Conceptually-based multivariate simulation of monthly runoff

P. Claps

Department of Hydraulics, Trasportation and Civil Infrastructures, Politecnico di Torino, Torino, Italy (claps@polito.it)

Abstract: This paper presents a multivariate extension of a parsimonious conceptually-based Auto Regressive-Moving Average (ARMA) stochastic model for monthly runoff. The multi-station model is a Contemporaneous-ARMA (CARMA), which considers separately the serial and space correlation of runoff. Serial correlation is reproduced in individual series by an ARMA model. The ARMA model residuals are uncorrelated in time but correlated in space. Spatial correlation of runoff is then reproduced by generating correlated series of residuals and using them to generate runoff through the individual ARMA models. In the conceptual framework, stochastic ARMA parameters are related to the parameters of a linear system, which represents the watershed filter that produces runoff. The system input is the effective rainfall, which is inversely estimated through the ARMA model residual. Application of the CARMA model in the conceptual framework consists in reproducing the spatial correlation on the effective rainfall rather than on the residuals. A suitable technique is also proposed for estimation of correlation in matrices with gaps. The performances of the model are discussed with regard to its application on a 9-station system in Southern Italy.

Keywords: Monthly Runoff; Conceptual model; multivariate; CARMA

1. INTRODUCTION

Simulation of simultaneous runoff series over several stations is an essential part of current practice of planning and management of water resources systems. For reasons of parsimony, it is nowadays widely accepted that multi-station runoff simulation can be achieved with models that preserve spatial correlation among stations (multi-site) and let serial correlation be reproduced by at-site univariate models (*Contemporaneous ARMA – CARMA* - models, [Salas *et al.,* 1980]). This representation means that runoff $d_{i,t}$ in station i at time t, depends on past values measured at the same station ($d_{i,t-1}$; $d_{i,t-2}$, etc.) but not on the other station's past values, as in the complete (vector) multivariate formulation. The (spatial) dependence of $d_{i,t}$ on the values occurred in the other stations is thus evaluated only on *contemporaneous* runoff values ($d_{i+1,t}$, $d_{i+2,t}$, etc.) by means of a correlation matrix of the residuals of the individual serial correlation models. This formulation reduces greatly the number of parameters to estimate with respect to the vector model, without significant loss of information. In addition, univariate models for serial correlation can be different from one station to another. This property has not yet been fully exploited in the literature, where the cumbersome tools for parameter estimation have left little room for deeper studies on the identification phase [see e.g. Salas et al., 1985].

To date, a full-featured contemporaneous periodic-ARMA model for monthly runoff is available [Rasmussen et al., 1996], in which serial and spatial correlations are reproduced season by season thanks to a large parameter set. In most practical cases, however, insufficient data prevents one to use the above, powerful but demanding, model. On the other hand, conceptually-derived models [Salas and Obeysekera, 1992; Claps et al., 1993] provide physical-like bases for supporting parsimonious model identification and estimation. They also present features allowing one to validate results of models applied on short or discontinuous time series. These approaches have been so far proposed only in the univariate form.

Conceptually-based models can help to overcome the following shortcomings of empirical models: i) reliance on long, continuous and contemporaneous runoff records; ii) sensitivity to normality of residuals; iii) uncertainties in identification of individual models for serial correlation. In this paper it will be shown how a CARMA model structure, that fits into the conceptually-based ARMA framework proposed by Claps et al. [1993], addresses the issues above.

2. NON GAUSSIAN CONCEPTUALLY BASED CARMA MODEL BUILDING

2. 1 Univariate model identification

In a CARMA formulation, serial correlation is reproduced in the individual stations by univariate models, that can differ among stations. In the present case, the univariate model building follows the methodology proposed by Claps et al. [1993], in which the type and the order of ARMA models are identified on the basis of a conceptualisation of the runoff process. Identification is made a-priori, as also proposed by Salas and Obeysekera [1992], considering runoff as the output of a linear system (the watershed) fed by the effective rainfall I_t (*i.e.* rainfall minus evapotranspiration). The system is made up of two groundwater terms, having over-year and over-month time constant, plus a quick (zero-lag) surface component.

In nature, the process I_t is skewed but uncorrelated. Therefore, the stochastic model obtained is an ARMA(2,2) model with non-gaussian residual:

$$d_t - \Phi_1\, d_{t-1} - \Phi_2\, d_{t-2} = \varepsilon_t - \Theta_1\, \varepsilon_{t-1} - \Theta_2\, \varepsilon_{t-2} \quad (1)$$

The zero-mean effective rainfall i_t ($=I_t - \mathrm{E}[I_t]$) is not known a-priori: it is obtained, as in inverse problems, from the residual term ε_t:

$$i_t = \varepsilon_t \,/\, \left[1 - ak\left(1-e^{-1/k}\right) - bq\left(1-e^{-1/q}\right)\right] \quad (2)$$

The four conceptual model parameters (a,b,k,q) are directly related [see Claps et al., 1993] to their stochastic counterparts ($\Phi_1,\Phi_2,\Theta_1,\Theta_2$). In the same paper, identification of runoff components and parameter estimation are described in detail. These phases take advantage of the different characters exhibited by the time series at aggregated scales: at the annual scale, for instance, groundwater components with under-year time constant are hidden; this allows one to better identify possible over-year components. In case only one (over-month) groundwater system is identified, the model reduces to an ARMA(1,1).

Relation (2) clarifies that the ARMA residual contains information on the input process that produces the catchment runoff. Therefore, to generate synthetic runoff series one can either reproduce the residual ε_t as it is, or (as chosen here) obtain the estimated effective rainfall series through (2) and generate it as a random variable with hydrological meaning. As such, I_t must be non negative and can have a finite probability at zero. Possible negative values resulting in the estimated I_t series are set to zero, while locally preserving the mean.

The distribution proposed for I_t is a compound square-root normal, with pdf as:

$$P(I=0) = P(0) = P_0; \quad (3)$$
$$f(I>0) = (1-P_0)\frac{1}{2r\sigma_r\sqrt{2\pi}}\exp\left[\frac{1}{2\sigma_r^2}(r-\mu_r)^2\right];$$

in which P_0 is the zero finite probability and μ_r, σ_r are the mean and standard deviation of the square-root transformed variable $r=\sqrt{I}$. Parameters of the continuous part I^+ ($=I>0$) of the distribution. are estimated by the method of moments on the original (untransformed) data, using the relations:

$$\mu_r = \sqrt{\mu_{I^+}}\cdot\left(1-\frac{\sigma^2_{I^+}}{8(\mu_{I^+})^2}\right); \quad \sigma_r^2 = \frac{\sigma^2_{I^+}}{4\mu_{I^+}} \quad (4)$$

The zero probability P_0 is evaluated as the sample frequency of zeros. This set of three parameters is assumed to vary month by month, with the possibility of a Fourier smoothing for more compact intra-annual representation.

To generate runoff series, synthetic series of I_t must be transformed in the residual ε_t by inverting (2), so to introduce it into the estimated ARMA model.

2. 2 Spatial correlation of effective rainfall

Based on the above premises, it can be more clear now that the focus of this paper is on the estimation of the spatial correlation of the effective rainfall, as the only step needed for the multivariate extension of the univariate conceptual ARMA model.

In a multisite CARMA model, spatial correlation is reproduced by generation of contemporaneous series of residuals, which must maintain the spatial correlation found in the original ARMA residuals obtained in each station. This can be made using a standard procedure [*e.g.* Salas et al., 1980], in which generic gaussian correlated data e_t can be generated using the matrix equation:

$$e_t = \mathbf{B}\, \boldsymbol{\xi}_t \quad (5)$$

where $\boldsymbol{\xi}_t$ is a nx1 vector of normal uncorrelated standardised values at time t and $\mathbf{B}$ is a nxn matrix. $\mathbf{B}$ is linked to the sample correlation matrix $\mathbf{G}$ of the original e_t series by the 'gramian' equation: $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$. Application of (5) to a non gaussian residual, like ε_t, requires a transformation of the variable (ε_t) to make it gaussian.

In the present case, the variable on which spatial correlation is reconstructed is the effective rainfall I_t, obtained through equation (2). Analysing I_t instead of ε_t gives the advantage of dealing with a variable with hydrological meaning, whose characteristics in a region may be related to prior climatic knowledge (e.g. total precipitation). This provides diagnostic features to the stochastic model and makes it possible, in principle, to deal with runoff generation in ungauged stations.

On the other hand, I_t is a (non-gaussian) compound variable, and this complicates considerably the transformation. It was decided not to transform the entire distribution but to deal independently with the spatial correlation for the zero – non zero occurrence of the variable and for the continuous part I^+ of the distribution. This has been done by adapting for the spatial case a procedure implemented by Chebaane et al. [1995] for reproducing the serial correlation of a univariate intermittent process. In practice, the generated multisite correlated sequences of I_t are obtained by first generating spatially correlated (0,1) sequences, with a two-state Markov chain; these sequences are then multiplied with generated values of the continuous distribution of I^+, by applying (5) to the square-root of the variable, to make it gaussian.

In the following, the multisite generation of effective rainfall will be discussed with reference to series with gaps, by considering separately the continuous and the intermittent part of the process. For the latter, the relations used derive directly from the transition probability method described in Chebaane et al. [1995].

3. CORRELATION MATRIX ESTIMATION IN SERIES WITH GAPS

Well-established procedures for statistical analysis with missing data tend to become quite sophisticated for multivariate cases and depend critically on the hypothesis of normality [see e.g. Schafer, 1997]. The approach presented here tries to tackle the missing data problem directly in the simulation phase without data reconstruction. Schemes for approximated, yet objective, estimation of parameters and correlation matrices are devised, preserving the mutual information among stations by considering dataset pairs instead of the entire matrix. It must be also considered that the proposed procedure is built to deal with datasets presenting long interruptions, or where information must be extracted from time series of reduced length.

In CARMA models, methods for dealing with missing data can be different for the two phases of reconstruction of serial and spatial correlation. For the serial correlation, the method used to deal with discontinuous series (see *e.g.* Table 1) is based on the assumption of a unique set of ARMA parameters, obtained by weighted average of those estimated in the continuous sub-series.

Effective rainfall values estimated in the subseries are transferred to the next stage without modifications. In practice, the matrix made up of the effective rainfall series resulting by the univariate estimation will contain all the gaps of the original data matrix. The next two subsections describe the techniques used for the estimation of spatial correlation on these gapped matrices.

3. 1 Spatial correlation of the effective rainfall: continuous part

Two main methods for estimating the correlation matrix **G** on series with gaps were found in literature. The first is referred to as *case deletion* [Schafer, 1997, p. 23] and works so as only the *cases* (rows) in which all of the stations have data are accepted and processed. Using data matrices of equal length ensures that the correlation matrix is positive definite, which is a sufficient condition to solve the ‘gramian’ equation $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$ (and have the matrix **B** available to generate spatially correlated variables using (5)). The problem with this simple and intuitive technique is that the resulting dataset can be too short to retain enough information about the spatial correlation structure.

The second method, the one adopted here, was proposed by *Basson et al.* [1994] and is based on the reproduction of the original correlation between couples of stations. In this method the *case deletion* is applied between each pair of stations, reducing greatly the number of deleted data. On the other hand, the data considered for a given station can vary depending on which other station is considered in turn. As a consequence of this procedure, it is not assured that the resulting correlation matrix is even positive *semidefinite*, which is a necessary condition for decomposition of $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$. If the computed correlation matrix comes out as negative definite, a *reconditioning* technique must be applied to make it at least positive semidefinite. This method ensures reasonable preservation of sample spatial correlation even when data are quite sparse, without requiring data infilling.

The choice of the reconditioning method can be made in a restricted lot, beginning with Fiering [1968] and arriving at Rasmussen et al. [1996]. The former is very simple and intuitive, while the latter is part of a wider method applied to families of correlation matrices required in periodic contemporaneous models. For the constant-parameter ARMA model considered here, the

Fiering method performed best, and was the one adopted throughout the procedure. Decomposition of the gramian equation $\mathbf{BB}^T = \mathbf{G}$ is achieved through the singular value decomposition (SVD) method [*e.g.* Press et al., 1986].

Based on the method by Basson et al. [1994] and on the reconditioning technique by Fiering [1968], spatially correlated non-zero values of effective rainfall are generated with the following steps:

1. The sample correlation matrix **G** is estimated on the square-root transformed (non-zero) values of estimated effective rainfall
2. If **G** is negative-definite, the reconditioning method by Fiering is applied.
3. Equation $\mathbf{BB}^T = \mathbf{G}$ is decomposed by the SVD method, implemented in the Matlab® environment.
4. Once **B** is computed, generated series of I^+ are obtained through (5), in which $\boldsymbol{e}_t$ is squared to preserve the original distribution.

3.2. Spatial correlation of the effective rainfall: Intermittent part

A two-state (0,1) occurrence variable X represents the intermittent part of the effective rainfall compound distribution. Estimation of the spatial correlation matrix on X is obtained through the transition probabilities (TP) method proposed by Chebaane et al. [1995]. This method is formulated for a 2-site case, and is applied here to pairs of stations, sequentially.

If $X_{S_0,\tau}$ is the variable X referred to the month τ in the station s_0 and $X_{S_1,\tau}$ is the contemporaneous variable referred to the station s_1, the transition probability (TP) $P_{ij} = P(X_{s_1,\tau} = j \mid X_{s_0,\tau} = i)$ is the conditional probability of having $j \in \{0,1\}$ in station s_1 given $i \in \{0,1\}$ in station s_0. Estimates of P_{ij} are obtained, for each month τ, directly from the sample, by counting the numbers n_{ij} of the actual $i \rightarrow j$ transitions and dividing them by the total number n_i of the starting states.

This bivariate procedure cannot be immediately extended to a general multivariate case. However, the application devised here is based on the presence of nested and adjacent basins (a quite frequent configuration in this kind of problems) and proceeds filling up the correlation matrix according to reasonable sequences of station pairs (upstream or downstream in the same basin and continuing towards adjacent basins). Practical applications on different sequences of station pairs showed that the final correlation matrix does not change much, independently of the choice of the sequence. However, this result is not necessarily expected in a more general context and is matter for further model development.

The presence of uneven datasets does not affect the nature of the transition matrix, but rather the inner congruence of transition probabilities. In particular, the following congruence equations must hold:

$$(P_0)_{S_1,\tau} = P_{00}(P_0)_{S_0,\tau} + P_{10}(P_1)_{S_0,\tau} \qquad (7.a)$$

$$(P_1)_{S_1,\tau} = P_{01}(P_0)_{S_0,\tau} + P_{11}(P_1)_{S_0,\tau} \qquad (7.b)$$

$$P_{00} + P_{01} = 1 \; ; \qquad P_{10} + P_{11} = 1 \qquad (7.c)$$

In the (7), $(P_0)_{S1,\tau}$ represents the marginal probability (MP) of zero effective rainfall for the month τ at the station S_1, while P_{10} is the TP P_{ij} with i=1 and j=0.

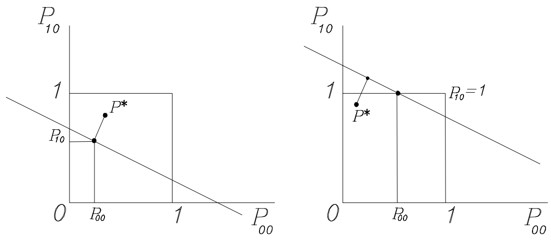
Uneven datasets will actually have different record lengths as, for instance, N_{S0} and N_{S1}. The case deletion makes it possible to compute the four P_{ij} TP's on even subsets of data, which will have length $N_{S0S1} \le min(N_{S0},N_{S1})$. On the other hand, marginal probabilities (*e.g.* $(P_0)_{S1,\tau}$) re-computed on these shorter subsets can change even considerably with respect to the original values. If MP's were retained to their original values, congruence would not be ensured in relations (7). In reconciling this incongruent situation it was chosen to save the original information related to marginal probabilities, modifying the estimation of TP's so as to respect conditions (7). Keeping the MP's to the original values, they will constrain the coefficients of the linear relations represented by (7.a) and (7.b). For instance, (7.a) represents the equation of a straight line in the (P_{00}, P_{10}) plane, that can be written as $P_{10} = a + bP_{00}$, with coefficients:

$$a = \frac{(P_0)_{s_1,\tau}}{(P_1)_{s_0,\tau}}; \qquad b = -\frac{(P_0)_{s_0,\tau}}{(P_1)_{s_0,\tau}} \qquad (8)$$

After the case deletion, the TP's P^*_{00}, P^*_{10} computed on the sample will be the coordinates of a point P^*, as indicated in Figure 1. The point P^* will not lie, in general, on the straight line (7.a), that has coefficients obtained through (8) using the original MP's. The P^* coordinates will be then corrected moving the point orthogonally towards the congruence line (Figure 1a). If the interception point falls outside the (1,1) congruence square, it must further move on the line until the square is reached (Figure 1b).

With respect to the case of equation (7.a), the corrected coordinates, obtained for the simpler situation depicted in Figure 1.a, result by:

$$P_{00} = \frac{P_{10}^{*} + \frac{1}{b} P_{00}^{*} - a}{b + \frac{1}{b}} \quad ; \quad P_{10} = a + b\, P_{00} \qquad (9)$$



(a) (b)

Figure 1. Definition sketch for correction of transition probabilities P_{ij}.

4. CASE STUDY

The procedures shown in the preceding paragraphs have been applied to a system of 9 gauging stations located in Basilicata (Southern Italy). Names and main characteristics of the basins considered are reported in Table 1. Data records available are highly variable in length and continuity: the longest period of continuous observation in all stations is of 6 years (see Table 1). Given the type of climatic regime, with one summer minimum and one winter maximum of precipitation and runoff, a water year beginning in October has been considered.

Owing to the high percentage of zeros in the effective rainfall series, correlation matrix of the continuous part of I^+ becomes insignificant in some months of the dry season. Consequently, the spatial correlation structure related to I^+ has been evaluated on a seasonal basis and considered to hold for all of the months within each season. Even after aggregation, it was necessary to apply the reconditioning method of Fiering [1968] to obtain positive semidefinite correlation matrices of I^+. This problem does not apply to the transition matrix used for the intermittent part of the process, which can be maintained with its monthly detail.

Summarising, generation of monthly effective rainfall derives from the marginal monthly distribution using constant (within a season) correlation for the continuous part and monthly correlation for the intermittent part of the process.

Final results obtained are shown in Figures 2-3, which relate to reproduction of the correlation structure of the effective rainfall process in the dry and wet seasons. Matrices reproduced in the figure present, in gray scale, the relative variations found between observed and generated correlations between all series pairs. Correlations are computed on complete effective rainfall series (i.e. including zeros) generated at monthly scale and aggregated seasonally.

Table 1. Main characteristics of the time series considered.

Code	*Station*	*Basin Area (Km^2)*	*Mean monthly runoff (mm)*	*Continuous sub-records*
1	Bradano 1	2743	7.60	1933- 1942
				1948- 1960
2	Bradano 2	459	12.50	1928- 1943
				1955- 1971
3	Basento 1	1405	21.67	1948- 1969
4	Basento 2	848	30.13	1927- 1943
				1948- 1966
5	Basento 3	42.4	48.75	1926- 1943
				1947- 1971
6	Agri 1	507	52.40	1926- 1942
				1946- 1962
7	Agri 2	174	69.07	1929- 1943
				1946- 1971
8	Sinni 1	1142	45.04	1950- 1976
9	Sinni 2	233	83.80	1930- 1942
				1948- 1967
				1970- 1980

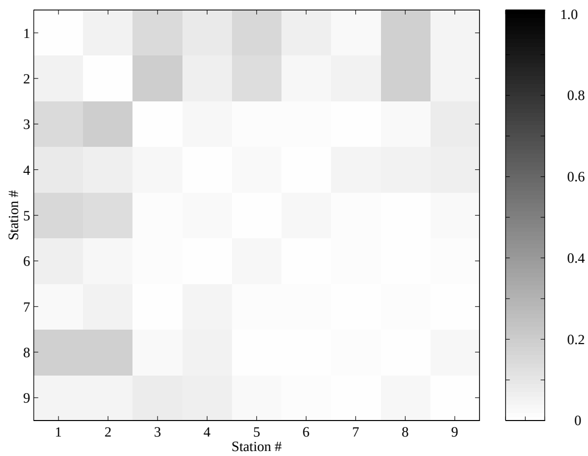


Figure 2. Fraction of the relative errors [abs($\rho_{gen} - \rho_{obs}$)/ρ_{obs}] in the reproduction of the spatial correlation of the effective rainfall in the wet season (November-April).

When considering spatial correlation of runoff, the results (not reported here) look very similar, because univariate linear models fed by spatially correlated inputs produce runoff series which preserve closely the spatial correlation structure embodied in the input process.

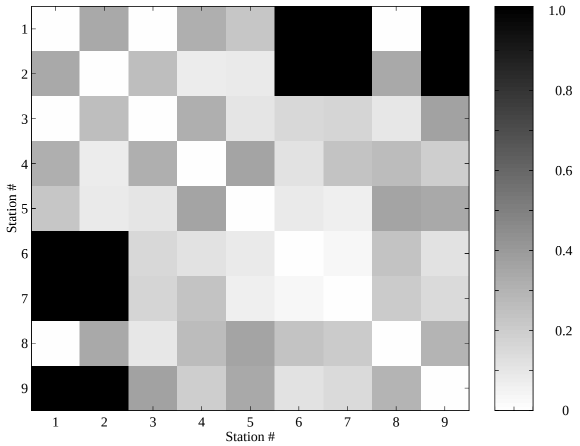


Figure 3. Fraction of the relative errors [abs($\rho_{gen}-\rho_{obs}$)/ρ_{obs}] in the reproduction of the spatial correlation of the effective rainfall in the dry season (May-October).

Overall, the proposed procedure performs quite well, presenting only less efficient results in the dry season with regard to stations located in opposite corners of the region. The inefficient reproduction of correlation in these cases, in terms of relative errors, is mitigated by the fact that for those stations absolute correlations are actually almost negligible in the dry season.

Moreover, the high number of zeros found in the dry season reduces greatly the importance of spatial correlation effects, as compared to the wet season. In fact, summer runoff volumes are only a small fraction of those in the wet season. For this reason, and in the framework of the Mediterranean climate, correct reproduction of the variability and correlation of the runoff process in the wet season must be considered crucial for planning and management purposes.

5. CONCLUSIONS

A procedure based on contemporaneous constant-parameters ARMA models is proposed for multisite generation of monthly runoff. As the conceptual analogy plays a role in all phases of model building [see Claps et al., 1993], it is suggested here to analyze and reproduce the spatial correlation using the estimated effective rainfall in place of the ARMA residual, using that variable as the innovation for multisite runoff generation. The aim is to better understand the variability of net rainfall in the space, and within seasons, and to constitute a base to investigate total to effective rainfall transformations.

Problems arising by this choice have been faced with an adapted technique that reproduces correlation of compound distributions. Further adjustments have been introduced for the analysis of series with gaps, trying to maximize the information available without data reconstruction.

The case study presented here shows that relative estimation errors increase with the increase of zero values and with the reduction of correlation. Future applications in regions with different climates are expected to provide additional indications in this sense. Further steps of model development will include full matrix estimation of the spatial correlation structure of intermittent variables.

6. ACKNOWLEDGEMENTS

The author thanks E. Straziuso, for assistance in computation and sharing of ideas, and M. Fiorentino for fruitful discussions. Funding was granted by GNDCI-CNR and COFIN 2000.

7. REFERENCES

Basson, M.S., R.B.Allen, G.G:S. Pegram, J.A. van Rooyen, *Probabilistic management of water resource and hydropower systems*, Water Resources Publications, 1994.

Chebaane, M., J. D. Salas, and D. C. Boes, Product periodic autoregressive precesses for modeling intermittent monthly streamflows, *Water Resour. Res.*, 31(6), 1513-1318, 1995.

Claps, P., F. Rossi and C. Vitale, Conceptual-stochastic modeling of seasonal runoff using Autoregressive Moving Average models and different time scales of aggregation, *Water Resour. Res.*, 2545-2559, 29(8), 1993.

Fiering M.B., Schemes for handling inconsistent matrices, *Water Resour. Res.*, 4(2), 291-297,1968.

Press, W.H., B.P. Flannery, S.A. Teukolsky, W.T. Wetterling, *Numerical recipes - The art of scientific computing,* Cambridge, 1986

Rasmussen, P.F., J.D. Salas, L. Fagherazzi, J-C. Rassaman and B. Bobée. Estimation and validation of contemporaneous Parma models for streamflow simulation, *Water Resour. Res.*, 32(10), 3151-3160, 1996.

Shafer, J.L., *Analysis of incomplete multivariate data*, Chapman and Hall, London, 1997

Salas, J.D., J.W. Delleur, V. Yevjevich and W.L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resour. Publ., Colorado, 1980.

Salas, J.D., G.Q. Tabios and P. Bartolini, Approaches to multivariate modeling of water resources time series, *Water Resources Bull.*, **21**(4), 683-708, 1985.

Salas, J.D. and J.T.B. Obeysekera, Conceptual basis of seasonal streamflows time series models, *J. Hydraul. Eng.*, ASCE,118, HY8, 1186-1194, 1992.