International Journal of Forecasting 30 (2014) 464–476

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Modeling and forecasting of Brazilian reservoir inflows via dynamic linear models

L.M. Marangon Lima [a,*], E. Popova [a], P. Damien [b]

[a] Department of Operations Research and Industrial Engineering, The University of Texas at Austin, United States

[b] Department of Information, Risk and Operations Management, The University of Texas at Austin, United States

ARTICLE INFO

Keywords:
Reservoir inflow forecasting
Seasonal models
Bayesian updating
Climate predictors

ABSTRACT

This work focuses on developing a forecasting model for the water inflow at an hydroelectric plant's reservoir for operations planning. The planning horizon is 5 years in monthly steps. Due to the complex behavior of the monthly inflow time series we use a Bayesian dynamic linear model that incorporates seasonal and autoregressive components. We also use climate variables like monthly precipitation, El Niño and other indices as predictor variables when relevant. The Brazilian power system has 140 hydroelectric plants. Based on geographical considerations, these plants are collated by basin and classified into 15 groups that correspond to the major river basins, in order to reduce the dimension of the problem. The model is then tested for these 15 groups. Each group will have a different forecasting model that can best describe its unique seasonality and characteristics. The results show that the forecasting approach taken in this paper produces substantially better predictions than the current model adopted in Brazil (see Maceira & Damazio, 2006), leading to superior operations planning.

1. Introduction

The available methods for hydrological forecasting fall into two classes: conceptual (or physical) methods, which correspond to the rainfall-runoff hydrological models, and data-driven methods such as regression, time series, and artificial neural networks models. Interest lies in either forecasting river streamflow for water management and flood control or forecasting natural inflow to hydropower reservoir for operation and scheduling, with the latter being the focus of this work.

The rainfall-runoff conceptual model is a hydrological model that transforms rainfall (precipitation) into runoff (streamflow) based on physical and empirical equations. The components involved in the transformation process are evaporation, infiltration, interception, soil moisture, land use, and various meteorological conditions, including air temperature and solar radiation (Collisschonn, Allasia, da Silva, & Tucci, 2007; Moradkhani, Hsu, Gupta, & Sorooshian, 2004).

Since conceptual models rely on an accurate knowledge of the physical mechanisms of the underlying streamflow at a particular location, the data-driven techniques gained more popularity in the field of hydrology over the last decade (Wang, 2006). Data-driven models are defined on the basis of connections between state variables (input, internal and output), with little knowledge of the physical behavior of the system (Solomatine, 2002) being needed. Therefore, the forecasting procedure can easily be extended and applied to different locations and conditions. Examples of data-driven models are statistical models like time series models and artificial neural networks.

* Corresponding author. Tel.: +1 55 35 3621 3630.
E-mail address: luana_marangon@yahoo.com.br (L.M.M. Lima).

http://dx.doi.org/10.1016/j.ijforecast.2013.12.009
0169-2070/

The most popular univariate time series models applied to inflow forecasting are the autoregressive moving average (ARMA) models and their variants (Box & Jenkins, 1976). These are built on the assumption of stationarity, that is, the statistical properties of the process are not a function of time. (For a Bayesian perspective, see Marriott & Newbold, 1998, and West, 2013). Therefore, they are more commonly used for forecasting annual streamflows. Streamflow series with time scales of less than a year (e.g. monthly, quartely) usually exhibit seasonality because the hydrologic phenomena vary from one season to another. According to Hipel and McLeod (1994) three types of models can be applied to these series: the seasonal autoregressive integrated moving average (SARIMA), periodic ARMA (PARMA) and deseasonalized ARMA models. The deseasonalized and periodic models are used for describing data that possess stationarity within each season (e.g. Chen, 1997; Maceira & Damazio, 2006; Mondal & Wasimi, 2006; Yurekli, Kurunc, & Ozturk, 2005). The SARIMA family of models can be fitted to data where the level and perhaps other features change within each season across the years (e.g. Bender & Simonovic, 1994; Noakes, McLeod, & Hipel, 1985).

A more general class of regression models, the dynamic linear models (DLMs), have the capability to deal with nonstationarity within a season (West & Harrison, 1997). Krishnaswamy, Halpin, and Richter (2001) introduced a Bayesian dynamic linear regression model as a useful tool for studying the dynamics of hydrology in systems which are subject to high natural variability and land-use change. The model was applied to the Terraba River basin in the southern part of Costa Rica. Kumar and Maity (2008) apply a Bayesian dynamic model to the Devil's Lake basin, located in North Dakota, USA. They claim that the major strength of this type of model lies in its quantification of prediction uncertainty, particularly under different climate change scenarios.

Migon and Monteiro (1997) propose a dynamic nonlinear Bayesian model for the Fartura river basin in Brazil, where the complex system of equations that defines the physical processes is replaced by a simple one that tries to mimic the runoff's behavior given current and past precipitations. An extension of this model to Brazil's Grande river basin is presented by Ravines, Schmidt, Migon, and Renno (2008). Other applications of the Bayesian dynamic model in the field of hydrology include those of Berger and Rios-Insua (1998), Krishnaswamy, Lavine, Richter, and Korfmacher (2000) and Rios-Insua, Salewicz, Muller, and Bielza (1997).

This paper presents a Bayesian DLM for forecasting the water inflow at the Brazilian hydropower reservoirs. The idea is to initially group the reservoirs by basin, then develop a model for each basin based on its particular characteristics. The model will be then used as an input to a multi-stage stochastic optimization problem that solves the hydrothermal planning. The planning horizon is five years ahead, meaning that we are dealing with long-term forecasting. We also want to incorporate relevant climate variables as predictors.

The data are non-stationary. The monthly mean inflows for the basins located in the southern part of Brazil show a tendency to increase. However, most importantly, we observe non-stationarity in the seasonal pattern of the time series; for instance, a delay in the wet season for some of the basins. By 'a delay in the wet season' we mean that the window with the peak water inflow, which used to be December to February, is now from January to March. Therefore, we are also dealing with non-stationarity within the season, which means that the series cannot be reduced to a stationary process by differencing, so we need to work with a general dynamic model.

We want to model the process in its original scale, i.e., without performing any transformation of the data in order to achieve normality. As was noted above, the forecasts from the DLM are fed into a stochastic optimization algorithm, which requires the forecasting approach to assume a linear error structure in the time series regression. The DLM approach in this paper does precisely that, unlike the approaches of Migon and Monteiro (1997) and Ravines et al. (2008).

The remainder of the paper is organized as follows. Section 2 presents the Brazilian framework, with its major river basins and hydroelectric capacity. Section 3 describes the basin time series and correlation analysis, while Section 4 describes the climate variables that will be used as predictors in the model. Section 5 offers a brief description of the Periodic Autoregressive model which is currently used in Brazil. In Section 6 we describe the Bayesian model for basin inflows. Section 7 details the forecasting results and model performance criteria for each basin. Section 8 concludes the work.

2. Brazilian framework

Brazil has many rivers that form twelve major drainage basins, as shown in Fig. 1, of which only ten have hydropower plants. The Parana basin has the highest hydroelectric potential, around 54 gigawatts [GW], which represents more than 50% of the total capacity. It can be further subdivided into six minor basins, based on its major rivers: Paranaiba, Grande, Tiete, Paranapanema, Parana and Iguacu. Table 1 shows the total installed capacity for each basin, which is the sum of the generation capacities of each of the hydro plants within the basin: treating the Parana as 6 sub-basins leads to a total of 15 basins.

There are around 140 hydroelectric power plants currently in operation, and these plants operate in a cascade scheme. In order to determine how much electricity each one will produce in the future, one needs to know how much water will be available in the reservoirs. The available historical data are the natural inflow for each generator on a monthly basis, starting from January 1931, and measured in cubic meters per second [m^3/s]. The 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. Since the reservoirs operate in a cascade scheme, decisions taken at the upstream reservoirs will influence the inflow of the downstream reservoirs. The available data exclude the upstream reservoir operation by summing the natural inflow of the reservoir upstream in the cascade and the incremental inflow. Consider an example with two reservoirs, represented by the two triangles depicted in Fig. 2.

Fig. 1. Major river basins in Brazil.

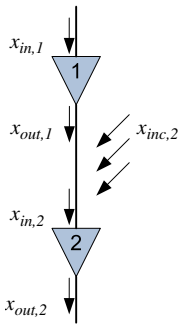


Fig. 2. Example of a reservoir cascade scheme with two reservoirs ($i = 1, 2$) represented by the two triangles. The arrows represent the water inflow in ($x_{in,i}$) and out ($x_{out,i}$) of the reservoirs and the water inflow captured between reservoirs 1 and 2 ($x_{inc,2}$).

Table 1
Major river basins: installed capacity in megawatts [MW].

	Basin	Capacity [MW]	%
1	Paranaiba	8,496	9.2%
2	Grande	7,619	8.2%
3	Tiete	6,830	7.4%
4	Parana	14,000	15.1%
5	Paranapanema	4,368	4.7%
6	Iguacu	7,264	7.9%
7	Paraguay	660	0.7%
8	Uruguay	5,572	6.0%
9	South Atlantic	1,765	1.9%
10	Southeast Atlantic	3,789	4.1%
11	East Atlantic	1,030	1.1%
12	Sao Francisco	10,577	11.4%
13	Parnaiba	237	0.3%
14	Tocantins	12,780	13.8%
15	Amazon	7,480	8.1%
	Total	92,467	100%

The natural inflow will be given by

$$x_{nat,1} = x_{in,1},$$
$$x_{nat,2} = x_{nat,1} + x_{in,2} - x_{out,1},$$

where $x_{nat,i}$ is the natural inflow of reservoir i (for $i = 1, 2$), $x_{in,i}$ and $x_{out,i}$ are the inflow in and out of reservoir i (for $i = 1, 2$), respectively, given by historical data from either a fluviometer station or the reservoir operation. (Typically, in Brazil, the outflow of a reservoir is given by the sum of turbined and spilled volumes.)

Instead of forecasting inflow for the 140 generators, we reduce the dimension of the problem by grouping the generators by basin. Consider, again, the example in Fig. 2: if we want to group reservoirs 1 and 2, we cannot simply sum their natural inflows. Instead, we sum the natural inflow at reservoir 1 with the incremental inflow at reservoir 2, which corresponds to the water captured in the area between reservoirs 1 and 2, and is given by

$$x_{inc,2} = x_{in,2} - x_{out,1}, \quad \text{or} \quad x_{inc,2} = x_{nat,2} - x_{nat,1}.$$

Therefore, before grouping the reservoirs by basin, we compute their incremental inflows, which are given by

$$x_{inc,i} = x_{nat,i} - \sum_{j \in \Omega_i} x_{nat,j},$$

where $x_{inc,i}$ is the incremental inflow of generator i, $x_{nat,i}$ is the natural inflow of generator i, and Ω_i is the subset of generators located immediately upstream of generator i on the cascade. The equivalent generator for each basin will have its natural inflow being equal to the sum of the incremental inflows of all reservoirs belonging to the basin.

3. Basin inflow preliminary data analysis

After grouping the 140 reservoirs by river basin, we reduce the dimension of the problem to 15 time series that we would like to forecast. The forecasting model is developed for these 15 basin inflow time series. We perform principal component analysis (PCA), not to cluster the basins, but to learn about the model variability and correlations between the basins. A Bayesian DLM is developed for each of the basins, and therefore, we actually forecast 15 time series. However, in order to simplify the discussion, we choose three representative basins, one from each component, for the purposes of describing the behavior of the series, conducting an autocorrelation analysis, and presenting forecasting performance results. Because the other series are similar, we omit them from the early discussions. However, forecast results for all 15 of the time series are given in the empirical analysis section.

The PCA results show that we have a first component with nine basins (Amazon, Paraguay, Tocantins, Parnaiba, Sao Francisco, Southeast Atlantic, Paranaiba, Grande and Tiete), a second component with five basins (Paranapanema, Iguacu, Parana, Uruguay and South Atlantic), and a third component with only the East Atlantic Basin. These three components explain 80% of the data variability. We can see from Fig. 3 that geography plays an important role in these data.

The behaviors of the time series of the various basins within each component are similar. Throughout, we choose one basin from each component to represent the others. We choose the Amazon basin for the first component; Uruguay for the second component; and East

Fig. 3. Principal component analysis of the 15 basins leads to three components.

Atlantic for the third component. The inflow time series for these three basins from January 1979 through December 2009 show different behaviors, as depicted in Fig. 4. The basins in the first component show a pattern similar to the Amazon time series in Fig. 4, while the basins in the second component show a disturbed pattern similar to the Uruguay time series in Fig. 4. The East Atlantic series in the third component has some extreme observations.

Fig. 5(a), (d) and (g) depict the inflow distribution as histograms, and indicate that the data might follow a lognormal distribution. Fig. 5(b), (e) and (h) depict the autocorrelation function (ACF) for up to 20 lags of the data. The autocorrelation plots indicate that the time series are nonstationary because the ACFs decay slowly with time, except for the Uruguay basin.

We can also confirm that the Amazon and East Atlantic basins have a seasonal component. The same trend is seen for the other basins in the first component, though we can see that the seasonal component is less pronounced for the Uruguay basin. Again, the same trend is observed for the other basins in the second component. This is a consequence of the climate characteristics. The south region is located below the Tropic of Capricorn and has a rainfall regime which is more evenly distributed throughout the year, whereas other parts of the country have well-defined wet and dry seasons. Since seasonality may lead to nonstationarity of the series, it is not a coincidence that the basins with exponential ACF decays are basins with smaller seasonal effects.

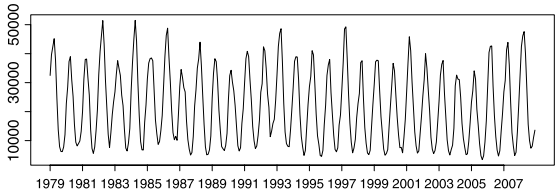
Fig. 5(c), (f) and (i) depict the partial autocorrelation (PACF) plots. These plots are useful for identifying the appropriate order of the autoregressive models. For instance, if we are dealing with a pure autoregressive model then the PACF should cut off sharply after lag p.

4. Potential climate predictors

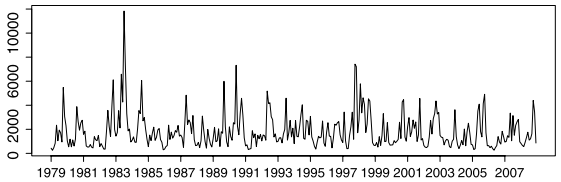
In this section we present the data analysis for the climate variables, as well as the correlations between these variables and the inflow time series of each basin. Some variables, like precipitation, are local and differ for each basin. Others, like the ocean indices, are the same for all basins.

The precipitation data presented here are the monthly accumulated precipitation for each basin in millimeters [mm]. The precipitation is measured at more than one rain gauge, which means that approximating the precipitation for the whole basin is not straightforward, due to the need for data interpolation. We chose to work with the data from the National Aeronautics and Space Administration (NASA) website. Basically, one can specify a box with the basin latitude and longitude coordinates and the website will give you the monthly accumulated rainfall for that area. The resulting data series is ready to be modeled and we do not need to perform any interpolation.

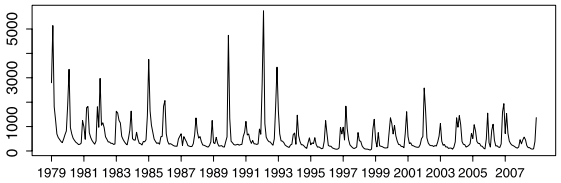
The problem lies in determining these boxes for each basin. Ideally, they should be as small as possible, and the precipitation of each would be averaged to get the basin's total precipitation. Here, we simplify things by instead determining a big box that covers the basin drainage



(a) Amazon basin inflow time series.

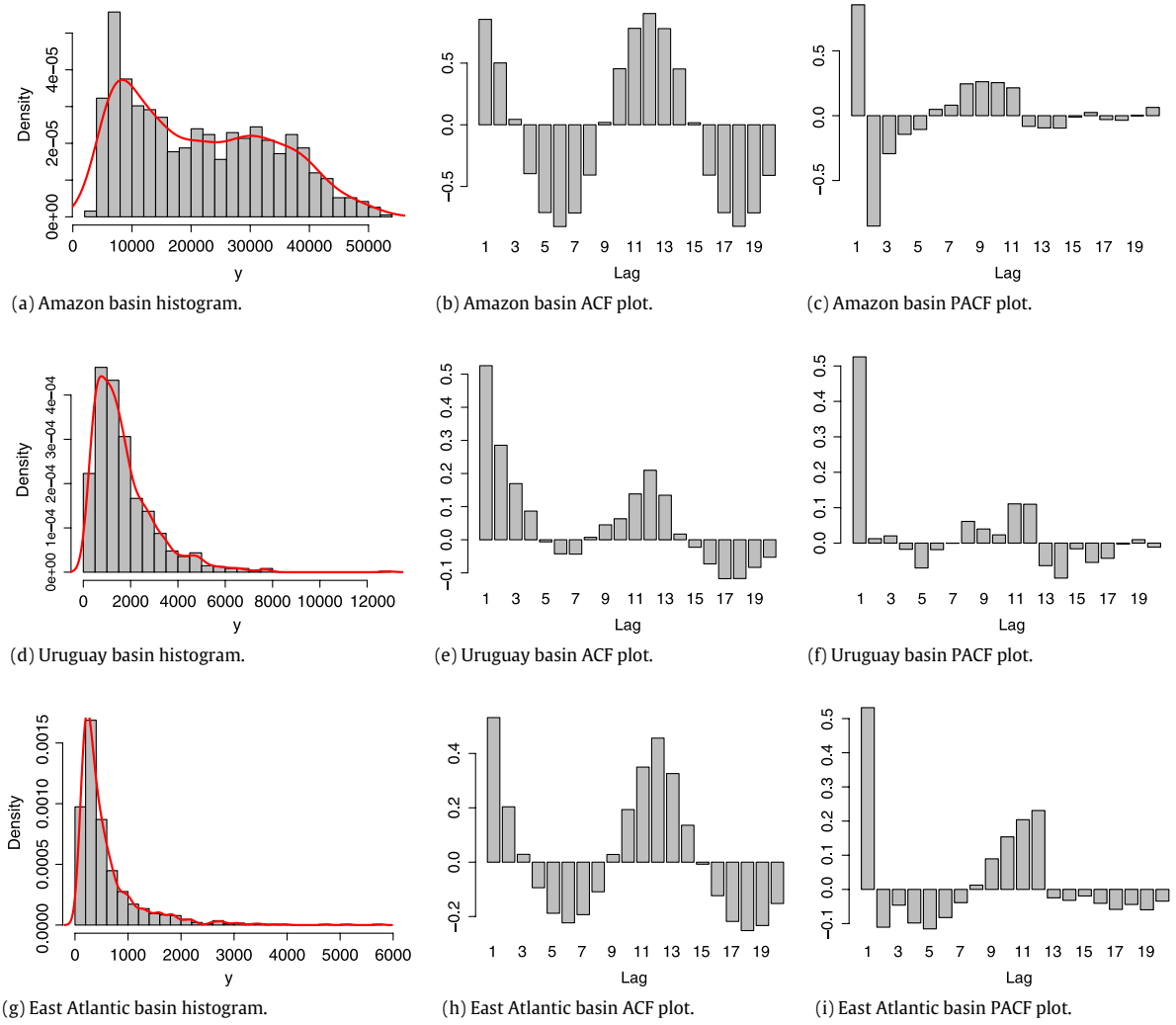


(b) Uruguay basin inflow time series.



(c) East Atlantic basin inflow time series.

Fig. 4. Historical data from January 1939 to December 2009 for the monthly water inflow time series in cubic meters per second [m^3/s].

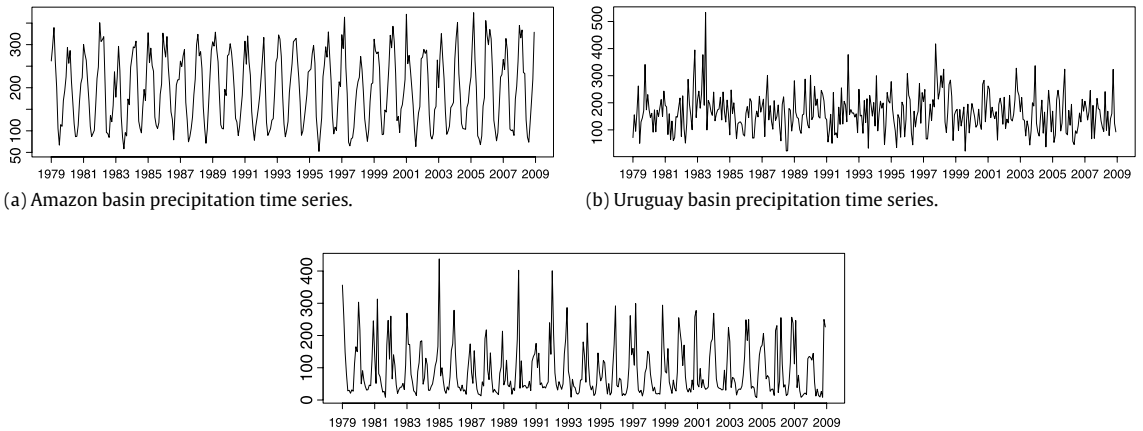


(a) Amazon basin histogram. (b) Amazon basin ACF plot. (c) Amazon basin PACF plot.

(d) Uruguay basin histogram. (e) Uruguay basin ACF plot. (f) Uruguay basin PACF plot.

(g) East Atlantic basin histogram. (h) East Atlantic basin ACF plot. (i) East Atlantic basin PACF plot.

Fig. 5. Basin water inflow series histogram, autocorrelation function (ACF) and partial autocorrelation function (PACF).



(a) Amazon basin precipitation time series.

(b) Uruguay basin precipitation time series.

(c) East Atlantic basin precipitation time series.

Fig. 6. Basin accumulated monthly precipitation time series from January 1979 to September 2009 in millimeters [mm].

area. The precipitation time series for one basin of each component are depicted in Fig. 6. The data span the period from January 1979 to September 2009. We can see that the precipitation time series present movements similar

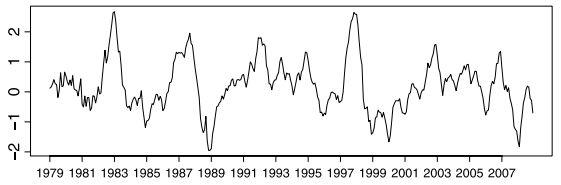


Fig. 7. NINO 3.4 index time series from January 1979 to December 2009.

Table 2
Basin water inflow series correlation with accumulated precipitation and NINO 3.4 index.

	Basin	Precipitation	NINO 3.4
1	Paranaiba	0.70	0.04
2	Grande	0.66	0.03
3	Tiete	0.68	0.03
4	Parana	0.46	0.20
5	Paranapanema	0.36	0.12
6	Iguacu	0.60	0.18
7	Paraguay	0.78	−0.11
8	Uruguay	0.69	0.26
9	South Atlantic	0.68	0.25
10	Southeast Atlantic	0.77	−0.01
11	East Atlantic	0.73	0.006
12	Sao Francisco	0.53	0.006
13	Parnaiba	0.82	−0.13
14	Tocantins	0.63	−0.08
15	Amazon	0.72	−0.05

to those in the inflow time series shown in Fig. 4, either smooth or disturbed.

Climatologists use the NINO indices to monitor the Pacific Ocean for signs of El Niño or La Niña. These indices are based on sea surface temperatures and correspond to the average value over a specified region. During El Niño events, sea surface temperatures are warmer than normal, while during La Niña the temperatures are usually cooler. The method for identifying El Niño or La Niña differs between researchers. A common one is based on the NINO 3.4 index, which is the preferred one since it has the strongest effect on the rainfall regime. An El Niño event will be identified if the 5-month running average exceeds +0.4 °C for six consecutive months. By analogy, a La Niña event is identified when the same average is below −0.4 °C (Trenberth, 1997).

The NINO 3.4 index is available through the International Research Institute for Climate and Society (IRI) dataset website. The dataset spans the period from January 1865 to January 2011. The NINO 3.4 time series from January 1979 to December 2008 is depicted in Fig. 7.

The correlations between climate variables and the basin time series are shown in Table 2. Note that all basins are highly positively correlated with precipitation. The NINO 3.4 index is positive and significantly correlated with the basins in the south of the country, namely Uruguay, South Atlantic, Iguacu and Paranapanema.

5. Periodic autoregressive model

In Brazil, hydrology scientists use the periodic autoregressive (PAR) model for forecasting inflows using the data described in the previous sections (see Maceira & Bezerra, 1997; Maceira & Damazio, 2006). We provide a brief description of the PAR model here, since we compare our Bayesian forecasting results to the PAR approach. According to Hipel and McLeod (1994), fitting a PAR model to a seasonal time series is like fitting a separate autoregressive (AR) model to each season of the year. Suppose that we have N years of data and let s be the seasons of the year, then, for monthly data, $s = 12$. Now let $z_{r,m}$ be the observation in the rth year and mth season, where $r = 1, 2, \ldots, N$, and $m = 1, 2, \ldots, s$. A PAR model for month m can be written as

$$\frac{z_{r,m} - \mu_m}{\sigma_m} = \sum_{i=1}^{p_m} \phi_i^{(m)} \frac{z_{r,m-i} - \mu_{m-i}}{\sigma_{m-i}} + a_{r,m},$$

where p_m is the order of the AR model for the mth month, $\phi_i^{(m)}$ is the AR coefficient for season m and lag i, and $a_{r,m}$ is white noise. Note that the observation $z_{r,m}$ is deseasonalized by subtracting the mean and dividing by the standard deviation. This approach is recommended before fitting a PAR model in order to satisfy stationarity conditions.

The PAR model is evaluated in three steps. The first is the identification of the model's order for each season using the partial autocorrelation function; the second is the estimation of the autoregressive coefficients using the Yule–Walker equations (Salas, Delleur, Yevjevich, & Lane, 1980); and the third is checking the adequacy of the fitted model by examining the properties of the residuals for each season. In particular, the residuals should be a white noise series, i.e., uncorrelated, normally distributed and homoscedastic (Hipel & McLeod, 1994).

6. Bayesian model for the inflows

The Bayesian model for the inflow is based on a dynamic linear model (DLM). Details of this model are provided by West and Harrison (1997). Here, only a brief description is given. We choose to work with DLMs because we want to develop a stochastic model directly from the undifferenced basin inflow data, and the DLM is able to model the dynamic change of the series. Throughout the section we use bold typeface to denote vectors and matrices. Matrices are always uppercase, vectors can be either lowercase or uppercase.

6.1. Dynamic linear models

At the outset, we simply develop the general DLM modeling framework, but in the next subsection we define the mathematical symbols to coincide with the actual data variables used in the empirical analysis.

Let $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]$ be a vector with the historical data of the observed series, in our case the water inflow time series for one of the basins. Now suppose that we want to predict Y_t for $t = N + 1$. Note that Y_t is a scalar variable. In order to estimate Y_t, we can rely on other variables that have a significant correlation with the observed series $\mathbf{Y}$, which could be, for instance, a previous observation of Y_t, say Y_{t-1}. These independent variables are also known as predictor variables. Let $\mathbf{F}_t'$ be a vector containing all of the predictor variables, and let

$\boldsymbol{\theta}_t$ be a vector containing the corresponding regression coefficients for all of the predictor variables at time t. The observation equation for the univariate normal DLM for time t is of the form

$$Y_t = \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t,\ \nu_t \sim N[0, V_t], \tag{1}$$

where $N[0, V_t]$ is the normal distribution with mean 0 and variance V_t, also known as the observational variance. For the univariate case, with n predictor variables, $\mathbf{F}_t'$ is a $(1 \times n)$ vector and $\boldsymbol{\theta}_t$ is a $(n \times 1)$ vector.

Note that we are dealing with the univariate DLM because Y_t is a scalar variable. We do not consider the correlation of the basin's water inflow during the modeling process. Therefore, we have one DLM for each basin.

This type of model is called dynamic because the state vector $\boldsymbol{\theta}_t$ changes with t, which is an important feature when modeling non-stationary time series, as is the case with the inflow data. The dynamic aspect of the model is given by the system equation

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim N[\mathbf{0}, \mathbf{W}_t], \tag{2}$$

where $\mathbf{G}_t$ is a known, $(n \times n)$ state evolution matrix, and $N[\mathbf{0}, \mathbf{W}_t]$ is a multivariate normal distribution with mean $\mathbf{0}$; $\mathbf{W}_t$ is an $(n \times n)$ evolution covariance matrix for $\boldsymbol{\theta}_t$. This equation captures the evolutionary changes in the regression parameters.

The evolution and observation equations may also be expressed for each t as

$$(Y_t|\boldsymbol{\theta}_t) \sim N[\mathbf{F}_t'\boldsymbol{\theta}_t, V_t],$$
$$(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \sim N[\mathbf{G}_t\boldsymbol{\theta}_{t-1}, \mathbf{W}_t].$$

Let D_t be the set containing all information up to and including time t. At time $t = 0$, the DLM is specified as

$$(\boldsymbol{\theta}_0|D_0) \sim N[\mathbf{m}_0, \mathbf{C}_0], \tag{3}$$

for some prior values $\mathbf{m}_0$, $\mathbf{C}_0$. At any subsequent time period t, *before* the datum Y_t is observed, the state equation $(\boldsymbol{\theta}_t|D_{t-1})$ is given by

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim N[a_t, R_t],$$

where

$$\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1},$$
$$\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t.$$

The one-step-ahead forecast or predictive distribution is given by

$$(Y_t|D_{t-1}) \sim N[f_t, Q_t],$$

where

$$f_t = \mathbf{F}_t'\mathbf{a}_t,$$
$$Q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + V_t.$$

After observing Y_t, we update the information set $D_t = \{D_{t-1}, y_t\}$, which will then be used to update the state vector; this update is essentially the posterior distribution of the state vector $(\boldsymbol{\theta}_t|D_t)$. These posterior quantities are computed via Bayes's theorem, and are given by

$$(\boldsymbol{\theta}_t|D_t) \sim N[m_t, C_t],$$

where

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t e_t,$$
$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t Q_t \mathbf{A}_t',$$
$$\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t Q_t^{-1},$$
$$e_t = y_t - f_t.$$

The DLM is represented by the quadruple $\{\mathbf{F}_t', \mathbf{G}_t, V_t, \mathbf{W}_t\}$. Usually $\mathbf{G}_t$ and the relevant values of the sequence $\mathbf{F}_t'$ are known. The evolution variance matrix $\mathbf{W}_t$ is also chosen by the modeler. Recall that $\mathbf{W}_t$ controls the stochastic variation of the model. If $\mathbf{W}_t = 0$, then there is no variation in the regression parameters, leading to a static regression model. Here, we specify $\mathbf{W}_t$ using the discount factor approach described by West and Harrison (1997), namely

$$\mathbf{W}_t = \frac{1-\delta}{\delta}\mathbf{G}\mathbf{C}_{t-1}\mathbf{G}'.$$

We show the details of the definition of $\mathbf{W}_t$, $\mathbf{F}_t'$ and $\mathbf{G}_t$ during the model specification process in Section 6.2.

The remaining element of the quadruple, V_t, is often unknown and large relative to the system variance $\mathbf{W}_t$. West and Harrison (1997) present a Bayesian learning procedure for an unknown observational variance, working in terms of the unknown precision parameter $\phi_t = 1/V_t$. A simple closed form Bayesian analysis is still available if we impose a particular structure on the $\mathbf{W}_t$ sequence and on the initial prior for $\boldsymbol{\theta}_0$. This structure enables a conjugate sequential updating procedure for ϕ_t in addition to $\boldsymbol{\theta}_t$. The conjugate analysis is based on gamma distributions for ϕ_t.

As West and Harrison (1997) describe, conditional on V_t being known, the DLM will be defined by

$$\text{Obs. eqn.: } (Y_t|\boldsymbol{\theta}_t) \sim N[\mathbf{F}_t'\boldsymbol{\theta}_t, V_t],$$
$$\text{Sys. eqn.: } (\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, V_t) \sim N[\mathbf{G}_t\boldsymbol{\theta}_{t-1}, V_t\mathbf{W}_t^*],$$
$$\text{Initial Information: } (\boldsymbol{\theta}_0|D_0, V_0) \sim N[\mathbf{m}_0, V_0\mathbf{C}_0^*].$$

Notice that all variances and covariances have V_t as a multiplier, providing a scale-free model in terms of the starred variances $\mathbf{C}_0^*$ and $\mathbf{W}_t^*$. For fixed values of V_t, the model coincides with the original model given by Eqs. (1)–(3), with the scale factor V_t simply being absorbed into these matrices.

If V_t is unknown, West and Harrison (1997) show that the normal distribution is replaced by Student-t distributions in the DLM. Therefore, the new DLM structure is given by

$$\text{Obs. eqn.: } (Y_t|\boldsymbol{\theta}_t) \sim N[\mathbf{F}_t'\boldsymbol{\theta}_t, V_t], \tag{4}$$
$$\text{Sys. eqn.: } (\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \sim T_{n_{t-1}}[\mathbf{G}_t\boldsymbol{\theta}_{t-1}, \mathbf{W}_t]. \tag{5}$$
Initial information:
$$(\boldsymbol{\theta}_0|D_0) \sim T_{n_{t-1}}[\mathbf{m}_0, \mathbf{C}_0], \tag{6}$$
$$(\phi_t|D_0) \sim G[n_0/2, n_0S_0/2], \tag{7}$$

with G denoting the Gamma distribution and $T_{n_{t-1}}[0, \mathbf{W}_t]$ being the Student-t distribution with n_{t-1} degrees of freedom. Note that, in specifying the prior, we must choose the prior estimate S_0 and the associated degrees of freedom n_0 in addition to m_0 and C_0; S_0 is a prior point estimate of the observational variance.

Now, at any subsequent time period t, *before* the datum Y_t is observed, the new state equation $(\boldsymbol{\theta}_t|D_{t-1})$ is given by

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim T_{n_{t-1}}[a_t, R_t],$$

where

$$\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1},$$
$$\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t.$$

The new one-step-ahead forecast or predictive distribution is given by

$$(Y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t], \tag{8}$$

where

$$f_t = \mathbf{F}_t'\mathbf{a}_t, \tag{9}$$

$$Q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + S_t. \tag{10}$$

The new posterior distribution of the state vector $(\boldsymbol{\theta}_t|D_t)$ is given by

$$(\boldsymbol{\theta}_t|D_t) \sim T_{n_{t-1}}[m_t, C_t],$$

where

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t e_t,$$
$$\mathbf{C}_t = \frac{S_t}{S_{t-1}}\left(\mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t'Q_t\right),$$
$$\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_tQ_t^{-1},$$
$$e_t = Y_t - f_t,$$

and

$$(\phi_t|D_t) \sim G[n_t/2, n_tS_t/2],$$

where

$$n_t = n_{t-1} + 1,$$
$$S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right).$$

If we are interested in forecasting k steps ahead, the forecasting distributions are given by

$$(\boldsymbol{\theta}_{t+k}|D_t) \sim T_{n_t}[\mathbf{a}_t(k), \mathbf{R}_t(k)], \tag{11}$$

$$(Y_{t+k}|D_t) \sim T_{n_t}[\mathbf{f}_t(k), \mathbf{Q}_t(k)], \tag{12}$$

where

$$f_t(k) = \mathbf{F}_{t+k}'\mathbf{a}_t(k), \tag{13}$$

$$Q_t(k) = \mathbf{F}_{t+k}'\mathbf{R}_t(k)\mathbf{F}_{t+k} + S_t, \tag{14}$$

$$\mathbf{a}_t(k) = \mathbf{G}_{t+k}\mathbf{a}_t(k-1),$$

$$\mathbf{R}_t(k) = \mathbf{G}_{t+k}\mathbf{R}_t(k-1)\mathbf{G}_{t+k}' + \mathbf{W}_{t+k}, \tag{15}$$

with starting values $\mathbf{a}_t(0) = \mathbf{m}_t$ and $\mathbf{R}_t(0) = \mathbf{C}_t$.

6.2. Basin inflow model specification

The univariate DLM is classified by the quadruple $\{\mathbf{F}_t, \mathbf{G}_t, V_t, \mathbf{W}_t\}$ and the initial information set D_0. In this section, we go through the DLM specification and define this quadruple and the prior distributions for $(\boldsymbol{\theta}_0|D_0)$ and $(\phi_0|D_0)$. Following the recommendations of West and Harrison (1997), we use the first two years (1979, 1980) of historical data to formulate these priors, where these prior estimates are obtained via ordinary least squares.

A great advantage of the Bayesian dynamic model is the possibility of combining regression, trend and seasonal components to create a single DLM. According to the partial autocorrelation function plots in Fig. 5, we will need an autoregressive component of order 1, 2, or 3 depending on the basin; hence, let l generically denote the order of the autoregressive model.

The seasonal component in the inflows has to be modeled. One approach is to define a number g which corresponds to the seasons of the year, and then categorize each time period t from 1 to g depending on the season to which t belongs. Note that g will depend on the basin inflow time series. For instance, if we consider $g = 4$ as representing the four seasons of the year, we would have a categorical variable, Z_t, such that

$$Z_t = \begin{cases} 1 & \text{if } t \text{ is January, February or March;} \\ 2 & \text{if } t \text{ is April, May or June;} \\ 3 & \text{if } t \text{ is July, August or September;} \\ 4 & \text{otherwise.} \end{cases}$$

Categorical variables with two levels may be entered as predictors in a regression directly. However, if $g > 2$, in order to obtain a meaningful regression, we need to create $g - 1$ binary or dummy variables, say $X_{i,t}$ for $i = 1, \ldots, g-1$, with two levels (0 or 1) to represent Z_t such that

$$X_{i,t} = \begin{cases} 1 & \text{if } t \text{ is in group } i; \\ 0 & \text{otherwise;} \end{cases} \qquad \text{for } i = 1, \ldots, g-1.$$

Now we can collate our predictor variables in one single vector to get $\mathbf{F}_t$. Suppose that we have l lagged variables from the autoregressive component, plus the $g-1$ dummies from the seasonal component and an intercept term. Our design vector, $\mathbf{F}_t$, will have $n = l+g$ components such that

$$\mathbf{F}_t' = \begin{bmatrix} 1 & Y_{t-1} & \ldots & Y_{t-l} & X_{1,t} & X_{2,t} & \ldots & X_{g-1,t} \end{bmatrix}.$$

The first component of $\mathbf{F}_t$ is 1 because it corresponds to the intercept term. The $\boldsymbol{\theta}_t$ coefficients will be the corresponding regression coefficients. In addition, if we include other regressors like the climate variables, we need to further augment vector $\mathbf{F}_t$ in order to incorporate these new variables, with a corresponding change to the dimension of the vector $\boldsymbol{\theta}_t$. The new number of predictor variables will be $n = l + g + c$, where c is the number of climate variables in the model. In this paper, we test only two possible climate predictors, namely precipitation and the NINO 3.4 index.

Rewriting the observation Eq. (4) in terms of the actual variables for the complete model, we have

$$\begin{aligned} Y_t = {} & \theta_{t,1} + \theta_{t,2}Y_{t-1} + \theta_{t,3}Y_{t-2} + \cdots + \theta_{t,l}Y_{t-l} \\ & + \theta_{t,l+1}X_{1,t} + \theta_{t,l+2}X_{2,t} \cdots + \theta_{t,l+g}X_{g-1,t} \\ & + \theta_{t,l+g+1}P_t + \theta_{t,l+g+2}N_t + \nu_t, \end{aligned}$$

where

Y_t = Predicted basin inflow at time t;

Y_{t-1} = Observed basin inflow at time $t-1$;

Y_{t-2} = Observed basin inflow at time $t-2$;

Y_{t-l} = Observed basin inflow at time $t-l$;

$X_{1,t}$ = Seasonal dummy 1 at time t;

$X_{2,t}$ = Seasonal dummy 2 at time t;

$X_{g-1,t}$ = Seasonal dummy $g-1$;

P_t = Precipitation at time t;

N_t = NINO 3.4 index at time t;

v_t = error term at time t.

The error term is given by $v_t \sim N[0, V_t]$. With respect to the observational variance V_t, we can see from Fig. 5 that the data are non-normal. We therefore decided to work with a variance law for the observational variance, given by

$$V_t = k(\mu_t)V_t,$$

where $\mu_t = \mathbf{F}_t'\boldsymbol{\theta}_t$ is the level of the series at time t (see Harrison & Stevens, 1971; Stevens, 1974). Since we are working with the reciprocal of V_t, i.e., ϕ_t, we have

$$V_t = k(\mu_t)\phi^{-1}.$$

Given the right-skewed histograms in Fig. 5, here we consider

$$k(\mu_t) = \mu_t^p,$$

with $p = 2$, since this is equivalent to a log transformation of the data, according to Migon, Gamerman, Lopes, and Ferreira (2005, chap. 19). Note that, by adopting a variance law, we actually allow V_t to change over time. The new variance law would change the updating Eq. (10) for the one-step ahead forecast and Eq. (14) for the k-step-ahead forecast, so that

$$Q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + \mu_t^p S_t,$$

$$Q_t(k) = \mathbf{F}_{t+k}'\mathbf{R}_t(k)\mathbf{F}_{t+k} + \mu_t^p S_t,$$

where $\mu_t = f_t$, with f_t being given by Eq. (9) for the one-step-ahead forecast or Eq. (13) for the k-step-ahead forecast.

For the system equations, we assume that the system matrix $\mathbf{G}_t$ is constant over time and equal to the identity matrix. Hence, the current state $\boldsymbol{\theta}_t$ is only dependent on the previous state $\boldsymbol{\theta}_{t-1}$, i.e., a simple random walk. Explicitly rewriting Eq. (5), we get

$$\begin{bmatrix}\theta_{t,1}\\ \theta_{t,2}\\ \vdots\\ \theta_{t,n}\end{bmatrix} = \begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}\begin{bmatrix}\theta_{t-1,1}\\ \theta_{t-1,2}\\ \vdots\\ \theta_{t-1,n}\end{bmatrix} + \begin{bmatrix}\omega_{t,1}\\ \omega_{t,2}\\ \vdots\\ \omega_{t,n}\end{bmatrix},$$

where

n = number of variables $(l+g+c)$;

$\theta_{1,t}$ = intercept term at time t;

$\theta_{2,t}$ = regression coefficient at time t associated with the observed inflow at time $t-1$;

$\theta_{t,n}$ = regression coefficient at time t associated with the NINO 3.4 index at time t;

$\omega_{t,1}$ = error term at t for regression coefficient $\theta_{t,1}$;

$\omega_{t,2}$ = error term at t for regression coefficient $\theta_{t,2}$;

$\omega_{t,n}$ = error term at t for regression coefficient $\theta_{t,n}$.

The error term vector given by $\boldsymbol{\omega}_t \sim N[\mathbf{0}, \mathbf{W}_t]$ represents purely random, unpredictable changes in level between times $t-1$ and t. For the evolution variance $\mathbf{W}_t$, we adopt the discount factor approach. A different δ can be specified for each set of predictors. Suppose

$$\mathbf{P_t} = \mathbf{GC_{t-1}G}',$$

and let δ_T, δ_S and δ_C be the discount factors associated with the autoregressive, seasonal and climate regression predictors, respectively. The evolution variance matrix is then defined as

$$\mathbf{W_t} = \text{block diag}\{\mathbf{P_{tT}}(\delta_\mathbf{T}^{-1} - \mathbf{1}), \mathbf{P_{tS}}(\delta_\mathbf{S}^{-1} - \mathbf{1}), \mathbf{P_{tC}}(\delta_\mathbf{C}^{-1} - \mathbf{1})\},$$

where $\mathbf{P}_{tT}$ is the upper left $(l+1)\times(l+1)$ block of $\mathbf{P_t}$ that corresponds to the autoregressive component, $\mathbf{P_{tS}}$ is the middle $(g-1)\times(g-1)$ block of $\mathbf{P_t}$ that corresponds to the seasonal component, and $\mathbf{P}_{tC}$ is the lower right c x c block of $\mathbf{P_t}$ that corresponds to the climate variable regression component, with c variables. The values of δ_T, δ_S and δ_C differ depending on the basin. As was described by West and Harrison (1997), these δs usually take values between 0.8 and 1.0, with smaller values anticipating greater changes in the model parameters at each stage and 1.0 being the static model. In our analysis, based on a little trial and error, we choose $\delta_T \approx 0.92$, $\delta_S \approx 0.96$ and $\delta_C \approx 0.90$. Note that δ_S is larger than both δ_T and δ_C, meaning that less information is being obtained about the seasonal parameters than about the trend and climate in each month.

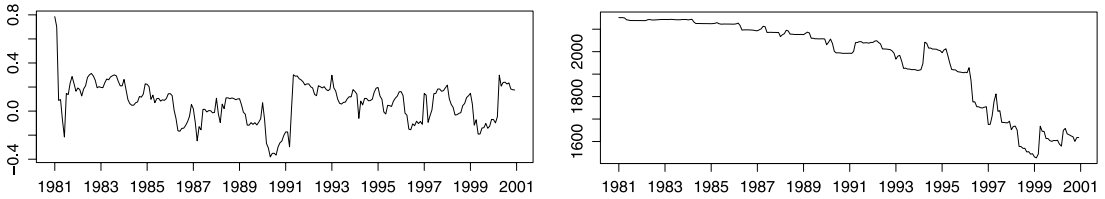
Recall that n is the number of predictor variables, meaning that at least n observations are necessary in order to fully specify the joint distributions. As was noted earlier, we use the first two years of historical data to obtain the prior moments $\mathbf{m_0}$ and $\mathbf{C_0}$ from Eq. (6) for the regression coefficients via a simple linear regression.

Because we are working with the reciprocal of V_t, ϕ_t, we also need to specify the moments n_0 and S_0 from Eq. (7). The degrees of freedom are usually given by $n_t = t - n$. Since we have already used two years of data for the prior, $n_0 = 24 - n$. We set S_0 to 0.1^2, to reflect our somewhat vague prior knowledge.

6.3. Retrospective reference analysis

- Data from January 1979 to December 1980 are used to construct the prior parameter values described in the previous section.
- Data from January 1981 to December 2000 are used in the estimation phase of the Bayesian model detailed in the previous section. This is the in-sample component of the model.
- The out-of-sample data are from January 2001 to December 2005. These were used to test the forecasting accuracy of the model. All of the results reported in Section 7 of the paper are based on these data.

In the estimation phase, we allow the model to run until December 2000 in order to minimize the effect of the first two years of prior data. That is, we proceed with the one-step-ahead forecast analysis and the sequential updating of the state equations until we reach December 2000. Fig. 8



(a) Lag 1 coefficient $\theta_{t,2}$. (b) Dummy 2 coefficient $\theta_{t,l+2}$.

Fig. 8. Evolution of regression coefficients $\boldsymbol{\theta}$ during the model estimation process.

shows how some of the $\boldsymbol{\theta}$ coefficients change during these 20 years of analysis for the Paranaiba basin. The spikes in the beginning are the results of the prior selection. Since we are not dealing with a stationary series, the dynamic model captures the evolutionary nature of the regression parameters better than a static model.

Based on the mathematical details presented earlier, we proceed to update and predict both the state vector and the inflow time series at each point in time. This process was executed for each basin using the appropriate numbers of lags and seasonal groups for each basin. Thus, the number of lags could be 1, 2 or 3, and the number of groups could be 1, 2, 3, 4, 6 or 12. Note that if $g = 1$, it means that that particular basin time series has no seasonality. If $g = 2$, we have two seasons within the year, with 6 consecutive months for each. If $g = 3$, each season will have 4 months, and so on, until we reach $g = 12$, where each month of the year corresponds to a different season.

The process for selecting the appropriate number of lags and seasonal dummies is based entirely on the practical considerations relating to Brazil's geography that were described in Sections 2 and 3. The starting point was $g = 4$ and $l = 1$. We observed that the lag 1 model has the best fit for all basins, mainly because adding more lagged variables to the model did not result in significant improvements to the one-step-ahead forecasting error. Indeed, it actually worsened the results for some basins, due to over fitting. With respect to the number of seasonal groups, the forecasts for the basins in the second principal component (collectively referred to as Uruguay) were better if we decreased g. This is consistent with the correlation analysis and the seasonal aspects of the basin time series presented in Section 3. Within the Uruguay group of basins, the forecasting errors show that, for the Parana, Paranapanema and Iguacu basins, the best model is attained with $g = 2$, whereas for the Uruguay and South Atlantic basins, the best model is with $g = 1$. For all of the other basins in the Amazon and East Atlantic groups of basins (the first and third principal components), the best model was achieved when $g = 6$.

7. Forecasting results

Recall that the Brazilian authorities require 5-year-ahead forecasts of inflows for each of the basins. The discussion in the preceding section was confined to the models' in-sample performances using the Bayesian DLM. Using the best models from the previous section, we are now ready to forecast five years into the future and compare such forecasts with actual data that were set aside for validating the models' out-of-sample forecasting accuracies. The forecasting horizon is from January 2001 to December 2005, meaning that the estimation analysis stops in December 2000. The predictive distributions in Eqs. (11) and (12) are derived for $k = 1$ to $k = 60$, i.e., a 5-year horizon.

To check the accuracy, we use two popular statistical metrics: the mean square relative error (MSRE), given by

$$\text{MSRE} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{f_t - y_t}{y_t}\right)^2,$$

and the mean absolute relative error (MARE), given by

$$\text{MARE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{f_t - y_t}{y_t}\right|,$$

where y_t is the realized inflow at time t, f_t is the forecasted value at t and T is the forecasting horizon.

Third metric that is commonly used in the context of inflow forecasting model performances is the Nash–Sutcliffe model efficiency coefficient (NSC) (Nash & Sutcliffe, 1970). It measures the accuracy of the model's prediction based on the observed mean of the inflow time series data. This metric is given by

$$\text{NSC} = 1 - \frac{\sum_{t=1}^{T}(y_t - f_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2},$$

where $\bar{y}$ is the observed mean. The NSC varies from $-\infty$ to 1, with 1 being the perfect model. If $NSC = 0$, the model performance is as good as the mean of the observed data, and if $NSC < 0$, the observed mean is a better predictor than that obtained by the model.

In what follows, recall that, based on the principal component analysis detailed in Section 3, the first component (labeled Amazon) comprises nine basins; the second component (labeled Uruguay) comprises five basins; and the third component (labeled East Atlantic) comprises only one basin.

Fig. 9 depicts the average MSRE by component, weighted by the basin installed capacities for three different models. *Model 1* corresponds to the DLM with only the autoregressive and seasonal components. Note that the errors are small for the first component (Amazon) basins, but high for the others. *Model 2* is the same as Model 1 but with the basin's precipitation added as a predictor variable. Note that the errors decrease in all cases, especially for the basins in the second (Uruguay) and

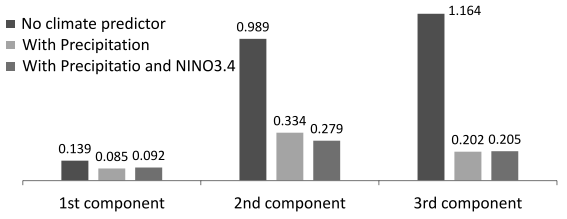


Fig. 9. Average mean square relative error (MSRE) by component, weighted by the basin installed capacity for three different models: Model 1 with no climate predictor; Model 2 with precipitation as a predictor; and Model 3 with precipitation and NINO3.4 as predictors.

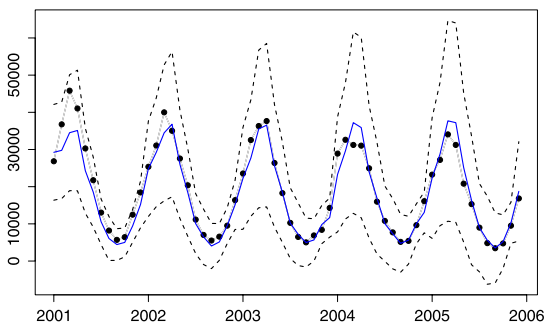
third (East Atlantic) principal components. *Model 3* is the same as Model 2 but also includes the NINO 3.4 series as a predictor. It is still possible to see some improvement for the second component basins; this is to be expected, since these basins are the ones that are most correlated with the NINO 3.4, as is shown in Table 2. However, for the first and third principal component basins, the NINO 3.4 variable does not improve the results. For these latter models, any incremental improvement might be a consequence of the El Niño and La Niña effects that are already represented in the basin's precipitation time series.

Some further analysis of the second component basins was conducted in order to improve the results shown in Fig. 9. We found that if we grouped the reservoirs in basins 4 and 6, we could reduce the errors to 0.555, 0.199 and 0.227 for Models 1, 2 and 3, respectively.

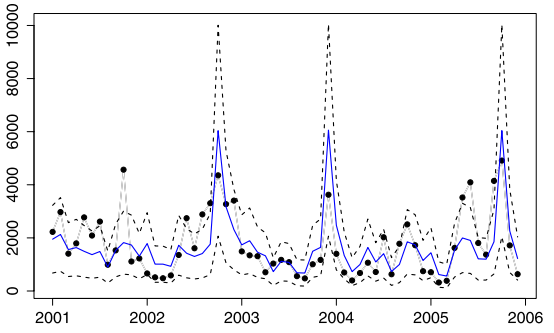
Because we are using a Bayesian approach, the model outcome is a distribution forecast rather than a point forecast. Fig. 10 depicts the forecast distribution analysis for the Amazon, Uruguay and East Atlantic basins using Model 2, i.e., the model with an autoregressive component of order 1, a seasonal component and only precipitation as a climate predictor. The blue line represents the point forecasts $f_t(k)$, which are the means of the forecasting distributions; the dotted lines provide the 95% prediction intervals for Y_{t+k} given D_t; and the black dots correspond to the observed inflow values. Note that even though the forecast for the Uruguay basin is not as good as those for the Amazon and East Atlantic basins, the actual inflows still fall within the 95% bounds, barring a few exceptions.

Next consider Table 3, which shows the 95% empirical coverage using 60 sets of predictive distributions, one for each of the 15 basins. The model performs very well, since the overall percent correct across all basins ranges from 86.7% to 100%. It should be noted that we have reported 95% prediction intervals because of convention, though there is nothing unique about 95% ranges from a decision-making or statistical perspective; see for instance Kadane (1990) and Raiffa and Schlaifer (1961).

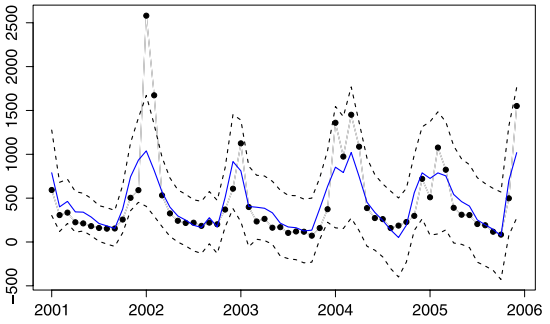
Table 4 compares the Bayesian Model 2's 5-year-ahead forecasts with the forecasts obtained using the periodic autoregressive (PAR) model, which is the model that is currently used in Brazil for all hydropower scheduling and operations decisions (Maceira & Damazio, 2006). The comparisons were made using the MSRE, MARE and NSC metrics. The results show that our Bayesian model produces substantially better forecasts than the PAR model for all of the basins. Indeed, the NSC values show that,



(a) Amazon basin.



(b) Uruguay basin.



(c) East Atlantic basin.

Fig. 10. 5-year-ahead forecasting distributions analysis for the water inflows using Model 2. The blue line represents the point forecasts $f_t(k)$, which are the means of the forecasting distributions; the dotted lines provide the 95% prediction intervals for Y_{t+k} given D_t; and the black dots correspond to the observed inflow values. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The out-of-sample forecast horizon is from January 2001 to December 2005, a total of 60 months. Hence, there are 60 unique, monthly 95% prediction intervals for each of the basins in the table. Panels (a), (b) and (c) of Fig. 10 show these 95% intervals for the Amazon, Uruguay and East Atlantic basins, respectively.

	Basin	Number of months true data points were included in the 95% prediction intervals	Overall correct (%)
1	Paranaiba	58 out of 60	96.7%
2	Grande	53 out of 60	88.3%
3	Tiete	56 out of 60	93.3%
4+6	Parana/Iguacu	56 out of 60	93.3%
5	Paranapanema	57 out of 60	95.0%
7	Paraguay	55 out of 60	91.7%
8	Uruguay	52 out of 60	86.7%
9	South Atlantic	53 out of 60	88.3%
10	Southeast Atlantic	58 out of 60	96.7%
11	East Atlantic	58 out of 60	96.7%
12	Sao Francisco	60 out of 60	100.0%
13	Parnaiba	58 out of 60	96.7%
14	Tocantins	59 out of 60	98.3%
15	Amazon	60 out of 60	100.0%

for the basins in the second and third components, the observed mean is a better forecast than those obtained from the PAR model. Clearly, this has severe policy implications for hydropower management in Brazil.

The largest MSRE from our analysis for the forecasting interval 2001 to 2005 is 0.294. One might wonder what happens for other time periods. Fig. 11 shows the average MSRE by component, weighted by the basin installed capacity, for various different time periods. The forecasting horizon is still five years ahead; therefore, 1996 on the x axis corresponds to the MSRE for the time period from January 1996 to December 2001; 1997 corresponds to the MSRE from January 1997 to December 2002, and so on. Note that the errors are approximately the same regardless of the rolling window adopted.

8. Conclusion

In this paper, a Bayesian dynamic linear model is developed for the hydropower reservoir water inflow forecasting in Brazil. First, the reservoirs were grouped by basin to reduce the dimension of the problem. The proposed models for each basin have the same basic DLM structure but distinct features that better represent the characteristics of each basin. Three DLMs were tested for each basin: the first had only an autoregressive and seasonal component, the second also incorporated the basin's precipitation as a predictor variable, and the third incorporated both precipitation and the NINO 3.4 index as predictor variables.

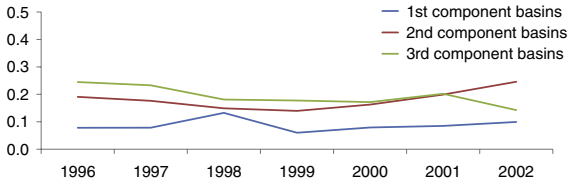


Fig. 11. Average mean square relative error (MSRE) by component, weighted by the basin installed capacity, for different time periods.

The MSRE results show that including the NINO 3.4 does not produce better forecasts, implying that precipitation alone accounts for the NINO 3.4 effect on the inflows. Other model performance criteria, such as the MARE and NSC, were also used to compare our results with the traditional periodic autoregressive model which is currently used in Brazil for long-term inflow forecasting. The Bayesian DLM with precipitation drastically outperforms the current protocol in Brazil. This should help decision makers to deal with scheduling and related operational issues in hydropower management better.

Moreover, one great advantage of the proposed model is the possibility of expert intervention via an assessment of subjective prior knowledge. One could use this technique to ensure that the associated forecasting model continues to produce reliable forecasts, even if sudden changes occurred in the basin inflow series in response to climate or land use changes.

For the climate variables, since we have been testing the model for some time in the past, we have used the observed data for period t from the data sets described in Section 4 as predictors. However, in reality, when we forecast five years ahead, we will need a five-year forecast of all of the climate variables as well. Therefore, our inflow forecasting model's performance is dependent on the precipitation

Table 4
Comparison of the forecasting performances of the proposed dynamic linear model (DLM) and the periodic autoregressive model (PAR), considering the mean square relative error (MSRE), mean absolute relative error (MARE) and Nash–Sutcliffe coefficient (NSC) criteria.

	Basin	MSRE		MARE		NSC	
		DLM	PAR	DLM	PAR	DLM	PAR
1	Paranaiba	0.038	0.196	0.153	0.330	0.803	0.332
2	Grande	0.145	0.330	0.261	0.422	0.704	0.261
3	Tiete	0.034	0.119	0.144	0.275	0.791	0.142
4+6	Parana/Iguacu	0.199	0.707	0.303	0.611	0.471	−0.806
5	Paranapanema	0.069	0.339	0.195	0.447	0.457	−1.914
7	Paraguay	0.049	0.075	0.180	0.203	0.741	0.452
8	Uruguay	0.294	2.292	0.434	0.995	0.413	−0.770
9	South Atlantic	0.222	1.334	0.344	0.780	0.595	−0.658
10	Southeast Atlantic	0.062	0.384	0.209	0.462	0.749	0.259
11	East Atlantic	0.202	2.511	0.364	1.047	0.652	−0.656
12	Sao Francisco	0.215	1.029	0.358	0.731	0.728	0.023
13	Parnaiba	0.025	0.226	0.122	0.353	0.795	0.214
14	Tocantins	0.049	0.277	0.179	0.403	0.863	0.630
15	Amazon	0.018	0.163	0.113	0.290	0.931	0.788

forecasting model. For future work, it would be interesting to couple our model with an atmospheric model, such as the ETA model. One could try to incorporate climate change scenarios into the model and see what would happen to the basin inflow, and consequently the assured future energy, for each reservoir.

One limitation of the model proposed in this paper is that it fails to account for spatial information. It would be interesting to see whether the quality of the forecasts would improve as a result of adding this extra feature into the DLM framework. However, it would mean that we would lose the elegant, closed form solution of the DLM approach in this paper, since the filtering equations would then have to be estimated via Markov chain Monte Carlo methods (Gamerman & Lopes, 2006).

Acknowledgments

This research is dedicated to the memory of Dr. Elmira Popova, whose untimely demise is a great loss to everyone who knew and worked with her. The authors thank the CAPES/Fulbright program in Brazil for financial support; the GES-DISC Interactive On line Visualization and Analysis Infrastructure (Giovanni) as part of the National Aeronautics and Space Administration NASA's Goddard Earth Sciences (GES) Data and Information Services Center (DISC); and the International Research Institute for Climate and Society (IRIS) Data Management Center for the climate variables data sets. We also thank the editor, associate editor and reviewers for their valuable comments and suggestions.

References

Bender, M., & Simonovic, S. (1994). Time-series modeling for long-range stream-flow forecasting. *Journal of Water Resources Planning and Management*, *120*, 857–870.

Berger, J. O., & Rios-Insua, D. (1998). Recent developments in Bayesian inference with applications in hydrology. In E. Parent, P. Hupert, B. Bobee, & J. Miquel (Eds.), *Statistical and Bayesian methods in hydrological science* (pp. 43–62). Paris: UNESCO Press.

Box, G. E. P., & Jenkins, G. M. (Eds.) (1976). *Time series analysis: forecasting and control* (2nd ed.). San Francisco: Holden-Day.

Chen, C. (1997). Robustness properties of some forecasting methods for seasonal times series: a Monte Carlo study. *International Journal of Forecasting*, *13*, 269–280.

Collisschonn, W., Allasia, D., da Silva, B. C., & Tucci, C. E. M. (2007). The MGB-IPH model for large-scale rainfall-runoff modelling. *Hydrological Sciences Journal*, *52*(5), 878–895.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman & Hall/CRC.

Harrison, P. J., & Stevens, C. F. (1971). A Bayesian approach to short-term forecasting. *Operations Research Quarterly*, *22*, 341–362.

Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*. Elsevier.

Kadane, J. B. (1990). A statistical analysis of adverse impact of employer decisions. *Journal of the American Statistical Association*, *85*, 925–933.

Krishnaswamy, J., Halpin, P. N., & Richter, D. D. (2001). Dynamics of sediment discharge in relation to land-use and hydro-climatology in a humid tropical watershed in Costa Rica. *Journal of Hydrology*, *253*, 91–109.

Krishnaswamy, J., Lavine, M., Richter, D. D., & Korfmacher, K. (2000). Dynamic modeling of long-term sedimentation in the Yadkin river basin. *Advances in Water Resources*, *23*, 881–892.

Kumar, D. N., & Maity, R. (2008). Bayesian dynamic modeling for nonstationarity hydroclimatic time series along with uncertainty quantification. *Hydrological Processes*, *22*, 3488–3499.

Maceira, M.E.P., & Bezerra, C.V. (1997). Stochastic streamflow model for hydroelectric systems. In *5th probabilistic methods applied to power systems-PMAPS*.

Maceira, M. E. P., & Damazio, J. M. (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.

Marriott, J., & Newbold, P. (1998). Bayesian comparison of ARIMA and stationary ARMA models. *International Statistical Review*, *66*, 323–336.

Migon, H. S., Gamerman, D., Lopes, H. F., & Ferreira, M. A. R. (2005). Dynamic models. In *Handbook of statistics*: *Vol. 25. Bayesian thinking: modeling and computation* (pp. 557–592).

Migon, H. S., & Monteiro, A. B. S. (1997). Rain-fall modeling: an application of Bayesian forecasting. *Stochastic Hydrology and Hydraulics*, *11*, 115–127.

Mondal, M. S., & Wasimi, S. A. (2006). Generating and forecasting monthly flows of the ganges river with PAR model. *Journal of Hydrology*, *323*, 41–56.

Moradkhani, H., Hsu, K., Gupta, H. V., & Sorooshian, S. (2004). Improved streamflow forecasting using self-organizing radial basis function artificial neural networks. *Journal of Hydrology*, *295*, 246–262.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I — a discussion on principles. *Journal of Hydrology*, *10*, 282–290.

Noakes, D. J., McLeod, A. I., & Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting*, *1*, 179–190.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Cambrigde, MA: MIT Press.

Ravines, R. R., Schmidt, A. M., Migon, H. S., & Renno, C. D. (2008). A joint model for rainfall-runoff: the case of Rio Grande basin. *Journal of Hydrology*, *353*, 189–200.

Rios-Insua, D., Salewicz, K. A., Muller, P., & Bielza, C. (1997). Bayesian methods in reservoir operations: the Zambezi River case. In S. French, & J. Smith (Eds.), *The practice of Bayesian analysis* (pp. 107–130).

Salas, J. D., Delleur, J. W., Yevjevich, V., & Lane, W. L. (1980). *Applied Modeling of Hydroelectric Series*. Water Resources Publications.

Solomatine, D. P. (2002) Data-driven modelling: paradigm, methods, experiences. In *Proceedings of the 5th international conference on hydroinformatics* (pp. 757–763).

Stevens, C. F. (1974). On the variability of demand for families of items. *Operations Research Quartely*, *25*, 411–419.

Trenberth, K. E. (1997). The definition of El Ninõ. *Bulletin of the American Meteorological Society*, *78*(12), 2771–2777.

Wang, W. (2006). *Stochasticity, nonlinearity and forecasting of streamflow processes*. IOS Press.

West, M. (2013). Bayesian dynamic modeling. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 145–166). Clarendon, England: Oxford University Press.

West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models*. Springer.

Yurekli, K., Kurunc, A., & Ozturk, F. (2005). Application of linear stochastic models to monthly flow data of Kelkit stream. *Ecological Modeling*, *183*, 67–75.

Luana M. M. Lima received her B.Sc. and M.Sc. degrees in Electrical Engineering, both from Federal University at Itajubá, Itajubá, Brazil, in 2005 and 2007, respectively. She got her Ph.D. in Operations Research and Industrial Engineering from the Mechanical Engineering Department at the University of Texas at Austin in 2011. Her areas of interest are electricity markets, including transmission and distribution regulation, and operations research, where her focus is on optimization techniques and Bayesian statistics.

Elmira Popova graduated with an MS in Mathematics from University of Sofia, Bulgaria, in 1985, and a Ph.D. in Operations Research from Case Western Reserve University, Cleveland, OH, in 1995. She joined the University of Texas at Austin in 1995 as an Assistant Professor in the Department of Mechanical Engineering, Graduate Program in Operations Research and Industrial Engineering. In 2008 she was named a Fulbright Scholar, and in 1999 she was given the Halliburton/Brown & Root young faculty excellence award in teaching and research. Dr. Popova specialized in stochastic processes, computational Bayesian statistics, and stochastic optimization. Her research has been funded by NSF, DNDO, NRC, DOE, STPNOC, EPRI, and several other industrial sponsors.

Paul Damien holds a Ph.D. in Mathematics from Imperial College, London, and is currently the B.M. Rankin Jr. Professor of Business, and Professor of Information, Risk and Operations Management at the McCombs School of Business at the University of Texas in Austin. He is also an Elected Fellow of the Royal Statistical Society of England. His research interests lie at the intersection of Bayesian methods and stochastic optimization. He has over 75 publications in leading academic journals, and industry reports.