Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 823046, 8 pages
http://dx.doi.org/10.1155/2014/823046

Research Article

Long Memory Models to Generate Synthetic Hydrological Series

Guilherme Armando de Almeida Pereira and Reinaldo Castro Souza

Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), 22451-900 Rio de Janeiro, RJ, Brazil

Correspondence should be addressed to Guilherme Armando de Almeida Pereira; gaap@ele.puc-rio.br

Received 21 April 2014; Revised 19 June 2014; Accepted 27 June 2014; Published 17 July 2014

Academic Editor: Nazim I. Mahmudov

In Brazil, much of the energy production comes from hydroelectric plants whose planning is not trivial due to the strong dependence on rainfall regimes. This planning is accomplished through optimization models that use inputs such as synthetic hydrologic series generated from the statistical model PAR(p) (periodic autoregressive). Recently, Brazil began the search for alternative models able to capture the effects that the traditional model PAR(p) does not incorporate, such as long memory effects. Long memory in a time series can be defined as a significant dependence between lags separated by a long period of time. Thus, this research develops a study of the effects of long dependence in the series of streamflow natural energy in the South subsystem, in order to estimate a long memory model capable of generating synthetic hydrologic series.

1. Introduction

It is known that, in Brazil, even with the growing diversification of its energetic matrix, approximately 85% of the potential to generate energy comes from hydroelectric plants. One of the chief characteristics of an energetic matrix like this one is its strong dependence on the rainfall regime. This causes an uncertainty in the operationalization of the system, making its planning far from trivial. The *Operador Nacional do Sistema Elétrico* (ONS) is in charge of the planning and operationalization of the Brazilian electric system. In these activities (area), the models adopted are the ones that simulate and/or optimize the operation; these models make use of the predicted and/or simulated natural streamflow as input to obtain outcomes that indicate the most adequate situations of storage, of water release from reservoir, and of hydropower generation in each time interval.

Within the diverse research related to the optimization phase we mention [1], which uses stochastic dynamic programming applied to the planning of long term electric power system operation. At the same time, estimating models that are capable of predicting and/or simulating the hydrological series are also of great importance to the optimal planning of the system. It is widely known that small advances in such models are capable of making possible the improvement of the system operation planning, something that is directly converted into investment savings, low tariffs, and a better use of the available system resources. This justifies the high investment that has been made by the sector.

In Brazil, medium term operation planning uses the computational tool called NEWAVE in which the planning is represented by a multistage stochastic linear programming problem whose objective is to minimize the total operation cost.

In order to do that, one of the main inputs is a set of synthetic hydrological series generated from the affluent natural energy history (ANE) (the ANE used is computed from the affluent natural streamflows and from the productibilities equivalent to the storage of 65% of the useful volume of the hydropower reservoirs [2]) of each of the four Brazilian subsystems (Southeast/Midwest, South, Northeast, and North). For such generation, the stochastic model used is an extension of the ARMA(p, q) model called PAR(p) (periodic autoregressive) [3–7]. The PAR(p) is used in time series that show a structure of autocorrelation that does not depend only on the time interval between the observations, but also on the observed period. This way, for each period, one AR(p) model is adjusted; that is, if it is a monthly series, 12 AR(p) models are adjusted with a p order not necessarily equal. The PAR(p) has been widely used for the generation

of synthetic series, but, recently, a search has begun for new statistical tools capable of generating synthetic hydrological series, among them [6–10].

Using methods capable of grasping the effects that the PAR(p) model is not able to estimate is one of the reasons for searching for new scenario-generating models. Among these effects mainly the long memory and/or cyclic ones are worth noting.

Long memory, or persistence in a time series, can be defined as the presence of dependence among observations very distant in time, different from the traditional models where the correlation among observations separated by a long period of time is considered nil or negligible. The ARFIMA model [11–14] has become one of the most popular tools to model series with this property. This model is an extension of models from the Box & Jenkins family, where the differentiation d can take fractional values. This adaptation is made in order to enable the capturing of long memory effects present in the time series.

The aim of this paper is the generation of synthetic hydrological series through long memory models applied to the affluent natural energy (ANE) series of the South subsystem. Using a nonparametric bootstrap test it was possible to demonstrate the existence of such effects in the analyzed series, thus justifying the use of this model. In the simulation of scenarios, we used the bootstrap technique in the residuals of the fitted models. Finally, we evaluated these scenarios through a set of statistical tests.

The remainder of the paper is organized as follows. Section 2 introduces the ARFIMA model. Section 3 describes the bootstrap techniques and also their two different uses in this research. Section 4 outlines the set of tests used to evaluate synthetic hydrological scenarios. Our case study is presented in Section 5 and in the last part; some final remarks will be made together with suggestions for further research.

2. ARFIMA Model

According to [15], the property of long memory is characterized by the fact that a spectral density is unbounded in the neighborhood of zero frequency. In this case, the spectral density behaves like

$$f(\omega) \sim C|\omega|^{-2d} \tag{1}$$

for $\omega \to 0$ and some positive constant C. The autocorrelation function decays hyperbolically

$$\rho(k) \sim |k|^{2d-1} \tag{2}$$

as $|k| \to \infty$. If $0 < d < 0.5$, we say that the process has long memory while if $d < 0$ we say that the process has intermediate memory. When $d = 0$, the process is short memory.

In order to represent these characteristics, [11, 16] developed the ARFIMA(p, d, q) model, which is a generalization of ARIMA models in the case where d assumes any real value. This is one of the most flexible and comprehensive long memory models in the literature.

We say that Z_t is an ARFIMA model if it satisfies

$$\phi(B)(1-B)^d Z_t = \theta(B)\, a_t, \tag{3}$$

where $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ and $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 B - \cdots - \theta_q B^q)$ are polynomials with all roots outside the unit circle. B is the back-shift operator and d is the fractional parameter that governs the memory of the process. a_t is a white noise process with zero mean and variance σ^2. The operator of fractional differentiation, $(1 - B)^d$, can be visualized in the following equation:

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = 1 - dB - \frac{d}{2!}(1-d)B^2 - \cdots. \tag{4}$$

The model is stationary and invertible if, and only if, $d \in (-0.5; 0.5)$ and display long memory property when $d \in (0; 0.5)$. The spectral density function of (3) is given by

$$f(\omega) = f_u(\omega)\left[2\sin\left(\frac{\omega}{2}\right)\right]^{-2d}, \quad \omega \in [-\pi, \pi], \tag{5}$$

where the function $f_u(\cdot)$ is the ARMA process' spectral density. More details about the ARFIMA models can be found in [11–14].

2.1. Estimation Method of ARFIMA Models. Several estimators exist in the literature for the parameters of ARFIMA and basically they can be divided into two categories (parametric methods and semiparametric methods). At first, all parameters are estimated simultaneously, usually based on the likelihood function. This class contains the estimators proposed by [17, 18]. On the other hand, in the semiparametric approach, the estimation is performed in two steps: first, we estimate d (using, e.g., the log-periodogram regression), and subsequently we estimate the autoregressive and the moving average parameters after the series have been differentiated using (4) and the estimated parameter $\widehat{d}$. In this class we can mention the estimators proposed by [19] and variations thereof such as those proposed by [12, 20] among others.

In this paper, we have adopted the semiparametric method proposed by [12] which is a variation of the method proposed by [19] named GPH. The GPH method can be obtained by taking the logarithm of (5). Consider

$$\ln f(\omega) = \ln f_u(\omega) - d \ln\left[2\sin\left(\frac{\omega}{2}\right)\right]^2 \tag{6}$$

which can be rewritten as

$$\ln f(\omega) = \ln f_u(0) - d \ln\left[2\sin\left(\frac{\omega}{2}\right)\right]^2 + \ln\left\{\frac{f_u(\omega)}{f_u(o)}\right\}. \tag{7}$$

The GPH method uses, as an estimator of the spectral density $f(\omega)$, the periodogram function, $I(\omega)$, given by

$$I(\omega) = \frac{1}{2\pi}\left[R(0) + 2\sum_{k=1}^{n-1} R(k)\cos(k\omega)\right], \tag{8}$$

where $R(\cdot)$ represents the sample autocovariance of Z_t and n is the sample size. Substituting in (7) ω for $\omega_j = 2\pi j/n$ and adding $\ln I(\omega_j)$ we obtain

$$\ln I\left(\omega_j\right) = \ln f_u(0) - d\ln\left[2\sin\left(\frac{\omega_j}{2}\right)\right]^2 + \ln\left\{\frac{f_u(\omega)}{f_u(o)}\right\} + \left\{\frac{I\left(\omega_j\right)}{f\left(\omega_j\right)}\right\}. \tag{9}$$

Considering the upper limit of j is equal to $g(n)$, which must be chosen satisfying $(g(n)/n) \to 0$ when $n \to \infty$, the term $\ln\{f_u(\omega)/f_u(o)\}$ can be considered negligible when compared with other terms. Therefore, we get an equation close to (9):

$$\ln I\left(\omega_j\right) \cong \ln f_u(0) - d\ln\left[2\sin\left(\frac{\omega_j}{2}\right)\right]^2 + \left\{\frac{I\left(\omega_j\right)}{f\left(\omega_j\right)}\right\}. \tag{10}$$

This equation is similar to a regression equation having the spectral density as the dependent variable. In other words, we have

$$Y_j \cong a + b_1 X_{1j} + \varepsilon_j, \quad \forall j = 1, \ldots, g(n), \tag{11}$$

where $Y_j = \ln I(\omega_j)$, $a = \ln f_u(0) - c$, $X_{1j} = \ln[2\sin(\omega_j/2)]^2$, $b_1 = -d$, $\varepsilon_j = \ln(I(\omega_j)/f(\omega_j)) + c$, $c = E - [\ln(I(\omega_j)/f(\omega_j))]$. Thus, the estimated parameter $\widehat{d}$ is obtained by the regression of $\ln I(\omega_j)$ (dependent variable) and $\ln[2\sin(\omega_j/2)]^2$ (independent variable) using ordinary least squares.

Because the periodogram is not a consistent estimator of the spectral density function, [12] suggests replacing the periodogram function by its smoothing version based on the Parzen lag window. Consider

$$f_{sp}(\omega) = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1}\lambda(k)\,R(k)\cos(k\omega). \tag{12}$$

The term $\lambda(k)$ is the Parzen lag window defined by

$$\lambda(s) = \begin{cases} 1 - 6\left(\frac{k}{m}\right)^2 + 6\left(\frac{|k|}{m}\right)^3, & |k| < \frac{m}{2}, \\ 2\left(1 - \left(\frac{|k|}{m}\right)\right)^3, & \frac{m}{2} \le |k| \le m, \\ 0, & |k| > m, \end{cases} \tag{13}$$

where $m = n^\beta$ and $0 < \beta < 1$. Again, the estimated parameter $\widehat{d}$ is obtained by regression between $\ln f_{sp}(\omega)$ and $\ln[2\sin(\omega/2)]^2$. In both these methods, the number of observations in the regression is determined by $g(n) = n^\alpha$, where α is a constant between zero and one and n is the length of the time series.

3. Bootstrap

Bootstrap is a computationally intensive, nonparametric statistical technique of resampling, introduced by [21, 22], and has the purpose of obtaining information about the characteristics of the distribution of some random variable. To do this, a probability distribution is approximated through an empirical function obtained from a finite sample. This technique is generally deployed when the concerned distribution is difficult, or even impossible, to be analytically evaluated or when just the asymptotic theory is available.

In this paper, bootstrap is applied in two distinct situations. The first application will aim to verify the statistical significance of the fractional parameter. Thus, bootstrap is used to approximate the probability distribution function of the parameter. In a second step, bootstrap will be employed in the residuals of the fitted model in order to simulate new time series, that is, simulate ANE's scenarios. Both approaches are described below.

3.1. Nonparametric Test. This nonparametric test is based on the bootstrap distribution of the parameter d and is intended to infer the statistical significance thereof. The bootstrap distribution of interest is obtained by applying bootstrap in the residuals of the regression (11) used for the parameter estimation. Several studies are being conducted regarding the use of bootstrap for estimation and approximation of the probability distribution of the parameter d; among them we can mention [15, 23, 24].

This procedure can be summarized as follows [15, 23].

(1) Using (10), which is similar to a regression equation (11), estimate $\widehat{d}$ by least squares and calculate the residuals $\widehat{e}_j$, $j = 1, \ldots, g(n)$.

(2) Resample $\widehat{e}_j$ with replacement to obtain a new sample of residuals $\widehat{e}_j^*$.

(3) Apply these residuals in (14) to obtain a sample $\ln f_{sp}^*(\omega_j)$. The terms without asterisk are held fixed:

$$\ln f_{sp}^*\left(\omega_j\right) = \ln f_u\left(\omega_j\right) - \widehat{d}\ln\left[2\sin\left(\frac{\omega_j}{2}\right)\right]^2 + \widehat{e}_j^*. \tag{14}$$

(4) Using $\ln f_{sp}^*(\omega_j)$ and $\widehat{d}\ln[2\sin(\omega_j/2)]^2$, calculate the new parameter $\widehat{d}$.

Steps (2)–(4) must be repeated B times in order to build a bootstrap distribution for the parameter d.

In possession of the bootstrap distribution, we are ready to make inferences about the desirable parameter. For this, confidence intervals for the parameter d will be constructed. The confidence interval to be used here is the one proposed by [22], based on the percentiles of the estimated bootstrap distribution.

Adopting $\widehat{G}$ as the accumulated distribution function of d, the percentile interval with coverage probability of $1 - 2\alpha$ is determined by the percentiles α and $1 - \alpha$ of the bootstrap distribution of d. This way, the lower bound is given by $\widehat{G}^{-1}(\alpha)$ and the upper bound is given by $\widehat{G}^{-1}(\alpha - 1)$; that is,

$$\left[\widehat{d}_{\text{lo}}(\alpha), \widehat{d}_{\text{up}}(\alpha - 1)\right] = \left[\widehat{G}^{-1}(\alpha), \widehat{G}^{-1}(\alpha - 1)\right]. \tag{15}$$

With this interval, the inference is simple. In case zero belongs to the interval, it is possible to say that the parameter statistically equals zero; in case it is out of it, it is assumed that the parameter is different from zero.

3.2. Bootstrap in the Residuals of the Fitted Model for the Simulation of Hydrological Scenarios. Bootstrap's second application in this paper is related to the simulation of ANE's synthetic series. For scenarios simulation, bootstrap will be carried out in the residuals of the fitted ARFIMA model. Based on a fitted ARFIMA model, random choices with replacement of the residuals are made and, for each error chosen, a new observation in the series is generated.

Model's equation can be obtained by solving the equation of differences expressed in (3) using the estimated parameters $\widehat{d}$, $\widehat{\phi}$, and $\widehat{\theta}$. As it could be observed in the equation mentioned, there are two polynomials with infinite order. In practical terms, when one has a historical series with n observations, only the first K terms of this polynomial are used, with $K \leq n$.

As previously described, and adopting K equal to 936, since all the available historical record will be used to the simulation, the model's equation is given by

$$Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_{935} Z_{t-935} + \varepsilon_t^*. \quad (16)$$

Thus for each time t, a residual with replacement was resampled and inject it into (16) to obtain a new realization of Z_t. This procedure was performed repeatedly until the desired size of the simulated scenario (T) was reached.

4. Evaluation of the Performance Model

It is desirable that the synthetic scenarios preserve the main characteristics of the historical series. This means that the utility of the model can be verified by its ability to reproduce some characteristics present in the time series. This way, with the aim of verifying whether the model used is capable of reproducing the statistical proprieties of the historical series, this section presents all the statistical tests that make up the assessment module of scenarios.

4.1. t-Test. The t-test is, for sure, the most important test to be verified. It compares monthly the mean of the scenarios with the mean of the historical record. That is, this test has as its objective the comparison between the scenarios' monthly mean and the historical record's monthly mean. In order to do so, the Januaries generated are compared with the Januaries of the historical record; the same goes, subsequently, for all months. The null and alternative hypotheses are given by

$$H_0 : \mu_{\text{hist}}^{(m)} = \mu_{\text{sce}}^{(m)},$$
$$H_1 : \mu_{\text{hist}}^{(m)} \neq \mu_{\text{sce}}^{(m)}. \quad (17)$$

In the above equation m indicates the month of the test. Thus, the null hypothesis says that, for a given month, the historical mean is statistically equal to the mean of the scenarios.

The analysis will be presented through the P values of the tests that must be above the adopted significance level, so that the null hypothesis is not rejected. To correctly perform a t-test for comparison of two means, the bootstrap t-test for equality of means proposed by [22] is used.

4.2. Adherence Analysis. The statistical tests used to verify the form of the probability distribution of the interest variables are now presented. The goal is to investigate whether variables from the synthetic scenarios and variables from historical time series have the same probability distribution. The tests used are the Kolmogorov-Smirnov and the Chi-Square ones. The former intends to determine whether the two samples have the same probability distributions. In this case, the objective is to verify if the probability distributions of the scenarios generated are equal to the distribution of the historical record. The null and alternative hypotheses are defined as follows:

$$H_0 : F_{\text{hist}}^{(m)}(x) = F_{\text{sce}}^{(m)}(x),$$
$$H_1 : F_{\text{hist}}^{(m)}(x) \neq F_{\text{sce}}^{(m)}(x). \quad (18)$$

This analysis is also done through the P values, and values above the significance level indicate that the null hypothesis cannot be rejected.

Initially, this test is applied to the distribution of each of the periods of the scenarios with the corresponding period in history, as well as the t-test presented earlier. For example, it tests if the probability distribution for the simulated months of January is equal to the distribution of historical January distribution. In addition, we also used the Kolmogorov-Smirnov test to check whether the distributions of variables *sequence sum* and *sequence intensity*, which will be introduced in Section 4.3, are statistically equal.

Regarding the Chi-Square test, this is used to evaluate how close the observed frequency is to the expected frequency. The null and alternative hypotheses are, respectively,

$$H_0 : \text{there is no difference between the observed and the expected frequencies,}$$
$$H_1 : \text{there is a difference between the observed and the expected frequencies.} \quad (19)$$

We apply this test in the variable *sequence length* (Section 4.3). Thus, the Chi-Square test is used to verify that the expected frequency of *sequence length* variable in the scenarios is statistically equal to the observed frequency of this variable in the original ANE.

4.3. Sequence Analysis. For the sequence analysis, new random variables were created to verify the capacity of the models in reproducing the frequencies observed in the historical record. The random variables introduced are related to the representation of critical periods, such as droughts and floods registered in the ANE. This way, the concepts of

Table 1: Variables to negative sequence analysis.

Random variable	Description	Calculus
Sequence length	Corresponds to the length of the intervals $(t_1 - t_2)$ and $(t_3 - t_4)$	$C = (t_2 - t_1)$
Sequence sum	Corresponds to the area below the limit during the sequence. In the previous figure, it is equivalent to the areas $A1$ and $A2$	$S = \sum_{i=t_1}^{t_2} (Z_i - \mu_i)$
Sequence intensity	Corresponds to the average value below the limit, that is, to the sequence sum divided by the respective sequence length	$I = \frac{S}{C} = \frac{\sum_{i=t_1}^{t_2} (Z_i - \mu_i)}{(t_2 - t_1)}$

Source: [25].

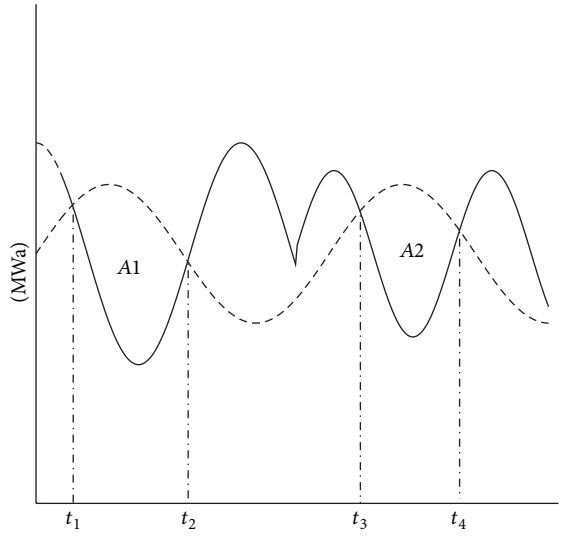

Figure 1: Representation of negative sequence. Source: adapted from [25].

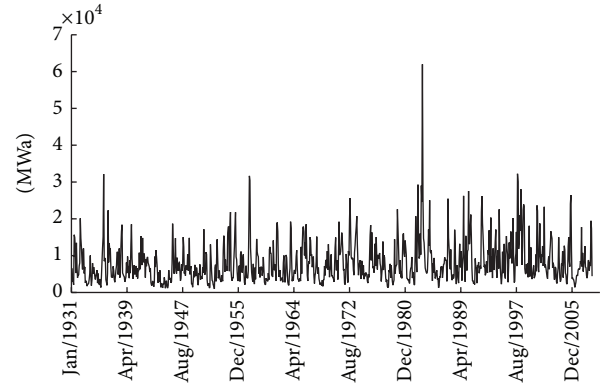

Figure 2: Affluent natural energy, South subsystem.

negative sequence and positive sequence are used. A negative sequence is defined by a long period of time in which the streamflows are continually below the predetermined values, while a positive sequence is determined by a period of time in which the streamflows are continually above the predetermined values. In this paper, the predetermined limits were the monthly averages.

The concept of sequence can be understood by observing Figure 1, where the continuous line represents an ANE hypothetical series and the dotted line represents a predetermined limit. As it is possible to be visualized, the intervals $(t_2 - t_1)$ and $(t_4 - t_3)$ represent negative sequences, while the interval $(t_3 - t_2)$ is an example of a positive sequence.

From each negative sequence found, three variables can be created both to the historical record and to the synthetic scenarios: *length*, *sum*, and *intensity* of negative sequence. With two samples of each variable, it is possible to verify whether the samples have the same distribution through the statistical adherence tests. The variable *sequence length* is evaluated by the Chi-Square test, while the *sum* and *intensity* variables are evaluated by the Kolmogorov-Smirnov test.

Similarly, to each positive sequence calculated, the *length*, *sum*, and *intensity* variables of positive sequence are obtained and tested to see whether the samples found (historical values and synthetic scenarios) have the same distribution. These variables, to negative sequence (and in a similar way to positive sequence), are defined as shown in Table 1.

5. Results

This section presents the obtained results. The ANE used, Figure 2, is a monthly one relative to the South subsystem, starting in January 1931 and ending in December 2008, totaling 936 observations. The ANE is computed from the natural streamflows and the possible production equivalent for the storage of 65% of the useful volume of the hydropower reservoirs [2].

5.1. Fitted Model. The estimated model will now be presented. Then, both the estimated fractional parameters and the short memory parameters can be observed. In addition to this the results of the nonparametric test used to determine the statistical significance of the fractional parameter are shown. In the end, in Figure 3, the autocorrelation function of the residuals can be visualized.

Table 3 contains the estimated value for the long memory parameter $\widehat{d}$ and the short memory parameters. The estimated parameters satisfy the stationary and invertible conditions and also the long memory propriety. To the definition of α and β, where $g(n) = n^{\alpha}$ (to regression) and $m = n^{\beta}$ (Parzen's window), we used $\alpha = 0.5$ and $\beta = 0.9$ [12, 15, 19].

Furthermore, to determine the statistical significance of the long memory parameter, a nonparametric bootstrap test was employed. Confidence intervals for different coverage

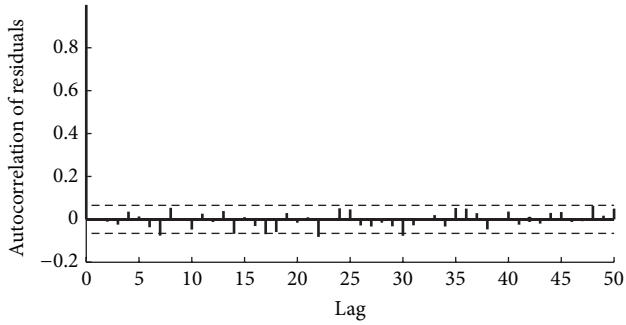

Figure 3: Autocorrelation function of the residuals.

Table 2: Confidence intervals to d.

	Lower bound	Upper bound
$\alpha = 0.05$	0.0509	0.2049
$\alpha = 0.025$	0.0363	0.2214
$\alpha = 0.005$	0.0123	0.2496

Table 3: Estimated parameters.

	$\widehat{d}$	$\widehat{\phi}_1$	$\widehat{\phi}_{12}$
South	0.1253	0.4008	0.1655
	$(0.0518)^2$	(0.0299)	(0.0324)

[2]Asymptotic standard deviation.

probability are displayed in Table 2. The number of bootstrap samples was ten thousand; that is, $B = 10000$. With these intervals, the analysis can be done in a simple way. If zero is contained in this interval, it can be said that the parameter is statistically equal to zero; otherwise it can be said that the parameter is statistically different from zero.

Through the analysis of the intervals, Table 2, it is proven that the estimated parameter is statistically different from zero, which indicates the existence of the effects of long memory in the series analyzed, thus justifying the use of ARFIMA.

After parameter $\widehat{d}$ estimation, following the semiparametric method of ARFIMA models construction, it is necessary to differentiate the time series and estimate the autoregressive and the moving average parameters via maximum likelihood. Regarding the selection of orders p and q of the AR and MA parts the BIC (Bayesian information criterion) was used.

The order identified was $p = P = 1$ and $q = Q = 0$; that is, the model identified is SARFIMA$(1, \widehat{d}, 0)(1, 0, 0)_{12}$. In Table 3, it is possible to see all the estimated parameters and their standard errors.

Finally, Figure 3 presents the autocorrelation function of the residuals. As can be seen, the errors are uncorrelated indicating that the proposed model was able to capture all of the existing structures of temporal dependence in the series of ANE South subsystem.

5.2. Scenarios Generation. The simulated ANE scenarios as well as their monthly averages (dotted black line) and historical averages (red line) are shown in Figure 4. In total, simulated 200 scenarios were simulated where each scenario has 60 months, which corresponds to 5 years. Through the graphical analysis, it can be stated that the averages of the synthetic scenarios are similar to the historical average and almost overlap each other. It can also be seen that the synthetic scenarios correctly reflect the hydrological periods; that is, the scenarios reproduce high ANE in rainy periods and low ANE in dry ones. This is a highly desirable feature of a simulation model, especially in regions where there is a large difference in available water between the seasons, as is the case of Brazil.

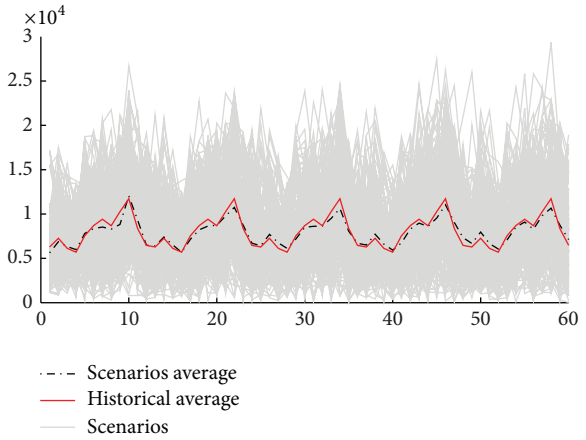

Figure 4: Scenarios.

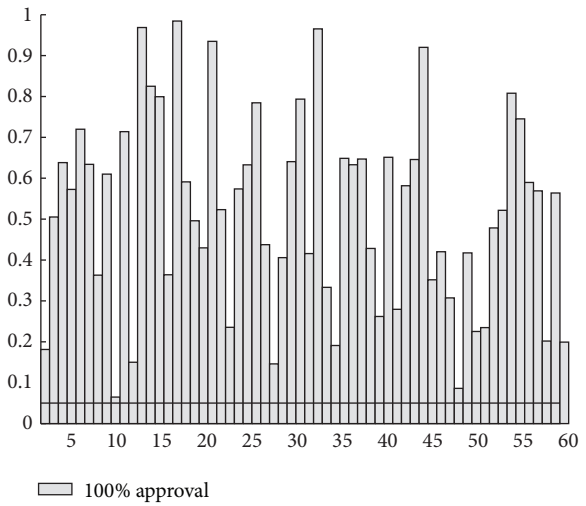

Figure 5: t-Test.

Regarding the statistical tests conducted to assess the simulated scenarios, Figures 5 and 6 present the P values of the t-test and the Kolmogorov-Smirnov test. All the analysis done took into account a significance level of 5%, which is represented in the figures by the continuous black line. So, P values above this line mean that the null hypothesis must not be rejected. It is worth emphasizing that the analysis done between the generated scenarios and the historical record was monthly; that is, it was checked whether each generated

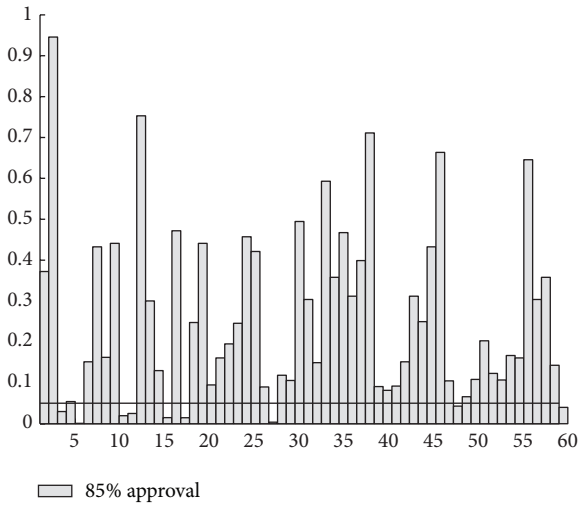

Figure 6: Kolmogorov-Smirnov's test.

Table 4: Scenarios assessment.

	Approval rating
t-Test	100%
Kolmogorov-Smirnov's test	85%

month had the interest variable statistically equal to the equivalent period in the historical record, in other words, whether the synthetic Januaries were equal to the Januaries from the historical series and the same for all the 60 periods generated.

In relation to the *t*-test, the most important one, the approval rating was 100%; that is, 100% of the 60 months tested had their mean equal when compared to the historical one. This shows that the proposed model is capable of satisfactorily reproducing the first moment of the historical series. Taking into account the Kolmogorov-Smirnov test, which verifies whether the months generated by the model come from the same distribution as the historical months, that is, whether both have equal probability distributions, the approval rating was 85%. Hence, it can be said that the synthetic scenarios and the historical series have, in most of the months, the same probability distribution. The results of both tests are presented in Table 4.

Complementing the analysis done, the results for assessing the capacity of the scenarios generated in reproducing the critical periods observed in the historical record are presented. The aim is to evaluate whether the scenarios reproduce each variable's probability distributions, comparing them to the respective historical distribution. The variables deployed were previously defined and are the following: *sequence length*, *sequence sum*, and *sequence intensity*. The tests done for the last two variables were the Kolmogorov-Smirnov ones, while the Chi-Square test was used for the first one.

The Kolmogorov-Smirnov test analysis takes place through the *P* values which must be superior to the significance level adopted (5%), in order to avoid the null hypothesis being rejected. Concerning the Chi-Square test, the analysis is done based on the statistical test that must be inferior to the critical value.

In Table 5, the results obtained for the sequence analysis are displayed. Regarding the positive sequence analysis, the *sum* and *length* variables of the simulated scenarios are statistically equal to the time series while the ENA intensity variable is different.

This indicates that the used model is capable of reproducing the critical rainy periods (high ANE) observed in the historical record. Taking into account the negative sequence analysis, the *sum* and *intensity* variables were adherent to the historical record. On the other hand, the *length* variable showed statistically significant differences between the historical record and the simulated scenarios. Based on the results obtained, it is possible to conclude that the proposed model is able to reproduce the critical periods found in the historical ANE.

6. Conclusions

The goal of this paper was to study the phenomenon of long dependence in the affluent natural energy series of the South subsystem, in order to create a model for generating synthetic series of ANE.

Bootstrap was used in two distinct purposes. At first, bootstrap was used in the preparation of a nonparametric test to verify the statistical significance of the fractional parameter. Thus, we constructed confidence intervals for the fractional parameter that allowed us to infer that the parameter is statistically significant. In the second time the bootstrap was performed on the residuals of the fitted ARFIMA model with the purpose of simulating new synthetic hydrological series.

Regarding the simulated scenarios, a set of tests to evaluate the simulated scenarios was employed. The aims of these tests were to investigate if the synthetic series preserve several existing features in the historical ANE.

In relation to the statistical tests, the *t*-test, the model obtained a very satisfactory approval rating (100%), and the Kolmogorov-Smirnov test had an approval rating of 85%. This indicates that the simulated scenarios maintain the patterns observed in the history.

In the sequence analysis, where the aim was to evaluate the capacity of the model in creating critical periods stricter than those observed in the historical record, the results can be considered acceptable. In the negative sequence test, three variables were tested and only one did not show adherence between the scenarios and the historical record. In the negative sequence test, the pattern is repeated, or, only one did not show adherence between the scenarios and the historical record. That said, it can be stated that the methodology used is capable of incorporating long dependence effects and generating synthetic series different from the historical record.

Finally, due to evidences of long memory presence and to the good performance in regard to the synthetic hydrological

Table 5: Sequence analysis (Kolmogorov-Smirnov and Chi-Square).

	Length Critical value: 3.84	Sum Min. P value: 0.05	Intensity Min. P value: 0.05
Negative sequence	6.10	0.598	0.586
Positive sequence	0.71	0.175	0.001

series generation, new studies about the long/cyclical memory phenomena must be done for the ANE series of the South subsystem. Further research on other long memory models and/or improvements in the model used could be conducted. Another line of research is related to the idea of incorporating such effects in periodic models, PAR(p) and PARMA(p, q).

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank Duke Energy for the financial support granted through R&D program of the Brazilian Electricity Regulatory Agency (ANEEL). This research is part of the results obtained from the R&D developed between PUC-Rio and Duke Energy.

References

[1] B. H. Dias, A. L. M. Marcato, R. C. Souza et al., "Stochastic dynamic programming applied to hydrothermal power systems operation planning based on the convex hull algorithm," *Mathematical Problems in Engineering*, vol. 2010, Article ID 390940, 20 pages, 2010.

[2] A. L. M. Marcato, *Hybrid representation of equivalents and individualized systems for the avegare stated period operation planning of power systems of great size [D.S. thesis]*, DEE, PUC, Rio, Brazil, 2002.

[3] H. A. Thomas and M. B. Fiering, "Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation," in *Design Water Resource Systems*, A. Mass, Ed., pp. 459–463, Harvard University Press, Cambrigde, Mass, USA, 1962.

[4] A. I. McLeod, "Parsimony, model adequacy and periodic correlation in time series forecasting," *International Statistical Review*, vol. 61, no. 3, pp. 387–393, 1993.

[5] A. I. McLeod, "Diagnostic checking of periodic autoregression models with application," *Journal of Time Series Analysis*, vol. 15, no. 2, pp. 221–233, 1994.

[6] F. L. C. Oliveira, P. G. C. Ferreira, and R. C. Souza, "A parsimonious bootstrap method to model natural inflow energy series," *Mathematical Problems in Engineering*, vol. 2014, Article ID 158689, 10 pages, 2014.

[7] F. L. C. Oliveira and R. C. Souza, "A new approach to identify the structural order of par (p) models," *Pesquisa Operacional*, vol. 31, pp. 487–498, 2011.

[8] L. C. D. Campos, *Periodic stochastic model based on neural networks [Ph.D. thesis]*, DEE, PUC-Rio, Rio de Janeiro, Brazil, 2010.

[9] R. M. Souza, *Modeling of periodic series via PAR(p) structures utilizing wavelet shrinkage [Ph.D. thesis]*, DEE, PUC-Rio, Rio de Janeiro, Brazil, 2014.

[10] P. G. C. Ferreira, *The stochasticity associated with Brazilian Electrical Sector and a new approach to generate natural inflow energy via periodic gama model, [D.Sc. Thesis]*, DEE, PUC-Rio, Rio de Janeiro, Brazil, 2013.

[11] J. R. M. Hosking, "Fractional differencing," *Biometrika*, vol. 68, no. 1, pp. 165–176, 1981.

[12] V. A. Reisen, "Estimation of the fractional difference parameter in the ARIMA(p, d, q) model using the smoothed periodogram," *Journal of Time Series Analysis*, vol. 15, no. 3, pp. 335–350, 1994.

[13] P. Doukham, G. Oppeenhein, and M. S. Taqqu, *Theory and Applications of Long-Range Dependence*, Birkhäuser, Basel, Switzerland, 2003.

[14] W. Palma, *Long-Memory Time Series: Theory and Methods*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, 2007.

[15] G. C. Franco and V. A. Reisen, "Bootstrap approaches and confidence intervals for stationary and nonstationary long-range dependence processes," *Physica A*, vol. 375, pp. 546–562, 2007.

[16] C. W. J. Granger and R. Joyeux, "An introduction to long-memory time series models and fractional differencing," *Journal of Time Series Analysis*, vol. 1, no. 1, pp. 15–29, 1980.

[17] F. Sowell, "Maximum likelihood estimation of stationary univariate fractionally integrated time series models," *Journal of Econometrics*, vol. 53, no. 1–3, pp. 165–188, 1992.

[18] R. Fox and M. S. Taqqu, "Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series," *The Annals of Statistics*, vol. 14, no. 2, pp. 517–532, 1986.

[19] J. Geweke and S. Porter-Hudak, "The estimation and application of long memory time series models," *Journal of Time Series Analysis*, vol. 4, no. 4, pp. 221–238, 1983.

[20] P. M. Robinson, "Semiparametric analysis of long-memory time series," *The Annals of Statistics*, vol. 22, no. 1, pp. 515–539, 1994.

[21] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[22] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY, USA, 1993.

[23] G. C. Franco and V. A. Reisen, "Bootstrap techniques in semiparametric estimation models for ARFIMA models: a comparison study," *Computational Statistics*, vol. 19, no. 2, pp. 243–259, 2004.

[24] G. C. Franco, V. A. Reisen, and F. A. Alves, "Bootstrap tests for fractional integration and cointegration: a comparison study," *Mathematics and Computers in Simulation*, vol. 87, pp. 19–29, 2013.

[25] D. D. J. Penna, *efinition of the streamflow scenario tree to long-term operation planning*, DEE, PUC-Rio, Rio de Janeiro, Brazil, 2009.