WATER RESOURCES RESEARCH, VOL. 24, NO. 5, PAGES 781–785, MAY 1988

A Representation of Spatial Cross Correlations in Large Stochastic Seasonal Streamflow Models

G. C. OLIVEIRA, J. KELMAN, AND M. V. F. PEREIRA

Centro de Pesquisas de Energia Elétrica, Rio de Janeiro, Brazil

JERY R. STEDINGER

Department of Environmental Engineering, Cornell University, Ithaca, New York

Pereira et al. (1984) present a special disaggregation procedure for generating cross-correlated monthly flows at many sites while using what are essentially univariate disaggregation models for the flows at each site. This was done by using a nonparametric procedure for constructing residual innovations or noise vectors with cross-correlated components. This note considers the theoretical underpinnings of that streamflow disaggregation procedure and a proposed variation and their ability to reproduce the observed historical cross correlations among concurrent monthly flows at nine Brazilian stations.

INTRODUCTION

The Brazilian hydroelectric system is perhaps unique in that multivariate stochastic streamflow sequences have been used routinely to test operating policies and to assist in capacity expansion decisions [*Lepecki and Kelman*, 1985; *Terry et al.*, 1986; *Pereira*, 1985]. *Pereira et al.* [1984] describe the models that have been employed and illustrate their use in capacity planning and hydropower system reliability analyses. The Brazilian hydroelectric system is also unique in its importance: 90% of Brazilian electric energy generation comes from hydro units. The installed hydroelectric capacity is over 35,000 MW. To generate stochastic streamflows to model this system, it was decided to generate concurrent monthly flows at some 100 stations.

Generation of concurrent monthly flows at so many stations poses special problems; if care is not exercised, the number of parameters in the multivariate disaggregation models often employed can easily outstrip the number of data points available for parameter estimation [*Lane*, 1982]. In general, the literature has proposed staged disaggregation procedures to deal with that problem [*Salas et al.*, 1980; *Loucks et al.*, 1981; *Stedinger and Vogel*, 1984]. However, for the Brazilian system a distinctly different approach was adopted. As described by *Pereira et al.* [1984], the annual flows generated for each site were disaggregated by a separate model using residual innovations or noise vectors whose components were cross-correlated with the components of the other vectors employed to generate monthly flows at all other sites. The use of cross-correlated residual innovations was intended to capture the correlation among concurrent monthly flows. Similar "diagonal" models are discussed by *Stedinger et al.* [1985].

An important feature of this approach is that it preserves the "univariate" character of the disaggregation model for flows at each station, thus avoiding the complexities introduced by going to large multivariate models. Also, if the operation of a proposed reservoir or power plant at a new site is to be investigated, one can proceed to generate additional series of synthetic flows for that site which are consistent with series already generated and available for all other sites. Thus one need not generate new streamflow series for every point in the entire system, only the new station of interest.

Because of the potential value of this modeling approach for large water resource and hydroelectric systems this note provides a discussion of the improvement provided by the use of cross-correlated residual innovations. A variation of the original nonparametric residual generation procedure proposed by *Pereira et al.* [1984] is also developed. It does a better job of capturing the cross correlations among concurrent monthly flows.

Paper number 7W4936.
0043-1397/88/007W-4936$05.00

OUTLINE OF THE METHODOLOGY

Notation

The basic disaggregation model which will be used for generating monthly flows at each site is that proposed by *Mejia and Rousselle* [1976]:

$$Y_t = AX_t + BZ_t + CV_t \tag{1}$$

where

- Y_t 12-dimensional vector of zero-mean translated monthly flows;
- X_t zero-mean translated annual flows;
- Z_t p-dimensional vector of zero-mean translated monthly flows from the preceding year;
- V_t 12-dimensional column vector of residuals (independent zero-mean unit variance random variables);
- A 12×1 coefficient matrix;
- B $12 \times p$ coefficient matrix;
- C 12×12 coefficient matrix.

X_t represents the annual flow during year t (minus its mean) and Y_t the monthly flows for the same year (minus their mean). Z_t represents the last p streamflows of the previous year $t - 1$.

Let S_{WU} represent the covariance between any two vectors W and U. Following *Mejia and Rousselle* [1976], the A, B, and C matricies were selected so as to reproduce the sample estimates of S_{YZ}, S_{YX}, and S_{YY}, approximately. In particular, CC' must equal a residual covariance matrix M which involves all four of those matricies if the generated Y vectors are to have a covariance matrix which approximates S_{YY}.

As *Kelman et al.* [1979], *Lane* [1982], and *Stedinger and Vogel* [1984] all observed, the Mejia-Rousselle model in general does not exactly reproduce the indicated statistics, as Mejia and Rousselle originally thought. *Lane* [1982] proposes a modified Mejia-Rousselle model which does reproduce S_{YY}. Lane's procedure was not used in the original study [*Pereira et al.*, 1984] or in this extension of that investigation, though it could have been. For univariate disaggregation procedures, *Kelman et al.* [1979] conclude that the errors incurred with the Mejia-Rousselle model are small in comparison to parameter uncertainty. If such errors are of concern, one can use Lane's modification.

Stedinger and Vogel [1984] proposed an alternative disaggregation procedure which avoids the problems with Mejia and Rousselle's model. However, it is not clear if the Stedinger-Vogel approach would work well with the nonparametric innovation generation scheme under investigation here.

Nonparametric Generation

Synthetic monthly streamflows were generated by sampling from the components of the historical residual vectors V_t obtained as a by-product of the model fitting process. *Pereira et al.* [1984] introduced this nonparametric approach. For each site the historical residuals $\hat{V}_t$ are obtained as the solution of

$$C\hat{V}_t = \hat{Y}_t - A\hat{X}_t - B\hat{Z}_t \qquad \text{for all } t = 1, 2, \cdots, m \tag{2}$$

where m is the length of the historical record.

The generated residual vectors V_t should have independent components, for the assumption made in the calculation of the C matrix is that the V_t vector's covariance matrix is the identity matrix (and thus its components are independently distributed with unit variances). Our nonparametric procedure generates innovation vectors V_t with independent components which have the empirical distribution of each component of the $\hat{V}_t$ vector for each station. This is done by independently drawing each component V_{it} of V_t from among the m observed values of $\hat{V}_{it}$. Let the value of this ith residual be that corresponding to year T_i. Then the core of the procedure is to select a set of indices $\{T_1, \cdots, T_{12}\}$ which are independent of one another and which take on the values 1–m with equal probability. Because there are m possible values for each T_i, and the corresponding V_{it}, there are $(m)^{12}$ possible values that the residual vector can assume. This is a large "population" given typical historical record lengths m of 30–50 years. This procedure is illustrated in Figure 1.

Representation of Cross Correlations Among Concurrent Flows

If the annual flows for each site are disaggregated independently, the only "source" of cross correlation among concurrent monthly flows in (1) would come from the annual values X_t. Thus the cross correlation between concurrent monthly flows would generally be smaller than those observed in the historical record.

To avoid this problem, *Pereira et al.* [1984] suggest that one use the same indices $T_1, T_2, \cdots, T_{12}$ that were previously employed at all other sites for that same year t. Note that the indices are the same, but the historical residual values to which they correspond vary from site to site. This scheme will reproduce cross-site correlations among each component of the residual vectors; thus it should improve the representation of the cross correlation between concurrently monthly flows

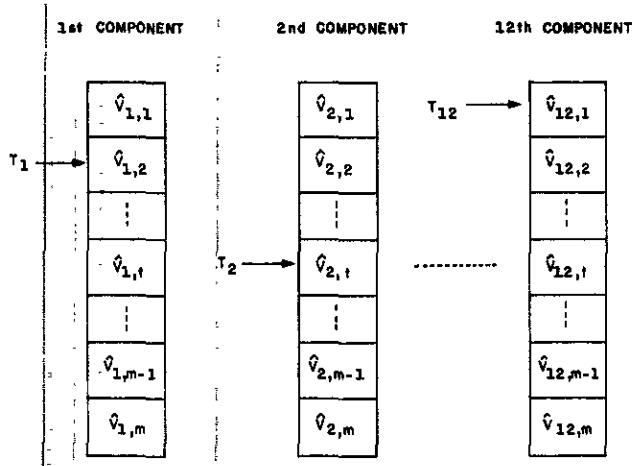

Fig. 1. Nonparametric generation of monthly flows for a historical record of m years.

QUANTIFICATION OF THE IMPROVEMENT

Notation

Some algebra allows quantification of the adequacy of the proposed residual sampling scheme. Let the monthly streamflow generation models for two sites k and h be represented as

$$Y_t^k = A^k X_t^k + B^k Z_t^k + C^k V_t^k \tag{3}$$

$$Y_t^h = A^h X_t^h + B^h Z_t^h + C^h V_t^h \tag{4}$$

The residual errors associated with these flows models are

$$W_t^k = C^k V_t^k = Y_t^k - A_t X_t^k - B^k Z_t^k \tag{5}$$

$$W_t^h = C^h V_t^h = Y_t^h - A^h X_t^h - B^h Z_t^h \tag{6}$$

The cross-covariance between W_t^k and W_t^h is related to the cross correlation between V_t^k and V_t^h such that

$$S_{ww}^{kh} = E(W^k W^{h\prime}) = C^k E(V^k V^{h\prime}) C^{h\prime} = C^k S_{vv}^{kh} C^{h\prime} \tag{7}$$

where the prime indicates the transpose operation.

Let $W' = [W^k, W^h]'$. Then

$$\text{Cov}\,[W] = \begin{bmatrix} S_{ww}^{kk} & S_{ww}^{kh} \\ S_{ww}^{hk} & S_{ww}^{hh} \end{bmatrix} \tag{8}$$

Interpretation of the Generation Scheme

Substituting (7) into (8), one obtains

$$\text{Cov}\,[W] = \begin{bmatrix} C^k S_{vv}^{kk} C^{k\prime} & C^k S_{vv}^{kh} C^{h\prime} \\ C^h S_{vv}^{hk} C^{k\prime} & C^h S_{vv}^{hh} C^{h\prime} \end{bmatrix} \tag{9}$$

By randomly selecting the different components of the generated V_t^k [see *Pereira et al.*, 1984], one insures that

$$S_{vv}^{kk} = S_{vv}^{hh} = I \tag{10}$$

where I is the identity matrix. In general, C^k is selected so that

$$C^k C^{k\prime} = M^k \tag{11}$$

where M^k is the 12×12 matrix of the historical covariances of the residuals. This choice insures that the generated flows reproduce, approximately, the observed covariance of the Y_t^k vectors.

Substituting (10) and (11) into (9), yields

$$\text{Cov}\,[W] = \begin{bmatrix} M^k & C^k S_{vv}^{kh} C^{h\prime} \\ C^h S_{vv}^{hk} C^{k\prime} & M^h \end{bmatrix} \tag{12}$$

Analogously, defining $V = [V^k, V^h]$, yields

$$\text{Cov}(V) = \begin{bmatrix} I & S_{vv}^{kh} \\ S_{vv}^{hk} & I \end{bmatrix} \tag{13}$$

In order to reproduce the cross correlation between monthly flows at stations k and h it is necessary to reproduce the residual cross-covariance matrix S_{vv}^{kh}. It can be seen that the independent sampling of the residual vectors for site corresponds to replacing S_{vv}^{kh} in (13) by a null matrix. The use of the same indices at all sites corresponds to replacing S_{vv}^{kh} by $\tilde{S}_{vv}^{kh}$, where

$$\tilde{S}_{vv}^{kh} = \text{diag}\,\{[S_{vv}^{kh}]_{ii}\} \tag{14}$$

In other words, the suggested generation scheme preserves the diagonal elements of the residual cross-covariance matrix S_{vv}^{kh}.

Equation (14) implies that S_{ww}^{kh} will be replaced by

$$\tilde{S}_{ww}^{kh} = C^k \tilde{S}_{vv}^{kh} C^{h'} \tag{15}$$

Calculation of the C Matrix

An important issue upon which the success of the nonparametric sampling schemes depends is selection of the C matrix. As in (11), C is obtained as the solution of $CC' = M$. However, this solution is not unique; several C matrices are suitable for the disaggregation scheme in (1).

Since C is used in the calculation of the historical residuals $\hat{V}_t$ (see equation (2)), different C will lead to different $\hat{V}_t$ vectors and hence to different approximations of the covariance matrix of the residuals. Some C may be "better" than others.

Pereira et al. [1984] originally calculated C by spectral decomposition of M so that

$$C = P\Lambda^{1/2} \tag{16}$$

where

P 12×12 matrix of eigenvectors of M;
Λ diagonal matrix of eigenvalues, diag (λ_i), with $\lambda_1 \geq \lambda_2, \cdots, \geq \lambda_{12} \geq 0$.

Pereira et al. [1984] also showed that the last eigenvalue is always zero so that one column of C can be deleted.

If C is calculated as in (16), then each $\hat{V}_{it}$ is the noise associated with the eigenvector corresponding to the ith largest eigenvalue of the matrix M; thus use of the same indices $T_1, \cdots, T_{12}$ at each site corresponds to reproduction of the cross correlation between the variation in the first, in the second, and up through the last eigenvectors spaces. However, there is no compelling reason why the variation in site ks ith eigenvector should have a particularly high cross correlation with variation in site hs ith eigenvector.

An alternative procedure is to decompose M into a lower-triangular C matrix using Cholesky decomposition. Then V_{it} corresponds to the unique innovation at each site for month 1, V_{2t} corresponds to the residual introduced for month 2, V_{3t} the innovation for month 3, and so forth. Thus reproduction of the cross correlation between concurrent V_{it} reproduces the cross correlation between the innovations which are introduced in each month to reproduce the variance of the flows in that month. This still does not insure that the cross correlation between concurrent flows are reproduced; however, the intersite correlations that are reproduced with a lower-triangular C are likely to be significant ones. These two options were tested in the case study discussed below.

An Example

The performance of the nonparametric residual generation schemes will be illustrated with the nine stations in the Brazilian hydroelectric system listed in Table 1. Those stations are located in four different river basins in the southeastern and southern regions of the country. The distance between the sites range from 250 to 1800 km.

A 4000-year sequence of synthetic monthly streamflows for the nine sites was obtained using three schemes for generating the residual innovation vectors: (1) independent residual vectors at all sites; (2) the use of the same streamflow index for each component of V based upon a C matrix calculated by spectral decomposition of M, with $\lambda_1 \geq \lambda_2 \geq, \cdots, \geq \lambda_{n-1}$; and (3) the variation of the procedure in (2) wherein C is a lower-triangular matrix.

Scheme (1) makes no attempt to capture the cross correlation among concurrent monthly flows at different stations. The only tie between the generated monthly flows will be the cross correlation among the generated annual flows. Scheme (2) proposed by *Pereira et al.* [1984] introduces cross correlation among the monthly flows within each year by preserving the cross correlation among each component of the innovation vectors V^k. Finally, our variation of the original residual generation scheme uses a lower-triangular C matrix so that each component of the V^k innovation vectors uniquely relates to the variation of the flow in the corresponding month. Thus scheme (3) is likely to do a better job of capturing the cross correlation among concurrent monthly flows that either scheme (2) or scheme (1).

The cross correlations among flows at stations 1, 5, and 6 and all other stations was calculated. For each pair of stations (k, h), and month i, the consistency between the historical correlations $\rho^H(k, h, i)$ and the calculated correlations $\rho^G(k, h, i)$ among the generated flows was summarized by the goodness to fit statistic:

$$\chi^2(k, h) = \sum_{i=1}^{12} [\rho^G(k, h, i) - \rho^H(k, h, i)]^2 \tag{17}$$

reported in Table 2. Figures 2 and 3 provided a visual comparison of the correlations for the worst and best cases (based upon the squared deviations). The analysis indicates:

Generation scheme (2) employed by *Pereira et al.* [1984] does substantially better than scheme (1) which generated the residuals independently.

Scheme (3), which employes a lower-triangular C matrix, is superior to the original procedure, scheme (2), which used a C matrix obtained by spectral decomposition of M. The improvement was not uniform for all stations.

TABLE 1 Characteristics of the Stations

Station Number	Station Name	River	Basin	Mean Inflow, m^3/s
1	Furnas	Grande	Paraná	912
2	A. Vermelha	Grande	Paraná	1929
3	S. Simão	Paraíba	Paraná	2241
4	R. Barbosa	Tietê	Paraná	581
5	Itaipu	Paraná	Paraná	9040
6	T. Marias	S. Francisco	S. Francisco	1453
7	Sobradinho	S. Francisco	S. Francisco	2200
8	S. Osório	Iguaçu	Iguaçu	926
9	Jacuí	Jacuí	Jacuí	181

Length of all the historical records is 40 years.

TABLE 2. Goodness of Fit for Monthly Spatial Correlation: Sum of Monthly Quadratic Deviations

	Station							
	2	3	4	5	6	7	8	9
	Station 1 × Station #							
Independently generated	2.186	0.923	1.022	0.994	1.132	0.419	0.981	0.280
C spectral decomposition	0.072	0.546	0.646	0.561	0.208	0.373	1.017	0.235
C lower triangular	0.016	0.116	0.160	0.137	0.132	0.172	0.799	0.175
	Station 5 × Station #							
Independently generated	1.420	0.738	2.260		0.328	2.016	0.227	0.629
C spectral decomposition	0.843	0.395	1.635		0.307	1.990	0.241	0.315
C lower triangular	0.204	0.108	0.301		0.175	0.720	0.208	0.261
	Station 6 × Station #							
Independently generated	1.092	1.303	0.486			0.997	0.486	0.697
C spectral decomposition	0.178	0.752	0.390			0.701	0.347	0.305
C lower triangular	0.125	0.102	0.319			0.372	0.330	0.273

In some cases, there was substantial disagreement between the historical cross correlation among concurrent monthly flows and the cross correlations among the generated monthly flows.

An unanswered question is how important is it to exactly reproduce the historical cross correlations among the generated monthly flows? First, the historical values are only estimates of limited precision. That observation is of limited comfort in that the nonparametric schemes consistently generate values less than their historical counterparts reflecting a consistent downward bias. With large reservoir systems, the cross correlation among concurrent monthly flows may not be critical because of their carryover capacity so that the cross correlation among seasonal and annual volumes may be more critical. However, for relatively small reservoir systems with little storage capacity the cross correlation among concurrent flows could be quite important. Thus the adequacy of the techniques considered here depends on both how well they do statistically in a particular instance and on the characteristics of the system being studied.

Conclusions

The generation of concurrent monthly flows at a large number of stations for reservoir and hydropower simulation poses special problems. The disaggregation scheme proposed by *Pereira et al.* [1984] has the advantage that separate univariate models can be used to generate monthly flows at each station separately. Thus one can avoid the complexity and

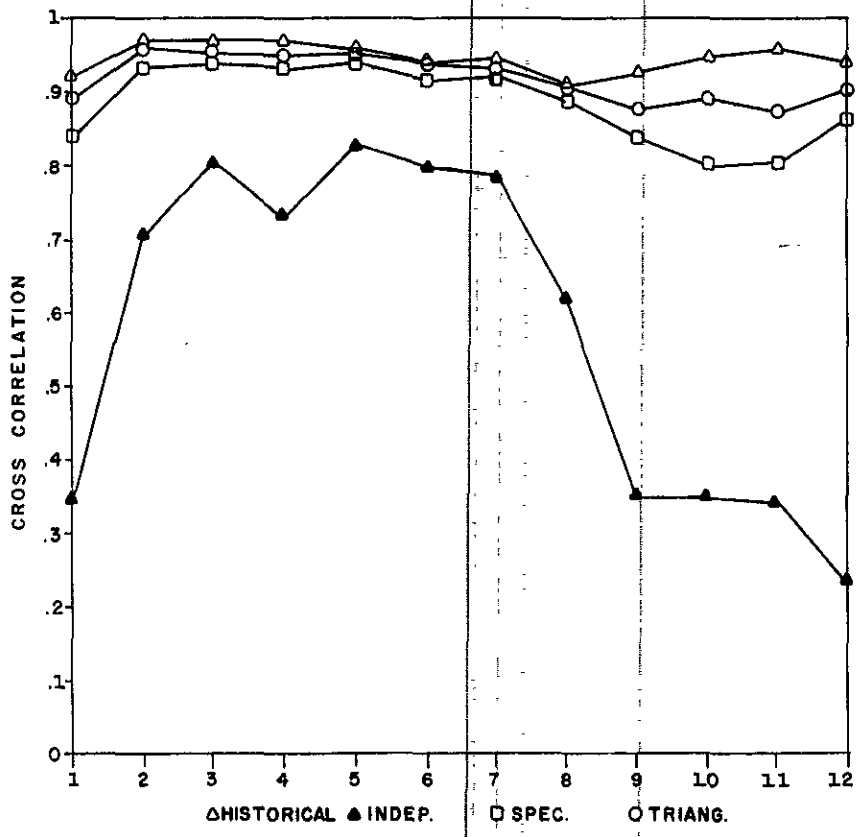

Fig. 2. Monthly cross correlations; stations 1 × 2.

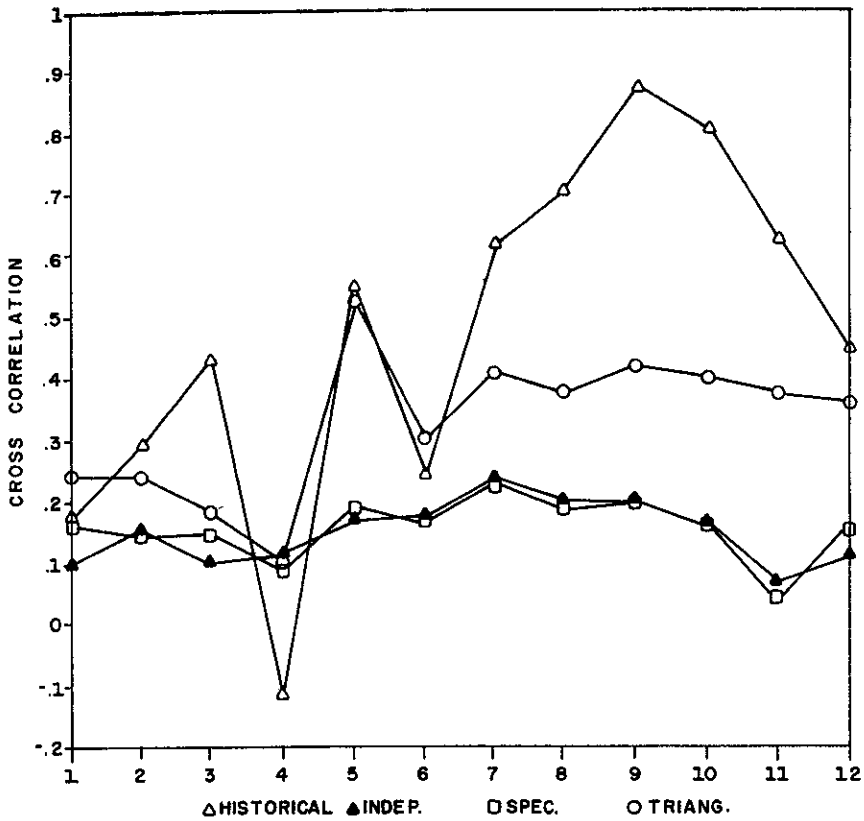

Fig. 3. Monthly cross correlations; stations 6 × 8.

drawbacks of large multivariate models. With that approach, it is also possible to generate a series of flows for a new station without the need to recompute flows for other stations which were generated earlier.

The performance of three residual generation schemes was illustrated by considering their performance for nine streamflow stations in southeastern and southern Brazil. A variation of the nonparametric procedure for generating residual vectors with cross-correlated components was shown to do the best job, though not a perfect job, of reproducing the historical cross correlation among concurrent flows at the nine stations.

REFERENCES

Kelman, J., G. C. Oliveira, and M. V. F. Pereira, Synthetic streamflow generation by disaggregation (in Portuguese), paper presented at the Fifth National Seminar on Production and Transmission of Electrical Energy (V SNPTEE), Recife, Brazil, 1979.

Lane, W. L., Corrected parameter estimates for disaggregation schemes, in *Statistical Analysis of Rainfall* and *Runoff*, edited by V. P. Singh, Water Resources Publications, Littleton, Colo., 1982.

Lepecki, J., and J. Kelman, Brazilian hydroelectric system, *Water Int.*, *10*(4), 156–161, 1985.

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

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

Pereira, M. V. F., Hydroelectric system planning, in *Expansion Planning of Electrical Power Systems: A Guide Book*, chapter 8, International Atomic Energy Agency, Vienna, 1985

Pereira, M. V. F., G. C. Oliveira, C. C. G. Costa, and J. Kelman, Stochastic streamflow models for hydroelectric systems, *Water Resour. Res.*, *20*(3), 379–390, 1984.

Salas, J. D., J. W Delleur, Y. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Series*, Water Resources Publications, Littleton, Colo., 1980.

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. P. Lettenmaier, and R. M. Vogel, Multisite ARMA (1, 1) and disaggregation models for annual streamflow generation, *Water Resour. Res.*, *21*(4), 497–509, 1985.

Terry, L. A., M. V. F. Pereira, T. A. Araripe Neto, P. H. Salles, and L F. A. C. Silva, Coordinating the energy generation of the Brazilian system, *Interfaces*, *16*(1), 65–82, 1986.

J. Kelman, G. C. Oliveira, and M. V. F. Pereira, Centro de Pesquisas de Energia Elétrica, P.O. Box 2754, Rio de Janeiro, Brazil.

J. R. Stedinger, Department of Environmental Engineering, Cornell University, Ithaca, NY 14853.

(Received April 27, 1987;
revised January 19, 1988;
accepted January 22, 1988.)