International Journal of Forecasting 39 (2023) 98–109

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Beta autoregressive moving average model selection with application to modeling and forecasting stored hydroelectric energy

Francisco Cribari-Neto [a,*], Vinícius T. Scher [a], Fábio M. Bayer [b]

[a] Departamento de Estatística, Universidade Federal de Pernambuco, Recife/PE, Brazil
[b] Departamento de Estatística and LACESM, Universidade Federal de Santa Maria, Santa Maria/RS, Brazil

ARTICLE INFO

Keywords:
βARMA model
Bootstrap
Forecasting
Information criterion
Model selection
Stored hydroelectric energy

ABSTRACT

We evaluate the accuracy of model selection and associated short-run forecasts using beta autoregressive moving average (βARMA) models, which are tailored for modeling and forecasting time series that assume values in the standard unit interval, (0, 1), such as rates, proportions, and concentration indices. Different model selection strategies are considered, including one that uses data resampling. Simulation evidence on the frequency of correct model selection favors the bootstrap-based approach. Model selection based on information criteria outperforms that based on forecasting accuracy measures. A forecasting analysis of the proportion of stored hydroelectric energy in South Brazil is presented and discussed. The empirical evidence shows that model selection based on data resampling typically leads to more accurate out-of-sample forecasts.

1. Introduction

The beta regression model was introduced by Ferrari and Cribari-Neto (2004) and has been extensively used with responses that assume values in the standard unit interval, (0, 1), such as rates, proportions, and concentration indices. It is assumed that the responses are independently distributed, and hence the model is not useful for time series modeling. An extension of the model for serially dependent random variables was introduced by Rocha and Cribari-Neto (2009, 2017). It incorporates autoregressive and moving average dynamics, allows for the inclusion of fixed covariates, and became known as the beta autoregressive moving average (βARMA) model. A novel feature of the model is that it accounts for the double bounded nature of the data and will never yield fitted values or out-of-sample forecasts that lie outside the standard unit interval. Additionally, the model accounts for the inherent non-constant variance pattern of random variables in the standard unit interval, such that the variance is smaller when the variable mean is close to zero or one, and is larger otherwise. The model can be used to produce out-of-sample forecasts of time series that assume values in (0, 1).

Even though the dynamic structure of the βARMA model is similar to that of the Gaussian ARMA model, there are some important differences between the two classes of models. For instance, unlike the latter, the errors in the former are not innovations that drive the stochastic process. Instead, the errors are, as we shall see, defined in a residual fashion, as in the class of generalized autoregressive moving average (GARMA) models proposed by Benjamin, Rigby, and Stasinopoulos (2003). Also, as noted above, the βARMA model is inherently heteroskedastic.

βARMA data modeling follows the standard Box–Jenkins approach, which consists of (i) model identification, (ii) parameter estimation, and (iii) diagnostic

* Corresponding author.
E-mail addresses: cribari@de.ufpe.br (F. Cribari-Neto), vinitscher@gmail.com (V.T. Scher), bayer@ufsm.br (F.M. Bayer).

https://doi.org/10.1016/j.ijforecast.2021.09.004
0169-2070/

analysis; for details regarding this approach, see Box, Jenkins, Reinsel, and Ljung (2015). Parameter estimation is carried out by conditional maximum likelihood based on the underlying assumption that the variable of interest at each point in time follows the beta law; see Rocha and Cribari-Neto (2009, 2017). Diagnostic analysis based on portmanteau testing inferences for fitted βARMA models was developed by Scher, Cribari-Neto, Pumi, and Bayer (2020). It remains to be established whether traditional model selection schemes work well when applied to βARMA data modeling and which model selection strategy is to be preferred, especially when the sample size is not large. The information criteria that are commonly employed for selecting models to be used for producing out-of-sample forecasts were not developed for dynamic models tailored to double bounded time series and warrant investigation in that context. As noted above, the βARMA dynamics are not driven by sequential realizations of white noise innovations as in Gaussian ARMA models. Given the different dynamic natures of the two processes, it is not clear that model selection strategies that perform best in traditional ARMA modeling will do so in βARMA modeling. It is thus important to assess the relative merits of different model selection strategies in the latter.

Practitioners may be tempted to resort to well-known model selection practices when selecting a model for forecasting double bounded time series. This is the case, for example, in Melchior, Zanini, Guerra, and Rockenbach (2021). The authors used the βARMA model to forecast mortality rates due to occupational accidents in three Brazilian states after performing model selection based on the well-known Akaike information criterion. A relevant question is: Can more accurate forecasts be obtained in most applications that deal with double bounded data by performing model selection based on alternative criteria? Our empirical results indicate that more accurate model selection may translate into more accurate double bounded out-of-sample forecasts.

We performed simulations to evaluate the finite sample performance of different information criteria. The numerical evidence we report shows that βARMA model selection becomes considerably more accurate when it is based on bootstrap resampling. In several cases, the frequency of correct model identification based on a bootstrap information criterion greatly exceeds that of the second-best performing criterion. We also evaluated the effectiveness of model selection based on measures of forecasting accuracy. The results favor a particular strategy based on directional forecasts, but they also indicate that more reliable model selection is achieved by using information criteria.

We present and discuss an empirical analysis in which the interest lies in forecasting future levels of stored hydroelectric energy in South Brazil. Climate change has added uncertainty to hydropower generation, and changing rainfall patterns and prolonged droughts have made it increasingly difficult to assess future river flows. As a result, the use of stored hydroelectric energy is increasingly important for hydropower generation. Interestingly, the most accurate forecasts were produced by βARMA models selected using bootstrap resampling. It is worth noticing that the same models were selected on the basis of the best performing strategies that employ measures of forecasting accuracy. The forecasts obtained from such models outperformed those computed from fitted models selected by alternative information criteria, in some cases by wide margins (e.g., over one-third). There is, thus, agreement between our empirical and numerical results.

The paper unfolds as follows. Section 2 briefly presents the βARMA model. In Section 3 we review some model selection criteria that can be used to determine the orders of the autoregressive and moving average βARMA dynamics. In Section 4 we report the results of extensive Monte Carlo simulations that were performed to evaluate the accuracy of different βARMA model selection strategies in small to moderate sample sizes. An empirical analysis is presented and discussed in Section 5. Finally, some concluding remarks are offered in Section 6.

2. A dynamic beta model

The βARMA model introduced by Rocha and Cribari-Neto (2009, 2017) is a dynamic model based on the beta regression model proposed by Ferrari and Cribari-Neto (2004). It is tailored for modeling random variables that assume values in $(0, 1)$ and evolve over time. It can be used, for example, to predict the future behavior of rates, proportions, and concentration indices.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a vector of n random variables such that, given the previous information set $\mathcal{F}_{t-1}$ (the smallest σ-algebra such that the variables $y_1, \ldots, y_{t-1}$ are measurable), y_t follows the beta law indexed by its conditional mean μ_t and a precision parameter, ϕ, for $t = 1, \ldots, n$. The conditional density of y_t given $\mathcal{F}_{t-1}$ is

$$f(y_t|\mathcal{F}_{t-1}) = \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi-1} \times (1-y_t)^{(1-\mu_t)\phi-1}, \quad 0 < y_t < 1, \tag{1}$$

$0 < \mu_t < 1$, $\phi > 0$, where $\Gamma(\cdot)$ is the gamma function. Here, $\mathbb{E}(y_t|\mathcal{F}_{t-1}) = \mu_t$ and $\mathrm{var}(y_t|\mathcal{F}_{t-1}) = \mu_t(1-\mu_t)/(1+\phi)$ are the conditional mean and conditional variance of y_t, respectively. For a given μ_t, the latter decreases as ϕ increases. Notice that the conditional variance of y_t is not constant; instead, it varies with μ_t. In particular, the conditional variance approaches zero as the conditional mean approaches zero or one.

The βARMA(p, q) model introduced by Rocha and Cribari-Neto (2009, 2017) assumes that y_t follows the above law with conditional mean, such that

$$g(\mu_t) = \alpha + \boldsymbol{x}_t'\boldsymbol{\beta} + \sum_{i=1}^{p} \varphi_i \left[g(y_{t-i}) - \boldsymbol{x}_{t-i}'\boldsymbol{\beta} \right] + \sum_{j=1}^{q} \theta_j r_{t-j}, \tag{2}$$

where $\boldsymbol{x}_t' \in \mathbb{R}^c$ is a set of non-random covariates at time t, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_c)' \in \mathbb{R}^c$ is a vector of parameters, and $g : (0, 1) \mapsto \mathbb{R}$ is a strictly monotonic and twice differentiable link function. Also, $\alpha \in \mathbb{R}$ is a scalar parameter and $p, q \in \mathbb{N}$ are the autoregressive and moving average orders associated with the $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ vectors of coefficients, respectively. Finally, r_t is an error term.

It is noteworthy the that βARMA model structure in (2) is similar to that of the class of GARMA models; see Benjamin et al. (2003). In both classes of models and unlike what happens in the Gaussian ARMA model, the error r_t is not an innovation that drives the stochastic process. Instead, r_t is defined in a residual fashion as $r_t = g(y_t) - g(\mu_t)$. Observe that, for fixed t, (2) only includes values of y_τ, and r_τ for $\tau < t$. Hence, both μ_t and r_t are $\mathcal{F}_{t-1}$-measurable.

βARMA parameter estimation can be performed by conditional maximum likelihood. The conditional log-likelihood function, given the first $a = \max\{p, q\}$ observations, is $\ell \equiv \ell(\boldsymbol{\nu}_k, \phi|\boldsymbol{y}) = \sum_{t=a+1}^{n} \log f(y_t|\mathcal{F}_{t-1})$, where $\boldsymbol{\nu}_k = (\alpha, \boldsymbol{\beta}', \boldsymbol{\varphi}', \boldsymbol{\theta}')'$ is the vector of mean parameters, and $f(y_t|\mathcal{F}_{t-1})$ is the beta density given in (1). For simplicity, we write $\ell(\boldsymbol{\nu}_k, \phi|\boldsymbol{y})$ as $\ell(\boldsymbol{\nu}_k|\boldsymbol{y})$ since ϕ is a fixed precision scalar. The model's score function and information matrix can be found in Rocha and Cribari-Neto (2017). Bias-corrected conditional maximum likelihood estimation was considered in Palm and Bayer (2018).

An extended version of the βARMA model that accommodates seasonal dynamics was proposed by Bayer, Cintra, and Cribari-Neto (2018) and a version of the model that includes fractional integration was introduced by Pumi, Valk, Bisognin, Bayer, and Prass (2019). Bayesian dynamic beta modeling was developed by Casarin, Valle, and Leisen (2012) and da Silva, Migon, and Correia (2011); the former considers Bayesian model selection for beta autoregressive processes. In what follows, we work with the standard, baseline βARMA model in the realm of frequentist statistical inference.

3. Model selection strategies

Model selection aims at selecting a statistical model from a set of candidate models on the basis of a given data set. The selected model is typically used for out-of-sample forecasting, provided that it yields a good data fit. The most commonly used model selection strategy is that based on criteria that penalize increases in the model's dimension. Typically, a set of models are fitted to the data, a given criterion is computed for each fitted model, and the model that displays the minimal criterion value is selected. Let $\boldsymbol{\nu}_k$ be the model's k-dimensional parameter vector. Its conditional maximum likelihood estimator is denoted by $\hat{\boldsymbol{\nu}}_k$, and the maximized log-likelihood function is written as $\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k)$. In what follows, we present model selection strategies based on (i) information criteria that penalize model augmentation, and (ii) forecasting accuracy measures.

The most commonly used criterion was introduced by Akaike (1974). It is obtained by minimizing the Kullback–Leibler distance between two densities and is known as the Akaike information criterion (AIC). The author showed that the model that minimizes minus two times the expected log-likelihood is the closest model to the true model according to the Kullback–Leibler information. He then used $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k)$ as an estimator of such a quantity, showed that its asymptotic bias is approximately equal to $-2k$ and arrived at the following information criterion:

$$\mathrm{AIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + 2k.$$

We note that $2k$, the bias correcting term, can be viewed as a penalization term, since it penalizes the model dimension augmentation when searching for the minimal criterion value. Based on data sets analyzed by Box and Jenkins, Ozaki (1978) showed that the use of Akaike's approach overcomes many difficulties with the identification procedure adopted in the authors' book. It was shown by Shibata (1976), however, that the AIC has a fixed overfitting probability asymptotically. As a consequence, the AIC tends to overestimate the model dimension. Several alternative criteria were then proposed, aiming at achieving more accurate model selection.

A criterion that incorporates a small sample correction and is asymptotically equivalent to the AIC was proposed by Sugiura (1978) and became known as $\mathrm{AIC_c}$; see also Hurvich and Tsai (1989). The corrected AIC is given by

$$\mathrm{AIC_c} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + 2k\left(\frac{n}{n-k-1}\right). \tag{3}$$

In essence, the new criterion includes an extra penalization term. According to Burnham and Anderson (2004), it should be preferred over the AIC unless $n/k > 40$ for the model with the largest value of k.

The Schwarz information criterion (SIC) was introduced by Schwarz (1978):

$$\mathrm{SIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + k\log n. \tag{4}$$

A novel feature of this criterion is that it is consistent. That is, the probability of selecting the true model tends to one as $n \to \infty$.

Model selection based on the SIC can be quite inaccurate in small samples, however. A modified version of the criterion was proposed by McQuarrie (1999). It incorporates a finite sample correction and can be expressed as

$$\mathrm{SIC_c} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + \frac{nk\log(n)}{n-k-1}.$$

The new criterion is asymptotically equivalent to the SIC and is expected to deliver superior model selection when the sample size is not large.

Hannan and Quinn (1979) focused on autoregressive model selection and proposed the Hannan–Quinn information criterion (HQIC):

$$\mathrm{HQIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + 2k\log(\log(n)).$$

Like the AIC and the SIC, its small sample behavior may be poor. A modified version of the criterion was introduced by McQuarrie and Tsai (1998):

$$\mathrm{HQIC_c} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + \frac{2nk\log(\log(n))}{n-k-1}. \tag{5}$$

The weighted-average information criterion (WIC) was proposed by Wu and Sepulveda (1998). It is based on the criteria given in (3) and (4). Let $A_c = 2kn/(n-k-1)$ and $B = k\log n$. The WIC can be expressed as

$$\mathrm{WIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + \frac{A_c^2 + B^2}{A_c + B}.$$

The above expression can also be obtained by combining (3) and (5). The authors showed that the WIC behaves

similarly to the AIC_c when the sample size is small, and similarly to the SIC in large samples. Like the SIC, the WIC is consistent.

Several authors considered the use of bootstrap resampling to estimate the penalty term used in the AIC. We present two bootstrap-based criteria. They are of the form

$$\text{EIC}i = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + \hat{B}_i,$$

$i = 1, 2$, where $\hat{B}_i$ is a bootstrap-based estimate of a penalty term that involves an expectation with respect to the distribution of the bootstrap sample. That is, the expected Kullback–Leibler discrepancy between the true and fitted models is estimated by means of data resampling. These criteria are said to be empirical because their penalty terms are estimated from the data using the bootstrap method. They are usually referred to as the empirical information criteria.

As noted by Shibata (1997, p. 379), a novel feature of the bootstrap estimation of B_i is that it is free from any expansion, whereas the AIC and related criteria are based on an expansion with respect to the model parameters. Hence, the bootstrap approach has wider applicability than the conventional bias correction. Additionally, we note that bootstrap bias correction is known to work well in other settings; see, e.g., Cribari-Neto, Frery, and Silva (2002) and Ospina, Cribari-Neto, and Vasconcellos (2006).

Let $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)'$ denote the bootstrap sample, and let N denote the number of bootstrap replications, so that we have N bootstrap samples, each denoted as $\boldsymbol{y}^*(j)$, $j \in \{1, \ldots, N\}$. The bootstrap estimates of $\boldsymbol{\nu}_k$ are $\hat{\boldsymbol{\nu}}_k^*(j)$, $j \in \{1, \ldots, N\}$, where $\hat{\boldsymbol{\nu}}_k^*(j)$ is obtained by maximizing $\log f(\boldsymbol{y}^*(j)|\boldsymbol{\nu}_k)$. In what follows, $\mathbb{E}_*$ is used to denote the expected value with respect to the distribution of $\boldsymbol{y}^*$.

Cavanaugh and Shumway (1997), in the context of Gaussian state space model selection, considered

$$B_1 = 2\left\{2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) - 2\,\mathbb{E}_*\left[\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k^*)\right]\right\},$$

which is estimated using bootstrap resampling. The corresponding model selection criterion is denoted as EIC1.

Shibata (1997) introduced the EIC2 bootstrap-based criteria. It uses

$$B_2 = 2\,\mathbb{E}_*\left\{2\log f(\boldsymbol{y}^*|\hat{\boldsymbol{\nu}}_k) - 2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k^*)\right\}.$$

As noted above, the bootstrap criteria use estimates of the penalty term obtained through data resampling. For instance, it can be shown that the EIC1 can be written as

$$\text{EIC1} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k) + 2\left\{\frac{1}{N}\sum_{j=1}^{N}\left[-2\log\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k^*(j))}{f(\boldsymbol{y}|\hat{\boldsymbol{\nu}}_k)}\right]\right\}.$$

EIC model selection for Gaussian autoregressive models was considered by Billah, Hyndman, and Koehler (2005). The authors also considered EIC model selection for exponential smoothing. It is worth noticing that bootstrap-based model selection can be carried out in several different ways. For instance, Fenga (2017) uses bootstrap-resampling for Gaussian ARMA model selection by computing a model selection criterion (e.g., AIC) for each fitted model in each bootstrap replication, identifying the model that minimizes the criterion for that bootstrap time series, and then finally selecting the model on the basis of its relative frequency over the all bootstrap samples.

A second model selection approach involves the use of forecasting accuracy measures. The underlying idea is to remove a set of observations from the end of the series, forecast them using different models, and select the model with the best forecasting performance. Specifically, the final s_f data points are removed, different models are fitted using the remaining $n - s_f$ observations, forecasts of the removed observations are produced, and a measure of forecasting accuracy is computed for each candidate model. The selected model is the one that displays the best forecasting performance. Below, we present numerical evidence for such a strategy. Some forecasting accuracy measures that can be used for model selection are (i) the mean absolute prediction error (MAPE), (ii) the root mean square error (RMSE), (iii) the mean directional forecast (MDF), and (iv) the rolling horizon weighted error (RHWE). The first two measures are well known and are routinely used to evaluate forecasting performance.

MDF-based model selection can be performed using a rolling window of n_r observations for parameter estimation and prediction. A sequence of $n-n_r-h$ out-of-sample h-step-ahead forecasts are produced and the corresponding forecasting errors are computed for each window, terminating at observation $T \in \{n_r, \ldots, n-h\}$. The commonly used MDF measures are: (i) mean directional accuracy (MDA) and (ii) mean directional forecast value (MDV):

$$\text{MDA}_h = \frac{1}{n-n_r-h}\sum_{t=n_r}^{n-h}\mathbb{1}(Z_t = 1),$$

$$\text{MDV}_h = \frac{1}{n-n_r-h}\sum_{t=n_r}^{n-h}(-1)^{1-Z_t}|(y_{t+h}-y_t)/y_t|,$$

where $\mathbb{1}(\cdot)$ is the indicator function, $Z_t = \mathbb{1}(W_t = \widehat{W}_t)$ is the directional forecast, $W_t = \mathbb{1}(y_{t+h} - y_t > 0)$ is the realized direction, and $\widehat{W}_t = \mathbb{1}(\hat{y}_t(h) - y_t > 0)$ is the predicted direction, with $\hat{y}_t(h)$ denoting the forecast of y_{t+h} produced at time t. The MDF measure is computed for each candidate model and the selected model is that with the largest MDA or MDV. For further details, see Blaskowitz and Herwartz (2009), Blaskowitz and Herwartz (2011), and Blaskowitz and Herwartz (2014).

The h-step-ahead RHWE measure of forecasting accuracy was proposed in Poler and Mula (2011) for performing model selection: $\text{RHWE}_h = \sum_s \sum_t |e_t^s|^{\pi(\delta_1)}\zeta_h(\delta_1)\lambda(\delta_2)$, where $\delta_1 = t - s$ is the forecast forward, $\delta_2 = n - s$ is the forecast age, $e_t^s = y_t - \hat{y}_s(t-s)$ is the error of the forecast of y_t produced at time s, $\pi(\delta_1) \geq 1$ is the error power according to the forecast forward, $\zeta_h(\delta_1)$ is the multiplicative error factor according to the forecast forward ($\sum_{\delta_1}\zeta_h(\delta_1) = 1$), and $\lambda(\delta_2)$ is the multiplicative error factor according to the forecast age ($\sum_{\delta_2}\lambda(\delta_2) = 1$), $s \in \{n - s_f, \ldots, n-1\}$, and $t \in \{s+1, \ldots, \min\{s+h, n\}\}$.

Model selection strategies for non-dynamic beta regression models were investigated by Bayer and Cribari-Neto (2015, 2017). In the next section, we investigate model identification for dynamic beta models.

4. Simulation study

The finite sample performance of the model selection strategies outlined in the previous section has already been evaluated under different regression and time settings. It is not clear, however, how such criteria perform when used for βARMA model selection. In order to fill that gap, we report the results from extensive Monte Carlo simulations that were carried out to assess the accuracy of different model selection schemes. In what follows, we first focus on model selection based on information criteria. Then, we consider model selection based on out-of-sample forecasting accuracy.

We consider pure autoregressive models, pure moving average models, and models with both dynamics. The sample sizes are $n \in \{50, 150, 250\}$ and $\phi = 120$. We also report results obtained under smaller precision: $\phi = 12$. All simulations were carried out using the R statistical computing environment (versions 4.0.0 and 4.0.4); see R Core Team (2021). The reported results were obtained using 5000 Monte Carlo replications and $N = 250$ bootstrap samples. This number of bootstrap replications is adequate, since data resampling was used to estimate the expected values, and not tail quantities, as in confidence intervals and hypothesis tests; see Efron and Tibshirani (1986, Section 9). Bootstrap resampling was performed parametrically. That is, we generated N bootstrap time series of size n from the fitted βARMA model after replacing the unknown parameters with their conditional maximum likelihood estimates.

Our Monte Carlo simulations are computationally challenging, since they entail a very large number of log-likelihood numerical optimizations. The simulations were run at the National Center of Supercomputing of the Universidade Federal do Rio Grande do Sul using a cluster of computers with 64 blades of processing, 15.97 TFLOPS, and 174-TB RAM that runs the SUSE Linux Enterprise Server operating system. We used parallel computing, and our simulations ran on three nodes with 24 clusters. By using parallel computing, we were able to reduce the execution time by approximately 89%.

Log-likelihood maximization was carried out using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton method with analytic first derivatives; see Nocedal and Wright (2006). Starting values for the parameters were selected as follows: (a) the starting values of all moving average parameters were zero, (b) the starting values of the autoregressive parameters were selected by regressing $g(y_t)$ on $g(y_{t-1}), \ldots, g(y_{t-p})$ using ordinary least squares, and (c) the starting value of ϕ was selected as in the beta regression model; see Ferrari and Cribari-Neto (2004). Beta random number generation was performed based on the Mersenne Twister uniform random generator. The logit link function was used in all data generating processes.

In what follows, we report the percentages of correct model selections achieved by using the following information criteria: AIC, AIC_c, SIC, SIC_c, HQIC, HQIC_c, WIC, and EIC1. We only report results for EIC1 because the results for EIC2 were very similar. For simplicity, we hereafter refer to EIC1 as EIC.

Table 1
Percentages of correct model selection, autoregressive models.

		βAR(1)	βAR(2)	βAR(3)	βAR(4)	βAR(5)	βAR(6)
$n = 50$	AIC	94.54	48.88	41.48	55.28	67.10	37.28
	AICc	95.32	47.10	38.32	50.98	63.69	27.17
	SIC	97.92	37.88	26.34	36.06	51.75	13.21
	SICc	**98.81**	34.02	20.30	30.22	40.31	15.91
	HQIC	96.06	44.62	36.12	46.14	61.98	25.95
	HQICc	97.02	42.14	31.08	43.38	56.34	16.61
	WIC	97.14	41.42	30.76	43.96	55.54	18.20
	EIC	98.31	**68.01**	**87.12**	**63.16**	**76.14**	**65.72**
$n = 150$	AIC	94.98	93.46	93.68	94.76	95.49	94.35
	AICc	95.24	93.80	94.14	94.78	96.21	95.22
	SIC	99.10	94.94	93.74	96.42	98.90	94.22
	SICc	**99.34**	94.64	93.32	96.02	**99.12**	93.36
	HQIC	97.44	95.14	95.00	95.79	97.32	96.30
	HQICc	97.78	95.42	95.16	96.47	97.64	96.41
	WIC	98.42	95.24	94.58	96.81	97.80	95.65
	EIC	98.62	**96.35**	**97.11**	**97.73**	98.53	**97.81**
$n = 250$	AIC	95.28	94.44	94.80	94.25	95.62	93.77
	AICc	95.56	94.76	95.02	94.32	96.05	94.45
	SIC	99.28	98.92	99.12	99.21	99.41	98.72
	SICc	**99.34**	**99.08**	**99.22**	**99.44**	**99.55**	**99.12**
	HQIC	97.72	97.52	97.68	97.22	97.84	97.10
	HQICc	97.76	97.72	97.86	97.84	98.24	97.24
	WIC	98.78	98.54	98.66	98.71	98.18	98.11
	EIC	97.78	97.61	98.7	98.91	99.45	98.23

At the outset, we focus on autoregressive processes. In particular, we consider βAR(p) models, $p \in \{1, \ldots, 6\}$. The parameter values are: βAR(1), $\varphi = 0.3$; βAR(2), $\boldsymbol{\varphi} = (0.2, 0.4)'$; βAR(3), $\boldsymbol{\varphi} = (0.2, -0.3, 0.4)'$; βAR(4), $\boldsymbol{\varphi} = (0.2, -0.5, 0.4, -0.4)'$; βAR(5), $\boldsymbol{\varphi} = (0.35, -0.4, 0.5, -0.45, 0.6)'$; and βAR(6), $\boldsymbol{\varphi} = (0.45, -0.52, 0.65, -0.35, 0.4, -0.5)'$. The percentages of correct model selection are reported in Table 1. For each data generating process and sample size, the best result is in boldface. When the true model was βAR(1), βAR(2) or βAR(3), we fitted autoregressive models up to order six. When data were generated from the βAR(4), βAR(5), and βAR(6) models, we fitted autoregressive models up to orders seven, eight, and nine, respectively.

The results in Table 1 show that the performance of all criteria improves as the sample size increases. When the sample size is small ($n = 50$), the criteria that incorporate finite sample corrections (AIC_c, SIC_c, and HQIC_c) do not typically outperform the corresponding unmodified criteria. All information criteria perform quite well, even when the sample size is very small, when the true data generating process is βAR(1). All correct model selection percentages lie between 94.54% and 98.81%. The best overall performer is the EIC. In some cases, it outperforms the competition by wide margins. For instance, when $n = 50$ and the true model is βAR(3), its rate of correct model selection is 87.12% whereas that of the runner-up is 41.48%. Overall, under autoregressive dynamics, all model selection criteria perform well when $n \geq 150$, their success rates exceeding 93%. When the sample size is small ($n = 50$), model selection only works very well when $p = 1$, i.e., when the true model is βAR(1). The global winner under autoregressive dynamics is the EIC, the bootstrap-based criterion, and the worst performers are the SIC and SIC_c.

Table 2
Percentages of correct model selection, moving average models.

		βMA(1)	βMA(2)	βMA(3)	βMA(4)	βMA(5)	βMA(6)
$n = 50$	AIC	61.62	42.19	39.43	39.11	36.84	18.52
	AICc	65.45	43.32	39.72	39.89	31.32	19.41
	SIC	76.40	47.62	34.92	32.47	32.91	26.15
	SICc	83.65	46.45	39.68	37.29	35.49	23.37
	HQIC	66.96	50.18	37.35	38.82	31.74	28.61
	HQICc	72.57	49.29	36.13	37.66	32.18	26.88
	WIC	71.90	49.12	36.65	36.54	31.93	27.21
	EIC	**97.20**	**65.64**	**62.56**	**61.70**	**60.17**	**61.25**
$n = 150$	AIC	92.39	91.47	87.65	81.61	87.04	87.62
	AICc	92.83	92.23	87.18	82.42	87.14	82.10
	SIC	97.41	93.37	95.57	91.66	89.51	81.75
	SICc	**97.93**	94.11	**96.52**	92.91	88.29	86.54
	HQIC	95.54	93.34	91.23	86.54	88.13	82.21
	HQICc	95.61	93.70	92.45	87.81	89.47	81.10
	WIC	96.20	93.72	94.22	88.19	80.80	87.80
	EIC	97.50	**96.16**	95.84	**92.97**	**91.33**	**88.80**
$n = 250$	AIC	92.32	93.28	90.78	83.78	88.31	89.27
	AICc	92.63	93.58	90.92	84.72	89.23	88.11
	SIC	97.35	97.49	97.52	95.31	89.99	90.49
	SICc	97.35	**98.62**	97.34	95.85	89.05	91.28
	HQIC	96.10	96.31	94.23	89.79	89.81	86.11
	HQICc	96.73	96.43	94.97	90.25	89.74	86.23
	WIC	97.28	97.29	96.21	93.13	88.33	89.81
	EIC	**97.50**	97.88	**97.90**	**95.95**	**90.12**	**93.54**

We now move to moving average processes. We consider models with $q \in \{1, \ldots, 6\}$. The true parameter values are: $\theta = 0.5$ for $q = 1$, $\boldsymbol{\theta} = (0.2, 0.4)'$ for $q = 2$, $\boldsymbol{\theta} = (0.3, 0.2, 0.6)'$ for $q = 3$, $\boldsymbol{\theta} = (0.2, 0.3, -0.4, 0.6)'$ for $q = 4$, $\boldsymbol{\theta} = (0.15, 0.2, 0.3, 0.45, 0.5)'$ for $q = 5$, and $\boldsymbol{\theta} = (0.13, 0.19, 0.25, 0.3, 0.35, 0.5)'$ for $q = 6$. When the true model was βMA(1), βMA(2), or βMA(3), we fitted moving average models up to order six. When data were generated from the βMA(4), βMA(5), and βMA(6) models, we fitted moving average models up to orders seven, eight, and nine, respectively. The percentages of correct model identification are presented in Table 2. The model selection strategies become more accurate as the sample size increases. Their performance deteriorates as q increases when the sample size is small ($n = 50$). For instance, the AIC and SIC rates of correct model identification drop from 61.62% and 76.40% to 18.52% and 26.15%, respectively, as the order of the moving average dynamics increases from one to six. Overall, the best performer is the EIC, the criterion that uses bootstrap resampling. In some cases, it outperforms the competing criteria by wide margins. For example, when $n = 50$ and $q = 1$, its rate of success is nearly 97%, whereas that of the runner-up (AIC) is slightly above 61%. The difference becomes even more dramatic when $q = 6$: 61.25% (EIC) vs. 18.52% (runner-up, AIC).

We now consider models that include both autoregressive and moving average dynamics. We use three data generating processes and two scenarios (parameter values) for each process. The scenarios are indicated by the subscripts a and b next to the model orders. The parameter values of the three models under the two scenarios are: (i) βARMA$(1, 1)_a$, $\varphi = 0.3$, $\theta = 0.4$; βARMA$(1, 1)_b$, $\varphi = 0.3$, $\theta = 0.5$; (ii) βARMA$(1, 2)_a$, $\varphi = 0.35$, $\boldsymbol{\theta} = (0.2, 0.5)'$; βARMA$(1, 2)_b$, $\varphi = 0.35$, $\boldsymbol{\theta} = (0.2, 0.6)'$; (iii) βARMA$(2, 1)_a$, $\boldsymbol{\varphi} = (0.1, 0.5)'$, $\theta = 0.4$; βARMA$(2, 1)_b$, $\boldsymbol{\varphi} = (0.1, 0.6)'$, $\theta = 0.4$. When searching for the best model, we considered all combinations of p and q with each ranging from 0 to 3, except for $(0, 0)$. It should be noted that this is a more challenging situation, since the search for the best fitting model includes pure AR, pure MA, and models with AR and MA components. It is thus expected that larger sample sizes are needed in order to achieve reliable model selection. The rates of correct model identification (expressed as percentages) for the different criteria are presented in Table 3.

Table 3
Percentages of correct model selection, autoregressive moving average models; the subscript next to the model order (a or b) identifies the scenario.

		$(1, 1)_a$	$(1, 1)_b$	$(1, 2)_a$	$(1, 2)_b$	$(2, 1)_a$	$(2, 1)_b$
$n = 50$	AIC	28.31	21.67	23.39	34.88	26.34	29.18
	AICc	28.82	23.59	25.51	35.59	26.99	29.02
	SIC	36.19	30.42	26.18	35.82	22.15	31.51
	SICc	37.45	32.73	26.45	35.96	22.60	33.29
	HQIC	35.24	24.16	23.15	34.18	23.47	32.05
	HQICc	36.52	26.14	23.98	34.77	23.89	33.63
	WIC	32.29	26.20	23.78	34.19	23.90	34.07
	EIC	**47.12**	**51.39**	**43.12**	**51.92**	**42.95**	**59.85**
$n = 150$	AIC	61.46	60.16	64.92	70.81	52.45	64.06
	AICc	61.84	61.04	66.89	72.25	54.50	64.50
	SIC	70.42	68.52	71.12	79.67	53.22	65.78
	SICc	70.91	69.18	72.06	80.92	54.18	65.89
	HQIC	65.69	63.40	63.54	75.42	49.61	65.12
	HQICc	66.09	64.21	64.89	75.88	51.19	65.98
	WIC	67.58	64.50	70.54	79.05	53.58	64.52
	EIC	**78.17**	**78.54**	**76.41**	**82.17**	**80.09**	**80.16**
$n = 250$	AIC	71.92	67.66	78.23	82.20	64.18	81.10
	AICc	73.09	67.89	78.86	83.01	64.89	81.99
	SIC	82.64	83.63	84.12	89.45	65.66	84.42
	SICc	83.47	85.07	85.19	89.81	66.05	85.06
	HQIC	78.52	80.21	83.25	85.42	72.45	83.12
	HQICc	79.69	80.35	83.88	86.07	73.09	84.03
	WIC	82.55	81.98	84.03	86.58	71.20	85.91
	EIC	**91.24**	**90.18**	**92.37**	**91.55**	**89.45**	**91.68**

As with pure autoregressive or pure moving average models, all model selection strategies become more accurate as the sample size increases. Again, the overall winner was the EIC. Indeed, it was the best performer in all simulations, i.e., for all combinations of model order and scenario. In some situations, the EIC outperformed the competition by very wide margins. For instance, when $n = 50$ and $(2, 1)_b$ (i.e., the βARMA$(2, 1)$ process and scenario b), its rate of success was 59.85%, whereas that of the second-best performer (WIC) was only 34.07%. Even when $n = 150$ ($n = 250$), the EIC was far superior in some cases: a success rate of 80.09% vs. 54.50% (89.45% vs. 73.09%) of the runner-up, which was the AIC_c (HC_c) under $(2, 1)_a$.

We computed the average percentage of correct model selection for each criterion and each sample size, and also the global figure; the latter was computed using all three sample sizes. The results are presented in Table 4. It is clear from these figures that the EIC has the best performance in all cases. When $n = 50$ it outperforms all alternative model selection strategies by wide margins; its average rate of correct model identification exceeds 64%, whereas that of the runner-up (AIC) is approximately 42%.

Table 4
Average percentages of correct model selection.

	$n = 50$	$n = 150$	$n = 250$	Global
AIC	41.45	81.57	86.17	69.73
AICc	40.62	81.90	86.61	69.71
SIC	38.66	85.29	91.82	71.92
SICc	38.55	85.84	92.21	72.20
HQIC	40.93	83.15	90.02	71.36
HQICc	40.01	83.73	90.45	71.39
WIC	39.71	84.40	91.40	71.83
EIC	**64.62**	**90.23**	**95.07**	**83.29**

Table 5
Percentages of correct model selection under wider model search.

	βAR(2)		βMA(2)	
	$A = 21$	Δ	$A = 21$	Δ
AIC	44.17	−49.29	43.99	−47.48
AICc	45.42	−48.38	45.99	−46.24
SIC	54.45	−40.49	58.91	−34.46
SICc	53.22	−41.42	59.63	−34.48
HQ	57.59	−37.55	54.05	−39.29
HQc	61.23	−34.14	55.04	−38.66
WIC	60.35	−34.92	57.52	−36.20
EIC	**78.97**	−17.38	**80.35**	−15.81

The relative advantage of the EIC over the competing criteria decreases as the sample size increases. For instance, when $n = 250$, the rates of success are above 95% for the EIC and over 92% for the second-best performer (SIC_c). The EIC overall frequency of correct model selection (nearly 83%) considerably exceeds those of all other criteria, the runner-up being the SIC_c (nearly 73%).

In Table 1 (2) we present the percentages of correct model specification for AR (MA) models obtained by only searching over AR (MA) models. The number of candidate models considered in the best fitting model search (A) ranged from six to nine depending on the order of the true model. We now consider the more challenging case in which the true data generating process is AR or MA and the model search is performed over AR, MA, and ARMA models. We consider the following true models: (i) βAR(2) with $\varphi = (0.2, 0.4)$ and (ii) βMA(2) with $\theta = (0.2, 0.4)$. In each case, $A = 21$ candidate models are fitted, namely, AR models with p ranging from 1 to 6, MA models with q ranging from 1 to 6, and ARMA models with p and q ranging from 1 to 3. The sample size is $n = 150$. The percentages of correct model specification are presented in Table 5. We also report the differences between the results obtained under a pure AR or MA model search and a wider model search (Δ). The following conclusions can be drawn from the figures in Table 5. First, the success rates of all model selection criteria are now smaller ($\Delta < 0$). This was expected, since more candidate models are considered in the search for the best fitting model. Second, the EIC is the best performer in both cases, i.e., under AR and MA data generating mechanisms. Third, under both dynamics, the EIC is the criterion with the smallest success rate reductions. For instance, under AR (MA) dynamics, its percentage of correct model determination dropped 17.38% (15.81%), whereas the corresponding figures for the alternative criteria ranged from 34.14% to 49.29% (34.46% to 47.48%). The AIC and the AICc are the criteria most impacted by the increase in the number of candidate models. Fifth, the performance ranks of the different criteria are the same as before.

The simulation results presented above were obtained using $\phi = 120$, which is the same precision value used in Scher et al. (2020). We now investigate the impact of a smaller precision on the rates of correct model specification. To that end, we consider $n = 150$ and three data generating processes: βAR(2), βMA(1), and βARMA$(1, 1)_b$. The percentages of correct model identification are presented in Table 6. For ease of comparison, the table also contains the differences in the figures obtained under the two scenarios (Δ). The results reported in Table 6 lead to several interesting conclusions. First, all model selection strategies become less accurate under small precision ($\Delta < 0$ in all cases). Second, βARMA model selection accuracies are more impacted than those of the βAR and βMA processes. Third, the EIC is the best performer under all data generating processes with $\phi = 12$. Fourth, the EIC went from runner-up to best performer under moving average dynamics when the precision parameter values were reduced. Fifth, the EIC displays the smallest losses in accuracy.

The numerical evidence presented above shows that all model selection criteria yield more accurate model identification as the sample size increases. More importantly, it reveals that there is much to be gained by resorting to bootstrap resampling when searching for the best model in the wider class of βARMA(p, q) models or when attention is restricted to the βAR(p) or βMA(q) processes. When the EIC, the bootstrap-based criterion, was not the best performer, it was a very close runner-up. By contrast, in several situations, the EIC not only performed the best but also did so by wide margins.

We now move to model selection based on out-of-sample forecasting accuracy. We use $n = 150$ and $h \in \{3, 6\}$. Model selection is based on MAPE, RMSE, MDF, and RHWE. Two out-of-sample MDF measures are also considered: MDA and MDV. We use a rolling window of 100 observations for parameter estimation and prediction. A sequence of $50 - h$ out-of-sample h-step-ahead forecasts are produced and the corresponding forecasting errors are computed for each window terminating at observation $T \in \{100, \ldots, 150 - h\}$. When computing the RHWE measure, we set $\pi(\delta_1) = 1$ and $s_f = 10$, the corresponding weights being $\zeta_3(\delta_1) = \{0.5, 0.33, 0.17\}$, $\zeta_6(\delta_1) = \{0.29, 0.24, 0.20, 0.14, 0.09, 0.04\}$, and $\lambda(\delta_2) = \{0.02, 0.04, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.16, 0.18\}$.

In what follows, we consider two scenarios when selecting a model based on an out-of-sample forecasting accuracy measure: (i) the model that displays the highest overall accuracy is selected (scenario 1), and (ii) the model that displays the highest accuracy among the models that pass a diagnostic test is selected (scenario 2). The motivation for carrying out a diagnostic test prior to forecasting is to exclusively consider models for which there is no evidence of model misspecification. We use the Q_4 test portmanteau diagnostic test proposed by Scher et al. (2020), which performs well when used with fitted βARMA models. The lag truncation parameter is $m =$

Table 6
Percentages of correct model selection in a small precision scenario ($\phi = 12$).

	βAR(2)		βMA(1)		βARMA(1, 1)$_b$	
	$\phi = 12$	Δ	$\phi = 12$	Δ	$\phi = 12$	Δ
AIC	81.92	−11.54	78.04	−14.35	20.42	−39.74
AICc	83.74	−10.06	79.24	−13.59	21.15	−39.89
SIC	93.62	−1.32	86.05	−11.36	24.88	−43.64
SICc	91.14	−3.50	86.78	−11.15	24.91	−44.27
HQ	91.28	−3.86	79.62	−15.92	22.36	−41.04
HQc	92.44	−2.98	85.04	−10.57	23.17	−41.04
WIC	93.06	−2.18	86.42	−9.78	23.51	−40.99
EIC	**95.32**	−1.03	**89.86**	−7.64	**53.52**	−25.02

13 and all testing inferences are carried out at the 5% significance level.

Table 7 contains the percentages of correct model selection obtained using the aforementioned out-of-sample forecasting accuracy criteria. For each criterion, the top and bottom rows correspond, respectively, to scenarios 1 and 2. The data generating processes are βAR(2), βMA(2), and βARMA(1, 1)$_b$. The true parameter values are as before, and $\phi = 120$. The results obtained with MDA and MDV were very similar, and for brevity we only present those relative to MDV. The results in Table 7 lead to interesting conclusions. First, model selections based on MAPE and RMSE are the least accurate strategies, with rates of correct model selection that range from 17.26% to 23.30%. Second, RHWE model selection is slightly more successful than that based on the previous criteria. Third, MDV model selection is considerably more accurate than all alternative forecasting-based strategies, with the corresponding rates of correct model identification fluctuating between 48.01% and 58% (approximately). In some cases, MDV model selection is more than twice as accurate as that based on the runner-up criterion. For instance, under βMA(2) dynamics and $h = 3$, the MDV rate of correct model selection is 56.38%, whereas that of the second-best performer (RHWE) is 28.16%. Fourth, the results obtained with $h \in \{1, 3, 6\}$ are similar, except for MDV in the βAR(2) model, where the rate of correct model selection is clearly smaller for $h = 1$. Fifth, model selection based on information criteria is considerably more accurate than that performed on the basis of measures of forecasting accuracy, especially under pure AR and pure MA dynamics. Interestingly, MDV model selection is slightly less accurate than that based on the AIC (the worst performing model selection criterion) under βARMA(1, 1) dynamics (58.06% with $h = 6$ vs. 60.16%). By contrast, under βAR(2) and βMA(2) dynamics, MDV model selection ($h = 6$) is considerably less accurate than that based on the AIC: 55.62% vs. 93.46% and 56.02% vs. 91.47%, respectively. Sixth, RHWE and MDV model selection always benefit from a prior screening based on the Q_4 diagnostic test, especially the former. The largest increase in the rate of successful model selection that follows from the diagnostic test screening is 2.93% (RWME, βMA(2), $h = 6$).

5. Forecasting stored hydroelectric energy

In what follows, we present and discuss an empirical analysis. Hydroelectricity is a renewable energy source, and it is widely used in Brazil. There are two types of reservoirs: accumulation reservoirs and water line reservoirs. The former are usually located at the headwaters of rivers, in places with high waterfalls, since their large size allows for the accumulation of substantial amounts of water that function as stock to be used in periods of drought. They also allow hydroelectric power plants to respond rapidly to fluctuations in the demand for electricity. It is also worth noticing that hydroelectric power (hydro) is environmentally friendly, since the hydroelectric lifecycle produces very small amounts of greenhouse gases, and hydro plants do not release pollutants into the air. Climate change, nonetheless, has added uncertainty to hydropower generation. Changing rainfall patterns and prolonged droughts have made it increasingly difficult to assess future river flows. As a result, the use of accumulation reservoirs has become crucial for hydropower generation. Stored energy is the energy value of the accumulated water, that is, how much energy (in megawatts each month) can be generated from the stored volume of water, expressed as a proportion of the total capacity. Stored energy forecasting is important for companies in charge of energy distribution. Our interest lies in modeling the proportion of stored hydroelectric energy (ONS, 2020) in South Brazil and producing out-of-sample forecasts. We also evaluate the impact of model selection on the accuracy of the out-of-sample forecasts.

The data are monthly averages from July 2000 to April 2018, thus spanning 214 months. The final six observations were removed from the data to be used for forecast evaluation. Hence, the effective sample size is $n = 208$, with the time series going up to October 2017. A shorter range of this time series was recently modeled by Scher et al. (2020). They used the AIC and focused on diagnostic (portmanteau) testing. In what follows, we use a longer time series and consider subsets of the data, namely, the first $n = 75$, $n = 150$, and $n = 208$ (complete data) observations. Some descriptive statistics are presented in Table 8: minimal value, maximal value, median, mean, variance, coefficient of asymmetry, and coefficient of excess kurtosis. There is negative asymmetry and negative excess kurtosis. The mean level of stored energy is 0.7016 and the maximal level is close to one (0.9862). The time series data plot (top panel), correlogram (bottom-left panel), and partial correlogram (bottom-right panel) can be found in Fig. 1. The sample autocorrelations do not decay slowly towards zero, and hence there is no indication of long memory behavior. Also, the sample partial autocorrelations show no evidence of seasonal fluctuations.

Table 7
Percentages of correct model selection based on out-of-sample forecasting model selection criteria; top and bottom rows are for scenarios 1 and 2, respectively.

Criterion	βAR(2)			βMA(2)			βARMA(1, 1)$_b$		
	$h=1$	$h=3$	$h=6$	$h=1$	$h=3$	$h=6$	$h=1$	$h=3$	$h=6$
MAPE	17.26	17.54	17.46	19.54	22.06	21.88	20.22	20.80	21.50
	18.20	19.09	17.56	20.19	21.95	22.90	20.18	22.74	21.95
RMSE	17.26	18.02	16.88	19.54	22.50	23.30	20.22	21.74	22.91
	18.20	18.88	17.62	20.19	24.23	24.68	20.18	24.21	24.24
RHWE	25.32	25.20	24.12	28.80	28.16	27.44	29.35	30.04	30.84
	27.74	26.33	26.25	29.00	30.00	30.37	31.75	32.16	31.61
MDV	48.01	56.58	55.62	56.18	56.38	56.02	57.14	57.05	58.03
	48.10	57.16	56.40	56.65	56.61	56.64	57.89	57.68	59.01

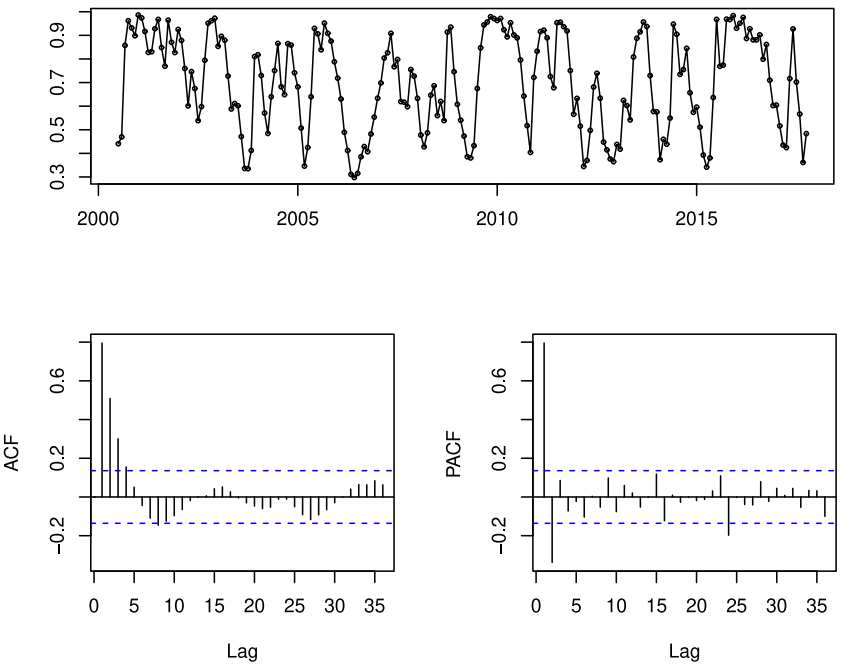

Fig. 1. Stored hydroelectric energy in South Brazil: time series data (top panel), correlogram (bottom-left panel), and partial correlogram (bottom-right panel).

Table 8
Descriptive statistics, stored hydroelectric energy in South Brazil, $n = 214$.

Min	Max	Median	Mean	Variance	Asymmetry	Kurtosis
0.2977	0.9862	0.7265	0.7016	0.0404	−0.2714	−1.2180

We note that the data contain several observations that are close to the upper standard unit interval limit.

As noted above, a novel feature of βARMA models is that they will never yield out-of-sample forecasts that lie outside $(0, 1)$. Such improper forecasts may be obtained, however, when using Gaussian ARMA models or an exponential smoothing algorithm. To illustrate this, we computed the first six out-of-sample forecasts from Gaussian ARMA models identified using the AIC and from the Holt algorithm for all subsamples of our time series with at least 24 observations (i.e., $n \geq 24$). In 12 such subsamples, there was at least one forecast that exceeded one.

We searched for the best fitting βARMA model using different model selection criteria. Our main interest lies in selecting a model to be used for out-of-sample forecasting. In practice, forecasts are only produced based on models that display good data fit, in particular, those based on models that pass diagnostic testing. Hence, the selected models were subjected to portmanteau diagnostic testing based on the Q_4 test statistic proposed by Scher et al. (2020). The lag truncation parameter value used in the test statistic was $m = \lceil \sqrt{n} \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. That is, $m \in \{9, 13, 14\}$ for $n \in \{75, 150, 208\}$. We tested the null hypothesis that the first m residual autocorrelations equal zero. Hence, under the null hypothesis, the data dynamics are fully captured by the fitted model. When the null hypothesis is rejected, by contrast, the residuals are serially correlated and thus there is evidence of model misspecification. Models for which the correct model specification is rejected at the 5% significance level by the portmanteau test are discarded. When that happens, the next best fitting model according to the model selection criterion is selected.

Even though we do not present results for the models that were discarded by the diagnostic analysis, we note that the out-of-sample forecasts produced by such models were typically less accurate than those obtained from the models we use in the empirical analysis that follows.

We consider three different data ranges. By Samples I and II, we mean that the time series consists of the first 75 and 150 observations, respectively. Sample III refers to the complete sample (i.e., all 208 data points). When the time series only includes the first 75 observations ($n = 75$, Sample I), the EIC selects the βAR(3) model and the βARMA(1, 1) model is chosen by all other criteria. Sample II includes the first 150 data points ($n = 150$). The model selected by the EIC is βAR(3). The βARMA(2, 1) model is chosen by all the other criteria. The final scenario we consider is Sample III, in which all 208 observations are used ($n = 208$). The EIC identifies the βARMA(2, 3) model. A different model is selected by the remaining criteria, namely, βARMA(1, 1).

The point estimates of α and ϕ for the models selected by the EIC (other criteria) in Samples I, II, and II are, respectively, (i) 0.2380 and 12.0681 (0.4311 and 10.4292), (ii) 0.2366 and 14.0793 (0.3727 and 13.2730), and (iii) 0.1526 and 11.7536 (0.3739 and 11.1327). The point estimates (standard errors in parentheses) of the AR and MA parameters are presented in Table 9. For each sample, the top model was selected by the EIC and the bottom model was identified using the alternative criteria.

Our interest lies in forecasting y_{n+h}, for $h \geq 1$. Forecasting accuracy is assessed using MAPEs. For $h \in \{1, \ldots, 6\}$, $\text{MAPE}(h) = h^{-1} \sum_{j=1}^{h} |y_{n+j} - \hat{y}_n(j)|$, where $\hat{y}_n(j)$ denotes the forecast of y_{n+j} made at time n. The MAPEs obtained using the βARMA models identified by the different model selection criteria are presented in Table 10. The smallest MAPE for each value of h in each sample in Table 10 is displayed in boldface. The figures in the table lead to interesting conclusions. At the outset, consider the smallest sample size ($n = 75$, Sample I). The βARMA identified by the EIC outperformed the corresponding model chosen by all other criteria at all forecasting horizons, and by wide margins. For instance, when $h \in \{2, 3\}$, the MAPEs of the forecasts made using the former were approximately 34% and 38% smaller than those obtained using the latter. When Sample II was used ($n = 150$), the dynamic beta model selected by the EIC (βAR(3)) outperformed that identified by all remaining criteria (βARMA(2, 1)) for $h \in \{1, 2, 3, 4, 6\}$. That is, it only fared worse for $h = 5$ and by a narrow margin (less than 4%). In some cases, the forecasts obtained from the βAR(3) model were considerably more precise than those from the competing beta model; e.g., the MAPE was nearly 20% smaller for $h = 1$. Next, we consider the situation in which the models were fitted using all data points ($n = 208$, Sample III). Here, the βARMA(2, 3) model identified by the EIC yielded the most accurate forecasts for $h \in \{1, 2, 3, 4, 5, 6\}$.

In this section, we evaluate the forecasting accuracy of βARMA models identified using different model selection criteria. Overall, the EIC is the winner. In most cases, this criterion was able to select the best performing βARMA model. Its use yielded considerable gains in forecasting accuracy in some situations. For instance, when the smallest sample size was used (Sample I, $n = 75$), the MAPEs of the βAR(3) model selected by the EIC for $h \in \{1, \ldots, 6\}$ were approximately 24%, 30%, 34%, 38%, 28%, and 18% smaller, respectively, than those obtained with the βARMA(1, 1) model that was selected by all other criteria.

It is worth noticing that in each sample (Samples I, II, and III), different βARMA models were identified by (i) the EIC and (ii) all other information criteria. That is, the EIC identified a model different from that selected by all other criteria. Recall that all identified models were sequentially subjected to portmanteau diagnostic testing. The selected model according to each criterion is the first model in the ordered (best to worst) list of models that passes diagnostic testing. It was only after diagnostic testing that one βARMA model was selected by the EIC and a different model was selected by the remaining information criteria.

We also carried out model selection on the basis of the out-of-sample model selection criteria described in Section 3. The rolling windows of observations in MDF (n_r) are 50, 100, and 150 for Samples I, II, and III, respectively, and the forecasting horizons are $h \in \{1, 3, 6\}$. As before, the Q_4 portmanteau diagnostic test was performed prior to forecasting. The RHWE was computed using the same parameters as in Section 4. For all three sample sizes, the RHWE and MDF measures selected the same model identified by the EIC. This result was obtained using both MDA and MDV. The MAPE and RMSE measures selected different models from the ones identified on the basis of the EIC and all other information criteria. The forecasts from such models were uniformly less accurate than those obtained from the model selected by the EIC (and RHWE and MDF).

As a final exercise, we investigated the sensitivity of the different model selection strategies to the presence of outliers in the data. To that end, we introduced outliers into the complete time series (Sample III, $n = 208$). At the outset, we introduced a single outlier into the data as follows: (i) we multiplied $y_{52} = 0.8649$ by $a \in \{0.75, 0.50, 0.25\}$, (ii) we multiplied $y_{104} = 0.5407$ by $a \in \{1.75, 1.50, 1.25, 0.75, 0.50, 0.25\}$, and (iii) we multiplied $y_{156} = 0.8649$ by $a \in \{0.75, 0.50, 0.25\}$. We chose to modify the values of cases 52, 104, and 156, transforming them into atypical data points, because they are located at 25%, 50%, and 75% of the time series length. Subsequently, we introduced three outliers into the data by replacing y_{52} and y_{156} with $0.25 \times y_{52}$ and $0.25 \times y_{156}$, respectively, and y_{104} with (i) $1.75 \times y_{104}$ and (ii) $0.25 \times y_{104}$. Model selection was not impacted by such outliers: in all cases, the EIC selected the βARMA(2, 3) model, and all other criteria selected the βARMA(1, 1) model, as with the unperturbed data. By contrast, the presence of outliers in the data noticeably impacted the short-term forecasts ($h \in \{1, 2\}$) and overall had little impact on the forecasts when $h \in \{3, \ldots, 6\}$. In particular, all one-step-ahead forecasts became less accurate when the time series included one or three outliers. In future research we shall further investigate the impact of model misspecification and data anomalies on βARMA model selection and forecasting.

Table 9
Point estimates (standard errors in parentheses).

Sample	Model	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
I	βAR(3)	0.8825	−0.3783	0.1767	–	–	–
		(0.0444)	(0.0721)	(0.0592)	–	–	–
	βARMA(1, 1)	0.4040	–	–	0.4317	–	–
		(0.0828)	–	–	(0.0957)	–	–
II	βAR(3)	0.9490	−0.3385	0.0871	–	–	–
		(0.0326)	(0.0511)	(0.0421)	–	–	–
	βARMA(2, 1)	0.2501	0.2250	–	0.6485	–	–
		(0.1227)	(0.1035)	–	(0.1160)	–	–
III	βARMA(2, 3)	0.2703	−0.4743	–	0.4778	−0.1512	0.1564
		(0.0812)	(0.0686)	–	(0.0920)	(0.0592)	(0.0609)
	βARMA(1, 1)	0.4944	–	–	0.2562	–	–
		(0.0484)	–	–	(0.0612)	–	–

Table 10
Mean absolute prediction errors.

Sample	Model	MAPE					
		$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
I	βAR(3)	**0.0990**	**0.0939**	**0.0801**	**0.0621**	**0.0611**	**0.0761**
	βARMA(1, 1)	0.1296	0.1339	0.1222	0.0995	0.0853	0.0930
II	βAR(3)	**0.0254**	**0.0836**	**0.0591**	**0.0542**	0.0666	**0.0788**
	βARMA(2, 1)	0.0316	0.0841	0.0647	0.0547	**0.0642**	0.0792
III	βARMA(2, 3)	**0.0064**	**0.0228**	**0.0738**	**0.0734**	**0.0604**	**0.0585**
	βARMA(1, 1)	0.0088	0.0369	0.0775	0.0748	0.0621	0.0586

6. Concluding remarks

The βARMA model is a dynamic model introduced by Rocha and Cribari-Neto (2009, 2017). It is tailored for use with time series that assume values in the standard unit interval, such as rates, proportions, and concentration indices. Parameter estimation is performed by conditional maximum likelihood. Diagnostic checking based on portmanteau tests in βARMA models was developed by Scher et al. (2020). Thus, model identification prior to out-of-sample forecasting in that class of models warranted investigation. That was our chief goal in this paper.

We considered βARMA model selection based on different information criteria. Since such criteria were not developed for dynamic models tailored to double bounded time series, it was important to investigate their usefulness in that context. We performed extensive and computer-intensive simulations to estimate the rates of correct model identification of several criteria for different sample sizes and by separately considering (i) autoregressive, (ii) moving average, and (iii) autoregressive and moving average dynamics. The numerical evidence we reported showed that all criteria yield more accurate model identification as the sample size increases. More importantly, our results showed that model selection can be made substantially more accurate in samples of small to moderate sizes by using bootstrap resampling. In some cases, the frequency of correct model identification was more than double that achieved by using criteria that are not resampling-based. We also considered model selection based on measures of forecasting accuracy. Our results showed that a measure based on directional forecasts leads to model selection that is more accurate than that obtained using alternative measures. Also, model selection guided by information criteria is more reliable than that guided by forecasting accuracy measures.

We also presented and discussed an empirical application. Our goal was to model and forecast the future behavior of the share of stored hydroelectric energy in South Brazil. We used different samples that corresponded to different sample sizes (75, 150, and 208 observations). Interestingly, in all cases the bootstrap model selection criterion identified a model that was different from that selected on the basis of all other criteria. Model selection based on directional forecasting accuracy agreed with that performed using bootstrap resampling. In nearly all situations, the forecasts obtained with the models selected with the aid of the bootstrap-based information criterion were more accurate than those yielded by the models identified by the competing information criteria, in some cases by a wide margin (e.g., over one-third).

Finally, a word of caution is in order. The simulation evidence we presented offers insight on the ability of different model selection strategies to recover the true model, and on their reliability. In our empirical analysis, by contrast, we focused on out-of-sample forecasting. Here, selecting the true model might be less important than assessing the quality of the forecasts. Interestingly, in both settings (simulations and real data analysis) the best results were obtained using the same model selection strategy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge financial support from Fundação de Apoio à Ciência e Tecnologia do Estado de Pernambuco (FACEPE), Brazil, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, and Centro Nacional de Supercomputação da Universidade Federal do Rio Grande do Sul (CESUP/UFRGS), Brazil. We thank an associate editor and three anonymous referees for comments and suggestions that led to a much-improved manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bayer, F. M., Cintra, R. J., & Cribari-Neto, F. (2018). Beta seasonal autoregressive moving average models. *Journal of Statistical Computation and Simulation*, *88*(15), 2961–2981.

Bayer, F. M., & Cribari-Neto, F. (2015). Bootstrap-based model selection criteria for beta regressions. *Test*, *24*(4), 776–795.

Bayer, F. M., & Cribari-Neto, F. (2017). Model selection criteria in beta regression with varying dispersion. *Communications in Statistics. Simulation and Computation*, *46*(4), 729–746.

Benjamin, M. A., Rigby, R. A., & Stasinopoulos, M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association*, *98*(461), 214–223.

Billah, B., Hyndman, R. J., & Koehler, A. B. (2005). Empirical information criteria for time series forecasting. *Journal of Statistical Computation and Simulation*, *75*(10), 831–840.

Blaskowitz, O., & Herwartz, H. (2009). Adaptive forecasting of the EURIBOR swap term structure. *Journal of Forecasting*, *28*(7), 575–594.

Blaskowitz, O., & Herwartz, H. (2011). On economic evaluation of directional forecasts. *International of Journal of Forecasting*, *27*(4), 1058–1065.

Blaskowitz, O., & Herwartz, H. (2014). Testing the value of directional forecasts in the presence of serial correlation. *International of Journal of Forecasting*, *30*(1), 30–42.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Hoboken: Wiley.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research*, *33*(2), 261–304.

Casarin, R., Valle, L. D., & Leisen, F. (2012). Bayesian model selection for beta autoregressive processes. *Bayesian Analysis*, *7*(2), 385–410.

Cavanaugh, J. E., & Shumway, R. H. (1997). A bootstrap variant of AIC for state-space model selection. *Statistica Sinica*, *7*(2), 473–496.

Cribari-Neto, F., Frery, A. C., & Silva, M. F. (2002). Improved estimation of clutter properties in speckled imagery. *Computational Statistics & Data Analysis*, *40*(4), 801–824.

Efron, B., & Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intrevals and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–77.

Fenga, L. (2017). Bootstrap order determination for ARMA models: A comparison between different model selection criteria. *Journal of Probability and Statistics*, *2017*, Article 1235979.

Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*, *41*(2), 190–195.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

McQuarrie, A. D. (1999). A small-sample correction for the Schwarz SIC model selection criterion. *Statistics & Probability Letters*, *44*(1), 79–86.

McQuarrie, A. D. R., & Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. London: World Scientific.

Melchior, C., Zanini, R. R., Guerra, R. R., & Rockenbach, D. A. (2021). Forecasting Brazilian mortality rates due to occupational accidents using autoregressive moving average approaches. *International Journal of Forecasting*, *37*(2), 825–837.

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). New York: Springer.

ONS (2020). Operador Nacional do Sistema Elétrico – Energia Armazenada. http://www.ons.org.br/historico/energia_armazenada.aspx.

Ospina, R., Cribari-Neto, F., & Vasconcellos, K. L. P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics & Data Analysis*, *51*(2), 960–981, Erratum: 55, 2011, 2445.

Ozaki, T. (1978). On the order determination of ARIMA models. *Applied Statistics*, *26*(3), 290–301.

Palm, B., & Bayer, F. M. (2018). Bootstrap-based inferential improvements in beta autoregressive moving average model. *Communications in Statistics. Simulation and Computation*, *47*(4), 977–996.

Poler, R., & Mula, J. (2011). Forecasting model selection through out-of-sample rolling horizon weighted error. *Expert Systems with Applications*, *38*(12), 14778–14785.

Pumi, G., Valk, M., Bisognin, C., Bayer, F. M., & Prass, T. S. (2019). Beta autoregressive fractionally integrated models. *Journal of Statistical Planning and Inference*, *200*, 196–212.

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Rocha, A. V., & Cribari-Neto, F. (2009). Beta autoregressive moving average models. *Test*, *18*(3), 529–545.

Rocha, A. V., & Cribari-Neto, F. (2017). Erratum to: Beta autoregressive moving average models. *Test*, *26*(2), 451–459.

Scher, V. T., Cribari-Neto, F., Pumi, G., & Bayer, F. M. (2020). Goodness-of-fit tests for βARMA hydrological time series modeling. *Environmetrics*, *31*(3), Article e2607.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*(1), 117–126.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, *7*(2), 375–394.

da Silva, C. Q., Migon, H. S., & Correia, L. T. (2011). Dynamic Bayesian beta models. *Computational Statistics & Data Analysis*, *55*(6), 2074–2089.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics. Theory and Methods*, *7*(1), 13–26.

Wu, T.-J., & Sepulveda, A. (1998). The weighted average information criterion for order selection in time series and regression models. *Statistics & Probability Letters*, *39*(1), 1–10.