flows at different sites and cross correlation between daily flows (during the estimation period) are reported in Table 1. The sites are strongly correlated with each other but not with the loss/gain site.

As already explained, for a given month a 3-month window is chosen (centering around the current month in the testing period), and the K nearest neighbors in the estimation period are selected based on Euclidean distance between the spatial monthly flow patterns for the corresponding season. Consider the disaggregation of the monthly flows for June 1991, a month in the 5-year model testing period (MD). The 3-month season considered for selecting monthly flow patterns is May, June, and July. Now, including the index site, we have five values for each month's flows for each of the 3 months in the window.

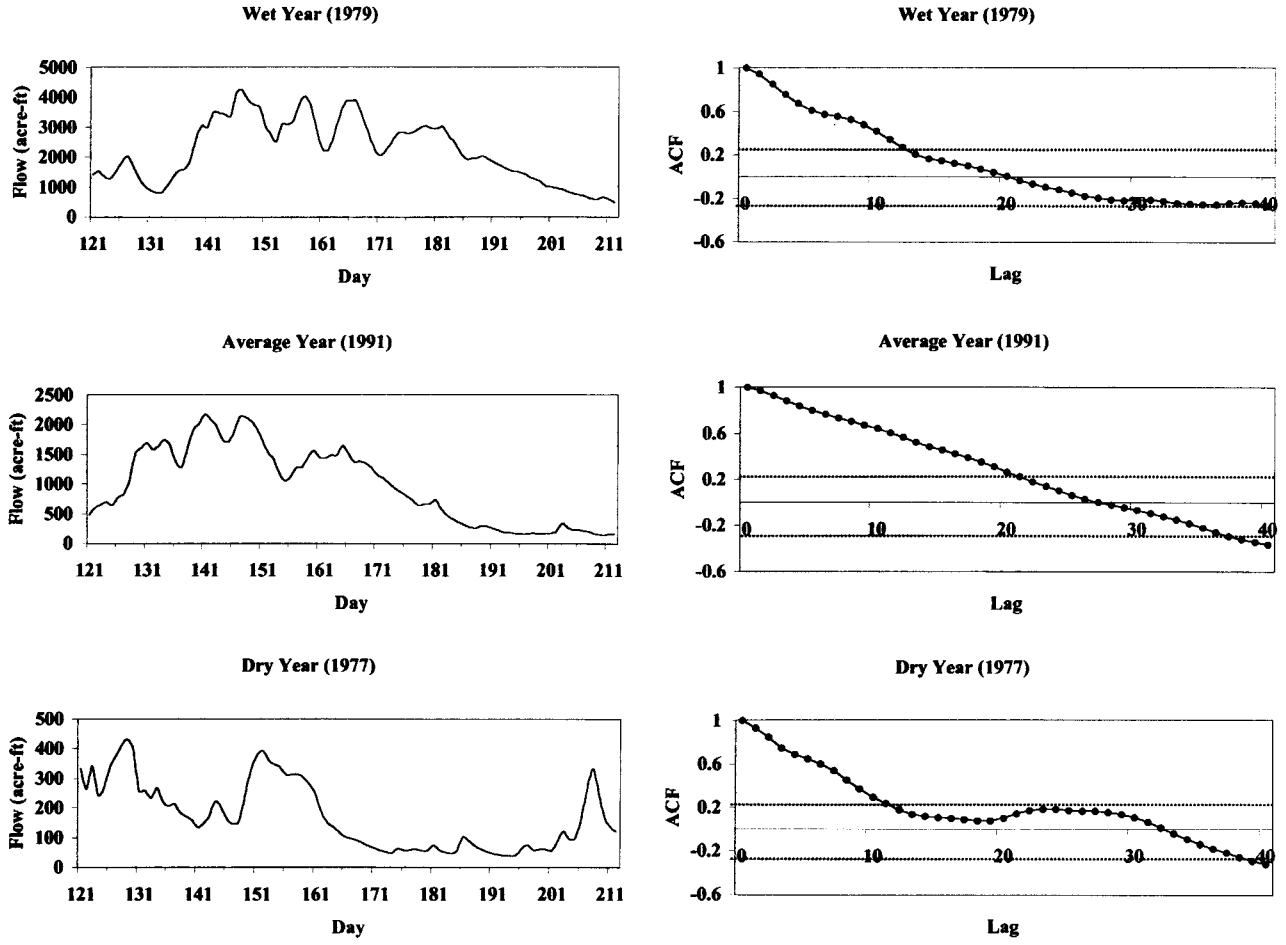

Figure 5. Time series and daily autocorrelation function for wet season (May–July) flows at gage 09342500 segregated by wet, average, and dry years. The state definition was based on water year volume. Note that daily flows are much more persistent in average than in wet or dry years for the wet season at this site.

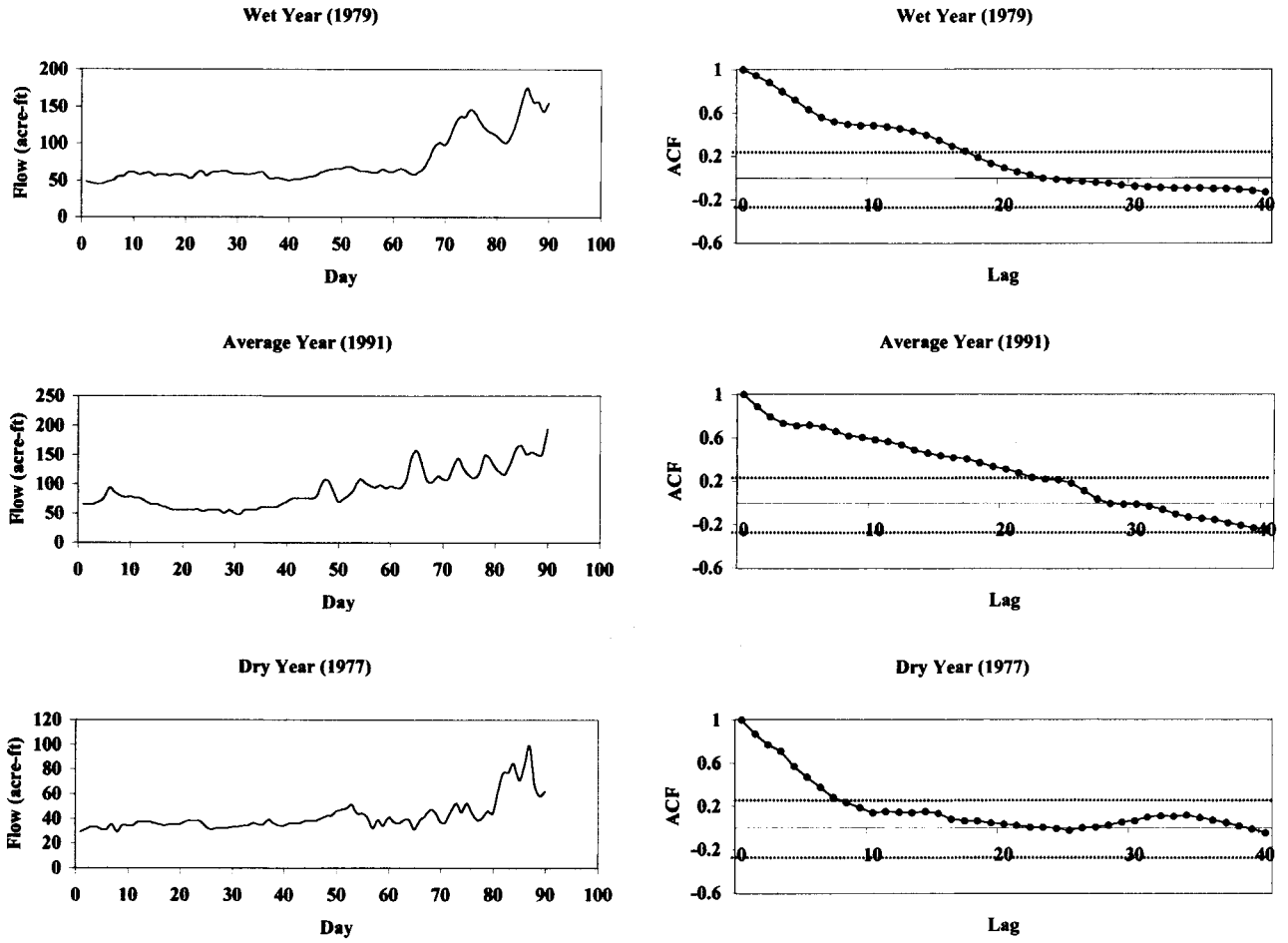

Figure 6. Time series and daily autocorrelation function for dry season (January–March) flows at gage 09342500 segregated by wet, average, and dry years. The state definition was based on water year volume. Note that daily flows are much more persistent in average than in wet or dry years for the wet season at this site. The serial dependence of daily flows is generally longer for the dry season than for the wet season shown in Figure 5. Interestingly, the dependence is much smaller in the dry year.

These 15 numbers are compared with the corresponding 15 flow values (same season, same site) for each of the years 1972 to 1986 in the model parameter estimation set MK. The K (e.g., 5) nearest neighbors of June 1991 in the historical data set are then selected as the years that are closest in terms of this 15 component distance metric to the May–July 1991 values. For this data set, 1984, 1981, 1976, 1974, and 1980 were the years for the five nearest neighbors of June 1991. The

Table 1. Cross Correlation of Monthly and Daily Data Between Different Sites for 1972–1987 Model Estimation Period

	Index Site 09346400	Gage 09342500	Gage 09344400	Gage 09345200	Gain/Loss Site
		Monthly Data			
Index site 09346400	1.00	0.96	0.93	0.76	0.31
Gage 09342500		1.00	0.98	0.74	0.03
Gage 09344400			1.00	0.71	−0.05
Gage 09345200				1.00	0.18
Gain/loss site					1.00
		Daily Data			
Index site 09346400	1.00	0.95	0.91	0.66	0.33
Gage 09342500		1.00	0.98	0.64	0.00
Gage 09344400			1.00	0.62	−0.07
Gage 09345200				1.00	0.14
Gain/loss site					1.00

Correlations >0.15 for monthly flows and >0.03 for daily flows are statistically significant at the 95% level.

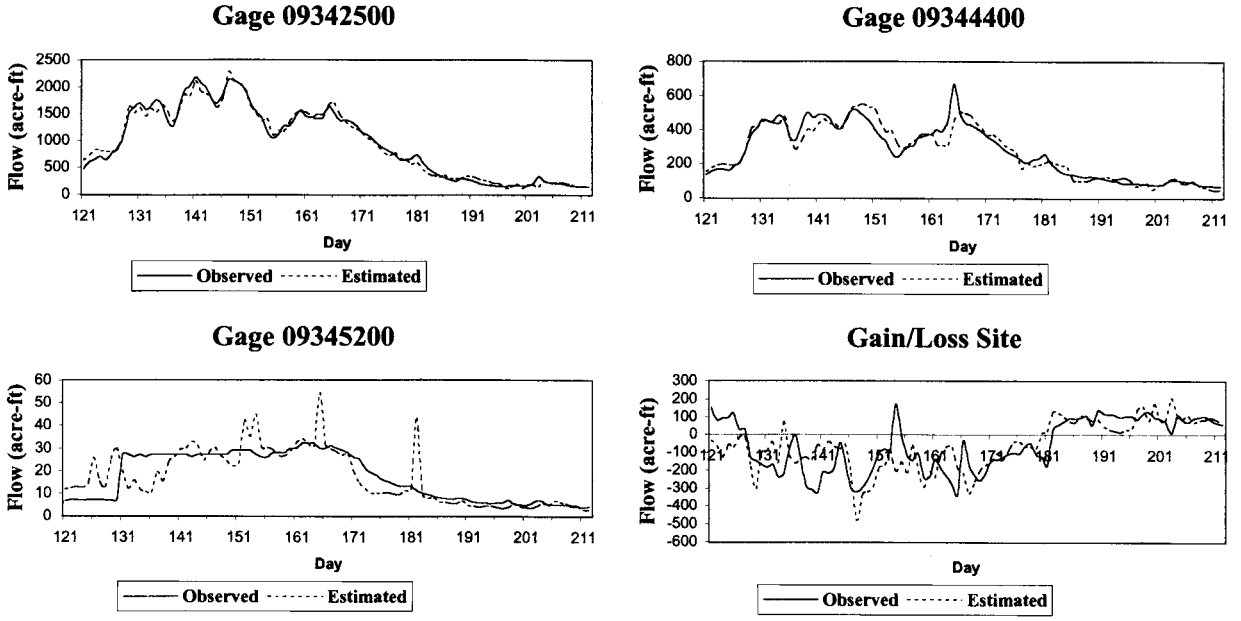

Figure 7. Estimated and observed flows for a wet season (May–July, 1991) for the four sites. Note the effects of apparent flow regulation at gage 09345200 starting about day 130 and the reproduction of the seasonal observed flow trends at all sites.

optimization problem is then defined in terms of the performance of the daily disaggregation proportions for the June 1991 flows in predicting the actual daily flows at each site for June of each of these 5 years. Recall that a weighted prediction error criterion is used for this purpose. The weights to be used for the prediction error for each day's flow at each site for each of the five past Junes selected are calculated next. The weight for a given day's prediction at a given site is calculated as the inverse of the Euclidean distance for a scaled daily flow for that day at the index site in June 1991 and June of the year of the kth neighbor and the monthly flow at the same site for June 1991 and the June for the kth neighbor.

For the results reported here only the constraints given in (5), (7), (9), and (12) were used. The time continuity constraints (equation (8)) and the pointwise error constraints (equation (10)) were not specified for the results reported here. We were interested in seeing how the algorithm would perform without these additional restraints. The solutions were within the bounds that we would have considered prescribing a priori. The computational burden of a linear programming problem is proportional to the square of the number of constraints specified. An adaptive approach where constraints are added if needed is consequently attractive. For a 30-day month with K taken to be five, 1320 (30 days $\times$ 4 sites for p and 5 neighbors $\times$ 4 sites $\times$ 30 days each for u and v) decision variables and 1594 (600 for equation (5), 34 for equation (7), 720 for equation (9), and 240 for equation (12) constraints were specified. The number of simplex iterations required for a 30-day month were approximately 1200.

The estimated and observed daily flows for a wet (May–July) and a dry (January–March) season for 1991 for all four sites are shown in Figures 7 and 8, respectively. The general seasonal trends in the observed flows are reproduced quite well in both cases. Fairly significant differences in the estimated and observed flows are apparent for a few days, particularly during the late March snowmelt period where the large contribution of the ungaged gain/loss site dominates the calculations. Interesting differences for gage 09345200 are also evident for the wet season, where the observed flows appear to show evidence of flow regulation that is captured somewhat differently by the disaggregated flows.

The performance of the algorithm over all 5 years of the testing period (1988–1992) assessed through the correlation between the observed and estimated daily flows for the 5-year period is presented in Table 2. A perusal of the diagonal elements in parentheses in Table 2 reveals that the daily flows estimated at each of the four sites correlate very strongly with those observed in this 5-year period. These correlations are much stronger than the raw cross-site correlations for observed daily flows for the same period, especially for the gain/loss site. Recall that only the monthly flows at each site and the daily flows at the index site are used from the 5-year period. The performance of the selected daily proportions is assessed using daily and monthly flow data from the five nearest neighbors of each month in the prior 15-year period. The cross correlations across sites for the 5-year period for the observed daily flows, for the estimated daily flows, and across the observed and estimated daily flows are also presented in Table 2. The cross correlations of the estimated flows are consistent with those for the observed flows.

Autocorrelation values of the estimated daily flows are also shown in Figure 9 for both the wet and dry seasons of 1991. They compare very well with the autocorrelation function plot of the observed data for the corresponding periods in Figure 5 and Figure 6. Recall that the serial correlation structure was not explicitly built into the disaggregation algorithm. Hence a satisfactory reproduction of these statistics and the high correlations between the estimated daily flows and the observed daily flows in the 5-year period reserved for algorithm validation provide an indication of the success of the algorithm.

The choice of the number of nearest neighbors to use reflects a bias-variance trade-off. As K increases, a much wider

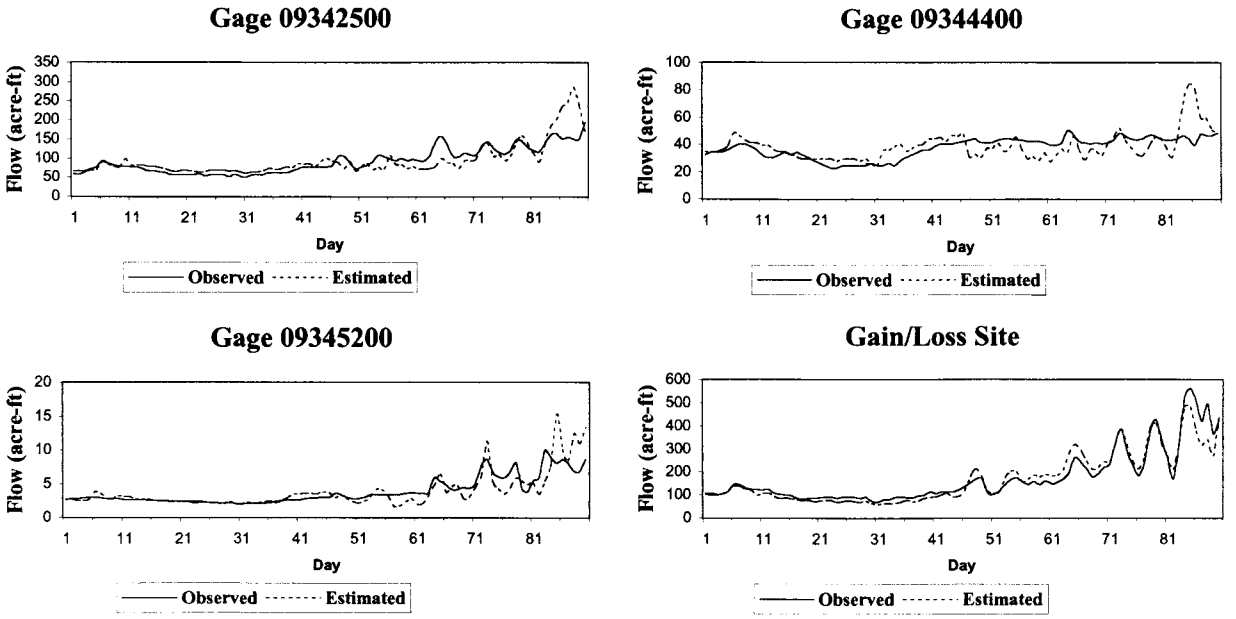

Figure 8. Estimated and observed flows for a dry season (January–March, 1991) for the four sites. Note reproduction of the seasonal observed flow trends at all sites and the compensating errors across the sites toward the end of the season. The gain/loss site has the majority of the flow in this period and hence dominates the error. Recall that it is weakly correlated with the other sites.

region of the state space of $\mathbf{y}_m$ is admitted into the estimation process. If the conditional probability distribution $f(\mathbf{x}_m|\mathbf{y}_m)$ is highly heterogeneous local estimates of this distribution will likely be biased. However, the increase in the sample size for estimation by increasing K while maintaining the same number of parameters (daily flow proportions) can translate into a reduced variance of estimate of the flow proportions. Since the number of constraints increases as K increases, the computational burden of the algorithm increases rapidly as K increases.

Table 2. Cross Correlations Among Observed and Estimated Flows for the Testing Period 1988–1992

	Gage 09342500	Gage 09344400	Gage 09345200	Gain/Loss Site
Correlations of Observed Daily Flows				
Gage 09342500	1.00	0.98	0.72	−0.25
Gage 09344400		1.00	0.69	−0.30
Gage 09345200			1.00	0.11
Gain/loss site				1.00
Correlations of Estimated Daily Flows				
Gage 09342500	1.00	0.95	0.69	−0.24
Gage 09344400		1.00	0.66	−0.25
Gage 09345200			1.00	0.09
Gain/loss site				1.00
Correlation Between Observed and Estimated Daily Flows				
Gage 09342500	(0.98)	0.95	0.69	−0.18
Gage 09344400	0.96	(0.96)	0.65	−0.22
Gage 09345200	0.70	0.68	(0.80)	0.16
Gain/loss site	−0.21	−0.25	0.10	(0.84)

Correlations >0.05 are statistically significant at the 95% level. Note that the cross-correlation structure of the observed and the estimated flows for this period is quite similar. Moreover, the correlations between the estimated and the observed flows (entries in parentheses) are very high, demonstrating that the daily flows predicted by the disaggregation model over the 5-year period that was not included in parameter estimation are very similar to those observed.

Recall that a different optimization problem is solved for each month to be disaggregated. A rule of thumb of $K = \sqrt{n}$ was proposed by *Lall and Sharma* [1996] in line with suggestions in the statistical literature for density estimation and classification. The efficacy of this rule was evaluated by varying K from 1 to 10 (the maximum sample size is 15 for any month corresponding to the 15 years of data used for estimation) with the

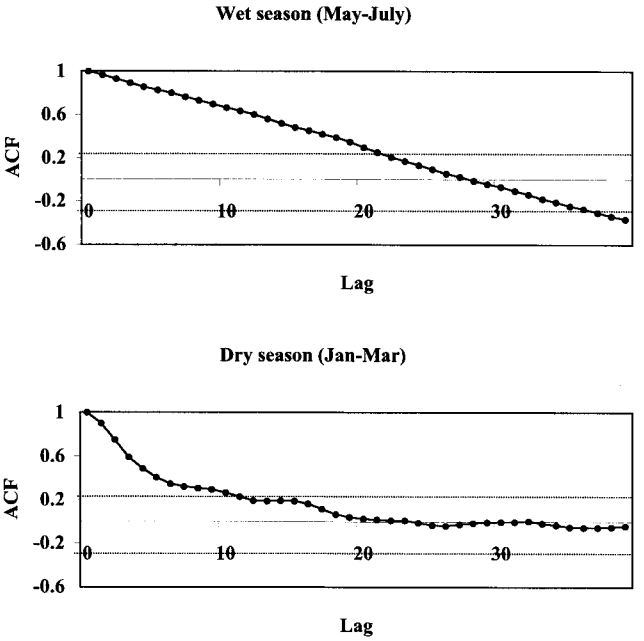

Figure 9. Autocorrelation function of the estimated daily flows at gage 09342500 for the wet season (May–July) and the dry season (January–March) for 1991. These acre-feet correspond well to those shown in Figures 5 and 6 for this gage.

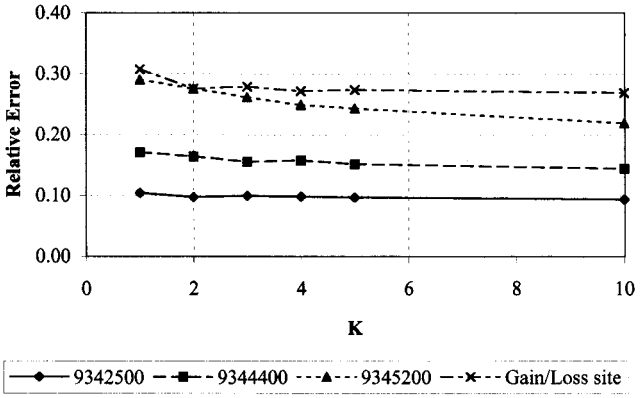

Figure 10. Relative error (mean absolute error/standard deviation of flow) of estimated daily flows for the 1988–1992 period for each site as a function of the number of nearest neighbors used. The relative errors decrease slowly as the number of neighbors increases. The relative error decreases by 5% to 10% for each order-of-magnitude increase (e.g., from 1 to 10) in K for the four sites.

example application presented here. Relative errors (ratio of mean absolute error to standard deviation of the data) were calculated for each site for different K values. As can be seen from Figure 10, relative error decreases as the K value increases. The relative error is the lowest for $K = 10$ for the cases investigated, but the gain over the rule of thumb is large only for the gain/loss site.

5. Discussion and Conclusions

The disaggregation algorithm presented here represents an operational solution to a difficult statistical estimation problem. Classical disaggregation algorithms rely on correlation measures to reproduce attributes considered important for streamflow series. Often such algorithms are very difficult to apply successfully to the high-dimensional situation considered here. Further, attributes such as flow continuity in time and space and natural bounds and ordering of flow magnitudes are hard to reproduce or even describe quantitatively. A purpose of disaggregation for the practitioner is the generation of subscale time series that essentially "looks like" real series and is reasonable for the situation studied. The general methodology exemplified here allows the user to interactively design such an estimation process. A variety of constraints can be explicitly imposed or just checked, depending on the bias of the investigator. While classical disaggregation algorithms consider the minimization of a global total error, the framework presented here can also bound the pointwise error and hence avoid the common problem of a solution set where the error is nonuniformly distributed over the solution space. The solution has interpretability in terms of whether some constraint sets are infeasible at the optimal solution or if they are slack. For instance, too tight a specification of the pointwise error bound could lead to an infeasible solution. The associated infeasible constraints can then be examined, and potential data errors or other factors leading to this situation can be investigated.

The approach presented here does not make prior assumptions about the correlation structure or about the associated probability distribution of the streamflow data. Traditional algorithms consider the use of globally (in state space) estimated correlation matrices and other parameters. Here the estimation process is very local, specific to each month's data that is processed and through the selection of a local neighborhood in the multivariate state space. These features allow for considerable flexibility in adapting to complex functional relationships. These are typical attributes of a nonparametric function estimator. However, a large number of parameters do need to be estimated, including some (e.g., K or the size of the seasonal window, percent error bounds, and flow continuity bounds) that are specified experimentally. The application strategy exemplified here for parameter selection is cross validatory; a subset of the MK data is reserved for validation (the last 5 years in our example were used as MD) and the balance used for estimation. The seasonal window and the number of neighbors to use can be selected through some experimentation at this stage. Subsequently, the full MK data set would be used to disaggregate data from an independent period.

The situation for which the formulation presented was developed included daily flow data at the downstream index gage. The reader may have wondered if this is always necessary. Clearly, the algorithm can be extended to address the case where no daily data are available in the period MD. One would then need to add decision variables to disaggregate the monthly flow at the index site to daily values. The lack of any daily data in the disaggregation period is likely to degrade the performance of the disaggregation scheme.

Traditional disaggregation models are often set up to generate stochastic disaggregated sequences. The algorithm presented here will lead to a single solution to the disaggregation problem. This was the preferred solution for the U.S. Bureau of Reclamation. However, a stochastic disaggregation strategy related to the algorithm presented here can also be readily developed. The essential difference in this case is that one needs to draw samples of daily flow vectors at each site for a month to be disaggregated from the conditional probability distribution $f(\mathbf{x}_m|\mathbf{y}_m)$ instead of estimating the conditional expectation $E[\mathbf{x}_m|\mathbf{y}_m]$ as was done earlier. As in the work by *Lall and Sharma* [1996], a resampling strategy can be used to generate the state space sample of $\mathbf{x}_m$ and $\mathbf{y}_m$ for estimating the disaggregated flows. Consider that the basic parameters (e.g., which constraints to use and the number of nearest neighbors K) of the disaggregation scheme have been decided by the expected value disaggregation scheme presented earlier. Now the algorithm can be modified to randomly draw K2 of the K nearest neighbors of the current month m^*. For instance, suppose we are disaggregating flows for a certain February and $K = 5$. If the five nearest neighbors correspond to the February conditions for historical years 1984, 1981, 1987, 1966, and 1977, then we may want to randomly draw K2 (e.g., 1 or 2 or 5) vectors with a replacement from this set of five. A discrete probability kernel [*Lall and Sharma*, 1996] for selecting from this set that gives a higher weight to the closer neighbors can be used for the purpose. The optimization algorithm is then implemented with these K2 neighbors, and the disaggregated flows are estimated. A choice of 1 for K2 seems intuitively useful. An experimental evaluation of these choices needs to be done.

The example application presented demonstrated the utility of the algorithm developed. Various statistical attributes, as well as attributes intuitively important to the user, were effectively reproduced. The ability to show representative traces (e.g., those associated with historical nearest neighbors of the current month) of daily flows that may be representative of the current situation allow the user to judge whether or not the

solution is likely to be good. The use of the linear programming framework allows the algorithm and its applications to be quite extensible and flexible. Computer programs implementing the algorithm are available on request from the authors. Extensions to consider conditioning on ENSO or on a low-frequency climate state are being pursued.

Acknowledgments. We acknowledge useful review comments provided by Peter Rasmussen. The first author wishes to thank the Department of Science and Technology, Government of India, for awarding a BOYSCAST fellowship to conduct research at the Utah Water Research Laboratory (UWRL), Utah State University, Logan. The UWRL is also thanked for its support of the research.

References

Bartolini, P., and J. D. Salas, Modeling of streamflow processes at different time scales, *Water Resour. Res*., *29*(8), 2573–2587, 1993.

Cleveland, W. S., and S. J. Devlin, Locally weighted regression: An approach to regression analysis by local fitting, *J. Am. Stat. Assoc*., *83*(403), 596–610, 1988.

Grygier, J. C., and J. R. Stedinger, Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res*., *24*(10), 1574–1584, 1988.

Harms, A. A., and T. H. Campbell, An extension to the Thomas-Fiering model for the sequential generation of streamflow, *Water Resour. Res*., *3*(3), 653–661, 1967.

Koutsoyiannis, D., A nonlinear disaggregation method with a reduced parameter set for simulation of hydrologic series, *Water Resour. Res*., *28*(12), 3175–3191, 1992.

Koutsoyiannis, D., and A. Manetas, Simple disaggregation by accurate adjusting procedures, *Water Resour. Res*., *32*(7), 2105–2117, 1996.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for time series resampling, *Water Resour. Res*., *32*(3), 679–693, 1996.

Lall, U., B. Rajagopalan, and D. G. Tarboton, A nonparametric wet/dry spell model for resampling daily precipitation, *Water Resour. Res*., *32*(9), 2803–2823, 1996.

Lane, W. L., *Applied Stochastic Techniques*, *Users Manual*, Eng. and Res. Cent., Bur. of Reclam., Denver, Colo., 1979.

Loucks, D. P., J. R. Stedinger, and D. A. Haith, *Water Resource Systems Planning and Analysis*, 559 pp., Prentice-Hall, Englewood Cliffs, N. J., 1981.

Mejia, J. M., and J. Rousselle, Disaggregation models in hydrology revisited, *Water Resour. Res*., *12*(2), 185–186, 1976.

Rajagopalan, B., and U. Lall, A k-nearest neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res*., *35*(10), 3089–3101, 1999.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, 484 pp., Water Resour. Publ., Highlands Ranch, Colo., 1980.

Santos, E. G., and J. D. Salas, Stepwise disaggregation scheme for synthetic hydrology, *J. Hydraul. Eng*., *118*(5), 765–784, 1992.

Srikanthan, R., Sequential generation of monthly streamflows, *J. Hydrol*., *38*, 71–80, 1978.

Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res*., *20*(1), 47–56, 1984.

Stedinger, J. R., D. Pei, and T. A. Cohn, A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations, *Water Resour. Res*., *21*(5), 665–675, 1985.

Tao, P. C., and J. W. Delleur, Multistation, multiyear synthesis of hydrologic time series by disaggregation, *Water Resour. Res*., *12*(6), 1303–1312, 1976.

Tarboton, D. G., A. Sharma, and U. Lall, Disaggregation procedures for stochastic hydrology based on nonparametric density estimation, *Water Resour. Res*., *34*(1), 107–119, 1998.

Valencia, D. R., and J. L. Schaake Jr., A disaggregation model for time series analysis and synthesis, *Rep. 149*, Ralph M. Parsons Lab., Mass. Inst. of Technol., Cambridge, 1972.

Valencia, D. R., and J. L. Schaake Jr., Disaggregation processes in stochastic hydrology, *Water Resour. Res*., *9*(3), 580–585, 1973.

D. N. Kumar, Department of Civil Engineering, Indian Institute of Technology, Kharagpur, 721302, India. (nagesh@civil.iitkgp.ernet.in)

U. Lall, Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT. (ulall@cc.usu.edu)

M. R. Petersen, Keller-Bliesner Engineering, 78 E. Center Street, Logan, UT 84321. (mrp@kelbli.com)

(Received May 20, 1999; revised February 22, 2000;
accepted February 23, 2000.)