A20975F

Research Journal

Stochastic Environmental Research and Risk Assessment

Springer

Originals

Indexed in Current Contents 14 (2) 79–160 (2000) *Printed on acid-free paper*

Stochastic Environmental Research and Risk Assessment 14 (2000) 123–138 © Springer-Verlag 2000

El Niño influence on streamflow forecasting

J. Kelman, A. de M. Vieira, J. E. Rodriguez-Amaya

Abstract. Stochastic models are often fitted to historical data in order to produce streamflow scenarios. These scenarios are used as input data for simulation/optimization models that support operational decisions for water resource systems. The streamflow scenarios are sampled from probability distributions conditioned on the available information, such as recent streamflow data. In this paper we introduce a procedure for further conditioning the probability distributions by considering the recent measurements of climatic variables, such as sea temperatures, that are used to describe the occurrence of El Niño. We adopt an auto-regressive model and use the "El Niño information" to refine the parameter estimation process for each time step. The corresponding methodology is tested for the monthly energy time series, "inflowing" to the power plants of Colombia. This is a linear combination of streamflow values for the 18 most important rivers of the country.

Key words: El Niño, streamflow forecasting, streamflow scenarios, Monte Carlo simulation.

1 Introduction

El Niño is a general term used to describe a set of concurrent and unusual climatic events in South Pacific, such as sea temperature increase and easterlies retreat, which are correlated to hydrological extreme events in different parts of

J. Kelman
COPPE, Federal University of Rio de Janeiro,
Rua Capitu 41, CEP 22750-040, Rio de Janeiro, Brazil
e-mail: Kelman@ruralrj.com.br

A. de M. Vieira
UERJ, State University of Rio de Janeiro
and Eletrobrás - Centrais Elétricas Brasileiras S.A.,
Rua Maria Angélica 114/401,
CEP 22470-200, Rio de Janeiro, Brazil
e-mail: Amvieira@gbl.com.br

J. E. Rodriguez-Amaya
ISA - Interconexión Eléctrica S.A.,
Medellin, Antioquia, Colombia
e-mail: Jerodriguez@isa.com.co

Several of the observations made by the anonymous reviewer were included in the final manuscript.

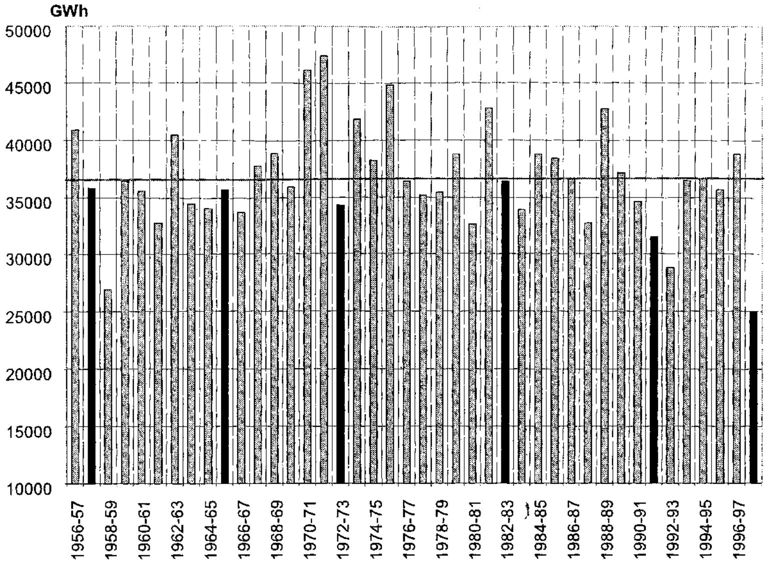

Fig. 1. Annual energy inflow to the Colombian hydro power system

the world. The term "El Niño" originally referred to relatively warm surface water that appears off the west coast of equatorial South America during the first few months of the calendar year due to an annual weakening of the trade winds. Now it means a wide-spread warming, compared to average, of the central and eastern equatorial Pacific Ocean. At the same time, sea surface temperatures in the western Pacific are cooler than average. During the most recent El Niño of 1997–98, sea surface temperature in the eastern Pacific were the warmest ever recorded (Liebmann 1998).

Because El Niño has received lately much attention from the general press, no effort will be made to provide a general description of the phenomenon. Good reviews of the correlation between the occurrence of El Niño and unusual meteorological and/or hydrological events are provided by Poveda and Mesa (1996) and by Piechota and Dracup (1996).

Among the most well-known impacts of El Niño on the west coast of South America, stands out the increase of precipitation on Ecuador and Peru and reduction of precipitation on most of Colombia. For Colombia, it means less water for the production of electric energy by the hydro power plants and, therefore, it means risk of energy shortage.

A lumped view of the hydrological variability effect on hydro energy availability can be obtained by a linear combination of streamflow values that maps a vector of streamflow values, for any given time interval, onto a single value, called "energy inflow to the system". The weights used in the linear combination are proportional to the mean hydraulic head of each hydro plant. That is, variability of the hydraulic head is neglected. Figure 1 shows the evolution of the energy inflow to the Colombian system[1], in GWh/year and correspond to the overall

[1] Energy inflow to the reservoirs is not the same as energy generated by the system, due to the regulating effect of the 19 major reservoirs in the country. The acumulated storage capacity is about 7000 Mm^3, which corresponds to about 14,000 GWh.

energy provided by the 18 most important rivers of the country[2]. Annual historical average is about 37,000 GWh and the annual consumption is about 44,000 GWh. In any year, the eventual difference between energy demand and hydro power production, if positive, is met by thermal generation or, when thermal plants are not available anymore, by shortages. "Strong" El Niño events (Quinn, as cited by Cadavid-Mazo et al. 1998) are marked dark in Fig. 1. Annual values correspond to events in the May–April time interval, which is the hydrological annual cycle in Colombia.

As it can be seen, the energy inflow during the last El Niño (May 97–April 98) was about 25,000 GWh, only 68% of the annual historical average. The return period for such observation is at least 100 years, considering all probability distributions usually fitted to annual data (Cadavid-Mazo et al. 1998). In the six years with occurrence of strong El Niño, since 1956, the energy inflows were smaller than the average. However, small values of energy inflow were also observed in several years without strong El Niño.

In Sect. 2 we introduce the use of climatic information for producing a streamflow forecast, as a weighted sum of historical data. We show how to calculate the weights as a function of the climatic observations. In Sect. 3 we review the periodic auto-regressive model, which has been used to produce streamflow scenarios based on recently observed streamflow data. For each river, and each time step, the streamflow forecast is simply the average of all values belonging to different streamflow scenarios. In Sect. 4 we show how to combine the two sources of information: (a) climatic data and (b) recently observed streamflow data. In Sect. 5 we show some results obtained with the model for the Colombian case. In Sect. 6 we present the conclusions.

2 Streamflow forecasting, using climatic information

Suppose that the time series of monthly streamflows for a particular river is given by

$$\mathrm{Z} = \begin{bmatrix} z(1,1), z(1,2), \ldots, z(1,m), \ldots, z(1,12) \\ z(2,1), z(2,2), \ldots, z(2,m), \ldots, z(2,12) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ z(n,1), z(n,2), \ldots., z(n,m), \ldots, z(n,12) \\ z(n+1,1), \ldots\ldots\ldots.z(n+1,m) \end{bmatrix}$$

where $z(y, m)$ stands for the streamflow in year y, month m. We have assumed for simplicity that $y = 1$ in the first year and that the historical record consists of n years with full data, plus m monthly observations in the year $n + 1$, which is the current year.

Assume that there is interest on making a forecast for the monthly streamflow many years from now, for example, for year $n + 50$, month $m + f$ (for simplicity, only the case $0 < m + f \leq 12$ will be considered). Because the interest is in what is going to occur in the remote future, the best alternative would be to use the naive forecast, which is the estimate of the marginal expected value, given by the arithmetic mean,

[2] Energy inflow to the reservoirs is calculated assuming mean hydraulic head at each power plant. That is, reservoir and tailwater fluctuations are neglected.

$$\hat{z}(n+50, m+f) = \frac{1}{n}\sum_{y=1}^{n} z(y, m+f) \tag{1}$$

However, if we were interested in making a forecast for the streamflow in the immediate future, say for month $m+f$, year $n+1$, then the best choice would be to use the conditioned expected value, given by the weighted mean,

$$\hat{z}(n+1, m+f) = \sum_{y=1}^{n} w(y) z(y, m+f) \tag{2}$$

where $w(y)$ is the "weight" allocated to the information of year y. Obviously Eq. (1) is a particular case of Eq. (2), for $w(y) = n^{-1}, \forall\, y$. The question is how to evaluate $w(y)$.

Suppose that the recent information related to El Niño, such as sea surface temperature, is given by the "El Niño predictor vector"

$$C(n+1, m) = [c(n+1, m), c(n+1, m-1), \ldots, c(n+1, m-k)]$$

and that equivalent information in any previous year y is given by

$$C(y, m) = [c(y, m), c(y, m-1), \ldots, c(y, m-k)]$$

In other words, the climatic condition is captured through the last $k+1$ observations of some variable capable of detecting the occurrence of El Niño. k will be called the "climatic lag".

Let us define $d(y)$ as the "distance" between $C(y, m)$and $C(n+1, m)$,

$$d(y) = ||C(y, m) - C(n+1, m)||, \quad \text{for } 0 < y < n+1 \tag{3}$$

The actual calculation of $d(y)$ can be done in several ways. We have adopted

$$d(y) = \sum_{i=0}^{k} |c(y, m-i) - c(n+1, m-i)| \tag{4}$$

The similitude between the current El Niño condition and the El Niño condition observed in year y is larger, the smaller is $d(y)$. Several analytical functions could be conceived to represent a decreasing relationship between $w(y)$ and $d(y)$. Among them we have selected

$$w(y) = \frac{e^{-\alpha d(y)}}{\sum_{i=1}^{n} e^{-\alpha d(i)}} \tag{5}$$

where α is a parameter. It should be noticed that if $\alpha = 0$, then $w(y) = n^{-1}$. That is, when $\alpha = 0$ the weighted mean (Eq. (2)) is reduced to the arithmetic mean (Eq. (1)). The parameter α can be estimated simulating the application of forecasts based on Eq. (2) for the time interval for which actual flows are known.

Suppose that in the past we were in month m and year y_h, where $0 < y_h \leq n$, trying to produce a forecast for $z(y_h, m+f)$, using Eq. (2). Furthermore, assume that in this particular situation, the entire time series Z would be available, with the exception of the data for year y_h. This exception is obviously necessary because otherwise the simulation would be senseless, as $z(y_h, m+f)$ would be actually known.

For each year y, such that $0 < y \leq n$ and $y \neq y_h$, it would be possible to calculate a distance $d(y, y_h)$ between the climatic conditions observed in year y and in year y_h, through the replacement of $n+1$ by y_h in Eq. (3).

The equivalent of Eq. (5) would be, in this case,

$$w(y, y_h) = \frac{e^{-\alpha d(y,y_h)}}{\sum_{i=1, i\neq y_h}^{n} e^{-\alpha d(i,y_h)}} \tag{6}$$

The equivalent of Eq. (2) would be, in this case,

$$\hat{z}(y_h, m+f) = \sum_{\substack{y=1 \\ y\neq y_h}}^{n} w(y, y_h) z(y, m+f) \tag{7}$$

The prediction error would be a function of α because of the role of Eq. (5) on the calculation of the forecast,

$$\xi(y_h, m, f; \alpha) = \hat{z}(y_h, m+f) - z(y_h, m+f) \tag{8}$$

An overall measure of the prediction error for lag-f streamflow forecasts is given by

$$g(f, \alpha) = \sum_{y_h=1}^{n} \sum_{m=1}^{12} \xi(y_h, m, f; \alpha)^2 \tag{9}$$

For each lag f, Eq. (9) can be minimized with regard to α. Let $\alpha^*(f)$ be this function,

$$\alpha^*(f) = \min_{\alpha}\{g(f, \alpha)\} \tag{10}$$

It is of particular interest to evaluate the reduction of the overall measure of prediction error, when using the proposed procedure, as compared to what would result from the use of naive forecasts (Eq. (1)),

$$h(f) = \frac{g(f, \alpha^*(f))}{g(f, 0)} \tag{11}$$

The smaller is $h(f)$ the larger is the benefit of using Eq. (2), rather than Eq. (1).

3 Streamflow forecasting, using previous streamflow information

Regression models, in general denominated periodic-auto-regressive of order p [PAR(p)] have been adopted to model streamflow with periodic variations of the mean, standard deviation and of the auto-correlation coefficients (for example, Salas et al. 1980; Maceira 1989).

Let Z_t, for $t = t(y, m) = (y-1)12 + m$. The variable Z_t is said periodically correlated if $E(Z_{t(y,m)})$ and $\mathrm{Cov}(Z_{t(y,m)}, Z_{t(y',m')})$ depend only on month m and forecast lag f. Let us define

$$\mu_m = E(Z_{t(y,m)}) \tag{12}$$

and

$$\gamma_{f,m} = \mathrm{Cov}(Z_{t(y,m)}; Z_{t(y',m')}) \tag{13}$$

where

$$y' = y \text{ and } m' = m + f, \quad \text{if } m + f < 13$$

$$y' = y + 1 \text{ and } m' = m + f - 12, \quad \text{otherwise}$$

Define vector p as p = $\{p_1, p_2, \ldots, p_{12}\}$ so that each element indicates the valid auto-regressive order for the corresponding month.

The PAR$(p_1, p_2, \ldots, p_{12})$ model can be described as:

$$\Phi^m(B)\left\{\frac{Z_{t(y,m)} - \mu_m}{\sigma_m}\right\} = a_t \tag{14}$$

where, μ_m is the expected value of Z_t for month m, σ_m is standard deviation of Z_t for month m, B is time operator $(B^i Z_t = Z_{t-i})$, p_m is order of the auto-regressive process for month m, Φ^m is auto-regressive operator of order p_m i.e., $\Phi^m(B) = (1 - \varphi_1^m B^1 - \varphi_2^m B^2 - \cdots - \varphi_{p_m}^m B^{p_m})$, a_t is time-independent variable with average zero and variance $\sigma^2_{a,m}$, often called "residual", not necessarily normally distributed.

Equation (14) can be rewritten like:

$$\left(\frac{Z_{t(y,m)} - \mu_m}{\sigma_m}\right) = \varphi_1^m\left(\frac{Z_{t(y,m)-1} - \mu_{m-1}}{\sigma_{m-1}}\right) + \cdots + \varphi_{p_m}^m\left(\frac{Z_{t(y,m)-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + a_t \tag{15}$$

The operational nuisance of dealing with highly skewed probability distributions for a_t (Todini 1980) is usually avoided by transforming the original streamflow time series into a normalized time series (Box and Cox 1964). However, a stochastic model adopting this procedure would not be compatible with efficient algorithms for setting the rules for reservoir operation, which assume linearity for the multiple regression of Z_t on the previous observations $Z_{t-1}, Z_{t-2}, \ldots$ (Pereira and Pinto 1991). With this constraint in mind, we have not used any transformation of the original variable. Instead, we have adopted an adaptable skewed probability distribution for a_t. At each time interval t a three parameter log-normal distribution is selected for a_t (zero expected value and variance equal to $\sigma^2_{a,m}$) in such a way that the probability of getting negative values for Z_t is minimized. It can be derived from Eq. (15) that Z_t will be positive if

$$a_t > \lambda_t = -\frac{\mu_m}{\sigma_m} - \varphi_1^m\left(\frac{Z_{t(r,m)-1} - \mu_{m-1}}{\sigma_{m-1}}\right) - \cdots - \varphi_{p_m}^m\left(\frac{Z_{t(r,m)-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) \tag{16}$$

λ_t would be an obvious choice for the lower bound of the domain of a_t. However, if $\lambda_t > 0$, the expected value of a_t could not possibly be zero. For this reason, the lower bound for the domain of a_t is set equal to ψ_t, defined

$$\psi_t = \min[\psi_{\max}, \lambda_t] \tag{17}$$

where $\psi_{\max}$ is a negative value, selected very close to zero.

The probability density function for a_t is

$$f_A(a_t) = \frac{1}{\sqrt{2\pi}\sigma_n(a_t - \psi_t)} \exp\left\{ -\frac{1}{2\sigma_y^2}[\ln(a_t - \psi_t)] - \mu_n \right\} \tag{18}$$

where

$$\mu_n = \frac{1}{2}\ln\left(\frac{\sigma_{a,m}^2}{\Delta(\Delta - 1)}\right) \tag{19}$$

$$\sigma_n^2 = [\ln \Delta]^{0.5} \tag{20}$$

and

$$\Delta = 1 + \frac{\sigma_{a,m}^2}{\psi^2} \tag{21}$$

Therefore the standard normal "innovation" for time step t would be

$$\varepsilon_t = [\ln(a_t - \psi_t) - \mu_n]/\sigma_n \tag{22}$$

There are several methodologies available for identifying the vector $p = \{p_1, p_2, \ldots, p_{12}\}$, as described by McLeod (1994). Most of them seek parameter parsimony. However, as demonstrated by Kelman (1987), parameter parsimony often results on stochastic models that "see" the worst drought of the historical record as a highly unlikely event. This is an undesirable feature because it means that the model would underestimate the probability of severe droughts. We have consider that parameter parsimony is a less valuable asset than "being on the safe side". Accordingly, a "crude approach" was adopted for identifying $p = \{p_1, p_2, \ldots, p_{12}\}$.

We calculate the variance of a_t, for each month m, assuming $p_m = 1, 2, \ldots, 6$. Let $\sigma_{a,m}^2(1), \sigma_{a,m}^2(2), \ldots, \sigma_{a,m}^2(6)$ be these variances. Rather than the usual procedure, that begins with the null hypothesis that $p_m = 0$ and the alternative hypothesis that $p_m > 0$, we begin with the null hypothesis that $p_m = 6$ and the alternative hypothesis that $p_m < 6$. Operationally, we accept $p_m = 6$, provided that $\sigma_{a,m}^2(6)/\sigma_{a,m}^2(5) < 0.975$. Otherwise, we accept $p_m = 5$, provided that $\sigma_{a,m}^2(5)/\sigma_{a,m}^2(4) < 0.975$. Otherwise...

The seasonal means and variances are estimated by:

$$\hat{\mu}_m = n^{-1}\sum_{y=1}^{n} Z_{(y-1)12+m}, \quad m = 1, 2, \ldots, 12 \tag{23}$$

$$\hat{\sigma}_m^2 = n^{-1}\sum_{y=1}^{n} (Z_{(y-1)12+m} - \hat{\mu}_m)^2, \quad m = 1, 2, \ldots, 12 \tag{24}$$

Auto-regressive parameters $\varphi_{m,i}, i = 1, 2, \ldots, p_m$ could be estimated by the Yule–Walker equations. However, because the estimator of the auto-correlation coefficient has large variance, particularly for large lags, often one is faced with numerical instability (Bras and Rodriguez-Iturbe 1985). One alternative is to estimate the auto-regressive parameters using the ordinary minimum squares method (Johnston 1963):

$$\hat{\varphi}_m = \begin{bmatrix} \varphi_{m,1} \\ \varphi_{m,2} \\ \vdots \\ \varphi_{m,p_m} \end{bmatrix} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{U} \tag{25}$$

where

$$\mathrm{U} = \begin{bmatrix} \frac{Z_m - \mu_m}{\sigma_m^2} \\ \frac{Z_{12+m} - \mu_m}{\sigma_m^2} \\ \cdots \\ \frac{Z_{12(n-1)+m} - \mu_m}{\sigma_m^2} \end{bmatrix}$$

$$\mathrm{X} = \begin{bmatrix} \frac{Z_{m-1} - \mu_{m-1}}{\sigma_{m-1}^2} & \frac{Z_{m-2} - \mu_{m-2}}{\sigma_{m-2}^2} & \cdots & \frac{Z_{m-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}^2} \\ \frac{Z_{12+m-1} - \mu_{m-1}}{\sigma_{m-1}^2} & \frac{Z_{12+m-2} - \mu_{m-2}}{\sigma_{m-2}^2} & \cdots & \frac{Z_{12+m-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}^2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{Z_{(n-1)12+m-1} - \mu_{m-1}}{\sigma_{m-1}^2} & \frac{Z_{(n-1)12+m-2} - \mu_{m-2}}{\sigma_{m-2}^2} & \cdots & \frac{Z_{(n-1)12+m-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}^2} \end{bmatrix} \tag{26}$$

The variance of the estimator $\hat{\varphi}_m$ is

$$\mathrm{Var}(\hat{\varphi}_m) = \sigma_{a,m}^2 (\mathrm{X}'\mathrm{X})^{-1} \tag{27}$$

and the variance of a_t, for each month is estimated by

$$\hat{\sigma}_{a,m}^2 = \frac{\mathrm{U}'\mathrm{U} - \varphi_m'\mathrm{X}'\mathrm{U}}{n - p_m} \tag{28}$$

We have dealt with the spatial streamflow dependence among the different rivers by imposing a contemporaneous cross correlation for the residuals. That is, if $a_t(j)$ is the residual of the jth river, expressed by Eqs. (14) and (15), and if r is the number of rivers (18 in the case of Colombia) then,

$$\aleph_t = \begin{bmatrix} \varepsilon_t(1) \\ \varepsilon_t(2) \\ \\ \varepsilon_t(r) \end{bmatrix} = \mathrm{L} \begin{bmatrix} \eta_t(1) \\ \eta_t(2) \\ \\ \eta_t(r) \end{bmatrix} \tag{29}$$

where $\eta_t(j)$ is a standard normal deviate, for all j, and $\eta_t(j)$ is independent of any other $\eta_t(i)$, provided $i \neq j$. L is the r by r "load" matrix, satisfying the following equation

$$LL' = \Sigma \tag{30}$$

where Σ is the covariance matrix that captures the spatial dependence among flows of different rivers. Usually, Σ is selected as the covariance matrix of the residuals. That is,

$$\Sigma = \hat{\text{cov}}(\aleph_t) \tag{31}$$

However, experience has shown that when Eq. (31) is used the cross correlation among annual flows of the synthetic sequences tend to be smaller than the corresponding values estimated from the historical record. This is an undesirable feature because it means that the model would underestimate the probability that severe droughts in different rivers would occur simultaneously.

The bias could be decreased by forcing higher values for some of the elements of matrix $\hat{\text{cov}}(\aleph_t)$. How much to increase these values would not be obvious because the "adjusted" matrix could turn out not positive definite. Let $H' = [h(1), h(2), \ldots h(r)]$ be the vector of annual flows for the r rivers. As experience has shown that the elements of $\hat{\text{cov}}(H)$ are in general larger than the corresponding elements of $\hat{\text{cov}}(\aleph_t)$, a reasonable alternative is to use $\hat{\text{cov}}(H)$, instead of $\hat{\text{cov}}(\aleph_t)$. That is, we have adopted

$$\Sigma = \hat{\text{cov}}(H) \tag{32}$$

4 Streamflow forecasting, using climatic and previous streamflow information

Equation 23 can be generalized in order to take into account that data for each year of the historical record has a different information value, translated by the weight $w(y)$. That is

$$\hat{\mu}_m = \sum_{y=1}^{n} w(y) z_{(y-1)s+m}, \quad m = 1, 2, \ldots, 12 \tag{33}$$

Analogously, Eqs. (24) and (25) can be generalized as

$$\hat{\sigma}_m^2 = \sum_{y=1}^{n} w(y) (z_{(y-1)s+m} - \hat{\mu}_m)^2, \quad m = 1, 2, \ldots, 12 \tag{34}$$

and

$$\hat{\varphi} = (X'VX)^{-1} X'VU \tag{35}$$

where V is a n by n diagonal matrix with $w(y), y = 1, 2, \ldots, n$ in the diagonal.

At each new month step, Eqs. (33)-(35) need to be applied again. In other words, we have adopted an adaptive procedure for the use of PAR(p) model, in which a new set of parameters is estimated each time there is new climatic or river flow information.

5 Case study

The severity of the 1997/98 drought in Colombia was most intense from September 97 to February 98. During this period the actual energy inflow was 7997 GWh, a merely 48% of the mean energy inflow for the period (16616 GWh). For this reason, we have studied this 6 months time interval, called "case 1997 (dry)". For the sake of comparison, we have also studied a wet 6 months time interval, called "case 1988 (wet)". During this period the actual energy inflow was 23471 GWh, or 141% of the mean energy inflow for the same period.

Table 1 shows eight of the most commonly referred climatic variables, measured over the Pacific Ocean, that have been related to the occurrence of El Niño or of La Nina. In order to select one of them to be used as the "El Niño predictor vector" C, defined in Sect. 2, we have individually fitted an ARIMA model to each of the eight climatic time series variables and to each of the 18 river flow time series. The purpose has been to get, in each case, the corresponding white noise time series, also called the "innovation" time series. As described in Sect. 3, the innovation time series is time-independent (Eq. (22)). Then the lagged cross-correlation between these series has been calcùlated, for all 8 × 18 possible pairs. Figure 2 shows a typical result, for "Sst Niño 1+2" as climatic variable and

Table 1. Climatic variables

Name	Period of record	Description
Soi	1951/1997	Southern oscillation index
Sst Niño 1+2	1950/1997	Sea surface temperature (0n–10s; 90w–80w)
Sst Niño 3	1950/1997	Sea surface temperature (5n–5s; 150w–90w)
Sst Niño 4	1959/1997	Sea surface temperature (5n–5s; 160e–150w)
Sst Niño 3.4	1950/1997	Sea surface temperature (5n–5s; 160e–90w)
V850_c	1979/1997	Trade wind index, 850 mb, Central Pacific (5n–5s; 160e–150w)
V850_e	1979/1997	Trade wind index, 850 mb, East Pacific (5n–5s; 135e–120w)
V850_w	1979/1997	Trade wind index, 850 mb, East Pacific (5n–5s; 135e–180w)

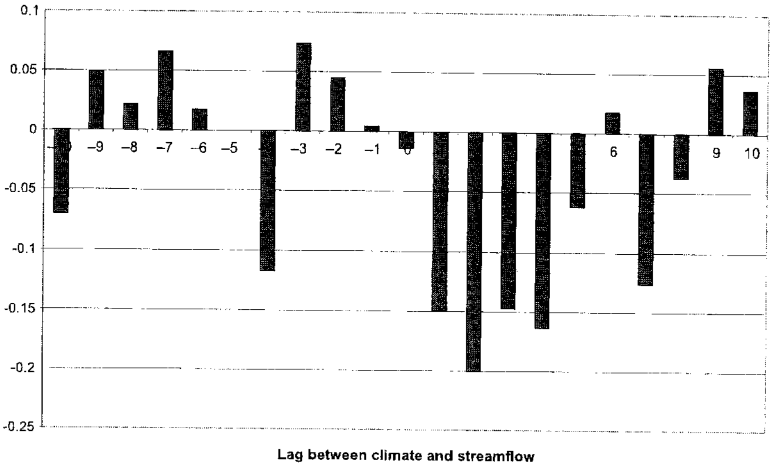

Fig. 2. Lagged cross correlation between the whitened sea surface temperature Niño 1+2 and the whitened streamflow data of Alto Anchicaya

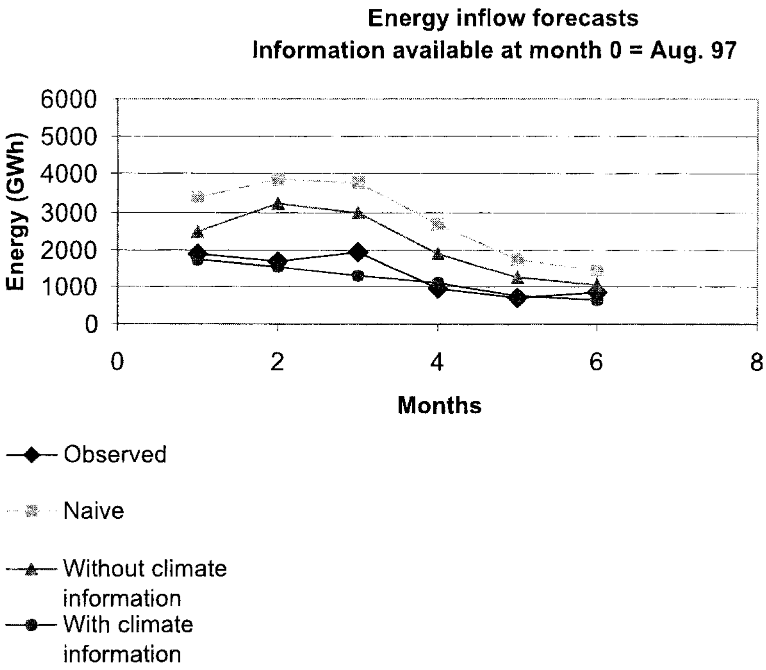

Fig. 3. Forecasts for energy inflow Sep. 97–Feb. 98 (case 1997-dry)

"Anchicaya" as the river flow variable. Anchicaya is just one out of the 18 rivers analyzed, but results for the other rivers are alike.

It can be seen that the cross correlation among residuals are rather small, very close to zero. Results for the other climatic time series have been even less encouraging. Nevertheless, for lags between 1 and 4 one can observe that the correlations, although small, are statistically different from zero: they are larger than the 95% critical value, which is 0.14. For this reason, we have selected Sst Niño 1+2 as the climatic time series. Also, for the "climatic lag" (Sect. 2), we have selected $k = 4$ (Eq. (4)).

Figure 3 shows the actual energy inflow for the Sep. 97–Feb. 98, case 1997-dry, and what would be the forecasted values for the conditions known at the end of August 97. Forecasts were done according to three models:

(a) naive model – the forecasted flow, for each month, is the mean value for the historical data available for each particular river.

(b) without climate – the forecasted flow, for each month and for each river, is obtained by the application of the PAR(p) model. That is, by successive application of Eq. (15), with residual a_t equal to zero, which is equivalent of getting the ensemble mean. Parameters are estimated without consideration for the climate information. That is, parameters are estimated by Eqs. (23)–(25).

(c) with climate – the same as (b), but parameters are estimated considering the climate information available at the end of August 97. That is, parameters are estimated by Eqs. (33)–(35).

It is quite obvious how poor the performance of the naive model would be, compared with the use of the PAR(p) model, with or without climate. It is also clear that it is much better to use climate information than not using it.

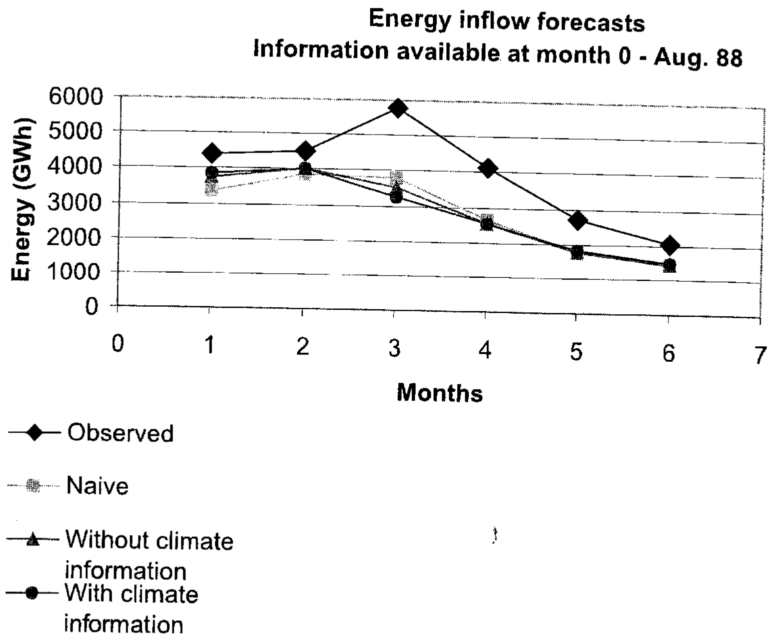

Fig. 4. Forecasts for energy inflow Sep. 88–Feb 89 (case 1988-wet)

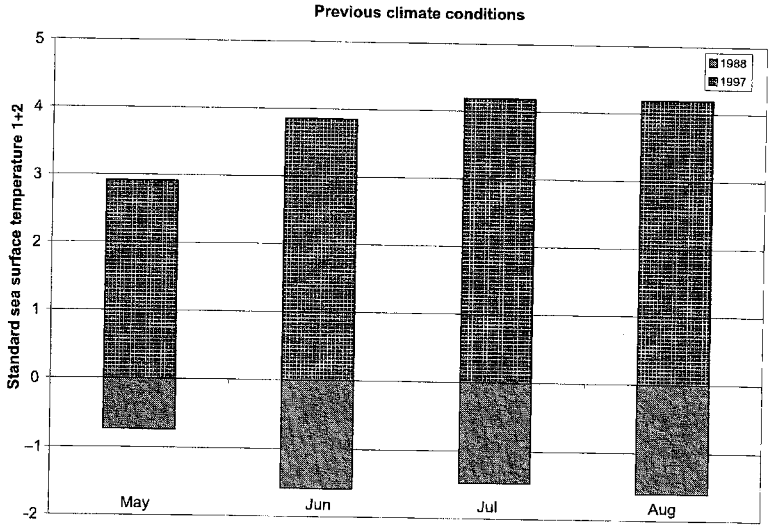

Fig. 5. Sea surface temperature Niño 1+2, for May–August

Figure 4 shows the same kind of graph, for the period of Sep. 88–Feb. 89, case 1988-wet. At first sight, it seems that there is no advantage in this case on the use of PAR(p) over the naive model for forecasting the mean flow, either with or without climate information.

One could wonder why the climatic information would have had such a substantial influence for predicting the mean energy inflow in 1997 and practically no influence in 1988. One possible explanation is given by Fig. 5. It shows the

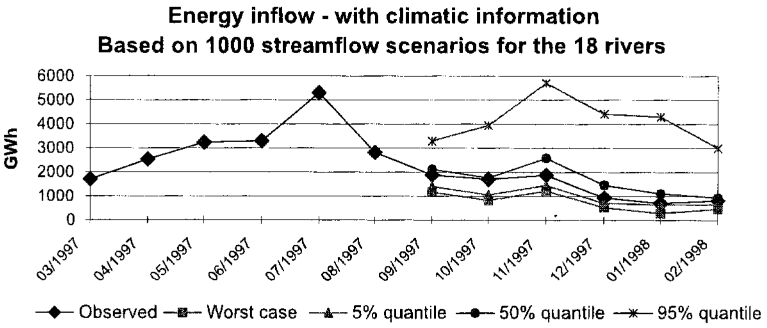

Fig. 6. Range of scenarios for energy inflow, with use of climatic SST 1+2 information. Case 1997 (dry)

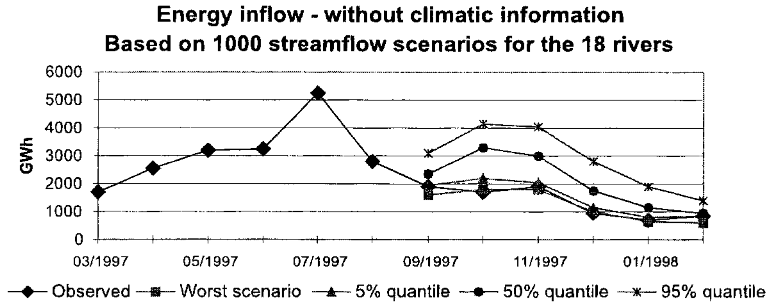

Fig. 7. Range of scenarios for energy inflow, without the use of any climatic information. Case 1997 (dry)

standard sea surface temperature Niño 1+2, for the May–August period, which would be the relevant climatic information available at the end of August. One can see that the intensity of El Niño in 1997 was much stronger than the intensity of La Nina in 1988.

Figure 6 shows the range of possible scenarios, ranked according to the total energy inflow for the period Sep. 97–Feb. 98. The scenarios were produced with the information available at the end of August of 1997: previous streamflow data, for the 18 rivers, and previous sea surface temperature Niño 1+2. The 50% quantile scenario (scenario ranked 500, out of 1000) shown in Fig. 6 should not be confused with the mean of "an infinite" number of scenarios, shown in Fig. 3. We can see in Fig. 6 that the observed values for Sep. 97–Feb. 98 - which for the sake of the production the streamflow scenarios were considered unknown - lie between the 5 and 50% quantile scenarios.

In Fig. 7 the scenarios were built using the previous streamflow data for the 18 rivers, as in Fig. 6, but without the use of any climatic information. Now the observed values for Sep. 97–Feb. 98 lie out of the band limited by the 5 and 50% quantile scenarios. In fact, in some months the observed time series is even smaller than the corresponding value for the worst (driest) scenario, out of 1000. Of course, this is evidence against the hypothesis that the observed time series was drawn from the same stochastic process that has produced the set of streamflow scenarios.

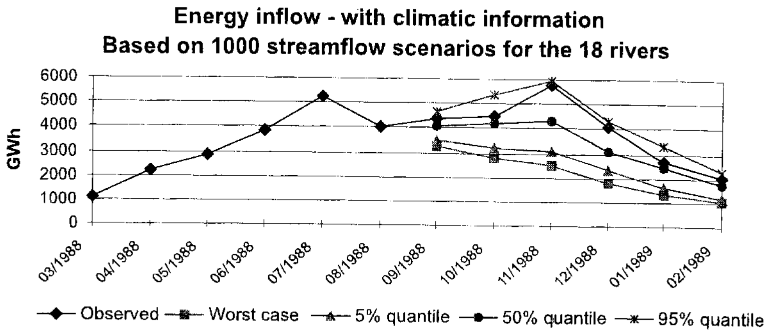

Fig. 8. Range of scenarios for energy inflow, with use of climatic SST 1+2 information. Case 1998 (wet)

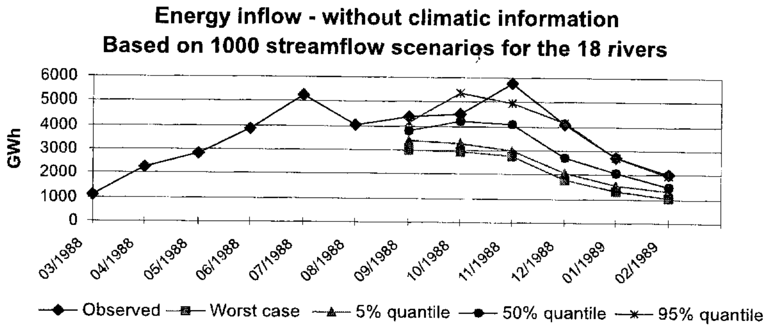

Fig. 9. Range of scenarios for energy inflow, without the use of any climatic information. Case 1998 (wet)

Figures 8 and 9 are the equivalent of Figs. 6 and 7, now for the 1988 case (wet). It can be seen that the observed values for Sep. 88–Feb. 89 lie in and out the band limited by the 5 and 50% quantile scenarios, respectively for scenarios built with (Fig. 8) and without (Fig. 9) the use of climatic information. By inspection one can see that in case 1988 (wet), as in case 1997 (dry), the set of scenarios built using climatic information has higher likelihood than the set built without the use of climatic information.

Figures 10 and 11 show the cumulative distribution function (cdf) for the total energy inflow for the September–February time interval, respectively for the case 1997 (dry) and 1988 (wet). Again, these cdfs were constructed based on sets of 1000 streamflow scenarios, for the 18 rivers, which were produced by the PAR(p) model, with and without using climatic information.

It can be seen in Fig. 10 that the probability of occurrence of what actually became reality would be close to zero, according to the cdf of total energy inflow derived from 1000 scenarios produced without climatic information. In reality there are only 8 scenarios, out of 1000 that had a total energy inflow for the period smaller than the observed value of 7997 GWh. In the context of hypothesis testing, this would be sufficient condition for rejecting this set of scenarios. On the other hand, the same probability would be close to 0.30, a very reasonable value, according to the cdf derived from 1000 scenarios produced with climatic information.

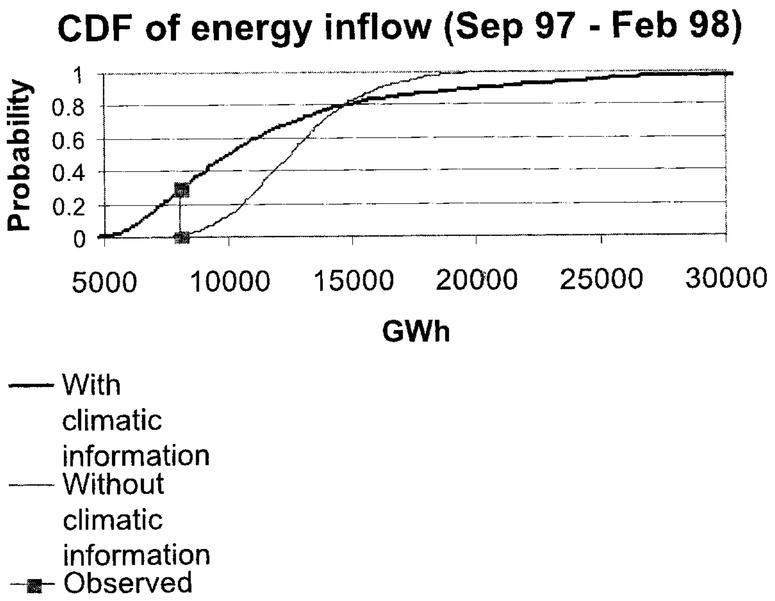

Fig. 10. Cumulative distribution function for the total energy inflow of case 1997 (dry)

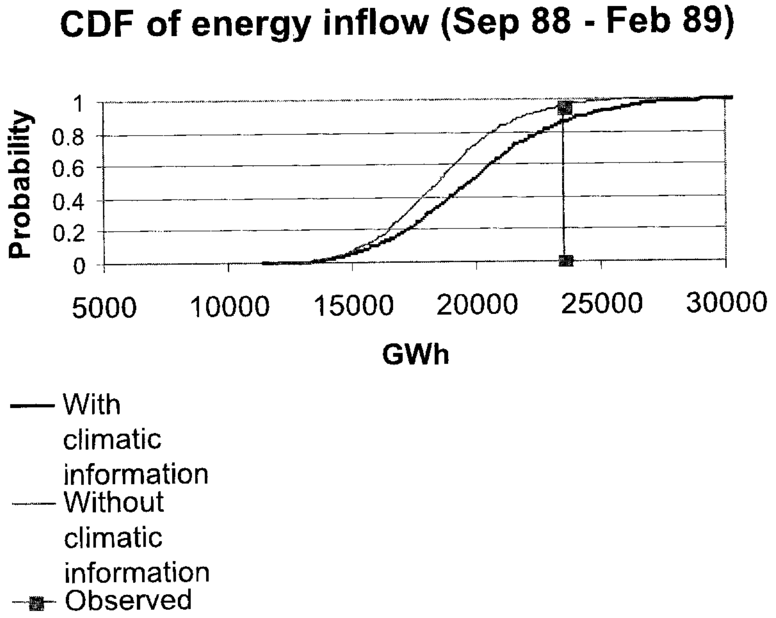

Fig. 11. Cumulative distribution function for the total energy inflow of case 1998 (wet)

It can be seen in Fig. 11 that the probability of occurrence of what actually became reality would be too close to 1 (actually 0.96), according to the cdf of total energy inflow derived from 1000 scenarios produced without climatic information. Again, in the context of hypothesis testing, this could be a reason for rejecting this set of scenarios. On the other hand, the same probability would be

0.86, according to the cdf derived from 1000 scenarios produced with climatic information.

Given the null hypothesis "the observed time series was drawn from the same stochastic process that produced the set of scenarios", and probability of Type I error of 10% for the test, the null hypothesis should be rejected for the set of scenarios built without the use of climatic information. On the other hand, the null hypothesis would not be rejected if parameters for the PAR(p) model were estimated with the use of climatic information.

6 Conclusions

It has been described a methodology for using climatic information for the estimation of parameters of a stochastic model for streamflows. The case study has shown that the use of this methodology can make a great difference, with practical implications. The reliability of river flow scenarios produced by the model, PAR(p), has been evaluated through the analysis of an aggregated time series, the inflow energy to the system. It has been shown that streamflow scenarios produced using climatic information were more likely to occur than the other way around.

References

Bras RL, Rodriguez-Iturbe I (1985) Random Functions and Hydrology. Addison-Wesley Publishing Co., Reading, USA

Box GEP, Cox DR (1964) An analysis of transformations. J. Roy. Statist. Soc., B, v. 26

Cadavid-Mazo E, Maya-Sánchez E, Chaparro-Villamizar N (1998) Experiencia colombiana en el evento cálido El Niño 1997–1998. Seminário Latino-Americano sobre os Impactos do El Niño/La Niña na Gestão de Recursos Hidricos em Sistemas Hidrelétricos, Rio de Janeiro, Brazil

Johnston J (1963) Econometrics Methods. McGraw-Hill, New York, USA

Kelman J (1987) in Modelos para Gerenciamento de Recursos Hídricos. Coleção ABRH de Recursos Hídricos. Vol. 1, Ed. Nobel, Brazil

Kelman J (1997) GESS – Gerador Estocástico de Séries Sintéticas, Kelman Consultoria, Rio de Janeiro, Brazil

Liebmann B (1998) An Overview of El Niño, La Niña, and the Southern Oscillation. Seminário Latino-Americano sobre os Impactos do El Niño/La Niña na Gestão de Recursos Hídricos em Sistemas Hidrelétricos, Rio de Janeiro, Brazil

Maceira MEP (1989) Operação Ótima de Reservatórios com Previsão de Afluências. Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro, Brazil

McLeod AI (1994) Diagnostic Checking of Periodic Autoregression Models with Application. J. Time Ser. Anal. 15(2), 221–233

Morrison DF (1967) Multivariate Statistical Methods, Mc-Graw-Hill Book Company

Pereira MVF, Pinto LMVG (1991) Multi-stage stochastic optimization applied to energy planning. Math. Prog. 52(2), 359–375

Piechota TC, Dracup JA (1996) Drought and regional hydrologic variation in the United States: Associations with the El Niño-Southern oscillation. Water Resources Research 32(5), 1359–1373

Poveda G, Mesa JO (1996) Caudales Medios Mensuales en 50 Ríos Colombianos durante El Niño y La Niña, XII Congreso Colombiano de Hidráulica y Hidrologia, Bogota, Colombia

Salas JD, Delleur JW, Yevjevich V, Lane WL (1980) Applied Modeling of Hydrologic Time Series. Water Resources Publication, Littleton, CO, USA

Todini E (1980) The preservation of skewness in linear disaggregation schemes. J. Hydrol. 47