Hydrology and Earth System Sciences, 6(4), *641–654* (2002) © EGS

Hydrology & Earth System Sciences

Multivariate synthetic streamflow generation using a hybrid model based on artificial neural networks

J.C. Ochoa-Rivera, R. García-Bartual and J. Andreu

Department of Hydraulic and Environmental Engineering, Universidad Politécnica de Valencia, Camino de Vera s/n, 46071 - Valencia, Spain

E-mail for corresponding author: jochoa@upvnet.upv.es

Abstract

A model for multivariate streamflow generation is presented, based on a multilayer feedforward neural network. The structure of the model results from two components, the neural network (NN) deterministic component and a random component which is assumed to be normally distributed. It is from this second component that the model achieves the ability to incorporate effectively the uncertainty associated with hydrological processes, making it valuable as a practical tool for synthetic generation of streamflow series. The NN topology and the corresponding analytical explicit formulation of the model are described in detail. The model is calibrated with a series of monthly inflows to two reservoir sites located in the Tagus River basin (Spain), while validation is performed through estimation of a set of statistics that is relevant for water resources systems planning and management. Among others, drought and storage statistics are computed and compared for both the synthetic and historical series. The performance of the NN-based model was compared to that of a standard autoregressive AR(2) model. Results show that NN represents a promising modelling alternative for simulation purposes, with interesting potential in the context of water resources systems management and optimisation.

Keywords: neural networks, perceptron multilayer, error backpropagation, hydrological scenario generation, multivariate time-series.

Introduction

It has been almost four decades since the initial contributions of time series analysis in hydrology and water resources were made. Since then the field has been nurtured by continuous theoretical improvements and applications to practical water resources problems, particularly for hydrological simulation and forecasting. Simulation and forecasting techniques of streamflow time-series allow practitioners and planners to explore possible realistic future scenarios of a given water resources system, thus helping in the decision-making process. Such techniques have been applied widely for determining the dimensions of hydraulic works for risk asessment in urban water supply systems and irrigation, optimal operation of reservoir systems, planning of new works and actions to optimise hydroelectric production, and others (Salas *et al*., 1985; Koutsoyiannis, 2000). Time-series prediction has been used in real-time operation, systems operation during drought periods, short term operation strategies in reservoirs and also for flood warning purposes (Salas *et al*., 2000).

It is well known that water resources systems simulation using only historical records of precipitation, discharge or both, introduces severe restrictions. For example, Loucks *et al*. (1981) remark the limited range of designs or alternative strategies that result when applying only historical data to simulate the future behaviour of a water resources system, while better operation rules and designs are obtained when they are tested with a variety of generated hydrological scenarios. Bras and Rodríguez-Iturbe (1985) emphasise the random nature of hydrological inputs, with a large degree of variability and uncertainty and state that, using only historical data as inputs to a water resources system, results in a scarcely documented response. Different designs, operational rules and strategies can be tested adequately in a more efficient and realistic way through the diversity of conditions resulting from synthetic series, as the past experience in operational and synthetic hydrology has extensively documented.

The research presented here fits into a broader project, in which a decision support system (DSS) (Andreu *et al*., 1996) has been used to estimate the risks of drought in the Tagus river basin. This is the longest river in the Iberian Peninsula. In the methodology applied, an essential phase is the generation of multiple future hydrological scenarios spanning several months in the future (between 24 and 60, depending on the basin); these are conditioned to the hydrological situation at the moment of the inquiry to the DSS. These multiple future scenarios are used to simulate the management of the water resource in the basin; the results are used in turn to estimate the statistics of future deficits (i.e. means, probabilities and distribution functions) and statistics of state variables (e.g. reservoir storage). Most of the rivers in the Iberian Peninsula experience droughts that last for several months. Hence, it is crucial that the synthetically generated future scenarios reproduce closely the statistics related to drought and storage. Presently, the DSS allows for synthetic data generation by means of classical stochastic models, such as autoregressive moving average (ARMA) models. The present work explores the possibilities of using neural networks (NN) as generators of future scenarios, with emphasis on the ability to reproduce the statistics related to drought and storage. This application relates to a study on a sub-basin of the Tagus river basin involving two reservoirs, the Entrepeñas with a capacity of 1639 hm^3 that regulates the river Guadiela, a tributary of the Tagus, and the Buendía, in the main course, with a capacity of 803 hm^3. The location of the reservoirs is indicated in Fig. 1. Both reservoirs are used for irrigation, production of hydroelectricity, and urban water supply (CHT, 1999). If the results of this study show that NN are useful, the work can be extended to the other sub-basins, where multivariate modelling with more than two sites will be necessary.

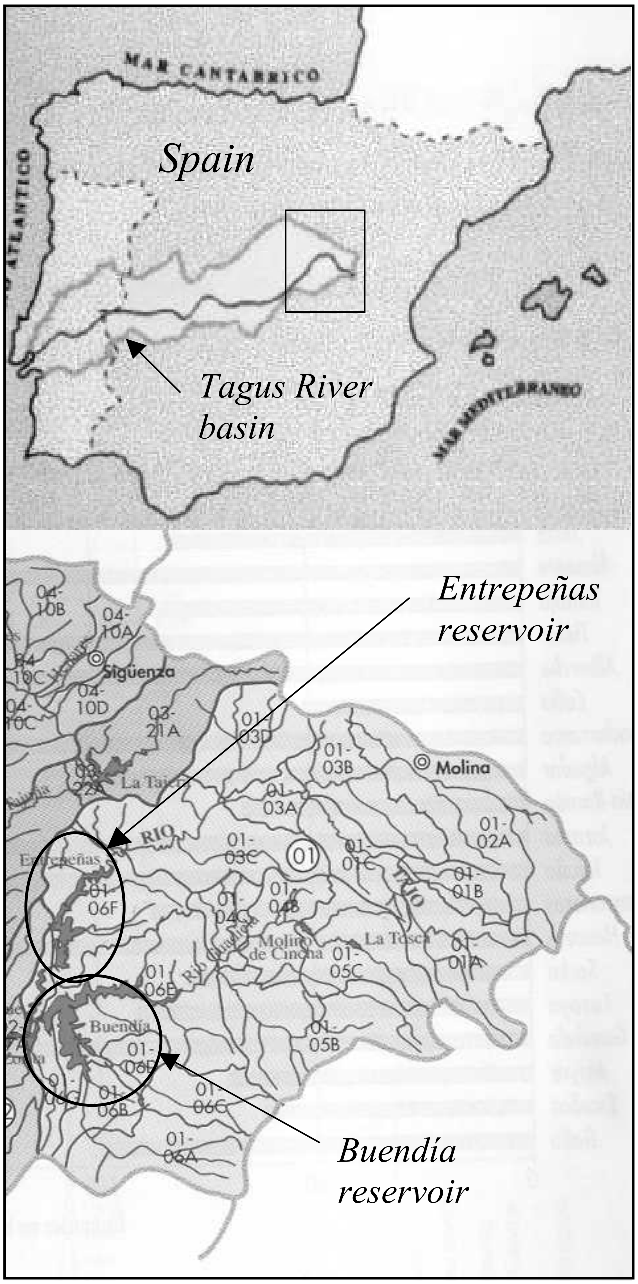

Fig. 1. *Location of Entrepeñas and Buendía reservoirs in the Tagus River basin (Spain)*

The present study uses an artificial NN approach for nonlinear modelling of multivariate streamflow time-series. This technique is used for synthetic generation of monthly inflows to the two reservoirs. The research follows the lines initiated by several authors in the past, some of whom compared, in practical case studies, linear stochastic models with NN models, while others used mixed models, i.e. NN plus a random noise. Lachtermacher and Fuller (1994) modelled annual streamflow series using multi-layer feed-forward error-backpropagation NN, with iterated multi-step prediction, where the single output of the model was used for subsequent forecasting. Then, a Box-Jenkins modelling approach was used to determine the appropriate number of inputs (previous values of past streamflows) in the NN. Raman and Sunilkumar (1995) built twelve different NNs, one for each month of the year, which were then used for streamflow generation for two reservoir sites. They compared the technique with results derived from a bivariate AR(2) model. The authors report a better performance of the NNs approach. Boogard *et al.* (1998) proposed a hybrid model that included an NN and an ARMAX model and applied it to predict the depth of a lake. Zealand *et al.* (1999) used a one-hidden-layer NN for short-term streamflow prediction; the model performed better than a stochastic-deterministic catchment model. Anmala *et al.* (2000) used recurrent NNs to predict streamflows in three different basins and the results were an improvement on empirical approaches. Salas *et al*. (2000) forecast streamflows with horizonts from one to four months using several NNs, all trained with the error-backpropagation learning algorithm, with successful results overall. Deo and Thirumalaiah (2000) tested different learning algorithms to train multi-layer feed-forward networks for time-series modelling of hourly discharges

The most extended techniques for synthetic streamflow generation and streamflow forecasting include simple and multiple linear regression, autoregressive moving average (ARMA) models, ARMA with exogenous variables (ARMAX) and ARMA and ARMAX models with periodic parameters. In all cases, a linear relationship between the relevant hydrological variables is assumed; this does not always yield the best results, and sometimes is even inadequate (Chakraborty *et al.,* 1992; Lachtermacher and Fuller, 1994; Raman and Sunilkumar, 1995; Lehtokangas *et al*., 1996). Classical nonlinear approaches typically require large amounts of exogenous information, which is not always available (Deo and Thirumalaiah, 2000). Some nonlinear and non-Gaussian techniques do not need exogenous information and behave better than linear models, as in the case of periodic gamma autoregressive processes PGAR (Fernandez and Salas, 1986), but they are univariate models. Different authors (Lapedes and Farber, 1988; Tang *et al*., 1991; Zealand *et al*., 1999; Imrie *et al*., 2000; and Salas *et al.,* 2000) have tested the capability of certain NN topologies to incorporate complex and non-linear hydrological relationships; they remark on their potentials and abilities as tools for hydrological forecasting.

The NN approach used here, the most widely referred to in the literature, is based on the well known one-hidden-layer fee-forward architecture trained with the error-backpropagation learning algorithm. The model developed has a deterministic component (NN), in addition to a normally distributed random noise, which takes into account the uncertainty typically affecting hydrological processes. The model is applied to generate monthly streamflow series which in turn are applied to real-time operation of the water resources system as mentioned above.

Data preprocessing

The data used in this research comprised two series of monthly inflows to the Buendía and Entrepeñas reservoirs for the period October 1940 to September 1993 (i.e. 53 years). The modelling technique applied is essentially data-driven, i.e. there is no preconceived notion about the existing relationships between the variables. For efficient operation of such models, previous work on data-conditioning, normalising, and scaling of the variables must be undertaken.

As a first step, skewness was removed from the original records by the transformation given by

$$X_{v\tau} = \log(Q_{v\tau} + c_\tau \overline{Q}_\tau) \tag{1}$$

with

$$c_\tau = a / g_\tau^2 \tag{2}$$

where $Q_{v\tau}$ is the monthly inflow (hm^3 month^{-1}) for month τ (τ=1,...,12) and year v (v=1,...,N_a); N_a is the number of years of record of the series; $\overline{Q}_\tau$ is the monthly average inflow for month τ; a is a dimensionless parameter of value 0.35 resulting from a regression analysis between g_t and c_τ; g_τ is the skewness coefficient for the set $Q_{1\tau}$, $Q_{2\tau}$, ..., $Q_{Na\tau}$; and $X_{v\tau}$ is the normalised inflow, for year v and month τ.

Equation (1) is the modified log-transformation suggested by Raman and Sunilkumar (1995), who used a single c_τ value, i.e. the same value for every month. But, in the present study, the optimal reduction in skewness was achieved using a different value for each month, as Eqn. (2) indicates. This equation was obtained by regression analysis of the monthly skewness coefficients, g_τ and their corresponding values of c_τ.

To account for periodicity, the resulting series after transformation of Eqn. (1) were standardised to improve learning efficiency and overall operation of the NN (Salas *et al*., 2000). The standardisation was applied on a monthly basis, through equation

$$Y_{v\tau} = \frac{X_{v\tau} - \overline{X}_\tau}{s_\tau} \tag{3}$$

where $\overline{X}_\tau$, s_τ are the sample mean and standard deviation of the normalised inflows for month τ; and $Y_{v\tau}$ is the standardised value for month τ and year v.

Finally, an additional transformation was performed on the Y series for a convenient scale of the data to be processed by the NN. The range of variation was reduced to the interval [0,1].

$$Z_t = \frac{Y_t - Y_m}{Y_M - Y_m} \tag{4}$$

being $Y_\tau = Y_{v\tau}$ with $t = 12(v\text{-}1)+\tau$; Y_M, the maximum value of the Y series; Y_m, the minimum value of the Y series; and Z_t, the scaled value to be presented to the NN. This procedure helps to avoid internal numerical instabilities of the NN learning process and operation (Lapedes and Farber, 1988; ASCE-TCAANNH, 2000).

Artificial neural network modelling

TOPOLOGY

The scheme employed is the one-hidden-layer feed-forward NN, trained with the popular error-backpropagation learning algorithm (Rumelhart *et al.*, 1986). This NN configuration has been used successfully in many water resources systems applications and shows generally good abilities to model hydrological time series (Chakraborty *et al.,* 1992; Gupta *et al*., 2000; ASCE-TCAANNH, 2000).

For the Buendía and Entrepeñas reservoirs monthly inflow time series, the input and output nodes of the network are set so that the target values to be predicted are the immediate next-month inflows, given the two past values of the series.This number of past values derives from a previous exploratory phase, in which several numbers of past values (lag-intervals) were tested. When the number of lag-intervals were higher than two lags, no significant improvement of the model performance was achieved. Note that each new lag-interval introduces twice the number of hidden nodes as the number of parameters in the NN. In this way, if the relation *improvement of the model performance/number of additional parameters* is considered for each number of lag-intervals tested, two lags are found to be the most appropriate number in this case study.

The inference is done simultaneously for the two sites, and therefore, a single network topology is used, as illustrated in Fig. 2. This approach implies a unique underlying non linear multivariate function involving two dependent variables, i.e., the one-step monthly inflow forecast for each of the reservoirs. An alternative to this procedure would be to design and train two independent networks, one for each of the sites. The former strategy has been adopted in this research, in favour of a simpler and more compact structure, being also the natural candidate for the interconnected hydrological system under consideration. On the other hand, the use of one single network in this case facilitates the overall formulation of the proposed hybrid model, providing an adequate framework for comparison to the traditional tools as the multivariate autoregressive models. Consequently, each *training exemplar*, or *pattern*, presented to the network consists of a *predictor section* with the input values [$x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$, $x_4^{(0)}$], and a *criterion section* with the target values [$y_1^{(2)}$, $y_2^{(2)}$]. As indicated in Fig. 1, a small number of nodes in the hidden layer was adopted as a result of optimal dimensioning of the NN after some numerical experiments. Optimal network geometry is a highly dependent problem and no general procedures or theories have been established (Maier and Dandy., 2000). In this study, only one-hidden-layer achitectures were tested, since sufficient degrees of freedom can always be provided by changing the number of nodes in the hidden layer (Hornik *et al*., 1989). The approach followed here to determine the optimal network dimension was to begin with the simplest possible architecture containing a single hidden unit, and train the network. Then, the number of hidden nodes was increased and the network retrained, repeating this stagewise process until no significant improvement in network performance, in terms of its predictive capabilities, was obtained. After this trial-and-error procedure, the best results were obtained with the network indicated in Fig. 1 and this was finally adopted.

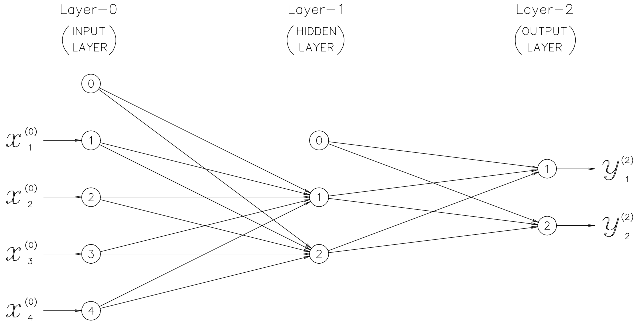

Fig. 2. *Artificial neural network topology*

FORMULATION

As usual, the segments in Fig. 1 connecting nodes between consecutive layers are associated with weights: $w_{ij}^{(l)}(n)$ will indicate the weight for the connection between node *i* of the layer *l*-1 and and node *j* of the layer l. For the one-hidden-layer topology, *l*=1, 2. The net incoming value to a node, also known as the *post synaptic potential* (*PSP*), is calculated as

$$v_j^{(l)} = \sum_{i=0}^{m_{l-1}} w_{ij}^{(l)} y_i^{(l-1)}; \quad l=1,2 \tag{5}$$

where $y_i^{(l-1)}$ is the output from node *i* in layer *l*-1; and $m_{l\text{-}1}$ is the number of nodes in layer *l*-1.

In this linear combination, the values $y_0^{(l)}$ are taken equal to 1, that is to say, the upper node or node 0 (Fig. 1) of the input layer and hidden layer are *fictitious nodes* producing unity as output, introducing a bias term. This makes weights $w_{0j}^{(l)}$ act as independent terms in the linear combination of Eqn. (5). They are usually referred to as *threshold parameters* for the node *j* in layer *l*.

The *activation function* in each node is a non-linear function transforming the PSP into an output or *activation value*,

$$y_j^{(l)} = f(v_j^{(l)}) \tag{6}$$

where $y_j^{(l)}$ is the output from node j (j=1, 2, ..., m_l) in layer l (l=1, 2), and $f(\ )$ is the non-linear operation performed in the node using a certain activation function.

In this study, *sigmoid functions* were used as activation functions, with output ranges [0,1] and [–1,1], given respectively by

$$f(v) = \frac{1}{1+\mathrm{e}^{-v}} \tag{7}$$

and

$$f(v) = \frac{2}{1+\mathrm{e}^{-v}} - 1 \tag{8}$$

These sigmoids or activation functions are very attractive since they are easily handled in the training process of the NN (ASCE-TCAANNH, 2000). The final results were obtained with the *bipolar sigmoid* of Eqn. (8), which gave the fastest training and lower prediction errors. Other recent applications (e.g. Zealand *et al.*, 1999; Salas *et al.*, 2000) also used Eqn. (8) successfully for the activation function of these NN architectures.

The nodes in the input layer, or layer 0, do not perform any operation. They are passive nodes, just allowing presentation of the inputs (x's values) to the network. Therefore, $y_j^{(0)} = x_j^{(0)}$ $(j = 1,...,4)$, while $y_0^{(0)}$ is as stated previously.

The normalised, standardised and scaled values of monthly inflows (Z series) are taken as inputs and target outputs of the NN:

$$\begin{aligned} x_1^{(0)} &= Z_{t-2}(b) \quad x_2^{(0)} = Z_{t-2}(e) \\ x_3^{(0)} &= Z_{t-1}(b) \quad x_4^{(0)} = Z_{t-1}(e) \\ y_1^{(2)} &= Z_t(b) \quad\ \ y_2^{(2)} = Z_t(e) \end{aligned} \tag{9}$$

where (b) stands for Buendía reservoir and (e) is for Entrepeñas reservoir.

The overall operation of the NN is a non-linear black box that transforms $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, x_4^{(0)}$ into $y_1^{(2)}, y_2^{(2)}$, through a non-linear function which results from the successive application of the single activation functions of the nodes in consecutive layers. It may be described analytically as shown below.

The final outputs of the NN are the activation values of nodes or *artificial neurons* in the output layer, given by

$$y_r^{(2)} = \frac{2}{1+\exp\left[-\sum_{q=0}^{2} w_{qr}^{(2)} y_q^{(1)}\right]} - 1, \quad r = 1, 2 \tag{10}$$

in which, $w_{qr}^{(2)}$ are the weights associated with segments between the hidden and output layer; and $y_q^{(1)}$ are the activation values of hidden nodes, given by

$$y_q^{(2)} = \frac{2}{1+\exp\left[-\sum_{p=0}^{4} w_{pq}^{(1)} y_p^{(0)}\right]} - 1, \quad q = 1, 2 \tag{11}$$

where

$$y_0^{(0)} = 1 \quad \text{and} \quad y_p^{(0)} = x_p^{(0)}, \quad \text{for} \quad p = 1,...,4$$

The resulting non-linear function transforming $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, x_4^{(0)}$ into $y_1^{(2)}, y_2^{(2)}$ can be obtained by combining Eqns. (10) and (11):

$$\begin{aligned} y_r^{(2)} &= 2\{1+\exp[-\sum_{q=0}^{2} w_{qr}^{(2)} (2[1 \\ &\quad + \exp(-\sum_{p=0}^{4} w_{pq}^{(1)} x_p^{(0)})]^{-1} - 1)]\}^{-1} - 1, \\ &\quad r = 1, 2 \end{aligned} \tag{12}$$

Therefore, the proposed NN can be interpreted as a non-linear regression (Stern, 1996), with the weights $w_{ij}^{(l)}$ playing the role of parameters to be estimated.

From this perspective, the problem of estimating $w_{ij}^{(l)}$ parameters can be regarded as a non-linear optimisation problem without restrictions. A convenient objective function can be the mean squared error function (MSE), which is given by

$$MSE = \frac{1}{2N_p} \sum_{n=1}^{N_p} \sum_{j=1}^{2} [d_j(n) - y_j^{(2)}(n)]^2 \tag{13}$$

in which N_p is the number of patterns or training exemplars shown to the network. In this study, N_p=634, extracted from the 53 years of monthly records; $y_j^{(2)}(n)$ are the output values produced by the NN (Eqn. 12), when the predictors $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, x_4^{(0)}$ of exemplar n are processed; $d_j(n)$, with j=1, 2, are the target values specified in the criterion section of the exemplars, that is, the values observed subsequently. In the present study, those target values are those observed subsequently in the Z series for the Buendía and Entrepeñas reservoirs.

TRAINING

The popular error-backpropagation algorithm was used to train the proposed NN. The method is not as fast as other techniques, and sometimes requires slow training sessions to achieve convergence (Cheng and Titterington, 1994; Stern, 1996). This is not a serious limitation when the NN architecture is not too complex, as in this case, while there are other benefits such as the robustness of the method.

The training process must aim to find optimal values of synaptic weights $w_{ij}^{(l)}$, producing minimum differences between target values and predicted or calculated ones, as indicated by

$$e_j(n) = d_j(n) - y_j^{(2)}(n), \quad j = 1,\ 2; \quad n = 1, 2, \ldots, N_p \tag{14}$$

where $e_j(n)$ for j=1, 2 are the errors for exemplar n. Note that $e_j(n)$ is included in Eqn. (13).

The training process was applied in a sequential mode or *exemplar-mode training*, so that the set of weights was modified successively after each exemplar processing as

$$w_{jk}^{(l)}(n+1) = w_{jk}^{(l)}(n) + \Delta w_{jk}^{(l)}(n) \tag{15}$$

where $\Delta w_{jk}^{(l)}(n)$ is the corresponding weight change. This procedure is very simple to implement and provides effective solutions in most cases (Haykin, 1999). To ensure effective learning by the network, the training patterns were presented to the network in a randomised way. Weight changes $\Delta w_{jk}^{(l)}(n)$ are evaluated by a gradient descent method; this, basically, calculates sensitivity of the current training error to changes in each of the network weights modifying it according to

$$\Delta w_{jk}^{(l)}(n) = \alpha[\Delta w_{jk}^{(l)}(n-1)] + \eta \delta_k^{(l)}(n) y_j^{(l-1)}(n) \tag{16}$$

where α is the momentum constant; η is the learning rate; and $\delta_k^{(l)}(n)$ is the local gradient for node k in layer l. Details of the procedure can be found in Haykin (1999). In the case study presented here, best training resulted from values α = 0 and η = 0.03. During each individual training session, α values (from 0 to 1) were kept constant, while η was progressively decreased from 0.5 at the beginning of training, to 0.01, to accelerate convergence during initial stages of the process and to avoid, in the final steps, damping of the error function trajectory around the minimum. To avoid local minima, the training procedure was repeated from independent initial conditions concerning weight values. The numerical experiments showed that changing the value of α affected the total training time only slightly but, in this case, resulted in no significant differences in the final weights of the calibrated network.

Table 1. Computed statistics for residual series

Statistics	*Buendía*		*Entrepeñas*	
	NN	*AR(2)*	*NN*	*AR(2)*
Mean	0.006	–0.004	0.005	–0.005
Std. Dv.	0.574	0.575	0.545	0.546
Skewness Coef.	0.654	0.654	0.594	0.582
Maximum	2.039	2.023	2.183	2.243
Minimum	–1.526	–1.642	–1.644	–1.878

NN with random component

In all cases the training was stopped after the error function showed neglegible decreases, with no restrictions to the number of epochs needed. The training process of the neural network was carried out using the generalised software package SERENA. This latter has been developed by the authors of this study as a part of a broader project.

Once the network was trained, a statistical analysis of prediction errors was performed over the Y series, i.e. normalised and standardised series after Eqn. (3). Figure 3 shows such prediction errors or *observed residual series*, ε_t, for both reservoirs. Table 1 summarises statistics computed for both residual series.

While the mean value is almost 0, as expected, a certain positive skewness of residuals is obtained in both cases. Figure 4 shows the corresponding frequency histograms.

Finally, the autocorrelation functions of the residual series and month-to-month correlations were computed, (Figs. 5 and 6). While the autocorrelation values and those of month-to-month correlations were negligible, a value of 0.861 for cross-correlation is obtained between the two residual series.

To build a model for synthetic generation of monthly inflows in the Buendía and Entrepeñas reservoirs, a *white noise generator*, ξ_t, was considered. ξ_t is a normally distributed and uncorrelated random signal with zero mean and standard deviation equal to 1. Then, the random component for both reservoir sites, ε'_t, is defined by

$$\{\varepsilon'\}_t = \mathbf{B}\{\xi\}_t \tag{17}$$

with

$$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \Sigma \tag{18}$$

where Σ is the matrix of variances (σ_b^2, σ_e^2) and covariances (σ_{be}, σ_{eb}) of the observed residual series, as indicated in expression (19). As Σ is the Gramian matrix of **B**, this last one is unknown and has to be found by solving the matrix Eqn. (18).

$$\Sigma = \begin{bmatrix} \sigma_b^2 & \sigma_{be} \\ \sigma_{eb} & \sigma_e^2 \end{bmatrix} \tag{19}$$

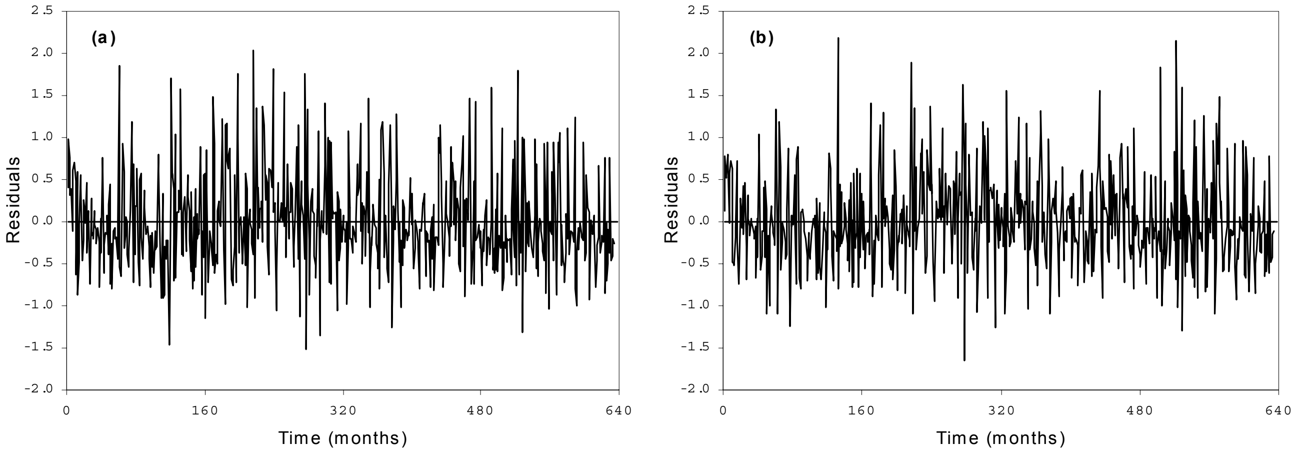

Fig. 3. *Residual series. (a) Buendía (b) Entrepeñas*

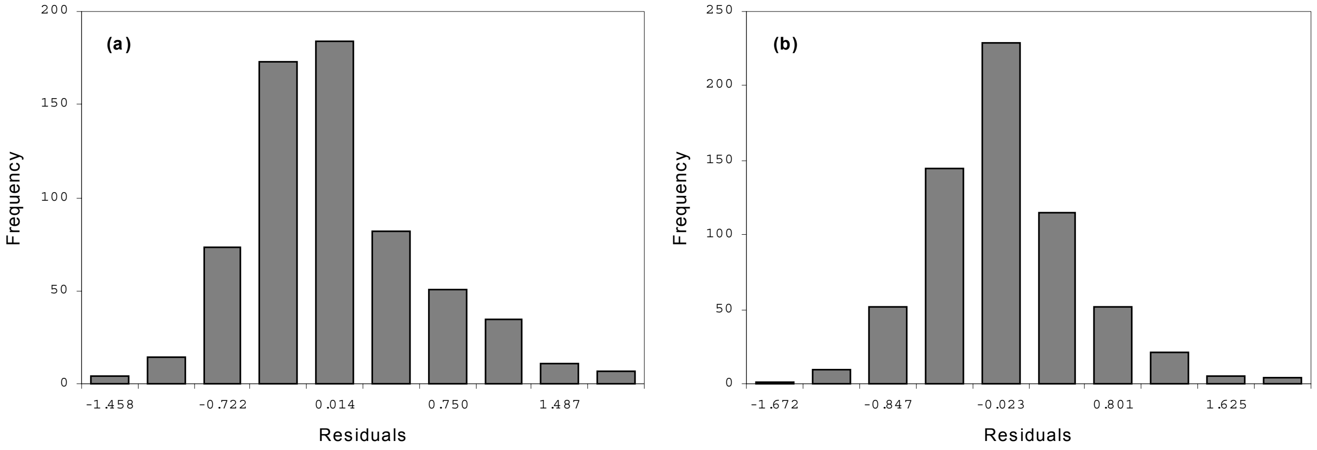

Fig.3. *Frequency histograms of residual series. (a) Buendía (b) Entrepeñas*

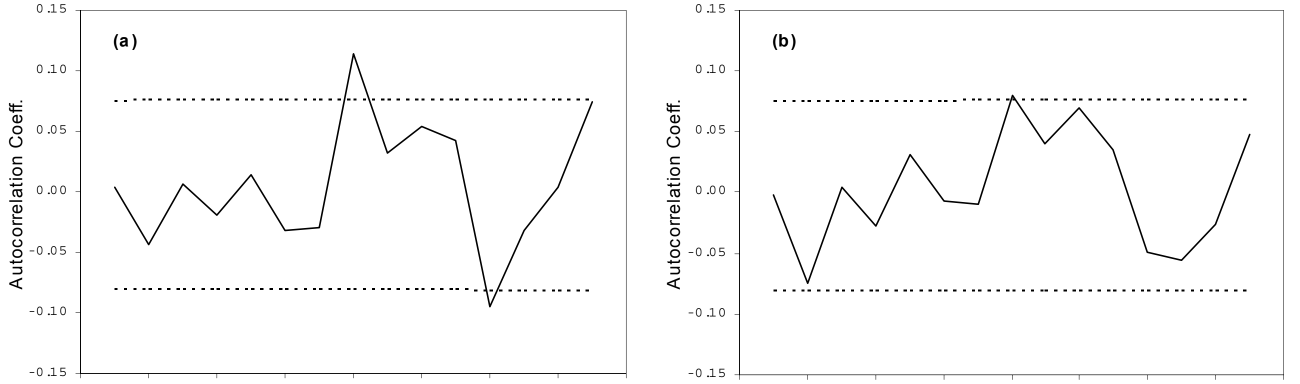

Fig. 4. *Autocorrelation functions of residual series. (a) Buendía (b) Entrepeñas*

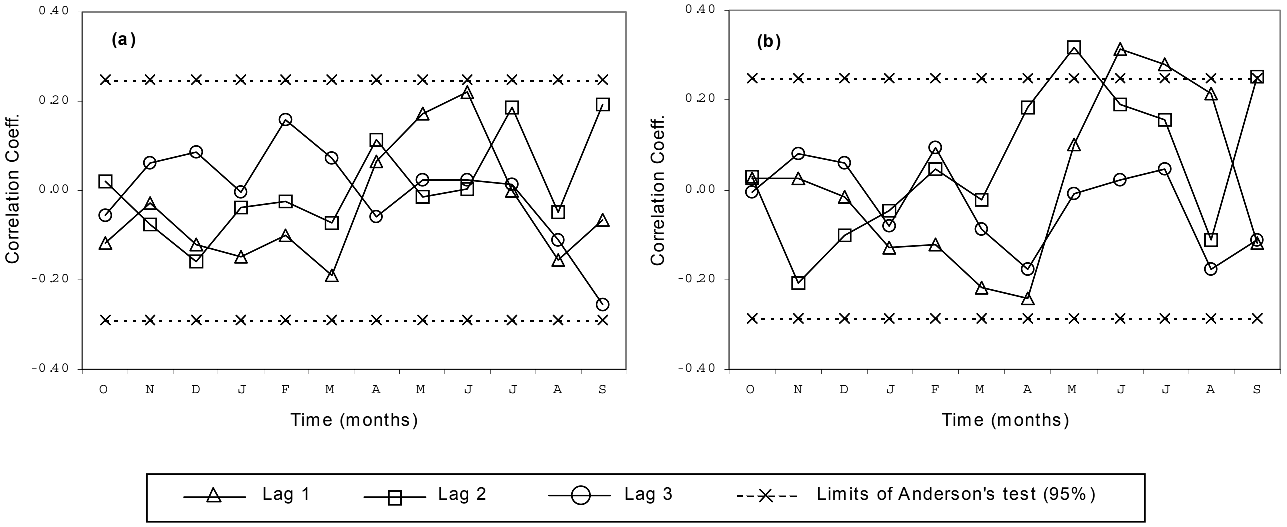

Fig. 6. *Month-to-month correlations of residual series. (a) Buendía (b) Entrepeñas*

Equations (17) and (18) result from the assumption of residuals distributed as a bivariate normal distribution.

Taking **B** as a lower triangular matrix, and forcing Σ to be a positive semi-definite matrix, Eqn. (18) has a unique solution. This technique is documented in detail in Bras and Rodríguez-Iturbe (1985).

Equation (17) defines the two stochastic components to be added to the previously formulated NN-deterministic approach, and can be re-written as

$$\begin{Bmatrix} \varepsilon_t^{'b} \\ \varepsilon_t^{'e} \end{Bmatrix} = \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix} \begin{Bmatrix} \xi_t^b \\ \xi_t^e \end{Bmatrix} \tag{20}$$

The two components are to be assembled over the normalised and standardised series (*Y* series). Consequently, the NN component, as given in a compiled form by Eqn. (12), needs to be de-scaled first. The final form of the model is the sum of both components, as given by

$$\begin{Bmatrix} Q_t^{'b} \\ Q_t^{'e} \end{Bmatrix} = F(\begin{Bmatrix} Y_t^{'b} \\ Y_t^{'e} \end{Bmatrix} + \begin{Bmatrix} \varepsilon_t^{'b} \\ \varepsilon_t^{'e} \end{Bmatrix}) \tag{21}$$

where $Q_t^{'b}$ and $Q_t^{'e}$ are the synthetic values produced by the model (hm^3 month^{-1}), $Y_t^{'b}$ and $Y_t^{'e}$ are the values produced by the NN scheme (including de-scaling after Eqn. (12)), and $\varepsilon_t^{'b}$ and $\varepsilon_t^{'e}$ are the corresponding stochastic components for the Buendía and Entrepeñas reservoirs, given by Eqn. (20). Function *F* represents the inverse of the preprocessing operations defined by Eqns. (3) and (1) respectively, that is,

$$X'_{v\tau} = (Y'_{v\tau} + \varepsilon'_{v\tau})s_\tau + \overline{X}_\tau \tag{22}$$

$$Q'_{v\tau} = 10^{X'_{v\tau}} - c_\tau \overline{Q}_\tau \tag{23}$$

with τ = t–12(v–1).

The proposed scheme is a hybrid stochastic-deterministic model for hydrological scenario synthetic generation, in terms of monthly series of reservoir inflows.

Multivariate statistical model

For comparison purposes, an autoregressive model of order 2, AR(2), was applied to the data series of Buendía and Entrepeñas. Model AR(2) represents the time dependence of a value Y_t of a period *t* as a function of the two previous values Y_{t-1} and Y_{t-2} corresponding to periods *t*-1 and *t*-2. Its formulation is given by

$$\{Y\}_t = \Phi_1\{Y\}_{t-1} + \Phi_2\{Y\}_{t-2} + \Theta_0\{\xi\}_t \tag{24}$$

being Y_t stationary time series normally distributed. In this study, those series were taken from the set of values stated in Eqn. (3), i.e. normalised and standardised inflows to the Buendía and Entrepeñas reservoirs. ξ_t is a random signal,

Table 2. Estimated values of AR(2) model parameters

Component	*Matrix Φ_1*	*Matrix Φ_2*	*Matrix Θ_0*
1,1	0.664	0.163	0.587
1,2	0.093	–0.097	0.000
2,1	0.030	–0.038	0.483
2,2	0.667	0.193	0.278

which is also normally distributed with mean zero and variance one. $\mathbf{\Phi}_1$, $\mathbf{\Phi}_2$ and $\mathbf{\Theta}_0$ are parameter matrices, which were estimated by the method of moments (Salas *et al.*, 1980}. Since the month-to-month correlations of the standardised inflow series are not significant statistically, as Fig. 7 indicates, parameter matrices were assumed to be constant.

The AR(2) modelling process consisted of two steps: firstly, the model was calibrated and, then, it was used for generating synthetic streamflow series in both reservoirs. Table 2 presents estimated values of AR(2) model parameters, and Table 1 shows the statistics of the residual series, which are very similar to those of the NN modelling. In particular, it can be seen that skewness coefficients of the residual series from both models are practically the same, showing that non-linear processing of the NN model does not induce any skew.

Evaluation of models' performance

The two multivariate models, after calibration, were used for synthetic generation of a total of 200 synthetic series of monthly inflows at both geographical sites, each series 53 years in length. All the series, synthetic and historical, were also aggregated using a time level of aggregation of one year.

For evaluation of the models' performance, certain relevant statistics were computed from both historical and synthetic series and then compared. This comparison is the right way to validate a model if it is intended for water resources systems planning and management management and planning (Jackson, 1975; Salas *et al.*, 1980; Stedinger and Taylor, 1982; Fernandez and Salas, 1986; Kendal and Dracup, 1991; Basson and van Rooyen, 2001). Therefore, considering that the proposed model is not a forecasting approach, the performance of the NN-based model must not be evaluated by using the classical procedure of splitting all available data into training and validation (and/or) test sets. The statistics were calculated over the monthly and annual series, and then *verification* and *validation* processes were done. As Stedinger and Taylor (1982) report, these are two important stages of the development and use of a stochastic streamflow model. Verification consists of demonstrating that the statistics explicitly involved in the model formulation are statistically the same for both generated and historical flows; validation of a streamflow model is the demonstration that such a model is capable of reproducing statistics which are not explicitly included in its formulation. In this study, verification statistics (means, standard deviations, lag-1 correlations and lag-2 correlations) were included in addition to some relevant validation statistics related to droughts and storage of the series. The computed statistics were grouped into four different categories:

- *Basic statistics*. Mean, standard deviation and skewness coefficient, all computed over the series in their original scale (hm^3 month^{-1}).

- *Series persistence statistics*. These statistics are related to autocorrelation and cross-correlation functions. Both were computed over the normalised and standardised series.

- *Drought statistics*. A given percentage of the mean discharge (Q_M) is taken as a threshold, so that each group

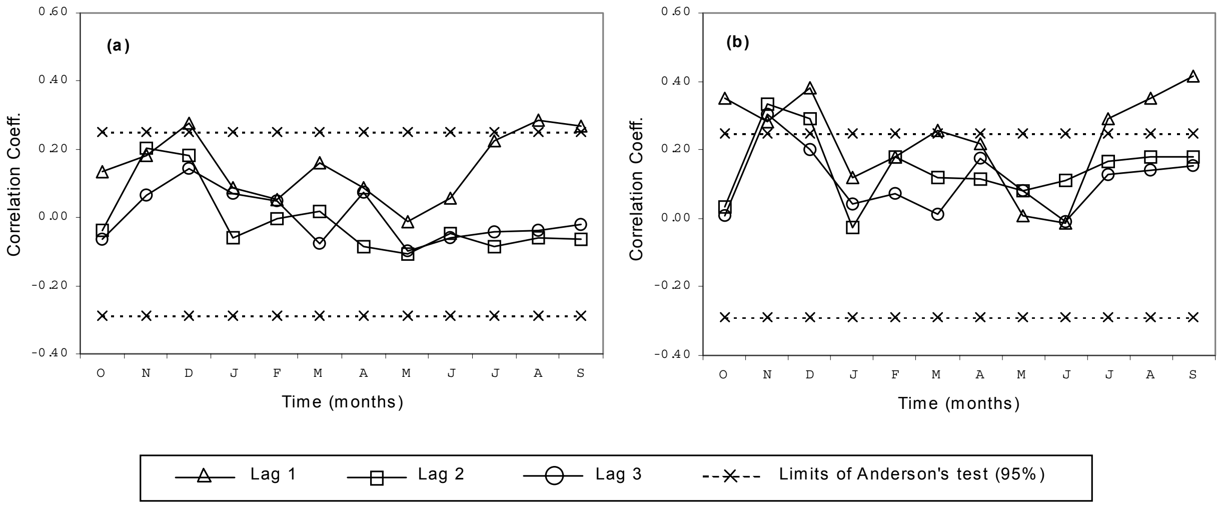

Fig. 7. *Month-to-month correlations of standardised inflow series. (a) Buendía (b) Entrepeñas*

Table 3. Relative root-mean-squared differences (RRMSD) of the synthetic statistics

Statistics	*Buendía*		*Entrepeñas*	
	NN	**AR(2)**	**NN**	**AR(2)**
Basic statistics of monthly series				
Mean	0.028	0.020	0.012	0.016
Standard deviation	0.129	0.077	0.113	0.042
Skewness coefficient	0.535	0.441	0.376	0.212
Series persistence statistics of monthly series				
Buendía	0.276	0.329	0.283	0.303
Entrepeñas	0.247	0.282	0.317	0.355
Maximum monthly drought statistics				
Frequency	0.166	0.211	0.144	0.226
Length	0.150	0.176	0.217	0.333
Intensity	0.116	0.119	0.132	0.112
Magnitude	0.144	0.250	0.119	0.331
Maximum annual drought statistics				
Frequency	0.191	0.265	0.201	0.285
Lenght	0.210	0.225	0.165	0.236
Intensity	0.112	0.025	0.040	0.166
Magnitude	0.052	0.114	0.195	0.189
Monthly storage statistics				
Reservoir capacity	0.267	0.283	0.468	0.504
Hurst coefficient	0.039	0.053	0.085	0.098
Annual storage statistics				
Reservoir capacity	0.320	0.345	0.501	0.540
Hurst coefficient	0.068	0.096	0.146	0.171

of consecutive values below it defines a single drought, with its duration, intensity (threshold minus minimum value) and magnitude (total volume below threshold). The frequency of droughts, together with the basic descriptors, have been computed for all series in both reservoirs, including annual and monthly series. Special attention is given to statistics of maximum droughts.

The first three categories of statistics were calculated according to Salas *et al*. (1980).

- *Storage statistics*. The hypothetical reservoir storage capacity to guarantee a given percentage of Q_M has been calculated for all the series during the period of 53 years. Again, statistics are computed from the monthly and annual series. Also, Hurst coefficients were computed and compared in all cases.

This group of statistics was computed according to Loucks *et al*. (1981) and Salas *et al*. (1980).

All the statistics in the four groups were computed for each of the 200 synthetic series and then averaged over the ensemble. Also, the statistical descriptors were computed from the historical records at Buendía and Entrepeñas reservoirs. These were compared, systematically, with averages computed for the synthetic series. All the results were plotted to give a general qualitative idea of each model's performance, and for each single graphic, the relative root-mean-squared difference (RRMSD) was obtained for both the AR(2) and NN-based model. Table 3 comprises all computed values of the RRMSD, while graphical results corresponding to the most relevant tests have been selected and presented (Figs. 8 to 11).

Results show that the AR(2) model reproduces the empirical standard deviations better than the NN-based model although the differences are not large. The correlation

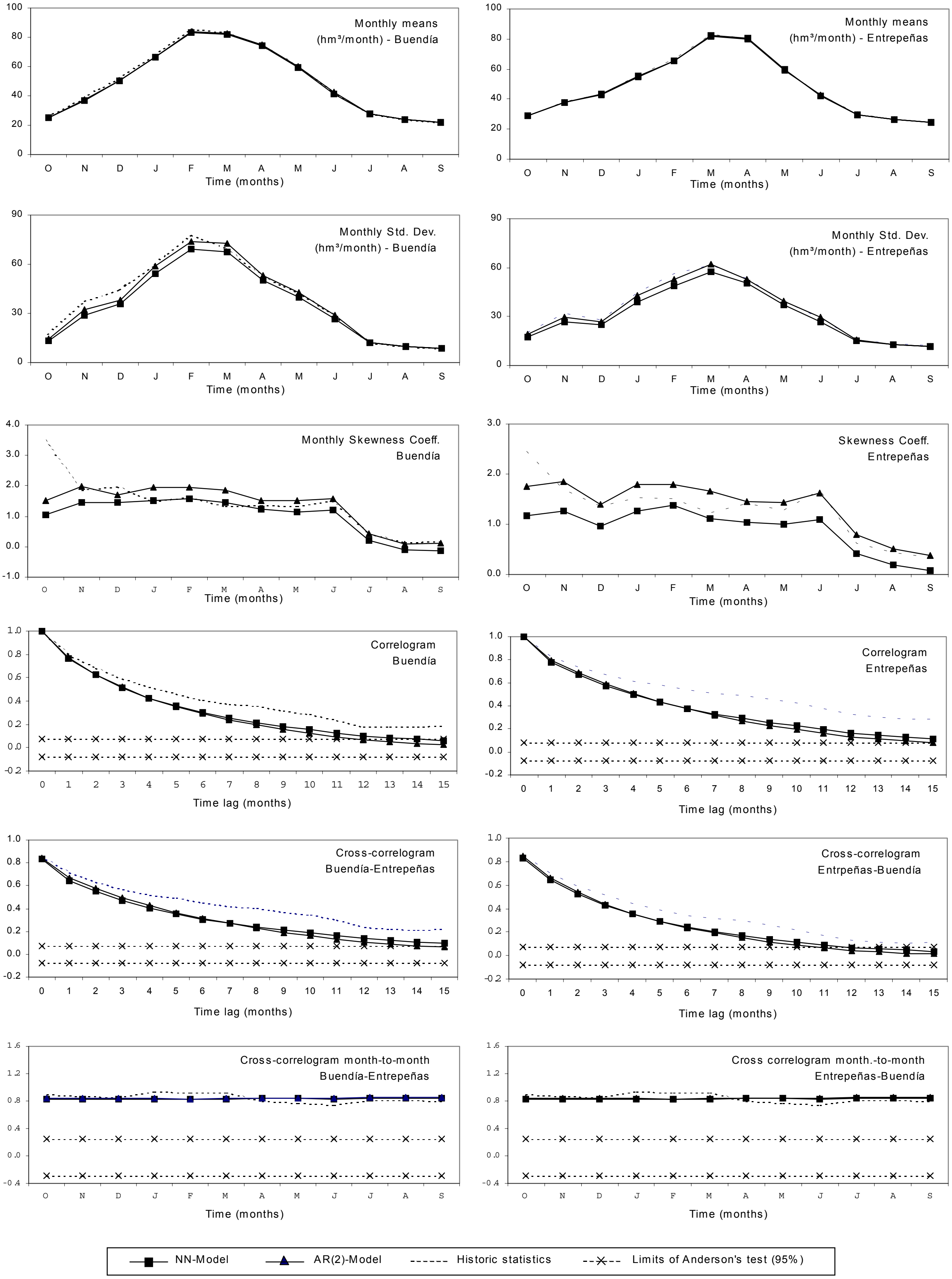

Fig. 8. *Basics statistics and series persistence statistics. Historical values* v. *synthetic values*

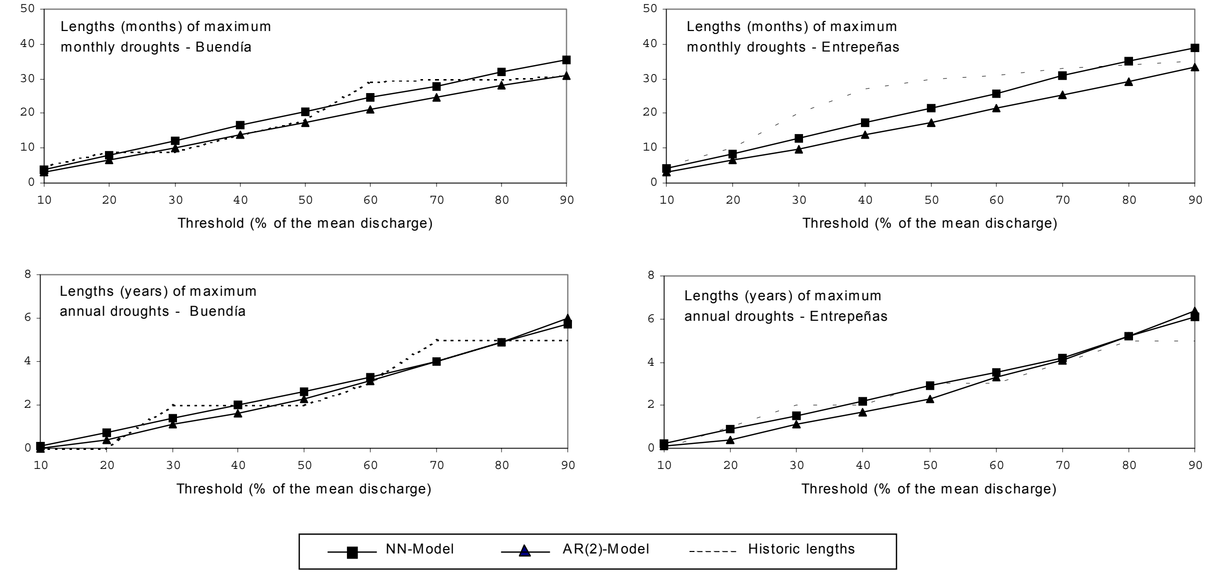

Fig. 9. Lenghts of maximum droughts. Historical values v. *Synthetic values*

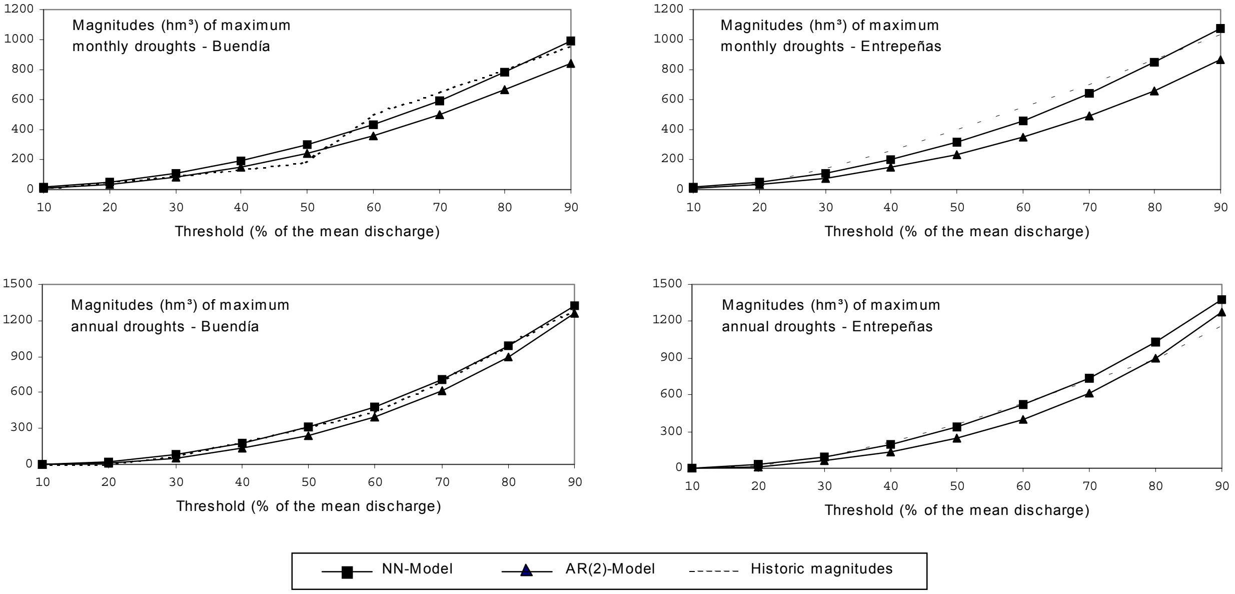

Fig. 10. Magnitudes of maximum droughts. Historical values v. *Synthetic values*

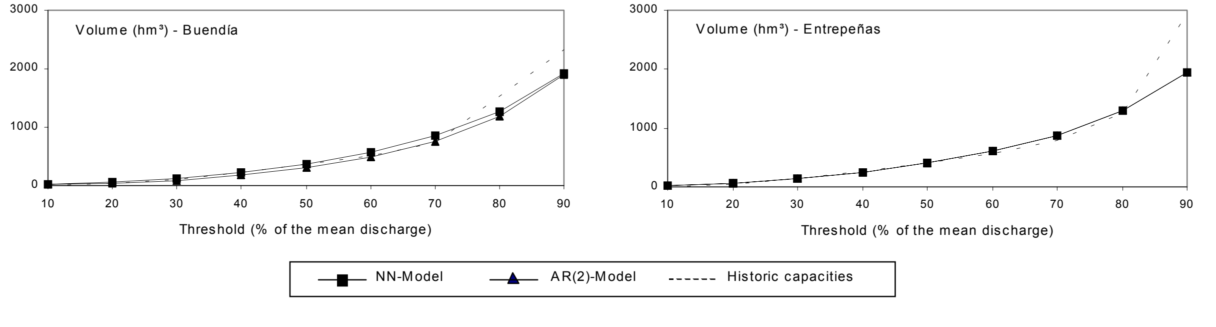

Fig. 11. *Monthly storage capacity. Historical values* v. *Synthetic values*

Table 4. Hurst coefficients

Reserv. Site	*Hist.*	*NN*	*AR(2)*
Buendía:			
Monthly value	0.76	0.73	0.72
Annual value	0.73	0.68	0.66
Entrepeñas:			
Monthly value	0.82	0.75	0.74
Annual value	0.82	0.70	0.68

persistence shown by the historical series is not reproduced adequately by any of the models, as Fig. 8 shows, and a similar conclusion stands for cross-correlations. Nevertheless, this underestimation of correlations is more evident in the AR(2) case, while the NN-based generator tends to keep better significant correlation values for larger lags, behaviour which is more acceptable than that of the AR(2) model. On the other hand, the same figure shows that month-to-month cross-correlations are equally well reproduced by both models.

The results concerning basic statistics show the NN model as favourite, particularly when the most relevant descriptors are considered, i.e. the duration and magnitude of droughts. Figures. 9 and 10 show results for maximum droughts. AR(2) underestimates the magnitude of droughts significantly, while the NN approach reproduces the historical drought statistics much better. When the droughts are identified from the annual series, similar conclusions can be reached. Maximum droughts resulting from different discharge thresholds, are reproduced better by the NN model. Figs. 9 and 10, for both reservoir sites, show empirical and synthetic statistics of duration and magnitude of droughts derived from monthly and annual series.

Storage statistics are well reproduced by both models for thresholds of 70% of Q_M or lower; results for both models are almost equivalent (Fig. 11). Concerning Hurst coefficients, Table 4 shows that in all cases the models underestimate empirical values, although the NN model gives lower differences.

Conclusions

A hybrid model is proposed as a practical tool for multivariate generation of monthly streamflow series at two geographical sites located in the Tagus River basin in Spain. The model consists of a deterministic core defined in terms of a one-hidden-layer feed-forward neural network with two hidden nodes, plus a stochastic part built with a multivariate white noise component. Historical discharge records were used to train and test the potentials of the model for generation of practical hydrological scenarios. The validation process of the developed model was carried out as usual with this kind of model: the comparison between historical and synthetic statistics which are relevant in the framework of water resources systems managment and planning. For comparative purposes, the well-known multivariate AR(2) model was also applied on an identical basis.

The combined stochastic-NN approach outperforms the purely stochastic AR(2) model, particularly for drought related statistics. It can be concluded that the proposed hybrid model is a viable alternative for future applications, in competition with linear purely stochastic autoregressive models. The present results demonstrate, qualitatively, its value for synthetic generation of monthly streamflow series in water resources system analysis. Its full potential should be explored in other similar studies that incorporate a larger number of related series at different locations. Other network architectures should be also explored, while different training techniques might be required for an efficient operation, as the number of locations is increased and the network becomes more complex. The authors are now working on new case studies, which include a larger number of stations. Preliminary results indicate that the hybrid technique proposed herein performs well when applied to synthetic generation of monthly streamflows at three sites.

Acknowledgements

The research grants from the Spanish Ministry of Education and Culture (CICYT project HID99-0656) and European Commission (INCOMED "WAMME" project ICA3.199.00014) are acknowledged. Likewise, authors do appreciate the thorough review work by the anonymous reviewers, which not only made it possible to enrich the paper, but also gave us key hints for future improvements of the methodology.

References

Andreu, J., Capilla, J. and Sanchís, E., 1996. AQUATOOL, a generalized decision-support system for water-resources planning and operational management. *J. Hydrol.,* **177**, 269–291.

Anmala, J., Zhang, B. and Govindaraju, R.S., 2000. Comparison of ANNs and empirical approaches for predicting watershed runoff. *J. Water Resour. Plan. Man.-ASCE,* **126**, 156–166.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (Rao Govindaraju), 2000. Artificial neural networks in hydrology. I: preliminary concepts. *J. Hydrolog. Eng.-ASCE,* **5**, 115–123.

Basson, M.S. and van Rooyen, J.A., 2001. Practical application of probabilistic approaches to the management of water resource systems. *J. Hydrol.,* **241**, 53–61.

Boogard, H.F.P., van den, Gautam, D.K. and Mynett, A.E., 1998. Auto-regressive neural networks for the modeling of time series. In: *Hydrodynamics 98*, Babovic and Larsen (Eds.), Balkema, Rotterdam, 741–748.

Bras, R.L. and Rodríguez-Iturbe, I., 1985. *Random functions and hydrology.* Addison-Wesley, Massachusetts.

CHT - Confederación Hidrográfica del Tajo, 1999. *La cuenca del Tajo en cifras.* Oficina de Planificación Hidrológica, CHT, Ministerio de Medio Ambiente, Madrid.

Chakraborty, K., Mehrotra, K., Mohan, C.K. and Ranka, S., 1992. Forecasting the behavior of multivariate time series using neural networks. *Neural Net.,* **5**, 961–970.

Cheng, B. and Titterington, D.M., 1994. Neural networks: a review from statistical perspective. *Statist. Sci.,* **9**, 2–54.

Deo, M.C. and Thirumalaiah, K., 2000. Real time forecasting using neural networks. In: *Artificial neural networks in hydrology*, R.S. Govindaraju and A.R. Rao (Eds.), Kluwer, Dordrecht, The Netherlands. 53–71.

Fernandez, B. and Salas, J.D., 1986. Periodic gamma autoregressive processes for operational hydrology. *Water Resour. Res***.,** **22**, 1385–1396.

Gupta, H.V., Hsu, K. and Sorooshian, S., 2000. Effective and efficient modeling for streamflow forecasting. In: *Artificial neural networks in hydrology*, R.S. Govindaraju and A.R. Rao (Eds.), Kluwer, Dordrecht, The Netherlands. 7–22.

Haykin, S., 1999. *Neural networks - A comprehensive foundation.* Prentice-Hall, New Jersey.

Hornik, K., Stinchcombe, M. and White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Net.,* **2**, 395–403.

Imrie, C.E., Durucan, S. and Korre, A., 2000. River flow prediction using artificial neural networks: generalisation beyond the calibration range. *J. Hydrol***.,** **233**, 138–153.

Jackson, B.B., 1975. The use of streamflow models in planning. *Water Resour. Res.,* **11**, 54–63.

Kendall, D.R. and Dracup, J.A., 1991. A comparison of index-sequential and AR(1) generated hydrologic sequences. *J. Hydrol.,* **122**, 335–352.

Koutsoyiannis, D., 2000 A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series *Water Resour. Res.,* **36**, 1519–1533.

Lachtermacher, G. and Fuller, J.D., 1994. Backpropagation in hydrological time series forecasting. In: *Stochastic and statistical methods in hydrology and environmental engineering - Time series analysis in hydrology and environmental engineering*, K.W. Hipel *et al.* (Eds.), Kluwer, Dordrecht, The Netherlands 229–242.

Lapedes, A. and Farber, R., 1988. How neural net works. In: *Neural information processing systems*, D. Z. Anderson (Ed.), American Institute of Physics, New York, 442–456.

Lehtokangas, M., Saarinen, J. and Kaski, K., 1996. A network of autoregressive processing units for time series modeling. *Appl. Math. Comput.,* **75**, 151–165.

Loucks, D.P., Stedinger, J.R. and Haith, D.A., 1981. *Water resources system planning and analysis.* Prentice Hall, New Jersey.

Maier, H.R. and Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Model. Software***,** **15**, 101–124.

Raman, H. and Sunilkumar, N., 1995. Multivariate modelling of water resources time series using artificial neural networks. *Hydrolog. Sci. J.,* **40**, 145–163.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning internal representations by error propagation. In: *Parallel Distributed Processing*, D. E. Rumelhart and J. L. McCleland (Eds.), Vol. 1, Chapter 8, MIT Press, Cambridge, MA.

Salas, J.D., Delleur, J.W., Yevjevich, V. and Lane, W.L., 1980. *Applied modeling of hydrologic time series*. Water Resources Publications. Littleton, Colorado.

Salas, J.D., Tabios III, G.Q. and Bartolini, P.,1985. Approaches to multivariate modeling of water resources time series, *Water Resour. Bull.,* **21**, 683–708.

Salas, J.D., Markus, M. and Tokar, A.S., 2000. Streamflow forecasting based on artificial neural networks. In: *Artificial neural networks in hydrology*, R. S. Govindaraju and A. R. Rao (Eds.), Kluwer, Dordrecht, The Netherlands. 23–51.

Stedinger, J.R. and Taylor, M.R., 1982. Synthetic streamflow generation 1. Model verification and validation, *Water Resour. Res.,* **18**, 909–918.

Stern, H.S., 1996. Neural networks in applied statistics. *Technometrics***,** **38**, 205–214.

Tang, Z., de Almeida, C. and Fishwick, P.A., 1991. Time series forecasting using neural networks vs. Box-Jenkins methodology, *Simulation,* **57**, 303–310.

Zealand, C.M., Burn, D.H. and Simonovic, S.P., 1999. Short term streamflow forecasting using artificial neural networks. *J. Hydrol.,* **214,** 32–48.