

## RESEARCH ARTICLE OPEN ACCESS

# A Matsuoka-Based GARMA Model for Environmental and Energy Systems: Theory, Estimation, and Applications

Guilherme Pumi<sup>1</sup>  | Danilo Hiroshi Matsuoka<sup>1</sup>  | Taiane Schaedler Prass<sup>1</sup>  | Bruna Gregory Palm<sup>2</sup> 

<sup>1</sup>Mathematics and Statistics Institute, Programa de Pós-Graduação em Estatística, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil |

<sup>2</sup>Department of Mathematics and Natural Sciences, Blekinge Institute of Technology, Karlskrona, Sweden

**Correspondence:** Guilherme Pumi ([guilherme.pumi@ufrgs.br](mailto:guilherme.pumi@ufrgs.br))

**Received:** 23 July 2025 | **Revised:** 9 February 2026 | **Accepted:** 1 April 2026

**Keywords:** non-Gaussian time series | partial maximum likelihood | regression models | time series analysis

## ABSTRACT

We propose a new time series model for continuous data supported on the open unit interval  $(0, 1)$ , motivated by applications in environmental and energy systems. The Matsuoka autoregressive moving average (MARMA) model combines the Matsuoka distribution—a uniparametric member of the canonical exponential family—as the conditional distribution with a flexible ARMA-type structure for the conditional mean. Parameters are estimated via partial maximum likelihood, allowing for random, time-dependent covariates and enabling standard asymptotic inference. To construct out-of-sample prediction intervals, we explore a bootstrap-based procedure that captures the uncertainty in the dynamic structure. A simulation study evaluates the finite-sample performance of the method. The model is applied to the monthly proportion of electricity generated in the United States from all sources, except conventional hydropower. This application highlights the model's utility in capturing serial dependence, ensuring predictions remain within bounds, and providing reliable forecast intervals—key features for robust energy system planning and environmental policy analysis.

**MSC2020 Classification:** 62M10, 62F12, 62E20, 62J12, 62J99

## 1 | Introduction

Most data appearing in natural sciences, including hydrology, climatology, and other environmental applications, consist of observations that are serially dependent over time. Capturing such temporal dependence calls for dynamic modeling frameworks that explicitly describe the stochastic evolution of the series. However, in many applications, the observed series are not only serially dependent, but also non-Gaussian and naturally bounded. This is particularly the case for variables supported on the open unit interval  $(0, 1)$ , including relative humidity, incidence rates, reservoir storage proportions, and energy shares. When models designed for real-valued data are applied to bounded variables, forecasts may fall outside the range of plausible values, leading to interpretational and practical difficulties

(Grande et al. 2022). For such data, classical Gaussian time series models, such as autoregressive moving average (ARMA) models, are often too restrictive. These limitations have motivated the development of non-Gaussian dynamic models specifically designed for time series supported on the open unit interval.

In environmental and hydrological applications, bounded time series often show behavior close to the limits of the unit interval. Variables such as reservoir storage proportions and energy or water supply shares may remain close to zero or close to one for long periods. In addition, their variability is usually not constant across the data, and often decreases as the process approaches physical or operational limits. This behavior reflects practical constraints of environmental systems, where rapid changes away from very high or very low levels are uncommon. For this reason,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Environmetrics* published by John Wiley & Sons Ltd.

it is useful to consider dynamic models that can capture both persistence near the boundaries and changes in variability across the support, in particular models that allow for asymmetric dispersion and reduced variability close to zero and one.

The motivation for this paper lies in a hydrological empirical problem, which involves modeling and forecasting the monthly proportion of electricity generated in the United States from all sources except conventional hydropower. Accurately modeling this proportion is essential for understanding the country's energy transition, as changes in the relative contribution of non-hydropower sources reflect broader trends in renewable integration and fossil fuel reliance. Such modeling, together with reliable forecasts and appropriate measures of uncertainty, is central to informed energy policy decisions and infrastructure planning. In particular, obtaining forecast prediction intervals provides valuable information for quantifying uncertainty in the presence of variable generation patterns and evolving demand.

One approach capable of handling double-bounded time series that has gained attention in the literature over the last decade is the so-called GARMA (generalized ARMA) models discussed in Benjamin et al. (2003). In a few words, GARMA modeling merges the strengths of ARMA modeling into a generalized linear model (GLM) framework, yielding a class of very flexible models that can be easily tailored to accommodate a wide variety of structures, including non-Gaussianity, bounds, asymmetries, and so forth. GARMA models are classified as observation-driven (Cox et al. 1981) and are defined by two components: the random component, which specifies the probability structure (conditional distributions) for the model, and the systematic component, which is responsible for the dependence structure present in the model.

In practice, original GARMA models only considered random components belonging to the canonical exponential family and an ARMA-like systematic component to model the conditional mean. However, as the literature on GARMA models for continuous double-bounded time series has grown over the years, so has the scope of these models. The interest nowadays lies in models for which the systematic component follows the usual approach of GLM with an additional dynamic term of the form

$$g(\mu_t) = \eta_t = \mathbf{X}_t' \boldsymbol{\beta} + \tau_t,$$

where  $g$  is a suitable link function,  $\mu_t$  is some quantity of interest related to the conditional distribution (usually the mean or median),  $\mathbf{X}_t$  denotes a vector of (possibly random and time dependent) covariates observed at time  $t$  with associated vector of coefficient  $\boldsymbol{\beta}$  and  $\tau_t$  is a term responsible for accommodating any serial dependence in  $\mu_t$ . The terms  $\mu_t$  and  $\tau_t$  vary according to the scope of the model and the intended application. For instance, the classical ARMA form and its variants are used in Rocha and Cribari-Neto (2009); Maior and Cysneiros (2018); Prass et al. (2025) and Bayer et al. (2017, 2018), while Pumi et al. (2019) and Benaduce and Pumi (2023) apply a long-range-dependent ARFIMA specification. More exotic nonlinear specifications can also be found, as in the case of the Beta autoregressive chaotic models of Pumi et al. (2021).

Another important feature in GARMA models is the nature of  $\mu_t$ . Originally,  $\mu_t$  denoted the model's conditional mean at time  $t$ ,

as in Beta-based models such as Rocha and Cribari-Neto (2009); Bayer et al. (2018); Pumi et al. (2019) and Pumi et al. (2021) or as in models for positive time series (Prass et al. 2025). But other possibilities were considered in the literature. For instance, for Kumaraswamy-based models (the KARMA model of Bayer et al. 2017),  $\mu_t$  denotes the model's median. For Unit-Weibull-based models (the UWARMA of Pumi et al. 2024), it represents any quantile, while for models for which the conditional distribution is a member of a symmetric family of distributions without a finite first moment,  $\mu_t$  denotes the model's point of symmetry (Benaduce and Pumi 2023).

The Matsuoka distribution is a uniparametric member of the canonical exponential family, supported on the interval  $(0, 1)$ . It was introduced and studied in Matsuoka et al. (2024), primarily motivated by applications in production frontier modeling. In this work, we propose a new GARMA model, called the Matsuoka autoregressive moving average (MARMA) model, in which the random component follows the Matsuoka distribution, while the systematic component models the conditional mean through an ARMA-like structure that may include exogenous covariates.

A natural question that arises when introducing a new model is: Why do we need yet another GARMA model for  $(0, 1)$ -bounded time series, particularly one based on a uniparametric distribution, given the existence of more highly parameterized alternatives? Several reasons justify this development. First, in terms of novelty, as far as we are concerned, the MARMA model is the first GARMA model for continuous  $(0, 1)$ -valued time series whose underlying distribution belongs, without restriction, to the canonical exponential family. Second, regarding tractability, employing a distribution from the canonical exponential family as the random component greatly simplifies key derivations, such as those for the score function and the information matrix. Third, from a theoretical point of view, models within this family benefit from a well-established asymptotic framework for partial maximum likelihood inference (Fokianos and Kedem 2004). Fourth, despite its simplicity, the Matsuoka distribution is remarkably flexible, allowing the MARMA model to serve as a parsimonious yet competitive alternative to more complex models—a notably important feature when dealing with real-world datasets with limited observations. Fifth, the density of the Matsuoka distribution is unimodal near 1, making it particularly well-suited for modeling proportions in scenarios where the data concentrate near 1. This property was explored in Matsuoka et al. (2024) in the context of production frontier modeling. In this context, firms' efficiency in competitive sectors tends to attain values close to 1, reflecting two main conditions: (i) it is an unattainable ideal for a firm to have full efficiency, with only a few operating near the frontier, and (ii) firms should not be overly concentrated far from the frontier. In the context of the particular energy source with typically high shares studied here, this also occurs.

Finally, it is well recognized that for a new statistical methodology to achieve practical relevance and widespread adoption, the availability of standard, user-friendly software implementations are essential. The proposed MARMA model is implemented in the R package BTSR (Prass and Pumi 2025), which is currently the gold-standard tool for GARMA and GARMA-like (that is, models with the same systematic component as GARMA models,

but a conditional distribution that is not a member of the exponential family in canonical form) modeling of (0, 1)-valued time series. As such, the model is readily accessible to end users and offers a viable alternative to the existing KARMA,  $\beta$ ARMA, and UWARMA models.

We propose a partial maximum likelihood estimation (PMLE) to estimate model parameters, allowing the inclusion of random time-dependent covariates in the model. The partial likelihood inference is a well-established method, and the associated theoretical properties in the context of MARMA models—particularly asymptotic results—are well developed (Fokianos and Kedem 2004). This facilitates the construction of asymptotic confidence intervals, hypothesis testing, and diagnostics. The frequentist approach based on PMLE, or more restrictively, the conditional maximum likelihood, is the most commonly used estimation method for GARMA and GARMA-like models. Bayesian inference is also a viable alternative for GARMA and GARMA-like models. See, for instance, Casarin et al. (2012), de Andrade et al. (2015), Grande et al. (2025), Lastra et al. (2025) and references therein.

We provide a standard approach for obtaining out-of-sample forecasts for the proposed model. Since deriving closed-form exact confidence intervals for out-of-sample forecasts may be impossible due to the complexity of the underlying stochastic processes, we explore a bootstrap-based method for obtaining such intervals. We consider a bootstrap scheme that takes advantage of the iterative nature of the systematic components in observation-driven models, incorporating the dependence structure uncertainty into the construction of prediction intervals. Although presented in the context of MARMA models, the method generalizes straightforwardly to any GARMA model (and even to any observation-driven model)—the method is used in the context of  $\beta$ ARMA and KARMA models in the application.

The paper is organized as follows. Section 2 introduces the proposed MARMA model. In Section 3, results regarding parameter estimation are derived, while Section 4 is dedicated to large sample inference, including hypothesis test, goodness-of-fit, model selection, and forecasting, and a bootstrap-based method for constructing prediction intervals. Finally, Section 5 presents a simulation study to explore the finite-sample performance of the proposed model parameter estimators. In Section 6 we present the main application of the paper, showcasing the use of the proposed model, highlighting the performance of the PMLE and the bootstrap-based method for constructing prediction intervals, and providing a comparison with other benchmark models. Section 7 concludes the paper.

## 2 | Matsuoka Distribution and the MARMA Model

The Matsuoka distribution is a uniparametric distribution on (0, 1) absolutely continuous with respect to the Lebesgue measure, with density given by

$$f(x; \kappa) := 2\sqrt{\frac{-\kappa^3 \ln(x)}{\pi}} x^{\kappa-1} I(x \in (0, 1)), \quad (1)$$

for  $\kappa > 0$ . We use the notation  $X \sim M(\kappa)$  to say that a random variable  $X$  follows the Matsuoka distribution with parameter  $\kappa > 0$ .

**Note on nomenclature.** The Matsuoka distribution has appeared under different names in the literature, which can lead to confusion. It was first introduced as the “log-gamma distribution” by Consul and Jain (1971), but this label was later adopted for other unrelated distributions in Hogg and Klugman (1984) and Halliwell (2021), making the term ambiguous. Grassia (1977) independently introduced the same distribution, derived its moments, and presented some real data applications, but did not provide a name for the distribution, which was later referred to as “Grassia 1” by Griffiths and Schafer (1981). To avoid ambiguity and acknowledge its role in the context of semiparametric estimation of the production frontier, Matsuoka et al. (2024) reintroduced it as the Matsuoka distribution. Throughout this paper, we use the latter name to refer specifically to the distribution with density given in (1), which is distinct from other distributions that may share similar historical labels.

Miscellaneous results for the Matsuoka distribution, including the derivation of moments, skewness and kurtosis,  $\alpha$ -expectiles, a closed-form expression for the stress-strength reliability, the Sharma-Mittal’s entropy can be found in Matsuoka et al. (2024). The density (1) displays a J-shape pattern for  $\kappa \leq 1$ , it is unimodal for  $\kappa > 1$ , but is never symmetric. The distribution is right-skewed for small values of  $\kappa$ , and gradually becomes increasingly concentrated near 1 as  $\kappa$  increases. This adaptability is crucial in applications, and we shall explore it later.

When  $\kappa$  is greater than approximately 1.72, the distribution becomes left-skewed, allowing for the modeling of left-skewed proportions with high resolution in  $\kappa$ . In other words, it is especially well suited to represent proportions concentrated near 1. Moreover, since  $\lim_{x \rightarrow 1} f(x; \kappa) = 0$  for all  $\kappa > 0$ , the distribution has no singularity at the upper boundary. This prevents the modeling of energy shares with unrealistic infinite spikes near one. In practice, it is unlikely for a region to rely exclusively on a single type of energy source. From a methodological standpoint, using distributions exhibiting such singularities may lead to overestimating the probability of high proportions. Distributions exhibiting singularities at the upper boundary include the Beta( $\alpha, \beta$ ) distribution, for  $\beta < 1$  and the Kumaraswamy in (0, 1) with parameters ( $a, b$ ), when  $b < 1$ .

From an applied environmental science perspective, this methodological distinction has direct consequences for inference and policy analysis. Models with singularities at the boundary (like the Beta with  $\beta < 1$ ) can lead to over-optimistic or alarmist conclusions. For example, when projecting the potential dominance of a specific clean energy source, an ill-fitting model might suggest a nonnegligible probability for a region to achieve a near-total dependence (e.g., 99.9%) on that source within a short timeframe, which is practically infeasible. Our chosen distribution, by avoiding this singularity, yields more conservative and realistic estimates for high proportions, ensuring that resulting analyses, such as forecasts, risk assessments, or policy targets, are grounded in plausible scenarios rather than mathematical artifacts. Therefore, by employing a distribution free from such

boundary singularities, our analysis remains robust against overestimating the probability of extreme energy shares, leading to more reliable interpretations of the energy transition dynamics captured in the data.

The cumulative distribution function associated with (1), is given by

$$F(x; \kappa) = \frac{2}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}, -\kappa \ln(x)\right) I(0 < x < 1) + I(x \geq 1), \quad (2)$$

for  $x \in \mathbb{R}$ , where  $\Gamma(k, t) = \int_t^\infty z^{k-1} e^{-z} dz$  for all  $k > 0$ , is the upper incomplete gamma function (see section 8.35 in Gradshteyn and Ryzhik 2007). From (2), it follows that the quantile function is given by  $F_\kappa^{-1}(0) = 0$ ,  $F_\kappa^{-1}(1) = 1$  and

$$F_\kappa^{-1}(Q) = \exp\left\{-\frac{1}{\kappa} \Gamma^{-1}\left(\frac{3}{2}, \frac{Q\sqrt{\pi}}{2}\right)\right\},$$

for  $Q \in (0, 1)$ , where  $\Gamma^{-1}(k, x)$  denotes the inverse of the upper incomplete gamma function. The moments are given by  $\mathbb{E}(X^k) = \left(\frac{\kappa}{\kappa+k}\right)^{\frac{3}{2}}$ , hence

$$\mathbb{E}(X) = \left(\frac{\kappa}{\kappa+1}\right)^{\frac{3}{2}} \quad \text{and} \quad \text{Var}(X) = \left(\frac{\kappa}{\kappa+2}\right)^{\frac{3}{2}} - \left(\frac{\kappa}{\kappa+1}\right)^3.$$

The Matsuoka distribution is a member of the 1-parameter regular exponential family, in the form  $f(x; \kappa) = h(x) \exp\{\eta t(x) - a(\eta)\}$ , with canonical parameter  $\eta = \kappa$ ,  $a(\eta) = -\frac{3}{2} \ln(\eta)$ ,  $h(x) = -\frac{2 \ln(x)}{x\sqrt{\pi}} I(0 \leq x \leq 1)$ , and natural complete sufficient statistic given by  $T(X) = \ln(X)$ . It can be shown that  $-\ln(X) \sim \text{Gamma}\left(\frac{3}{2}, \frac{1}{\kappa}\right)$ , so that  $\mathbb{E}(\ln(X)) = -\frac{3}{2\kappa}$ .

Let  $\{Y_t\}_{t \in \mathbb{Z}}$  be a stochastic process taking values in  $(0, 1)$  and let  $\{X_t\}_{t \in \mathbb{Z}}$  be a set of  $r$ -dimensional exogenous covariates to be included in the model. These can be either random or deterministic and time-dependent, or any combination of these. Let  $\mathcal{F}_t$  denote the information ( $\sigma$ -field) available to the observer at time  $t$ , that is,  $\mathcal{F}_t := \sigma\{X_{t+1}, Y_t, X_t, Y_{t-1}, \dots\}$ , where, by convention,  $X_t$  denotes the observed values at time  $t$  for deterministic covariates, and at time  $t-1$  for stochastic ones. Let  $g : (0, 1) \rightarrow \mathbb{R}$  be a twice differentiable bijective link function. The proposed Matsuoka autoregressive moving average (MARMA) class of models is observation-driven, for which the random component is implicitly defined by assigning  $Y_t | \mathcal{F}_{t-1} \sim M(\kappa_t)$ . Upon observing that  $\mu_t := \mathbb{E}(Y_t | \mathcal{F}_{t-1}) = \left(\frac{\kappa_t}{1+\kappa_t}\right)^{\frac{3}{2}}$ , we follow the GLM approach by setting

$$\eta_t := g(\mu_t) = \alpha + X_t' \beta + \sum_{i=1}^p \phi_i [g(Y_{t-i}) - X_{t-i}' \beta] + \sum_{j=1}^q \theta_j r_{t-j}, \quad (3)$$

where  $\eta_t$  is the linear predictor,  $\alpha$  is an intercept,  $\beta = (\beta_1, \dots, \beta_r)'$  is the parameter vector related to the covariates,  $\phi = (\phi_1, \dots, \phi_p)'$  and  $\theta = (\theta_1, \dots, \theta_q)'$  are the AR and MA coefficients, respectively. The error term in (3) is defined in a recursive fashion by setting  $r_t := g(Y_t) - g(\mu_t)$ . The proposed class of models, hereafter denoted MARMA( $p, q$ ), is defined by setting  $Y_t | \mathcal{F}_{t-1} \sim M(\kappa_t)$  together with a systematic component given by (3). In view of (3), it is clear that  $\eta_t$  and  $\mu_t$  are  $\mathcal{F}_{t-1}$ -measurable.

Observe that we can write  $\mu_t = u(\kappa_t)$ , where  $u(x) := \left(\frac{x}{1+x}\right)^{3/2}$  with inverse  $u^{-1}(y) = \frac{y^{2/3}}{1-y^{2/3}}$ , so that (3) is defined in the traditional GLM fashion. This is the approach of Benjamin et al. (2003) and Fokianos and Kedem (2004), but it is slightly different from the approach commonly used in GARMA-like models for which the distribution is not a member of the canonical exponential family, like the KARMA, UWARMA e  $\beta$ ARMA. In these cases, it is usually simpler to parameterize the distribution in terms of the measure of interest  $\mu_t$ . This can be explicitly done in this case, by saying that  $Y_t$  follows a Matsuoka distribution parameterized by  $\mu_t$ , writing  $Y_t \sim M(\mu_t)$ , when  $Y_t \sim M(\kappa_t)$  for  $\kappa_t = \left(\frac{\mu_t}{1+\mu_t}\right)^{3/2}$ . The end result is the same, though, and in the case of the traditional GARMA, it is a matter of preference.

Regarding the choice of link function, the logit, loglog, and cloglog (complementary loglog) are traditional in the GARMA, often preferred due to the availability in the package BTRSR. Since the Matsuoka distribution belongs to the exponential family in canonical form, the implied canonical link  $g(y) = \frac{y^{2/3}}{1-y^{2/3}}$  may be used as well, but its absence in the BTRSR hinders its application. Parametric alternatives can also be considered, as discussed in Pumi et al. (2020), Manchini et al. (2024), and Mendes et al. (2025).

In a distributional sense, being uniparametric in principle means the Matsuoka distribution is less flexible compared to other commonly used distributions on  $(0, 1)$ , such as the Beta and Kumaraswamy distributions, especially when all parameters are treated as free. However, in GARMA modeling with a given  $\mu_t$ , the situation is more nuanced. To understand why, consider the distance between two densities  $f$  and  $g$  on  $(0, 1)$  given by  $d(f, g) := \int_0^1 |f(x) - g(x)| dx$ .

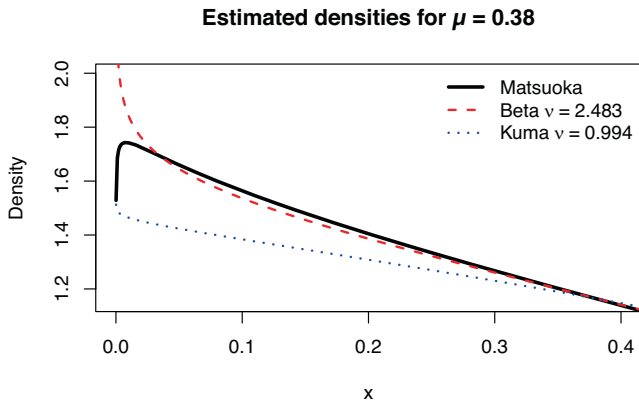
For fixed  $\mu \in (0, 1)$ , let  $f_M(\cdot; \mu)$  denote the Matsuoka distribution parameterized by  $\mu$ , and  $f_B(\cdot; \mu, \nu)$  and  $f_K(\cdot; \mu, \nu)$  denote the Beta and Kumaraswamy distributions parameterized by their mean and median, respectively, and a free shape parameter  $\nu > 0$ . For fixed  $\mu \in (0, 1)$ , consider

$$\hat{\nu}_B := \operatorname{argmin}_{\nu > 0} \{d(f_M(\cdot; \mu), f_B(\cdot; \mu, \nu))\} \quad \text{and}$$

$$\hat{\nu}_K := \operatorname{argmin}_{\nu > 0} \{d(f_M(\cdot; \mu), f_K(\cdot; \mu, \nu))\}.$$

Under this distance metric,  $f_B(\cdot; \mu, \hat{\nu}_B)$  and  $f_K(\cdot; \mu, \hat{\nu}_K)$  are the Beta and Kumaraswamy densities closest to  $f_M(\cdot; \mu)$ . Now, for  $\mu = 0.38$ , the Matsuoka density is unimodal, whereas  $f_B(\cdot; \mu, \hat{\nu}_B)$  and  $f_K(\cdot; \mu, \hat{\nu}_K)$  can only approximate  $f_M(\cdot; \mu)$  in an inverse J-shape form, with  $\hat{\nu}_B \approx 2.483$  and  $\hat{\nu}_K \approx 0.994$ . This is illustrated in Figure 1. This implies that if  $\mu_t$  fluctuates around 0.38 and the true distribution is Matsuoka, both the Beta and Kumaraswamy distributions will provide poor approximations for the dynamics, modeling as it was unlimited in the vicinity of zero.

One disadvantage of the Matsuoka distribution's uniparametric nature concerns the mean-variance relationship. While the conditional variance in  $\beta$ ARMA depends on both  $\mu_t$  and the dispersion parameter  $\nu$ , in MARMA it depends only on  $\mu_t$ . This reduced flexibility could be disadvantageous in some



**FIGURE 1** | Plot of the best fit for the Beta and Kumaraswamy densities and the Matsuoka density, for  $\mu = 0.38$ .

applications. The Matsuoka’s fixed mean-variance relationship is advantageous, for instance, when the data-generating process naturally exhibits a strong functional link between location and scale, or when parsimony is preferred to avoid overfitting in small samples. This is the case in our empirical analysis, where the MARMA specification proves sufficiently flexible to model the dataset more parsimoniously than its competitors while providing superior out-of-sample forecast performance. Conversely, in settings where the mean and variance can vary independently (e.g., when volatility clustering occurs without mean shifts), a two-parameter model like  $\beta$ ARMA or KARMA may be more appropriate.

Another important detail in applied settings is the asymmetry of the Matsuoka conditional variance, which attains its maximum at  $\mu \approx 0.36048$ . It also converges to 0 as  $\mu$  approaches the boundaries of the interval. This implies that when the MARMA process nears 0 or 1, it is more likely to remain there than to revert to the middle of (0, 1). This contrasts with the  $\beta$ ARMA model, whose conditional variance is symmetric for any fixed  $v$ . Consequently, the MARMA model may be advantageous in contexts where peak variability is expected significantly before the midpoint of the unit interval or when the data are concentrated near 0 or 1. Conversely, the  $\beta$ ARMA model is preferable when symmetric uncertainty around  $\mu_t = 0.5$  is more plausible or when the data is away from 0 or 1.

### 3 | Parameter Estimation

In this section, we propose the use of the partial maximum likelihood approach for parameter estimation in the context of MARMA models. Let  $\{Y_t\}_{t \in \mathbb{Z}}$  be a MARMA( $p, q$ ) model with associated  $r$ -dimensional covariates  $\{X_t\}_{t \in \mathbb{Z}}$ . Let  $\gamma := (\alpha, \beta, \phi, \theta)' \in \Omega$ , where  $\Omega \subset \mathbb{R}^{r+p+q+1}$  denotes the parameter space. Considering a sample  $\{(Y_t, X_t)\}_{t=1}^n$ , conditionally on a set of initial conditions  $\mathcal{F}_0$  (see Section 4.3), the partial log-likelihood function is given by

$$\ell(\gamma) = \sum_{t=1}^n \ell_t(\gamma), \quad (4)$$

where

$$\begin{aligned} \ell_t(\gamma) &:= \ln(2\pi^{-1/2}) + \frac{3}{2} \ln(\kappa_t) + \frac{1}{2} \ln(-\ln(Y_t)) + (\kappa_t - 1) \ln(Y_t) \\ &= \ln(2\pi^{-1/2}) + \frac{1}{2} \ln(-\ln(Y_t)) + \ln(\mu_t) - \frac{3}{2} \ln(1 - \mu_t^{2/3}) \\ &\quad + \left( \frac{\mu_t^{2/3}}{1 - \mu_t^{2/3}} - 1 \right) \ln(Y_t), \end{aligned}$$

since  $\kappa_t = u^{-1}(\mu_t)$  and  $\mu_t$  is specified by (3). The partial maximum likelihood estimator (PMLE) of  $\gamma$  is defined by

$$\hat{\gamma} = \operatorname{argmax}_{\gamma \in \Omega} \{\ell(\gamma)\}. \quad (5)$$

In practice, computing the partial log-likelihood function requires initialization. In the R package BTSR,  $r_t$  is initialized as  $r_t = 0$  for  $t < 1$ . By default, covariates are initialized with  $X_t = \frac{1}{p} \sum_{j=1}^p X_j$  for  $t < 1$ , while the initial values for  $Y_t$  are set to  $g^{-1}(0)$ . The package allows for the customized initialization of these values.

### 4 | Large Sample Inference

The asymptotic theory for the proposed PMLE in the context of MARMA( $p, q$ ) models falls into the general theory of Fokianos and Kedem (2004), since the Matsuoka distribution is a member of the exponential family in canonical form. To simplify the presentation, consider model (3) without any covariates. Let  $Y_1, \dots, Y_n$  be a sample from a MARMA( $p, q$ ) model and define  $Z_{t-1} := (1, g(Y_{t-1}), \dots, g(Y_{t-p}), r_{t-1}, \dots, r_{t-q})'$ , so that (3) becomes  $\eta_t = Z_{t-1}' \gamma$ , where  $\gamma = (\alpha, \phi', \theta)'$ . The conditions for the consistency and asymptotic normality of the PMLE for  $\gamma$  are presented in Fokianos and Kedem (2004), which are reproduced here for completeness and to fix the notation:

1. The true parameter  $\gamma_0$  belongs to an open set  $\Omega \subseteq \mathbb{R}^{p+q+1}$  and  $Z_{t-1}$  almost surely lies on a compact subset  $\Gamma \subset \mathbb{R}^{p+q+1}$ , such that  $P(\sum_{t=1}^n Z_{t-1}' Z_{t-1} > 0) = 1$ .
2. The link function  $g$  is twice continuously differentiable with inverse  $g^{-1}$  satisfying  $\partial g^{-1}(x) / \partial x \neq 0$  and so that  $Z_{t-1}' \gamma$  belongs almost surely to the domain of  $g^{-1}$ , for all  $\gamma \in \Omega$  and all  $t$ .
3. There exists a probability measure  $\lambda$  on  $\mathbb{R}^{p+q+1}$  such that  $\int_{\mathbb{R}^{p+q+1}} \mathbf{v} \mathbf{v}' \lambda(d\mathbf{v})$  is positive definite and such that

$$\frac{1}{n} \sum_{t=1}^n I(Z_{t-1} \in A) \xrightarrow{P} \lambda(A), \quad \text{as } n \rightarrow \infty, \text{ at } \gamma_0.$$

A detailed discussion on these assumptions and their implications can be found in Section 5 of Fokianos and Kedem (2004). In particular, Condition 3 calls for a type of ergodic theorem in the sense that if  $h$  is a continuous bounded function defined on  $\Gamma$ ,

$$\frac{1}{n} \sum_{t=1}^n h(Z_{t-1}) \xrightarrow{P} \int_{\mathbb{R}^{p+q+1}} h(\mathbf{v}) \lambda(d\mathbf{v}),$$

which implies (A3), with  $I(\gamma_0)$  positive definite and, thus, invertible. Under Assumptions 1 to 3, for large  $n$  an almost surely

unique PMLE  $\hat{\gamma}$  exists and satisfies  $\hat{\gamma} \xrightarrow{P} \gamma_0$ . Furthermore, (A3) is satisfied and

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \rightarrow N_{p+q+1}(0, I^{-1}(\gamma_0)), \quad (6)$$

in distribution, as  $n \rightarrow \infty$ , where  $I$  is given in (A4) and  $N_m(\mathbf{0}, \Sigma)$  denotes the  $m$ -variate normal distribution with mean vector  $\mathbf{0} = (0, \dots, 0)' \in \mathbb{R}^m$  and variance-covariance matrix  $\Sigma$ . The proof of these results is based on a careful analysis of the asymptotic behavior of the conditional information matrix along with a central limit theorem for the properly normalized partial score vector  $U(\gamma)$ . Closed forms for the partial score and conditional information matrix in the context of MARMA models are provided in the Appendix. Further details can be found in Fokianos and Kedem (1998, 2004).

### 4.1 | Hypothesis Testing

Let  $Y_1, \dots, Y_n$  be a sample from a MARMA( $p, q$ ) model with covariates  $X_1, \dots, X_n$ , with  $X_k \in \mathbb{R}^r$ , and let  $\gamma_0 := (\gamma_1^0, \dots, \gamma_{p+q+r+1}^0)'$  and  $\hat{\gamma} := (\hat{\gamma}_1, \dots, \hat{\gamma}_{p+q+r+1})'$  denote the true parameter vector and the PMLE estimate, respectively. The central limit theorem (6) provides a familiar framework for the construction of asymptotic confidence intervals and test statistics similar to those used in the i.i.d. context, since the approximation  $\sqrt{K_n(\hat{\gamma})}(\hat{\gamma}_j - \gamma_j^0) \approx N(0, 1)$  holds for all large enough  $n$ , where  $K_n(\hat{\gamma})^{jj}$  denotes the  $j$ th diagonal element of  $K_n(\hat{\gamma})^{-1}$  (see equation (A2) in the appendix). Level  $\delta$  confidence intervals for  $\gamma_j^0$  can be obtained straightforwardly as  $\hat{\gamma}_j \pm z_{1-\delta/2} / \sqrt{K_n(\hat{\gamma})^{jj}}$ , where  $z_{1-\delta/2}$  is the  $(1 - \delta/2)$ th quantile from a standard normal distribution. Tests of the form  $H_0 : \gamma_j = \gamma_j^*$  for some prespecified  $\gamma_j^*$  can be carried on using the following Wald's  $z$  statistics

$$z = \frac{\hat{\gamma}_j - \gamma_j^*}{\sqrt{K_n(\hat{\gamma})^{jj}}},$$

which is approximately standard normally distributed under the null hypothesis and for all sufficiently large  $n$ . Other traditional tests, such as Rao's score, likelihood ratio, among others, are constructed analogously and asymptotically follow the same distribution as their counterparts under independence. See Fahrmeir (1987) and Section 6 of Fokianos and Kedem (2004).

### 4.2 | Residuals and Goodness-of-Fit

Residual analysis and goodness-of-fit tests are crucial for any time series analysis. However, contrary to ARMA models, the error term in specification (3) is constructed iteratively so that no information about its distribution is available. In fact, since  $r_t = g(Y_t) - g(\mu_t)$ , unless  $g$  is the identity function,  $\mathbb{E}(r_t)$  cannot be computed, and it is likely to be nonzero. Hence, the obvious candidate for a residual analysis  $\hat{r}_t = g(Y_t) - g(\hat{\mu}_t)$ , where  $\hat{\mu}_t$  is obtained from the PMLE estimate  $\hat{\gamma}$ , is not a good one. Observe that the error term defined as  $e_t := Y_t - \mu_t$ , where  $\mu_t = \mathbb{E}(Y_t | \mathcal{F}_{t-1})$ , implies that  $\{(e_t, \mathcal{F}_{t-1})\}_{t \in \mathbb{Z}}$  is a martingale difference. Hence, if the MARMA model is well specified, the simple residual defined by  $\hat{e}_t := Y_t - \hat{\mu}_t$  should behave as a martingale difference with respect to  $\mathcal{F}_{t-1}$ . This can be tested using a martingale

difference test. Another commonly used approach is based on the so-called quantile residuals, defined as

$$e_t^{(q)} := \Phi^{-1}(F(Y_t | \mathcal{F}_{t-1})), \quad (7)$$

where  $F(\cdot | \mathcal{F}_{t-1})$  denotes the cumulative distribution function associated with the model's random component and  $\Phi^{-1}$  denotes the standard normal quantile function. In the present setting, when the model is correctly specified, the quantile residuals obtained from (7) after plugging in the PMLE into (2), should asymptotically follow a standard normal distribution when the model is correctly specified (see Lemma 2.1 in Kalliovirta 2012). These goodness-of-fit procedures will be further explored in the simulations.

### 4.3 | Model Selection and Forecasting

Forecasting follows the same approach as in other GARMA-like models such as the KARMA, UWARMA, and  $\beta$ ARMA models. Let  $Y_1, \dots, Y_n$  be a sample of a MARMA( $p, q$ ) model with associated covariates  $X_1, \dots, X_n$ . To obtain  $h$ -step ahead forecasts  $\hat{Y}_{n+1}, \dots, \hat{Y}_{n+h}$ , we assume that future values  $X_{n+1}, \dots, X_{n+h}$  are available or can be obtained (by forecasting, for instance). With the PMLE  $\hat{\gamma}$  in hand, starting at  $t = 1$ , we recursively obtain

$$\hat{\eta}_t = \hat{\alpha} + \hat{X}_t' \hat{\beta} + \sum_{i=1}^p \hat{\phi}_i [g(\hat{Y}_{t-i}) - \hat{X}_{t-i}' \hat{\beta}] + \sum_{k=1}^q \hat{\theta}_k \hat{r}_{t-k}, \quad (8)$$

with  $\hat{\mu}_t = g^{-1}(\hat{\eta}_t)$ , for  $t \geq 1$ ,  $\hat{r}_t = (g(\hat{Y}_t) - \hat{\eta}_t)I(1 \leq t \leq n)$ ,

$$\hat{Y}_t = \begin{cases} g^{-1}(0), & p > 0, t < 1, \\ Y_t, & 1 \leq t \leq n, \\ \hat{\mu}_t, & t > n, \end{cases} \quad \text{and} \quad \hat{X}_t = \begin{cases} \frac{1}{p} \sum_{i=1}^p X_i, & p > 0, t < 1, \\ X_t, & t \geq 1. \end{cases} \quad (9)$$

The sequence  $\hat{\mu}_1, \dots, \hat{\mu}_n$  comprises the in-sample forecasted values whereas the  $h$ -step ahead forecasted values are obtained by setting  $\hat{Y}_{n+k} = \hat{\mu}_{n+k}$  for  $k \in \{1, \dots, h\}$ . Different values for  $\hat{Y}_t$  and  $\hat{X}_t$  when  $t < 0$  in (9) may be chosen. Our experiments suggest that as long as these initial values are "reasonable" and the sample size is not too small, the effects of these default values on the forecast are negligible. Values of  $r_t$  for  $t \notin \{1, \dots, n\}$  were taken as 0, but any reasonable values can be used with negligible impact on the forecasts.

In view of (A3) and (6), the delta method allows for the construction of an approximate level  $\delta$  in-sample forecasting interval through (Fokianos and Kedem 2004)

$$CI(\mu_t; \delta) = \hat{\mu}_t \pm z_{1-\frac{\delta}{2}} \sqrt{\frac{Z_{t-1}' K_n(\hat{\gamma})^{-1} Z_{t-1}}{ng'(\hat{\mu}_t)^2}}.$$

Out-of-sample forecasting intervals in the context of  $\beta$ ARMA models were studied in Palm et al. (2023), where 5 different approaches, 3 bootstrap-based, are investigated. Although the methods presented there can be, in principle, adapted to the present context, here we shall consider a different approach

based on the recurrence nature of (8). Let  $Y_1, \dots, Y_n$  be a sample from an MARMA( $p, q$ ) model and  $\hat{\gamma}$  be the PMLE. Suppose we are interested in obtaining  $h$ -step ahead forecasting intervals for  $Y_t$ . The idea is to generate  $m$  bootstrap samples for  $Y_{n+1}, \dots, Y_{n+h}$  from an MARMA( $p, q$ ) model considering  $\hat{\gamma}$  as parameter. To accomplish that, we start by reconstructing the sequences  $\hat{\mu}_1, \dots, \hat{\mu}_{n+1}$  and  $\hat{r}_1, \dots, \hat{r}_n$  using the PMLE  $\hat{\gamma}$  through (8) as before. For each  $b \in \{1, \dots, m\}$ , the algorithm initiates by sampling  $\hat{Y}_{n+1}^{(b)}$  from a Matsuoka distribution with parameter  $\hat{\mu}_{n+1}$ , computed via (8), and then updates  $\hat{r}_{n+1}^{(b)} = g(\hat{Y}_{n+1}^{(b)}) - g(\hat{\mu}_{n+1})$ . Next, the following two steps are performed sequentially, for  $k \in \{2, \dots, h\}$ :

1. Update  $\hat{\mu}_{n+k}^{(b)}$  through (8) using  $Y_1, \dots, Y_n, \hat{Y}_{n+1}^{(b)}, \dots, \hat{Y}_{n+k-1}^{(b)}$  and  $\hat{r}_1, \dots, \hat{r}_n, \hat{r}_{n+1}^{(b)}, \dots, \hat{r}_{n+k-1}^{(b)}$ .
2. Sample  $\hat{Y}_{n+k}^{(b)}$  from a Matsuoka distribution with parameter  $\hat{\mu}_{n+k}^{(b)}$  and update  $\hat{r}_{n+k}^{(b)} = g(\hat{Y}_{n+k}^{(b)}) - g(\hat{\mu}_{n+k}^{(b)})$ .

From these steps, we obtain a collection  $\left\{ \hat{Y}_{n+1}^{(b)}, \dots, \hat{Y}_{n+h}^{(b)} \right\}_{b=1}^m$  of bootstrap samples. For each  $k \in \{1, \dots, h\}$ , a level  $\delta$  prediction interval for  $Y_{n+k}$  is obtained from the  $(1 - \delta/2)$ th and  $\delta/2$ th sample quantiles calculated from  $\hat{Y}_{n+k}^{(1)}, \dots, \hat{Y}_{n+k}^{(m)}$ .

The presented bootstrap scheme differs from those considered in Palm et al. (2023) for two main reasons. First, because it is formulated directly at the level of the recursive systematic component of GARMA models, rather than relying on residual or block resampling schemes. Second, because it propagates uncertainty through the full dynamic structure by recursively simulating future observations and innovations. We explore the finite-sample properties of the presented approach in a Monte Carlo simulation in Section 5.

Model selection in the context of GARMA models may be conducted using information criteria such as the AIC, BIC, and HQC, which are calculated as usual based on the maximized partial likelihood. Bayesian approaches via Reversible Jump Markov Chains have also been considered in the literature (Casarin et al. 2012; Lastra et al. 2025), but we shall not delve into this matter in this paper.

## 5 | Monte Carlo Simulation

In this section, we explore the finite-sample performance of the proposed PMLE approach for parameter estimation in MARMA models. Our goal is to study point and interval estimation, as well as residual analysis. The latter will be explored using the approaches delineated in Section 4.2. The simulation was carried out using R version 4.3.1 (R Core Team 2023).

### 5.1 | Point Estimation

To evaluate the accuracy and reliability of the estimator across different parameter configurations and sample sizes, considering both mean and median estimates along with their dispersion. Particular attention is given to potential biases and skewness in small samples and the impact of parameter values on estimation

performance. Additionally, we investigate the asymptotic normality of the PMLE, developed in Section 4, by analyzing the distributional behavior of the estimates as the sample size increases.

#### 5.1.1 | Data Generating Process

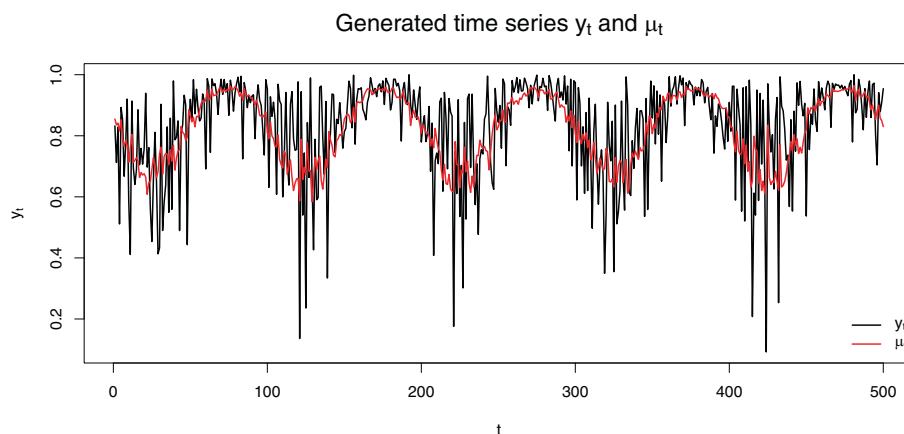
We generate samples of size  $n \in \{100, 200, 500\}$  from an MARMA(1, 1) model for all combinations of parameters  $\alpha \in \{0.5, 1\}$  and  $(\phi, \theta) \in \{(0.2, -0.8), (-0.8, 0.2), (-0.4, -0.2), (0.4, 0.2)\}$ , plus  $(\alpha, \phi, \theta) = (0.5, 0.8, -0.2)$  and considering a sinusoidal covariate given by  $X_t = \sin(\pi t/50)$  with coefficient  $\beta = -0.5$ . As the link function, we consider  $g(x) = \log(-\log(1 - x))$ , known in the literature as cloglog. The link function was selected for convenience. In simulated studies, however, the choice of link function does not affect estimation, since the fitted model is correctly specified. The combination  $(\alpha, \phi, \theta) = (1, 0.8, -0.2)$  tends to produce samples with very high positive values of  $\eta_t$ , which, after applying the inverse link function, often result in values numerically indistinguishable from 1. This leads to numerical difficulties when simulating from this parameter configuration. Therefore, we do not consider this set of parameters in our analysis. We emphasize, however, that the model remains theoretically well-defined under this configuration, and the difficulty is purely numerical. Similar numerical issues are common in GARMA and GARMA-like models, such as KARMA,  $\beta$ ARMA, and UWARMA, although they are rarely reported. See Casarin et al. (2012) and Pumi et al. (2024) for further discussion. Under these specifications, the underlying model is given by

$$\eta_t := g(\mu_t) = \alpha - 0.5 \sin(\pi t/50) + \phi [g(Y_{t-1}) + 0.5 \sin(\pi(t-1)/50)] + \theta r_{t-1},$$

where  $r_t = g(Y_t) - g(\mu_t)$ . A burn-in period of size 100 is applied to generate the time series. A total of 1000 replicas of each scenario were generated. Routines to sample from a MARMA( $p, q$ ) process and to perform estimation via PMLE are available from the R package BTSR (Prass and Pumi 2025). We used the package's default values for all tasks related to simulation, estimation, and forecasting, except where otherwise stated. For reference, a typical realization of such a process is presented in Figure 2 along with the respective  $\mu_t$ .

#### 5.1.2 | Simulation Results – Point Estimation

Table 1 summarizes the simulation results. For each set of parameters, we present the mean (left), median (center, in italics), and standard deviations (in parentheses) calculated from the 1000 replicas. The table shows that parameter  $\beta$  is remarkably well estimated in most cases. Parameter  $\alpha$  is also well estimated, especially when  $n = 500$ . A comparison between the mean and median estimates of  $\theta$  and  $\phi$  shows that for  $n = 100$ , the estimates are slightly skewed. This can be better visualized in the boxplots in Figure 3. For most parameter combinations, estimation for  $n = 500$  is fairly accurate, except for  $(\alpha, \phi, \theta) \in \{(0.5, 0.8, -0.2), (1, 0.2, -0.4)\}$  (cases 5 and 6 in Figure 3). Under the stronger autoregressive dynamics imposed by  $(\alpha, \phi, \theta) = (0.5, 0.8, -0.2)$ , the estimation procedure becomes considerably more challenging. Although point estimation improves as  $n$



**FIGURE 2** | A typical example of a time series considered in the simulation study. The plot was generated considering  $n = 500$ ,  $\alpha = 0.5$ ,  $\beta = -0.5$ ,  $\phi = 0.2$ ,  $\theta = -0.4$ .

increases, the parameters  $\alpha$ ,  $\theta$ , and  $\phi$  tend to be either jointly well estimated or jointly poorly estimated. This induces a bimodal pattern in their empirical distributions, as reflected in the boxplots.

For  $n = 100$ , the performance is typically fair, but considerable bias is observed in some parameter configurations—a commonly seen feature in partial maximum likelihood estimation of GARMA and GARMA-like models, such as the  $\beta$ ARMA and the KARMA. Hence, one should be aware of the possible presence of bias in MARMA (and other GARMA and GARMA-like) modeling in small samples. Interestingly,  $(\alpha, \phi, \theta) = (0.5, 0.2, -0.4)$  (case 1 in Figure 3) is well estimated. The estimates for  $\alpha = 1$  for this particular combination (case 6 in Figure 3) were poor as the peaks of the generated time series get very close to 1 in this scenario, causing the optimization of the log-likelihood to fail to converge. It is also perceptible that the variance of the PMLE differs depending on the parameter combination.

### 5.1.3 | Simulation Results – Joint Behavior

To investigate the asymptotic normality of the PMLE, we examine pairwise scatter plots and marginal behavior (histograms and boxplots) presented in Figure 4 for the case  $\alpha = 1$ ,  $\beta = -0.5$ ,  $\phi = -0.4$ ,  $\theta = -0.2$ . Other cases are presented in Figures S1–S8 in the Supporting Information and reveal similar behavior. In general, the results suggest that the empirical distributions approach a Gaussian shape as  $n$  increases, while the right column of Figure 4 indicates a strong dependence among the estimates of  $\alpha$ ,  $\phi$ , and  $\theta$ . In sharp contrast, the estimates of  $\beta$  seem uncorrelated with the other parameters, as seen in the plots in the left column of Figure 4. The boxplots become more symmetrical as  $n$  increases, with histograms resembling the shape of a normal distribution.

To complement the graphical analysis, we also applied two formal multivariate normality tests, namely the Henze-Zirkler and the Energy test, both implemented in the MVN package in R (Korkmaz et al. 2014). Visual inspection revealed the presence of several outliers in the scatter plots, and these tests are known to be sensitive to multivariate outliers—measured in terms of the adjusted quantile method based on Mahalanobis distance,

as implemented in the MVN package. Therefore, we reran the tests after removing up to 10% of the most extreme observations. The test results can be found in Table 2. Our findings show a marked improvement in the test results for sample sizes  $n > 100$  after the removal of outliers, except perhaps for  $(\phi, \theta)$ , which exhibited the highest number of discrepant points. For  $n = 100$ , results remained less satisfactory, which is not surprising given the smaller sample size.

## 5.2 | Parametric Bootstrap Confidence Intervals

In this section, we construct parametric bootstrap confidence intervals for individual parameters. For this exercise, the DGP is the same as in Section 5.1.1. To construct the parametric bootstrap confidence intervals, we first compute the fitted parameter vector  $\hat{\gamma}$  for each generated time series. Using these estimates, we then generate 1000 bootstrap samples, refit the model to each, and obtain corresponding bootstrap estimates  $\check{\gamma}_1, \dots, \check{\gamma}_{1000}$ . Finally, the  $100(1 - \delta)\%$  bootstrap confidence intervals are derived by taking the  $\delta/2$  and  $1 - \delta/2$  quantiles of the bootstrap distribution. We consider  $\delta = 0.05$ , and the quantiles necessary were obtained using function `colQuantiles` from R package `matrixStats` (Bengtsson 2025), which provides optimized and fast matrix quantile calculation.

When constructing a parametric bootstrap confidence interval, an additional difficulty arises: after generating the original time series using the true parameters  $\gamma$ , we must first estimate  $\hat{\gamma}$  and then generate samples from this estimate. This is straightforward when  $|\phi|$  is small; even with considerable bias,  $|\phi|$  remains relatively small, allowing stable sampling. The situation changes drastically for larger values, such as  $|\phi| = 0.8$ . Here, the estimated  $|\phi|$  is often driven to values close to 1, making sampling challenging. The generated time series then frequently produce values too close to 0 or 1, leading to numerical instability (see the discussion on the variance of the Matsuoka distribution in Section 2).

To partially counter this effect, after sampling from  $\hat{\gamma}$ , we tested whether the generated time series contained values too close to 0 or 1. If it did, a new sample was generated and tested again until

**TABLE 1** | Simulation results – point estimates based on 1000 replicas of each scenario. For each  $n$ ,  $\alpha$ ,  $\beta$ ,  $\phi$ , and  $\theta$ , the presented values correspond to the mean (left), the median (center, in italic), and the standard deviation (right, in parentheses).

$n$	$\alpha = 0.5$			$\beta = -0.5$			$\phi = 0.2$			$\theta = -0.4$		
100	0.539	<i>0.513</i>	(0.238)	-0.500	<i>-0.499</i>	(0.043)	0.145	<i>0.185</i>	(0.392)	-0.391	<i>-0.443</i>	(0.454)
200	0.514	<i>0.499</i>	(0.184)	-0.502	<i>-0.502</i>	(0.029)	0.180	<i>0.199</i>	(0.300)	-0.398	<i>-0.432</i>	(0.310)
500	0.501	<i>0.500</i>	(0.114)	-0.501	<i>-0.501</i>	(0.018)	0.200	<i>0.201</i>	(0.187)	-0.408	<i>-0.417</i>	(0.179)
$n$	$\alpha = 0.5$			$\beta = -0.5$			$\phi = -0.8$			$\theta = 0.2$		
100	0.513	<i>0.515</i>	(0.061)	-0.486	<i>-0.488</i>	(0.042)	-0.745	<i>-0.764</i>	(0.087)	0.147	<i>0.164</i>	(0.108)
200	0.512	<i>0.518</i>	(0.054)	-0.487	<i>-0.489</i>	(0.037)	-0.761	<i>-0.770</i>	(0.050)	0.164	<i>0.172</i>	(0.061)
500	0.518	<i>0.520</i>	(0.030)	-0.490	<i>-0.489</i>	(0.020)	-0.774	<i>-0.779</i>	(0.027)	0.177	<i>0.181</i>	(0.032)
$n$	$\alpha = 0.5$			$\beta = -0.5$			$\phi = -0.4$			$\theta = -0.2$		
100	0.484	<i>0.488</i>	(0.064)	-0.497	<i>-0.498</i>	(0.043)	-0.372	<i>-0.376</i>	(0.151)	-0.227	<i>-0.234</i>	(0.186)
200	0.481	<i>0.488</i>	(0.058)	-0.490	<i>-0.497</i>	(0.045)	-0.410	<i>-0.395</i>	(0.116)	-0.173	<i>-0.208</i>	(0.162)
500	0.477	<i>0.492</i>	(0.056)	-0.485	<i>-0.497</i>	(0.044)	-0.431	<i>-0.408</i>	(0.095)	-0.139	<i>-0.188</i>	(0.157)
$n$	$\alpha = 0.5$			$\beta = -0.5$			$\phi = 0.4$			$\theta = 0.2$		
100	0.609	<i>0.578</i>	(0.190)	-0.490	<i>-0.486</i>	(0.082)	0.296	<i>0.322</i>	(0.187)	0.271	<i>0.269</i>	(0.171)
200	0.552	<i>0.539</i>	(0.119)	-0.493	<i>-0.492</i>	(0.058)	0.351	<i>0.360</i>	(0.118)	0.233	<i>0.230</i>	(0.116)
500	0.519	<i>0.517</i>	(0.066)	-0.497	<i>-0.497</i>	(0.038)	0.383	<i>0.382</i>	(0.066)	0.210	<i>0.211</i>	(0.066)
$n$	$\alpha = 0.5$			$\beta = -0.5$			$\phi = 0.8$			$\theta = -0.2$		
100	2.500	<i>2.539</i>	(0.274)	-0.468	<i>-0.468</i>	(0.033)	0.021	<i>0.006</i>	(0.106)	0.479	<i>0.489</i>	(0.113)
200	1.782	<i>2.356</i>	(0.828)	-0.473	<i>-0.473</i>	(0.025)	0.301	<i>0.077</i>	(0.322)	0.264	<i>0.395</i>	(0.244)
500	0.700	<i>0.693</i>	(0.124)	-0.483	<i>-0.483</i>	(0.016)	0.722	<i>0.724</i>	(0.048)	-0.089	<i>-0.088</i>	(0.052)
$n$	$\alpha = 1$			$\beta = -0.5$			$\phi = 0.2$			$\theta = -0.4$		
100	1.252	<i>1.239</i>	(0.258)	-0.498	<i>-0.499</i>	(0.028)	-0.001	<i>0.009</i>	(0.211)	-0.237	<i>-0.249</i>	(0.276)
200	1.204	<i>1.181</i>	(0.250)	-0.499	<i>-0.498</i>	(0.018)	0.037	<i>0.057</i>	(0.202)	-0.260	<i>-0.284</i>	(0.229)
500	1.115	<i>1.097</i>	(0.213)	-0.499	<i>-0.499</i>	(0.011)	0.108	<i>0.124</i>	(0.173)	-0.321	<i>-0.342</i>	(0.179)
$n$	$\alpha = 1$			$\beta = -0.5$			$\phi = -0.8$			$\theta = 0.2$		
100	0.963	<i>0.975</i>	(0.082)	-0.494	<i>-0.496</i>	(0.039)	-0.737	<i>-0.769</i>	(0.120)	0.130	<i>0.160</i>	(0.143)
200	0.980	<i>0.989</i>	(0.060)	-0.491	<i>-0.494</i>	(0.032)	-0.767	<i>-0.781</i>	(0.060)	0.166	<i>0.176</i>	(0.078)
500	0.989	<i>1.000</i>	(0.053)	-0.490	<i>-0.494</i>	(0.027)	-0.781	<i>-0.787</i>	(0.030)	0.184	<i>0.186</i>	(0.038)
$n$	$\alpha = 1$			$\beta = -0.5$			$\phi = -0.4$			$\theta = -0.2$		
100	0.943	<i>0.949</i>	(0.117)	-0.500	<i>-0.500</i>	(0.031)	-0.319	<i>-0.329</i>	(0.174)	-0.292	<i>-0.294</i>	(0.196)
200	0.970	<i>0.971</i>	(0.074)	-0.500	<i>-0.501</i>	(0.021)	-0.358	<i>-0.363</i>	(0.111)	-0.246	<i>-0.250</i>	(0.130)
500	0.988	<i>0.988</i>	(0.043)	-0.500	<i>-0.500</i>	(0.014)	-0.383	<i>-0.384</i>	(0.065)	-0.218	<i>-0.220</i>	(0.078)
$n$	$\alpha = 1$			$\beta = -0.5$			$\phi = 0.4$			$\theta = 0.2$		
100	1.451	<i>1.534</i>	(0.354)	-0.488	<i>-0.486</i>	(0.043)	0.144	<i>0.095</i>	(0.203)	0.387	<i>0.426</i>	(0.181)
200	1.230	<i>1.214</i>	(0.302)	-0.494	<i>-0.493</i>	(0.030)	0.269	<i>0.279</i>	(0.173)	0.290	<i>0.296</i>	(0.155)
500	1.070	<i>1.055</i>	(0.168)	-0.498	<i>-0.498</i>	(0.020)	0.360	<i>0.368</i>	(0.096)	0.224	<i>0.224</i>	(0.092)

an “acceptable” one was obtained. To prevent infinite loops, the procedure was halted after 50 attempts; if the replication failed, it was recorded as NA, and the simulation advanced to the next replica. For all scenarios where  $|\phi| \leq 0.4$ , bootstrap samples from

$\hat{\gamma}$  were generated without issue. In contrast, for  $|\phi| = 0.8$ , the failure rate was substantially higher, reaching 78% across 1000 replicas, despite our fail-safe. Consequently, these cases were excluded from the final results.

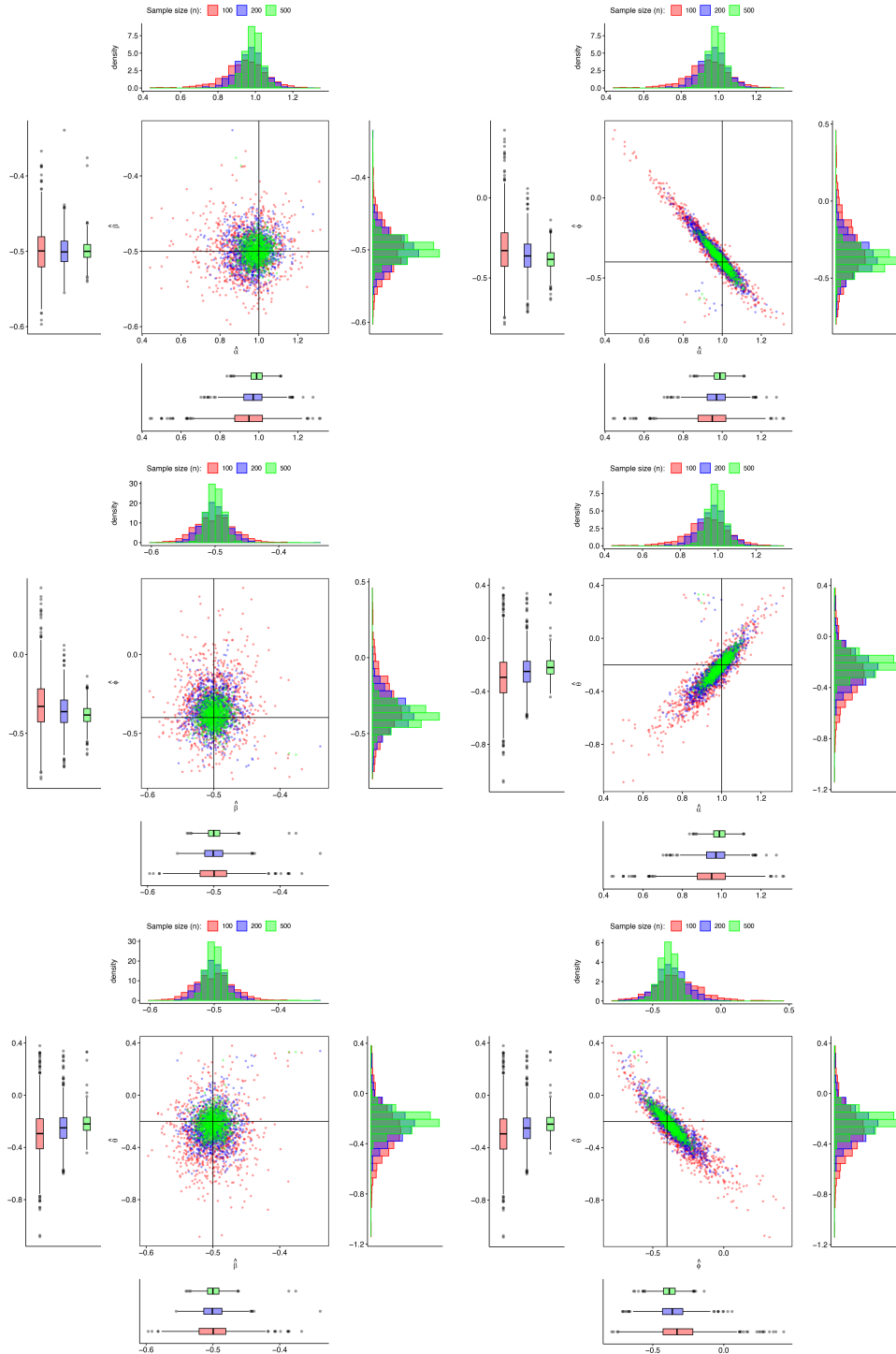


**FIGURE 3** | Boxplots of the simulation results for all parameters with  $\alpha = 0.5$  (top) and  $\alpha = 1$  (bottom), keeping  $\beta = -0.5$  fixed. The results for parameters  $(\phi, \theta)$  are labeled as follows. Cases 1 and 6:  $(0.2, -0.4)$ , cases 2 and 7:  $(-0.8, 0.2)$ , cases 3 and 8:  $(-0.4, -0.2)$ , cases 4 and 9:  $(0.4, 0.2)$ , case 5:  $(0.8, -0.2)$ . Vertical blue lines indicate the true parameter value.

**5.2.1 | Simulation Results**

Table 3 presents empirical convergence rates (in %) for the 95% confidence intervals for each combination of parameters, while Figure 5 complements the results, presenting the boxplot of the upper (U) and lower (L) confidence limits for each scenario characterized by case. From Table 3, the true value of  $\alpha$  appears to

be the single most influential factor in the results. Performance is markedly better and more stable when  $\alpha = 0.5$  compared to  $\alpha = 1.0$ , which can be explained by the higher variance observed in the estimation of  $\alpha$  compared to other parameters (see Table 1). As expected, coverage generally improves as  $n$  increases from 100 to 500, especially for problematic cases under  $\alpha = 1.0$ . The intercept parameter  $\beta$  consistently shows the best and most



**FIGURE 4** | Pairwise joint and marginal behavior of the estimated values for  $\alpha = 1$ ,  $\beta = -0.5$ ,  $\phi = -0.4$ ,  $\theta = -0.2$ . Solid lines in the scatter plot represent the true values.

stable coverage across all scenarios, suggesting it is the easiest parameter to estimate. This reinforces the findings presented in Section 5.1.2 (Table 1).

For  $\alpha = 0.5$  (cases 1, 3, and 4 in Figure 5), the performance is generally good. For  $\phi = 0.2$  and  $\theta = -0.4$  (case 1), the bootstrap intervals are strongly conservative for small values of  $n$ , converging appropriately to the 93%–96% range at  $n = 500$ . For

$\phi = -0.4$  and  $\theta = -0.2$  (case 3), moderate undercoverage emerges, particularly for  $\alpha$  and  $\theta$ , which worsens slightly with  $n$  (e.g.,  $\alpha$  coverage drops to 82.4% at  $n = 500$ ). This is due to persistent estimation bias in these parameters under this configuration (see Table 1). For the case with  $\phi = 0.4$  and  $\theta = 0.2$  (case 4), coverage starts low at  $n = 100$  (83%–92%) but improves steadily with  $n$ , reaching satisfactory levels (90%–96%) at  $n = 500$ .

**TABLE 2** | Pairwise bivariate normal testing of the estimated values for  $\alpha = 1$ ,  $\beta = -0.5$ ,  $\phi = -0.4$ ,  $\theta = -0.2$ . Presented are the  $p$ -values of the Henze-Zirkler and the Energy test, considering all replicas (top line) and after removal of (at most) 10% of the most extreme multivariate outliers.

$n$	Henze-Zirkler test						Energy test					
	$(\alpha, \beta)$	$(\alpha, \phi)$	$(\alpha, \theta)$	$(\beta, \phi)$	$(\beta, \theta)$	$(\phi, \theta)$	$(\alpha, \beta)$	$(\alpha, \phi)$	$(\alpha, \theta)$	$(\beta, \phi)$	$(\beta, \theta)$	$(\phi, \theta)$
100	0.0000 *	0.0000	0.0000	0.0008	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.0018	0.0001	0.0000	0.1604	0.1799	0.0000	0.0000	0.0000	0.0000	0.0090	0.0490	0.0000
200	0.0121	0.0000	0.0000	0.0544	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.2046	0.1746	0.0083	0.5161	0.3480	0.0000	0.0790	0.0920	0.0000	0.1700	0.1390	0.0000
500	0.0034	0.0000	0.0000	0.0025	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.3437	0.7728	0.7770	0.3648	0.7382	0.0033	0.1740	0.5750	0.3130	0.2920	0.4200	0.0110

Note: \* 0.0000 means  $p$ -value  $< 0.00005$ .

**TABLE 3** | Coverage percentages of 95% confidence intervals (in %).

$\alpha = 0.5$					$\alpha = 1.0$				
$n$	$\alpha = 0.5$	$\beta = -0.5$	$\phi = 0.2$	$\theta = -0.4$	$n$	$\alpha = 1.0$	$\beta = -0.5$	$\phi = 0.2$	$\theta = -0.4$
100	99.9%	89.6%	100.0%	99.6%	100	100.0%	94.8%	99.9%	100.0%
200	99.9%	89.9%	99.8%	98.9%	200	99.9%	93.8%	99.9%	100.0%
500	96.4%	94.0%	96.1%	95.4%	500	94.8%	93.9%	95.4%	95.9%
$n$	$\alpha = 0.5$	$\beta = -0.5$	$\phi = -0.4$	$\theta = -0.2$	$n$	$\alpha = 1.0$	$\beta = -0.5$	$\phi = -0.4$	$\theta = -0.2$
100	89.7%	95.0%	98.7%	99.3%	100	86.5%	92.8%	87.9%	85.8%
200	85.8%	91.2%	95.1%	91.7%	200	89.4%	92.3%	90.5%	89.7%
500	82.4%	84.7%	86.1%	82.2%	500	90.3%	93.5%	91.5%	91.1%
$n$	$\alpha = 0.5$	$\beta = -0.5$	$\phi = 0.4$	$\theta = 0.2$	$n$	$\alpha = 1.0$	$\beta = -0.5$	$\phi = 0.4$	$\theta = 0.2$
100	83.4%	91.8%	86.9%	91.3%	100	29.1%	85.3%	29.1%	43.8%
200	88.7%	93.0%	89.5%	91.7%	200	63.9%	89.3%	64.5%	69.4%
500	90.2%	95.6%	91.2%	92.4%	500	82.0%	93.0%	82.4%	87.4%

For  $\alpha = 1$  (cases 6, 8, and 9), the performance is significantly worse, reflecting the substantial estimation bias reported in Table 1, particularly at small  $n$ . The case  $\phi = 0.2$  and  $\theta = -0.4$  (case 6) shows a similar pattern as case 1, converging well to values close to the nominal at  $n = 500$ . For  $\phi = -0.4$  and  $\theta = -0.2$  (case 8), the bias reflected in Table 1 translates into significant undercoverage across all parameters at  $n = 100$  (86%–93%), with only marginal improvement as  $n$  increases. Case 9,  $\phi = 0.4$  and  $\theta = 0.2$ , is the most critical one. For  $n = 100$ , coverage for  $\alpha$ ,  $\phi$ , and  $\theta$  is catastrophically low (29.1%, 29.1%, 43.8% respectively), reflecting the high bias in parameter estimation found in Table 1. While coverage improves with  $n$ , it remains under 90% even at  $n = 500$  for  $\alpha$  and  $\phi$  (82.0%, 82.4%). Figure 5 showcases these problems in Case 9. This specific combination of a larger  $\alpha$  with positive  $\phi$  and  $\theta$  appears highly problematic, likely due to data being generated near the support boundaries, leading to inflated estimation bias.

### 5.3 | Goodness-of-Fit Tests

This section examines the finite-sample performance of goodness-of-fit tests based on the simple and quantile residuals discussed in Section 4.2. For the simple residuals,  $\hat{\epsilon}_t = Y_t - \hat{\mu}_t$ , where  $\hat{\mu}_t$  is obtained using the estimated parameter values. If

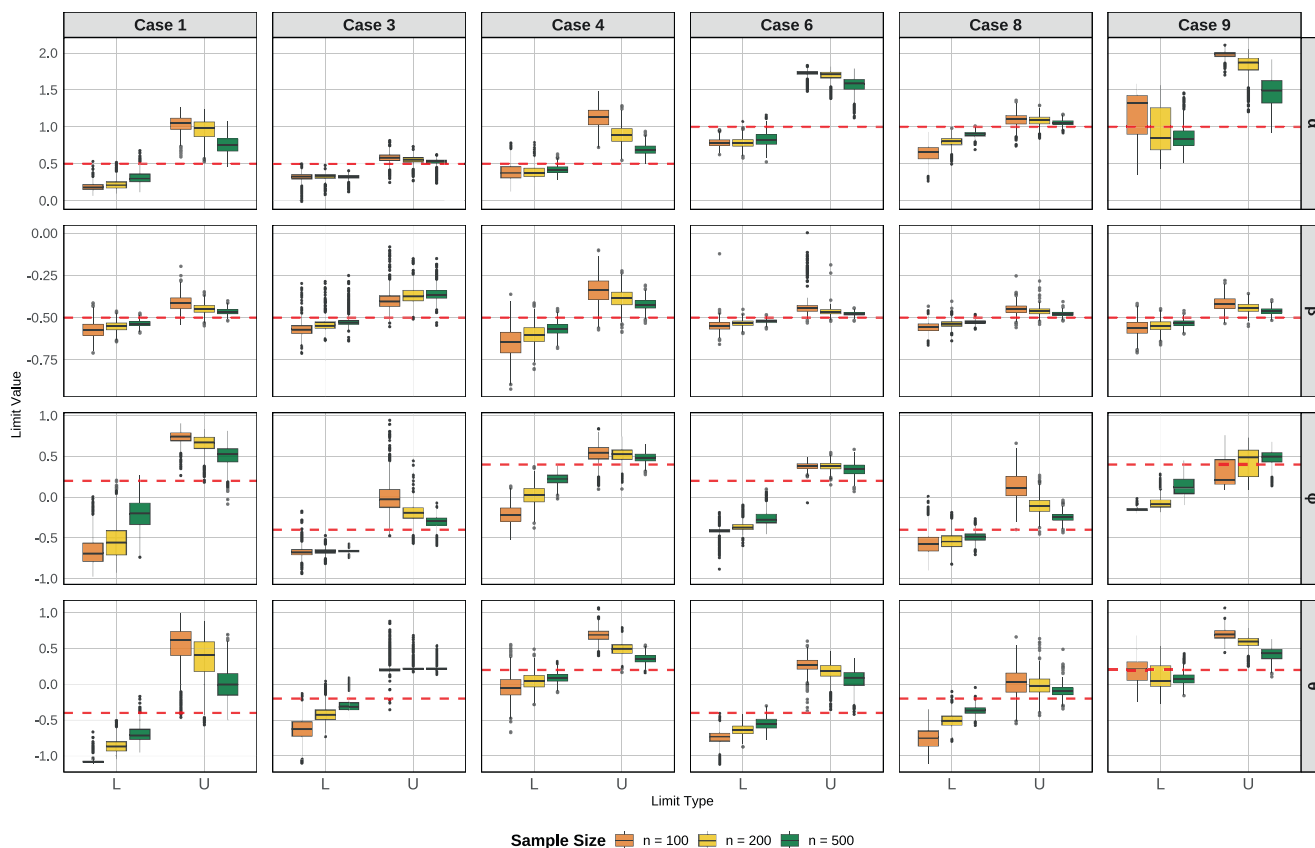
the model is well specified,  $\hat{\epsilon}_t$  should approximately behave as a martingale difference with respect to the process's history, up to estimation error. This can be tested using any martingale difference test, such as the wild bootstrap automatic variance ratio test (WB for short) proposed by Choi (1999), which is adopted here. In our experience, we found that the WB presents a good balance between power and computational speed. The finite-sample performance of this and other methods is discussed in Charles et al. (2011). The method is implemented in the R package `vrtest` (Kim 2014). As for the quantile residuals, under the correct model specification, they should asymptotically follow a standard normal distribution. For testing purposes, we apply five commonly used normality tests: the Anderson-Darling (AD), Cramér-von Mises (CvM), Kolmogorov-Smirnov (KS) implemented in the R package `nortest` (Gross and Ligges 2015), and the Shapiro-Francia (SF), available from base R. Further details on these tests can be found in Thode (2002).

#### 5.3.1 | Data Generating Process

For this exercise, we generate samples of size  $n \in \{100, 200, 500\}$  from an MARMA(1,1) model for parameters  $\alpha \in \{0.5, 1\}$  and  $(\phi, \theta) \in \{(0.2, -0.8), (-0.8, 0.2), (-0.4, -0.2), (0.4, 0.2)\}$ , as well as

### Confidence Interval Limits: L (Lower) and U (Upper)

Red line: true parameter value. Colors represent sample size.



**FIGURE 5** | Boxplots of the upper (U) and lower (L) bootstrap confidence intervals. The red lines represent the true parameter value. The models are represented by:  $\alpha = 0.5$  and  $(\phi, \theta)$  as follows: Case 1 (0.2, -0.4); Case 3 (-0.4, -0.2); Case 4 (0.4, 0.2); For  $\alpha = 1$ , Case 6 (0.2, -0.4), Case 8 (-0.4, -0.2); and Case 9 (0.4, 0.2); Cases 2, 5 and 7 presented  $|\phi| = 0.8$  and were excluded.

for  $(\alpha, \phi, \theta) = (0.5, 0.8, -0.2)$ . Similarly to Section 5.1, the combination  $(\alpha, \phi, \theta) = (1, 0.8, -0.2)$  tends to yield numerical issues due to values extremely close to 1. Covariates are not considered, and  $g(\cdot)$  is the cloglog function. Under these specifications, the underlying model is given by

$$\eta_t := g(\mu_t) = \alpha + \phi g(Y_{t-1}) + \theta r_{t-1}, \quad g(x) = \log(-\log(1-x)).$$

A burn-in period of size 100 is applied to generate the time series. A total of 1000 replicas of each scenario were generated, and tests were evaluated at a 5% significance level. To perform the WB test, we consider 500 bootstrap samples using Mammen’s two-point distribution (see package `vrtest`’s documentation for details).

#### 5.3.2 | Simulation Results

Table 4 summarizes the simulation results. All tests performed well, with rejection rates close to the nominal level of 0.05 for most parameter configurations. The normality tests yielded broadly similar results, as expected. However, two notable exceptions were observed: for  $\alpha = 1$ ,  $\phi = -0.8$  and  $\theta = 0.2$ , the SF test exhibited somewhat higher-than-expected rejection rates. In the case of  $\alpha = 0.5$ ,  $\phi = 0.8$ , and  $\theta = -0.2$ , the SF test performed extremely poorly. It is known that the Shapiro-Francia test is very sensitive to skewness, especially when compared

to the alternative tests applied. Upon manual inspection, we found that the simulated time series in the latter scenario were highly concentrated near 1, which not only made the estimation more challenging—as evidenced in Table 1—but also led to more skewed residuals compared to other scenarios. As a result, the Shapiro-Francia test showed severely inflated rejection rates for this particular scenario, standing out among the considered tests.

### 5.4 | Prediction Intervals

In this exercise, we explore the finite-sample performance of the bootstrap approach discussed in Section 4.3 for constructing prediction intervals.

#### 5.4.1 | Data Generating Process

In this study, for simplicity, we fix  $\alpha = 1$ . The remaining parameters follow the same settings as in the goodness-of-fit exercise. We generate MARMA(1, 1) models without any covariates, with the following configurations:

- i. **Model 1** with parameter  $\phi = -0.8$  and  $\theta = 0.2$ ;
- ii. **Model 2:**  $\phi = -0.4$  and  $\theta = -0.2$ ;

**TABLE 4** | Simulation results – martingale difference, and normality tests. For each  $n$ ,  $\alpha$ , and  $(\phi, \theta)$ , the results presented correspond to the proportion of tests that rejected the null hypothesis for each specific test.

$n$	$(\phi, \theta)$	$\alpha = 0.5$					$\alpha = 1$				
		WB	AD	CvM	KS	SF	WB	AD	CvM	KS	SF
100	$\phi = 0.2 \theta = -0.4$	0.03	0.06	0.06	0.05	0.06	0.03	0.04	0.05	0.05	0.04
200		0.01	0.05	0.05	0.06	0.05	0.02	0.06	0.06	0.05	0.06
500		0.00	0.05	0.06	0.05	0.06	0.01	0.05	0.05	0.05	0.04
100	$\phi = -0.8 \theta = 0.2$	0.04	0.06	0.05	0.06	0.08	0.06	0.09	0.08	0.06	0.17
200		0.06	0.06	0.05	0.05	0.09	0.06	0.07	0.07	0.06	0.15
500		0.06	0.04	0.04	0.03	0.05	0.07	0.05	0.05	0.04	0.12
100	$\phi = -0.4 \theta = -0.2$	0.03	0.05	0.05	0.05	0.06	0.02	0.05	0.05	0.06	0.06
200		0.05	0.06	0.06	0.06	0.07	0.03	0.06	0.06	0.06	0.08
500		0.09	0.05	0.05	0.04	0.10	0.03	0.04	0.04	0.05	0.08
100	$\phi = 0.4 \theta = 0.2$	0.04	0.05	0.05	0.06	0.05	0.04	0.07	0.07	0.06	0.06
200		0.03	0.05	0.06	0.06	0.05	0.05	0.07	0.07	0.06	0.06
500		0.04	0.05	0.05	0.05	0.05	0.07	0.06	0.06	0.05	0.06
100	$\phi = 0.8 \theta = -0.2$	0.18	0.13	0.12	0.11	0.17	—	—	—	—	—
200		0.15	0.19	0.16	0.12	0.57	—	—	—	—	—
500		0.00	0.09	0.08	0.06	0.99	—	—	—	—	—

iii. **Model 3:**  $\phi = 0.2$  and  $\theta = -0.4$ ;

iv. **Model 4:**  $\phi = 0.2$  and  $\theta = 0.4$ .

Top: each block presents the coverage of the bootstrap prediction interval for a given  $n$  (column) and  $\delta$  (row) as a function of the forecasting horizon  $h$ , with nominal level indicated by the blue lines. Bottom: each block presents the average interval's width of the bootstrap prediction interval for  $\delta \in \{0.01, 0.05, 0.1\}$ , for a given  $n$  (column) and model (rows), as a function of the forecasting horizon  $h$ . We consider cloglog as the link function, and samples of size  $n \in \{100, 200, 500\}$ . Under these specifications, the underlying model is given by

$$\eta_t := g(\mu_t) = 1 + \phi g(Y_{t-1}) + \theta r_{t-1}, \quad g(x) = \log(-\log(1-x)).$$

A burn-in period of size 100 was applied to generate the time series, and 1000 replicas were considered.

For each time series, we estimate the parameter vector  $\gamma = (\alpha, \phi, \theta)$  via PMLE and use this estimate to create  $m = 500$  bootstrap samples, as explained in Section 4.3, considering  $h = 50$  as the forecasting horizon. Confidence intervals of level  $\delta \in \{0.1, 0.05, 0.01\}$  are obtained as explained in Section 4.3. The R package `BTSR` was used to obtain the bootstrap samples, while the base R function `quantile` (using defaults) was employed to get the lower and upper boundaries of the confidence intervals.

### 5.4.2 | Simulation Results

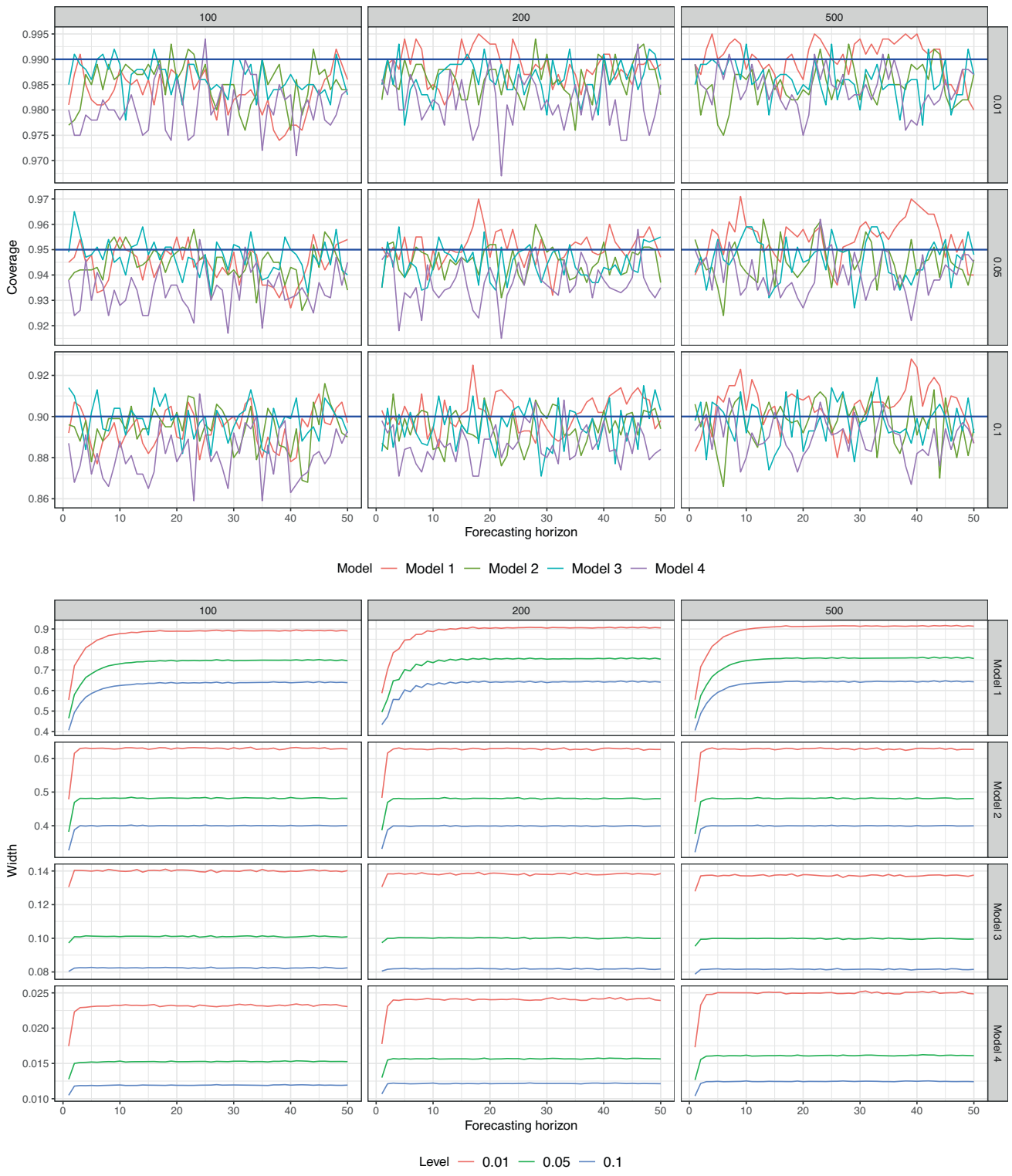
Figure 6 presents the simulation results. In the top panels, each block depicts the coverage of the bootstrap prediction confidence interval as a function of the forecasting horizon  $h$ , for each combination of sample size  $n$  (columns) and level  $\delta$  (rows). The blue

lines indicate the nominal coverage level. In the bottom panels, each block displays the average width of the bootstrap prediction interval for  $\delta \in \{0.01, 0.05, 0.1\}$ , again for each  $n$  (column) and model (row), as a function of  $h$ .

The top panels show that coverage is generally close to the nominal level across all forecasting horizons and sample sizes. Also, the sample size does not appear to considerably affect the prediction interval's width across forecasting horizons, which tends to stabilize in value after a few steps. The average interval width depends strongly on the model specification, with Model 4 yielding the most liberal intervals and Model 1 the most conservative. Model 1's parameter configuration tends to generate trajectories near the upper boundary of 1. As a result, the upper limits of the prediction intervals are constrained near 1, while the lower limits extend further downward to preserve the nominal coverage. This conservativeness is a desirable property in this context, as it prevents unrealistic predictions outside the  $(0, 1)$  interval. The contrast in behavior between Models 1 and 4 is also evident in their coverage performance, as seen in the top panels of Figure 6.

## 6 | Application to Real Data

This section evaluates the performance of the proposed model in forecasting the proportion of net electricity generated by all means but conventional hydroelectric power in the United States. Accurate short- and medium-term forecasts of this proportion are relevant for planning reserve capacity and assessing the reliability of an increasingly renewable-dependent power system. Conventional hydropower maintains a pivotal role in the United States' renewable energy portfolio, representing approximately 6.2% of total (utility-scale) electricity generation and 28.7% of renewable energy production (U.S. Energy Information



**FIGURE 6** | Simulation results for the prediction confidence interval exercise. Top: each block shows the coverage of the bootstrap prediction interval as a function of the forecasting horizon  $h$ , for a given sample size  $n$  (columns) and level  $\delta$  (rows). Blue lines indicate the nominal level. Bottom: each block shows the average width of the bootstrap prediction interval for  $\delta \in \{0.01, 0.05, 0.1\}$ , for each  $n$  (column) and model (rows), as a function of  $h$ .

Administration (EIA 2023).<sup>1</sup> As the most established renewable energy source, it provides critical baseload power and grid stability, complementing variable renewable sources such as wind and solar energy. The technology’s capacity for rapid output adjustment and energy storage through pumped-storage systems

further enhances its value in maintaining a reliable electricity supply.

The advantages of conventional hydropower are multifaceted. It offers low operational costs and an extended lifespan of 50 to

100 years, with minimal greenhouse gas emissions after the initial construction phase. Additionally, hydropower infrastructure often serves secondary purposes, including irrigation support, flood control, and recreational opportunities, thereby contributing to broader water resource management objectives. As the renewable energy sector evolves, it will continue to play a vital, though increasingly specialized, role in achieving a sustainable and resilient energy system.

Given the critical yet constrained role of conventional hydropower in the U.S. renewable energy mix outlined, there is a compelling need to investigate net electricity generation from all other sources to better understand the nation's evolving energy transition. While hydropower provides stable baseload power, its growth potential is limited by environmental, geographic, and climate-related challenges. Consequently, the expansion of renewable capacity and grid reliability increasingly depends on non-hydropower sources, such as wind, solar, and emerging technologies. By isolating non-hydropower generation, we obtain critical insights into the evolving balance between traditional and emerging energy technologies. This focused approach offers policymakers and utilities a clearer understanding of America's true energy production capabilities beyond its established hydropower infrastructure.

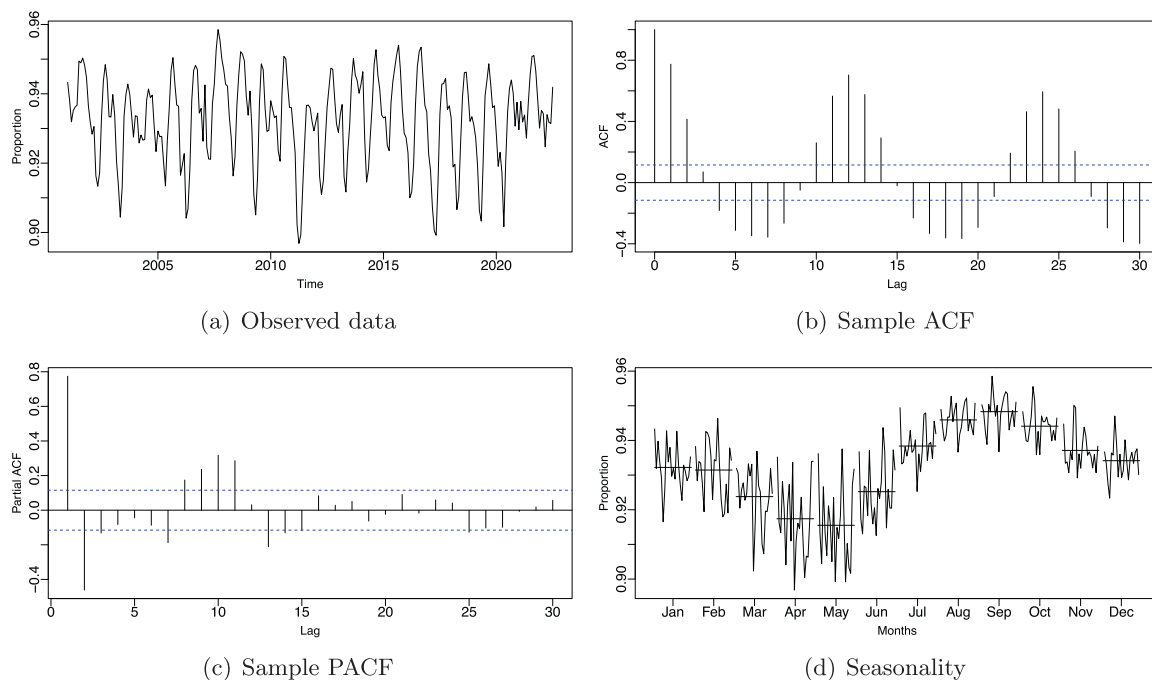
## 6.1 | Data Description

The data used for this analysis are the monthly proportions of net electricity generated by all means (i.e., coal, petroleum liquids, petroleum coke, natural gas, other gases, nuclear, other renewables (total), hydro-electric pumped storage and "other") but conventional hydroelectric power for all sectors in the United States from January 2001 to January 2025. The data is freely

available from the U.S. Energy Information Administration's (EIA) website<sup>2</sup> (U.S. Energy Information Administration 2025). The last 30 observations, from August 2022 to January 2025, were reserved for out-of-sample forecasting purposes. Hence, only the data from January 2001 to July 2022 were used to fit the model, resulting in a sample of size  $n = 259$  referred to as the "observed time series". The values from the observed time series range from 0.8968 to 0.9585, with a mean of 0.9326, a median of 0.9343, and a standard deviation of 0.0127. This strong concentration near the upper boundary of the unit interval matches the modeling scenario that motivated the proposed MARMA specification. A time series plot is presented in Figure 7a, whereas Figure 7b and c show the respective sample autocorrelation and partial autocorrelation function (ACF and PACF, respectively), and Figure 7d presents a seasonal plot.

The plots clearly reveal an annual seasonal pattern, arising from interactions among natural resource availability (sunlight, wind, water), weather-driven demand fluctuations (heating and cooling), and human activity cycles (holiday travel, harvesting, maintenance shutdowns). This motivates the explicit consideration of seasonal dynamics in the systematic component of the model. Hydropower production, for instance, follows an annual cycle with peaks in spring and rainy seasons but declines during dry periods and winter. Conversely, electricity demand typically surges in winter (due to heating needs) and in summer (for cooling). Different energy sources may increase or decrease their output as needed to meet energy demand.

Two primary methodologies exist for modeling seasonal time series. The first employs deterministic covariates, incorporating sine and cosine functions to capture annual seasonal patterns. In contrast, the second approach models the data directly without relying on such covariates. In our



**FIGURE 7** | Monthly proportion of net electricity generated by all means but conventional hydroelectric power for all sectors in the United States. (a) Observed time series, (b) Sample ACF, (c) Sample PACF, (d) Seasonality.

analysis, we consider both approaches for completeness. To showcase the versatility of the proposed model, we compare the MARMA against established benchmark models—the  $\beta$ ARMA and KARMA—which consider bi-parametric conditional distributions and are recognized for their flexibility. The  $\beta$ ARMA model was the first widely known GARMA-like model, whereas the KARMA is popular in the hydrological literature, often being praised as a more robust alternative to the  $\beta$ ARMA. We also include a comparison with traditional SARMA models, implemented by applying a logit transformation to the data, conducting the analysis, and back-transforming the results to the  $(0, 1)$  domain.<sup>3</sup>

## 6.2 | Parameter Estimation, Inference and Model Identification

Figure 7 shows an obvious 12-month cycle, and since we are working with monthly data, the autoregressive and moving average orders were constrained to a maximum of 12 lags (i.e.,  $p \leq 12$  and  $q \leq 12$ ), allowing some of the coefficients to be absent from the model. The seasonality in the data is addressed using two different approaches: the first one explores the AR and MA structures in the systematic component can accommodate the seasonal cycles without the aid of external regressors, similar to classical SARMA modeling; in the second one, deterministic harmonic covariates  $X_t = (\sin(2\pi t/12), \cos(2\pi t/12))'$  are added to the systematic component. This evaluation examines the model's adaptability in these different contexts. We also considered the logit, loglog, and cloglog link functions to identify the most appropriate specification for the conditional mean.

Model selection employed a bidirectional stepwise approach. To guide this process, we defined two distinct significance thresholds: a predictor was removed if its  $p$ -value exceeded 0.11 ( $\alpha_{\text{remove}}$ ), while a previously excluded predictor was reentered if its  $p$ -value fell below 0.10 ( $\alpha_{\text{enter}}$ ). These relaxed thresholds, higher than the conventional 0.05, aim to retain potentially meaningful variables and prevent the model from becoming overly simplistic. This asymmetric design also prevents infinite cycling between model states. Iteration continued until no further significant changes occurred, with decisions at each step informed by the Wald test (Section 4.1). The procedure starts with full MARMA(12, 12) specification.

For parameter estimation, we used the BTR package and implemented a dual-optimization strategy to ensure numerical stability. Two configurations were run in parallel: one prioritized the L-BFGS-B algorithm with Nelder-Mead as a fallback, while the other reversed this order. This ensured that if one optimizer failed to converge, the selection process could continue using the alternative. To maintain consistency, this entire selection and estimation framework - including the stepwise procedure, thresholds, and dual-optimization strategy - was identically applied to the  $\beta$ ARMA and KARMA models for comparison.

For SARMA modeling, the initial step involved utilizing the `auto.arima` function from the R package `forecast` (Hyndman and Khandakar 2008) to derive a preliminary model. We set the maximum lag parameters for AR, MA, SAR, and SMA

(`max.p`, `max.q`, `max.P`, and `max.Q` respectively) to 12, without differencing (i.e., `d` and `D` were set to 0), and forced the algorithm to explore all possible models (i.e., `stepwise` was set to `FALSE`). Furthermore, `approximation` was set to `FALSE`, and `nmodels` to  $12^4 = 20,736$ . See the package's documentation for details.

The model selected by this initial procedure was a simple stationary ARMA(2, 1) model, which subsequently failed all goodness-of-fit tests. This outcome reflects a primary limitation of the `auto.arima` function: its search is restricted to complete models.<sup>4</sup> For instance, it does not consider an AR(3) model where  $\phi_1 = \phi_2 = 0$ . Following this initial step, we proceeded with a Box and Jenkins approach, guided by the ACF and PACF, until we arrived at the final model, for which all coefficients were significant, and the residuals were uncorrelated, as indicated by the Ljung-Box test at the 5% level.

In-sample goodness-of-fit was assessed using the root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean directional accuracy (MDA). The MDA metric is particularly relevant for this analysis as it evaluates the model's capacity to correctly predict the direction of changes in the time series—crucial for capturing trend reversals. Higher MDA values indicate a stronger alignment between the forecasted and observed directional movements. The results are discussed in the sequel.

### 6.2.1 | Model Complexity and Numerical Stability

We evaluated 36 model configurations considering MARMA, KARMA, and  $\beta$ ARMA models, three link functions (logit, loglog, cloglog), two approaches for modeling the seasonality (with/without harmonic regressors), and two optimization algorithms (L-BFGS-B and Nelder-Mead available from the BTR package) and the SARMA modeling. The stability and final specifications of the selected models for the MARMA, KARMA, and  $\beta$ ARMA models are reported in Table B1 (Appendix B). The table first addresses algorithmic reliability, reporting any convergence “Failures” during the stepwise procedure and the final estimation “Status”. It then profiles the complexity and significance of the final models through four key metrics: the ARMA order ( $p, q$ ), the total number of regressors ( $r$ ), the subset of those that are statistically significant ( $s$ ), and the highest  $p$ -value ( $p_{\text{max}}$ ) found among all estimated parameters.

In terms of numerical stability, MARMA is the most reliable one, with only one moderate optimization issue (a single *Fallback*). KARMA crashes systematically with all but one configuration (104–274 intermediate failures), while  $\beta$ ARMA consistently succeeds with L-BFGS-B but struggles with Nelder-Mead (2/6 complete failures). No optimizer universally dominates; MARMA achieves balanced performance across both. The SARMA modeling presented no numerical issues. Model complexity follows the stability pattern. MARMA consistently selects simple autoregressive structures with 4–5 significant parameters ( $s$ ), and typically one regressor ( $r = 1$ ) or none, unlike benchmarks which often retain the full set ( $r = 2$ ). KARMA favors high-order AR terms ( $p = 11$  or  $12$ ) while  $\beta$ ARMA requires substantial MA components ( $q = 12$  in 5 out of 12 cases) and

from  $s = 8$  to  $s = 12$  parameters in most cases. As for the fitted SARMA model, the final model was in fact an ARMA(2, 12) with a total of  $s = 6$  significant parameters with  $p_{max} = 0.0024$ : for the MA part, only  $\theta_1$ ,  $\theta_{11}$ , and  $\theta_{12}$  are statistically significant, whereas both AR parameters were significant. This implies that the fitted SARMA model is more parsimonious than KARMA and  $\beta$ ARMA and requires higher-order MA terms, similar to  $\beta$ ARMA.

As for stationarity, the fitted SARMA model is stationary, as indicated by the roots of its characteristic function having an absolute value of 1.3694. Although stationarity conditions for GARMA and GARMA-like models based on continuous distributions are not known apart from a few particular cases (Pumi et al. 2021; Benaduce and Pumi 2023), it is customary to check for the presence of unitary roots in the model's AR part. Table B2 (Appendix B) reports the absolute value for the characteristic root closest to the unit for each of the 36 estimated models. Overall, MARMA with regressors tends to produce roots further away from the unit circle, whereas KARMA and  $\beta$ ARMA more often exhibit near-unit-root behavior, reinforcing the robustness-versus-instability pattern suggested by the failure and complexity diagnostics. Moreover, this highlights the versatility of the MARMA, which was able to model the time series more parsimoniously than competitors while avoiding unit roots.

### 6.2.2 | Goodness-of-Fit and In-Sample Accuracy

Table 5 reports goodness-of-fit metrics (AIC, BIC, HQC, Log-Lik) and in-sample accuracy statistics (RMSE, MAPE, MDA) for the models with the best result (highlighted in bold)

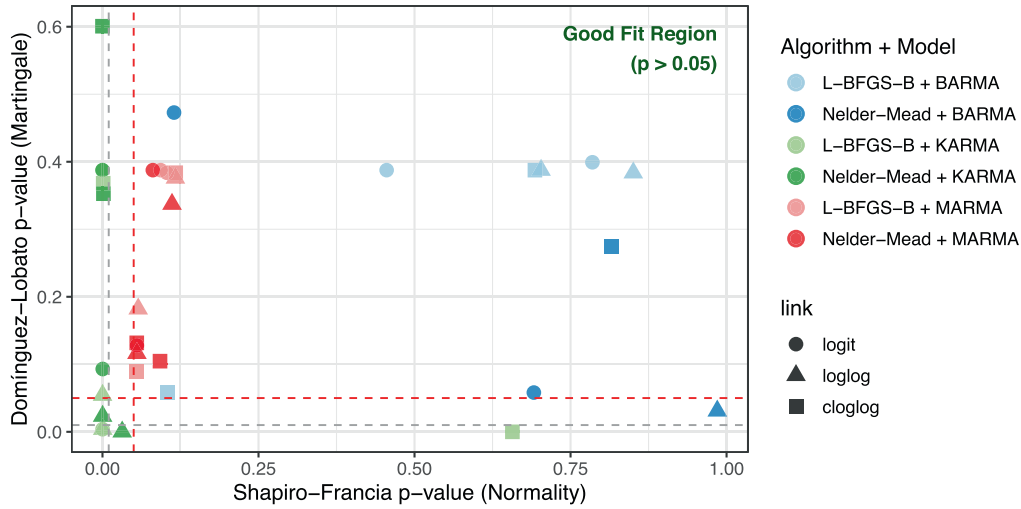
within each model family. Additionally, the column DF/SF indicates whether each model has passed the Martingale Difference (DF) and Shapiro-Francia (SF) tests. Complete results for the other 36 GARMA models can be found in Table B3 (Appendix B).

A direct inspection of the information criteria reveals a distinct pattern: the benchmark models (KARMA and  $\beta$ ARMA) consistently achieve significantly higher log-likelihood values (implying lower AIC, BIC, and HQC values) compared to the MARMA specifications. For instance, the best-fitting  $\beta$ ARMA model (loglog + no-regressors) attained an AIC of  $-1851.0$ , whereas the optimal MARMA configuration (loglog + no-regressors) reached  $-1012.4$ . It is also worth noting that, within each model family, the results were relatively close across different link functions (see Table B3), suggesting limited influence of the link choice on the overall goodness-of-fit. The worst information criteria values are the SARMA's by a good margin.

To contextualize the numerical results, Figure 8 provides scatter plots comparing two test  $p$ -values: the Shapiro-Francia (SF) normality test for quantile residuals (horizontal axis) and the Dominguez-Lobato (DL) martingale difference test for simple residuals (vertical axis) for the GARMA models. In the plot, points are color-coded according to the combination of model family and optimization algorithm, while the link function is identified by the shape of the markers. The dashed vertical and horizontal lines represent the significance levels of  $\alpha = 0.05$  in red and  $\alpha = 0.01$  in gray, defining a "good fit region" in the upper-right quadrant: models located within this area fail to reject both null hypotheses, indicating that they satisfy the assumptions of normality and uncorrelatedness required for valid inference. Since the fitted values of all models presented a

**TABLE 5** | Goodness-of-fit and in-sample accuracy measures for the models with the best result (highlighted in bold) for each metric within each model family. **Reg?** denotes inclusion ("Yes") or exclusion ("No") of harmonic regressors; **Alg.** indicates the optimization algorithm (L = L-BFGS-B; NM = Nelder-Mead); **DL/SF** summarizes the diagnostic results for Dominguez-Lobato (Martingale) and Shapiro-Francia (Normality) tests, where symbols  $\checkmark$  and  $\times$  indicate pass (i.e.,  $p$ -value  $\geq 0.05$ ) or failure ( $p$ -value  $< 0.05$ ), respectively.

Model	Link	Reg?	Alg.	Valid? DL/SF	Goodness-of-fit				In-sample accuracy		
					AIC	BIC	HQC	LogLik	RMSE	MAPE	MDA
<b>MARMA</b>	logit	No	L	$\checkmark/\checkmark$	-1012.2	-998.0	-1013.4	510.1	<b>0.0070</b>	<b>0.0060</b>	0.7132
	loglog	No	L	$\checkmark/\checkmark$	-1012.2	<b>-1001.6</b>	-1013.1	509.1	0.0084	0.0074	0.5775
	loglog	No	NM	$\checkmark/\checkmark$	<b>-1012.4</b>	-998.2	<b>-1013.6</b>	510.2	0.0072	0.0060	0.6977
	cloglog	Yes	L	$\checkmark/\checkmark$	-1012.2	-997.9	-1013.3	510.1	0.0078	0.0065	<b>0.7403</b>
	cloglog	Yes	NM	$\checkmark/\checkmark$	-1012.2	-998.0	-1013.3	510.1	0.0077	0.0064	<b>0.7403</b>
	cloglog	No	NM	$\checkmark/\checkmark$	-1010.6	-992.8	-1012.0	<b>510.3</b>	0.0071	0.0061	0.7132
<b>KARMA</b>	logit	Yes	L	$\times/\times$	-1808.9	<b>-1784.0</b>	-1810.9	911.5	0.0070	0.0059	0.7093
	logit	No	NM	$\checkmark/\times$	-1795.9	-1771.0	-1797.9	905.0	0.0072	0.0061	<b>0.7364</b>
	loglog	No	L	$\times/\times$	<b>-1813.9</b>	-1781.9	<b>-1816.4</b>	<b>915.9</b>	<b>0.0069</b>	<b>0.0059</b>	0.7248
<b><math>\beta</math>ARMA</b>	logit	Yes	L	$\checkmark/\checkmark$	-1828.9	-1786.2	-1832.3	926.4	<b>0.0065</b>	0.0055	0.7597
	logit	No	L	$\checkmark/\checkmark$	-1752.8	-1717.3	-1755.7	886.4	0.0065	0.0055	<b>0.7829</b>
	loglog	Yes	L	$\checkmark/\checkmark$	-1829.9	-1790.8	-1833.1	926.0	0.0065	<b>0.0054</b>	0.7287
	loglog	No	NM	$\times/\checkmark$	<b>-1851.0</b>	<b>-1822.5</b>	<b>-1853.2</b>	<b>933.5</b>	0.0066	0.0057	0.7597
<b>SARMA</b>	logit	No	—	$\checkmark/\times$	-465.91.	-441.01	-425.03	240.0	0.0062	0.0052	0.7259



**FIGURE 8** | Residual Diagnostics:  $p$ -values from Normality versus Martingale Difference tests.

pronounced outlier at  $t = 1$  due to initialization, the tests were conducted after removing this observation.

Visual inspection of Figure 8 combined with the results in Table B3, provides some interesting findings. Regarding serial dependence, MARMA and  $\beta$ ARMA models consistently satisfy the martingale difference null hypothesis (DL test) across all configurations ( $\beta$ ARMA fails at 5% but passes at 1% in one configuration). Interestingly, KARMA models passed the test for 7 out of 12 configurations, despite  $\mu_t$  representing the median in that framework. However, regarding normality, the KARMA models systematically failed the SF test ( $p < 0.003$ ) across 11 out of 12 configurations, presenting significant departure from normality. In contrast, the MARMA and the  $\beta$ ARMA models present comparable performance, with valid residuals observed at all configurations. SARMA’s residuals failed the SF test ( $p$ -value 0.0241) but passed the DL test.

Regarding in-sample accuracy, the SARMA presented slightly lower values of RMSE and MAPE compared to other GARMA models. Among GARMA models, the  $\beta$ ARMA is the one presenting the overall best results in terms of RMSE, MAPE, and MDA values. The MARMA exhibited higher error metrics, probably as a consequence of its parsimony. Therefore, the choice between MARMA, SARMA, and  $\beta$ ARMA extends beyond mere validity to efficiency. While the  $\beta$ ARMA requires a highly complex structure ( $s \approx 7-12$  parameters) and the inclusion of deterministic regressors to model the data, the MARMA model does so with a significantly more parsimonious structure ( $s \approx 3-5$ ) and often without the need for exogenous covariates. Notably, the MARMA configuration with the lowest in-sample RMSE (0.0070) was the logit model without deterministic regressors. In energy planning applications, this parsimony enables clearer interpretation of seasonal effects and simpler model updates as new data becomes available.

### 6.3 | Out-of-Sample Forecasting Performance

Arguably, the most important objective in time series analysis is forecasting, out-of-sample performance the ultimate benchmark

for model comparison. In this section, we consider out-of-sample forecasts for the 30 holdout observations (August 2022–January 2025). Forecasts are obtained using only information available up to the estimation sample end (December 2023), considering all model configurations from Section 6.1.

More explicitly, for each fitted model,  $h$ -step-ahead predictions  $\hat{Y}_{n+h}$ , with  $h \in \{1, \dots, 30\}$ , are computed using only information from the observed sample. The hold-out sample is used exclusively to calculate the accuracy measures. Cumulative forecast accuracy up to horizon  $h$  is assessed using the RMSE, MAPE, and MDA, defined as follows for completeness. Let  $\{Y_{n+h}\}_{h=1}^{30}$  and  $\{\hat{Y}_{n+h}\}_{h=1}^{30}$  denote the observed (hold-out) and predicted values, respectively, then we define

$$\text{RMSE}(h) = \sqrt{\frac{1}{h} \sum_{k=1}^h (Y_{n+k} - \hat{Y}_{n+k})^2},$$

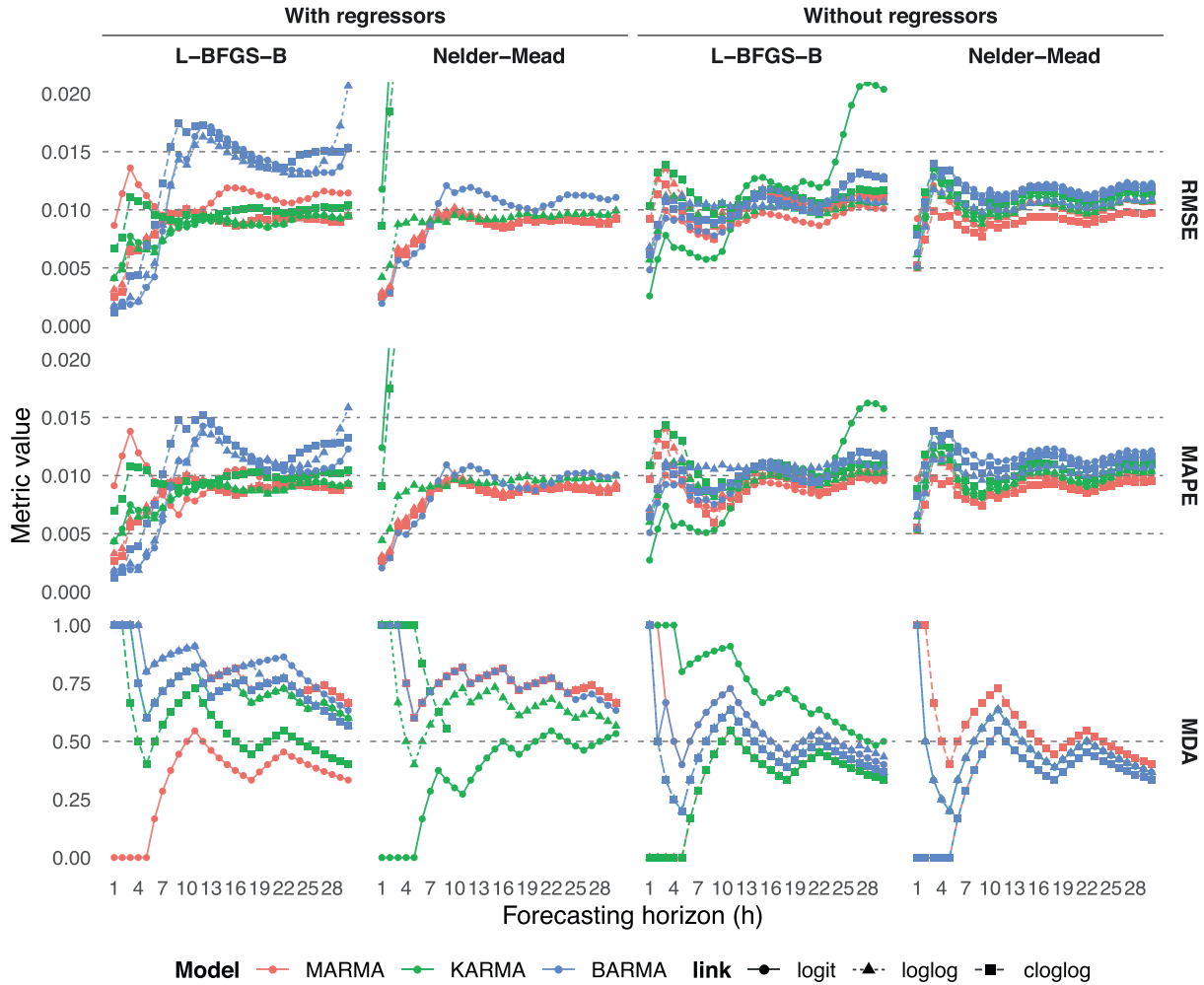
$$\text{MAPE}(h) = \frac{1}{h} \sum_{k=1}^h \left| \frac{Y_{n+k} - \hat{Y}_{n+k}}{Y_{n+k}} \right|,$$

and

$$\text{MDA}(h) = \frac{1}{h} \sum_{k=1}^h I(\text{sign}(Y_{n+k} - Y_n) = \text{sign}(\hat{Y}_{n+k} - Y_n)),$$

for  $h \in \{1, \dots, 30\}$ . Hence, for any  $h$ ,  $\text{RMSE}(h)$ ,  $\text{MAPE}(h)$  and  $\text{MDA}(h)$  evaluate cumulative performance over  $h$  out-of-sample forecasts.

Considering every fitted GARMA model in Section 6.1 addresses four fundamental questions regarding model specification effects on forecasting performance. First, the influence of link functions (logit, loglog, cloglog), optimization algorithms (L-BFGS-B, Nelder-Mead), and regressor inclusion on predictive accuracy within each model family. Second, whether models with superior in-sample performance (by likelihood criteria and error metrics) maintain superiority in out-of-sample forecasting over the 30-month hold-out. Third, whether residual diagnostic failures (martingale difference, normality) affect forecast performance.



**FIGURE 9** | Out-of-sample RMSE, MAPE, and MDA (accumulated values) computed over the  $h$ -step-ahead forecasts values, with  $h \in \{1, \dots, 30\}$ , across all MARMA, KARMA, and  $\beta$ ARMA configurations (model-link-algorithm combinations).

Fourth, performance differences across forecast horizons, categorized as short (1–6 months), medium (7–12 months), and long (> 12 months).

Figure 9 presents out-of-sample forecasting performance across horizons  $h \in \{1, \dots, 30\}$  for 36 model configurations.  $RMSE(h)$ ,  $MAPE(h)$ , and  $MDA(h)$  metrics are shown in separate panels (top to bottom). Model families (MARMA, KARMA,  $\beta$ ARMA) are distinguished by color, while link functions (logit, loglog, cloglog) are represented through linetypes and point shapes. Panels are organized by regressor inclusion (with/without) and optimization method (L-BFGS-B, Nelder-Mead). Horizontal reference lines appear at 0.005, 0.010, and 0.015 for RMSE/MAPE and at 0.5 for MDA. Missing panels indicate convergence failures requiring algorithm fallback. RMSE and MAPE axes are truncated at 0.02 to maintain visualization clarity, as extreme values from two KARMA configurations would otherwise dominate the scale. Similar plot considering the SARMA model along with the best and worst MARMA, KARMA, and  $\beta$ ARMA is presented in Figure B2 (Appendix B).

Figure B1 (Appendix B) presents the same information as Figure 9 in bar plot format, with bars filled by link +

algorithm combination (light colors = L-BFGS-B, dark colors = Nelder-Mead), faceted by model family and regressors, providing a complementary visualization for detailed comparative analysis. Table 6 reports the mean and standard deviation of  $RMSE(h)$  values across horizons, serving as a baseline to contextualize the variability observed in the figures. For this table, L-BFGS-B (Nelder-Mead) entries marked “Failed” reflect convergence issues that lead to the same model as Nelder-Mead (L-BFGS-B). These are the same scenarios suppressed from Figures 9 and B1. Among the 36 configurations, we highlight the top performer per family: MARMA (loglog/cloglog, Nelder-Mead, regressors), KARMA (logit/loglog, L-BFGS-B, regressors),  $\beta$ ARMA (logit, Nelder-Mead, regressors). For reference, the mean and standard deviation of the  $RMSE(h)$  for the SARMA model are 0.0111 and 0.0022, respectively.

### 6.3.1 | The Influence of Link, Algorithm, and Regressors

First, the influence of link functions (logit, loglog, cloglog), optimization algorithms (L-BFGS-B, Nelder-Mead), and regressor inclusion on predictive accuracy is evident from the RMSE,

**TABLE 6** | Mean and standard deviation (in parentheses) of the 30 RMSE( $h$ ) values, by model family, regressor inclusion, algorithm, and link function. The smallest/largest mean, by model family, is highlighted in **bold/italics**.

Model	Reg?	Algorithm	Link		
			logit	loglog	cloglog
MARMA	yes	L-BFGS-B	0.0109 (0.0011)	0.0086 (0.0016)	0.0085 (0.0018)
		Nelder-Mead	0.0085 (0.0017)	0.0086 (0.0017)	<b>0.0083 (0.0018)</b>
	no	L-BFGS-B	0.0090 (0.0009)	<i>0.0110 (0.0011)</i>	0.0103 (0.0011)
		Nelder-Mead	0.0101 (0.0008)	0.0099 (0.0012)	0.0089 (0.0009)
KARMA	yes	L-BFGS-B	<b>0.0084 (0.0013)</b>	<b>0.0084 (0.0014)</b>	0.0097 (0.0008)
		Nelder-Mead	<i>0.1395 (0.1393)</i>	0.0090 (0.0012)	NaN*
	no	L-BFGS-B	0.0115 (0.0051)	0.0098 (0.0010)	0.0113 (0.0009)
		Nelder-Mead	0.0104 (0.0013)	0.0101 (0.0010)	0.0111 (0.0010)
$\beta$ ARMA	yes	L-BFGS-B	0.0120 (0.0051)	0.0120 (0.0050)	<i>0.0130 (0.0046)</i>
		Nelder-Mead	<b>0.0097 (0.0027)</b>	Failed	Failed
	no	L-BFGS-B	0.0103 (0.0020)	0.0104 (0.0008)	0.0105 (0.0015)
		Nelder-Mead	0.0117 (0.0012)	0.0103 (0.0012)	0.0116 (0.0011)

Note: \* NaN indicates that the model failed to produce predictions after  $h = 10$ .

MAPE, and MDA trajectories in Figure 9, from the detailed RMSE bars in Figure B1, and from the mean/SD RMSE summary in Table 6. The proposed MARMA family presents the overall best performance (means 0.0083–0.0110). Its lowest RMSE is obtained with cloglog + Nelder-Mead + regressors (0.0083). The configuration exhibiting the poorest mean performance (0.0110) is loglog without regressors using L-BFGS-B. No-regressor configurations systematically display lower variability (SDs 0.0008–0.0012) than those with regressors (SDs 0.0011–0.0018). The results show that the inclusion of regressors consistently reduces RMSE across all MARMA configurations, despite increasing variability.

KARMA shows the widest performance range (means 0.0084–0.1395). Its best configurations—logit/loglog with regressors + L-BFGS-B (0.0084)—rival MARMA's accuracy, but others degrade severely, such as logit + regressors + Nelder-Mead (0.1395). The cloglog link with Nelder-Mead and regressors produces forecasts only up to  $h = 10$ . For every link-algorithm combination where valid predictions exist for both cases, regressor inclusion yields lower RMSE.  $\beta$ ARMA shows intermediate accuracy (means 0.0097–0.0130). In the two configurations where Nelder-Mead fails with regressors (loglog and cloglog links), no valid comparison is possible; however, in all remaining cases where the model converges, configurations with regressors uniformly outperform their no-regressor counterparts (except for the logit link with Nelder-Mead). The best result is achieved with logit + regressors + Nelder-Mead (0.0097).

MARMA, KARMA, and  $\beta$ ARMA display markedly different patterns across the three forecasting metrics over the 30-step horizon (Figure 9). In terms of RMSE, MARMA consistently shows the lowest and most stable trajectories, with values ranging from 0.0025 to 0.0136. The best-performing MARMA curves lie clearly below those of  $\beta$ ARMA (0.0011–0.0207) and, especially, KARMA (0.0026–0.3650). The high mean RMSE of KARMA is driven by a few configurations with extreme errors, notably the logit link

with regressors under Nelder-Mead optimization. With an RMSE of 0.0111, SARMA ranks among the weaker performers. It only surpasses the worst KARMA model and is outperformed by every MARMA model, though it does exceed all but the best  $\beta$ ARMA configuration.

The behavior of MAPE mirrors this ranking. MARMA maintains the lowest percentage errors (0.0026–0.0140, mean 0.0089), followed by  $\beta$ ARMA (0.0012–0.0158, mean 0.0103). KARMA shows the highest MAPE on average (0.0027–0.2270, mean 0.0174), with pronounced deterioration in specific link-regressor combinations, confirming that configurations with elevated RMSE also yield higher percentage errors.

Regarding directional accuracy, MDA exhibits clear differences across model families and specifications. MARMA with regressors and loglog/cloglog link achieves the most consistent performance (mean MDA = 0.741, SD = 0.048). Without regressors, MARMA's accuracy decreases substantially (mean MDA = 0.421–0.466). KARMA shows intermediate performance with notable variability. With regressors, mean MDA ranges from 0.549 (cloglog) to 0.664 (loglog), while without regressors it drops to 0.394–0.443.  $\beta$ ARMA with regressors delivers the highest directional accuracy overall, particularly with logit/loglog link (mean MDA = 0.768). However, without regressors,  $\beta$ ARMA's performance declines to levels similar to MARMA (mean MDA = 0.421–0.486). Across all models, the inclusion of regressors improves mean MDA by 0.20–0.35 points, with the largest gains observed for  $\beta$ ARMA ( $\Delta \approx 0.28$ –0.31) and the most stable improvement for MARMA ( $\Delta \approx 0.27$ –0.32). The loglog link generally provides the best MDA performance when regressors are included, while cloglog shows more variability across configurations. SARMA performs competitively with the best models up to horizon  $h = 12$ , but its performance deteriorates beyond  $h = 14$ , eventually falling below 0.5 after  $h = 26$ .

### 6.3.2 | In-Sample Versus Out-of-Sample Performance

Second, models presenting the best in-sample performance by maximum log-likelihood and information criteria do not consistently remain superior in the 30-month hold-out when evaluated by RMSE. Table 5 identifies  $\beta$ ARMA loglog without regressors under Nelder-Mead as the overall in-sample winner (highest log-likelihood: 933.5, lowest AIC:  $-1851.0$ ), while for MARMA the best AIC ( $-1012.4$ ) is achieved by loglog without regressors under Nelder-Mead, and the best in-sample RMSE (0.0070) by logit without regressors under L-BFGS-B. For KARMA, loglog without regressors under L-BFGS-B has the best AIC ( $-1813.9$ ) and best in-sample RMSE (0.0069). However, out-of-sample results in Table 6 reveal significant rank reversals.

For MARMA, neither the best AIC configuration (loglog + no-regressors + Nelder-Mead, out-of-sample mean RMSE 0.0099) nor the best in-sample RMSE configuration (logit + no-regressors + L-BFGS-B, out-of-sample mean RMSE 0.0090) achieves the lowest out-of-sample RMSE. Instead, MARMA cloglog with regressors under Nelder-Mead, which ranks only fifth in AIC within its family, delivers the best out-of-sample performance (RMSE 0.0083).

Similarly,  $\beta$ ARMA's best in-sample configuration (loglog + no-regressors + Nelder-Mead) shows poorer out-of-sample accuracy (RMSE 0.0103), being outperformed by configurations that ranked lower in-sample, particularly logit with regressors under Nelder-Mead (RMSE 0.0097). SARMA exemplifies this pattern even more starkly: despite achieving the best in-sample performance (RMSE 0.0062, MAPE 0.0052), it presents poor out-of-sample forecasting accuracy.

KARMA exhibits the most extreme discrepancies. Its best in-sample configuration (loglog + no-regressors + L-BFGS-B) achieves only moderate out-of-sample performance (RMSE 0.0098), while configurations with much poorer in-sample fits, such as logit with regressors under Nelder-Mead (in-sample RMSE 0.0101), produce catastrophic out-of-sample forecasts (RMSE 0.1395). Notably, KARMA logit and loglog with regressors under L-BFGS-B achieve the best out-of-sample RMSE (0.0084) despite not being the in-sample optimal configurations.

Thus, in-sample optimality poorly predicts hold-out forecasting superiority for KARMA,  $\beta$ ARMA, and SARMA, with second-ranked models often outperforming nominal "winners", consistent with overfitting in maximum-likelihood selection. MARMA, however, shows robust performance regardless of the evaluation criterion.

### 6.3.3 | Goodness-of-Fit Versus Forecasting Performance

Third, residual diagnostic failures (martingale difference, normality) do not systematically hinder out-of-sample forecasts. Table B3 shows diagnostic validation (DL/SF) varies widely across configurations, with  $\checkmark$  indicating passed tests ( $p \geq 0.05$ ) and  $\times$  failures, yet Table 6 reveals no clear pattern linking diagnostic status to RMSE performance.

Top out-of-sample performers include MARMA cloglog + regressors + Nelder-Mead (RMSE 0.0083, DL  $\checkmark$ /SF  $\checkmark$ ) alongside MARMA logit + regressors + Nelder-Mead (RMSE 0.0085, DL  $\checkmark$ /SF  $\checkmark$ ). Similarly, KARMA's best RMSE configuration (logit + regressors + L-BFGS-B, 0.0084) fails both DL and SF ( $\times$ / $\times$ ), while another strong KARMA performer (loglog + regressors + L-BFGS-B, RMSE 0.0084) passes DL but fails SF ( $\checkmark$ / $\times$ ).

$\beta$ ARMA's top out-of-sample performer (logit + regressors + Nelder-Mead, RMSE 0.0097) passes both diagnostic tests ( $\checkmark$ / $\checkmark$ ), yet other  $\beta$ ARMA configurations with similar diagnostic status show substantially higher RMSE. Diagnostic failures thus appear unrelated to forecasting ability: models passing both DL/SF do not systematically outperform those failing one or both, suggesting that the MARMA framework produces reliable hold-out forecasts even when in-sample residuals exhibit mild martingale difference or normality violations.

### 6.3.4 | Performance Across Forecast Horizons

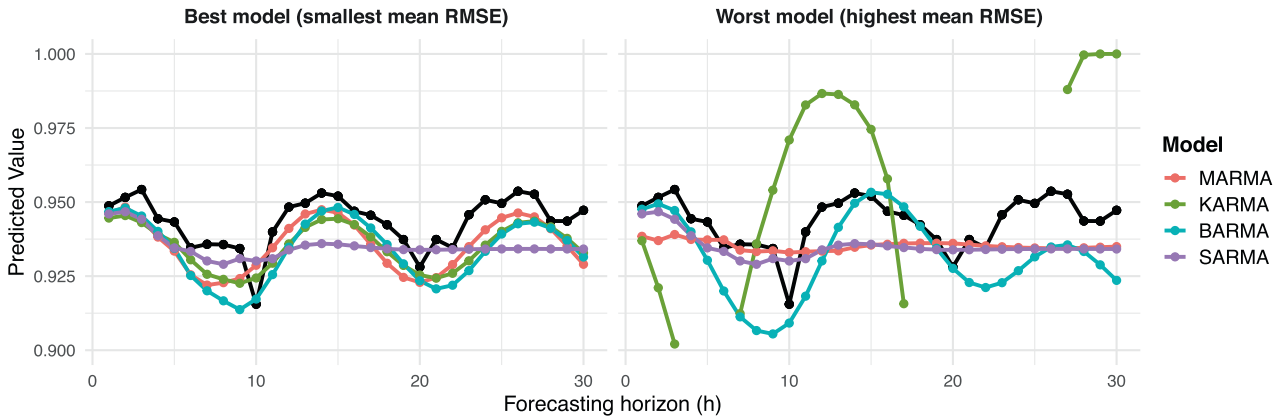
Fourth, performance differences across forecast horizons reveal MARMA's consistent superiority. Figure 9 trajectories show MARMA maintaining the lowest RMSE across most horizons, with values ranging from 0.0025 to 0.0136 across all configurations—the narrowest range among the three families.  $\beta$ ARMA displays higher and more volatile trajectories, spanning 0.0011 to 0.0207, often reaching 0.015–0.018 at longer horizons. In terms of RMSE, the SARMA model is competitive between lags 6 and 12, falling behind after that.

KARMA exhibits extreme variability: while its minimum RMSE (0.0026) is competitive, its maximum (0.3650) is an order of magnitude higher than other models, driven by two pathological cases: logit + regressors + Nelder-Mead and cloglog + regressors + Nelder-Mead (predictions unavailable beyond  $h = 10$ ). This wide dispersion contrasts sharply with MARMA's stability.

Table 6 confirms MARMA's superior consistency: its mean RMSE ranges from 0.0083 to 0.0110, outperforming  $\beta$ ARMA (0.0097–0.0130) and KARMA (0.0084–0.1395). The latter's extreme maximum reflects the same unstable configurations visible in Figure 9. The results suggest that MARMA presents a robust forecasting stability regardless of horizon length, which was unmatched by the GARMA benchmarks, which lose accuracy as uncertainty accumulates or exhibit high volatility. This superior stability is particularly noteworthy given that the MARMA model is based on a uniparametric conditional distribution.

## 6.4 | Prediction Intervals

Figure 10 presents the observed time series and out-of-sample forecasts based on the MARMA, KARMA, and  $\beta$ ARMA families. For this figure, we consider the models reported in Table 7, which correspond to the best (left) and worst (right) performing models for each family, in terms of out-of-sample mean RMSE (see Table 6). When multiple candidate models exist for a family, we select the model whose AR polynomial roots are furthest from 1 (see Table B2). For visual clarity, the y-axis is limited to the  $[0, 1]$



**FIGURE 10** | Observed time series (black) and fitted values (out-of-sample forecast) for the SARMA model along with the best (left) and worst (right) performing MARMA, KARMA, and  $\beta$ ARMA models.

**TABLE 7** | Summary for the best (left) and worst (right) performing MARMA, KARMA and  $\beta$ ARMA models, in terms of out-of-sample RMSE: link function, optimization algorithm, estimated coefficients and respective standard errors (in parenthesis), RMSE(30), MAPE(30) and MDA(30).

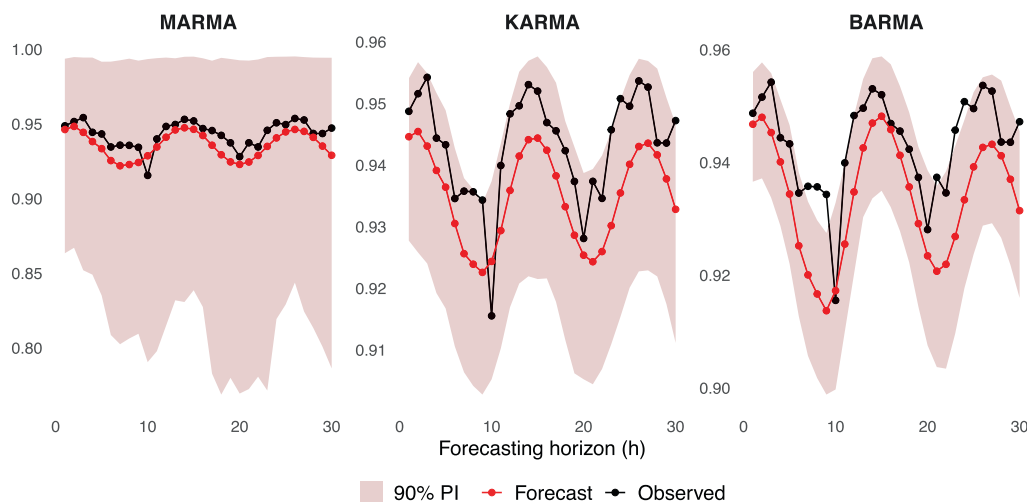
	Model					
	MARMA		KARMA		$\beta$ ARMA	
Link	cloglog	loglog	logit	logit	logit	cloglog
Alg.	Nelder-Mead	L-BFGS-B	Nelder-Mead	Nelder-Mead	L-BFGS-B	L-BFGS-B
$\alpha$	0.882 (0.255)	-2.264 (0.572)	2.616 (0.081)	3.383 (0.181)	2.476 (0.083)	1.111 (0.041)
$\beta_1$	-0.068 (0.031)	—	-0.150 (0.014)	—	-0.190 (0.012)	-0.072 (0.005)
$\beta_2$	—	—	0.058 (0.013)	—	0.080 (0.011)	0.035 (0.005)
$\phi_1$	0.414 (0.245)	0.489 (0.195)	0.634 (0.038)	1.461 (0.078)	0.443 (0.031)	0.467 (0.042)
$\phi_2$	—	—	-0.360 (0.032)	-0.574 (0.066)	—	—
$\phi_4$	—	—	—	-0.228 (0.044)	—	—
$\phi_6$	—	—	—	—	0.110 (0.035)	0.070 (0.039)
$\phi_7$	—	-0.327 (0.105)	—	-0.359 (0.050)	-0.120 (0.048)	-0.103 (0.049)
$\phi_8$	—	—	—	—	-0.135 (0.050)	-0.183 (0.049)
$\phi_9$	-0.289 (0.095)	—	-0.258 (0.013)	-0.163 (0.080)	-0.325 (0.050)	-0.251 (0.049)
$\phi_{10}$	—	—	—	0.296 (0.077)	0.092 (0.035)	-0.121 (0.037)
$\phi_{12}$	—	—	—	-0.710 (0.041)	—	—
$\theta_1$	—	—	—	—	—	0.145 (0.084)
$\nu$	—	—	145.40 (7.045)	74.582 (3.704)	1412.28 (124.15)	1007.66 (88.593)
RMSE	0.009	0.012	0.009	0.335	0.011	0.015
MAPE	0.009	0.011	0.009	0.201	0.010	0.013
MDA	0.667	0.333	0.600	0.533	0.633	0.567

interval, as the worst-performing KARMA model produced forecasts far outside the range of the observed curve. Table 7 reports the optimization algorithm, link function, significant parameters, RMSE(30), MAPE(30), and MDA(30) for each model.

From Table 7, we can highlight some patterns emerging across the best and worst configurations within each model family. First, almost always a predictor (covariate or autoregressive term) that appears in more than one specification retains the same sign

across all models in which it is selected, with the exception of  $\alpha$ . The first-order autoregressive coefficient  $\phi_1$  is the only one significant in every reported specification and is consistently positive. Notably, a moving-average (MA) term was selected only for the worst  $\beta$ ARMA model.

Regarding out-of-sample performance, a sharp contrast is evident. The RMSE(30) and MAPE(30) of the worst-performing configurations for MARMA and  $\beta$ ARMA remain comparable to



**FIGURE 11** | Observed time series (black), predicted values (red) forecasts obtained with the fitted MARMA, KARMA and  $\beta$  ARMA (left to right) models and the respective 90% bootstrap prediction intervals (PI).

their best counterparts and to the best KARMA model. However, the worst-performing KARMA specification (logit, RMSE = 0.335) exhibits a forecasting error an order of magnitude larger. This outlier is the reason the  $y$ -axis in Figure 10 was limited to the  $[0.9, 1]$  interval for visual clarity, and it highlights a potential robustness issue with the KARMA family under certain specifications. Conversely, the MARMA models show stable performance even in their least favorable case, delivering forecasting accuracy on par with the best alternatives from other families and suggesting robustness of the proposed approach.

Figure 11 displays the 30-step-ahead forecasts for the best-performing model for each family. The black line represents the observed out-of-sample values, the red line shows the corresponding point forecasts, and the shaded area depicts the associated 90% bootstrap prediction intervals (PI), obtained using the methodology presented in Section 5.4, calculated from 1000 bootstrap samples. The point forecasts for all models track the general level and seasonal pattern of the observed series reasonably well. However, the prediction intervals for the KARMA and  $\beta$ ARMA models fail to cover the observed values at several horizons, indicating that their estimated uncertainty is too narrow. From an operational perspective, overly narrow prediction intervals may lead to an underestimation of uncertainty in future generation mixes, which can adversely affect planning and risk management in energy systems. In contrast, the wider bands of the MARMA model provide better empirical coverage for this forecasting exercise. The difference in band width is related to the conditional variance of the three models: in all cases, the PIs closely follow the band created using  $\pm 2\hat{\sigma}_{n+h}$ , where  $\hat{\sigma}_t^2$  is obtained by plugging  $\hat{\mu}_t$  (and  $\hat{\nu}$ ) into the expression of  $\text{Var}(Y_t | \mathcal{F}_{t-1})$  for each model.

## 7 | Conclusion

The present work proposed the Matsuoka autoregressive moving average (MARMA) model for time series taking values in the

interval  $(0, 1)$ . The idea is to consider a GARMA specification based on the fact that the inference for the proposed model was conducted using partial maximum likelihood. The Matsuoka distribution belongs to the exponential family in canonical form, so the asymptotic theory can be studied in detail. Inferential methods such as confidence intervals, residual analysis, model selection, and forecasting were also explored. In particular, a novel bootstrap-based method for constructing prediction intervals was introduced and studied. The proposed method can be applied to any GARMA model.

We provided a simulation study to assess the finite sample performance of the PMLE and the method for constructing prediction intervals. We also assess goodness-of-fit tests, including a martingale difference test and normality tests for residuals. Overall, the results were satisfactory, indicating that the estimators behaved as expected. Furthermore, the overall finite-sample performance of the proposed bootstrap prediction confidence interval was good, as the empirical coverages were close to the nominal levels for all forecasting horizons.

The proposed MARMA model was used to model and forecast the proportion of net electricity generated by all means but conventional hydroelectric power in the United States. To provide grounds for comparison and to showcase the proposed model's strengths, we also considered the KARMA,  $\beta$ ARMA, and logit-transformed SARMA models as benchmarks in the application. The results showcase the superior out-of-sample forecasting performance and parsimony of the proposed MARMA model. Compared to benchmark models, MARMA achieved the lowest RMSE and MAPE over a 30-month forecasting horizon, highlighting its predictive advantage. Remarkably, this performance was attained with only four significant parameters-fewer than those required by all competing models-emphasizing its flexibility and parsimony despite its uniparametric structure. While all models produced reasonably accurate forecasts, MARMA stood out for capturing the seasonal dynamics effectively with minimal complexity.

## Acknowledgments

The Article Processing Charge for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614).

## Funding

G. Pumi and T.S. Prass gratefully acknowledge the financial support received by the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - Bolsa de Produtividade em Pesquisa - Proc. 303281/2025-1 (Pumi) and 305886/2025-8 (Prass).

## Conflicts of Interest

The authors declare no conflicts of interest.

## Data Availability Statement

The data that support the findings of this study are available in the U.S. Energy Information Administration at <https://www.eia.gov/electricity/data/browser/topic/>. These data were derived from the following resources available in the public domain: U.S. Energy Information Administration, <https://www.eia.gov/electricity/data/browser/topic/>.

## Endnotes

<sup>1</sup><https://www.eia.gov/energyexplained/hydropower/>. retrieved 04/04/2025.

<sup>2</sup><https://www.eia.gov/electricity/data/browser/topic/>., retrieved 04/04/2025.

<sup>3</sup>The authors wish to emphasize that they do not recommend, support, or advocate for the use of unbounded-support models with transformed bounded time series. This comparison is included solely at the referee's request.

<sup>4</sup>We first applied `auto.arima` with default settings, which selected an ARMA(2,1) model. To mitigate the risk of incomplete model search under default settings, we adopted the more comprehensive approach presented in our methodology. While both approaches ultimately selected the same initial ARMA(2,1) model, we retain our configurations in the analysis to document that thorough exploration was conducted.

## References

Bayer, F. M., D. M. Bayer, and G. Pumi. 2017. "Kumaraswamy Autoregressive Moving Average Models for Double Bounded Environmental Data." *Journal of Hydrology* 555: 385–396.

Bayer, F. M., R. J. Cintra, and F. Cribari-Neto. 2018. "Beta Seasonal Autoregressive Moving Average Models." *Journal of Statistical Computation and Simulation* 88, no. 15: 2961–2981.

Benaduce, H. S., and G. Pumi. 2023. "SYMARFIMA: A Dynamical Model for Conditionally Symmetric Time Series With Long Range Dependence Mean Structure." *Journal of Statistical Planning and Inference* 225: 71–88.

Bengtsson, H. 2025. "matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)," R Package Version 1.5.0.

Benjamin, M., R. Rigby, and D. Stasinopoulos. 2003. "Generalized Autoregressive Moving Average Models." *Journal of the American Statistical Association* 98, no. 461: 214–223.

Casarin, R., L. Dalla Valle, and F. Leisen. 2012. "Bayesian Model Selection for Beta Autoregressive Processes." *Bayesian Analysis* 7, no. 2: 385–410.

Charles, A., O. Darné, and J. H. Kim. 2011. "Small Sample Properties of Alternative Tests for Martingale Difference Hypothesis." *Economics Letters* 110, no. 2: 151–154.

Choi, I. 1999. "Testing the Random Walk Hypothesis for Real Exchange Rates." *Journal of Applied Econometrics* 14, no. 3: 293–308.

Consul, P. C., and G. C. Jain. 1971. "On the Log-Gamma Distribution and Its Properties." *Statistische Hefte* 12: 100–106.

Cox, D. R., G. Gudmundsson, G. Lindgren, et al. 1981. "Statistical Analysis of Time Series: Some Recent Developments [With Discussion and Reply]." *Scandinavian Journal of Statistics* 8, no. 2: 93–115.

de Andrade, B. S., M. G. Andrade, and R. S. Ehlers. 2015. "Bayesian GARMA Models for Count Data." *Communications in Statistics-Case Studies, Data Analysis and Applications* 1, no. 4: 192–205.

Fahrmeir, L. 1987. "Asymptotic Testing Theory for Generalized Linear Models." *Statistics* 18, no. 1: 65–76.

Fokianos, K., and B. Kedem. 1998. "Prediction and Classification of Non-Stationary Categorical Time Series." *Journal of Multivariate Analysis* 67: 277–296.

Fokianos, K., and B. Kedem. 2004. "Partial Likelihood Inference for Time Series Following Generalized Linear Models." *Journal of Time Series Analysis* 25, no. 2: 173–197.

Gradshteyn, I. S., and I. M. Ryzhik. 2007. *Table of Integrals, Series, and Products*. 7th ed. Academic Press.

Grande, A. F., G. Pumi, and G. B. Cybis. 2022. "Granger Causality and Time Series Regression for Modeling the Migratory Dynamics of Influenza Into Brazil." *Sort* 46, no. 2: 161–188.

Grande, A. F., G. Pumi, and G. B. Cybis. 2025. "Bayesian Analysis of Beta Autoregressive Moving Average Models." *Journal of Statistical Computation and Simulation* 95, no. 16: 3577–3594.

Grassia, A. 1977. "On a Family of Distributions With Argument Between 0 and 1 Obtained by Transformation of the Gamma and Derived Compound Distributions." *Australian Journal of Statistics* 19, no. 2: 108–114.

Griffiths, D., and C. Schafer. 1981. "Closeness of Grassia's Transformed Gammas and the Beta Distribution." *Australian Journal of Statistics* 23, no. 2: 240–246.

Gross, J., and U. Ligges. 2015. "nortest: Tests for Normality," R Package Version 1.0–4.

Halliwel, L. J. 2021. "The Log-Gamma Distribution and Non-Normal Error." *Variance* 13: 173–189.

Hogg, R. V., and S. A. Klugman. 1984. *Loss Distributions*. John Wiley & Sons.

Hyndman, R. J., and Y. Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 26, no. 3: 1–22.

Kalliovirta, L. 2012. "Misspecification Tests Based on Quantile Residuals." *Econometrics Journal* 15, no. 2: 358–393.

Kim, J. H. 2014. "vrtest: Variance Ratio Tests and Other Tests for Martingale Difference Hypothesis," R Package Version 0.97.

Korkmaz, S., D. Goksuluk, and G. Zararsiz. 2014. "MVN: An R Package for Assessing Multivariate Normality." *R Journal* 6, no. 2: 151–162.

Lastra, K. Z., G. Pumi, and T. S. Prass. 2025. "Order Selection in GARMA Models for Count Time Series: A Bayesian Perspective." *Journal of Applied Statistics* 52, no. 14: 2720–2744.

Maior, V., and F. Cysneiros. 2018. "SYMARMA: A New Dynamic Model for Temporal Data on Conditional Symmetric Distribution." *Statistical Papers* 59: 75–97.

Manchini, C. E. F., D. R. Canterle, G. Pumi, and F. M. Bayer. 2024. "Beta Autoregressive Moving Average Model With the Aranda-Ordaz Link Function." *Axioms* 13: 806.

Matsuoka, D. H., G. Pumi, H. Torrent, and M. Valk. 2024. "A Two-Step Approach to Production Frontier Estimation and the Matsuoka's Distribution," arXiv, 2311.06086.

Mendes, F. H. d. P. e. S., D. E. Turatti, and G. Pumi. 2025. "Mitigating the Choice of the Duration in DDMS Models Through a Parametric Link." *Journal of Applied Statistics* 52, no. 6: 1219–1238.

Palm, B. G., F. M. Bayer, and R. J. Cintra. 2023. "Prediction Intervals in the Beta Autoregressive Moving Average Model." *Communications in Statistics-Simulation and Computation* 52, no. 8: 3635–3656.

Prass, T. S., and G. Pumi. 2025. "BTSR: Bounded Time Series Regression," R Package Version 1.0.0.

Prass, T. S., G. Pumi, C. G. Taufemback, and J. H. Carlos. 2025. "Positive Time Series Regression Models: Theoretical and Computational Aspects." *Computational Statistics* 40: 1185–1215.

Pumi, G., T. S. Prass, and R. R. Souza. 2021. "A Dynamic Model for Double-Bounded Time Series With Chaotic-Driven Conditional Averages." *Scandinavian Journal of Statistics* 48, no. 1: 68–86.

Pumi, G., T. S. Prass, and C. G. Taufemback. 2024. "Unit-Weibull Autoregressive Moving Average Models." *Test* 33: 204–209.

Pumi, G., C. Rauber, and F. M. Bayer. 2020. "Kumaraswamy Regression Model With Aranda-Ordaz Link Function." *Test* 29: 1051–1071.

Pumi, G., M. Valk, C. Bisognin, F. M. Bayer, and T. S. Prass. 2019. "Beta Autoregressive Fractionally Integrated Moving Average Models." *Journal of Statistical Planning and Inference* 200: 196–212.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rocha, A. V., and F. Cribari-Neto. 2009. "Beta Autoregressive Moving Average Models." *Test* 18, no. 3: 529–545.

Thode, H. C. 2002. *Testing for Normality*. CRC Press.

U.S. Energy Information Administration. 2025. "Electricity Data Browser," U.S. Energy Information Administration Website Online Data Tool for U.S. Electricity Statistics.

U.S. Energy Information Administration (EIA). 2023. "Hydropower Explained," Online. U.S. Energy Information Administration.

### Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.

## Appendix A

### The Partial Score Vector

In principle, the partial score vector  $\frac{\partial \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$  can be used to obtain  $\hat{\boldsymbol{\gamma}}$  by solving the system  $\frac{\partial \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \mathbf{0}$ . To obtain the partial score vector, in view of (4), we only need to derive  $\frac{\partial \ell_t(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ . In what follows, all equalities are to be understood to hold almost surely. Observe that

$$\begin{aligned} \frac{\partial \ell_t(\boldsymbol{\gamma})}{\partial \gamma_j} &= \frac{\partial \ell_t(\boldsymbol{\gamma})}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \gamma_j} \\ &= \frac{1}{g'(\mu_t)} \left( \frac{2 \ln(Y_t)}{3(1 - \mu_t^{2/3})^2 \mu_t^{1/3}} + \frac{1}{(1 - \mu_t^{2/3}) \mu_t} \right) \frac{\partial \eta_t}{\partial \gamma_j}, \end{aligned}$$

where the last equality follows since  $\frac{\partial \mu_t}{\partial \eta_t} = \frac{1}{g'(\mu_t)}$ . Since  $\eta_t$  given in (3) has exactly the same specification as that of the KARMA model, the derivatives  $\frac{\partial \eta_t}{\partial \gamma_j}$  follow the same recursions as the ones presented in section 3 in Bayer et al. (2017), namely

$$\begin{aligned} \frac{\partial \eta_t}{\partial \alpha} &= 1 - \sum_{j=1}^q \theta_j \frac{\partial \eta_{t-j}}{\partial \alpha}, \\ \frac{\partial \eta_t}{\partial \beta_l} &= X_{tl} - \sum_{i=1}^p \phi_i X_{(t-i)l} - \sum_{j=1}^q \theta_j \frac{\partial \eta_{t-j}}{\partial \beta_l}, \\ \frac{\partial \eta_t}{\partial \phi_k} &= g(Y_{t-k}) - \mathbf{X}'_{t-k} \boldsymbol{\beta} - \sum_{j=1}^q \theta_j \frac{\partial \eta_{t-j}}{\partial \phi_k}, \quad \text{and} \\ \frac{\partial \eta_t}{\partial \theta_s} &= r_{t-s} - \sum_{j=1}^q \theta_j \frac{\partial \eta_{t-j}}{\partial \theta_s} \end{aligned} \tag{A1}$$

for  $l \in \{1, \dots, r\}$ ,  $k \in \{1, \dots, p\}$  and  $s \in \{1, \dots, q\}$ , where  $X_{tl}$  denotes the  $l$ -th component of  $\mathbf{X}_t$ . Let  $D_{\boldsymbol{\gamma}}$  be the  $n \times (p + q + r + 1)$  matrix whose  $(i, j)$ th elements are given by

$$[D_{\boldsymbol{\gamma}}]_{i,j} = \frac{\partial \eta_i}{\partial \gamma_j}, \quad \mathbf{h} := \left( \frac{\partial \ell_1(\boldsymbol{\gamma})}{\partial \mu_1}, \dots, \frac{\partial \ell_n(\boldsymbol{\gamma})}{\partial \mu_n} \right)'$$

and  $T$  be a diagonal matrix given by

$$T := \text{diag} \left\{ \frac{\partial \mu_1}{\partial \eta_1}, \dots, \frac{\partial \mu_n}{\partial \eta_n} \right\} = \text{diag} \left\{ \frac{1}{g'(\mu_1)}, \dots, \frac{1}{g'(\mu_n)} \right\}.$$

With these definitions, the partial score vector can be compactly written as

$$U(\boldsymbol{\gamma}) := \frac{\partial \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = D'_{\boldsymbol{\gamma}} T \mathbf{h}.$$

It is clear from (A1) that the PMLE cannot be obtained analytically, so we have to resort to numerical optimization to accomplish that.

### Conditional Information Matrix

In this section, we obtain the conditional information matrix in closed form, which will be useful later on when deriving the asymptotic properties of the partial maximum likelihood estimator for the proposed model. Equalities in this section are to be understood to hold almost surely. Let

$$\begin{aligned} H_t(\boldsymbol{\gamma}) &:= -\frac{\partial^2 \ell_t(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}, \quad \text{and} \\ H(\boldsymbol{\gamma}) &:= -\frac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = -\sum_{t=1}^n \frac{\partial^2 \ell_t(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = \sum_{t=1}^n H_t(\boldsymbol{\gamma}). \end{aligned}$$

Notice that  $H(\boldsymbol{\gamma})$  and  $\ell(\boldsymbol{\gamma})$  both depend on  $n$ . However, for simplicity and since no confusion will arise, we shall drop the dependence on  $n$  on the notation. Let  $I_n(\boldsymbol{\gamma}) := \mathbb{E}(H(\boldsymbol{\gamma}))$  be the information matrix corresponding to the sample of size  $n$  and let

$$I^{(n)}(\boldsymbol{\gamma}) := -\frac{1}{n} \sum_{t=1}^n \mathbb{E} \left( \frac{\partial^2 \ell_t(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right) = -\frac{1}{n} \mathbb{E} \left( \frac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right),$$

so that  $I_n(\boldsymbol{\gamma}) = nI^{(n)}(\boldsymbol{\gamma})$ . Now, observe that

$$I^{(n)}(\boldsymbol{\gamma}) = -\frac{1}{n} \sum_{t=1}^n \mathbb{E} \left( \mathbb{E} \left( \frac{\partial^2 \ell_t(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \middle| \mathcal{F}_{t-1} \right) \right) = \frac{1}{n} \mathbb{E}(K_n(\boldsymbol{\gamma})),$$

with

$$K_n(\boldsymbol{\gamma}) := -\sum_{t=1}^n \mathbb{E} \left( \frac{\partial^2 \ell_t(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \middle| \mathcal{F}_{t-1} \right). \tag{A2}$$

The matrix  $K_n(\boldsymbol{\gamma})$  is known as the conditional information matrix corresponding to the sample of size  $n$ . Under some regularity conditions (see the discussion in Section 4),

$$\begin{aligned} \frac{1}{n} H(\boldsymbol{\gamma}) - I^{(n)}(\boldsymbol{\gamma}) &\xrightarrow{P} 0 \quad \text{and} \\ \frac{1}{n} K_n(\boldsymbol{\gamma}) - I^{(n)}(\boldsymbol{\gamma}) &\xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (\text{A3})$$

Furthermore,  $I^{(n)}(\boldsymbol{\gamma}) \rightarrow I(\boldsymbol{\gamma})$ , where

$$I(\boldsymbol{\gamma}) = \lim_{n \rightarrow \infty} I^{(n)}(\boldsymbol{\gamma}) = \lim_{n \rightarrow \infty} -\frac{1}{n} \mathbb{E} \left( \frac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right), \quad (\text{A4})$$

which is the analogous of the  $I_1(\boldsymbol{\gamma})$  matrix for i.i.d. samples. We shall derive  $K_n(\boldsymbol{\gamma})$  in closed form. First notice that

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_j} &= \sum_{i=1}^n \frac{\partial}{\partial \mu_i} \left( \frac{\partial \ell_i(\boldsymbol{\gamma})}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_j} \right) \frac{d \mu_i}{d \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_j} \\ &= \sum_{i=1}^n \left[ \frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial \mu_i^2} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_j} + \frac{\partial \ell_i(\boldsymbol{\gamma})}{\partial \mu_i} \frac{\partial}{\partial \mu_i} \left( \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_j} \right) \right] \frac{d \mu_i}{d \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_j}. \end{aligned}$$

Since  $\frac{\partial \ell_i(\mu_i, \varphi)}{\partial \kappa_i} = \frac{3}{2\kappa_i} + \ln(Y_i)$  it follows that  $\mathbb{E} \left( \frac{\partial \ell_i(\mu_i, \varphi)}{\partial \mu_i} \middle| \mathcal{F}_{i-1} \right) = 0$ , and by the  $\mathcal{F}_{i-1}$ -measurability of  $\mu_i$  and  $\eta_i$ , it follows that

$$\mathbb{E} \left( \frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i \partial \boldsymbol{\gamma}_j} \middle| \mathcal{F}_{i-1} \right) = \mathbb{E} \left( \frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial \mu_i^2} \middle| \mathcal{F}_{i-1} \right) \left[ \frac{\partial \mu_i}{\partial \eta_i} \right]^2 \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_i} \frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_j},$$

where  $\frac{\partial \eta_i}{\partial \boldsymbol{\gamma}_k}$  is given in (A1). Hence, we only need to obtain  $\mathbb{E} \left( \frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial \mu_i^2} \middle| \mathcal{F}_{i-1} \right)$ .

$$\begin{aligned} \frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial \mu_i^2} &= \frac{\partial}{\partial \mu_i} \left[ \frac{2 \ln(Y_i)}{3(1 - \mu_i^{2/3})^2 \mu_i^{1/3}} + \frac{1}{(1 - \mu_i^{2/3}) \mu_i} \right] \\ &= \frac{2}{3} \ln(Y_i) \frac{\partial}{\partial \mu_i} \left[ (1 - \mu_i^{2/3})^{-2} \mu_i^{-1/3} \right] + \frac{\partial}{\partial \mu_i} \left[ \frac{1}{(1 - \mu_i^{2/3}) \mu_i} \right] \\ &= -\frac{2}{3} \ln(Y_i) \left[ \frac{1 - 5\mu_i^{2/3}}{3(1 - \mu_i^{2/3})^3 \mu_i^{4/3}} \right] + \frac{5\mu_i^{2/3} - 3}{3(1 - \mu_i^{2/3})^2 \mu_i^2}. \end{aligned}$$

Now, since  $\mathbb{E}(\ln(Y_i) | \mathcal{F}_{i-1}) = -\frac{3}{2\kappa_i} = -\frac{3(1 - \mu_i^{2/3})}{2\mu_i^{2/3}}$  and  $\mu_i$  is  $\mathcal{F}_{i-1}$ -measurable,

$$\begin{aligned} \mathbb{E} \left( \frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial \mu_i^2} \middle| \mathcal{F}_{i-1} \right) &= \frac{1 - \mu_i^{2/3}}{\mu_i^{2/3}} \left[ \frac{1 - 5\mu_i^{2/3}}{3(1 - \mu_i^{2/3})^3 \mu_i^{4/3}} \right] \\ &\quad + \frac{5\mu_i^{2/3} - 3}{3(1 - \mu_i^{2/3})^2 \mu_i^2} \\ &= -\frac{2}{3(1 - \mu_i^{2/3})^2 \mu_i^2}. \end{aligned}$$

By letting  $E_\mu$  be the  $n \times n$  diagonal matrix for which the  $k$ th diagonal elements is given by

$$[E_\mu]_{k,k} := -\mathbb{E} \left( \frac{\partial^2 \ell_k(\boldsymbol{\gamma})}{\partial \mu_k^2} \middle| \mathcal{F}_{k-1} \right) = \frac{2}{3(1 - \mu_k^{2/3})^2 \mu_k^2},$$

and  $D_\gamma$  and  $T$  as before, we obtain

$$K_n(\boldsymbol{\gamma}) = D_\gamma' T E_\mu T D_\gamma.$$

## Appendix B

Table B1 provides a comprehensive assessment of numerical stability and model complexity throughout the model selection and final estimation process. The ‘‘Model’’, ‘‘Link’’, and ‘‘Reg?’’ columns identify each of the 36 configurations spanning MARMA, KARMA,  $\beta$ ARMA families, three link functions (logit, loglog, cloglog), and two regressor scenarios (Yes = with harmonics; No = without). The ‘‘Alg.’’ column specifies the initial optimization algorithm (L-BFGS-B or Nelder-Mead). The ‘‘Failures’’ column classifies convergence failures during intermediate backward and forward selection steps into four tiers: Stable (0 failures), Low (1–20 failures), Moderate (21–100 failures), and High (exceeding 100 failures). Numbers in parentheses report specific contingency interventions, where the backup algorithm (Nelder-Mead for L-BFGS-B failures; L-BFGS-B for Nelder-Mead failures) was temporarily activated to continue the stepwise search. The ‘‘Status’’ column summarizes final estimation outcomes: ‘‘Converged’’ indicates the initial algorithm successfully estimated the final selected model; ‘‘Fallback’’ means permanent switching to the backup optimizer was required; ‘‘Fail’’ flags cases where fallback produced a model identical to that obtained when using the alternative algorithm as the initial optimizer. Model complexity appears in the rightmost columns: autoregressive order ( $p$ ), moving average order ( $q$ ), regressors retained ( $r$ ), significant parameters ( $s$ ), and highest  $p$ -value ( $p_{max}$ ) among parameters.

Table B2 reports the absolute characteristic root closest to the unit root for each of the 36 estimated ARMA processes, organized by model family (MARMA, KARMA,  $\beta$ ARMA), regressor inclusion (yes/no), optimization algorithm (L-BFGS-B, Nelder-Mead), and link function (logit, loglog, cloglog). Highlighted values indicate near-unit-root behavior and potential instability in the AR dynamics, while ‘‘-’’ indicates no AR component was estimated.

Table B3 presents a comprehensive comparison of in-sample goodness-of-fit statistics (AIC, BIC, HQC, Log-Likelihood) and predictive accuracy metrics (RMSE, MAPE, MDA) for all estimated configurations. The results are stratified by the inclusion of harmonic regressors and the optimization algorithm employed, allowing for a direct assessment of their impact on model performance. Complementing these fit statistics, the diagnostic column (‘‘DL/SF’’) summarizes the residual validation status at the 5% significance level: the symbol  $\checkmark$  indicates that the model satisfies the respective assumption (Dominguez-Lobato for the martingale difference property and Shapiro-Francia for normality), while  $\times$  indicates a rejection of the null hypothesis. The best result for each metric within each model family is highlighted in bold, isolating the optimal configuration for that specific class regardless of the link function or estimation method used.

Figure B1 presents detailed out-of-sample RMSE by forecast horizon ( $h \in \{1, \dots, 21\}$ ) for all 36 model configurations in bar plot format. Bars are grouped by configuration (link function + optimization algorithm), with light colors representing L-BFGS-B and dark colors representing Nelder-Mead, faceted across six panels by model family (MARMA, KARMA,  $\beta$ ARMA) and regressor inclusion (with/without). Horizontal dashed reference lines at RMSE = 0.005, 0.010, and 0.015 to help comparison against performance thresholds.

Figure B2 present out-of-sample forecast accuracy measures for the SARMA model and the best MARMA, KARMA and  $\beta$ ARMA.

**TABLE B1** | Summary of Estimation Complexity and Numerical Stability for All Combinations of Model Family (MARMA, KARMA,  $\beta$ ARMA), Link Function (Logit, Loglog, Cloglog), Optimization Algorithm (L-BFGS-B, Nelder-Mead), and Covariate Specification (Reg? = Yes With Harmonic Regressors; No Without). “Failures” Reports the Frequency of Convergence Failures During Stepwise Selection, Classified as Stable (0 Failures), Low (1–20), Moderate (21–100), or High (> 100). “Status” Summarizes the Outcome of the Estimation Routine (Converged, Fallback, or Fail), While the Complexity of the Final Selected Model is Described by Its Order ( $p, q$ ), Number of Regressors ( $r$ ), Number of Significant Parameters ( $s$ ), and the Highest  $p$ -Value Among Model Parameters ( $p_{max}$ ).

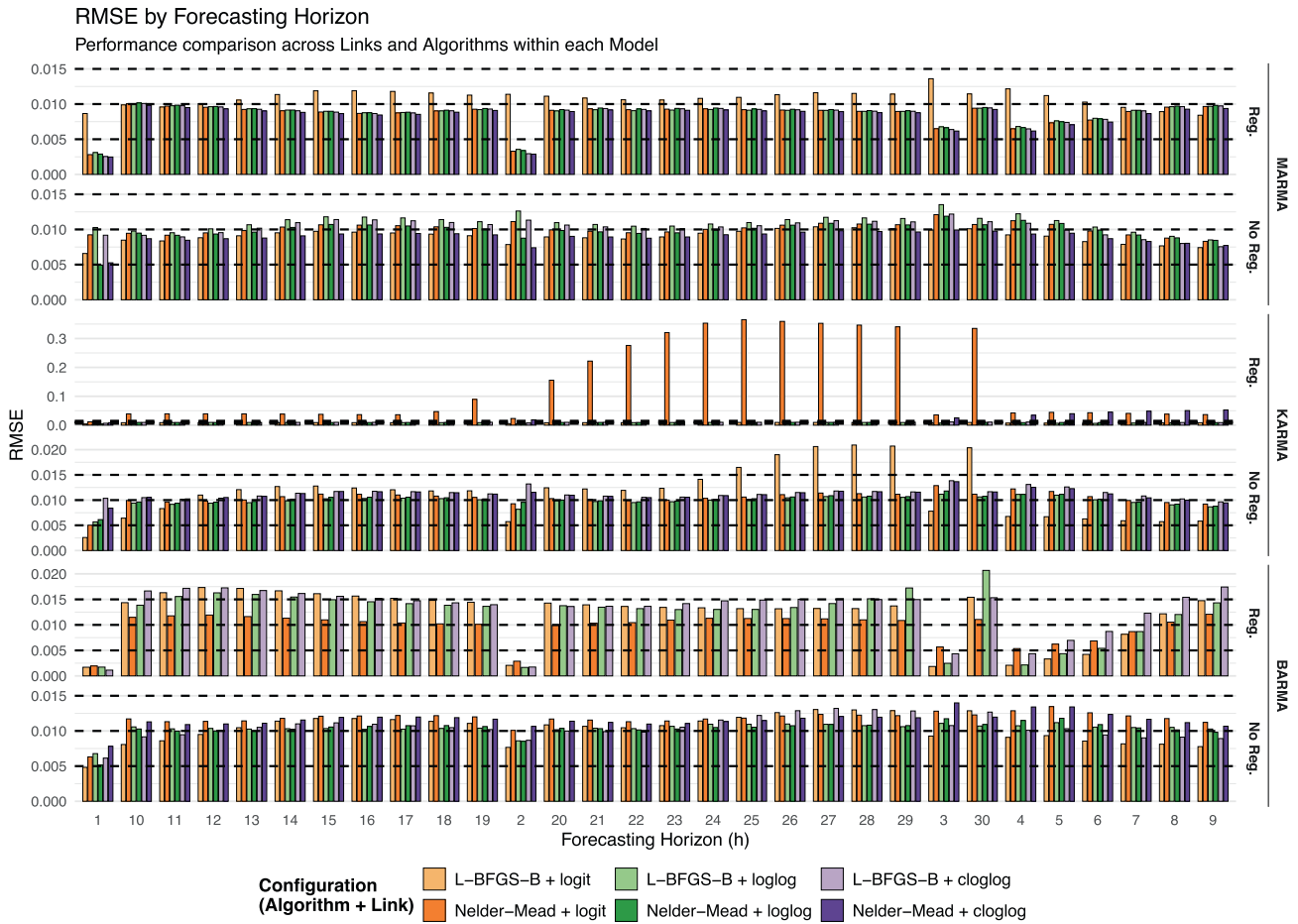
Model	Link	Reg?	Alg.	Stepwise inst.	Status	$p$	$q$	$r$	$s$	$P_{max}$
MARMA	logit	Yes	L-BFGS-B	Moderate (51)	Converged	8	0	0	3	0.0376
			Nelder-Mead	Moderate (65)	Converged	9	0	1	4	0.0868
		No	L-BFGS-B	Low (12)	Converged	7	0	0	4	0.0455
			Nelder-Mead	Low (9)	Converged	7	0	0	4	0.0988
	loglog	Yes	L-BFGS-B	Low (8)	Converged	9	0	1	4	0.0399
			Nelder-Mead	Moderate (44)	Converged	9	0	1	4	0.0919
		No	L-BFGS-B	Moderate (60)	Fallback	7	0	0	3	0.0122
			Nelder-Mead	Moderate (24)	Converged	7	0	0	4	0.0131
	cloglog	Yes	L-BFGS-B	Moderate (58)	Converged	9	0	1	4	0.1066
			Nelder-Mead	Moderate (35)	Converged	9	0	1	4	0.0911
		No	L-BFGS-B	Moderate (31)	Converged	7	0	0	3	0.0063
			Nelder-Mead	Moderate (32)	Converged	11	0	0	5	0.0742
KARMA	logit	Yes	L-BFGS-B	High (274)	Fallback	9	0	2	7	< 0.001
			Nelder-Mead	High (229)	Fallback	12	0	0	9	0.0428
		No	L-BFGS-B	Moderate (72)	Converged	11	12	0	14	0.0184
			Nelder-Mead	High (136)	Converged	12	0	0	7	0.0010
	loglog	Yes	L-BFGS-B	High (251)	Fallback	12	0	2	8	0.0234
			Nelder-Mead	High (192)	Converged	12	0	1	8	< 0.001
		No	L-BFGS-B	High (188)	Fallback	11	0	0	9	0.0878
			Nelder-Mead	High (135)	Converged	11	0	0	10	0.0878
	cloglog	Yes	L-BFGS-B	High (274)	Fallback	11	0	1	7	0.0091
			Nelder-Mead	High (118)	Fallback	12	12	2	15	0.1091
		No	L-BFGS-B	High (192)	Fallback	7	0	0	6	< 0.001
			Nelder-Mead	High (104)	Converged	9	0	0	6	< 0.001
$\beta$ ARMA	logit	Yes	L-BFGS-B	Stable	Converged	10	12	2	12	0.0679
			Nelder-Mead	High (135)	Converged	10	0	2	10	0.0125
		No	L-BFGS-B	Low (7)	Converged	7	12	0	10	0.0644
			Nelder-Mead	High (159)	Fallback	7	1	0	9	0.0943
	loglog	Yes	L-BFGS-B	Stable	Converged	9	12	2	11	0.0594
			Nelder-Mead	High (188)	Fail	9	12	2	11	0.0594
		No	L-BFGS-B	Stable	Converged	11	0	0	8	0.0204
			Nelder-Mead	High (152)	Converged	11	0	0	8	< 0.001
	cloglog	Yes	L-BFGS-B	Stable	Converged	10	1	2	11	0.0864
			Nelder-Mead	High (188)	Fail	10	1	2	11	0.0864
		No	L-BFGS-B	Stable	Converged	8	12	0	9	0.0152
			Nelder-Mead	High (135)	Converged	9	0	0	7	< 0.001

**TABLE B2** | Absolute value of the AR characteristic root ( $r$ ) closest to 1, by model family, regressor inclusion, algorithm and link function. “Failed” indicates estimation failure. Values satisfying  $0.95 < |r| < 1.05$  (near-unit-root behavior) are highlighted in *red italics*.

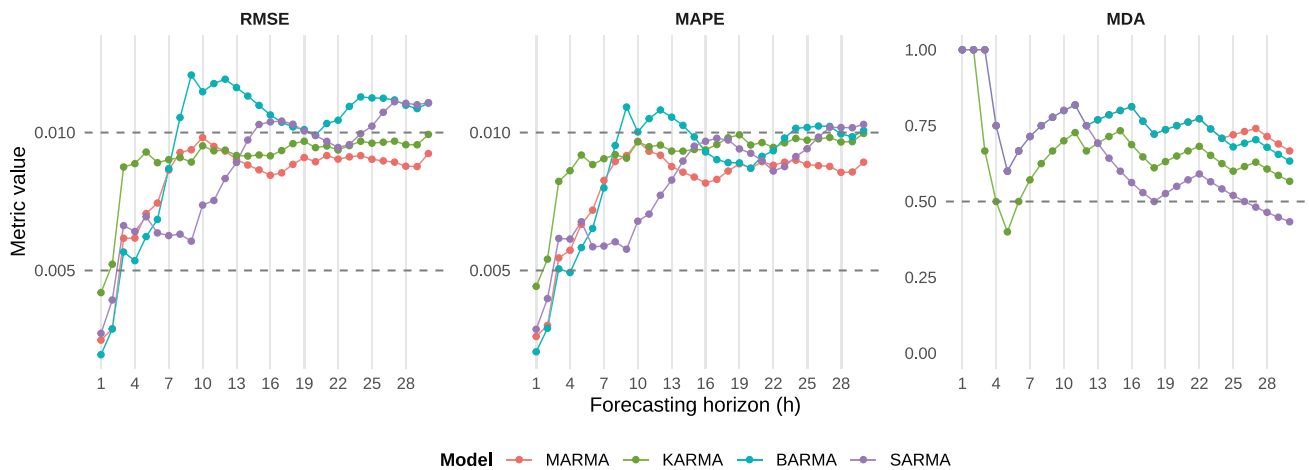
Model	Reg?	Algorithm	Link		
			logit	loglog	cloglog
MARMA	yes	L-BFGS-B	1.1027	1.0597	1.1026
		Nelder-Mead	1.0761	1.0808	1.0942
	No	L-BFGS-B	<i>1.0306</i>	1.0777	1.0718
		Nelder-Mead	1.0512	1.0565	<i>1.0227</i>
KARMA	yes	L-BFGS-B	<i>1.0279</i>	<i>1.0088</i>	1.1182
		Nelder-Mead	<i>0.9974</i>	<i>1.0123</i>	<i>0.9745</i>
	No	L-BFGS-B	<i>1.0294</i>	<i>1.0380</i>	1.0749
		Nelder-Mead	1.0519	<i>1.0382</i>	1.0708
$\beta$ ARMA	yes	L-BFGS-B	<i>1.0121</i>	<i>1.0210</i>	<i>1.0441</i>
		Nelder-Mead	1.0577	Failed	Failed
	No	L-BFGS-B	<i>1.0183</i>	<i>0.9910</i>	<i>0.9859</i>
		Nelder-Mead	<i>1.0398</i>	<i>1.0135</i>	<i>1.0370</i>

**TABLE B3** | Comparison of Goodness-Of-Fit Metrics, In-Sample Accuracy, and Diagnostic Validation for All Model Configurations. the Best Result for Each Metric Within Each Model Family is Highlighted in **Bold**; the Worst Result for Each Metric Within Each Family is Shown in *italic*. “Reg?” Denotes Inclusion (“Yes”) or Exclusion (“No”) of Harmonic Regressors; “Alg.” Indicates the Optimization Algorithm (L = L-BFGS-B; NM = Nelder-Mead); “DL/SF” Summarizes the Diagnostic Results for Dominguez-Lobato (Martingale) and Shapiro-Francia (Normality) Tests, Where Symbols ✓ And × Indicate Pass ( $p$ -Value  $\geq 0.05$ ) or Failure ( $p$ -Value  $< 0.05$ ), Respectively.

Model	Link	Reg?	Algo	Valid? DL/SF	Goodness-of-fit				In-sample accuracy				
					AIC	BIC	HQC	LogLik	RMSE	MAPE	MDA		
MARMA	logit	Yes	L	✓/✓	-1011.3	-1000.7	-1012.2	<i>508.7</i>	<i>0.0086</i>	<i>0.0076</i>	<i>0.5504</i>		
			NM	✓/✓	-1012.2	-997.9	-1013.3	510.1	0.0077	0.0064	0.7326		
		No	L	✓/✓	-1012.2	-998.0	-1013.4	510.1	<b>0.0070</b>	<b>0.0060</b>	0.7132		
			NM	✓/✓	-1012.0	-997.8	-1013.1	510.0	0.0076	0.0065	0.6667		
	loglog	Yes	L	✓/✓	-1012.1	-997.8	-1013.2	510.0	0.0076	0.0063	0.6899		
			NM	✓/✓	-1012.1	-997.9	-1013.3	510.1	0.0078	0.0065	0.7364		
		No	L	✓/✓	-1012.2	<b>-1001.6</b>	-1013.1	509.1	0.0084	0.0074	0.5775		
			NM	✓/✓	<b>-1012.4</b>	-998.2	<b>-1013.6</b>	510.2	0.0072	0.0060	0.6977		
	cloglog	Yes	L	✓/✓	-1012.2	-997.9	-1013.3	510.1	0.0078	0.0065	<b>0.7403</b>		
			NM	✓/✓	-1012.2	-998.0	-1013.3	510.1	0.0077	0.0064	<b>0.7403</b>		
		No	L	✓/✓	-1011.9	-1001.3	-1012.8	509.0	0.0084	0.0074	<i>0.5504</i>		
			NM	✓/✓	<i>-1010.6</i>	<i>-992.8</i>	<i>-1012.0</i>	<b>510.3</b>	0.0071	0.0061	0.7132		
KARMA	logit	Yes	L	✗/✗	-1808.9	<b>-1784.0</b>	-1810.9	911.5	0.0070	0.0059	0.7093		
			NM	✓/✗	-1568.8	-1536.8	-1571.4	793.4	0.0101	0.0083	<i>0.6163</i>		
			No	L	✓/✗	-1626.1	-1576.3	-1630.1	827.1	0.0088	0.0073	0.6783	
		No	NM	✓/✗	-1795.9	-1771.0	-1797.9	905.0	0.0072	0.0061	<b>0.7364</b>		
			loglog	Yes	L	✓/✗	-1804.7	-1776.3	-1807.0	910.4	0.0070	0.0060	0.7054
				NM	✗/✗	-1797.7	-1769.2	-1800.0	906.8	0.0073	0.0062	0.7248	
	cloglog	Yes	L	✗/✗	<b>-1813.9</b>	-1781.9	<b>-1816.4</b>	<b>915.9</b>	<b>0.0069</b>	<b>0.0059</b>	0.7248		
			NM	✗/✗	-1798.6	-1763.0	-1801.5	909.3	0.0070	0.0060	0.7287		
			No	L	✗/✓	-1723.2	-1698.3	-1725.2	868.6	0.0088	0.0072	0.7132	
		No	NM	✓/✗	<i>-1052.1</i>	<i>-998.7</i>	<i>-1056.4</i>	<i>541.0</i>	<i>0.0180</i>	<i>0.0150</i>	0.6705		
			L	✓/✗	-1769.3	-1747.9	-1771.0	890.6	0.0076	0.0066	0.7287		
			NM	✓/✗	-1771.3	-1749.9	-1773.0	891.6	0.0076	0.0066	0.7093		
βARMA	logit	Yes	L	✓/✓	-1828.9	-1786.2	-1832.3	926.4	<b>0.0065</b>	0.0055	0.7597		
			NM	✓/✓	-1843.7	-1808.1	-1846.5	931.8	0.0066	0.0057	0.7636		
			No	L	✓/✓	-1752.8	-1717.3	-1755.7	886.4	0.0065	0.0055	<b>0.7829</b>	
		No	NM	✓/✓	-1745.0	-1713.0	-1747.5	881.5	0.0068	0.0057	0.7364		
			loglog	Yes	L	✓/✓	-1829.9	-1790.8	-1833.1	926.0	0.0065	<b>0.0054</b>	<i>0.7287</i>
				NM									Fail – Fallback – same model as L-BFGS-B
	cloglog	Yes	L	✓/✓	<i>-1734.7</i>	<i>-1706.3</i>	<i>-1737.0</i>	<i>875.4</i>	0.0069	<i>0.0060</i>	0.7519		
			NM	✗/✓	<b>-1851.0</b>	<b>-1822.5</b>	<b>-1853.2</b>	<b>933.5</b>	0.0066	0.0057	0.7597		
			No	L	✓/✓	-1817.7	-1778.6	-1820.9	919.9	<i>0.0070</i>	0.0059	0.7752	
		No	NM									Fail – Fallback – same model as L-BFGS-B	
			L	✓/✓	-1767.4	-1735.4	-1770.0	892.7	0.0068	0.0059	0.7481		
			NM	✓/✓	-1826.9	-1802.0	-1828.9	920.4	<i>0.0070</i>	<i>0.0060</i>	0.7636		



**FIGURE B1** | Out-of-sample RMSE by forecast horizon ( $h \in \{1, \dots, 30\}$ ) across all model configurations (link + algorithm). Horizontal dashed lines mark performance thresholds at  $RMSE = 0.005, 0.010,$  and  $0.015$ .



**FIGURE B2** | Out-of-sample RMSE, MAPE and MDA by forecast horizon ( $h \in \{1, \dots, 30\}$ ) for the SARMA model and the best MARMA, KARMA and  $\beta$ ARMA. Horizontal dashed lines mark performance thresholds at 0.005 and 0.010 for RMSE and MAPE, and 0.5 for MDA.