Revista Colombiana de Estadística
Special Issue: Colombian and Brazil Graduate Programs
December 2025, volume 48, issue 3, pp. 529–552
http://doi.org/10.15446/rce.v48n3.123655

Nonlinear Macroeconomic Time Series Forecasting with TSARX: An Empirical Comparison Across Three Economies

Pronósticos de series de tiempo macroeconómicas no lineales con TSARX: un estudio comparativo en tres economías

Sebastián Árbelaez-Quíntero[1,a], Sergio A. Calderón V.[1,b], Joaquín González Borja[2,c]

[1]Department of Statistics, Faculty of Science, Universidad Nacional de Colombia, Bogotá, Colombia

[2]Department of Mathematics and Statistics, Faculty of Science, Universidad del Tolima, Ibagué, Colombia

Abstract

The nonlinearity and seasonality of the business cycle account for the majority of short-run movements in quarterly or monthly macroeconomic time series.

The multiplicative seasonal threshold autoregressive model with exogenous inputs (TSARX) combines threshold dynamics, multiplicative seasonality, and external regressor, the model is a special case of a general non-multiplicative TAR model, but there is limited evidence on how it performs in forecasting relative to simpler models and machine-learning algorithms.

We evaluate the performance out-of-sample of TSARX models using seasonally unadjusted macroeconomic time series for three economies (Colombia, the United States of America, and the United Kingdom) and three key variables (gross domestic product(GDP), unemployment rate, and inflation). For each country–variable pair, we compare the TSARX models with four competitor models: a TAR model, a linear seasonal autoregressive model (SAR), Holt-Winters exponential smoothing (ES), and long short-term memory (LSTM) neural networks. Multi-step forecasts at horizons from one to four periods ahead are produced under a rolling-window design, and accuracy is assessed using Mean Squared Error (MSE) and Diebold-Mariano tests (DM) for equal predictive ability.

[a]M.Sc. E-mail: sarbelaezq@unal.edu.co
[b]Ph.D. E-mail: sacalderonv@unal.edu.co
[c]Ph.D. E-mail: jgonzalezb@ut.edu.co

Across 36 series–country–horizon combinations, the TSARX achieves the lowest MSE in three cases and is often statistically indistinguishable from the best benchmark; in many situations simpler models remain difficult to beat. These findings show that additional nonlinear and seasonal structure does not guaranty superior forecasts and that the benefits of the TSARX models are context- and horizon-dependent.

***Keywords*:** Macroeconomic time series; Nonlinearity; Predictive ability; Seasonality; Threshold models.

Resumen

La no linealidad y la estacionalidad del ciclo económico explican la mayor parte de los movimientos de corto plazo en las series de tiempo macroeconómicas trimestrales o mensuales.

El modelo autorregresivo con umbrales y estacionalidad multiplicativa con regresores exógenos (TSARX) combina dinámica por regímenes, estacionalidad multiplicativa y variables explicativas externas, el cual es un caso especial de un modelo TAR no multiplicativo general, pero existe poca evidencia sobre su desempeño en pronósticos frente a modelos más simples y algoritmos de aprendizaje automático.

En este trabajo evaluamos el desempeño fuera de muestra de los modelos TSARX utilizando series de tiempo macroeconómicas sin desestacionalizar de tres economías (Colombia, Estados Unidos de América y Reino Unido) y tres variables clave (producto interno bruto, tasa de desempleo e inflación). Para cada combinación país–variable comparamos los modelos TSARX con cuatro modelos competidores: un modelo TAR, un modelo autorregresivo estacional lineal (SAR), el suavizamiento exponencial de Holt y Winters (ES) y redes neuronales de memoria de largo y corto plazo (LSTM). Se generan pronósticos multi paso a horizontes de uno a cuatro períodos adelante bajo un esquema de ventana rodante, y la precisión se evalúa mediante el Error Cuadrático Medio (MSE) y pruebas de Diebold y Mariano (DM) de igualdad de capacidad predictiva.

En 36 combinaciones serie–país–horizonte, los modelos TSARX obtienen el menor MSE en tres casos y usualmente es estadísticamente indistinguible del mejor modelo de referencia; en muchas situaciones modelos más simples siguen siendo difíciles de superar. Estos resultados muestran que añadir estructura no lineal y estacional no garantiza mejores pronósticos y que las ganancias de los modelos TSARX dependen del contexto y del horizonte.

***Palabras clave*:** Capacidad predictiva; Estacionalidad; Modelos con umbrales; No linealidad; Series de tiempo macroeconómicas.

1. Introduction

Threshold autoregressive (TAR) models provide a flexible and interpretable way to represent nonlinear dynamics in time series. Since the seminal contributions of Tong & Lim (1980) and Tong (1990), TAR and related threshold models have been used to capture regime changes, asymmetric adjustments, and limit-cycle behavior in economics and finance; see, among many others Tsay (1989);

Chan (1993); Hansen (2011); Tsay (1998). In the presence of pronounced seasonality, multiplicative seasonal extensions of TAR models are particularly attractive because they allow for regime-dependent dynamics within and across seasons. The multiplicative seasonal self-exciting threshold autoregressive models of De Gooijer & Vidiella-i Anguera (2003) and the multiplicative seasonal threshold models with exogenous inputs (TSARX) proposed by González & Nieto (2020) are examples of such specifications, designed to accommodate both nonlinear regime switching and seasonal interaction effects.

Nonlinear models of this type are especially relevant for macroeconomic time series that exhibit recurrent but regime-dependent behaviour, such as expansions and recessions, inflationary and disinflationary regimes, or high and low unemployment states. Empirical evidence suggests that GDP growth, inflation and unemployment rate often display threshold effects, asymmetries and state-dependent persistence that are difficult to reconcile with purely linear representations (Franses & van Dijk, 2000; Milas et al., 2006; Lee & Wang, 2012). At the same time, many macroeconomic time series also show strong and evolving seasonal patterns, whose explicit modelling can improve the description of the data and the understanding of underlying economic mechanisms (Ghysels et al., 2006).

Forecasting is equally central in economic applications. Forecasts of gross domestic product (GDP), unemployment rate and inflation inform monetary and fiscal policy, business planning and financial decisions (Montgomery et al., 1998; Clements & Krolzig, 1998; Dixon et al., 2020; Yilmaz & Arabaci, 2021). A vast literature compares forecasting methods for macroeconomic time series, including classical linear models such as autoregressive integrated moving average (ARIMA) and seasonal ARIMA models, nonlinear models such as smooth-transition and threshold autoregressions, and a wide range of machine-learning techniques. In many empirical applications, the time series are first seasonally adjusted and then modelled with methods that ignore the original seasonal structure. However, when the goal is to forecast the raw, non-adjusted series, models that explicitly handle seasonality and potential nonlinearity deserve careful consideration.

Building on Master Thesis (Arbeláez Quintero, 2022), this paper conducts an empirical assessment of the TSARX model for macroeconomic forecasting using monthly and quarterly seasonally unadjusted time series from three economies (Colombia, the United States of America and the United Kingdom) and three key variables (GDP, unemployment rate and inflation). We treat TSARX as a parsimonious nonlinear benchmark that combines multiplicative seasonality, threshold dynamics and exogenous regressors, and we compare its forecasting performance with that of four alternative models: a standard TAR model, a linear seasonal autoregressive (SAR) model, Holt-Winters exponential smoothing (ES) and long short-term memory (LSTM) neural networks.

The focus of this article is therefore *forecasting-oriented*: our main objective is to evaluate the out-of-sample predictive performance of TSARX relative to these competing models for multi-step forecasting of nonlinear seasonal macroeconomic time series. Rather than proposing a new statistical model or estimation method, we aim to clarify in which situations the additional nonlinear and seasonal struc-

ture embodied in TSARX translates into measurable gains in forecast accuracy and in which situations simpler models remain preferable.

The contribution of the paper is threefold. First, it provides a systematic comparison of TSARX with both traditional seasonal models (TAR, SAR and ES) and a modern nonlinear benchmark (LSTM) across three countries and three macroeconomic variables, using a common rolling-window design and multiple forecast horizons. Second, it documents in a transparent way the cases in which TSARX delivers improvements in Mean Squared Error (MSE) and in Diebold–Mariano tests (DM) for equal predictive accuracy, and the cases in which simpler seasonal specifications are difficult to beat. Third, it offers a detailed empirical illustration of TSARX in a setting where the interplay between nonlinearity and seasonality is economically plausible and where the choice between seasonally adjusted and unadjusted data is itself of interest to practitioners.

The remainder of the paper is organised as follows. Section 2 describes the data, the competing models, the forecast construction, and the evaluation criteria. Section 3 reports the forecasting results for each country and variable. Finally, Section 4 summarises the main findings and discusses implications for macroeconomic forecasting practice.

2. Methodology

This section presents the data and variables, summarizes the forecasting models considered in the study, describes the forecast construction, and explains the evaluation criteria.

2.1. Data Description

We conduct a comparative study using time series from Colombia, the United States of America (USA), and the United Kingdom (UK). We select three macroeconomic variables that are central in policy and business decisions: Gross Domestic Product (GDP) or a monthly proxy, unemployment rate and inflation. For each variable and country we work with seasonally unadjusted time series, preserve the original seasonality and standard transformations, and record data sources.

GDP. The GDP is a core macroeconomic indicator measured by national statistical offices and central banks. For quarterly series, GDP exhibits pronounced business-cycle dynamics and often strong seasonal patterns due to production cycles, fiscal calendars and holidays. For monthly time series, industrial production or activity indices are used as high-frequency proxies for GDP, and calendar effects and institutional customs can also shift seasonal dynamics (e.g., Easter effects).

Unemployment rate. The unemployment rate measures the share of the Economically Active Population (EAP) that is unemployed. It is usually based on labour force surveys conducted by national statistics offices, and it typically exhibits seasonality driven by school calendars, harvest periods and institutional hiring cycles. In Colombia, the EAP definition follows (DANE, 2015).

Inflation. The Consumer Price Index (CPI) tracks price changes in a representative consumption basket. Monthly inflation, computed as the monthly percentage change in the CPI, is a key indicator of the purchasing power of households and firms. It often shows seasonal patterns related to regulated prices, food items, and seasonal demand (e.g., holidays, school-year start). Central banks typically monitor the CPI and publish inflation series monthly because they inform monetary policy and short-term cycles.

Colombian time series were retrieved from the National Administrative Department of Statistics (Departamento Administrativo Nacional de Estadística, DANE) and the Central Bank of Colombia (Banco de la República, Banrep); we also report units, frequencies, and threshold variables (needed for TAR/TSARX). For the USA and the UK, GDP, unemployment rate and inflation time series were obtained from official statistical agencies and central banks. In all cases, we work with the longest available spans subject to the requirement of excluding the most disruptive phase of the COVID-19 pandemic.

For Colombia, we proxy monthly GDP using the Monthly Economic Activity Index (Indicador de Seguimiento a la Economía, ISE) produced by DANE as a high-frequency measure of real activity; see DANE (2016). The Colombian ISE time series contains 182 monthly observations (Feb 2005–Mar 2020); the first 167 (Feb 2005–Dec 2018) are used for estimation and the last 15 for forecast evaluation. We model monthly percentage changes, $X_t = \{\ln(\text{ISE}_t) - \ln(\text{ISE}_{t-1})\} \times 100\%$. As potential threshold variables, we consider the lagged growth rate and an interest-rate spread, both constructed from Banrep data; see Vaca (2018).

The USA GDP sample comprises 142 quarterly observations (1980Q1–2015Q2). We work with the log-difference of real GDP at quarterly frequency, and consider as potential threshold variables the lagged growth rate and the 10-year Treasury yield minus the effective federal funds rate. For the UK we adopt a similar specification based on real GDP and an interest-rate spread, over a sample of 135 quarterly observations.

Additional details on the samples and transformations used for unemployment rate and inflation are analogous and omitted for brevity, more details in Arbeláez Quintero (2022). For all series we align the end of the sample at March 2020 (or the corresponding quarter) to avoid COVID-19 regime-break effects thereafter.

Figures 1 to 3 show the graphs of the target time series and the selected threshold variables. The presence of seasonality and asymmetries in GDP, unemployment rate, and inflation is evident for the three countries under study. Meanwhile, the threshold variables exhibit stable behavior over time.

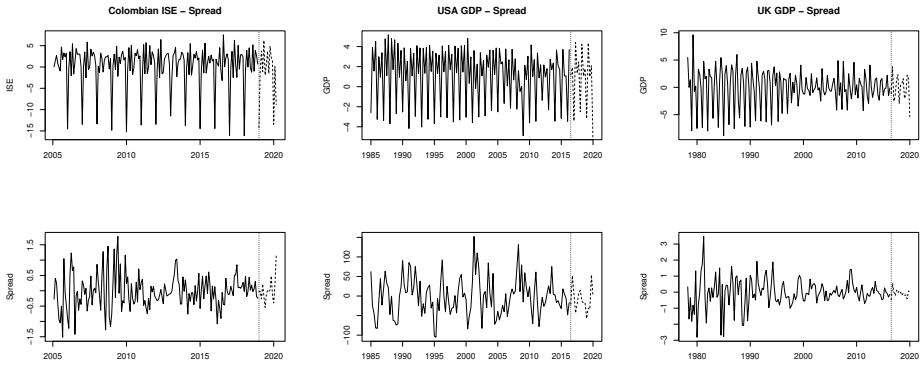

FIGURE 1: GDP time series and threshold variables for Colombia, USA, and UK. The solid segment indicates the estimation period; the dashed segment indicates the forecast evaluation period.

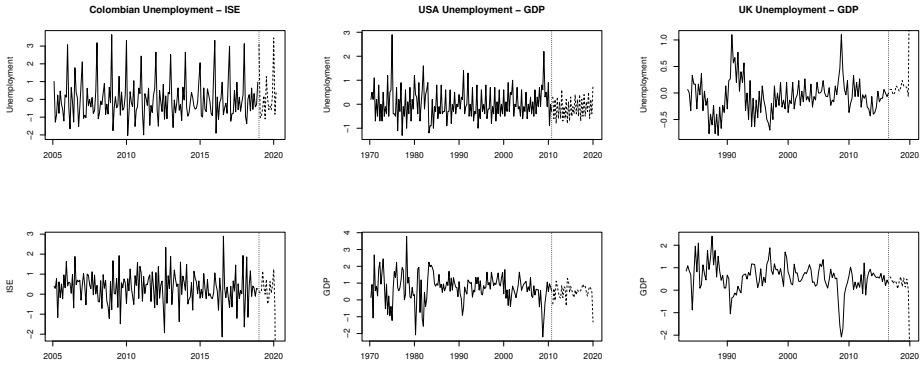

FIGURE 2: Unemployment rate time series and threshold variables for Colombia, USA, and UK. The solid segment indicates the estimation period; the dashed segment indicates the forecast evaluation period.

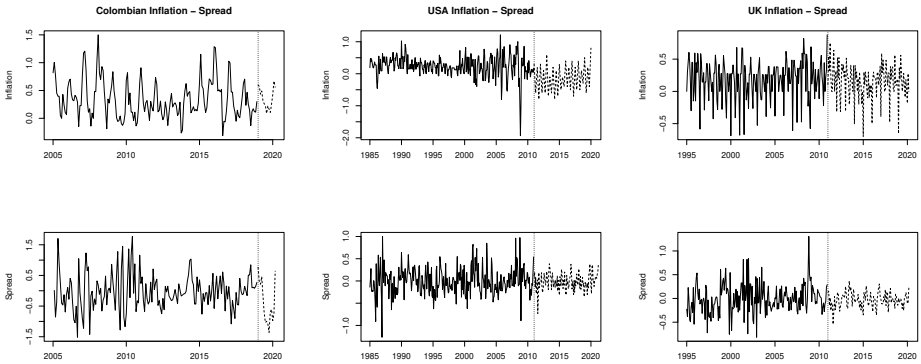

FIGURE 3: Inflation time series and threshold variables for Colombia, USA, and UK. The solid segment indicates the estimation period; the dashed segment indicates the forecast evaluation period.

Table 1 summarizes the frequency and length of time series training observations by variable and countries, used in our empirical analysis.

TABLE 1: Datasets information.

Variable	Country	Frequency	Length training time series
GDP	Colombia	Monthly	167
	USA	Quarterly	120
	UK	Quarterly	112
Unemployment rate	Colombia	Monthly	161
	USA	Quarterly	163
	UK	Quarterly	131
Inflation	Colombia	Monthly	168
	USA	Monthly	313
	UK	Monthly	192

2.2. Models

We briefly outline the models used to compare forecasting accuracy against TSARX on seasonal macroeconomic time series.

In what follows, $\{X_t\}$ denotes the target time series (GDP, unemployment rate or inflation), $\{Z_t\}$ denotes the threshold or exogenous process that drives regime changes, and $\hat{X}_{t+h|t}$ denotes the h-step-ahead forecast of X_{t+h} based on information up to time t. Upper-case letters are used for observed time series, whereas lower-case Greek letters are reserved for unknown parameters; innovations are denoted by $\{\epsilon_t\}$ and are typically assumed to be independent with mean zero.

2.2.1. TAR Model

The threshold autoregressive (TAR) model (Tong & Lim, 1980; Tong, 1990; Tsay, 1989) relates a univariate time series $\{X_t\}$ to a scalar threshold variable $\{Z_t\}$ via a piecewise linear autoregression

$$X_t = a_0^{(j)} + \sum_{i=1}^{k_j} a_i^{(j)} X_{t-i} + h^{(j)}\epsilon_t, \qquad \text{if } r_{j-1} < Z_{t-d} \le r_j, \quad t \in \mathbb{Z},$$

where $\mathbb{Z}$ is the set of integers numbers, the thresholds $\{r_j\}$ define regimes $j = 1, \ldots, l$, with $r_0 = -\infty$ and $r_l = +\infty$. Structural parameters are the number of regimes l, the threshold vector $\mathbf{r} = (r_1, \ldots, r_{l-1})'$, the delay d, and the per-regime autoregressive orders $k_1, \ldots, k_l$. Within regime j, the process behaves like a linear AR(k_j) model with regime-specific coefficients $\{a_i^{(j)}\}$ and scale $h^{(j)}$ ($i = 1, \ldots, k_j; j = 1, \ldots, l$), and $\{\epsilon_t\}$ is typically assumed to be i.i.d. $N(0,1)$ or to follow another zero-mean innovation distribution.

2.2.2. TSARX Model

The TSARX model (González & Nieto, 2020) extends the TAR specification by incorporating multiplicative seasonality and exogenous regressors. It relates the target time series $\{X_t\}$ to an exogenous threshold process $\{Z_t\}$ as

$$\begin{aligned} X_t = a_0^{(j)} &+ \sum_{i=1}^{k_j} a_i^{(j)} X_{t-i} + \sum_{u=1}^{K_j} b_u^{(j)} X_{t-su} - \sum_{i=1}^{k_j}\sum_{u=1}^{K_j} a_i^{(j)} b_u^{(j)} X_{t-i-su} \\ &+ \sum_{v=1}^{q_j} c_v^{(j)} Z_{t-v} + h^{(j)}\epsilon_t, \qquad \text{if } r_{j-1} < Z_{t-d} \le r_j,\ t \in \mathbb{Z}, \end{aligned}$$

where s is the seasonal period (12 for monthly time series and 4 for quarterly time series), and k_j, K_j and q_j denote, respectively, the non-seasonal autoregressive order, the seasonal autoregressive order and the order of the exogenous component in regime j ($j = 1, \ldots, l$). The innovation process $\{\epsilon_t\}$ is assumed i.i.d. with mean zero and unit variance and independent of $\{Z_t\}$. As in the TAR model, the thresholds $\{r_j\}$, the delay d and the per-regime orders (k_j, K_j, q_j) are structural parameters, while the coefficients $\{a_i^{(j)}\}$, $\{b_u^{(j)}\}$, $\{c_v^{(j)}\}$ and scales $h^{(j)}$ ($i = 1, \ldots, k_j; u = 1, \ldots, K_j; v = 1, \ldots, q_j; j = 1, \ldots, l$) are regime-specific non-structural parameters. In our empirical application $\{Z_t\}$ is taken to be an observed exogenous macroeconomic indicator (such as an interest-rate spread), so that no additional dynamic structure is imposed on the threshold process.

2.2.3. SAR Model

We include as a linear benchmark a seasonal autoregressive (SAR) model for the transformed time series X_t, of the form

$$X_t = a_0 + \sum_{i=1}^{k} a_i X_{t-i} + \sum_{u=1}^{K} b_u X_{t-su} + h\epsilon_t,$$

where s denotes the seasonal period, k and K are the non-seasonal and seasonal autoregressive orders, respectively, and $\{\epsilon_t\}$ is white noise. This model is akin to the autoregressive component of a seasonal ARIMA specification, but without differencing.

2.2.4. Exponential Smoothing (ES)

As a classical benchmark for seasonal forecasting we use additive Holt-Winters exponential smoothing (Brown, 1959; Holt, 1957; Winters, 1960), which in the (Hyndman & Athanasopoulos, 2021) notation corresponds to the ETS(A, A, A) model. The observed series $\{X_t\}$ is decomposed into a local level ℓ_t, a local trend b_t and a seasonal component s_t with period s (12 for monthly time series and 4 for quarterly time series). The state-space formulation of the additive error, additive trend, additive seasonality model is

$$\begin{aligned} X_t &= \ell_{t-1} + b_{t-1} + s_{t-s} + \epsilon_t, \text{where} \\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha\epsilon_t, \\ b_t &= b_{t-1} + \beta\epsilon_t, \\ s_t &= s_{t-s} + \gamma\epsilon_t, \end{aligned}$$

and $\{\epsilon_t\}$ is a zero-mean error term, and $\alpha, \beta, \gamma \in (0, 1)$ are smoothing parameters controlling the adaptation of the level, trend and seasonal component, respectively.

Given the estimated states at time t, h-step-ahead point forecasts are obtained by extrapolating the level and trend and repeating the seasonal pattern:

$$\hat{X}_{t+h|t} = \ell_t + hb_t + s_{t+h-s^*}, \qquad h = 1, 2, \ldots,$$

where s^* is chosen so that $t + h - s^*$ indexes the appropriate position within the seasonal cycle (e.g., $s^* = s$ for $h \leq s$, $s^* = 2s$ for $s < h \leq 2s$, etc.).

In our empirical application, we estimate the initial states $(\ell_0, b_0, s_{-s+1}, \ldots, s_0)$ and the smoothing parameters (α, β, γ) by minimizing the in-sample one-step-ahead squared forecast errors on the training sample, following the recommendations in Hyndman & Athanasopoulos (2021). The resulting ES model provides a simple linear benchmark that explicitly accounts for both trend and seasonality.

2.2.5. LSTM Networks

As a nonlinear benchmark, we consider long short-term memory (LSTM) networks, a class of recurrent neural networks designed to model sequences with potentially long-range temporal dependence (Hochreiter & Schmidhuber, 1997; Dixon

et al., 2020). From a time-series perspective, LSTMs can be viewed as nonlinear autoregressive models with a latent state that evolves dynamically and controls how past information is propagated over time (Dixon et al., 2020, Chapter 6).

Let $\{X_t\}$ denote the target time series and let x_t be the input vector at time t, typically containing recent lags of X_t (and, if desired, other covariates). An LSTM layer maintains two internal sequences: a hidden state $\mathbf{h}_t \in \mathbb{R}^m$ and a cell state $\mathbf{c}_t \in \mathbb{R}^m$, where m is the number of units in the layer. At each time t the LSTM updates these states through three gates and a candidate state:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_i x_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i)\,, \\
\mathbf{f}_t &= \sigma(W_f x_t + U_f \mathbf{h}_{t-1} + \mathbf{b}_f)\,, \\
\mathbf{o}_t &= \sigma(W_o x_t + U_o \mathbf{h}_{t-1} + \mathbf{b}_o)\,, \\
\tilde{\mathbf{c}}_t &= \tanh(W_c x_t + U_c \mathbf{h}_{t-1} + \mathbf{b}_c)\,,
\end{aligned}$$

where $\sigma(\cdot)$ denotes the logistic function applied element-wise, $W_\cdot$ and $U_\cdot$ are weight matrices and $\mathbf{b}_\cdot$ are bias vectors. The cell and hidden states are then updated as

$$\begin{aligned}
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t),
\end{aligned}$$

where $\odot$ denotes element-wise multiplication. The forget gate $\mathbf{f}_t$ controls how much of the previous cell state is retained, the input gate $\mathbf{i}_t$ regulates the contribution of the candidate state $\tilde{\mathbf{c}}_t$, and the output gate $\mathbf{o}_t$ determines how much of the internal state is exposed through the hidden state $\mathbf{h}_t$.

For forecasting, we feed the LSTM with sequences of length L constructed from the observed time series and use the final hidden state $\mathbf{h}_t$ to produce an h-step-ahead point forecast through a linear output layer,

$$\hat{X}_{t+h|t} = \mathbf{w}'\mathbf{h}_t + b,$$

where $\mathbf{w} \in \mathbb{R}^m$ and $b \in \mathbb{R}$ are output-layer parameters. All parameters of the LSTM (weights, biases, and output layer) are estimated by minimizing the MSE between forecasts and realizations on the training sample, using backpropagation through time and stochastic gradient-based optimization.

Hyperparameters are chosen via a grid search over a set of reasonable architectures (number of layers, number of neurons, lag length), and activation functions, among others. However, we considered an architecture with a single hidden layer to control model complexity and reduce the risk of overfitting, while still retaining sufficient approximation capacity to capture the relevant nonlinearities in the series. Additionally, the lag length was set based on Partial Auto-Correlation Function (PACF). We use 5 folds and select the specification with the lowest training MSE. This tuning procedure is applied separately for each country-variable pair, so that the LSTM benchmark is adapted to the dynamics of each series while remaining comparable to the statistical models in terms of forecast evaluation.

2.3. Forecast Construction

We treat each variable as a univariate time series $\{X_t\}_{t=1}^{T}$. For each country-variable pair, we split the sample into a training period and an evaluation period. The first t_0 observations are used to initialize parameter estimation and obtain the first set of forecasts, while the remaining observations are reserved for forecast evaluation. The exact values of t_0 and the number of evaluation points depend on the variable and sampling frequency and are described in the data subsection.

Forecasts are produced using a rolling-window scheme. For a given forecast horizon $h \in \{1, 2, 3, 4\}$ and forecast origin $j = t_0, \ldots, t_0 + \ell - 1$, where ℓ denotes the number of evaluation points for that horizon, we estimate each model on the window $\{X_{j-t_0+1}, \ldots, X_j\}$ and then compute the h-step-ahead forecast $\hat{X}_{j+h|j}$. The forecast error at horizon h from origin j is

$$e_{j+h|j} = X_{j+h} - \hat{X}_{j+h|j},$$

and the collection of errors $\{e_{j+h|j} : j = t_0, \ldots, t_0 + \ell - 1\}$ is used to evaluate forecasting performance. All models are estimated and evaluated under exactly the same rolling scheme for a given time series and horizon.

For quarterly time series, we consider horizons measured in quarters, while for monthly series, we consider horizons measured in months. In order to avoid the strong structural break associated with the COVID-19 pandemic, all samples are truncated in March 2020 so that the evaluation period excludes the most disruptive part of the pandemic. This design follows the recommendations in Tashman (2000); Zivot & Wang (2006) for rolling-origin forecast evaluation in time-series contexts.

2.4. Evaluation Criteria

In evaluating the performance of predictive models, We compare models using the MSE and the DM test for equal predictive accuracy. For a given horizon h and time series, the formula for MSE of a model is defined as

$$\mathrm{MSE}(h) = \frac{1}{\ell} \sum_{j=t_0}^{t_0+\ell-1} \left(e_{j+h|j}\right)^2,$$

where $e_{j+h|j} = X_{j+h} - \hat{X}_{j+h|j}$ is the forecast error defined in the previous subsection and ℓ is the number of forecast origins used for evaluation at horizon h. Lower values of $\mathrm{MSE}(h)$ indicate better average point forecast performance at that horizon. This rolling backtest is a standard design for time series forecast evaluation and can be viewed as a special case of time series cross-validation (Tashman, 2000; Zivot & Wang, 2006).

To formally compare two competing models, say model 1 and model 2, we use the DM test of Diebold & Mariano (1995). Let $e_{j+h|j,1}$ and $e_{j+h|j,2}$ denote their forecast errors at horizon h from origin j, and define the loss differential

$$d_j = (e_{j+h|j,1})^2 - (e_{j+h|j,2})^2, \qquad j = t_0, \ldots, t_0 + \ell - 1.$$

The null hypothesis of equal predictive ability is

$$H_0 : \mathbb{E}(d_j) = 0,$$

which states that the two models have the same expected squared forecast error at horizon h. The sample mean of the loss differential is

$$\bar{d} = \frac{1}{\ell} \sum_{j=t_0}^{t_0+\ell-1} d_j,$$

and an estimator of $\mathrm{Var}(\bar{d})$ that is robust to h-dependence in $\{d_j\}$ is obtained via a Newey-West long-run variance estimator (Diebold & Mariano, 1995). The DM statistic is then

$$DM(h) = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}},$$

which under H_0 and suitable regularity conditions satisfies

$$DM(h) \xrightarrow{d} \mathcal{N}(0,1).$$

We report the decision based on p-values associated with $DM(h)$ for each pairwise comparison (Section 3.). Small p-values (e.g., below conventional 5% or 10% thresholds) indicate statistically significant differences in forecast accuracy. Given the relatively small number of forecast origins and the large number of pairwise comparisons, we interpret these p-values as descriptive evidence of predictive differences rather than as the outcome of formal multiple-testing procedures.

In addition to numerical criteria, we inspect residual diagnostics based on normality, serial autocorrelation, and stability tests summarised in appendix. Although our main focus is on MSE and DM test results, these diagnostics help to assess whether the fitted models are broadly compatible with the observed data.

3. Results

Estimation, forecast construction, and model comparison were carried out following standard practice in the time series literature. For each country and macroeconomic variable, we specified the four competing models introduced in Subsection 2.2 using the data described in Subsection 2.1, identified the (non-structural) parameters, and conducted residual–based diagnostic checks. Detailed model specifications and diagnostics are reported in appendix. This section focuses on the out-of-sample forecasting performance of the competing models, evaluated using the rolling multi-step design and accuracy measures described in Subsections 2.3 and 2.4.

The appendix summarizes, for each time series and model, the selected number of regimes, autoregressive orders, seasonal and threshold lags, delay values, as well as the main LSTM hyperparameters (input lag, number of neurons, and activation function); see Tables A1, A3, and A5. It also reports the outcomes of

the residual–based diagnostic tests in Tables A2, A4, and A6: the Jarque-Bera test (normality), Ljung-Box test (no serial autocorrelation), CUSUM test (parameter stability and correct specification), and CUSUMSQ test (absence of marginal heteroskedasticity). Bold p-values indicate rejection of the corresponding null hypothesis at the 5% level. All the computations are performed using the R software.

GDP

For Colombia, we analyze the Monthly Economic Activity Index (ISE) as a high-frequency proxy for GDP. Model specifications for the GDP (ISE) series in Colombia, the USA, and the UK are reported in Table A1. For TSARX, we select two or three regimes, with low autoregressive and seasonal orders and threshold delays between 2 and 3 periods. The TAR model also uses two or three regimes, but with relatively high autoregressive orders in Colombia (e.g., $k = 26$), reflecting the persistence of the ISE time series. The SAR specifications employ seasonal autoregressive orders $k = 2$ with seasonal period $K = 2$ across countries. The LSTM architectures use relatively long input lags (5 to 26 observations) and between 75 and 100 neurons with ReLU activation.

Residual diagnostics for GDP are summarised in Table A2. For Colombia and the UK, TSARX and TAR yield residuals that are approximately normal and uncorrelated, with no evidence of parameter instability or marginal heteroskedasticity according to the CUSUM and CUSUMSQ tests; they therefore provide an overall adequate in-sample fit. In Colombia, SAR departs from normality (JB test $p < 0.01$) but otherwise passes the autocorrelation and stability checks, while LSTM residuals depart from normality and indicate variance instability (CUSUMSQ test). In the USA, TSARX and TAR show no evidence of autocorrelation or parameter instability but both exhibit marginal heteroskedasticity; in addition, TAR departs from normality. SAR presents normality issues in Colombia and the UK, autocorrelation in the UK and USA, and heteroskedasticity in the USA. LSTM residuals depart from normality in all three countries, display autocorrelation in the UK, and indicate heteroskedasticity in the USA. Overall, TSARX achieves a broadly satisfactory in-sample fit across countries, although in the USA no model fully captures the variance dynamics.

All forecasting exercises use rolling-origin multi-step-ahead predictions at horizons $h = 1, \ldots, 4$. Table 2 reports the MSE by country, model, and forecast horizon. For Colombia (ISE), SAR delivers the lowest MSE at horizon $h = 1$, ES at $h = 2$ and $h = 4$, and LSTM at $h = 3$. For the USA, SAR performs best at horizons $h = 1$ and $h = 2$, whereas TAR and TSARX achieve the lowest MSE at $h = 3$ and $h = 4$, respectively. For the UK, LSTM is the most accurate model at horizons $h = 1$, $h = 3$, and $h = 4$, while SAR performs best at $h = 2$. These results confirm that even for a single macroeconomic variable, the ranking of models depends on the country and forecast horizon.

To assess whether these MSE differences are statistically significant, Table 3 summarises DM test results using TSARX as the reference model. Most pairwise comparisons are not statistically distinguishable (N.D.). The exceptions are: (i)

in the USA and the UK at $h = 1$, TSARX is significantly more accurate than ES; (ii) in Colombia at $h = 1$, TSARX is significantly more accurate than LSTM; and (iii) in the UK at $h = 2$, LSTM is significantly more accurate than TSARX. Thus, for GDP/ISE series, TSARX is competitive relative to the benchmarks, but the gains or losses in MSE are rarely statistically decisive.

Table 2: MSE for each forecast horizon and country for GDP. For Colombia, GDP is proxied by the ISE.

		MSE			
Country	Model	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Colombia	TSARX	9.6342	12.9723	13.2598	14.0732
	TAR	9.7868	13.4796	13.6463	14.9369
	SAR	**9.6068**	13.1489	13.4553	14.6034
	ES	10.5208	**11.8631**	12.7714	**13.7887**
	LSTM	13.2098	13.0887	**12.2006**	15.1599
USA	TSARX	0.6403	0.6320	0.6343	**0.6684**
	TAR	0.5949	0.6412	**0.6206**	0.6685
	SAR	**0.5785**	**0.6228**	0.6845	0.7261
	ES	6.8005	0.6352	0.7321	0.7289
	LSTM	0.8307	0.6643	1.0613	0.8262
UK	TSARX	1.0447	1.8736	1.5484	1.8211
	TAR	1.2111	2.1411	1.8057	2.0320
	SAR	1.0930	**1.8081**	1.5685	1.7569
	ES	2.1245	1.8164	1.5570	1.5756
	LSTM	**0.6726**	18.4725	**0.6638**	**0.9996**

Table 3: DM test conclusions (GDP). N.D.: no statistically significant difference.

Country	Compared models	Forecast horizon			
		$h = 1$	$h = 2$	$h = 3$	$h = 4$
Colombia	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	**TSARX**	N.D.	N.D.	N.D.
USA	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	**TSARX**	N.D.	N.D.	N.D.
	TSARX-LSTM	N.D.	N.D.	N.D.	N.D.
UK	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	**TSARX**	N.D.	N.D.	N.D.
	TSARX-LSTM	N.D.	**LSTM**	N.D.	N.D.

Unemployment rate

Model specifications for the unemployment rate time series are reported in Table A3. The TSARX models again rely on two or three regimes with low autoregressive and seasonal orders and delays between 2 and 3 periods. The TAR models exhibit relatively high autoregressive orders for Colombia (e.g., $k = 25$)

and more parsimonious orders in the USA and UK. SAR specifications use seasonal orders between $k = 1$ and $k = 2$, and the LSTM architectures employ input lags between 5 and 13 observations, with 30 to 100 neurons and either tanh or ReLU activations.

Residual diagnostics for unemployment are presented in Table A4. For Colombia, TSARX, SAR, and LSTM show no evidence of non-normality or autocorrelation (JB test and LB test p-values comfortably above 0.05), and pass the CUSUM stability test; however, all three display variance instability according to CUSUMSQ test. TAR passes all four residual checks and thus provides the most satisfactory in-sample fit in Colombia. In the USA, TSARX and TAR pass all diagnostic tests, whereas SAR and LSTM exhibit non-normal and autocorrelated residuals (small p-values), with SAR also showing variance instability. In the UK, TSARX and SAR pass all tests, TAR departs from normality, and LSTM departs from both normality and no-autocorrelation, with evidence of variance instability.

Table 4 reports the MSE by country, model, and forecast horizon for the unemployment rate. For Colombia, TSARX delivers the lowest MSE at horizons $h = 1$ and $h = 4$, TAR at $h = 2$, and ES at $h = 3$. For the USA, LSTM performs best at one step ahead, while ES yields the smallest MSE at horizons $h = 2$, $h = 3$, and $h = 4$. For the UK, TAR achieves the lowest MSE at horizons $h = 1$ and $h = 4$, whereas SAR is the most accurate model at $h = 2$ and $h = 3$. Again, the ranking of models is sensitive to the country and forecast horizon.

Table 4: MSE for each forecast horizon and country for unemployment rate.

Country	MSE				
	Model	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Colombia	TSARX	**0.3047**	0.8089	0.4012	**0.3701**
	TAR	0.3701	**0.3295**	0.4425	0.4719
	SAR	0.3897	0.3920	0.4727	0.5714
	ES	0.3319	0.3519	**0.3780**	2.2416
	LSTM	1.7519	1.3101	0.6247	0.6620
USA	TSARX	0.0661	0.0438	0.0396	0.0434
	TAR	0.0558	0.0535	0.0535	0.0538
	SAR	0.0656	0.0448	0.0378	0.0416
	ES	0.0425	**0.0428**	**0.0300**	**0.0321**
	LSTM	**0.0411**	0.0817	0.0861	0.0989
UK	TSARX	0.0151	0.0151	0.0120	0.0219
	TAR	**0.0073**	0.0088	0.0092	**0.0090**
	SAR	0.0082	**0.0087**	**0.0085**	0.0103
	ES	0.0154	0.0149	0.0150	0.0113
	LSTM	0.0145	0.0192	0.0182	0.0341

The DM comparisons in Table 3, with TSARX as the benchmark, show that most contrasts are not statistically distinguishable. The main exceptions are: (i) in Colombia at $h = 4$, TSARX significantly outperforms SAR; and (ii) in the USA, TSARX significantly outperforms LSTM at horizons $h = 2$, $h = 3$, and $h = 4$. No statistically significant differences among the models are detected for the UK across horizons. Overall, for unemployment rate, TSARX remains a competitive alternative in multi-step forecasting—particularly relative to LSTM in the USA—while ES often attains the smallest MSE in that country.

Table 5: DM test conclusions (unemployment rate). N.D.: no statistically significant difference.

Country	Compared models	Forecast horizon (h steps ahead)			
		$h = 1$	$h = 2$	$h = 3$	$h = 4$
Colombia	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	**TSARX**
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	N.D.	N.D.	N.D.	N.D.
USA	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	N.D	**TSARX**	**TSARX**	**TSARX**
UK	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D	N.D.	N.D.	N.D.
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	N.D.	N.D.	N.D.	N.D.

Inflation

Model specifications for the inflation time series are summarised in Table A5. As in the previous cases, the TSARX models use two or three regimes with low autoregressive and seasonal orders and short threshold delays. The TAR models feature moderate to high autoregressive orders (e.g., $k = 18$ for the UK), SAR specifications employ low seasonal orders, and the LSTM architectures use input lags of 13 or 26 observations with between 15 and 100 neurons and ReLU activation.

Residual diagnostics for inflation are reported in Table A6. For Colombia, TSARX and LSTM show no evidence of non-normality or autocorrelation and pass the CUSUM stability test; however, LSTM suggests variance instability according to CUSUMSQ test. TAR and SAR present departures from normality but otherwise pass the autocorrelation and stability checks. For the USA, TSARX, TAR, and SAR exhibit non-normal residuals and indications of variance instability (CUSUMSQ test), with TAR and SAR also showing autocorrelation; LSTM departs from normality but passes the autocorrelation and stability tests. For the UK, TSARX passes all checks, SAR fails variance stability only (CUSUMSQ test), TAR departs from normality and displays both autocorrelation and variance instability, and LSTM shows autocorrelation but otherwise passes the remaining diagnostics.

Table 6 reports the MSE by country, model, and forecast horizon for inflation. For Colombia, SAR delivers the lowest MSE at horizon $h = 1$, TAR at $h = 2$, and LSTM at $h = 3$ and $h = 4$. For the USA, ES achieves the smallest MSE at all horizons, from $h = 1$ to $h = 4$. For the UK, TAR performs best at horizons $h = 1$ and $h = 4$, whereas ES yields the lowest MSE at $h = 2$ and $h = 3$. Thus, for inflation, ES is particularly competitive in the USA and UK, while LSTM dominates at longer horizons in Colombia.

TABLE 6: MSE for each forecast horizon and country for inflation.

Country	MSE				
	Model	$h=1$	$h=2$	$h=3$	$h=4$
Colombia	TSARX	0.0225	0.0179	0.0206	0.0206
	TAR	0.0274	**0.0116**	0.0142	0.0175
	SAR	**0.0147**	0.0213	0.0250	0.0263
	ES	0.0241	0.0221	0.0207	0.0226
	LSTM	0.0156	0.0145	**0.0123**	**0.0127**
USA	TSARX	0.0609	0.0805	0.0798	0.0762
	TAR	0.0703	0.1041	0.1205	0.0738
	SAR	0.0554	0.0728	0.0733	0.0697
	ES	**0.0523**	**0.0542**	**0.0543**	**0.0509**
	LSTM	0.0753	0.0962	0.1044	0.0866
UK	TSARX	0.0328	0.0321	0.0317	0.0322
	TAR	**0.0318**	0.0325	0.0319	**0.0314**
	SAR	0.0332	0.0324	0.0318	0.0321
	ES	0.0322	**0.0310**	**0.0310**	0.0317
	LSTM	0.6709	0.6172	0.7155	0.6307

Finally, Table 7 presents DM test conclusions for inflation. Most contrasts relative to TSARX are again not statistically distinguishable. Notable exceptions are: (i) in the USA, TSARX is significantly more accurate than TAR at $h = 2$ and $h = 3$, and more accurate than LSTM at all horizons; and (ii) in the UK, TSARX is significantly more accurate than LSTM at all horizons. By contrast, TSARX versus ES is classified as N.D. in both the USA and the UK, despite ES attaining the smallest MSE there. This underscores that differences in point forecast accuracy need not translate into statistically significant gains. When DM tests are inconclusive, model choice should also consider parsimony, interpretability, and residual diagnostics.

TABLE 7: DM test conclusions (inflation). N.D.: no statistically significant difference.

Country	Compared models	Forecast horizon			
		$h=1$	$h=2$	$h=3$	$h=4$
Colombia	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	N.D.	N.D.	N.D.	N.D.
USA	TSARX-TAR	N.D.	**TSARX**	**TSARX**	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	**TSARX**	**TSARX**	**TSARX**	**TSARX**
UK	TSARX-TAR	N.D.	N.D.	N.D.	N.D.
	TSARX-SAR	N.D.	N.D.	N.D.	N.D.
	TSARX-ES	N.D.	N.D.	N.D.	N.D.
	TSARX-LSTM	**TSARX**	**TSARX**	**TSARX**	**TSARX**

4. Conclusions

This paper has presented an empirical assessment of the multiplicative seasonal threshold autoregressive model with exogenous inputs (TSARX) for nonlinear macroeconomic time series forecasting. Using seasonally unadjusted monthly and quarterly time series of GDP, unemployment rate, and inflation for Colombia, the United States of America, and the United Kingdom, and comparing TSARX with four seasonal benchmarks (TAR, SAR, Holt-Winters exponential smoothing (ES), and LSTM networks), we evaluated multi-step forecasting performance at horizons from one to four periods ahead under a common rolling-window design.

The results show that TSARX is not a uniformly superior forecasting method in this setting. Across 36 series–country–horizon combinations, TSARX achieves the lowest Mean Squared Error (MSE) in 3 cases, while the benchmark models (TAR, SAR, ES, and LSTM) attain the smallest MSE in the remaining cases. According to Diebold-Mariano tests (DM), TSARX is often statistically indistinguishable from the best-performing benchmark, and in some instances significantly outperforms LSTM or TAR. However, in many situations simpler seasonal models, particularly Holt–Winters exponential smoothing (ES), remain difficult to beat. This finding is consistent with evidence from large-scale forecast competitions, where it is well documented that no single model dominates across a wide variety of time series and evaluation schemes (Hyndman & Koehler, 2006; Makridakis et al., 2020).

At the same time, our empirical exercises highlight that explicitly modelling both nonlinear (threshold) dynamics and multiplicative seasonality can be beneficial in particular macroeconomic contexts, especially for some unemployment rate and inflation time series and at specific forecast horizons. TSARX offers a flexible yet interpretable framework in which regime-dependent behaviour and seasonal effects can be analysed jointly, and our results document when this additional structure translates into improvements in forecast accuracy and when it does not.

From a practical perspective, our findings suggest that TSARX should be viewed as a useful member of a forecasting toolkit rather than as a general replacement for established seasonal models. For applications similar to those considered here, we recommend comparing TSARX with simpler benchmarks using an evaluation design akin to the one implemented in this paper, rather than adopting it by default. Future work should extend the comparison to density and interval forecasts and to alternative accuracy measures such as the Continuous Ranked Probability Score (CRPS) and the Mean Absolute Error (MAE), explore other nonlinear and machine-learning algorithms and alternative seasonal structures, and consider longer samples and additional macroeconomic variables.

[Received: September 2025 — Accepted: November 2025]

References

Arbeláez Quintero, S. (2022), 'Pronósticos en series de tiempo no lineales: aplicación del modelo tsarx y comparación con modelos para datos estacionales', Tesis de Maestría, Repositorio Universidad Nacional de Colombia. https://repositorio.unal.edu.co/handle/unal/83875

Brown, R. G. (1959), *Statistical Forecasting for Inventory Control*, McGraw-Hill.

Brown, R. L., Durbin, J. & Evans, J. M. (1975), 'Techniques for testing the constancy of regression relationships over time', *Journal of the Royal Statistical Society: Series B (Methodological)* **37**(2), 149–192.

Chan, K. S. (1993), 'Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model', *The Annals of Statistics* **21**(1), 520–533.

Clements, M. P. & Krolzig, H.-M. (1998), 'A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP', *The Econometrics Journal* **1**(1), C47–C75.

DANE (2015), 'Metodología general: Gran encuesta integrada de hogares (geih)', Documento metodológico.

DANE (2016), 'Metodología general: Indicador de seguimiento a la economía (ise)', Documento metodológico.

De Gooijer, J. G. & Vidiella-i Anguera, A. (2003), 'Nonlinear stochastic inflation modelling using seasetars', *Insurance: Mathematics and Economics* **32**(1), 27–36.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **13**(3), 253–263.

Dixon, M. F., Halperin, I. & Bilokon, P. (2020), *Machine Learning in Finance: From Theory to Practice*, Springer, Cham.

Edgerton, D. & Wells, C. (1994), 'Critical values for the CUSUMSQ statistic in medium and large sized samples', *Oxford Bulletin of Economics and Statistics* **56**(3), 355–365.

Franses, P. H. & van Dijk, D. (2000), *Nonlinear Time Series Models in Empirical Finance*, Cambridge University Press, Cambridge.

Ghysels, E., Osborn, D. R. & Rodrigues, P. M. M. (2006), Forecasting seasonal time series, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting', Vol. 1, Elsevier, pp. 659–711.

González, J. & Nieto, F. H. (2020), 'Bayesian analysis of multiplicative seasonal threshold autoregressive processes', *Revista Colombiana de Estadística* **43**(2), 251–285.

Hansen, B. E. (2011), 'Threshold autoregression in economics', *Statistics and Its Interface* **4**, 123–127.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**(8), 1735–1780.

Holt, C. C. (1957), Forecasting seasonals and trends by exponentially weighted averages, *in* 'Office of Naval Research Memorandum', Carnegie Institute of Technology, Pittsburgh.

Hyndman, R. J. & Athanasopoulos, G. (2021), *Forecasting: Principles and Practice*, 3rd edn, OTexts, Melbourne. https://otexts.com/fpp3/

Hyndman, R. J. & Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**(4), 679–688.

Lee, Y.-M. & Wang, K.-M. (2012), 'Searching for a better proxy for business cycles: With supports using us data', *Applied Economics* **44**(11), 1433–1442.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2020), 'The m4 competition: 100,000 time series and 61 forecasting methods', *International Journal of Forecasting* **36**(1), 54–74.

Milas, C., Rothman, P. & van Dijk, D. (2006), *Nonlinear Time Series Analysis of Business Cycles*, Elsevier, Amsterdam.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S. & Tiao, G. C. (1998), 'Forecasting the U.S. unemployment rate', *Journal of the American Statistical Association* **93**(442), 478–493.

Tashman, L. J. (2000), 'Out-of-sample tests of forecasting accuracy: An analysis and review', *International Journal of Forecasting* **16**(4), 437–450.

Tong, H. (1990), *Non-linear Time Series: A Dynamical System Approach*, Vol. 6 of *Oxford Statistical Science Series*, Oxford University Press, Oxford.

Tong, H. & Lim, K. S. (1980), 'Threshold autoregression, limit cycles, and cyclical data', *Journal of the Royal Statistical Society: Series B (Methodological)* **42**(3), 245–292.

Tsay, R. S. (1989), 'Testing and modeling threshold autoregressive processes', *Journal of the Royal Statistical Society: Series B (Methodological)* **51**(1), 231–246.

Tsay, R. S. (1998), 'Testing and modeling multivariate threshold models', *Journal of the American Statistical Association* **93**(443), 1188–1202.

Vaca, P. A. (2018), Analysis of the forecasting performance of the threshold autoregressive model, Master's thesis, Universidad Nacional de Colombia.

Winters, P. R. (1960), 'Forecasting sales by exponentially weighted moving averages', *Management Science* **6**(3), 324–342.

Yilmaz, F. M. & Arabaci, O. (2021), 'Should deep learning models be in high demand, or should they simply be a very hot topic? a comprehensive study for exchange rate forecasting', *Computational Economics* **57**(1), 217–245.

Zivot, E. & Wang, J. (2006), *Modeling Financial Time Series with S-PLUS*, 2nd edn, Springer, New York, NY.

Appendix

In this appendix, we present the identification of the structural parameters of the AR, TAR, TSARX, and LSTM models. We also report model validation based on residual diagnostics, assessing whether the assumptions of normality (Jarque-Bera test), absence of serial autocorrelation (Ljung-Box test), correct model specification (CUSUM test), and no marginal heteroskedasticity (CUSUMSQ test) are satisfied. Also, theoretical considerations regarding the statistical tests employed are provided.

Model Specification and Residual-Based Model Validation

TABLE A1: Identification of the structural parameters of the candidate models for GDP.

Country	TSARX				TAR			SAR		LSTM		
	l	k	K	d	l	k	d	k	K	Lag	Neurons	Activation
Colombia	2	2	2	2	2	26	2	2	2	26	100	Relu
USA	2	2	2	3	2	5	0	2	2	5	75	Relu
UK	3	2	2	3	3	10	3	2	2	5	100	Relu

TABLE A2: Some statistical tests for residuals of the fitted models for GDP.

Country	Model	JB test (p-value)	LB test (p-value)	CUSUM test	CUSUMSQ test
Colombia	TSARX	0.8872	0.4967	✓	✓
	TAR	0.9676	0.4903	✓	✓
	SAR	**0.0042**	0.14730	✓	✓
	LSTM	**0.0126**	0.1025	✓	×
USA	TSARX	0.9774	0.3216	✓	×
	TAR	**0.0194**	0.0673	✓	×
	SAR	0.1607	**0.0339**	✓	×
	LSTM	<**0.0001**	0.1211	✓	×
UK	TSARX	0.8724	0.3994	✓	✓
	TAR	0.6458	0.3785	✓	✓
	SAR	<**0.0001**	**0.0150**	✓	✓
	LSTM	<**0.0001**	**0.0352**	✓	✓

Note: A ✓indicates that the corresponding null hypothesis is not rejected, i.e., correct model specification (CUSUM test) and no marginal heteroskedasticity (CUSUMSQ test).

Table A3: Identification of the structural parameters of the candidate models for unemployment rate.

Country	TSARX				TAR			SAR		LSTM		
	l	k	K	d	l	k	d	k	K	Lag	Neurons	Activation
Colombia	2	2	2	2	2	25	2	2	2	13	100	tanh
USA	2	2	2	3	3	5	0	2	1	5	100	Relu
UK	3	2	2	3	3	5	0	2	2	5	30	Relu

Table A4: Some statistical tests for residuals of the fitted models for unemployment rate.

Country	Model	JB test (p-value)	LB test (p-value)	CUSUM test	CUSUMSQ test
Colombia	TSARX	0.7367	0.5108	✓	×
	TAR	0.6590	0.6032	✓	✓
	SAR	0.4763	0.7698	✓	×
	LSTM	0.7296	0.7774	✓	×
USA	TSARX	0.1111	0.1314	✓	✓
	TAR	0.3486	0.1854	✓	✓
	SAR	**0.0011**	**0.0048**	✓	×
	LSTM	**0.0066**	**0.0210**	✓	✓
UK	TSARX	0.3532	0.5690	✓	✓
	TAR	**0.0013**	0.4577	✓	✓
	SAR	0.1439	0.2241	✓	✓
	LSTM	**0.0001**	**0.0001**	✓	×

Note: A ✓indicates that the corresponding null hypothesis is not rejected, i.e., correct model specification (CUSUM test) and no marginal heteroskedasticity (CUSUMSQ test).

Table A5: Identification of the structural parameters of the candidate models for inflation.

Country	TSARX				TAR			SAR		LSTM		
	l	k	K	d	l	k	d	k	K	Lag	Neurons	Activation
Colombia	2	1	2	2	2	13	1	1	2	13	15	Relu
USA	2	2	2	3	2	25	3	2	2	26	30	Relu
UK	3	2	1	1	3	18	1	1	1	13	100	Relu

TABLE A6: Some statistical tests for residuals of the fitted models for inflation.

Country	Model	JB test (p-value)	LB test (p-value)	CUSUM test	CUSUMSQ test
Colombia	TSARX	0.1649	0.7337	✓	✓
	TAR	**0.0063**	0.6782	✓	✓
	SAR	**0.0378**	0.3826	✓	✓
	LSTM	0.1943	0.7677	✓	×
USA	TSARX	**<0.001**	0.0587	✓	×
	TAR	**<0.001**	**0.0088**	✓	×
	SAR	**<0.001**	**0.0251**	✓	×
	LSTM	**0.0002**	0.2181	✓	✓
UK	TSARX	0.6025	0.5191	✓	✓
	TAR	**0.0262**	**0.0022**	✓	×
	SAR	0.6573	0.1219	✓	×
	LSTM	0.7492	**0.0140**	✓	✓

Note: A ✓indicates that the corresponding null hypothesis is not rejected, i.e., correct model specification (CUSUM test) and no marginal heteroskedasticity (CUSUMSQ test).

Diagnostic Tests

For completeness, we briefly summarize the diagnostic tests used in the empirical analysis. Let $\{\hat{\epsilon}_t\}$ denote the residuals from a fitted model. Standard diagnostic tests can be applied to $\{\hat{\epsilon}_t\}$, such as the Ljung-Box test to determine autocorrelation, the Jarque-Bera test to verify normality, the CUSUM test to determine the correct specification of the model, and the CUSUMSQ test to check for marginal heteroscedasticity.

Jarque-Bera normality test. The Jarque-Bera (JB) test assesses whether the residuals are consistent with a normal distribution by jointly considering skewness and kurtosis. If $\hat{S}$ and $\hat{K}$ denote the sample skewness and sample kurtosis of $\{\hat{\epsilon}_t\}$, the test statistic is

$$JB = \frac{n}{6}\left(\hat{S}^2 + \frac{(\hat{K}-3)^2}{4}\right),$$

where n is the sample size. Under the null hypothesis of normality and for large n, JB is asymptotically distributed as a chi-square random variable with two degrees of freedom. Large values of JB (small p-values) indicate departures from normality.

Ljung-Box autocorrelation test. The Ljung-Box (LB) test is used to detect residual autocorrelation at multiple lags. For a chosen maximum lag m and sample autocorrelations $\hat{\rho}_k$ of the $\{\hat{\epsilon}_t\}$, the test statistic is

$$Q_{\mathrm{LB}}(m) = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k},$$

where n is the sample size.

Under the null hypothesis that the residuals are white noise and for large n, $Q_{\mathrm{LB}}(m)$ is approximately chi-square distributed with $(m-p)$ degrees of freedom, where p denotes the number of parameters effectively estimated in the model. Large values of $Q_{\mathrm{LB}}(m)$ (small p-values) suggest serial autocorrelation in the residuals.

CUSUM and CUSUMSQ stability tests. The CUSUM and CUSUMSQ tests monitor the cumulative behaviour of residuals to detect structural instability in the regression relationship (Brown et al., 1975; Edgerton & Wells, 1994). Let $\{\hat{u}_t\}$ denote recursive residuals from a fitted model. The CUSUM process is defined as the cumulative sum of recursive residuals, suitably scaled by an estimate of their variance, and the CUSUMSQ process is defined as the cumulative sum of squared recursive residuals. In both cases the test compares the observed cumulative process with boundaries that correspond to a chosen significance level under the null hypothesis of parameter constancy.

If the CUSUM or CUSUMSQ trajectory remains within the critical bands, the data are consistent with stable parameters over time. Systematic excursions outside the bands provide evidence of structural change.