

INTRODUCING A CAUSAL PAR(p) MODEL TO EVALUATE THE INFLUENCE OF CLIMATE VARIABLES IN RESERVOIR INFLOWS: A BRAZILIAN CASE

Paula Medina Maçaira^{1*}, Fernando Luiz Cyrino Oliveira¹,
Pedro Guilherme Costa Ferreira², Fernanda Villa Nova de Almeida¹
and Reinaldo Castro Souza³

Received June 10, 2016 / Accepted March 16, 2017

ABSTRACT. The Brazilian electricity energy matrix is essentially formed by hydraulic sources which currently account for 70% of the installed capacity. One of the most important characteristics of a generation system with hydro predominance is the strong dependence on the inflow regimes. Nowadays, the Brazilian power sector uses the PAR(p) model to generate scenarios for hydrological inflows. This approach does not consider any exogenous information that may affect hydrological regimes. The main objective of this paper is to infer on the influence of climatic events in water inflows as a way to improve the model's performance. The proposed model is called "causal PAR(p)" and considers exogenous variables, such as El Niño and Sunspots, to generate scenarios for some Brazilian reservoirs. The result shows that the error measures decrease approximately 3%. This improvement indicates that the inclusion of climate variables to model and simulate the inflows time series is a valid exercise and should be taken into consideration.

Keywords: Reservoir inflow modelling, Periodic models, Climate predictors.

1 INTRODUCTION

The Brazilian electricity generation is mainly composed by hydroelectric plants owned by multiple players; the Brazilian National Interconnected System (NIS) integrate power and transmission lines from the South, Southeast, Midwest, Northeast and part of the North region. Only 1.7% of the country's electricity production capacity relies outside the NIS, in small isolated systems located mainly in the Amazon region [21].

Planning the Brazilian energy sector means, basically, making decisions about the dispatch of hydroelectric and thermoelectric plants, with the risk of financial losses or energy rationing, as happened such strong in 2001 [25], affecting almost all Brazilian regions.

*Corresponding author.

¹Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), 22451-900 Rio de Janeiro, RJ, Brasil. E-mails: paulamacaira@aluno.puc-rio.br; cyrino@puc-rio.br; fernanda.vnalmeida@gmail.com

²Fundação Getúlio Vargas (FGV), 22231-000 Rio de Janeiro, RJ, Brasil. E-mail: pedro.guilherme@fgv.br

³Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), 22451-900 Rio de Janeiro, RJ, Brasil. E-mail: reinaldo@ele.puc-rio.br

One of the main characteristics of the hydraulic generation system is its strong dependence on hydrological regimes. Thus, the dispatch operation planning has to define generation goals for both hydroelectric and thermal plants along the study horizon, considering the electricity demand, the plants and electrical operating constraints [24].

Considering the dependence on the hydrological regimes, the existing uncertainty of the Brazilian power planning requires an appropriate and consistent stochastic modelling of hydrological series. Therefore, it is possible to identify how important it is to build models to generate hydrological scenarios, in order to optimize the system operation performance, adding reliability to the system and reducing its costs [18]. This optimization process has a stochastic variable: natural inflow.

There are, basically, two approaches to predict the natural inflow: physical and statistical models, where the first one includes the rainfall-runoff hydrological model and the second covers data-driven methods such as time series. To perform monthly forecasts and simulation the classical Periodic Autoregressive (p) model [29], has been widely used. This type of model adjusts the series using the estimated parameters of the historical data [13], and does not consider any exogenous information that could affect the hydrological regimes and, consequently, the electricity generation. Several examples of the application of PAR(p) can be found in the literature, see for example [16] who generate and forecast monthly inflows of the Ganges River with the PAR model. A quick literature search also returns an extensive set of univariate models applied to reservoir inflows (e.g. [27]).

However, models that incorporate explanatory variables, specifically climate variables, have been only recently developed. Some examples of these works in chronological order, are: [32] studied the relation between the Sea Surface Temperature (SST) pattern over the Atlantic and the Pacific Oceans and the variability of water availability in the Amazon Basin; [8] introduced a procedure for further conditioning the inflow probability distributions by considering the recent measurements of climatic variables, [28] developed a semiparametric approach for forecasting inflow at multiple gaging locations on climate precursors; [14] applied Artificial Neural Network to model the complex relationship between inflow and climatic phenomenon; [10] investigated the potential of the Bayesian dynamic modelling approach through an application to forecast a hydrologic time series using relevant climate index information; [11] included climate information in a periodic auto-regressive model in order to provide monthly inflow forecasts for 54 hydropower sites in Brazil; and [12] applied Bayesian Dynamic Models to model and forecast the water inflow for Brazilian hydropower reservoirs, and concluded that the incorporation of climate variables such as rainfall precipitation and El Niño variables, increased the accuracy of both modelling and prediction.

Considering the context and the relevance of the subject, this paper aims to investigate and propose methodological advances in time series modelling and stochastic simulation to generate synthetical hydrological scenarios to model the Brazilian hydrothermal dispatch.

The proposed method, called causal PAR(p), intends to include, exogenously, the meteorological phenomena influence by adjusting a Dynamic Regression (Autoregressive Distributed Lags Model) to the traditional PAR(p) residuals and climate series, incorporating the regression coefficient in the traditional modelling. Several preliminary analyses will be conceived in order to obtain a greater understanding of the involved series and its applicability into the proposed method.

Besides this main goal, which is to present a new approach to model the inflow series, this work also intends to generate synthetical scenarios that better represent the original historical series as well as confidence intervals for out-of-sample forecasts.

The paper is organized as follows: section 2 presents the theoretical background, with a brief description of the Periodic Autoregressive model, the mathematical details of the proposed approach (causal PAR(p)), and a short description of the Bootstrap technique, used to generate synthetical scenarios. Section 3 describes the input variables, their connection with reservoir inflows the underlying system and the Brazilian case study; section 4 presents a exploratory analysis of the available variables. The results from the traditional model and the proposed approach are shown on section 5, and section 6 sums up the work and summarizes its conclusion and final remarks.

2 THEORETICAL BACKGROUND

The proposed framework to include the climate series behaviour in the PAR(p) modelling, named as causal PAR(p), is composed using two main techniques: PAR(p) and Dynamic Regression Model. These techniques are briefly presented in what follows.

2.1 Traditional PAR(p)

Periodic Autoregressive models can also be referred to PAR(p), where p corresponds to the order of the model, in other words, the number of autoregressive terms identified in the model. The PAR(p) model fits to each series period an AR(p) model. Generally, p is a vector, $p = [p_1, p_2, \dots, p_{12}]$, where each element provides the order for each period (month, in case of monthly series). For more details about the PAR(p) model see [7].

PAR(p) model is mathematically described as follows:

$$\begin{aligned} \left(\frac{Z_t - \mu_m}{\sigma_m}\right) &= \varphi_1^m \left(\frac{Z_1 - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \left(\frac{Z_2 - \mu_{m-2}}{\sigma_{m-2}}\right) + \dots + \\ &\dots + \varphi_{p_m}^m \left(\frac{Z_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + a_t \end{aligned} \tag{1}$$

where,

Z_t is the seasonal series of period S .

S is the number of periods ($S = 12$ to monthly series).

T is the time index, $t = 1, 2, \dots, SN$, function of the T year ($T = 1, 2, \dots, N$) and the m period ($m = 1, 2, \dots, S$).

N is the number of years.

μ_m is the seasonal average of the period m .

σ_m is the seasonal standard deviation of the period m .

φ_m is the i -th autoregressive coefficient of the period m .

p_m is the order of the autoregressive operator of the period m .

a_t is the series of independent noises with average zero and variance $\sigma_a^{2(m)}$.

2.2 Dynamic Regression Model

A Dynamic Regression model can be described by the following general equation:

$$a_t = \sum_{i=1}^k \beta_i(L)X_{i,t} + \frac{1}{a(L)}\varepsilon_t \tag{2}$$

Where a_t is the dependent (or output) variable; $X_{i,t}$ are the explanatory (or inputs) variables; $\beta_i(L) = \frac{b_i(L)}{a(L)}$ and $a(L), b_1(L), \dots, b_k(L)$ are finite order lag polynomials of degrees r, s_1, \dots, s_k , respectively, and ε_t is assumed to be white noise. Such a formulation can be seen in [23] and for more mathematical details see [5].

2.3 Causal PAR(p)

The step by step sequence to perform the causal PAR(p) follows the three steps described below:

1. Estimate the traditional PAR(p) model;
2. Find the significant explanatory variables by applying the Dynamic Regression model; and
3. Estimate the causal PAR(p) model.

In the first step the traditional PAR(p) is estimated and the residuals series are extracted to be used in the second step, that fits the Dynamic Regression model. In this step, the exogenous variables are one of the inputs and the residuals generated by the traditional PAR(p) are the outputs. Then, the coefficients obtained in step two and the exogenous variables are inserted in the mathematical formulation of the traditional PAR(p), generating the causal PAR(p).

$$\begin{aligned} \left(\frac{Z_t - \mu_m}{\sigma_m}\right) &= \varphi_1^m \left(\frac{Z_1 - \mu_{m-1}}{\sigma_{m-1}}\right) + \varphi_2^m \left(\frac{Z_2 - \mu_{m-2}}{\sigma_{m-2}}\right) + \dots \\ &\dots + \varphi_{p_m}^m \left(\frac{Z_{t-p_m} - \mu_{m-p_m}}{\sigma_{m-p_m}}\right) + \beta_1 X_{1,t} + \dots + \beta_i X_{i,t} + e_t \end{aligned} \tag{3}$$

Where, $X_{i,t}$ are the exogenous variables, β_i the coefficient and e_t is the white noise.

Therefore the innovation at the proposed approach is to introduce exogenous variables in the PAR(p) modelling. Further studies would consider the development of some kind of periodic transfer function.

Afterwards, the synthetical scenarios generation are carried out applying the Bootstrap technique to the residual series as well as in obtaining confidence interval. The methodology is detailed in the next section.

2.4 Synthetical scenarios' generation and confidence interval

In order to simulate synthetical scenarios the Bootstrap technique is used. This technique, first developed by [6], is a method of sampling with replacement the observations of a random sample that allows the assessment of the variability of an estimator. Such technique generates as many new samples as one wishes, called "Bootstrap sample", usually with the same size of the original sample. In the context of time series, there are basically two ways to apply this technique: Bootstrap in the residuals and the method called Moving Blocks [9].

In this paper Bootstrap is used in the residuals, due to the fact that for all the studied series it is possible to extract residuals, thus ensuring the hypotheses of a random sample (i.e., independent and identically distributed observations) a required condition to apply Bootstrap.

A formal description of the method is: consider R_1, \dots, R_N the random sample and B the number of residuals series to be generated. B residual series are drawn with replacement from the original sample, generating B Bootstrap residuals series of size N each: r^1, \dots, r^B , where $r^i = r_1^i, \dots, r_N^i, i = 1, \dots, B$.

Anderson et al. [2] computed Gaussian prediction intervals for the estimated parameters of a periodic autoregressive moving average (PARMA) models. In the present study the 5% confidence interval is developed from the simulated series by calculating the quantiles 2.5% and 97.5%.

3 DATA ANALYSIS

In Brazil, the fifteen major river basins have an installed capacity of approximately 92 Giga Watts [GW]. The Parana river basin has the highest hydroelectric potential (around 54 GW). In these major rivers there are around 190 hydroelectric power plants currently in operation [22], and these plants operate in a cascade scheme. As an illustration Figure 1 displays the cascade scheme for Paranaíba and Grande basins with 19 reservoirs, represented by triangles, and 15 hydroelectric with no reservoir (circles).

This way decisions taken at the upstream reservoirs will impact the inflow of the downstream reservoirs. The historical data available is the natural inflow⁴ for each reservoir, on a monthly

⁴The natural inflow is the average incoming water per unit of time at each generator's reservoir from affluent rivers, lakes and its own drainage area.

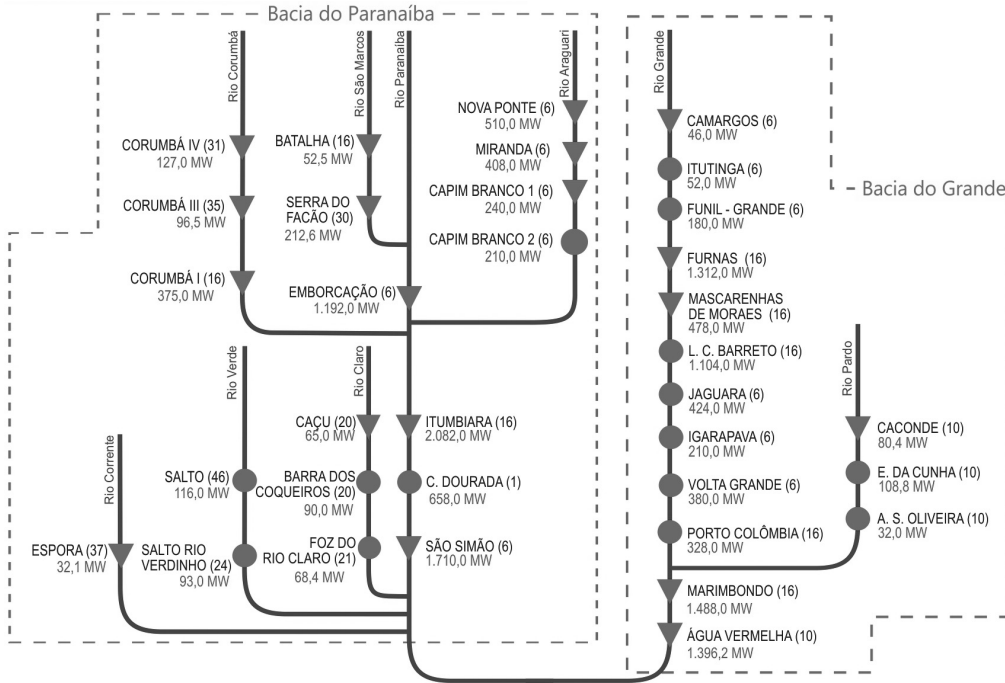


Figure 1 – Example of a cascade scheme. Source: Adapted from [21].

basis, starting in January 1931 and ending in December 2014, measured in cubic meters per second [m^3/s].

The climate variables were selected through a literature search [17]. However, the authors intend to broaden this search for other variables to be considered in the model building. The selected variables are basically related to El Niño and the Sunspots numbers; the variables representing El Niño/La Niña phenomenon are: Southern Oscillation Index (SOI), Equatorial SOI, Niño variations and ONI.

The Southern Oscillation Index (SOI) is calculated based on the difference between the atmospheric pressure at sea level in the regions of Tahiti (in the Western Pacific) and Darwin (Australia, Western Pacific). The Equatorial SOI measures the average difference of atmospheric pressure at sea level between two regions centered on the equator: Indonesia and East Pacific.

The range to indicate the presence or absence of El Niño/La Niña is the same for both the SOI index and Equatorial SOI. Consecutive periods of negative figures indicate El Niño phenomenon occurrence; meanwhile consecutive positive figures denote the presence of La Niña and values close to zero indicate a normal situation, where none of the two phenomenon occur. The official historical monthly series of these indices are provided by the National Oceanic and Atmospheric Administration (NOAA). SOI index data starts in 1951 and Equatorial SOI in 1949.

The sea surface temperature anomaly is a proxy for El Niño and La Niña. Thus, this index is used to classify and quantify such phenomena in four Niño regions: Niño 1+2, Niño 3, Niño 4 and Niño 3.4, defined as follows by NOAA in 2014. Through the location of the Niño regions it is possible to conclude that regions Niño 1+2 and Niño 3 better identify temperature anomalies for the Eastern Pacific Ocean sea surface and region Niño 4 for the Western Pacific. The Niño 3.4 region is centralized in the Pacific, which allows a better understanding of anomalies across it. Therefore, currently the Niño 3.4 region is the official measure used to represent SST (Sea Surface Temperature). However, depending on the study, other regions may be a better alternative.

The threshold for the normal state of this index is between -0.5°C and $+0.5^{\circ}\text{C}$. The criteria commonly used to define an El Niño phenomenon consists of five consecutive averages of SST anomalies above $+0.5^{\circ}\text{C}$. Similarly, for La Niña, this criterion remains, but now the SST anomaly should be below -0.5°C . The time series for all regions are provided by NOAA, on a weekly and monthly basis. For the weekly series, the data starts in 1990 and ends at the current week and the monthly data starts at 1982 and ends up at the current month.

The Oceanic Niño Index (ONI) measures the average sea surface temperature anomalies for the region Niño 3.4, removing the existing warming trend on it. The ONI uses multiple periods based on thirty years to perform the calculation for five successive years. The base periods uses a fifteen year interval, for the lower and upper bound, for example, for 1950 and 1955 the base period considered starts in 1936 and ends in 1965. The El Niño and La Niña are indicated in the same manner as the SST index, the time series is monthly and is provided by NOAA.

Sunspots comprehends solar surface regions of high magnetic field, which have considerably lower temperature than its surroundings and thus appears as a dark area. The magnetic flux amount on the sun surface varies over eleven year periods, known as sunspot and solar cycles. During this cycle there is a minimum and a maximum magnetic flux, which is not only difficult to identify the sunspots and but also they appear almost all the time. The cycle reaches its maximum approximately every eleven years, therefore the observed cycle duration corresponds to eleven years.

The daily and monthly number of sunspots calculation is accomplished with the Relative Index American number of sunspots. This index indicates the solar phenomenon occurrence taking into account their relationship with the Earth, including geomagnetic variations and ionosphere effects. The Solar Division from American Association of Variable Star Observers coordinates the data collection program and the analysis of this phenomenon. Thus, the National Geophysical Data Center (NGDC), provides the historical data from the number of sunspots per month since 1749 and forecasts have been produced until December 2019.

4 EXPLORATORY ANALYSIS

In order to provide a better understanding of the series used in this research, this section presents an exploratory analysis that includes: evolution of the series through time using line's graphic;

variables’ probability distribution using the histogram; descriptive statistics and correlation between the reservoir series and the climate variables.

Also, to reduce the problem dimension (currently with 192 reservoir inflow series), were selected the eight basins that present the biggest correlation with each one of the climate variables. As an example, for the SOI Standard variable the reservoir Curuá-una was the one with the highest correlation among all 192 possibles. Table 1 presents the selected hydroelectric power plants and the Pearson’s correlation⁵ with the climate variables.

Table 1 – Correlation between reservoir inflow series and climate variables.

	SOI Std.	Eq. SOI	Niño 1+2	Niño 3	Niño 4	Niño 3,4	ONI	Suns.
Monjolinho	-0.165	-0.272	0.353	0.327	0.170	0.268	0.281	-0.035
São José	-0.160	-0.294	0.302	0.332	0.175	0.290	0.308	-0.018
Curuá-una	0.254	0.288	-0.202	-0.220	-0.213	-0.234	-0.246	-0.011
Balbina	0.252	0.294	-0.176	-0.219	-0.230	-0.251	-0.269	-0.023
Lajes	0.059	0.057	-0.004	-0.029	-0.066	-0.046	-0.051	-0.177
Quebra-queixo	-0.106	-0.242	0.369	0.293	0.134	0.218	0.226	0.115
Itauba	-0.152	-0.286	0.320	0.340	0.165	0.286	0.300	0.004
Jauru	-0.102	-0.161	0.116	0.056	-0.035	0.024	0.037	0.211

Note that the biggest values were found between the reservoirs Quebra Queixo and Monjolinho with variable Niño 1+2 (0.369 and 0.353, respectively). This does not mean that in the climate variables selecting process this particular variable necessarily will be picket up in the model formulation. This is due to the fact that the variable selection is carried out via a backward process that selects only the significant variables.

The descriptive statistics for each one of the selected reservoirs is shown in Table 2. See that Balbina contains the higher values for mean, median, standard deviation and quartiles. On the other hand Lajes is the one with the lowest values.

Observing Figure 2, which shows the reservoirs time series plot, it is possible to identify, a strong periodicity in all inflow series.

In the first and third columns of Figure 3 the histogram of each reservoir series is presented. It seems quite clear that there is a strong asymmetry indicating that the data might follow a Weibull distribution, which is the distribution that usually models natural events. The second and fourth columns shows the Autocorrelation Functions (ACF) and from them it is possible to confirm the presence of periodicity by observing significantly picks each six months.

Moving now to the climate variables, in Table 3 it is presented the descriptive statistics for each climate variables available. Note that the values presented in the table rely on the same range

⁵The results were significant at 1% level

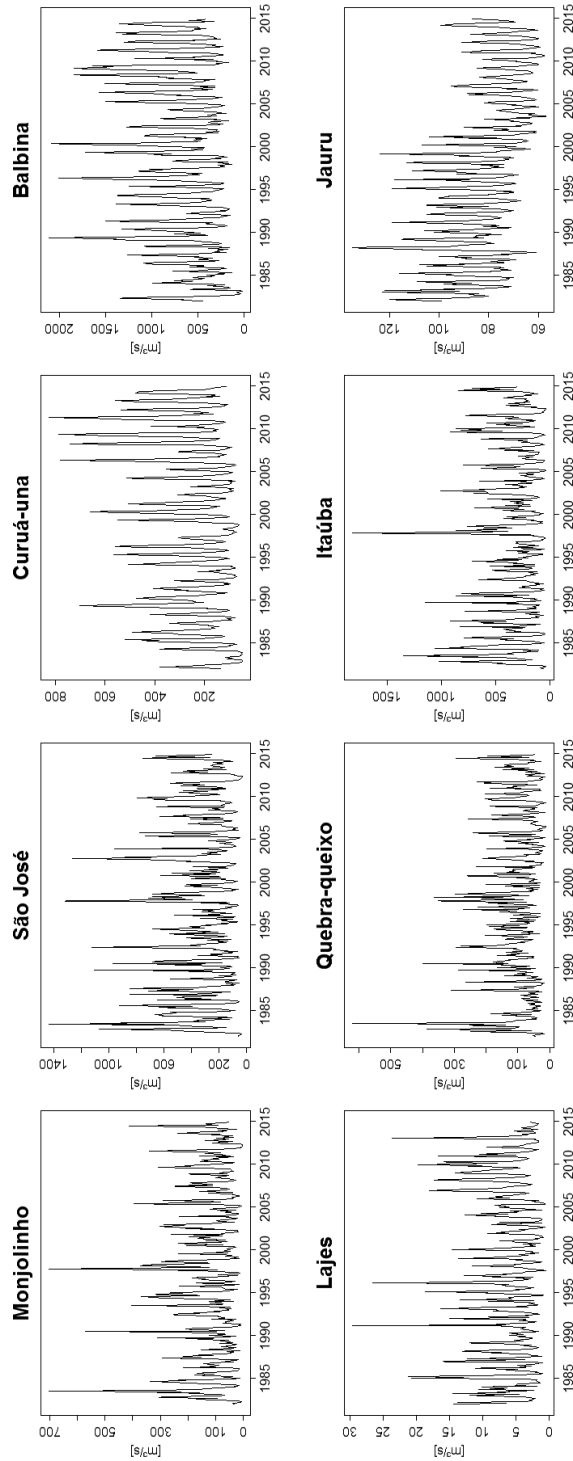


Figure 2 – Reservoirs inflow, time series from January 1982 up to December 2014. Source: The authors.

Table 2 – Descriptive analyses of the reservoirs inflows.

	Mean	Median	Std. Dev.	Min.	Max.	1st quartile	3rd quartile
Monjolinho	114.83	85.61	99.59	2.63	702.00	46.30	153.34
São José	302.43	231.05	236.66	24.93	1435.00	135.00	411.15
Curuá-una	229.00	178.21	151.57	44.81	826.80	115.60	304.70
Balbina	640.56	510.97	411.69	19.00	2115.20	321.90	878.90
Lajes	5.71	4.37	4.51	0.409	29.70	2.50	7.29
Quebra-queixo	96.89	74.25	74.50	11.00	621.00	46.9	125.25
Itauba	303.67	238.97	240.72	35.83	1816.00	133.8	391.25
Jauru	79.85	76.86	14.79	56.851	135.00	69.00	88.00

Table 3 – Descriptive statistics of the climate variables.

	Mean	Median	Std. Dev.	Min.	Max.	1st quartile	3rd quartile
SOI Standard	0.025	0.000	1.006	-3.600	2.900	-0.600	0.700
Equatorial SOI	0.017	0.100	1.030	-3.500	3.000	-0.500	0.700
Niño 1+2	0.081	-0.180	1.210	-2.100	4.620	-0.740	0.590
Niño 3	0.046	-0.130	0.994	-2.070	3.620	-0.605	0.562
Niño 4	0.035	0.190	0.721	-1.870	1.670	-0.510	0.600
Niño 3.4	0.034	0.005	0.970	-2.380	2.950	-0.602	0.595
ONI	0.053	-0.010	0.836	-1.780	2.370	-0.520	0.552
Sunspots	60.117	56.600	44.311	1.700	158.500	19.050	91.225

except for the Sunspots variable. In fact the first seven variables measure exactly the same events: El Niño/La Niña, while the latter variable measures the sunspots events.

One important point to check is the presence of significant correlations between the various El Niño series. Table 4 shows the Pearson correlations⁶, and as expected, there is a negative correlation between the Equatorial SOI and the SOI index with the others, as for those, El Niño is represented by negative values, while for other variables, this same phenomenon is represented by positive values. Also note that ONI and Niño 3.4 show the highest correlation, a consistent result with their concepts, since both measure the sea surface temperature anomaly in the same region.

Figure 4 shows each climate variable evolution through time, highlighting that for the variables representing El Niño events, the extreme values coincide with stronger El Niño or La Niña occurrences. It is also observed the strong 11 year cycle presented in the time series of Sunspots.

⁶The results were significant at 1% level

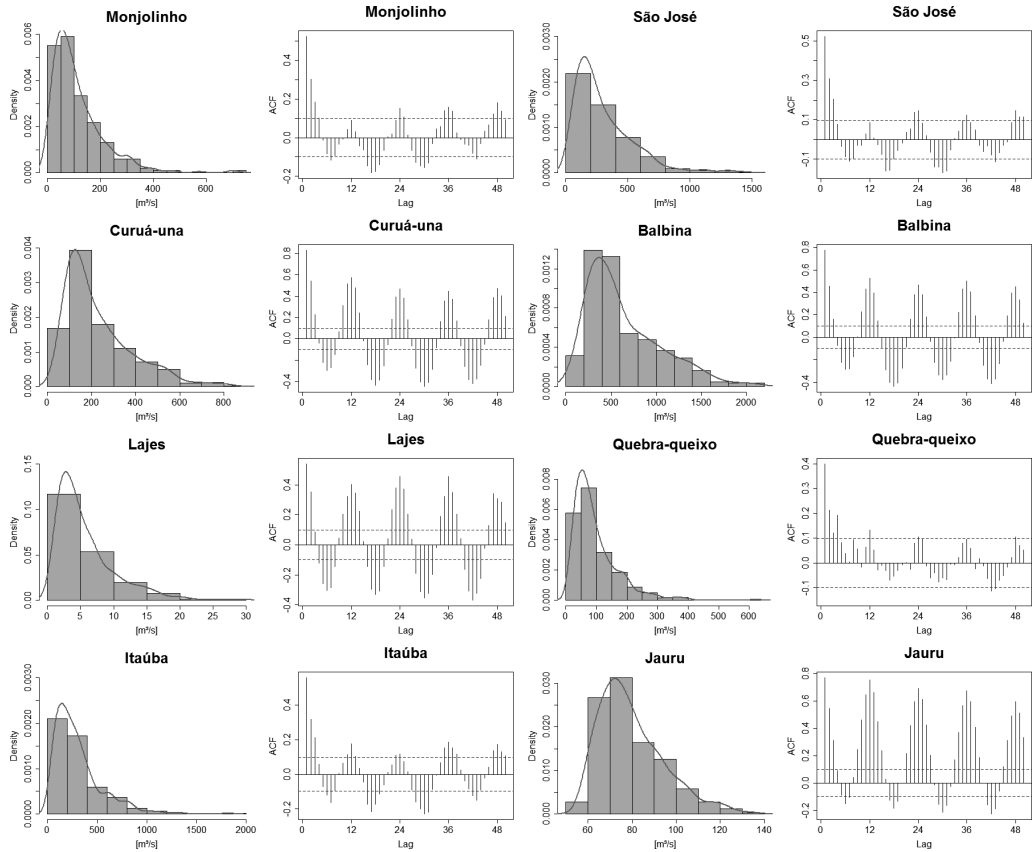


Figure 3 – Reservoir inflows histogram, Autocorrelation Functions (ACF) and Partial Autocorrelation Functions (PACF). Source: The authors.

Table 4 – Correlation between El Niño proxies.

	SOI Standard	Equatorial SOI	Niño 1+2	Niño 3	Niño 4	Niño 3.4	ONI
SOI Standard	1.000	0.797	-0.470	-0.674	-0.691	-0.746	-0.744
Equatorial SOI	0.797	1.000	-0.655	-0.826	-0.745	-0.851	-0.854
Niño 1+2	-0.470	-0.655	1.000	0.822	0.410	0.642	0.633
Niño 3	-0.674	-0.826	0.822	1.000	0.729	0.942	0.923
Niño 4	-0.691	-0.745	0.410	0.729	1.000	0.882	0.879
Niño 3.4	-0.746	-0.851	0.642	0.942	0.882	1.000	0.985
ONI	-0.744	-0.854	0.633	0.923	0.879	0.985	1.000

The last analysis presented is a map that contains the localization of the selected reservoirs and the climate variables measured area, that is where the El Niño/La Niña proxy is measured. From

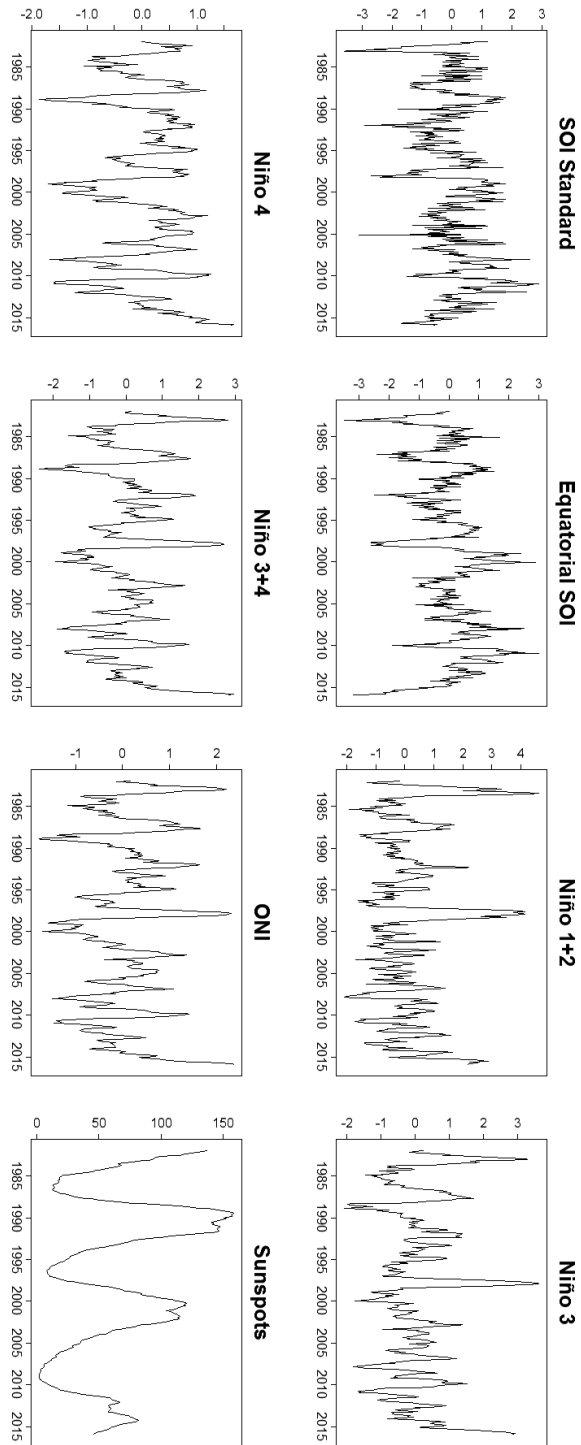


Figure 4 — Climate variables, time series from January 1982 up to December 2014. Source: The authors.

Figure 5 one can see that four of the selected reservoirs are in the same region (South), one at the Southeast, one at the Midwest and two at the North. Figure 5 displays each of the Niño regions, where Niño 3.4 is in the middle of Niño 3 and Niño 4, and the ONI index is measured in the Niño 3 region.

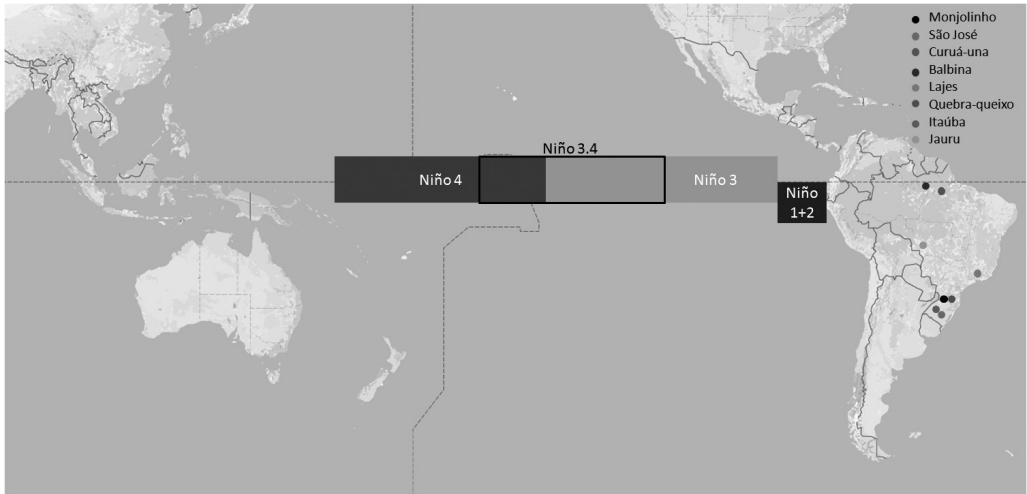


Figure 5 – Niño regions and basins location. Source: The authors.

In Figure 6 the SOI regions are represented, i.e.: the measurement areas for the SOI Standard and the Equatorial SOI index, together with the reservoir localization.

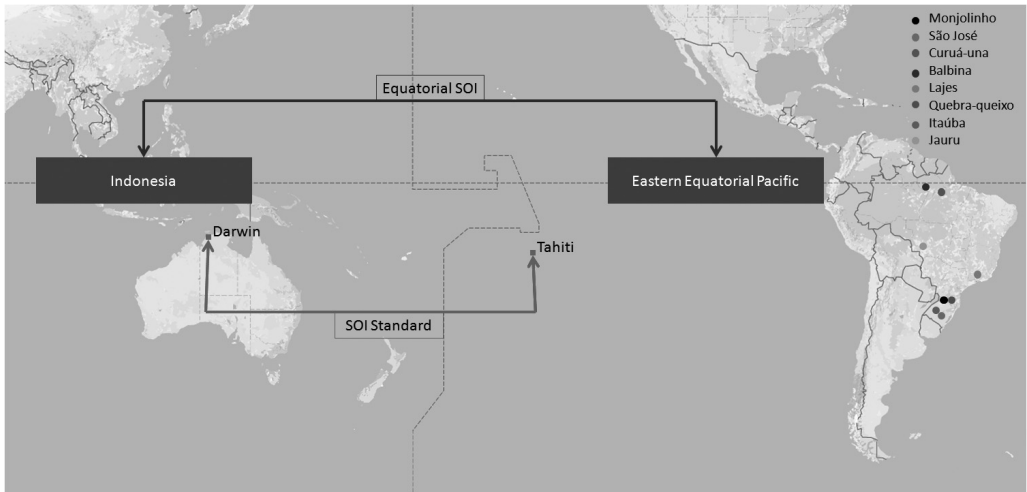


Figure 6 – SOI regions and basins location. Source: The authors.

5 RESULTS

In order to evaluate the proposed methodology, the causal PAR(*p*) approach was applied to model the eight reservoirs inflow series previously presented. The set of possible input variables are the variables representing El Niño/La Niña and Sunspots phenomenon. Since the proposed approach has been developed to improve the current model it will be shown comparisons between the traditional PAR(*p*) and the causal. The last result obtained is the generation of scenarios with both approaches and a comparison with real values.

In order to carry out an out-of-sample analysis, the last year of the available data (2014) was omitted from the initial analysis. So, the modelling set is the data ranging from January 1982 up to December 2013, resulting in 384 observations or 32 years. It is worth mentioning that all the results were generated using the R software [26] and the packages **pear** [15], **TSA** [4], **tseries** [30] and **dynlm** [33].

To estimate the models performance, three metrics were considered: the Mean Absolute Scaled Error (MASE), given by

$$MASE = \frac{1}{n} \sum_{t=1}^n \left(\frac{|Y_t - F_t|}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|} \right), \quad (4)$$

the Mean Absolute Percentage Error (MAPE), given by

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{Y_t - F_t}{Y_t} \right|, \quad (5)$$

and the Root Mean Squared Error (RMSE), given by

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - F_t)^2}, \quad (6)$$

where Y_t is the observed value, F_t is the fitted value and n is the sample size.

5.1 Modelling reservoirs inflow series

As detailed before, the first step of the model is to fit the traditional PAR(*p*) to each of the reservoir inflow series. There are many ways to select the best PAR(*p*) model, i.e. the model's order, [31], for example, applied a genetic algorithm, while [19] uses the Bootstrap technique. Using information criterion is a common approach to select the best model, more specifically Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC) [1]. Both criteria are based on the likelihood function and a penalization term for the number of parameters in the model, however AIC is better for prediction as it is asymptotically equivalent to cross-validation, while BIC is best for explanation as it allows consistent estimation of the underlying data generating process. Since the ultimate goal is to analyse the out-of-sample model performance, in this work it is used the AIC criterion to select the PAR(*p*) orders.

To better explain the causal PAR(p) model fitting step by step, Monjolinho reservoir is used as example series in what follows. In Table 5 the model orders obtained and their respective weights are displayed.

Table 5 – Monjolinho PAR(p) model orders and coefficients.

Period	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
p	2	5	1	1	1	1	6	1	1	8	1	1
ϕ_1	0.30	0.39	0.76	0.88	1.04	0.59	-0.08	0.28	0.51	-0.17	0.64	0.27
ϕ_2	0.23	-0.03	-	-	-	-	0.83	-	-	0.78	-	-
ϕ_3	-	-0.07	-	-	-	-	-0.97	-	-	0.14	-	-
ϕ_4	-	0.38	-	-	-	-	1.57	-	-	0.46	-	-
ϕ_5	-	-0.16	-	-	-	-	-0.56	-	-	-0.23	-	-
ϕ_6	-	-	-	-	-	-	-0.57	-	-	-0.23	-	-
ϕ_7	-	-	-	-	-	-	-	-	-	-1.83	-	-
ϕ_8	-	-	-	-	-	-	-	-	-	1.66	-	-

From the previous step the residuals series is obtained and with these numerous tests is performed using dynamic regression to select the best climate variables. The idea is to select only the significant climate variables (p -value < 0.05) in the fitting exercise. By that all the explanatory variables are considered as candidates and, those that show no statistical significance are eliminated one at a time until the final formulation is reached.

Using again the Monjolinho, Table 6 displays the causal variables selecting process, where the “x” means the variable is selected and “-” means it was removed. In Model 1 all the variables are included since the backward approach is performed, in Model 2 the most insignificant variable were removed from the previous Model and so on until the last model where only significant variables remain: SOI Standard, Equatorial SOI and Equatorial SOI (-1). The “(-1)” means that a lagged variable is used.

Thus, Table 7 shows which variables were selected to join the causal PAR(p) modelling, for each reservoir inflow series. Note that some of them consider lags and also it is possible to visualize that the variable Equatorial SOI was the most used, whereas Niño 1+2 and Niño 3 were not at all picked by any reservoir.

Moving to Step 3 of the model building, the causal model must be estimated in accordance with the explanatory variables selected. In Table 8 the three error measurements statistics obtained to both, PAR(p) and causal PAR(p) models are displayed. Note that the inclusion of climate variables leads to a small reduction in the error statistics, i.e., 1% in the MASE, 5% in the MAPE and 2% in the RMSE.

Table 6 – Monjolinho causal variables process.

Model	1	2	3	4	5	7	8	9	10	11	12	13	15	16
Intercept	x	x	-	-	-	-	-	-	-	-	-	-	-	-
SOI Standard	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Equatorial SOI	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Niño 1+2	x	x	x	x	x	-	-	-	-	-	-	-	-	-
Niño 3	x	x	x	x	x	x	x	x	-	-	-	-	-	-
Niño 4	x	x	x	-	-	-	-	-	-	-	-	-	-	-
Niño 3.4	x	x	x	x	x	x	x	x	x	-	-	-	-	-
ONI	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sunspots	x	x	x	x	x	x	x	x	x	x	x	-	-	-
SOI Standard (-1)	x	x	x	x	-	-	-	-	-	-	-	-	-	-
Equatorial SOI (-1)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Niño 1+2 (-1)	x	-	-	-	-	-	-	-	-	-	-	-	-	-
Niño 3 (-1)	x	x	x	x	x	x	x	-	-	-	-	-	-	-
Niño 4 (-1)	x	x	x	x	x	x	-	-	-	-	-	-	-	-
Niño 3.4 (-1)	x	x	x	x	x	x	x	x	x	x	x	x	x	-
ONI (-1)	x	x	x	x	x	x	x	x	x	x	x	x	-	-
Sunspots (-1)	x	x	x	x	x	x	x	x	x	x	-	-	-	-

5.2 Generating synthetical scenarios

The synthetical scenarios generation is performed, with the residuals series, using the Bootstrap technique. First, the BDS test [3] confirms that the residuals obtained by both, traditional and causal PAR(p) models are white noise (p-value greater than 0.05).

The first procedure in this section is the generation of in-sample scenarios and the respective confidence intervals for the period 1982 up to 2013. Table 9 shows the error statistics comparing the estimated means of the scenarios generated using the traditional PAR(p) and the causal PAR(p) with the actual value of the series. Note that, in general, the causal PAR(p) slightly outperforms the traditional PAR(p): 3% in MASE, MAPE and RMSE.

Next, for the generation of synthetical reservoir inflow scenarios, one should ensure that all simulated values are positive; as a negative value does not make sense. The initial approach does not guarantee this constraint, so the strategy employed was to resample with replacement the corresponding residuals until a positive value for it is obtained, similar to the proposal of [20].

Table 10 presents the coverage of the 95% empirical confidence interval using the series' historical (in-sample values). Note that the proposed approach performs rather better, since the overall coverage across all reservoirs ranges from 94.53% to 95.57% with an average of 94.79%, better than the traditional PAR(p) that ranges from 92.45% to 93.49% (average 92.97%).

Table 7 – Causal PAR(*p*) coefficients.

	Monjolinho	São José	Curuá Una	Balbina	Lajes	Quebra queixo	Itauba	Jauru
Intercept	–	–	–	–	–	–	–	–
SOI Standard	17.77	–	–	–	–	12.69	–	–
Equatorial SOI	–11.34	–	12.71	34.54	–	– 14.31	–	–1.30
Niño 1+2	–	–	–	–	–	–	–	–
Niño 3	–	–	–	–	–	–	–	–
Niño 4	–	– 39.28	–	–	–	–4.30	–	–
Niño 3.4	–	–	–	–	–	–	–	–1.55
ONI	20.49	89.41	–	–	–	–	63.32	–
Sunspots	–	–	–	–	0.02	–	–	–
SOI Standard (–1)	–	14.26	–	–	–	–	15.73	–
Equatorial SOI (–1)	1.47	– 10.73	–	–	–	–	– 19.83	–0.13
Niño 1+2 (–1)	–	–	–	–	–	–	–	–
Niño 3 (–1)	–	–	–	–	–	–	–	–
Niño 4 (–1)	–	– 25.11	–	–	–	–6.94	– 68.33	–
Niño 3.4 (–1)	–	–	–	–	–	11.40	–	0.11
ONI (–1)	–	–	–	–	–	–	17.70	–
Sunspots (–1)	–	–	–	–	–0.03	–	–	–

Table 8 – Fitted errors with PAR(*p*) and causal PAR(*p*).

	MASE		MAPE		RMSE	
	PAR(<i>p</i>)	causal PAR(<i>p</i>)	PAR(<i>p</i>)	causal PAR(<i>p</i>)	PAR(<i>p</i>)	causal PAR(<i>p</i>)
Monjolinho	0.77	0.75	75.41	70.84	66.82	64.94
São José	0.80	0.78	57.00	54.23	174.38	169.82
Curuá-una	0.52	0.51	14.91	15.05	50.19	48.51
Balbina	0.62	0.60	28.57	25.32	169.22	165.58
Lajes	0.66	0.65	43.37	41.52	2.83	2.82
Quebra-queixo	0.73	0.72	57.38	55.70	53.55	52.35
Itauba	0.79	0.78	53.53	50.78	167.15	164.27
Jauru	0.58	0.57	4.97	4.88	5.95	5.90
Average performance	0.68	0.67	41.89	39.79	86.26	84.27

Table 9 – Comparison between the average of the generated scenarios and the in-sample forecasts.

	MASE		MAPE		RMSE	
	PAR(<i>p</i>)	causal PAR(<i>p</i>)	PAR(<i>p</i>)	causal PAR(<i>p</i>)	PAR(<i>p</i>)	causal PAR(<i>p</i>)
Monjolinho	0.81	0.79	105.28	102.64	67.48	65.64
São José	0.84	0.83	69.94	67.18	176.06	171.14
Curuá-una	0.52	0.51	15.60	14.95	50.35	48.45
Balbina	0.62	0.61	30.78	28.44	169.94	165.90
Lajes	0.69	0.69	52.22	51.93	2.85	2.85
Quebra-queixo	0.75	0.73	64.19	62.69	53.59	52.21
Itauba	0.83	0.81	64.65	63.02	168.32	164.23
Jauru	0.58	0.57	4.94	4.88	5.93	5.91
Average performance	0.71	0.69	50.95	49.47	86.82	84.54

Table 10 – Accounting for the number of observations covered by the 95% confidence interval in-sample.

	N. of obs. (out of 384)		Overall correct (%)	
	PAR(<i>p</i>)	causal PAR(<i>p</i>)	PAR(<i>p</i>)	causal PAR(<i>p</i>)
Monjolinho	357	364	92.97	94.79
São José	356	367	92.71	95.57
Curuá-una	355	363	92.45	94.53
Balbina	359	366	93.49	95.31
Lajes	359	364	93.49	94.79
Quebra-queixo	357	364	92.97	94.79
Itauba	359	364	93.49	94.79
Jauru	356	363	92.71	94.53
Average performance	357	364	92.97	94.79

Hence, the average obtained for each month with the synthetical scenarios is checked against the corresponding average of the test set, i.e., year 2014. Table 5.2 contains the out-of-sample statistics for both approaches, again the causal PAR(*p*) performs better for all error statistics.

These results revealed that the causal PAR(*p*), produces better fit and coverage in comparison with the traditional PAR(*p*). This results is aligned and corroborates the findings of [12]. The improvement provided by the proposed approach compared with Lima’s work relies on the fact that in this study eight climate variables were tested for each reservoir while Lima’s study includes only two variables (Niño 3.4 and precipitation).

Table 11 – Comparison between the average of the generated scenarios and the out-of-sample values of the series.

	MASE		MAPE		RMSE	
	PAR(p)	causal PAR(p)	PAR(p)	causal PAR(p)	PAR(p)	causal PAR(p)
Monjolinho	0.71	0.61	54.75	53.19	85.44	83.98
São José	0.72	0.62	37.30	36.01	160.88	159.84
Curuá-una	1.04	0.94	23.46	21.69	67.72	66.65
Balbina	0.68	0.58	17.47	15.79	156.95	155.24
Lajes	5.11	5.14	151.08	150.01	4.40	4.39
Quebra-queixo	0.68	0.58	57.65	56.57	63.18	62.17
Itauba	0.79	0.69	31.24	30.00	200.30	199.35
Jauru	0.88	0.78	6.55	5.77	7.10	6.01
Average performance	1.33	1.24	47.44	46.13	93.25	92.20

6 CONCLUSIONS AND FINAL REMARKS

The PAR(p) is currently used to estimate the operational costs of the Brazilian hydro-thermal optimal dispatch, and it does not take into account any possible exogenous information that may possibly affect the hydrological regimes and therefore the power generation. Causal PAR(p) is a novel approach that proposes the inclusion of explanatory variables exogenously, in the PAR(p) model, using Dynamic Regression. In this study, the inclusion of climate variables related to the El Niño and Sunspots variable were considered to model the reservoirs inflow series. This new approach was able to generate better results when compared with the current PAR(p).

Another important contribution of this research is the literature review made to identify possible climate variables that could influence the hydrological regime, resulting in eight possible variables; seven related to the El Niño phenomenon and one representing the Sunspots.

The major contribution of this paper is the upgrading of the current approach by including climate variables, that result in better results than the traditional PAR(p). Depending on the location of the reservoir different climatic variables were considered in the model fitting.

Finally, the causal PAR(p) was used to generate the synthetic scenarios and its comparison with the measurements each month of 2014 shows that the averages generated with the simulated scenarios reproduce quite well the future values.

In general the proposed approach performs slightly better than the currently used methodology. As a further improvement of this research one can envisage the construction of a model similar to a periodic transfer function, in such way that the explanatory variables are related to the original series and not only with the residuals generated by the PAR(p). Besides that, the search for other variables that could influence the hydrological regime is another possible improvement. Another

possible future step is to aggregate the reservoirs by major basins, aiming to analyse if the climate variables can further improve the results by increasing the range of inflow data.

As a final word, although the causal PAR(p) produces only minor improvements this result is very important as it may result in different electricity prices when introduced in the optimization model used in the Brazilian hydrothermal dispatch.

ACKNOWLEDGEMENTS

The authors would like to thank the R&D program of the Brazilian Electricity Regulatory Agency (ANEEL) for the financial support (PD-0387-0315/2015). They also thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for the doctoral financial support. F. L. Cyrino Oliveira also thanks the support of the National Council of Technological and Scientific Development (CNPq) (research project 443595/2014-3) and FAPERJ (research projects E-26/202.806/2015 and E-26/201.912/2015).

REFERENCES

- [1] AHO K, DERRYBERRY D & PETERSON T. 2014. Model selection for ecologists the worldviews of aic and bic. *Ecology*, **95**: 631–636.
- [2] ANDERSON PL, MEERSCHAERT MM & ZHANG K. 2012. Forecasting with prediction intervals for periodic autoregressive moving average models. *Journal of Time Series Analysis*, **34**: 187–193.
- [3] BROCK WA, DECHERT WD & SCHEINKMAN JA. 1987. A test for independence based on the correlation dimension. Department of Economics, University of Wisconsin at Madison, University of Houston and University of Chicago.
- [4] CHAN K-S & RIPLEY B. 2012. TSA: Time Series Analysis. R package version 1.01. <https://CRAN.R-project.org/package=TSA>
- [5] COCHRANE D & ORCUTT GH. 1949. Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, **44**: 32–61.
- [6] EFRON B. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**: 1–26.
- [7] HIPEL KW & MCLEOD AI. 1994. Time Series Modelling of Water Resources and Environmental Systems. Elsevier.
- [8] KELMAN J, VIEIRA AM & RODRIGUEZ-AMAYA JE. 2000. El niño influence on streamflow forecasting. *Stochastic Environmental Research and Risk Assessment*, **14**: 123–138.
- [9] KUENSCH HR. 1989. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, **17**: 1217–1241.
- [10] KUMAR DN & MAITY R. 2008. Bayesian dynamic modelling for nonstationary hydroclimatic time series forecasting along with uncertainty quantification. *Hydrological Processes*, **22**: 3488–3499.
- [11] LIMA CHR & LALL U. 2010. Climate informed monthly streamflow forecasts for the brazilian hydropower network using a periodic ridge regression model. *Journal of Hydrology*, **380**: 438–449.
- [12] LIMA MLM, POPOVA E & DAMIEN P. (2014). Modeling and forecasting of Brazilian reservoir inflows via dynamic linear models. *International Journal Forecasting*, **30**: 464–474.

- [13] MACEIRA MEP & DAMÁZIO JM. 2006. Use of the Par(p) Model in the Stochastic Dual Dynamic Programming Optimization Scheme Used in the Operation Planning of the Brazilian Hydropower System. *Probability in the Engineering and Informational Sciences*, **20**: 143–156.
- [14] MAITY R & KUMAR DN. 2008. Basin-scale stream-flow forecasting using the information of large-scale atmospheric circulation phenomena. *Hydrological Processes*, **22**: 643–650.
- [15] MCLEOD AI. 1994. Diagnostic checking periodic autoregression models with applications. *Journal of Time Series Analysis*, **15**: 221–233.
- [16] MONDAL MS & WASIMI SA. 2006. Generating and forecasting monthly flows of the ganges river with PAR model. *Journal of Hydrology*, **323**: 41–56.
- [17] NOAA. 2016. February. Climate Prediction Center. National Oceanic and Atmospheric Administration. www.cpc.ncep.noaa.gov/data/indices/
- [18] OLIVEIRA FLC. 2010. Nova abordagem para geração de cenários de aflúncias no planejamento da operação energética de médio prazo. Pontifícia Universidade Católica do Rio de Janeiro.
- [19] OLIVEIRA FLC & SOUZA RC. 2011. A new approach to identify the structural order of par (p) models. *Pesquisa Operacional*, **31**: 487–498.
- [20] OLIVEIRA FLC, SOUZA RC & MARCATO ALM. 2015. A time series model for building scenarios trees applied to stochastic optimisation. *International Journal of Electrical Power & Energy Systems*, **67**: 16–38.
- [21] ONS. 2014. October. Operador Nacional do Sistema Elétrico. www.ons.com.br
- [22] ONS. 2015. Atualização de Séries Históricas de Vazões, período 1931 a 2014. Technical report, Operador Nacional do Sistema Elétrico.
- [23] OTEXTS. 2016. February. Forecasting: principles and practice, advanced forecasting methods, dynamic regression models. Texts: online, open-access textbooks. <https://www.otexts.org/fpp/9/1>
- [24] PEREIRA MVF. 1989. Optimal stochastic operations scheduling of large hydroelectric systems. *International Journal of Electric Power and Energy Systems*, **11**: 161–169.
- [25] PINGUELLI LR, FIDELIS NS, GIANNINI MP & DIAS LL. 2013. Evolution of Global Electricity Markets: New Paradigms, New Challenges, New Approaches. Academic Press. 435–459.
- [26] R CORE TEAM. 2015. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- [27] RAVINES RR, SCHMIDT AM, MIGON HS & RENNO CD. 2008. A joint model for rainfall-runoff: the case of Rio Grande basin. *Journal of Hydrology*, **353**: 189–200.
- [28] SOUZA FILHO FA & LALL. 2003. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: applications of a multivariate, semiparametric algorithm. *Water Resources Research*, **39**(11): 1–13.
- [29] TERRY LA, PEREIRA MVF, NETO TA, SILVA LFCA & SALES PRH. 1986. Coordinating the Energy Generation of the Brazilian National Hydrothermal Electrical Generating System. *Interfaces*, **16**: 16 p.
- [30] TRAPLETTI A & HORNIK K. 2015. tseries: Time Series Analysis and Computational Finance. R package version 0.10-34. <http://CRAN.R-project.org/package=tseries>

- [31] URSU E & TURKMAN KF. 2012. Periodic autoregressive model identification using genetic algorithms. *Journal of Time Series Analysis*, **33**: 398–405.
- [32] UVO CB & GRAHAM NE. 1998. Seasonal runoff forecast for northern South America: a statistical model. *Water Resources Research*, **34**(12): 3515–3524.
- [33] ZEILEIS A. 2014. dynlm: Dynamic Linear Regression. R package version 0.3-3. <http://CRAN.R-project.org/package=dynlm>