Modelling processes with inadequate data: The case of multivariate streamflow simulation models

Pierluigi Claps, Eugenio Straziuso

Dipartimento di Ingegneria e Fisica dell'Ambiente, Universita' della Basilicata

C/da Macchia Romana - 85100 Potenza (Italy) - ***e-mail:****claps@unibas.it*

Abstract

Multivariate streamflow simulation models available in the literature seem to have reached a more than acceptable level with regard to the representation of the space-time statistical features of the process.However, using these models with few or patched data prevents one taking advantage of their full potential, not to mention the case in which data need to be generated in ungauged stations. Dealing with inadequate multivariate datasets can require a greater effort than the construction of the simulation model itself. The usual approach of reconstructing missing data prior to model application has been mainly proposed for the precipitation process and can be considered as supported by sufficiently robust statistical methods. However, application of these methods is quite onerous and produces new data that are model-dependent. Moreover, to patch missing data for the runoff process, statistical procedures must be supported by hydrological arguments, owing to the nonlinear dependence between rainfall and runoff and to the serial correlation that affects streamflows. In this paper, a different view is proposed to deal with incomplete multivariate data, based on the goal of maximising the use of the existing information with procedures compatible with the standard application of the simulation model. So, missing data is not reconstructed and application of the different modules of the stochastic model is accompanied by techniques that guarantee feasibility and congruence of the solutions. The model in which these procedures are introduced is a conceptually-based contemporaneous ARMA with periodic-independent residual. Peculiar problems and solutions arise with regard to the characteristics of the residual process, which is treated as an estimate of the effective rainfall and is reproduced by a compound probability distribution. Model application in a 9-station system located in Basilicata (Southern Italy) showed the performances of the procedures proposed.

Introduction

In the framework of water resources systems control and management, multivariate time series models represent an important tool that allows one to simulate the system performance under different hydrologic conditions and operating rules. Models for time series generation of monthly runoff have reached an acceptable level of refinement, both from the viewpoint of the model structure and from the estimation and validation tools. The multivariate PARMA model (*Salas et al.*, 1980, *Salas et al.*, 1985, *Rasmussen et al.*, 1996) presents sufficient generality and flexibility to be considered as the reference case for such kinds of application, even because it can be easily restricted to less general, yet simpler, model forms whenever necessary.

Therefore, notwithstanding the impending overparametrization implicit in the standard form of the PARMA model, the task of generating spatially correlated runoff at the monthly scale when data is fully available and sufficient in number can be considered as technically achieved.

However, in many regions of the world hydrological data sets are not particularly complete, not to mention the cases in which the data is unavailable at all. In cases of extremely poor data, the entire task of simulating streamflow series is nonsensical, but there are many cases in which the historical series are incomplete but not so much to preclude a multivariate analysis. On the other hand, for all the cases in which the quality of available data is poor, models like the multivariate PARMA can present serious problems at a standard application level. In addition, there seems to be no way to apply this model when simulated data in ungauged stations are needed.

In this paper it is shown how a simpler constant-parameter conceptually-based model, proposed in univariate form by *Claps et al.* (1993), when applied in multivariate form can introduce conditions for a reasonable treatment of inadequate samples and also is potentially suitable for dealing with data generation in ungauged basins.

Methods for dealing with inadequate data

Procedures for statistical analysis with missing data are well-established in some applications and less developed in other cases often found in the technical practice. A major distinction is made between univariate and multivariate cases (*Schafer*, 1997), the latter requiring sophisticated statistical techniques. The starting point of these methods is the very basic procedure of the so-called *case deletion*, that allows one to deal with a complete dataset by deletion of all cases in which some observation is missing. It is easy to realise that in multivariate datasets this procedure can lead one to discard a large amount of information. Moreover, a substantial bias is introduced in the sample, because of the systematic differences between the analysed and the observed samples (*Little and Rubin*, 1987). We can also consider that there is a minimum length of the sample required to achieve reliable estimates of the serial correlation structure, and the application of the *case deletion* principle can easily lead to samples of insufficient length for model estimation.

Most of the recent methods proposed in the hydrological literature refer to the principle of *data infilling*, for which updated methods can be found in *Basson et al.* (1994), *Bennis et al.* (1997), *Makhuvha et al.* (1997) and Pegram (1977).

The following considerations arise from the review of literature:

1. Sophisticated and efficient methods can be applied to the estimation of missing hydrological data, even in the multivariate cases, on normal and uncorrelated variables. In many cases precipitation series can approach these conditions.
2. When data is not normal and/or autocorrelated the ensemble of techniques required to face these problems become varied and actual applications become critical.
3. Some approaches to the *repairing* of streamflow data require an hydrological (deterministic or statistical) procedure to save the nonlinear aspects of the rainfall-runoff transformation in the reconstructed series (see *Basson et al.*, 1994, chapt. 3). However, this means that the hydrologic model used to infill the data must be estimated on an incomplete dataset as well.
4. When data are missing in a complex pattern, implementation of ad hoc schemes that preserve the most important aspects of the joint distribution of variables can be very demanding.
5. In any case, patched data must be considered 'not real' when determining measures of uncertainty (standard errors, etc.) with standard, complete-data methods (*Schafer,* 1997). This means that an additional uncertainty resulting from the error associated to the estimates of new data must be accounted for in subsequent risk analyses.

Considering the previously-cited conditions that make data reconstruction difficult, it seems that additional investigations are required to produce practical procedures for dealing with multivariate streamflow records with complex patterns of missing data.

The case study presented in this paper involves a dataset with the above characteristics. Additionally, the treatment of missing data is applied to a non-normal and compound-distributed variable that also represents a problem with regard to the application of consolidated techniques.

The approach presented here tries to answer the missing data problem directly in the simulation model, by devising schemes for objective, yet approximated, estimation of parameters and correlation matrices in the multivariate framework. The advantages of operating in this way are that there is a unique and compact approach to model estimation in a framework of missing data, which does a correct job in the sense of point #5 above and includes the hydrologic phase (point #3) mentioned by *Basson et al.* (1994). Considering that the characteristics of the hydrologic variables of interest are far from being normal, the bases of this approach look sufficiently justified.

Multivariate model choice with inadequate data

As mentioned in the introduction, reconstruction of missing rainfall data requires less effort than that needed to infill or patch series of streamflows, due to the reduced impact of autocorrelation and the absence of hydro-geological considerations that bias the streamflow structure even in adjacent watersheds.

Given that, a multivariate model that must operate in the presence of missing data will perform much better if it allows one to reconstruct the rainfall (or a surrogate of it) independently or in advance with respect to the reconstruction of runoff. The practice of multivariate modelling of seasonal streamflows in fact makes this goal relatively easy to achieve, because it has been widely accepted that models that reproduce independently the serial and the spatial correlation of data (*multisite* or *contemporaneous* models) are fully justified by hydrological considerations and provide more than accurate reproduction of the space-time correlation structure of the process. *Salas et al.* (1985) and *Rasmussen et al.* (1996), among others, provide extensive discussions on the model selection process for the hydrological variable considered with reference to contemporaneous ARMA (CARMA) formulations.

Of the different approaches to treat the estimation of the spatial correlation structure in contemporaneous models (see e.g. *Stedinger et al.*, 1985; *Salas et al.*, 1980; *Rasmussen et al.*, 1996) the most suitable one with respect to requirements in terms of missing data is the one proposed by *Salas et al.* (1980), in which the streamflow spatial correlation matrix is obtained indirectly from the corresponding matrix computed on the residuals. In the case considered by *Salas et al.* (1980) residuals do not have the meaning of hydrological inputs but the rationale sounds correct in the sense discussed earlier. An additional important advantage of this approach (made possible by the decoupling of the estimation of the space and time correlations) is that stochastic models for reproduction of the serial correlation can differ from one station to another. This is a really important matter when there is the need and the possibility of discriminating between the univariate model structure among the series.

When dealing with patched datasets, the multivariate CARMA model allows one to choose whether or not to rebuild data, while other multivariate structures require complete datasets from the beginning. The CARMA model is the one used in this paper, with a variant made up of a conceptually-based framework that modifies the identification and estimation phases and that makes explicit reference to a hydrological input to the watershed system.

Advantages of a conceptually-based contemporaneous model

Claps et al. (1993) proposed a conceptually-based framework for identification of univariate constant-parameter ARMA models of monthly runoff. One of the peculiar features of that model is the parsimony of parameters achieved by considering independent sources of monthly correlation, such as deep groundwater and seasonal groundwater systems, whose effects are evaluated at independent aggregation scales. Model identification and estimation are strictly conditioned by this structure. In particular, in the identification phase the only option is related to the recognition of the presence of a deep (over-year) groundwater component, that leads to a more general ARMA(2,2) model instead of an ARMA(1,1). Another feature of the conceptual framework is that the model residual is formally identified with the effective rainfall, that results from being inversely estimated once the model is identified. To maintain this correspondence, no transformation or deseasonalisation is applied to the data, such that a non-gaussian periodic-independent residual is obtained (PIR-ARMA model).

The difficulties arising in the probabilistic modelling of the residual are compensated for by the possibility of validating parameter estimates that have conceptual meaning and by the possibility of allowing streamflow generation in ungauged sites. The first issue is particularly useful when data are scarce and uncertain. Conditions that are necessary for the second task, with reference to an ungauged station are: a) The effective rainfall must be estimated in adjacent basins and relations with the total rainfall must be established; b) ARMA parameter values are to be related to the estimates in adjacent basins by means of relations with hydro-geological features of the basins.

The extension of the conceptually-based approach to the multisite case, as proposed by *Straziuso* (1997) and *Straziuso et al.* (1998), presents the model preserving the advantages just discussed, with the additional feature that when residuals are considered in their spatially correlated structure this is reinforced by their conceptual meaning. In other words, it is more effective to look for the spatial correlation in the effective rainfall process than in the runoff, which is highly affected by serial correlation.

In the following, the process of generating multisite series of monthly runoff will be reduced to the task of generating spatially correlated residuals, or effective rainfall, leaving as a trivial final step the final reproduction of serially correlated streamflows in each site with the specific ARMA stochastic model.

Multivariate process of the effective rainfall

In the conceptually-based approach the effective rainfall process behaves as the ARMA model residual. The advantages of introducing a conceptual meaning for the residual are partly counterbalanced by the disadvantage of having to work with its compound (and skewed) marginal distribution. The Bessel distribution originally introduced to model this variable (*Claps et al.*, 1993) was replaced by a more treatable compound square-root normal (*csr-normal*), which has probability density function as:

$$f_I(i) = P(0) = P_0; \qquad\qquad i = 0 \tag{1}$$

$$f_I(i) = (1-P_0)\frac{1}{2r\sigma_r\sqrt{2\pi}}\exp\left[\frac{1}{2\sigma_r^2}\left(r-\mu_r\right)^2\right]; \qquad r = \sqrt{i} \qquad\qquad i > 0$$

This distribution behaves well with respect to the Bessel (*Straziuso,* 1997) particularly when the parameters are estimated with the method of moments directly applied to the transformed form of the normal dostribution (*Lloyd*, 1980, p. 154):

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

These relations apply to the continuous part I^+ of the distribution, while the zero finite probability *P(0)* coincides with the sample frequency of zeros.

Reproduction of the spatial correlation structure of this variable requires that the continuous part (or the whole distribution) be reduced to a normal distributed variable e_t so that the classical scheme for generation of normal correlated variables (see e.g. *Salas et al.*, 1980) can be applied. This scheme is based on the matrix equation $\boldsymbol{e_t} = \mathbf{B}\ \boldsymbol{\xi_t}$, in which $\boldsymbol{\xi_t}$ is a vector of normal uncorrelated standardised values corresponding to the innovation for the n stations at time *t*, while **B** is the n×n matrix carrying the information content of the spatial correlation. In fact, this latter is explicitly contained in the correlation matrix **G** with which **B** is connected by means of the *gramian* equation $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$.

A conformal transformation between the csr-normal and a gaussian distribution was tried by *Straziuso* (1997), based on some results given by *Bell* (1987) on analytical and numerical methods to transform the correlation matrix estimated on the gaussian (transformed) variable in the correlation matrix relative to the untransformed variable. In this attempt, the compound nature of the distribution did not provoke specific problems in the application of the transformation. What made this method practically inapplicable, because of the poor results obtained, was the extreme variability of the P(0) values between months and between stations.

A more complex, yet more natural, mechanism for reproduction of the spatial correlation for the csr-normal was selected using the properties of correlated intermittent processes, considered in the family of Discrete AR and ARMA models (*Jacobs and Lewis,* 1978). In particular we looked at an evolution of this family, as the product Periodic DAR(1) model developed by *Chebaane et al.* (1995) to reproduce intermittent monthly flows processes.

The PDAR(1) model by *Chebaane et al.* (1995) reproduces zero-non zero sequences of a univariate correlated intermittent process, and is proposed for application in the time domain. We have used a non periodic formulation of this scheme in a spatial domain, that starts from the same vector equation (product model):

$$\mathbf{Y}_\tau = \mathbf{X}_\tau \, \mathbf{Z}_\tau \qquad (3)$$

in which **Y** is the vector at time τ of the process to reproduce, **X** is a Bernoulli stochastic process of (0,1) occurrences and **Z** is the vector of the continuous part of the distribution.

Evaluation of the correlation structure of the continuous part is coherent with the general case cited above (*Salas et al.*, 1980) while for the intermittent process *Chebaane et al.* (1995) provide the estimation tools with respect to a formulation for a 2-site process that is equivalent to:

$$X_{s_1,\tau} = V_\tau \cdot X_{s_0,\tau} + (1 - V_\tau) U_\tau \qquad (4)$$

In this relation the correlation is evaluated between the value $X_{S_0,\tau}$ of the variable X at month τ in the 'independent' station s_0 and the value $X_{S_1,\tau}$ at the same time in the 'dependent' station s_1.

V_τ and U_τ are mutually independent binary processes with Binomial distribution and parameters:

$$\gamma_\tau = P\{V_\tau = 1\} \qquad \delta_\tau = P\{U_\tau = 1\}; \qquad (5)$$

with $0 \le \gamma_\tau \le 1$ and $0 \le \delta_\tau \le 1$ while $\eta_{s_0,\tau} = P\{X_{s_0,\tau} = 1\}$ and $\eta_{s1,\tau} = P\{X_{s_1,\tau} = 1\}$ characterise the process $\mathbf{X}_{S,\tau}$.

Using the transition matrix estimation method reported in *Chebaane et al.* (1995), one obtains the dependence of γ_τ and δ_τ on the conditional probabilities $P_{ij} = P(X_{s_1,\tau} = j | X_{s_0,\tau} = i)$:

$$\gamma_\tau = P_{00} + P_{11} - 1 \qquad \delta_\tau = \frac{[1 - P_{00}]}{\{2 - [P_{00} + P_{11}]\}} \qquad (6)$$

in which P_{00} and P_{11} are obtained directly on the sample, with reference to the month τ, by counting the numbers n_{00} and n_{11} of the actual 0→0 and 1→1 transitions relative to the total number n_0 and n_1 of starting states.

Chebaane et al. (1995) introduced the product model with reference to a periodic univariate AR(1) model, in which the intermittent scheme is sequentially applied to couples of consecutive months. The transposition of this scheme in a spatial correlation framework leads to a bivariate contemporaneous model, that is not immediately extendible to the general multivariate case. On the other hand, the bivariate form does not prevent one from obtaining realistic results, given that simulation of contemporaneous streamflow data must obey some constraints imposed by the parental relations existing between river basins drained by the river sections of interest. In the case study examined, the presence of nested and adjacent basins allowed us to establish reasonable sequences of couples of stations (upstream in one basin and continuing towards adjacent basins).

Application of different sequences of couples showed that the final correlation matrices resemble each other very well, independently of choice of sequence. Even the correlation values resulting for pairs not included in a sequence does not change significantly if the pair is subsequently included in a different sequence. This result derives from the substantial homogeneity of the spatial field of the effective rainfall process and is not guaranteed in other, more variable, contexts.

Correlation matrix of effective rainfall estimation with patched data

Continuous part

In the framework of multivariate contemporaneous models the handling of missing data without reconstruction requires two distinct approaches within the estimation of serial and spatial correlations. For serial correlation we need to estimate unique parameter vectors for discontinuous series. Since we apply univariate linear stochastic models to the continuous subsets of the series, final values of the parameters can be obtained by weighted averages of the subset estimates, with the record length as the weight. These parameters are used in the generation step, while it is important to underline that the effective rainfall series are estimated independently for each subset and are not modified when the final stochastic parameter estimates are obtained. Therefore, at the end of the univariate model estimation on all of the record subsets we have a patched matrix of estimated effective rainfall data, that contains also a significant fraction of zero values. Referring to the product model of equation (3), evaluation of the spatial correlation structure is independently performed for the continuous and for the intermittent parts. In the following we will start by evaluating techniques for robust estimation of the correlation matrix of a continuous process in the presence of uneven and/or intermittent datasets.

In the literature we found two main methods for dealing with patched matrices in estimating the spatial correlation. The first was referred to as *case deletion* and works so as only the *cases* 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 allow one to solve the gramian equation $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$ and generate spatially correlated random numbers (*Salas et al.,* 1980). The problem with this simple and intuitive technique is that the resulting even dataset is frequently too short to retain enough information about the spatial correlation structure.

The second method found was proposed by *Basson et al.*(1994 pp.163-164) and is more effective in reproducing the original correlations between couples of stations. In this method the *case deletion* is only applied between each pair of stations, reducing to a minimum extent the number of deleted data. This means that 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 certain 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 semidefinite, a *reconditioning* method must be applied to make it at least positive semidefinite.

This latter method ensures reasonable preservation of sample spatial correlation even when data are quite sparse, and is certainly coherent with the approach presented here, which avoids infilling the data when the gaps are large and systematic.

The choice of the reconditioning method is from a restricted lot. The methods found in literature were those proposed by *Fiering* (1968), *Crosby and Maddock* (1970), *Mejia and Millàn* (1974) and *Rasmussen et al.* (1996). Each of the related techniques has advantages in terms of practicability and drawbacks in terms of alteration of the original correlation figures. We compared the performances of the algorithms by *Fiering* and by *Rasmussen et al.* The former is very simple and intuitive, and so preferable to the other two methods presented in the seventies, and the latter is part of a wider method to recondition families of correlation matrices within the framework of periodic contemporaneous models. For the data considered here, the method by *Fiering* performed best, and was the one adopted throughout the procedure. The technique consists of modifying the eigenvalue matrix $\mathbf{\Lambda}$ by setting to zero the lowest (negative) value and distributing the error uniformly on the other eigenvalues, leaving their sum unchanged. The only constraint on the recomputed correlation matrix $\mathbf{G}_1=\mathbf{\Theta}^{\mathbf{T}}\mathbf{\Lambda}_1\mathbf{\Theta}$, with $\mathbf{\Theta}$ as the eigenvector matrix, will be that the main diagonal must have unit values. After having imposed this constraint, eigenvalues and eigenvectors are again computed until all eigenvalues are all non-negative. As a result of application of this procedure one obtains a positive semidefinite matrix **G** that is only slightly different from the original matrix.

Decomposition of the gramian equation $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$ is achieved through the Singular Value Decomposition (SVD) method (also called *eigenvalues* method) originally coded by *Wilkinson and Reinsch* (1971) and largely accessible in numerical methods books (*e.g. Press et al., 1986*). This methods is computationally efficient and does not fail with zero and negative eigenvalues. The accuracy resulting from this method when large matrices are considered depends essentially on the algorithm used for computation of eigenvalues and eigenvectors. In the cases in which **G** is positive definite, decomposition of the above equation can result also by the Gram-Schmidt orthogonalization procedure, first introduced by *Young and Pisano* (1968) for this factorization problem. This latter algorithm presents computational instabilities for large matrices, consequently being substantially less appealing than the SVD for practical applications, in which positive semidefinite **G** matrices are commonly encountered.

Having chosen the method by *Basson et al.* (1994) and the reconditioning technique by *Fiering* (1968) the generation of spatially correlated non-zero values of effective rainfall proceeds with the following steps:

1. Estimation of the sample correlation matrix **G** from the square-root transformed (non-zero) values of estimated inputs
2. If **G** is negative-definite, the reconditioning method by Fiering is applied.
3. Equation $\mathbf{BB}^{\mathrm{T}} = \mathbf{G}$ is decomposed by the SVD method, implemented in the Matlab© environment (*The Mathworks,* 1997).
4. Once **B** is computed, generated values of the continuous part of the inputs result from equation $\boldsymbol{e_t} = \mathbf{B}\ \boldsymbol{\xi_t}$, in which $\boldsymbol{e}_t$ vectors are squared to respect the original distribution.

Intermittent part

Parameter estimation for the intermittent part of the effective rainfall process is obtained through evaluation of transition probabilities computed between couples of stations. In this case the presence of uneven datasets does not affect the nature of the transition matrix, but rather the inner congruence of probabilities. In the following it is shown how this problem has been addressed.

Some congruence equations hold in the scheme for estimation of the intermittent process parameters by means of the transition matrix. They are:

$$P_{00} + P_{01} = 1\ ; \qquad P_{10} + P_{11} = 1 \tag{7}$$

and

$$(P_0)_{s_1,\tau} = P_{00}(P_0)_{s_0,\tau} + P_{10}(P_1)_{s_0,\tau} \tag{8}$$

$$(P_1)_{s_1,\tau} = P_{01}(P_0)_{s_0,\tau} + P_{11}(P_1)_{s_0,\tau} \tag{9}$$

in which $(P_0)_{S1,\tau}$ is the marginal probability of zero effective rain for the month τ and the station S_1. Relations (8) and (9) derive from the following equation, applied to a generic month τ:

$$(P_0)_{S1} = \frac{(n_0)_{S1}}{N} = \frac{(n_{00})_{S1} + (n_{10})_{S1}}{N} = \frac{(n_{00})_{S1}}{(n_0)_{S0}}\frac{(n_0)_{S0}}{N} + \frac{(n_{10})_{S1}}{(n_1)_{S0}}\frac{(n_1)_{S0}}{N} = P_{00}(P_0)_{S0} + P_{10}(P_1)_{S0} \tag{10}$$

where, in this case, the probabilities are interpreted as maximum likelihood estimates based on observed frequencies.

When dealing with missing data, application of this scheme can produce problems related to the total number of data in the two stations. Based on the streamflow data, we have two relatively independently observed datasets, with lengths N_0 and N_1 , on which we compute the numbers n_{ij} of actual transitions on the two subsets of contemporaneous data, of length $N \leq min(N_0,N_1)$. The reduction of the length of original series can produce significant modifications of the marginal probabilities (P_0) and (P_1), resulting in the estimation of model parameters that will not allow one to reproduce the (0,1) marginal occurrences in the individual stations. Conversely, it is not possible to estimate transitions without reducing the data to the contemporaneous observations. Doing so, we cannot use directly the original marginal values of n_0 and n_1 in evaluating P_{ij}, because this will prevent the fulfilment of the congruence in relations (8) and (9).

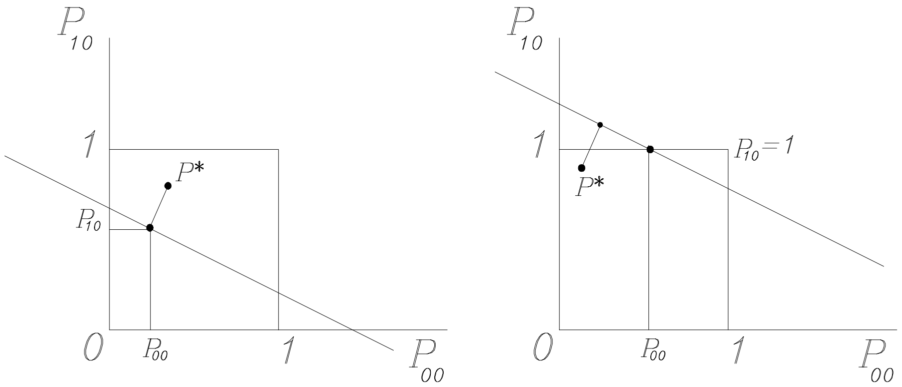
What we propose for saving most of the original information contained in the marginal zero and 1 probability is to impose the congruence, as represented by relations (8) and (9), on the estimated transition probabilities. In fact, one can consider that relation (8) is a linear equation in the (P_{00}, P_{10}) plane, that can be written as $P_{10} = a + bP_{00}$ with coefficients:

$$b = -\frac{(P_0)_{S0,\tau}}{(P_1)_{S0,\tau}}\ ; \qquad a = \frac{(P_0)_{S1,\tau}}{(P_1)_{S0,\tau}} \tag{11}$$

If we indicate with P* the point in the (P_{00}, P_{10}) plane corresponding to the sample transition probabilities $\mathrm{P}^*_{00}, \mathrm{P}^*_{10}$, in general this point will not lie on the line corresponding to the equation written above, because of the *case deletion* applied. The correction proposed, that preserves the marginal probabilities, consists in moving the point P* orthogonally with respect to the congruence line until it has reached the line itself (see figure 1a). The rationale is to minimise the distance between the originally estimated point and the corrected point, coherently with respect to the conditions that both P_{00} and P_{10} must lie in the interval (0,1). If this is not the case, such as when the interception point falls outside the (1,1) congruence square, the new point will still move on the line until this condition is met (figure 1b).

Congruence of the ($P_{00,}$ P_{10}) pair with relations (8) and (9) is obtained with the expressions:

$$\mathrm{P}_{00} = \frac{\mathrm{P}_{10}^{*} + \frac{1}{b}\mathrm{P}_{00}^{*} - a}{b + \frac{1}{b}} \; ; \qquad \mathrm{P}_{10} = a + b\,\mathrm{P}_{00} \qquad (12)$$



(a) (b)

Figure 1. *Definition sketch for correction of transition probabilities.*

Case study

The procedures shown in the preceding paragraphs have been applied to a real world case, consisting of a system of 9 gauging stations located in the Basilicata region (Southern Italy). Names and main characteristics of the basins considered are reported in table 1. The dataset available for the stations are highly variable in length and continuity and the maximum number of years in which the 9 stations have contemporaneous data is 6 (see figure 2).

Owing to the high percentage of zeros in the effective rainfall series, in some months of the dry season the correlation matrix of the continuous part of the input distribution has little significance, because the number of contemporaneous non-zero values can be very low. This problem does not apply to the transition matrix used for the intermittent part of the process. As a consequence of this outcome, the spatial correlation structure related to the continuous part of the distribution was evaluated on a seasonal basis, and was considered constant for all of the months within each of the two seasons. By contrast, the marginal probability distributions and the correlation structure of the intermittent part of the process were maintained with monthly detail.

Even on a seasonal basis it was necessary to apply the reconditioning method of *Fiering et al.* to make the correlation matrices of the continuous part of the input positive semidefinite. Table 2 reports the differences found between empirical and reconditioned matrices for the dry season (May-October). Station numbers are the codes shown in Table 1.

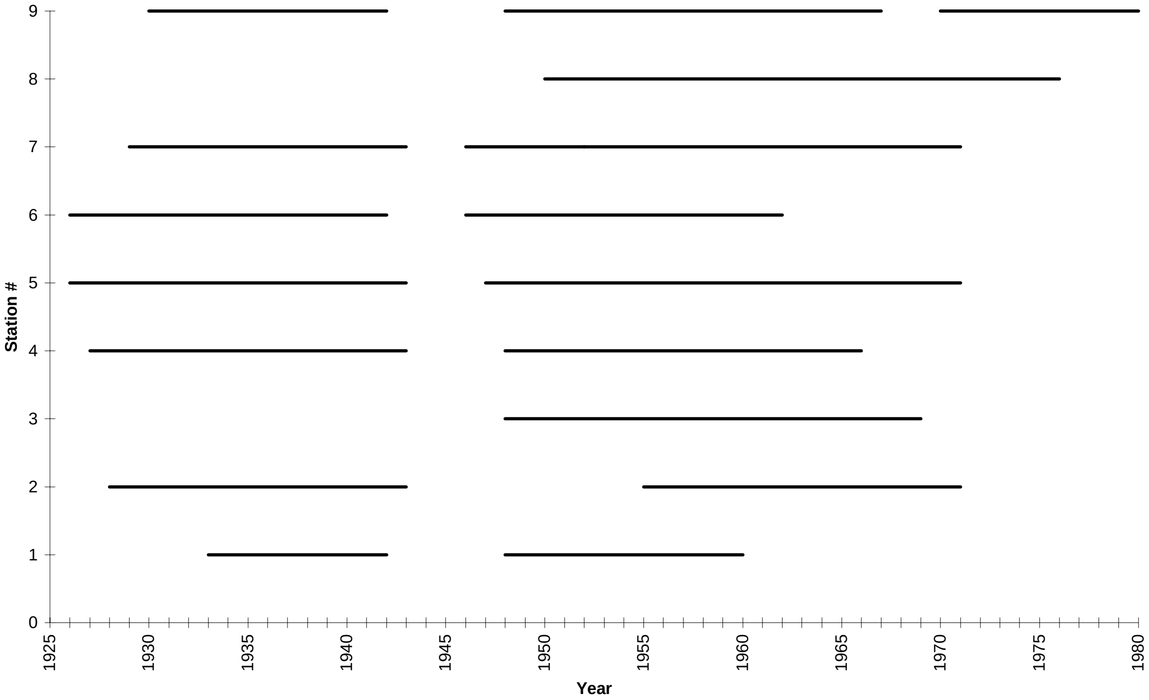


Figure 2. *Graphical representation of the pattern of contemporaneous data analysed.*

Code	*Station*	*Area* *(Km^2)*	*Mean monthly runoff (mm)*
1	Bradano a Tavole Palatine	2743	7.60
2	Bradano a Ponte Colonna	459	12.50
3	Basento a Menzena	1405	21.67
4	Basento a Gallipoli	848	30.13
5	Basento a Pignola	42.4	48.75
6	Agri a Tarangelo	507	52.40
7	Agri a Le Tempe	174	69.07
8	Sinni a Valsinni	1142	45.04
9	Sinni a Pizzutello	233	83.80

Table 1. *Main characteristics of the gauging stations considered.*

As an example of application of the correction method for transition probabilities, the results of the corrections introduced in the transition matrix computed for the month of October are reported in Table 3.

The final results obtained in terms of reproduction of the correlation structure of the effective rainfall process are shown in figures 3-4. The matrices reproduced in the figure present in grey scale the relative variations found between observed and generated correlations between series at all the stations pairs. The series considered are those of the complete effective rainfall process (including the zeros) aggregated on the two seasons chosen, wet and dry. The quality of results presented for the effective rainfall was also obtained with respect to the runoff process. In the runoff, the reproduction of the spatial correlation depends on the efficiency of representation of the serial correlation, which is quite well reproduced in the univariate scheme

Corrected

Station #	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	1.00	0.81	0.80	0.29	0.36	0.14	0.01	0.65	0.01
2	0.81	1.00	0.69	0.66	0.36	-0.16	0.03	0.52	0.00
3	0.80	0.69	1.00	0.56	0.58	0.46	0.46	0.55	0.31
4	0.29	0.66	0.56	1.00	0.64	0.23	0.63	0.43	0.45
5	0.36	0.36	0.58	0.64	1.00	0.60	0.61	0.34	0.34
6	0.14	-0.16	0.46	0.23	0.60	1.00	0.81	0.50	0.64
7	0.01	0.03	0.46	0.63	0.61	0.81	1.00	0.47	0.59
8	0.65	0.52	0.55	0.43	0.34	0.50	0.47	1.00	0.41
9	0.01	0.00	0.31	0.45	0.34	0.64	0.59	0.41	1.00

Observed

Station #	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	1.00	0.91	0.80	0.24	0.35	0.21	0.13	0.64	0.07
2	0.91	1.00	0.74	0.78	0.39	-0.30	0.17	0.59	0.08
3	0.80	0.74	1.00	0.54	0.57	0.49	0.49	0.52	0.32
4	0.24	0.78	0.54	1.00	0.63	0.28	0.66	0.40	0.45
5	0.35	0.39	0.57	0.63	1.00	0.63	0.61	0.33	0.33
6	0.21	-0.30	0.49	0.28	0.63	1.00	0.81	0.54	0.65
7	0.13	0.17	0.49	0.66	0.61	0.81	1.00	0.49	0.58
8	0.64	0.59	0.52	0.40	0.33	0.54	0.49	1.00	0.42
9	0.07	0.08	0.32	0.45	0.33	0.65	0.58	0.42	1.00

Table 2. *Corrected (Fiering algorithm) and observed interstation correlations for the continuous part of the effective rainfall distribution. Dry season (May-October) .*

Corrected $\mathbf{P}_{00}$

Station #	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0.286	1.000	0.023	0.545	0.167	0.189	0.090	0.074	0.085
2	0.511	0.452	0.099	0.521	0.108	0.169	0.057	0.194	0.054
3	0.033	0.541	0.095	0.027	0.006	0.011	0.003	0.501	0.003
4	0.415	1.000	0.024	0.235	0.121	0.138	0.000	0.046	0.000
5	0.508	1.000	0.003	0.502	0.049	0.501	0.500	0.004	0.500
6	0.344	0.679	0.007	0.335	0.335	0.094	0.274	0.017	0.260
7	1.000	1.000	0.001	0.002	1.000	1.000	0.026	0.002	0.951
8	0.059	0.578	0.254	0.049	0.010	0.020	0.005	0.154	0.005
9	1.000	1.000	0.002	0.003	1.000	1.000	1.000	0.003	0.024

Observed $\mathbf{P}_{00}$

Station #	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0.286	1.000	0.000	0.500	0.167	0.167	0.167	0.000	0.167
2	0.429	0.452	0.071	0.571	0.143	0.143	0.071	0.143	0.071
3	0.000	0.500	0.095	0.000	0.000	0.000	0.000	0.500	0.000
4	0.375	1.000	0.000	0.235	0.125	0.125	0.000	0.000	0.000
5	0.500	1.000	0.000	0.500	0.049	0.500	0.500	0.000	0.500
6	0.333	0.667	0.000	0.333	0.333	0.094	0.333	0.000	0.333
7	1.000	1.000	0.000	0.000	1.000	1.000	0.026	0.000	1.000
8	0.000	0.500	0.250	0.000	0.000	0.000	0.000	0.154	0.000
9	1.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.024

Table 3. *Corrected and observed probabilities for the 0→0 transitions.*

The proposed procedure performs well under almost all conditions, with less efficient results in the dry season and with stations located in the extreme positions of the area. The inefficient reproduction of the correlation in these cases, in terms of relative errors, is mitigated by the fact that absolute correlation is almost negligible in the dry season for those stations. In addition, the high number of zeros found in the dry season makes spatial correlation effects less important than in the wet season, because the runoff volumes in the dry season are a fraction of those in the wet season, which dominate the annual flow distribution

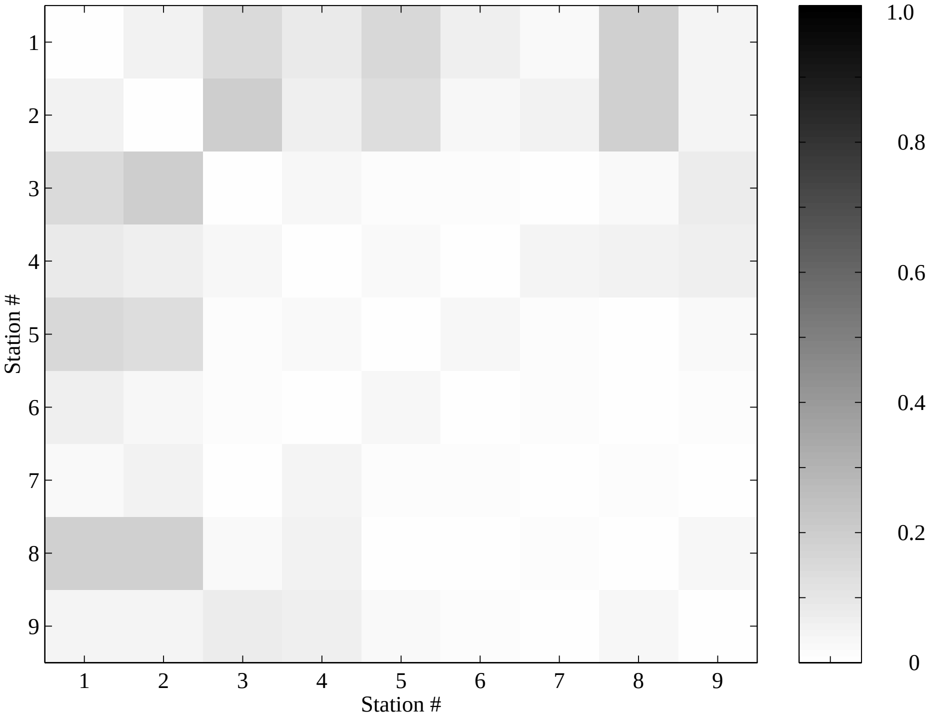


Figure 3. *Percent of the relative errors [abs(ρ_{gen}–ρ_{obs})/ρ_{obs}] in the reproduction of the spatial correlation of the effective rainfall in the wet season (November-April).*

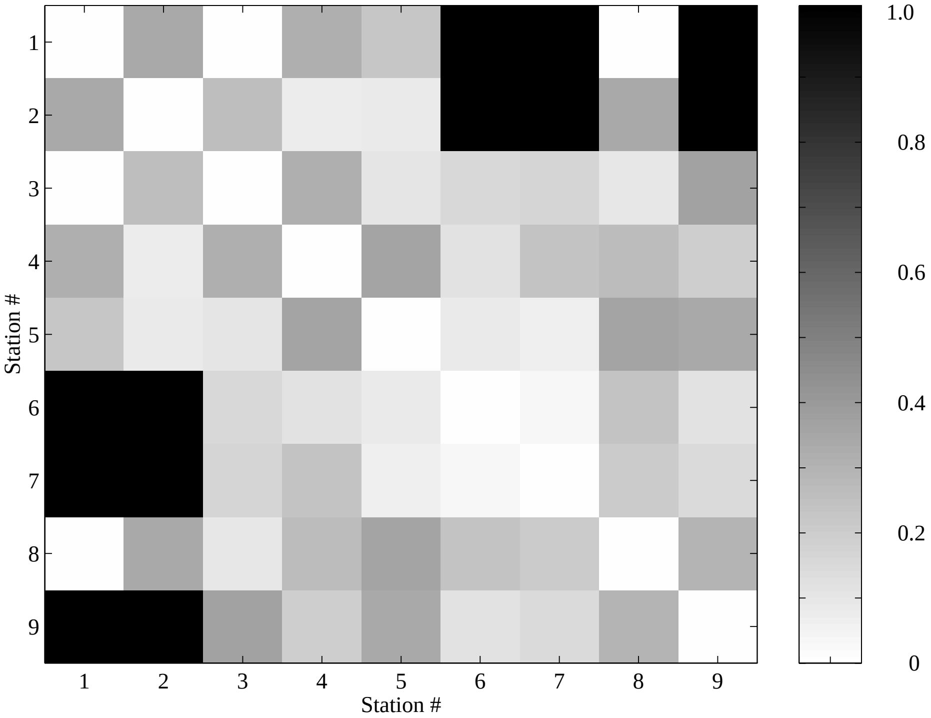


Figure 4. *Percent of the relative errors [abs(ρ_{gen}–ρ_{obs})/ρ_{obs}] in the reproduction of the spatial correlation of the effective rainfall in the dry season (May-October).*

Final Remarks

In the framework of the conceptually-based analysis of multivariate monthly streamflow data, the issue of model application with incomplete datasets is addressed in this paper. The main features of the procedure presented here are: (i) the attempt to make the best use of the existing data without reconstructing unobserved runoff records and (ii) the specific points related to the modelling of intermittent processes with incomplete data. Missing records are treated in a different way with reference to the reproduction of serial correlation and of spatial correlation; this is made possible by the use of the reduction of the complete multivariate model to a multisite formulation.

As regards the reproduction of the spatial structure of the process, the stochastic model residual is considered and its physically-consistent meaning as an estimate of the effective rainfall provides useful indications on the results arising from the applications. This is especially true with reference to the relations between the number of zeros in the input (effective rainfall) sequences and interstation correlation.

The practical implementation of the proposed procedure suggests the following considerations:

- The use of a conceptually-based modelling framework helps in the interpretation of results and in the reduction of errors. In this context, relative errors emerging in the reproduction of correlation in the dry season are to be ascribed to the peculiar characteristics of the input process in that season and are certainly less important than comparable errors that may arise in the wet season.
- To maintain the monthly detail in all of the modules of the stochastic procedures looks difficult at the application level and is probably not always completely justified, particularly if one is interested in finding physically consistent indications in the results. This was the case of the spatial correlation in the continuous part of the input distribution, that sometimes cannot be computed for lack of relevant data and that is characterised by low values and insignificant variations between the consecutive dry months.

The ensemble of procedures proposed tends to address an objective way of treating incomplete multivariate datasets, producing even more meaningful results, considering the physically-consistent meaning of the residual analysed. We believe that the results obtained provide an interesting contribution to the discussion on a topic of great practical relevance in multivariate runoff simulation in the context of incomplete data.

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.

Bell, T. L, A Space-time stochastic model of rainfall for satellite remote sensing studies, *J. Geophys. Res.*, 92(D8), 9631-9643, 1987.

Bennis, S, F. Berrada and N, Kang, Improving single variable and multivariable techniques for estimating missing hydrological data, *Jour. Hydrol.*, 191, 87-105,1997

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

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

Crosby O.S. and T. Maddock, Estimating coefficients of a flow generator for monotone samples of data, *Water Resour. Res.*, 6 (**4**): 1079-1086, 1970.

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

Jacobs, P. A. and Lewis, P. A. W. Discrete time series generated by mixtures, 1. Correlation and run properties, *J. R. Stat. Soc. B.* , 40 (1),94-105,1978.

Little, J.A. and D.B. Rubin, *Statistical analysis with missing data*, Wiley series in probability and mathematical statistics, New York, 1987.

Lloyd, E.H., *Probability*, in: W. Ledermann (Eds.): *Handbook of Applicable Mathematics* Vol. II, John Wiley and Sons, New York,1980.

Makhuvha T., G.G.S. Pegram, R.S. Sparks,W.S Zucchini, Patching rainfall data using regression methods, Parts 1 and 2, *Journal of Hydrology*, 198, 289-318, 1997.

Mejia J.M. and J. Millàn, Una metodologìa para tratar el problema de matrices inconsistentes en la generatiòn multivariada de series hydrològicas, *VI congr. Lat. Am. de Hidràulica,* Bogotà, Colombia, 1974.

Pegram G.G.S., Patching rainfall data using regression methods. 3. Grouping, patching and outlier detection. *Journal of Hydrology*, 198, 319-344, 1997.

Press, W.H., B.P. Flannery, S.A. Teukolsky, W.T. Wetterling, *Numerical recipes - The art of scientific computing,* Cambridge Univ. Press, 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 Resources Research*, 32(10), 3151-3160, 1996.

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

Stedinger, J. R., Lettenmaier, D. P. and Vogel, R. M.. Multisite ARMA(1,1) and Disaggregation Models for Annual Streamflow Generation, *Water Resources Research*, 21(4), 497-509, 1985.

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

Straziuso E., *Multivariate analysis of river flows*, PhD dissertation, Università della Basilicata, 1997 (in Italian).

Straziuso E., P. Claps, M. Fiorentino, A model for generation of contemporaneous streamflows at a monthly scale, Proc. XXVI Conf. in Hydraulics and Hydraulic Structures, Catania (Italy), Vol. II, 365-376, 1998 (in Italian).

Wilkinson J.H and C. Reinsch (eds.), *Handbook of automatic computation*, Vol. 2, Springer-Verlag, 1971

Young, G. K. and Pisano, W. C., Operational Hydrology Using Residuals, *J. Hydraul. Div., ASCE (*now *J. Hydraul. Eng.),* 94 (HY4), 1968.

The Mathworks, *Matlab* ver. 5, Natick, MA - USA- 1997.