Department of Economics- FEA/USP

NEWS IV: A model with news and implied volatility for enhanced volatility prediction

Vitor Gentini

Márcio Issao Nakane

Working Paper Series Nº 2026-25

DEPARTMENT OF ECONOMICS, FEA-USP
WORKING PAPER Nº 2026-25

NEWS IV: A model with news and implied volatility for enhanced volatility prediction

Vitor Gentini (vitorgentini@hotmail.com)

Márcio Issao Nakane (minakane@usp.br)

Abstract:

This paper examines whether implied volatility and textual news jointly improve volatility forecasting. We propose the NEWS IV model, which combines the realized-variance components of the HAR model with option-implied variance and news topics extracted via Latent Dirichlet Allocation within a flexible machine learning framework. Using data for the Ibovespa ETF and major Brazilian stocks, we evaluate predictive performance relative to HAR-type benchmarks across multiple horizons. We show that the model augmented with implied volatility and news delivers performance comparable to standard models at the daily horizon and improves forecasts at weekly and monthly horizons. The results reveal a clear horizon-dependent pattern. Implied volatility plays a central role in short-term predictions, while news-based variables become increasingly relevant at longer horizons. These findings highlight the complementary informational content of market expectations and textual data for understanding volatility dynamics.

Keywords: Volatility forecasting, news analysis, implied volatility, realized volatility, Brazilian equity market

JEL Codes: C22, C53, C58, G17,

NEWS IV: A MODEL WITH NEWS AND IMPLIED VOLATILITY FOR ENHANCED VOLATILITY PREDICTION

VÍTOR GENTINI
Department of Economics, University of São Paulo and EPGE Brazilian School of Economics and Finance

MÁRCIO ISSAO NAKANE
Department of Economics, University of São Paulo

This paper examines whether implied volatility and textual news jointly improve volatility forecasting. We propose the NEWS IV model, which combines the realized-variance components of the HAR model with option-implied variance and news topics extracted via Latent Dirichlet Allocation within a flexible machine learning framework. Using data for the Ibovespa ETF and major Brazilian stocks, we evaluate predictive performance relative to HAR-type benchmarks across multiple horizons. We show that the model augmented with implied volatility and news delivers performance comparable to standard models at the daily horizon and improves forecasts at weekly and monthly horizons. The results reveal a clear horizon-dependent pattern. Implied volatility plays a central role in short-term predictions, while news-based variables become increasingly relevant at longer horizons. These findings highlight the complementary informational content of market expectations and textual data for understanding volatility dynamics.

KEYWORDS: Volatility forecasting, news analysis, implied volatility, realized volatility, Brazilian equity market.

1. INTRODUCTION

Volatility is a fundamental object in financial economics. It determines option prices, governs portfolio allocation, shapes risk management practices, and enters regulatory capital requirements. Because volatility is latent and time-varying, forecasting it accurately remains a central empirical challenge. Over the last two decades, advances in high-frequency data have enabled increasingly precise measurement of realized volatility, while derivatives markets have provided forward-looking measures through implied volatility. At the same time, the digitalization of information has made textual news a measurable and potentially powerful driver of market dynamics. Yet we still lack a unified understanding of how these distinct sources of information (historical volatility, option-implied expectations, and structured news) interact in forecasting volatility across horizons.

Vítor Gentini: vitor.gentini@fgv.edu.br
Márcio Issao Nakane: minakane@usp.br
We thank Rodrigo De-Losso, Marcelo Fernandes, Flávia Graminho, Leandro Maciel and Flávio Ziegelmann for their insightful comments and feedbacks.
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

This paper studies whether and how implied volatility and textual news jointly improve volatility forecasts. We propose a model, NEWS IV, that integrates option-implied variance and news topics extracted via Latent Dirichlet Allocation (LDA) with the realized-variance dynamics of the HAR model, and we estimate it using random forests to accommodate non-linearities and high dimensionality. We evaluate its performance relative to the heterogeneous autoregressive (HAR) class of models and their extensions across daily, weekly, and monthly horizons.

Our empirical setting is the Brazilian equity market, focusing on the BOVA11 ETF and five of the most liquid stocks with available implied volatility data. Brazil serves as a high-variance laboratory for studying volatility dynamics. While its market capitalization is smaller than that of the U.S., it is characterized by structurally higher return volatility and frequent exposure to macroeconomic and political shocks. These exogenous shocks, combined with a concentrated index and heavy commodity dependency, result in a volatility profile marked by extreme skewness and frequent jumps, as evidenced by the 2018 truck drivers' strike and the COVID-19 pandemic. The high frequency of these non-linear shocks makes Brazil a uniquely suited environment to test our model. If forward-looking option data and textual news contain predictive signals, their impact should be most discernible in a market where historical volatility alone often fails to capture sudden, event-driven shifts.

Our work builds on and extends the high-frequency volatility forecasting literature centered on realized variance (Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002a), and Barndorff-Nielsen and Shephard (2002b)). The heterogeneous autoregressive (HAR) model of Corsi (2009) has become the workhorse specification for capturing the long-memory properties of volatility. Subsequent extensions include treatment of jumps via bipower variation (HARCJ, Andersen et al. (2007)), leverage effects (LHARCJ, Corsi and Reno (2012)), and realized semivariances (SHAR, Patton and Sheppard (2015)). These models have been shown to improve forecast accuracy across asset classes and markets.

A parallel literature examines the predictive content of implied volatility. Because implied volatility embeds risk-neutral expectations of future uncertainty, it is a natural forward-looking predictor of the underlying quadratic variation. Bekaert and Hoerova (2014) show that implied volatility significantly predicts S&P 500 realized volatility using a general-to-specific approach. Kambouroudis et al. (2021) find that implied volatility adds predictive content beyond HAR for international equity indices using model confidence set procedures, while Christensen et al. (2023) find that machine learning models often prioritize implied volatility over historical measures when forecasting realized variation. Evidence for the Brazilian market is more limited and mixed. Astorino et al. (2017) document predictive gains from implied volatility for the Ibovespa index, while Asgharian et al. (2023) report weaker improvements when extending HAR with VIX-type measures. In these empirical settings, implied volatility serves as a forward-looking regressor to capture the latent quadratic variation through its observable realized counterparts.

More recently, textual analysis has emerged as a promising tool in volatility forecasting. Following Antweiler and Frank (2004), financial research has increasingly incorporated sentiment and news-based measures. Audrino et al. (2020) show that augmenting HAR models with sentiment variables from news and social media improves volatility forecasts for U.S. equities. Bodilsen and Lunde (2025) extend HAR using macroeconomic and firm-specific news sentiment via a MIDAS approach and document improved performance for U.S. stocks. In Brazil, Fernandes and Pereira (2025) extend LHARCJ with news-based uncertainty indices constructed from major newspapers and find forecast improvements for both the BOVA11 ETF and large Brazilian firms.

Our contribution differs from and complements this literature in three important ways.

First, we integrate implied variance and structured news topics within a single forecasting framework and estimate the model using random forests. Rather than augmenting a HAR-type specification with a single additional regressor, we construct a model that nests the HAR information set within a high-dimensional set of option-implied variance and LDA-derived news topics, with the implied-variance terms available at every split. We show that this model matches or exceeds the predictive performance of HAR-type models at horizons beyond one day. At the 22-day horizon, NEWS IV reduces mean squared forecast errors by approximately 58% on average across the six assets relative to the HAR benchmark, and by approximately 69% across the five individual stocks.

Second, we document a clear horizon-dependent pattern in the relative importance of information sources. Implied variance plays a dominant role in short-term forecasts, consistent with its interpretation as a forward-looking measure of uncertainty. In contrast, textual news topics become increasingly important at weekly and monthly horizons. This finding suggests that news captures slower-moving volatility components linked to macroeconomic, political, and sectoral developments, features particularly salient in emerging markets.

Third, we show that nonlinear modeling is crucial when integrating high-dimensional textual information. While random forests applied to standard HAR variables do not materially improve forecasts, flexible machine learning methods become essential when combining implied variance with dozens of news topics. The forecasting gains arise not only from richer information but also from modeling interactions between option-implied expectations and evolving news narratives.

Studying Brazil is not merely a matter of data availability. Emerging markets are characterized by elevated macroeconomic uncertainty, policy volatility, political risk, and concentrated sectoral exposure. These features generate greater variation in volatility drivers than is typically observed in large developed markets. In such an environment, forward-looking option information and structured news may contain especially valuable signals. By documenting substantial forecasting gains in this setting, we provide evidence that the joint role of implied volatility and news is economically meaningful in markets where volatility is both high and episodic.

More broadly, our findings contribute to understanding how markets aggregate information. Implied volatility reflects risk-neutral expectations embedded in option prices, while news captures evolving narratives and informational flows. The interaction between these two sources helps explain volatility dynamics across horizons.

The remainder of the paper is organized as follows. Section 2 describes the RV framework and the main HAR class models. Section 3 describes the exogenous variables and additional models used in this work. Section 4 presents the estimation and evaluation methods, Section 5 discusses the results, and Section 6 concludes.

2. REALIZED VARIANCE FRAMEWORK AND HAR CLASS MODELS

This section presents an overview of realized variance, with particular attention to its continuous and jump components, realized semivariance, the leverage effect, and their relation to the HAR-class models.

2.1. *Realized variance and HAR*

We model the log-price process $(p_t)_{t\geq 0}$ as a semimartingale with stochastic volatility and jumps:

$$p_t = p_0 + \int_0^t \mu_s \, ds + \int_0^t \sigma_s \, dW_s + J_t, \tag{1}$$

where μ_s is the drift, σ_s is the instantaneous volatility, W_t is a standard Brownian motion, and J_t is a pure-jump process.

This specification is consistent with the absence of arbitrage in frictionless markets and is standard in the realized volatility literature.

The quadratic variation of p_t over $[0, t]$ is

$$[p]_t = \int_0^t \sigma_s^2 \, ds + \sum_{0<s\leq t} (\Delta J_s)^2. \tag{2}$$

Since the empirical analysis is conducted at the daily frequency, the object of interest is the daily increment of quadratic variation:

$$QV_t \equiv [p]_t - [p]_{t-1} = \int_{t-1}^t \sigma_s^2 \, ds + \sum_{t-1<s\leq t} (\Delta J_s)^2. \tag{3}$$

which cannot be directly observed, but can be consistently estimated, under suitable conditions (see Andersen et al. (2001), Barndorff-Nielsen and Shephard (2002a) and Barndorff-Nielsen and Shephard (2002b)), employing the realized variance (RV) defined as:

$$RV_t = \sum_{i=1}^{n_t} r_{t,i}^2. \tag{4}$$

where $r_{t,i} = p_{t,i} - p_{t,i-1}$ denotes the i-th intraday return within day t, and n_t represents the number of such returns.

In the absence of market microstructure noise, and under standard regularity conditions, realized variance converges in probability to the quadratic variation as the intraday sampling frequency increases:

$$RV_t \xrightarrow{p} QV_t \quad \text{as} \quad n_t \to \infty, \tag{5}$$

In practice, however, observed prices are contaminated by market microstructure noise (arising from the bid-ask bounce, price discreteness, and asymmetric information), so that realized variance computed at the highest available frequencies is biased and, in general, no longer consistent for QV_t (Hansen and Lunde, 2006, Zhang et al., 2005). Because such frictions are pervasive, we do not sample at the finest available frequency but at a coarser one that attenuates their effect. This study employed 5-minute transaction prices during trading hours (10:00 to 17:00) to calculate the RV.[1] The samples were gathered from Bloomberg and span from July 25, 2011 to December 29, 2022 (the last trading session of that year) for the most active stocks traded in B3 with implied volatility available, namely the financial institutions Bradesco (BBDC4) and Itaú (ITUB4), the basic materials industries CSN (CSNA3) and Gerdau (GGBR4), and the state-controlled oil company Petrobras (PETR4), and from April 9, 2014 to December 29, 2022 for the BOVA11 Exchange Traded Fund (ETF)[2], for the reasons set out below.

The implied volatility series requires further comment, because its coverage differs markedly across assets. Our index instrument is not an Ibovespa index option. It is BOVA11, an exchange-traded fund replicating the Ibovespa, and the implied volatility we use is extracted from the option chain written on that ETF. Liquidity in Ibovespa index options therefore does not carry

[1]Although the literature proposes other QV_t estimators, outperforming the 5-minute sample RV has proved quite challenging (Liu et al., 2015).
[2]ETF from Blackrock that replicates the Ibovespa index composition.

over to the BOVA11 chain. Options on an index-tracking ETF are materially less liquid in Brazil than options on the most heavily traded single stocks (which is precisely the criterion by which our five stocks were selected), and the Brazilian index option market is itself documented as a low-liquidity market (Astorino et al., 2017). Sparser coverage for BOVA11 is thus the expected pattern rather than an anomaly.

It also matters what a missing value is. The Bloomberg field is a model-based point on a fitted implied volatility surface, at a 30-day tenor and 100% forward moneyness, not the volatility of an individual trade. A missing value means that no valid surface point was published after quote selection, filtering, calibration and maturity interpolation. It therefore signals insufficient quote coverage of the option chain on that date. It does not indicate an absence of trading in the underlying, nor an unusual level of volatility. The missingness mechanism is a property of the depth of the option book rather than of the variable being forecast, so it does not induce selection on the outcome.

The resulting pattern is concentrated in time rather than spread across the sample. For BOVA11, 331 of the 332 unavailable days precede April 2014, with runs of up to 38 consecutive sessions. From April 2014 onward the series has a single missing day in 2,161 sessions. Each of the five individual stocks has at most three missing days over the full eleven and a half years, and four of the five have none. There is no clustering in high-volatility episodes, and no missing value at all after 2017.

We address this in two ways. First, for BOVA11 alone we restrict the sample to begin on April 9, 2014. Deletion by itself is inadequate when roughly half of the early observations are absent, because the aggregated regressors are averages over fixed windows of trading sessions. Deleting rows and averaging over whichever observations remain would place a multi-month average in a variable labelled as a twenty-two-day average. On the unrestricted BOVA11 series, 316 of the 2,477 nominal twenty-two-observation implied variance windows would span more than twenty-two trading sessions, the worst spanning 129, and the nominal one-session lag would reach as far as 39 sessions back. That is a measurement error rather than a missing value, and the restriction removes it. No other asset requires such a restriction. Second, for the few remaining unavailable days we delete rather than impute. Implied variance is a forward-looking quantity extracted from option prices, and imputing it on days when the option market did not support its construction would manufacture information that the market did not provide, precisely on the days when the chain was least informative. Four days across all six assets are affected.

All lags, leads and rolling averages are aligned to the trading calendar and are computed strictly. A window that spans a session without data is discarded rather than silently averaged over fewer days. This is the conservative choice (the alternative would retain the observation while storing, say, a nineteen-day mean in a variable labelled as a twenty-two-day average), and Table I reports its cost. Panel A of that table reports the sessions on which a variable is unavailable, Panel B the observations that cannot be used because the models employ lags of up to twenty-two sessions, and Panel C the resulting sample. Because every regressor is a strictly lagged window, a session on which a variable is unavailable is itself retained (its own regressors are built from earlier sessions), while the twenty-two sessions that follow it cannot be used. Panel B therefore accounts for the whole of the difference between the trading sessions in the sample period and the resulting sample.

The exclusions are therefore concentrated in the early part of the estimation sample, and the evaluation period is unaffected. Every regressor matrix and every forecast target is complete for all six assets over the entire out-of-sample window, so no forecast, loss, loss ratio, or model confidence set result reported below rests on a deleted or imputed observation. What the missing values do affect is the length of BOVA11's estimation sample, which is roughly 30% shorter than that of the individual stocks.

Table II displays daily summary statistics for each asset's RV. All assets show high skewness and kurtosis.

Figure 1 illustrates the RV time series for each asset. The COVID-19 crisis marks the highest volatility for all analyzed assets, except PETR4, where the peak of volatility occurred dur-

TABLE I: Missing observations and observations lost to the lag structure

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Trading sessions with news and implied volatility data	2,159	2,831	2,830	2,831	2,831	2,831
Panel A: Sessions with a missing variable						
Realized variance	0	0	0	0	0	0
Implied variance	1	0	3	0	0	0
News topic proportions	1	1	1	1	1	1
Total	*2*	*1*	*4*	*1*	*1*	*1*
Panel B: Observations lost to the lag structure and to unavailable data						
Warm-up: first 22 sessions	22	22	22	22	22	22
Windows spanning a missing implied variance	22	0	25	0	0	0
Windows spanning the missing news session	22	22	22	22	22	22
Total	*66*	*44*	*69*	*44*	*44*	*44*
Panel C: Resulting sample						
Observations	2,093	2,787	2,761	2,787	2,787	2,787
Estimation period	1,618	2,312	2,286	2,312	2,312	2,312
Out-of-sample period	475	475	475	475	475	475

Note: This table reports, for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4, the trading sessions excluded from the analysis and the reason for each exclusion. The sample period runs from July 25, 2011 (the first day of the news corpus) to December 29, 2022, the last trading session of that year, except for BOVA11, whose sample begins on April 9, 2014 for the reasons discussed in the text. The first row counts the trading sessions for which the news and implied volatility source provides a record. It excludes November 20, 2020 for all assets, and additionally January 23, 2017 for BOVA11 and January 3, 2014 for CSNA3, on which no record exists. Panel A reports the sessions on which a variable is unavailable. Realized variance is available for every session in the sample period, since the intraday price data cover the entire trading calendar. Implied variance is missing on January 23, 2017 for BOVA11 and on January 3, 6, and 8, 2014 for CSNA3. The news topic proportions are missing on November 20, 2020 for all assets. Panel B reports the observations that cannot be used. Since all regressors are strictly lagged, a session on which a variable is unavailable does not itself invalidate the twenty-two sessions before it, but rather the twenty-two sessions that follow it. Panel C reports the resulting sample for all assets .

TABLE II

SUMMARY STATISTICS OF DAILY RV

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Observations	2093	2787	2761	2787	2787	2787
Mean	1.55	3.46	9.03	6.64	3.15	5.46
Std. Dev.	2.95	4.71	8.61	7.96	4.22	11.70
Median	1.05	2.58	6.68	4.73	2.38	3.22
1st Quartile	0.70	1.79	4.61	3.25	1.66	1.99
3rd Quartile	1.68	3.83	10.23	7.26	3.44	5.67
Skewness	13.07	12.25	4.72	8.07	12.56	21.96
Kurtosis	223.91	207.16	35.46	99.84	217.12	715.75

Note: This table shows the summary statistics of the RV for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 in the whole sample. Except the number of observations, skewness, and kurtosis, all values are multiplied by 10^4. Kurtosis is reported in excess of 3.

ing Brazil's truck driver strike in 2018.[3] A notable surge in volatility is also evident in 2017, following the political controversy known as "Joesley Day".[4]

Corsi (2009) developed a constrained AR(22) model called heterogeneous autoregressive (HAR) that effectively captures the long memory exhibited in the autocorrelation function displayed in Figure 2 and the fat tails seen in the high kurtosis of the variables in Table II.

Following Corsi (2009), we model realized volatility using heterogeneous components that capture persistence at different horizons. The h-step-ahead realized volatility target is defined as the average of future realized variances:

$$RV_{t+h|t} = \frac{1}{h}\sum_{i=1}^{h} RV_{t+i}. \tag{6}$$

To capture volatility dynamics at different frequencies, we construct backward-looking averages of realized volatility. For any variable Z_t, define the k-period lagged average as:

$$Z_t^{(k)} = \frac{1}{k}\sum_{i=0}^{k-1} Z_{t-i}, \tag{7}$$

[3]In May 2018, Brazilian truck drivers went on strike over fuel prices. PETR4, a state-controlled oil company, accounted for 76% of diesel sales in Brazil in 2017, according to the authors' calculations (Petrobras, 2018, Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, 2018).

[4]On May 17, 2017, the release of an audio recording between then-President Michel Temer and businessman Joesley Batista was poorly received by financial markets.

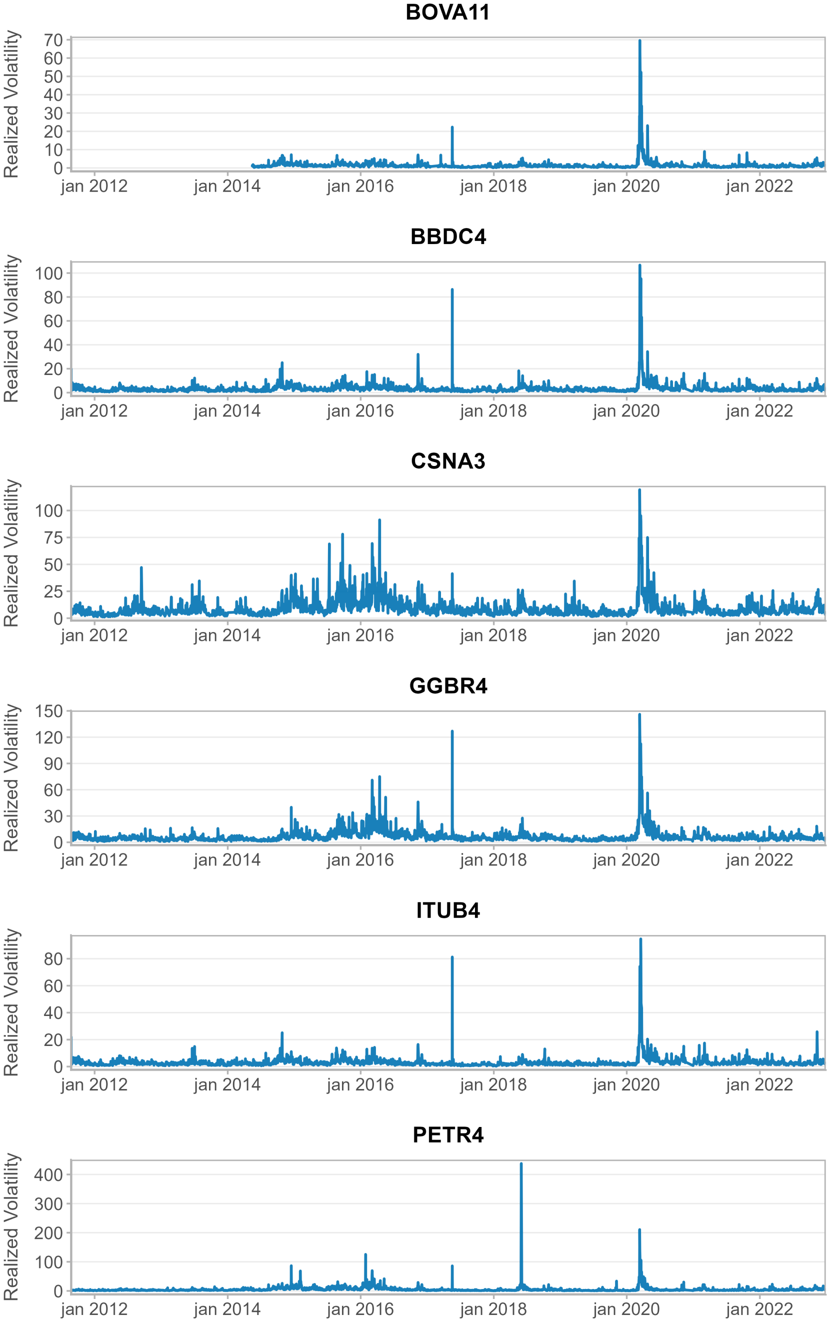

FIGURE 1.—Realized volatility time series plot for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4. All values are multiplied by 10^4 and each asset features a distinct y-axis scale.

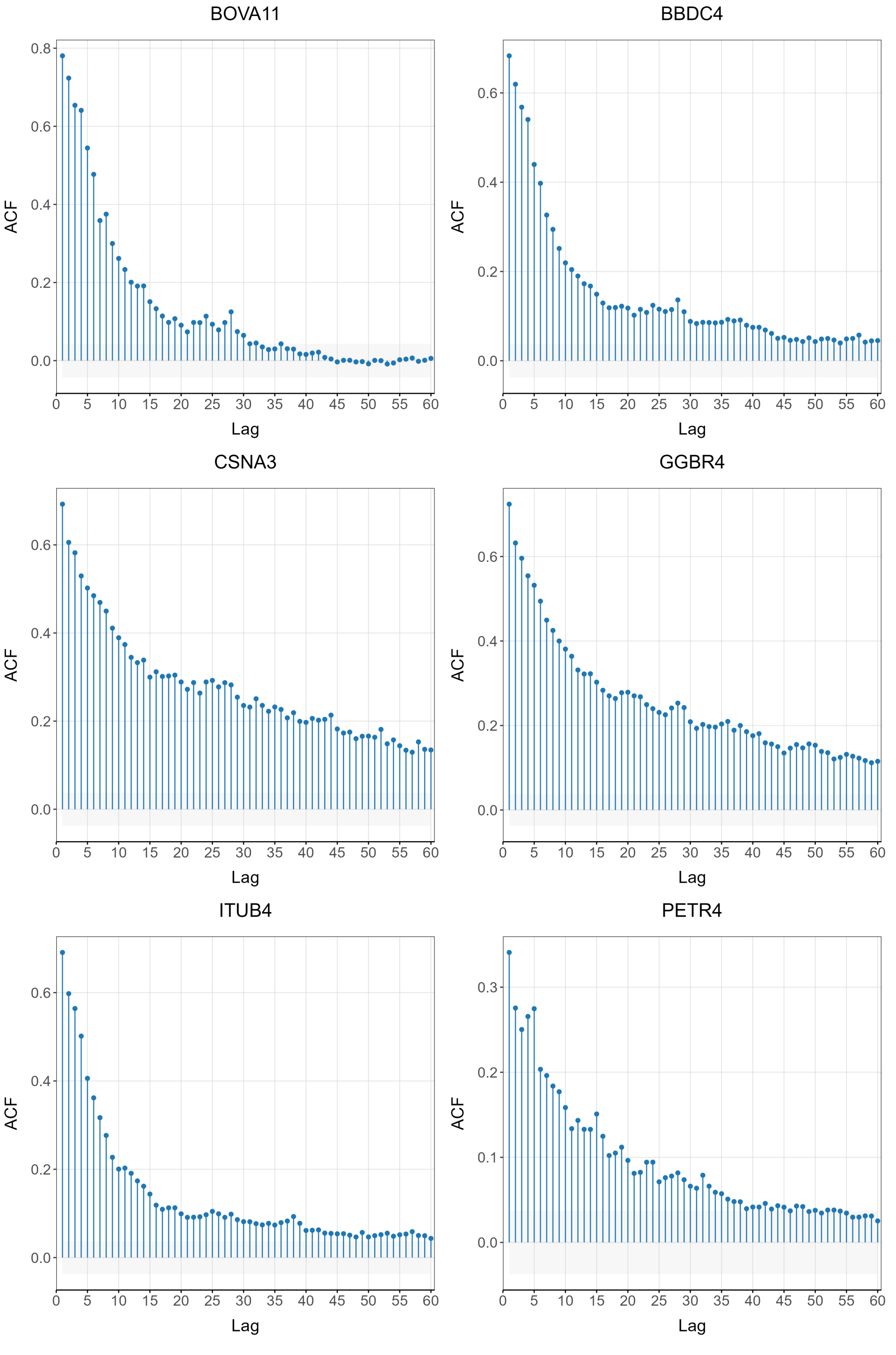

FIGURE 2.—Autocorrelation function of realized variance plot of the daily realized variance for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 until lag 60. BOVA11 and PETR4 feature distinct y-axis scales.

which ensures that all regressors are constructed exclusively from information available at time t.

In particular, we define the daily, weekly, and monthly realized volatility components respectively as:

$$RV_t^{(d)} = RV_t, \quad RV_t^{(w)} = RV_t^{(5)}, \quad RV_t^{(m)} = RV_t^{(22)}. \tag{8}$$

The HAR model is then specified as:

$$RV_{t+h|t} = f(RV_t^{(d)}, RV_t^{(w)}, RV_t^{(m)}) \tag{9}$$

Corsi (2009) identified that the HAR model yields lower forecast MSE and MAE than the AR and ARFIMA models for the USD/CHF exchange rate, the S&P500 index futures, and the 30-year US Treasury Bond futures.

The convention of defining lagged averages as in Equation (7) is applied not only to realized volatility, but also to implied variance, and news-based variables throughout the paper.

2.2. *Jumps and HARCJ*

Barndorff-Nielsen and Shephard (2004) proposed the bipower variation estimator

$$BPV_t = \frac{\pi}{2} \sum_{i=2}^{n_t} |r_{t,i-1}|\, |r_{t,i}|, \tag{10}$$

which, in the absence of microstructure noise and under standard regularity conditions, converges in probability to the continuous component of daily quadratic variation:

$$BPV_t \xrightarrow{p} \int_{t-1}^{t} \sigma_s^2 ds. \tag{11}$$

A natural empirical proxy for daily jump variation is then given by

$$JV_t = \max\{RV_t - BPV_t, 0\}. \tag{12}$$

Using these components, Andersen et al. (2007) developed the HARCJ specification written as

$$RV_{t+h|t} = f\left(BPV_t, BPV_t^{(5)}, BPV_t^{(22)}, JV_t, JV_t^{(5)}, JV_t^{(22)}\right). \tag{13}$$

in which $BPV_t^{(k)}$ and $JV_t^{(k)}$ are calculated analogously to equation (7).

Andersen et al. (2007) show improved R^2's for the HARCJ across exchange rates, bonds, and the S&P 500 index relative to the HAR benchmark.

Tables III and IV present the summary statistics for BPV_t and JV_t, for each asset. BPV_t and JV_t, like RV_t, are highly skewed and leptokurtic.

TABLE III

SUMMARY STATISTICS OF THE DAILY BPV

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Observations	2093	2787	2761	2787	2787	2787
Mean	1.41	3.21	8.07	5.97	2.90	4.90
Std. Dev.	2.67	4.05	7.65	6.62	3.52	6.82
Median	0.97	2.39	6.02	4.31	2.21	3.01
1st Quartile	0.63	1.64	4.05	2.90	1.51	1.83
3rd Quartile	1.52	3.58	9.22	6.69	3.23	5.36
Skewness	13.16	10.97	4.52	6.78	11.02	6.82
Kurtosis	220.30	176.17	32.09	73.95	181.84	72.79

Note: This table shows the summary statistics of the BPV for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 in the whole sample. Except the number of observations, skewness, and kurtosis, all values are multiplied by 10^4. Kurtosis is reported in excess of 3.

TABLE IV

SUMMARY STATISTICS OF THE DAILY JUMP VARIATION

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Observations	1518	1941	2158	2131	1995	1967
Mean	0.24	0.46	1.40	1.00	0.43	0.94
Std. Dev.	0.68	1.43	2.40	3.33	1.64	9.97
Median	0.11	0.25	0.80	0.53	0.22	0.31
1st Quartile	0.05	0.11	0.39	0.25	0.10	0.14
3rd Quartile	0.22	0.47	1.56	1.07	0.43	0.65
Skewness	12.02	20.48	10.74	22.97	26.08	37.86
Kurtosis	181.01	521.52	206.38	623.92	823.90	1537.13

Note: This table shows the summary statistics of the daily jump variation JV_t of Equation (12), which proxies the jump contribution to quadratic variation, for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 in the whole sample. Statistics are computed over days with a strictly positive jump-variation estimate ($JV_t > 0$), which accounts for the smaller number of observations relative to Table III. Except the number of observations, skewness, and kurtosis, all values are multiplied by 10^4. Kurtosis is reported in excess of 3.

2.3. *Leverage effect and LHARCJ*

The negative correlation between past returns and current volatility, a phenomenon known as the leverage effect, is well documented in the literature[5] and was incorporated into the HARCJ model (Equation(13)) by Corsi and Reno (2012) leading to the LHARCJ:

$$RV_{t+h|t} = f(BPV_t, BPV_t^{(5)}, BPV_t^{(22)}, JV_t, JV_t^{(5)}, JV_t^{(22)}, L_t, L_t^{(5)}, L_t^{(22)}), \qquad (14)$$

where $L_t = \min[R_t, 0]$ and R_t is the return of the asset on day t.

Table V displays the summary statistics of daily leverage effects for each asset.

Using a DM test with HMSE loss function, Corsi and Reno (2012) showed that LHARCJ offers superior predictive performance over HARCJ and HAR models for forecast horizons ranging from 1 day to 1 month for the S&P 500 futures.

2.4. *Realized semivariance and SHAR*

Following Barndorff-Nielsen et al. (2010), realized variance can be decomposed into positive and negative semivariances, defined as:

[5] E.g. Glosten et al. (1993), Bollerslev et al. (2009a).

TABLE V

SUMMARY STATISTICS OF THE DAILY LEVERAGE EFFECT

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Observations	1006	1369	1394	1398	1359	1331
Mean	-1.18	-1.55	-2.63	-2.03	-1.47	-2.20
Std. Dev.	1.28	1.56	2.38	1.89	1.32	2.42
Median	-0.90	-1.14	-2.06	-1.57	-1.14	-1.59
1st Quartile	-1.58	-2.10	-3.58	-2.70	-2.03	-2.87
3rd Quartile	-0.38	-0.56	-1.03	-0.75	-0.53	-0.75
Skewness	-4.50	-3.62	-2.99	-2.77	-2.23	-4.51
Kurtosis	36.48	25.39	19.13	14.47	9.22	39.43

Note: This table shows the summary statistics of the leverage effect for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 in the whole sample. Except the number of observations, skewness, and kurtosis, all values are multiplied by 10^2. Kurtosis is reported in excess of 3.

$$
\begin{aligned}
RS_t^+ &= \textstyle\sum_{i=1}^{n_t} r_{t,i}^2 \mathbf{1}\{r_{t,i} > 0\}, \\
RS_t^- &= \textstyle\sum_{i=1}^{n_t} r_{t,i}^2 \mathbf{1}\{r_{t,i} < 0\}.
\end{aligned} \tag{15}
$$

In the absence of microstructure noise and under standard regularity conditions, realized semivariances converge in probability to the corresponding components of quadratic variation associated with positive and negative returns, respectively (Barndorff-Nielsen et al., 2010).

These measures capture the contribution of positive and negative intraday returns to total realized variance. Building on this decomposition, Patton and Sheppard (2015) propose HAR-type models that decompose recent realized variance into its signed components and find that the predictive content lies predominantly in the negative semivariance.

We adopt their preferred realized semivariance specification, which replaces the daily realized-variance lag with the negative semivariance while retaining the weekly and monthly averages of realized variance:

$$
RV_{t+h|t} = f\left(RS_t^-, RV_t^{(5)}, RV_t^{(22)}\right). \tag{16}
$$

This specification requires no decomposition of jump variation into signed components and has the same number of regressors as the HAR model.

Using DM tests with the QLIKE loss function, Patton and Sheppard (2015) find that this restricted specification outperforms the standard HAR for the S&P 500 ETF at all horizons (1, 5, 22, and 66 days) and performs better for individual stocks than the less parsimonious two-semivariance specification.

Table VI presents the summary statistics of the daily negative semivariance for each asset. As with the previous realized measures, the distributions are right-skewed and leptokurtic.

TABLE VI

SUMMARY STATISTICS OF THE DAILY NEGATIVE SEMIVARIANCE

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Observations	2093	2787	2761	2787	2787	2787
Mean	0.80	1.74	4.46	3.31	1.58	2.81
Std. Dev.	1.61	2.50	4.45	4.04	2.29	8.98
Median	0.53	1.27	3.29	2.34	1.17	1.56
1st Quartile	0.32	0.86	2.16	1.53	0.78	0.95
3rd Quartile	0.85	1.96	5.20	3.69	1.77	2.84
Skewness	14.89	13.55	5.24	7.84	14.60	37.79
Kurtosis	308.22	262.84	45.67	93.26	297.97	1734.43

Note: This table shows the summary statistics of the daily negative realized semivariance (RS_t^-) for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 in the whole sample. Except the number of observations, skewness, and kurtosis, all values are multiplied by 10^4. Kurtosis is reported in excess of 3.

3. EXOGENOUS REGRESSORS AND NEWS IV

This section presents the exogenous regressors and NEWS IV model.

3.1. *News sentiment*

3.1.1. *Data*

The textual dataset consists of the title combined with the initial paragraph from 678,616 articles, comprising news reports and opinion pieces, published on the *Valor Econômico* website between July 25, 2011, and December 31, 2022 (Cardoso and Nakane, 2024).

Our choice to restrict the extraction to the title and first paragraph relies on two key considerations:

1. Journalistic Structure: Modern news writing strictly adheres to the *inverted pyramid* structure (Scanlan, 2000, Pöttker, 2003), where the core narrative, essential facts, and key eco-

nomic triggers are concentrated at the very top (the headline and lead paragraph) (Hamborg et al., 2018). Subsequent paragraphs typically contain secondary details, background context, or elaboration. Filtering for the title and lead paragraph maximizes the signal-to-noise ratio, isolating high-density informational content while excluding peripheral narrative fluff.

2. Data Accessibility and Operational Feasibility: Unlike full-text articles, which are frequently behind strict paywalls or require proprietary subscriptions, the title and lead paragraph of articles on *Valor Econômico* are freely accessible on the web.

During text preprocessing, standard natural language processing techniques were applied to clean the raw corpus before topic modeling. A 2-gram model was implemented to identify meaningful multi-word key phrases (e.g., "Banco Central" or "taxa SELIC"). Standard Portuguese stop words were removed, as were low-frequency terms occurring in less than 0.1% of all articles, eliminating sparse lexical noise (Cardoso and Nakane, 2024). Combining the headline and lead paragraph yields concise text entries averaging 57 words per article, ensuring a coherent, focused document corpus for topic extraction.

3.1.2. *LDA*

Introduced by Blei et al. (2003), Latent Dirichlet Allocation (LDA) is an unsupervised algorithm designed to categorize words from texts into topics of related word.[6]

Figure 3 illustrates the LDA generative model. The hyperparameter α defines the Dirichlet prior for the proportion of topics in each document (θ_d), the hyperparameter η defines the Dirichlet prior for the proportion of words in each topic (β_k), $z_{d,v}$ is the assignment of topics for the term n^{th} in document d, and $w_{d,v}$ is the observed word.

[6]This approach does not bundle words with multiple meanings in specific contexts, e.g., “virus” could refer to computers or biological concepts.

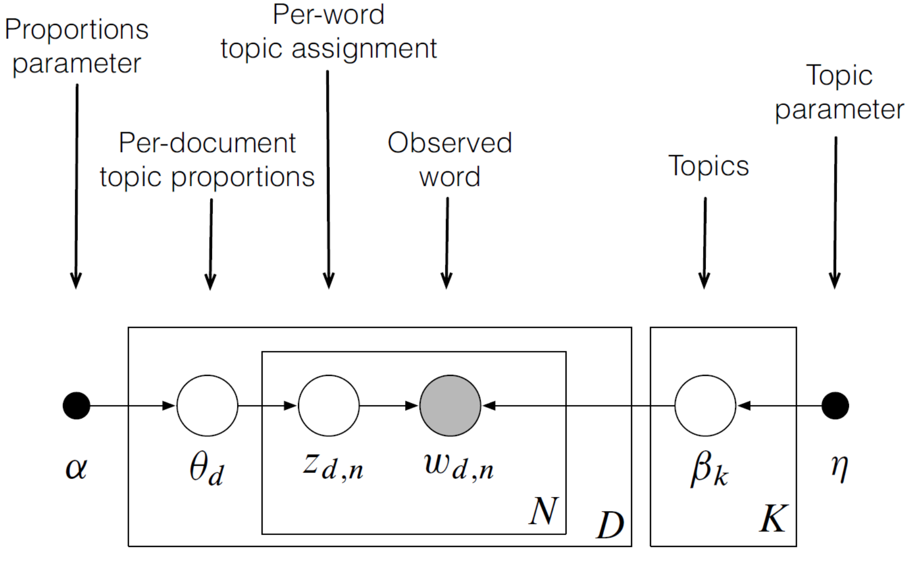

FIGURE 3.—LDA generative procedure. Sourced from Blei (2012)

Following Cardoso and Nakane (2024), we set the number of topics to $K = 40$ (chosen to minimize the perplexity score), with Dirichlet hyperparameter priors $\alpha = 1.25$ and $\eta = 0.055$. The LDA procedure estimates each document's topic proportions (θ_d) yielding the X_t (40 x 1) vector of daily topic proportion.

To prevent look-ahead bias in the evaluation period, the topic-word distributions β_k ($k = 1, \ldots, 40$) are estimated strictly once using the pre-evaluation corpus ending on December 31, 2020. Topic distributions for articles published during the out-of-sample period (2021-2022) are inferred conditional on these fixed pre-evaluation topic-word distributions and evaluation-period documents do not update the LDA model. Topic proportions for articles in the 2011-2020 estimation sample are in-sample representations under this common pre-evaluation LDA fit, rather than outputs from a recursively re-estimated topic model. Thus, later documents within the training partition may affect the representation of earlier training documents, but no evaluation-period document affects the fitted text representation.

For each trading day t, we aggregate the estimated topic proportions across all articles published on that day, producing a 40-dimensional vector $X_t = [x_{1,t}, x_{2,t}, \ldots, x_{40,t}]'$, where $\sum_{k=1}^{40} x_{k,t} = 1$. The resulting topics span macroeconomic policy, corporate announcements, sector-specific developments, global market conditions, and domestic political events (see A for full topic labels and top keywords).

Our approach to incorporating news flow into volatility models differs conceptually and methodologically from other available frameworks. The paper closest to ours is Fernandes and Pereira (2025) and comparing our approach to theirs highlights key trade-offs in data selection and textual representation.

Fernandes and Pereira (2025) analyze full-text articles to capture granular details. Full texts often introduce substantial noise, repetition, and boilerplate sentences. By leveraging the journalistic inverted pyramid, our model focuses exclusively on headlines and leads, capturing the primary economic shock while keeping data processing efficient. The main trade-off is that highly specific firm-level details buried deep in long articles might be omitted, though Boudoukh et al. (2012) empirical study for S&P 500 firms shows that extracting text signals at the headline/lead level significantly sharpens the signal-to-noise ratio for forecasting aggregate market shocks and latent narrative shifts.

Fernandes and Pereira (2025) construct uncertainty indices by counting specific keywords related to economic policy or specific asset names using a Bag-of-Words dictionary. This targeted approach is intuitive, but relies on predefined lexicons that may fail to capture emerging, unanticipated narratives (e.g., unexpected political scandals, sector-specific regulatory shifts, or global supply chain bottlenecks). In contrast, LDA operates in an unsupervised manner across (in our case) 40 continuous thematic dimensions. It automatically identifies evolving macroeconomic, fiscal, sectoral, and political narratives across the market without human bias or fixed dictionary constraints.

Because LDA generates a high-dimensional vector of (in our case) 40 continuous topic probabilities, it naturally pairs with non-linear machine learning architectures like Random Forests. This allows NEWS IV to capture non-linear interactions between option-implied expectations (IVAR) and evolving textual narratives, a feature that standard linear regressions built on single dictionary counts cannot easily accommodate.

3.2. *Implied volatility*

Implied volatility is a forward-looking measure extracted from option prices. We denote by IV_t the 30-day implied volatility of the asset and by

$$\mathrm{IVAR}_t = IV_t^2. \tag{17}$$

the corresponding implied variance.

IVAR differs from the corresponding expectation under the physical measure by the variance risk premium (Bollerslev et al., 2009b, Bekaert and Hoerova, 2014), which makes it a biased yet informative market-based predictor of future volatility.

In this paper, we use the 30-day implied volatility series provided by Bloomberg, which is constructed using a model-based methodology and interpolated across maturities to obtain a constant-horizon measure.[7] Since our forecasting object is realized variance, we work throughout with the implied variance IVAR_t defined above rather than with IV_t itself.

Because Bloomberg quotes implied volatility on an annualized basis, we report the implied variance on a daily basis by dividing by 252 trading days, so that it is directly comparable with the realized measures reported above. Table VII reports the summary statistics of the daily implied variance for each asset. The distributions exhibit noticeable skewness and excess kurtosis.

TABLE VII
SUMMARY STATISTICS OF THE DAILY IMPLIED VARIANCE

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Observations	2093	2787	2761	2787	2787	2787
Mean	2.70	4.42	11.55	7.79	4.06	8.51
Std. Dev.	2.60	3.64	7.08	6.11	3.60	8.27
Median	2.16	3.59	9.86	6.23	3.28	6.00
1st Quartile	1.69	2.84	7.66	4.62	2.62	4.30
3rd Quartile	2.86	4.77	12.76	8.49	4.32	9.80
Skewness	8.69	7.12	4.02	5.21	10.90	4.68
Kurtosis	105.35	76.44	29.16	46.24	224.02	33.22

Note: This table shows the summary statistics of the IVAR (implied variance, in decimal units) for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 in the whole sample. Except the number of observations, skewness, and kurtosis, all values are multiplied by 10^4. Kurtosis is reported in excess of 3.

3.3. *NEWS IV model*

Similarly to Corsi (2009) in the HAR model (Equation (9)), we can calculate the averages of X_t and IVAR_t over periods of 1, 5 and 22 days to establish the NEWS IV model outlined in Equation (18):

$$RV_{t+h|t} = f(\mathrm{IVAR}_t, \mathrm{IVAR}_t^{(5)}, \mathrm{IVAR}_t^{(22)}, X_t, X_t^{(5)}, X_t^{(22)}, RV_t, RV_t^{(5)}, RV_t^{(22)}), \tag{18}$$

where the terms $\mathrm{IVAR}_t^{(k)}$, $X_t^{(k)}$ and $RV_t^{(k)}$ are calculated using Equation (7). The realized volatility components of Equation (9) are retained as candidate predictors, so that the model

[7]For further details on the methodology, see Quantitative Analytics Bloomberg L.P. (2019).

nests the HAR information set and the contribution of the implied variance and news variables is measured beyond it. The two groups are not treated alike by the estimator. As described in sub-subsection 4.1.1, only the implied variance averages are evaluated at every node, while the realized volatility components compete for selection with the news topics.

We estimate NEWS IV exclusively using a random forest approach. The primary motivation is to capture nonlinear effects and interactions between implied variance and news topics, which a linear specification could accommodate only through a prohibitive number of pre-specified interaction terms. A complementary benefit is that the random selection of candidate variables at each split acts as implicit regularization (Mentch and Zhou, 2020), which keeps the estimator well-behaved in the high-dimensional setting of Equation (18), where OLS and WLS estimators perform poorly.

4. FORECAST AND EVALUATION

This section details the estimation techniques as well as the loss functions and the model confidence set (MCS) procedure used for assessing pseudo-out-of-sample forecast performances.

4.1. *Forecast*

The models are estimated using OLS (Corsi and Reno, 2012), WLS with inverse OLS fitted values as weights to address the residual heteroskedasticity associated with RV level (Patton and Sheppard, 2015) and by random forest (RF) described in detail in sub-subsection 4.1.1.

Forecasts are generated separately for $h \in \{1, 5, 22\}$ using fixed-length rolling windows Hansen et al. (2011).

As the topics underwent training from July 25, 2011 to December 31, 2020 (Cardoso and Nakane, 2024) and we need 22 trading sessions after the end of the topic-training corpus on December 31, 2020 to avoid the look-ahead bias, our out-of-sample period starts at February 4, 2021. Table VIII displays the number of training data points for each asset.

4.1.1. *Random forest*

Introduced by Breiman (2001), the random forest (RF) averages the predictions of multiple decision trees with each tree grown from an individual bootstrap sample. At each node of the

TABLE VIII

TRAINING SAMPLE SIZE IN TRADING DAYS

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
Training sample size	1618	2312	2286	2312	2312	2312

Note: This table shows the training sample size in trading days for BOVA11, BBDC4, CSNA3, GGBR4, ITUB4, and PETR4.

decision tree, a subset of features is randomly selected without replacement from the full set of features [8]. The feature with the split threshold that achieves the greatest reduction in MSE between the parent and child nodes is then selected for further splitting.

For B decision trees built using different bootstraps from the training sample and the information set Z_t, the RF prediction is expressed as in Equation (Equation 19):

$$\hat{f}(Z_t) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(Z_t, \theta^b), \tag{19}$$

with θ^b denoting the features chosen at random at each node and $\hat{f}^b(Z_t, \theta^b)$ representing the prediction of the bth bootstrap-trained decision tree.

Evidence from the high-frequency volatility research suggests that implied volatility (IV) effectively predicts volatility (see Bekaert and Hoerova (2014), Kambouroudis et al. (2021), Christensen et al. (2023)). Moreover, when only a few predictors carry most of the signal while many carry little, sampling a random subset of candidates at each node can leave the informative variables unavailable at a large share of splits (Hastie et al., 2009), so that a larger number of candidate variables is preferable in such settings (Probst et al., 2019). In our setting this is quantitatively important because with 126 candidate predictors and the default choice of $\lfloor\sqrt{126}\rfloor = 11$ candidates per node, none of the three implied-variance regressors would be available at approximately 75% of the splits. Therefore, the variables IVAR_{t-1}, $\text{IVAR}^{(5)}_{t-1}$, and $\text{IVAR}^{(22)}_{t-1}$ present in NEWS IV are evaluated in every node of the forest along with a random subset of averaged news topics.

[8]This per-node (per-split) sampling of candidate splitting variables follows the canonical algorithm of Breiman (2001) and the mtry parameter of the ranger implementation (Wright and Ziegler, 2017). It differs from the random subspace method of Ho (1998), in which a single subset of features is drawn once per tree.

We set the number of trees at 1000, above the package default of 500, for two reasons. First, under the mean squared error, which is the criterion by which the trees are grown, the expected loss of the ensemble decreases monotonically with the number of trees (Probst and Boulesteix, 2018). Second, the variable importance rankings reported in Section 5 stabilise more slowly than the forecasts themselves (Probst et al., 2019).

The other hyperparameters are the default values of the `ranger` package[9] (Wright and Ziegler, 2017) and Table IX lists all of them.

A further implementation detail worth making explicit is that the variables forced into every split are considered *in addition to* the $\lfloor\sqrt{p}\rfloor$ predictors drawn at random. As a result, each node of the NEWS IV forest evaluates 14 candidate predictors rather than 11.

TABLE IX

HYPERPARAMETERS OF RANDOM FOREST

Hyperparameter	**Value/Setting**	**Source**
Candidate variables per split	$\lfloor\sqrt{p}\rfloor$	`ranger` default
Always-split variables	IVAR_{t-1}, $\text{IVAR}^{(5)}_{t-1}$, $\text{IVAR}^{(22)}_{t-1}$	Explicit choice
Minimum node size	5	`ranger` default
Maximal tree depth	Unlimited	`ranger` default
Split rule	Variance reduction (MSE)	`ranger` default
Bootstrap	With replacement, sample fraction 1	`ranger` default
Number of trees	1000	Explicit choice (default 500)

Note: This table lists the hyperparameters for random forest training. p denotes the number of regressors of the specification being estimated. The always-split variables are candidates in addition to the $\lfloor\sqrt{p}\rfloor$ variables drawn at random, so each node of the NEWS IV forest evaluates 14 candidates. Variable importance is computed from the impurity reduction and does not affect the fitted forecasts.

4.2. *Evaluation*

4.2.1. *Model confidence set*

Given a set of all competing models ($\mathcal{M}^0$), the MCS procedures construct a subset of top-performing models ($\mathcal{M}^*$) that fail to reject the null hypothesis of equal predictive ability ($H_0 : E(d_{ij,t}) = 0$) where $d_{ij,t} \equiv L_{i,t} - L_{j,t}$ represents the difference between the loss functions of the competing models i and j at time t [10].

[9] Further information can be found in the Ranger documentation.

[10] We note that membership in the MCS does not correspond to a ranking of point loss estimates. The elimination algorithm of Hansen et al. (2011) compares each model's loss differential against the remaining set to its own

Following Bodilsen and Lunde (2025), this article employs the MCS procedure at the confidence level 90%, using a block bootstrap with block size matched to a h-step forecast over 10,000 iterations using the range statistic[11] and the robust[12] loss functions for volatility forecast QLIKE

$$\text{QLIKE} = \log\left(\widehat{RV}_{t+h|t}\right) + \frac{RV_{t+h|t}}{\widehat{RV}_{t+h|t}} \tag{20}$$

and the mean of squared errors (MSE), where the squared error (SE) is calculated as

$$\text{SE} = (RV_{t+h|t} - \widehat{RV}_{t+h|t})^2, \tag{21}$$

as described by Patton (2011)[13] to evaluate the volatility predictions of NEWS IV and the HAR type models described in Section 2.

QLIKE is asymmetric and penalizes underpredictions more heavily, whereas SE is symmetric and more sensitive to tail observations. Although both are robust loss functions, QLIKE exhibits higher power than SE (Patton, 2011).

4.2.2. *Variable importance*

The importance of each variable is assessed using the mean decrease in impurity (MDI), which quantifies its contribution to the decrease in MSE during forest training (Hastie et al., 2009). For clarity, MDIs associated with news topics were combined into a single "Topics" category, and the importance values were adjusted to ensure that the total MDIs sum up to 100.

5. OUT-OF-SAMPLE FORECASTS AND VARIABLE IMPORTANCE

All forecast comparisons are based on loss ratios relative to the HAR model estimated by OLS, so that values below one indicate better predictive performance relative to the HAR(OLS) benchmark.

bootstrap-estimated sampling variance, so a model with a nominally larger ratio can remain in the set if its performance is not statistically distinguishable from the best model.

[11]See Hansen et al. (2011) for details.

[12]Robust in the sense that the forecasting ranking obtained using a volatility proxy is the same yield using the true volatility.

[13]The MCS results for the other biased loss functions commonly used to ranking volatility forecasts (MSE-LOG, MSE-SD, HMSE, MAE, MAE-LOG, MAE-SD, MAPE) described in Patton (2011) are displayed in B.

Figure 4 shows the NEWS IV forecast for one day ahead compared to the observed RV. The gray lines depict forecasts using HAR-type variables. Due to the inclusion of the pandemic period in the NEWS IV training sample, forecasting overshooting occasionally occurs.

Tables X and XI display MCS outcomes for QLIKE and MSE in the daily forecast. The models within the best set ($\mathcal{M}^*$) at the 90% confidence level are highlighted. The results indicate that the predictive performance of the NEWS IV model, as measured by QLIKE and MSE loss ratios relative to the HAR(OLS) benchmark, is comparable to or better than that of HAR-type models for individual stocks, highlighting that implied variance and news topics contain incremental predictive information beyond HAR-type dynamics. Consistent with Patton and Sheppard (2015) and Clements and Preve (2021), the models using WLS outperform the OLS models. This result is consistent with the presence of heteroskedasticity in the residuals that is related to the level of realized volatility, as discussed in Andersen et al. (2003) (Patton and Sheppard, 2015). The inferior performance of RF models with HAR-type variables suggests that nonlinearities in HAR-type variables do not materially improve predictive performance at the daily horizon.

TABLE X
QLIKE RESULTS FOR H=1

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.000
HARCJ (OLS)	1.017	**1.000**	**1.000**	**1.000**	**1.001**	0.996
LHARCJ (OLS)	1.034	1.005	**0.999**	1.001	1.002	1.014
SHAR (OLS)	1.000	**1.001**	**0.999**	**1.000**	1.003	1.004
HAR (WLS)	**0.998**	**0.999**	**1.000**	**1.000**	**1.000**	**0.993**
HARCJ (WLS)	**0.998**	**0.999**	**1.000**	**1.000**	**0.999**	**0.994**
LHARCJ (WLS)	**0.991**	**0.998**	**0.999**	**0.999**	**0.998**	**0.991**
SHAR (WLS)	**0.997**	**1.000**	**0.999**	**1.000**	1.003	0.998
NEWS IV (RF)	1.015	**0.999**	**0.999**	**1.001**	**0.999**	**0.994**
HAR (RF)	1.011	1.006	1.003	1.003	1.004	**0.995**
HARCJ (RF)	**1.002**	1.004	1.003	**1.001**	1.004	**0.996**
LHARCJ (RF)	**0.994**	**1.001**	1.002	**1.000**	**1.002**	**0.990**
SHAR (RF)	1.012	1.005	1.005	1.004	1.008	**0.995**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

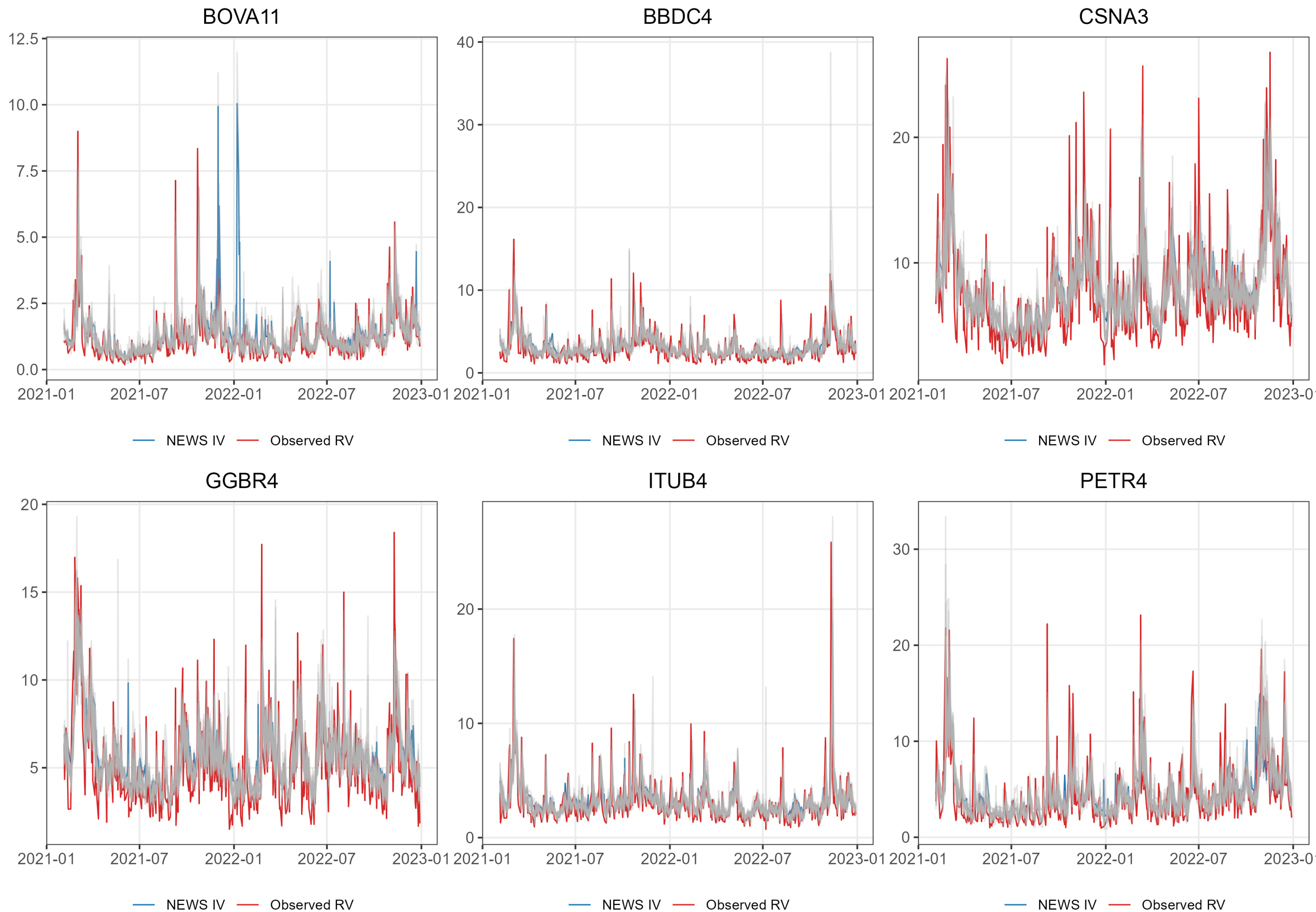

Figure 4.: Comparison between the observed realized volatility and the forecasts in 1 day ahead horizon. The gray lines reference the models estimated with the HAR-type variables.

TABLE XI

MSE RESULTS FOR H=1

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
HARCJ (OLS)	**1.091**	**1.014**	**0.999**	**1.008**	**0.995**	1.032
LHARCJ (OLS)	1.130	1.117	**0.995**	**1.045**	**0.967**	1.259
SHAR (OLS)	**0.960**	**1.001**	**0.968**	**1.016**	**1.077**	1.051
HAR (WLS)	**0.998**	**0.985**	**0.996**	**0.996**	**0.993**	**0.959**
HARCJ (WLS)	**0.984**	**0.992**	**0.994**	**1.003**	**0.949**	**1.008**
LHARCJ (WLS)	**0.919**	**0.930**	**0.964**	**0.968**	**0.912**	**0.910**
SHAR (WLS)	**0.951**	**0.992**	**0.965**	**0.998**	**1.069**	**1.002**
NEWS IV (RF)	**1.801**	**0.971**	**0.951**	**1.058**	**0.914**	**0.936**
HAR (RF)	1.084	1.181	1.144	**1.111**	1.052	**1.064**
HARCJ (RF)	**1.011**	1.089	1.072	**1.051**	**1.080**	**1.039**
LHARCJ (RF)	**0.966**	**1.768**	1.062	**1.014**	**1.021**	**0.949**
SHAR (RF)	1.082	1.115	1.167	**1.247**	**1.408**	**1.007**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

Figure 5 illustrates the variable importance of each asset in the daily forecast. News topics and the lagged implied variance IVAR_{t-1} together account for the greater part of the reduction in mean squared error within the random forest model, with the lagged realized variances contributing the remainder, and at the daily horizon the first two are closely matched. News topics lead for BBDC4, CSNA3 and ITUB4, while IVAR_{t-1} leads for GGBR4 and PETR4, in both cases only marginally. The exception is BOVA11, for which IVAR_{t-1} is clearly dominant. The lagged variable IVAR_{t-1} ranks among the two leading predictors for every asset, while further IVAR lags show less influence on the overall SE reduction. The lagged realized variances account for 11% of the reduction on average, ranging from 6% for PETR4 to 14% for CSNA3, with RV_{t-1} the most influential of the three for every asset.

Figure 6 examines the significance of news topics by highlighting the 10 most crucial topics for each asset in the daily forecast. The analysis reveals that the contribution of individual news topics to forecast performance varies across assets, even between companies within the same sector such as BBDC4 and ITUB4 (financial) and CSNA3 and GGBR4 (basic materials). Moreover, Figure 6 indicates that BBDC4, ITUB4, and PETR4 exhibit certain topics of high

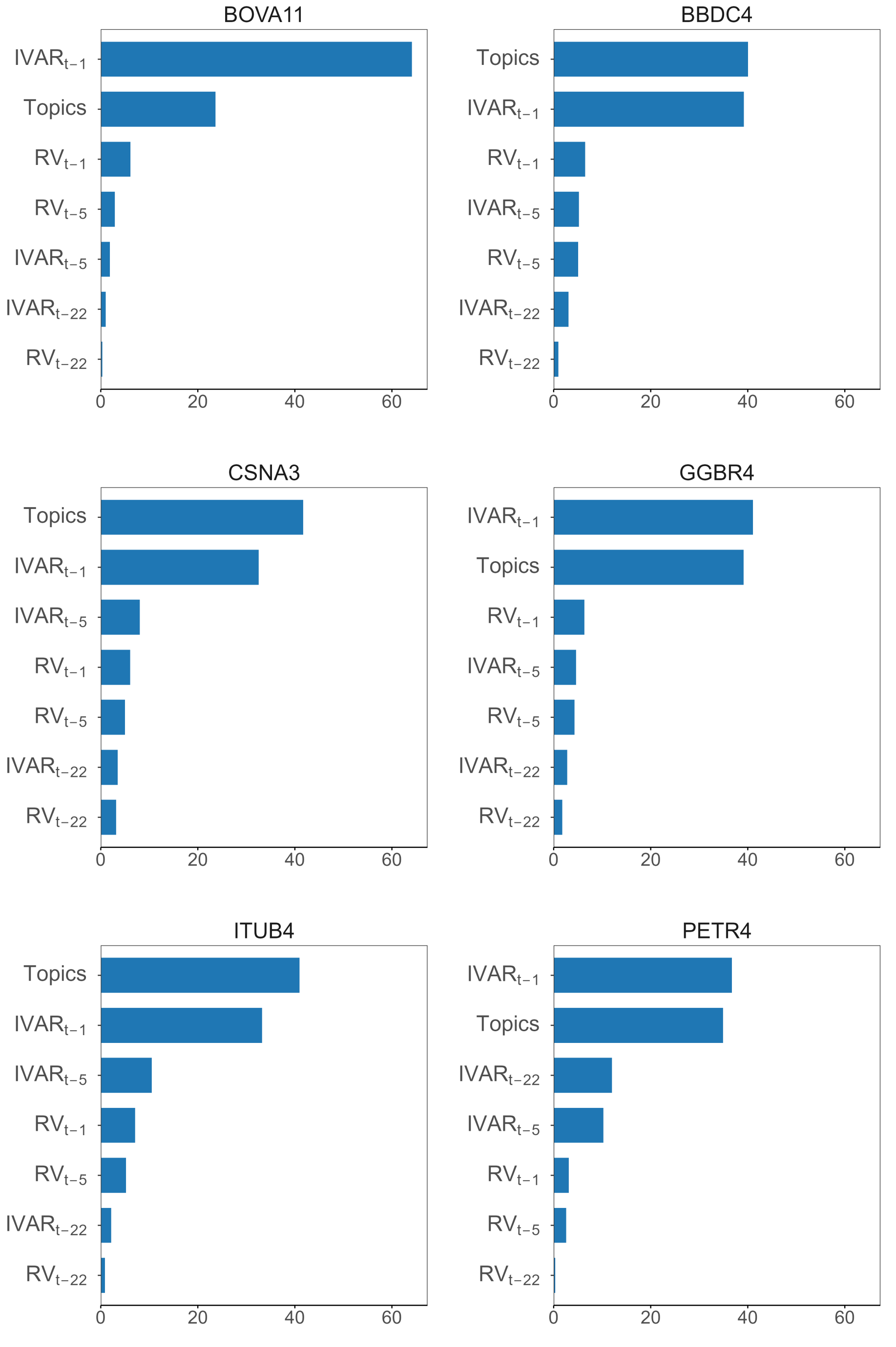

FIGURE 5.—Variable importance for NEWS IV in 1 day ahead forecast horizon. The variables importance were scaled in a way that the MDIs of all variables sums 100.

relevance, while for BOVA11, CSNA3 and GGBR4, the importance of the topics seems to be more evenly distributed.

In the context of a 5-day forecast, Figure 7 shows that the forecasts and the observed RV are smoother than in the daily observations. Table XII reveals that NEWS IV yields lower loss ratios relative to the HAR(OLS) benchmark for all six assets under the QLIKE criterion. Similarly, Table XIII indicates that NEWS IV exhibits superior relative predictive performance for BBDC4, CSNA3, GGBR4, ITUB4, and PETR4 when evaluating MSE, the exception being the case of BOVA11. Regarding the heteroskedasticity of residuals relative to the RV level, the results vary by asset. WLS outperforms OLS in BOVA11, but the evidence is mixed for PETR4, which may reflect the averaging out of measurement noise at longer horizons, as suggested by Bollerslev et al. (2016).

TABLE XII

QLIKE RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	1.000	1.002	1.001	1.000	1.003	0.999
LHARCJ (OLS)	1.034	1.006	1.002	1.002	1.008	1.019
SHAR (OLS)	0.998	0.999	1.000	1.000	1.000	1.001
HAR (WLS)	0.995	1.000	1.000	1.000	1.000	0.997
HARCJ (WLS)	0.994	1.000	1.000	1.000	1.000	0.999
LHARCJ (WLS)	0.988	0.999	1.000	0.999	0.999	1.000
SHAR (WLS)	0.992	0.999	1.000	1.000	1.000	0.997
NEWS IV (RF)	**0.986**	**0.989**	**0.994**	**0.994**	**0.987**	**0.983**
HAR (RF)	1.009	0.998	1.002	1.001	1.004	1.001
HARCJ (RF)	0.988	0.999	1.000	1.001	1.000	1.001
LHARCJ (RF)	**0.980**	0.999	0.999	0.999	0.999	0.999
SHAR (RF)	1.006	1.000	1.002	1.002	1.003	1.000

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

In the analysis of the importance of variables, as illustrated in Figure 8, news topics account for an increasing share of the reduction in prediction error as the forecast horizon lengthens

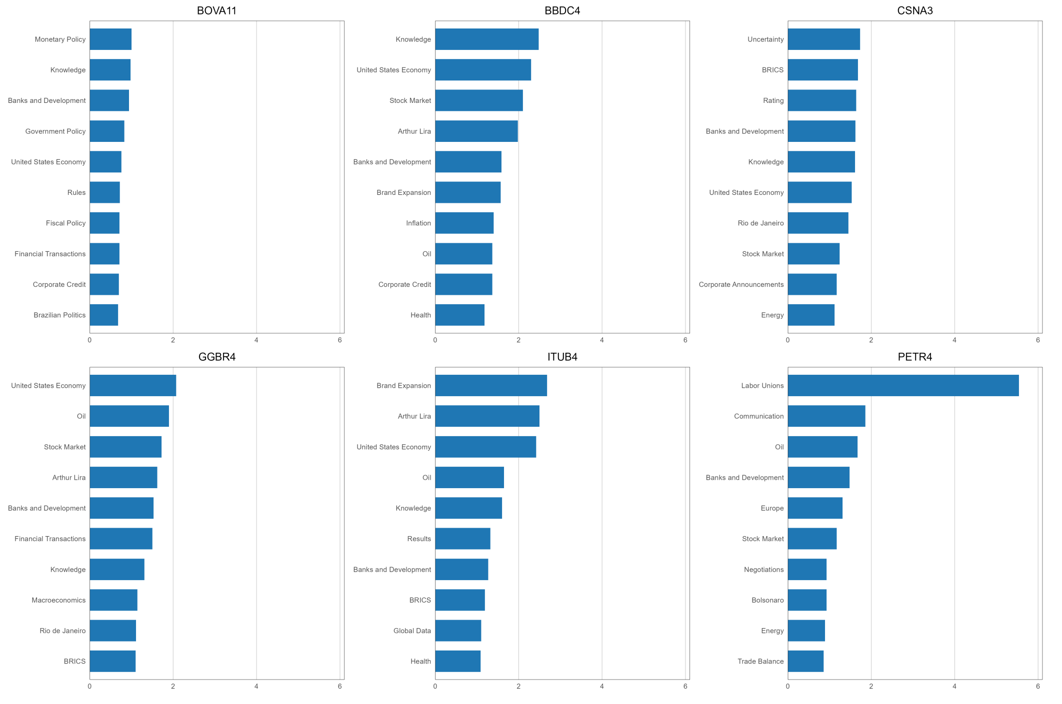

FIGURE 6.—News topics importance for NEWS IV in 1 day ahead forecast horizon. For clarity, only the 10 most important topics are displayed.

TABLE XIII

MSE RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.937**	1.065	1.035	1.006	1.131	1.119
LHARCJ (OLS)	1.376	1.396	1.141	1.131	1.167	1.935
SHAR (OLS)	**0.837**	0.950	1.009	1.012	1.047	1.025
HAR (WLS)	0.940	0.990	1.007	0.989	1.009	0.995
HARCJ (WLS)	**0.870**	1.016	1.031	0.993	0.953	1.158
LHARCJ (WLS)	**0.790**	0.977	1.036	0.959	0.889	1.153
SHAR (WLS)	**0.813**	0.943	1.015	0.989	1.023	0.981
NEWS IV (RF)	**3.588**	**0.580**	**0.580**	**0.650**	**0.560**	**0.640**
HAR (RF)	**1.552**	0.964	1.155	1.084	1.248	2.140
HARCJ (RF)	**0.837**	1.080	0.970	0.989	1.012	1.567
LHARCJ (RF)	**0.729**	1.983	0.932	0.934	0.886	2.190
SHAR (RF)	**1.404**	1.003	1.180	1.106	1.744	1.996

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

for all assets. Figure 9 illustrates that the topic “Banks and Development”[14] has now gained significant prominence across all assets.

[14]Important words for the topic “Banks and Development” comprise bank, plan, program, projects, development, goal, infrastructure, institution, financing, and performance (Cardoso and Nakane, 2024).

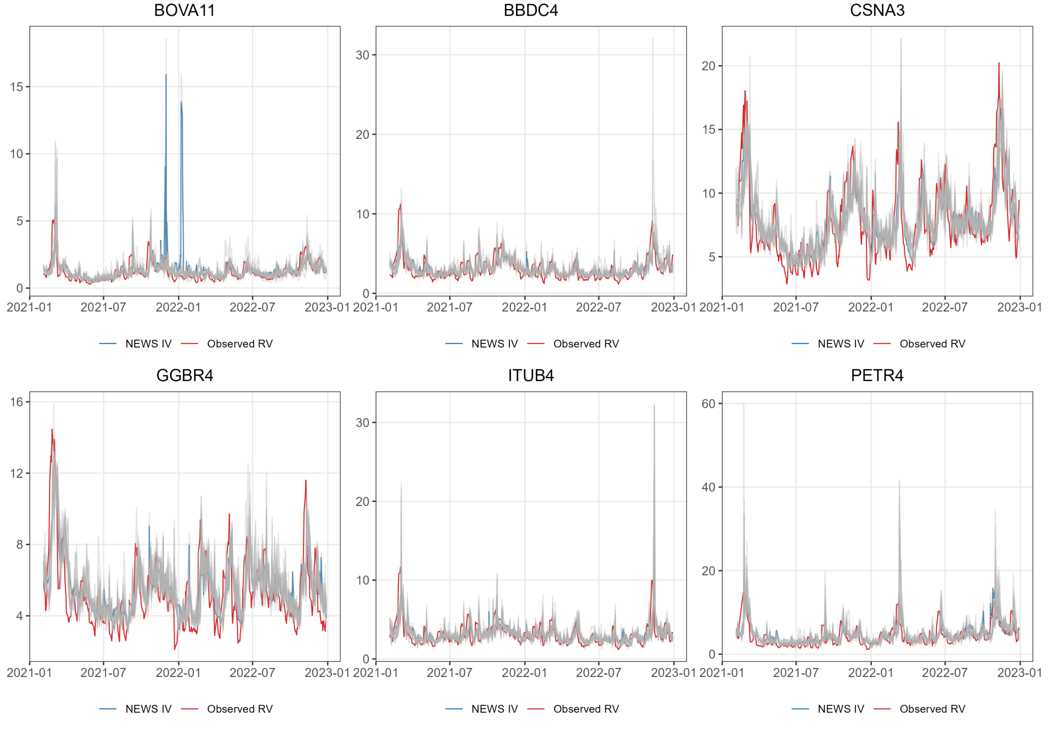

FIGURE 7.—Comparison between the observed realized volatility and the forecasts in 5 day ahead horizon. The gray lines reference the models estimated with the HAR-type variables.

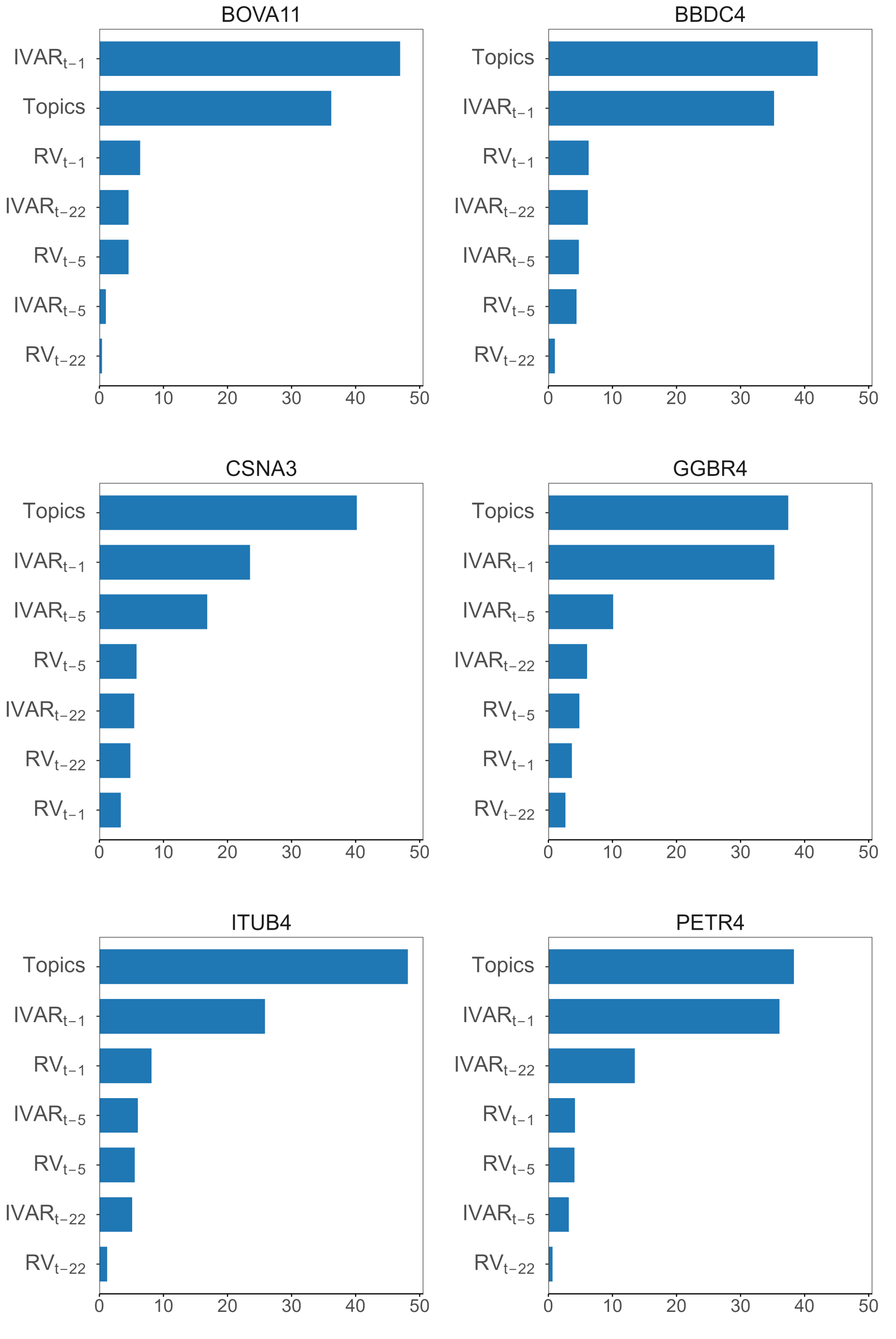

FIGURE 8.—Variable importance for NEWS IV in 5 day ahead forecast horizon. The variables importance were scaled in a way that the MDIs of all variables sums 100.

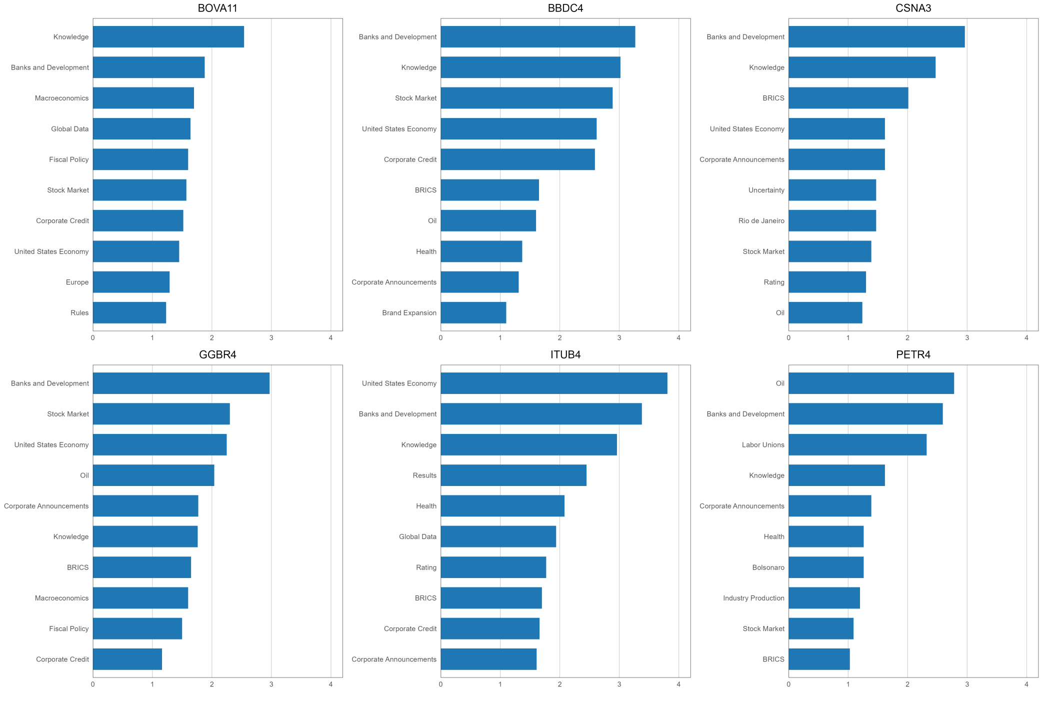

FIGURE 9.—News topics importance for NEWS IV in 5 day ahead forecast horizon. For clarity, only the 10 most important topics are displayed.

In the monthly forecast, Figure 10 shows that the NEWS IV forecast and the observed RV are even smoother than the weekly observations and that HAR-type variables models tend to overshoot the forecast. Table XIV shows that NEWS IV delivers lower loss ratios than the HAR(OLS) benchmark for all six assets under the QLIKE criterion, and Table XV shows the same for all six assets under MSE, where the margins are substantial for the individual stocks. BOVA11, which is the exception under MSE at the weekly horizon, also falls below the benchmark at the monthly horizon, though by a much narrower margin than the other assets. This asset-specific pattern is consistent with the shorter estimation sample available for the ETF discussed in Table I.

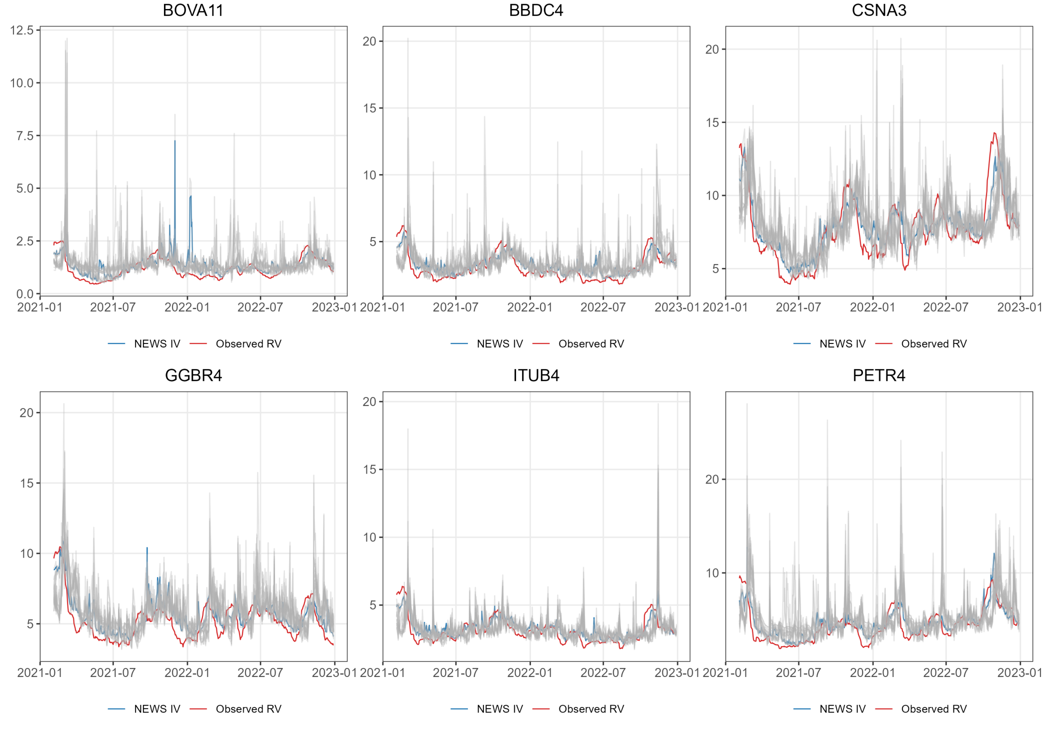

FIGURE 10.—Comparison between the observed realized volatility and the forecasts in 22 day ahead horizon. The gray lines reference the models estimated with the HAR-type variables.

TABLE XIV

QLIKE RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.993**	1.000	1.000	1.000	1.002	1.000
LHARCJ (OLS)	1.014	1.004	1.000	1.001	1.002	1.001
SHAR (OLS)	0.999	1.000	1.000	1.000	1.000	1.000
HAR (WLS)	0.994	1.000	1.000	1.000	1.001	0.997
HARCJ (WLS)	**0.989**	1.000	1.000	1.000	1.002	0.998
LHARCJ (WLS)	0.998	1.002	1.000	1.000	1.001	1.000
SHAR (WLS)	0.993	0.999	1.000	1.000	1.001	0.997
NEWS IV (RF)	**0.964**	**0.988**	**0.991**	**0.991**	**0.990**	**0.978**
HAR (RF)	1.015	1.007	1.001	1.004	1.007	1.003
HARCJ (RF)	**0.986**	1.004	0.998	1.003	1.003	1.012
LHARCJ (RF)	**0.986**	1.006	0.998	1.001	1.002	1.008
SHAR (RF)	1.015	1.005	1.001	1.004	1.006	1.000

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE XV

MSE RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.860**	1.018	0.952	1.001	1.156	1.168
LHARCJ (OLS)	1.691	1.610	0.975	1.075	1.185	1.560
SHAR (OLS)	**0.935**	0.981	1.008	1.003	1.009	1.011
HAR (WLS)	**0.994**	1.040	1.021	1.001	1.161	0.998
HARCJ (WLS)	**0.870**	1.047	0.971	0.999	1.208	1.264
LHARCJ (WLS)	**1.160**	1.399	1.022	0.964	1.127	1.410
SHAR (WLS)	**0.897**	1.001	1.030	0.996	1.145	0.997
NEWS IV (RF)	**0.978**	**0.345**	**0.253**	**0.332**	**0.326**	**0.312**
HAR (RF)	**3.577**	2.122	1.334	1.587	1.860	2.202
HARCJ (RF)	**1.360**	1.946	0.813	1.396	1.632	2.917
LHARCJ (RF)	**1.818**	2.483	0.898	1.243	1.414	2.628
SHAR (RF)	**3.422**	1.581	1.190	1.574	2.092	1.926

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

Figure 11 shows that the influence of variables on news topics has increased relative to weekly forecasts, underscoring their role in forecasting long-term volatility. Figure 12 illustrates that the topics “Knowledge”[15] and “Lula”[16] are the most important news’ topics in the MDI sense for the majority of assets (BOVA11, BBDC4, CSNA3, and ITUB4). For the government-controlled oil company PETR4, “Knowledge” is again the leading topic and “Lula” - referring to the Brazilian president - occupies the fourth position. Notably, the topic “Oil”[17], which is the single most important topic for PETR4 at the weekly horizon, no longer appears among its ten most important topics at the monthly horizon, consistent with commodity-specific news mattering for near-term rather than long-run volatility.

[15] Important words for the topic “Knowledge” comprise vaccination, social, study, digital, increase, population, education, needs, specialists and problems (Cardoso and Nakane, 2024).

[16] Important words for the topic “Lula” comprise former, Lula, PT (Workers’ Party), Lula Silva, governor, elections, electoral, campaign, TSE (Superior Electoral Court) and party (Cardoso and Nakane, 2024).

[17] Important words for the topic “Oil” comprise production, oil, price, record, commodities, fuels, volume, south, as per and tons (Cardoso and Nakane, 2024).

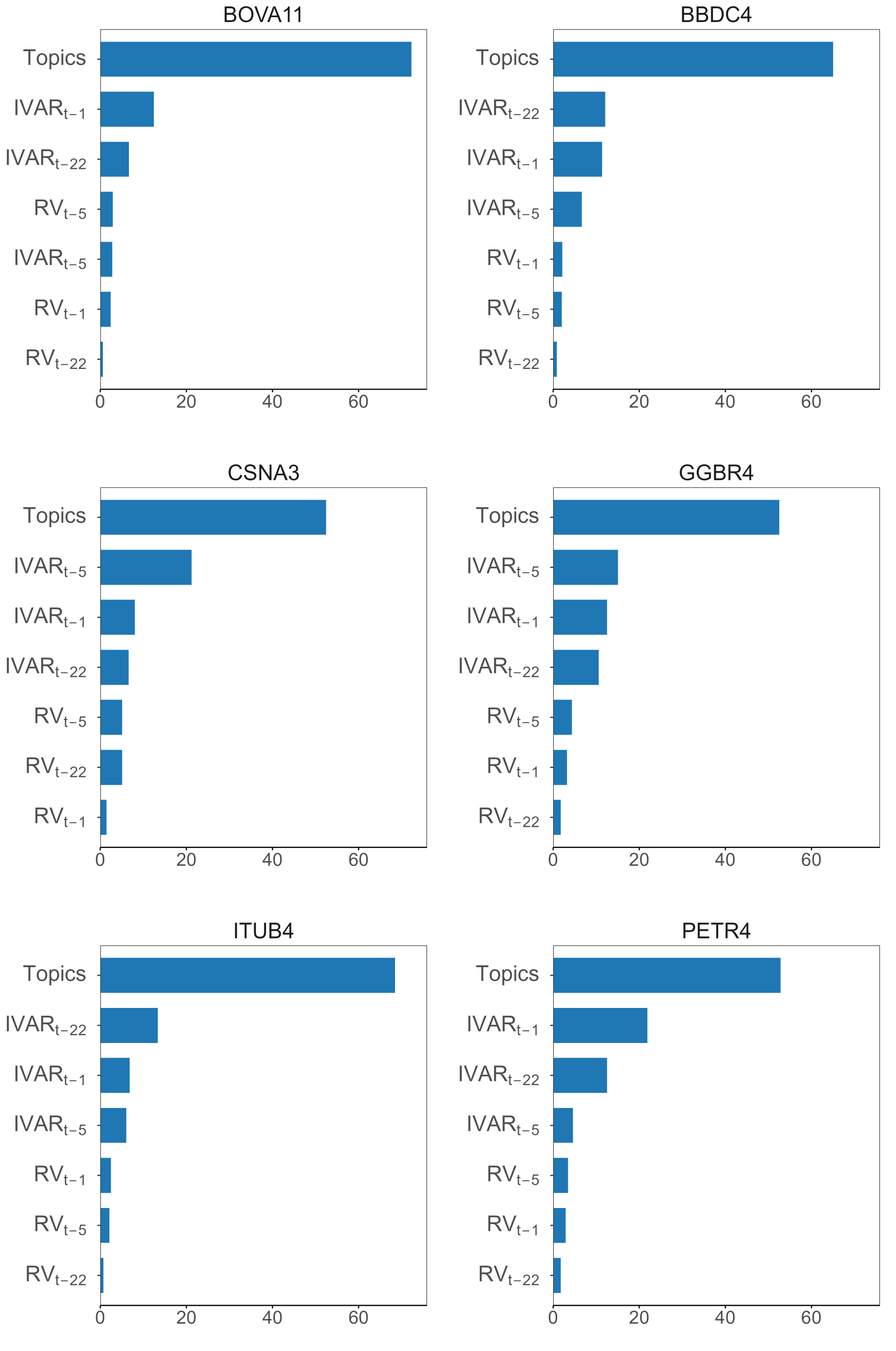

FIGURE 11.—Variable importance for NEWS IV in 22 day ahead forecast horizon. The variables importance were scaled in a way that the MDIs of all variables sums 100.

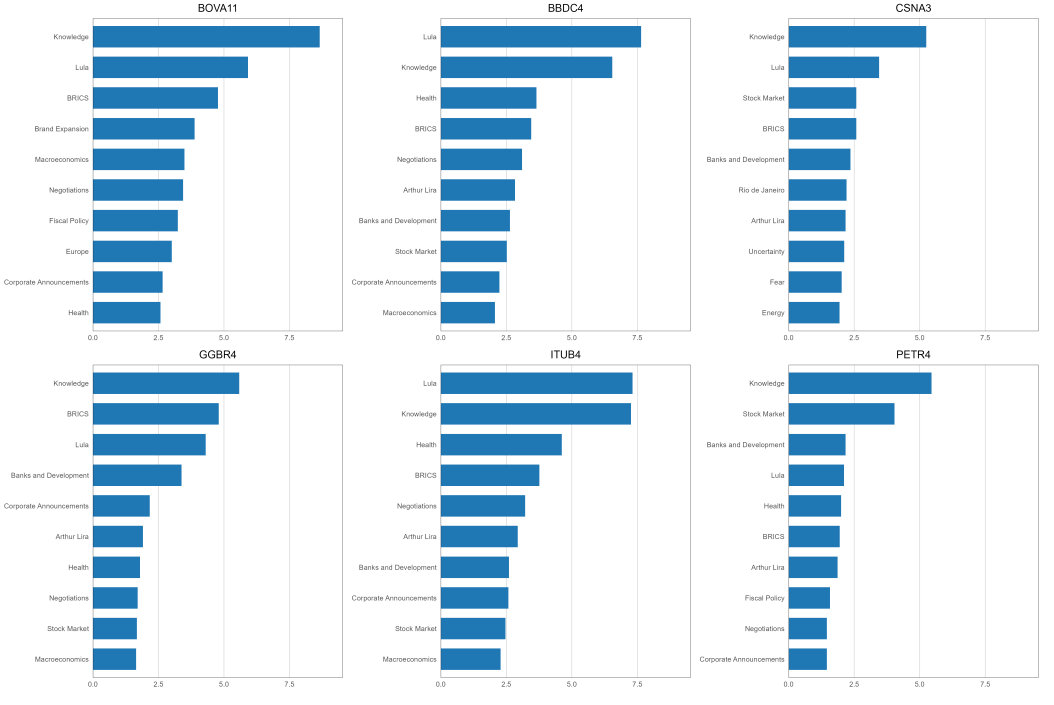

FIGURE 12.—News topics importance for NEWS IV in 22 day ahead forecast horizon. For clarity, only the 10 most important topics are displayed.

Overall, the relative predictive performance of the NEWS IV model, measured by loss ratios relative to the HAR(OLS) benchmark, improves at longer forecast horizons. This improvement is associated with a greater contribution of news-based variables and a reduced reliance on implied variance lags. The comparatively weaker results for BOVA11 may be partly explained by its shorter estimation sample, reported in Table I, which begins in April 2014 because implied volatility is not usable for the ETF before that date. This leaves the random forest roughly 30% fewer training observations than for the individual stocks (1,618 against 2,312), even though the evaluation period itself is complete for every asset.

6. CONCLUSION

This paper studies whether forward-looking option-based information and textual news jointly improve the forecasting of realized volatility. We propose the NEWS IV model, which integrates implied variance and LDA-derived news topics with the realized-variance components of the HAR model within a flexible machine learning framework, and evaluate its predictive performance relative to standard HAR-type models across multiple horizons.

Our results show that a model in which implied variance and news topics carry the predictive content, with the realized volatility components contributing a small and horizon-declining share, delivers substantial improvements in predictive performance at longer horizons. In particular, NEWS IV yields significantly lower forecast losses relative to the HAR benchmark at the weekly and monthly horizons, while performing comparably at the daily horizon. These findings indicate that the informational content of news and option-implied measures is especially relevant for capturing medium- and long-run components of volatility.

The analysis of variable importance highlights a clear horizon-dependent pattern. Implied variance contributes most to short-term forecasts, consistent with its interpretation as a forward-looking measure of near-term uncertainty. In contrast, news topics account for an increasing share of the reduction in forecast error as the prediction horizon lengthens, suggesting that textual information captures slower-moving drivers of volatility related to macroeconomic conditions, policy developments, and firm-specific news.

Taken together, these results provide evidence that different sources of information contribute to volatility forecasting at different horizons. Option prices embed timely market-based assessments of near-term risk, while textual news conveys broader informational flows that affect volatility dynamics over longer periods. The interaction between these sources is particularly informative in an environment such as the Brazilian market, where volatility is high and closely linked to macroeconomic and political developments.

These improvements are economically meaningful, particularly at longer horizons where accurate volatility forecasts are critical for risk management and portfolio allocation decisions.

From a methodological perspective, our findings also underscore the importance of flexible modeling approaches when incorporating high-dimensional textual data. While nonlinear methods do not substantially improve performance when applied to traditional HAR variables alone, they become essential when combining implied volatility with a large set of news-based predictors.

Future research could explore alternative approaches to modeling nonlinearities, as well as extend the analysis to multivariate settings, including the joint modeling of volatility and covariance for portfolio allocation. In addition, investigating the role of alternative textual data sources and option maturities may provide further insights into the interaction between market expectations and informational flows in volatility dynamics.

REFERENCES

Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (2018). Anuário estatístico brasileiro do petróleo, gás natural e biocombustíveis 2018. [10]

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4):701–720. [2, 14]

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453):42–55. [2, 5]

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625. [27]

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3):1259–1294. [3]

Asgharian, H., Christiansen, C., and Hou, A. J. (2023). The effect of uncertainty on stock market volatility and correlation. *Journal of Banking & Finance*, 154:106929. [3]

Astorino, E. S., Chague, F., Giovannetti, B. C., and Silva, M. E. d. (2017). Variance premium and implied volatility in a low-liquidity option market. *Revista Brasileira de Economia*. [3, 7]

Audrino, F., Sigrist, F., and Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, 36(2):334–357. [3]

Barndorff-Nielsen, O. E. and Shephard, N. (2002a). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(2):253 – 280. [2, 5]

Barndorff-Nielsen, O. E. and Shephard, N. (2002b). Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(2):167–241. [2, 5]

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37. [13]

Barndorff-Nielsen, O. E., Kinnebrock, S., and Shephard, N. (2010). Measuring downside risk – realized semivariance. In *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*. Oxford University Press. [15, 16]

Bekaert, G. and Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. *Journal of Econometrics*, 183(2):181–192. [2, 21, 24]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. [18]

Bodilsen, S. T. and Lunde, A. (2025). Exploiting news analytics for volatility forecasting. *Journal of Applied Econometrics*, 40(1):18–36. [3, 26]

Bollerslev, T., Kretschmer, U., Pigorsch, C., and Tauchen, G. (2009a). A discrete-time model for daily S&P 500 returns and realized variations: Jumps and leverage effects. *Journal of Econometrics*, 150(2):151–166. Recent Development in Financial Econometrics. [15]

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18. [31]

Bollerslev, T., Tauchen, G., and Zhou, H. (2009b). Expected stock returns and variance risk premia. *The Review of Financial Studies*, 22(11):4463–4492. [21]

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2012). Which news moves stock prices? a textual analysis. NBER Working Paper Series 18725, National Bureau of Economic Research. [20]

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. [23, 24]

Cardoso, G. R. and Nakane, M. I. (2024). What's in a headline? News impact on the Brazilian economy. Working Papers, Department of Economics 2024_12, University of São Paulo (FEA-USP). [17, 18, 20, 23, 33, 40, 48]

Christensen, K., Siggaard, M., and Veliyev, B. (2023). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, 21(5):1680–1727. [3, 24]

Clements, A. and Preve, D. P. (2021). A practical guide to harnessing the HAR volatility model. *Journal of Banking & Finance*, 133:106285. [27]

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196. [2, 10, 13, 22]

Corsi, F. and Reno, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics*, 30(3):368–380. [2, 15, 23]

Fernandes, M. and Pereira, M. A. P. (2025). Forecasting realized volatility using news flow. *Quarterly Review of Economics and Finance*, 104:102040. [3, 20, 21]

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801. [15]

Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., and Gipp, B. (2018). Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions. In *Transforming Digital Worlds*, Lecture Notes in Computer Science. Springer. [18]

Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161. [6]

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497. [23, 25, 26]

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag, New York, NY, 2nd edition. [24, 26]

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844. [24]

Kambouroudis, D. S., McMillan, D. G., and Tsakou, K. (2021). Forecasting realized volatility: The role of implied volatility, leverage effect, overnight returns, and volatility of realized volatility. *Journal of Futures Markets*, 41(10):1618–1639. [3, 24]

Liu, L. Y., Patton, A. J., and Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1):293–311. [6]

Mentch, L. and Zhou, S. (2020). Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36. [23]

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256. [26]

Patton, A. J. and Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *The Review of Economics and Statistics*, 97(3):683–697. [2, 16, 17, 23, 27]

Petrobras (2018). Earnings release 2018. [10]

Probst, P. and Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181):1–18. [25]

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3):e1301. [24, 25]

Pöttker, H. (2003). News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511. [17]

Quantitative Analytics Bloomberg L.P. (2019). Equity Implied Volatility Surface. Technical report, Bloomberg. [22]

Scanlan, C. (2000). *Reporting and Writing: Basics for the 21st Century*. Oxford University Press, New York, NY, 3rd edition. [17]

Wright, M. N. and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17. [24, 25]

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411. [6]

APPENDIX A: MAIN WORDS PER TOPIC

Table A.1 outlines the labels and key terms for each topic Cardoso and Nakane (2024).

TABLE A.1: Topic labels and key terms

Label	**Important Words (Portuguese)**	**Important Words (English)**
United States Economy	eua, estados unidos, fed, americano, mercados, espera, americana, foco, globais, riscos	USA, United States, Federal Reserve, American, markets, wait, American (feminine), focus, global, risks
Knowledge	vacinação, social, estudo, digital, aumentar, população, educação, precisa, especialistas, problemas	vaccination, social, study, digital, increase, population, education, needs, specialists, problems
Fiscal Policy	fiscal, recursos, pagamento, estados, medidas, reduzir, orçamento, dívida, pública, auxílio	tax, resources, payment, states, measures, reduce, budget, debt, public, aid
Fear	estado, guerra, segurança, direito, decidiu, autoridades, icms, valores, especial, Minas Gerais	state, war, security, right, decided, authorities, tax, values, special, Minas Gerais
Communication	acordo, serviços, rede, uso, negociações, acesso, via, negociação, serviço, internet	agreement, services, network, use, negotiations, access, route, negotiation, service, internet
Stock Market	queda, alta, dólar, forte, bolsa, ibovespa, sessão, movimento, exterior, york	drop, rise, dollar, strong, stock market, ibovespa, session, movement, abroad, New York

Continued on next page

Table A.1 – Continued from previous page

Label	Important Words (Portuguese)	Important Words (English)
Financial Transactions	mercado, compra, ativos, venda, investidores, financeiro, aquisição, papéis, comprar, vender	market, purchase, assets, sale, investors, financial, acquisition, securities, buy, sell
Inflation	alta, preços, índice, março, dezembro, outubro, novembro, abril, maio, fevereiro	rise, prices, index, March, December, October, November, April, May, February
Labor Unions	nacional, união, São Paulo, entidade, cidades, paulista, democracia, trabalhadores, reajuste, concessão	national, union, São Paulo, entity, cities, from São Paulo, democracy, workers, adjustment, concession
Oil	produção, petróleo, preço, recorde, commodities, combustíveis, volume, sul, conforme, toneladas	production, oil, price, record, commodities, fuels, volume, south, as per, tons
Rules	dias, medida, prazo, mudança, mudanças, tema, regras, Brasília, transição, poderá	days, measure, term, change, changes, theme, rules, Brasília, transition, may
Macroeconomics	setor, aumento, crescimento, investimento, prevê, pib, expansão, desempenho, anual, ritmo	sector, increase, growth, investment, predicts, GDP, expansion, performance, annual, pace
Corporate Announcements	empresa, companhia, informou, anunciou, ceo, administração, funcionários, comunicado, realizada, anuncia	company, corporation, informed, announced, CEO, administration, employees, statement, held, announces

Continued on next page

Table A.1 – Continued from previous page

Label	Important Words (Portuguese)	Important Words (English)
Trade Balance	bilhões, us, bilhão, bi, total, ação, positivo, comercial, euros, negativo	billions, US, billion, billion (abbr.), total, share/action, positive, trade, euros, negative
Corporate Credit	empresas, crédito, operações, maiores, bancos, clientes, caixa, dinheiro, companhias, custo	companies, credit, operations, largest, banks, clients, cash, money, companies, cost
Rating	risco, sp, relatório, afirma, nota, segue, vê, avaliação, avalia, alto	risk, SP, report, states, note, follows, see, evaluation, assesses, high
Research	pesquisa, principais, impacto, europa, mostra, pressão, semanas, econômico, instituto, divulgada	research, main, impact, Europe, shows, pressure, weeks, economic, institute, released
Monetary Policy	inflação, Banco Central, BC, acima, IBGE, expectativa, abaixo, divulgado, meta, média	inflation, Central Bank, CB, above, IBGE (Brazilian Institute of Geography and Statistics), expectation, below, disclosed, target, average
Lula	ex, Lula, pt, Lula Silva, governador, eleições, eleitoral, campanha, TSE, partido	former, Lula, PT (Workers' Party), Lula Silva, governor, elections, electoral, campaign, TSE (Superior Electoral Court), party

Continued on next page

Table A.1 – Continued from previous page

Label	**Important Words (Portuguese)**	**Important Words (English)**
Brazilian Politics	projeto, senado, lei, câmara, congresso, proposta, análise, casa, aprovou, aprovação	project, senate, law, chamber, congress, proposal, analysis, house, approved, approval
Brazilian Presidents	presidente, executivo, atual, vice, domingo, eleito, cargo, comando, chefe, entrevista	president, executive, current, vice, Sunday, elected, position, command, chief, interview
Uncertainty	economia, global, política, brasileira, crise, cenário, recuperação, econômica, internacional, ambiente	economy, global, politics, Brazilian, crisis, scenario, recovery, economic, international, environment
Justice	federal, STF, decisão, processo, justiça, Supremo Tribunal Federal, público, pedido, ministros, defesa	federal, Supreme Federal Court, decision, process, justice, Supreme Federal Court, public, request, ministers, defense
Global Data	dados, mundo, divulgação, demanda, quase, principal, manter, devido, continua, Japão	data, world, disclosure, demand, almost, main, maintain, due, continues, Japan
Negotiations	início, operação, negócios, volta, frente, busca, destaques, vista, estratégia, passou	beginning, operation, business, return, front, search, highlights, view, strategy, passed
Europe	Ucrânia, ficar, série, começa, locais, evitar, Reino Unido, vida, começou, qualquer	Ukraine, stay, series, begins, places, avoid, United Kingdom, life, started, any

Continued on next page

Table A.1 – Continued from previous page

Label	Important Words (Portuguese)	Important Words (English)
Copom Meetings	juros, taxa, redução, reunião, corte, taxas, longo, semana passada, mantém, espaço	interest, rate, reduction, meeting, cut, rates, long, last week, keeps, space
Results	milhões, trimestre, período, lucro, resultado, receita, anterior, registrou, lucro líquido, resultados	millions, quarter, period, profit, result, revenue, previous, recorded, net profit, results
Rio de Janeiro	rio, gestão, região, cidade, Rio de Janeiro, centro, áreas, modelo, construção, receber	river, management, region, city, Rio de Janeiro, center, areas, model, construction, receive
Industry Production	passado, janeiro, setembro, vendas, queda, julho, agosto, junho, indústria, comparação	past, January, September, sales, drop, July, August, June, industry, comparison
BRICS	Brasil, país, China, países, brasileiro, comércio, mundial, Argentina, Índia, UE	Brazil, country, China, countries, Brazilian, trade, global, Argentina, India, EU
Russia	Rússia, pessoas, brasileiros, tempo, evento, possibilidade, causa, sábado, curtas, encontro	Russia, people, Brazilians, time, event, possibility, cause, Saturday, shorts, meeting
Bolsonaro	Bolsonaro, Petrobras, conta, Twitter, estatal, família, redes sociais, fala, alvo, fica	Bolsonaro, Petrobras, account, Twitter, state-owned, family, social media, speaks, target, stays
Government Policy	governo, ministro, afirmou, ministério, secretário, secretaria, fazenda, defende, Guedes, deixar	government, minister, stated, ministry, secretary, secretariat, treasury, defends, Guedes, leave

Continued on next page

Table A.1 – Continued from previous page

Label	Important Words (Portuguese)	Important Words (English)
Energy	energia, diretor, agência, leilão, contratos, contrato, gás, geração, responsável, consumidores	energy, director, agency, auction, contracts, contract, gas, generation, responsible, consumers
Institutional Investing	capital, investimentos, fundo, participação, vale, plataforma, controle, negócio, restrições, acionistas	capital, investments, fund, participation, worth, platform, control, business, restrictions, shareholders
Arthur Lira	feira, terça, semana, futuro, agenda, manhã, melhor, voltou, Lira, força	fair, Tuesday, week, future, schedule, morning, better, returned, Lira, strength
Health	ações, saúde, casos, oferta, tecnologia, vacinas, vacina, parceria, informações, planos	actions, health, cases, offer, technology, vaccines, vaccine, partnership, information, plans
Banks and Development	banco, plano, programa, projetos, desenvolvimento, objetivo, infraestrutura, instituição, financiamento, atuação	bank, plan, program, projects, development, goal, infrastructure, institution, financing, performance
Brand Expansion	cerca, base, produtos, marca, chegou, custos, capacidade, linha, chega, milhão	about, base, products, brand, arrived, costs, capacity, line, arrives, million

APPENDIX B: OTHER LOSS FUNCTIONS

This appendix presents the results with the MSE-LOG, MSE-SD, HMSE, MAE, MAE-LOG, MAE-SD, and MAPE loss functions, defined below.

MSE-LOG: $\left(\log RV_{t+h|t} - \log \widehat{RV}_{t+h|t}\right)^2$

MSE-SD: $\left(\sqrt{RV_{t+h|t}} - \sqrt{\widehat{RV}_{t+h|t}}\right)^2$

HMSE: $\left(\frac{RV_{t+h|t}}{\widehat{RV}_{t+h|t}} - 1\right)^2$

MAE: $\left|RV_{t+h|t} - \widehat{RV}_{t+h|t}\right|$

MAE-LOG: $\left|\log RV_{t+h|t} - \log \widehat{RV}_{t+h|t}\right|$

MAE-SD: $\left|\sqrt{RV_{t+h|t}} - \sqrt{\widehat{RV}_{t+h|t}}\right|$

MAPE: $\left|\frac{RV_{t+h|t}}{\widehat{RV}_{t+h|t}} - 1\right|$

B.1. *1 day forecast horizon*

TABLE B.1

MSE-LOG RESULTS FOR H=1

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	**1.000**	**1.000**	**1.000**
HARCJ (OLS)	0.996	0.996	1.009	**1.001**	**1.007**	1.007
LHARCJ (OLS)	**0.946**	**0.944**	**0.971**	**0.968**	**0.963**	**0.953**
SHAR (OLS)	1.009	1.025	0.987	**0.995**	1.066	**0.999**
HAR (WLS)	1.149	1.006	0.999	**1.000**	**1.006**	1.194
HARCJ (WLS)	1.101	0.998	1.010	**1.003**	**1.033**	1.053
LHARCJ (WLS)	1.095	**0.942**	0.971	**0.969**	**0.959**	**0.959**
SHAR (WLS)	1.110	1.029	0.991	**0.997**	1.102	1.047
NEWS IV (RF)	1.024	**0.964**	**0.937**	**0.973**	**0.940**	**0.963**
HAR (RF)	1.078	1.100	1.124	1.067	1.072	1.042
HARCJ (RF)	1.030	1.066	1.059	1.036	1.034	1.032
LHARCJ (RF)	**0.946**	**0.985**	1.015	**1.004**	**0.980**	**0.942**
SHAR (RF)	1.091	1.115	1.167	1.096	1.147	**1.036**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.2

MSE-SD RESULTS FOR H=1

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	**1.000**	**1.000**	1.000	**1.000**	1.000	**1.000**
HARCJ (OLS)	**1.003**	**0.999**	1.012	**1.013**	0.992	1.020
LHARCJ (OLS)	**0.921**	**0.949**	**0.974**	**0.978**	**0.943**	**0.952**
SHAR (OLS)	0.988	1.018	**0.973**	**0.992**	1.066	**0.993**
HAR (WLS)	**0.984**	**0.995**	0.999	**0.999**	0.995	**1.000**
HARCJ (WLS)	**0.982**	**0.993**	1.010	**1.006**	0.986	1.011
LHARCJ (WLS)	**0.909**	**0.933**	**0.971**	**0.969**	**0.942**	**0.929**
SHAR (WLS)	**0.976**	1.015	**0.975**	**0.992**	1.060	**0.968**
NEWS IV (RF)	1.141	**0.970**	**0.940**	**1.005**	**0.917**	**0.961**
HAR (RF)	1.070	1.126	1.126	**1.068**	1.063	1.050
HARCJ (RF)	1.008	1.079	1.067	**1.039**	1.056	1.058
LHARCJ (RF)	**0.932**	1.031	1.031	**0.998**	**0.995**	**0.914**
SHAR (RF)	1.080	1.112	1.169	1.119	1.201	**1.030**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.3

HMSE RESULTS FOR H=1

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	1.000	**1.000**	**1.000**	1.000	**1.000**	1.000
HARCJ (OLS)	**0.794**	**1.032**	**0.997**	1.037	**1.097**	0.872
LHARCJ (OLS)	1.102	1.159	**1.014**	**1.103**	**1.076**	0.941
SHAR (OLS)	1.001	**1.037**	**1.014**	1.021	**1.073**	1.084
HAR (WLS)	**0.941**	**0.982**	**0.987**	**0.970**	**0.991**	**0.690**
HARCJ (WLS)	**0.919**	**1.011**	**0.980**	1.016	**1.009**	0.782
LHARCJ (WLS)	**0.959**	**0.973**	**0.976**	**0.996**	**0.957**	0.752
SHAR (WLS)	**0.917**	**1.012**	**0.999**	**0.972**	**1.061**	0.785
NEWS IV (RF)	2.635	**1.074**	1.114	**1.121**	**1.038**	0.829
HAR (RF)	**0.999**	**1.312**	1.181	**1.089**	**1.107**	**0.681**
HARCJ (RF)	**1.027**	**1.119**	**1.080**	**1.106**	1.161	0.720
LHARCJ (RF)	**0.935**	**1.094**	**1.093**	**1.079**	**1.054**	**0.619**
SHAR (RF)	**1.066**	**1.120**	1.226	**1.178**	**1.323**	**0.699**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.4

MAE RESULTS FOR H=1

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	1.000	**1.000**	**1.000**	**1.000**	**1.000**	1.000
HARCJ (OLS)	**0.968**	**1.007**	**0.997**	**1.007**	**1.025**	1.002
LHARCJ (OLS)	1.086	1.100	**0.990**	**1.008**	**1.027**	1.042
SHAR (OLS)	0.994	1.008	**0.989**	**0.996**	**1.034**	1.032
HAR (WLS)	**0.984**	**0.986**	**1.000**	**0.994**	**0.994**	**0.932**
HARCJ (WLS)	**0.974**	**0.990**	**0.996**	**1.001**	**0.997**	**0.967**
LHARCJ (WLS)	**0.978**	**0.985**	**0.980**	**0.976**	**0.978**	**0.926**
SHAR (WLS)	**0.978**	**0.991**	**0.990**	**0.985**	**1.031**	**0.960**
NEWS IV (RF)	1.202	**1.004**	**1.000**	1.052	**1.001**	**0.931**
HAR (RF)	1.036	1.059	1.075	**1.031**	**1.037**	**0.956**
HARCJ (RF)	**1.016**	**1.033**	1.041	**1.010**	1.052	**0.968**
LHARCJ (RF)	**0.989**	**1.095**	1.036	**1.000**	**1.010**	**0.914**
SHAR (RF)	1.043	**1.051**	1.087	**1.060**	1.108	**0.929**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.5

MAE-LOG RESULTS FOR H=1

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	**1.000**	**1.000**	1.000	**1.000**	1.000	**1.000**
HARCJ (OLS)	**1.007**	**0.999**	1.004	**1.002**	1.003	1.010
LHARCJ (OLS)	**0.988**	**0.972**	**0.986**	**0.986**	**0.974**	**0.984**
SHAR (OLS)	**1.009**	1.009	**0.997**	**0.998**	1.032	**1.004**
HAR (WLS)	1.079	**1.002**	**1.000**	**1.000**	1.006	1.102
HARCJ (WLS)	1.054	**1.000**	1.004	**1.002**	1.015	1.033
LHARCJ (WLS)	1.057	**0.972**	**0.987**	**0.985**	**0.959**	**0.987**
SHAR (WLS)	1.075	1.011	**0.999**	**0.998**	1.048	1.034
NEWS IV (RF)	**1.025**	**0.982**	**0.970**	**0.983**	**0.967**	**0.976**
HAR (RF)	1.059	1.049	1.052	**1.029**	1.046	**1.024**
HARCJ (RF)	1.027	1.034	1.021	**1.006**	1.020	1.020
LHARCJ (RF)	**0.988**	**0.994**	**0.999**	**0.994**	**0.984**	**0.977**
SHAR (RF)	1.062	1.060	1.080	1.048	1.075	**1.016**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.6

MAE-SD RESULTS FOR H=1

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	**1.000**	**1.000**	**1.000**	**1.000**	1.000	**1.000**
HARCJ (OLS)	**0.993**	**1.001**	**1.001**	**1.003**	1.000	1.020
LHARCJ (OLS)	**0.982**	**0.988**	**0.981**	**0.985**	**0.969**	0.990
SHAR (OLS)	**1.010**	**1.010**	**0.992**	**0.991**	1.037	**1.007**
HAR (WLS)	**0.991**	**0.997**	**1.001**	**0.999**	0.997	**0.997**
HARCJ (WLS)	**0.992**	**0.998**	**1.003**	**1.001**	0.995	1.012
LHARCJ (WLS)	**0.976**	**0.976**	**0.982**	**0.981**	**0.966**	**0.974**
SHAR (WLS)	**0.999**	**1.006**	**0.995**	**0.990**	1.032	**0.992**
NEWS IV (RF)	1.071	**0.987**	**0.977**	**1.009**	**0.974**	**0.982**
HAR (RF)	1.047	1.063	1.060	**1.019**	1.038	1.025
HARCJ (RF)	**1.017**	1.039	1.031	**1.001**	1.031	1.029
LHARCJ (RF)	**0.984**	**1.013**	**1.011**	**0.986**	**0.991**	**0.963**
SHAR (RF)	1.057	1.059	1.082	**1.044**	1.087	**1.004**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.7

MAPE RESULTS FOR H=1

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	**1.000**	**1.000**	1.000	**1.000**	1.000
HARCJ (OLS)	**0.889**	**1.009**	**0.996**	1.012	**1.016**	0.911
LHARCJ (OLS)	1.013	1.083	**0.988**	**1.005**	**1.011**	0.863
SHAR (OLS)	1.013	1.014	**1.002**	0.996	**1.032**	1.036
HAR (WLS)	0.963	**0.985**	**0.998**	**0.987**	**0.997**	0.823
HARCJ (WLS)	0.961	**0.995**	**0.994**	1.004	**1.012**	0.862
LHARCJ (WLS)	0.972	**0.996**	**0.983**	**0.979**	**0.990**	0.837
SHAR (WLS)	0.967	**0.993**	**1.000**	**0.978**	**1.031**	0.870
NEWS IV (RF)	1.278	**1.038**	1.041	1.078	**1.049**	0.857
HAR (RF)	1.002	**1.053**	1.066	**1.022**	**1.048**	**0.803**
HARCJ (RF)	1.015	**1.028**	1.044	**1.016**	1.075	0.833
LHARCJ (RF)	0.977	**1.034**	1.049	**1.008**	**1.032**	**0.775**
SHAR (RF)	1.020	**1.026**	1.085	**1.035**	1.103	**0.790**

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

B.2. *5 day forecast horizon*

TABLE B.8

MSE-LOG RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.986	1.011	0.996	0.993	1.026	1.000
LHARCJ (OLS)	0.916	0.956	0.976	0.949	0.969	0.991
SHAR (OLS)	0.986	0.999	1.000	0.996	1.010	1.011
HAR (WLS)	1.097	0.991	0.999	0.998	1.008	0.976
HARCJ (WLS)	1.145	1.012	0.995	0.994	1.011	0.974
LHARCJ (WLS)	0.999	0.957	0.974	0.948	0.949	0.944
SHAR (WLS)	1.069	0.990	0.999	0.995	1.018	1.018
NEWS IV (RF)	**0.708**	**0.587**	**0.556**	**0.586**	**0.550**	**0.585**
HAR (RF)	1.159	0.963	1.052	0.991	1.085	1.046
HARCJ (RF)	0.939	0.955	0.946	0.957	0.997	1.015
LHARCJ (RF)	**0.815**	0.890	0.871	0.894	0.938	0.961
SHAR (RF)	1.140	1.023	1.091	1.017	1.083	1.024

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.9

MSE-SD RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.995	1.050	1.017	1.015	1.011	1.063
LHARCJ (OLS)	0.903	1.020	1.020	0.976	0.952	1.116
SHAR (OLS)	0.950	0.979	1.001	0.992	1.003	0.985
HAR (WLS)	0.955	0.986	1.007	1.001	0.989	1.026
HARCJ (WLS)	0.945	1.017	1.018	1.011	0.987	1.054
LHARCJ (WLS)	**0.852**	0.968	1.008	0.964	0.929	1.057
SHAR (WLS)	0.917	0.971	1.009	0.994	0.991	1.013
NEWS IV (RF)	**1.030**	**0.576**	**0.568**	**0.603**	**0.543**	**0.582**
HAR (RF)	1.184	0.964	1.108	1.030	1.083	1.213
HARCJ (RF)	0.901	0.992	0.970	0.983	0.964	1.141
LHARCJ (RF)	**0.790**	1.006	0.911	0.916	0.902	1.150
SHAR (RF)	1.148	1.008	1.137	1.060	1.204	1.162

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.10

HMSE RESULTS FOR H=5

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.731**	1.067	1.039	1.021	1.210	0.978
LHARCJ (OLS)	1.204	1.439	1.145	1.138	1.257	1.246
SHAR (OLS)	0.921	0.977	1.020	1.004	1.045	1.029
HAR (WLS)	0.821	0.942	0.969	0.918	1.010	0.741
HARCJ (WLS)	**0.744**	0.971	0.989	0.940	1.010	0.896
LHARCJ (WLS)	0.753	0.993	1.004	0.904	0.932	0.908
SHAR (WLS)	0.746	0.916	0.984	0.909	1.025	0.744
NEWS IV (RF)	**3.991**	**0.586**	**0.547**	**0.622**	**0.591**	**0.463**
HAR (RF)	0.996	0.913	1.077	1.026	1.202	0.913
HARCJ (RF)	**0.656**	1.026	0.902	1.016	1.080	0.825
LHARCJ (RF)	**0.575**	1.120	0.892	0.938	0.907	0.785
SHAR (RF)	0.963	0.953	1.122	1.077	1.701	0.859

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.11

MAE RESULTS FOR H=5

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.903**	1.028	1.000	1.001	1.064	0.998
LHARCJ (OLS)	1.213	1.156	1.041	1.025	1.111	1.189
SHAR (OLS)	0.960	0.973	1.000	0.988	1.002	1.012
HAR (WLS)	0.955	0.989	0.996	0.977	1.017	0.924
HARCJ (WLS)	0.909	1.000	0.996	0.979	1.017	0.987
LHARCJ (WLS)	0.903	0.997	0.998	0.953	0.987	1.001
SHAR (WLS)	0.915	0.965	0.995	0.965	1.015	0.927
NEWS IV (RF)	**1.062**	**0.766**	**0.742**	**0.804**	**0.785**	**0.743**
HAR (RF)	1.050	0.964	1.054	1.024	1.081	1.086
HARCJ (RF)	0.895	1.009	0.971	0.976	1.021	1.035
LHARCJ (RF)	**0.838**	1.079	0.950	0.940	0.979	1.023
SHAR (RF)	1.031	0.984	1.054	1.016	1.088	1.075

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.12

MAE-LOG RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.997	1.003	0.996	0.997	1.010	1.004
LHARCJ (OLS)	0.970	0.979	0.985	0.972	0.985	1.005
SHAR (OLS)	0.991	0.994	0.995	1.001	1.012	1.006
HAR (WLS)	1.053	0.994	0.999	0.999	0.997	0.987
HARCJ (WLS)	1.111	1.002	0.997	0.997	1.001	0.984
LHARCJ (WLS)	1.017	0.978	0.986	0.972	0.969	0.969
SHAR (WLS)	1.035	0.988	0.994	1.001	1.010	1.005
NEWS IV (RF)	**0.804**	**0.767**	**0.728**	**0.772**	**0.745**	**0.757**
HAR (RF)	1.080	0.966	1.012	1.011	1.038	1.012
HARCJ (RF)	0.964	0.968	0.966	0.968	1.002	1.003
LHARCJ (RF)	0.898	0.938	0.931	0.937	0.968	0.973
SHAR (RF)	1.064	0.994	1.032	0.994	1.035	0.997

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.13

MAE-SD RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.979	1.020	0.999	1.000	1.009	1.030
LHARCJ (OLS)	0.944	1.006	0.996	0.973	0.973	1.062
SHAR (OLS)	0.973	0.984	0.996	0.993	1.007	0.994
HAR (WLS)	0.969	0.992	1.003	0.998	0.998	1.006
HARCJ (WLS)	0.958	1.007	1.002	0.999	1.002	1.027
LHARCJ (WLS)	0.919	0.988	0.996	0.971	0.968	1.038
SHAR (WLS)	0.946	0.981	0.998	0.993	1.006	1.005
NEWS IV (RF)	**0.854**	**0.762**	**0.737**	**0.786**	**0.750**	**0.763**
HAR (RF)	1.048	0.961	1.040	1.024	1.040	1.057
HARCJ (RF)	0.929	0.982	0.976	0.983	1.000	1.047
LHARCJ (RF)	**0.862**	0.972	0.947	0.949	0.965	1.027
SHAR (RF)	1.034	0.989	1.053	1.013	1.042	1.035

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.14

MAPE RESULTS FOR H=5

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.835**	1.028	0.996	1.007	1.061	0.945
LHARCJ (OLS)	1.098	1.147	1.040	1.015	1.102	0.999
SHAR (OLS)	0.980	0.984	1.005	0.992	1.005	1.013
HAR (WLS)	0.912	0.973	0.984	0.957	1.013	0.838
HARCJ (WLS)	**0.866**	0.988	0.978	0.966	1.029	0.894
LHARCJ (WLS)	0.883	1.001	0.989	0.940	0.999	0.902
SHAR (WLS)	0.884	0.958	0.986	0.947	1.014	0.846
NEWS IV (RF)	**1.043**	**0.777**	**0.733**	**0.799**	**0.811**	**0.670**
HAR (RF)	0.961	0.935	1.025	1.004	1.074	0.876
HARCJ (RF)	**0.840**	0.985	0.954	0.970	1.043	0.867
LHARCJ (RF)	**0.792**	1.003	0.941	0.936	0.995	0.838
SHAR (RF)	0.950	0.955	1.033	0.997	1.092	0.859

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

B.3. *22 day forecast horizon*

TABLE B.15

MSE-LOG RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.983	0.998	0.971	0.945	1.025	1.019
LHARCJ (OLS)	0.976	0.970	0.948	0.889	0.951	1.070
SHAR (OLS)	0.987	0.990	1.001	0.996	1.007	1.009
HAR (WLS)	1.272	0.983	0.991	0.985	0.962	0.977
HARCJ (WLS)	1.357	0.990	0.970	0.937	0.988	0.980
LHARCJ (WLS)	1.995	0.967	0.953	0.879	0.921	1.032
SHAR (WLS)	1.293	0.977	0.992	0.983	0.970	0.980
NEWS IV (RF)	**0.331**	**0.287**	**0.236**	**0.232**	**0.241**	**0.272**
HAR (RF)	1.239	1.159	1.026	1.040	1.165	1.074
HARCJ (RF)	0.840	1.019	0.749	0.902	1.052	1.131
LHARCJ (RF)	0.786	1.036	0.713	0.845	0.957	1.091
SHAR (RF)	1.228	1.154	1.004	1.037	1.160	1.026

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.16

MSE-SD RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.941**	1.031	0.979	0.978	1.072	1.153
LHARCJ (OLS)	**0.967**	1.079	0.963	0.919	1.002	1.259
SHAR (OLS)	**0.963**	0.980	1.004	0.995	1.003	0.987
HAR (WLS)	**0.993**	1.016	1.006	1.020	1.036	1.061
HARCJ (WLS)	**0.947**	1.039	0.986	0.994	1.087	1.172
LHARCJ (WLS)	**0.955**	1.080	0.982	0.927	1.006	1.260
SHAR (WLS)	**0.959**	0.998	1.011	1.014	1.035	1.051
NEWS IV (RF)	**0.500**	**0.298**	**0.238**	**0.249**	**0.254**	**0.278**
HAR (RF)	**1.595**	1.373	1.100	1.192	1.272	1.403
HARCJ (RF)	**0.901**	1.223	0.752	1.034	1.164	1.597
LHARCJ (RF)	**0.910**	1.391	0.754	0.968	1.032	1.526
SHAR (RF)	**1.554**	1.254	1.033	1.197	1.351	1.293

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.17

HMSE RESULTS FOR H=22

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.851**	1.036	0.906	1.007	1.184	1.020
LHARCJ (OLS)	1.112	1.483	0.961	1.062	1.191	0.992
SHAR (OLS)	0.987	0.988	1.019	1.006	1.032	1.014
HAR (WLS)	0.831	0.959	0.952	0.891	1.145	0.829
HARCJ (WLS)	**0.742**	0.994	0.869	0.894	1.232	0.958
LHARCJ (WLS)	**0.841**	1.269	0.932	0.873	1.153	0.944
SHAR (WLS)	**0.800**	0.931	0.969	0.888	1.166	0.844
NEWS IV (RF)	**0.743**	**0.417**	**0.225**	**0.330**	**0.416**	**0.238**
HAR (RF)	**3.398**	2.415	1.305	1.342	2.104	1.547
HARCJ (RF)	**1.067**	1.762	0.783	1.249	2.071	2.663
LHARCJ (RF)	**1.163**	2.205	0.880	1.121	1.772	1.886
SHAR (RF)	**2.853**	1.626	1.181	1.333	2.416	1.191

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.18

MAE RESULTS FOR H=22

	BOVA11	BBDC4	CSNA3	GGBR4	ITUB4	PETR4
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.914**	1.019	0.974	0.999	1.077	1.043
LHARCJ (OLS)	1.243	1.189	0.962	0.996	1.099	1.111
SHAR (OLS)	0.989	0.990	1.010	0.997	0.997	1.006
HAR (WLS)	0.964	0.972	0.989	0.962	1.045	0.947
HARCJ (WLS)	**0.901**	0.992	0.965	0.963	1.088	1.023
LHARCJ (WLS)	1.055	1.117	0.968	0.944	1.069	1.063
SHAR (WLS)	0.946	0.957	0.999	0.957	1.036	0.951
NEWS IV (RF)	**0.666**	**0.582**	**0.480**	**0.547**	**0.606**	**0.513**
HAR (RF)	1.213	1.193	1.002	1.130	1.216	1.145
HARCJ (RF)	**0.924**	1.126	0.817	1.079	1.118	1.311
LHARCJ (RF)	**0.952**	1.212	0.848	1.022	1.101	1.246
SHAR (RF)	1.206	1.140	0.967	1.132	1.222	1.074

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.19

MAE-LOG RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.994	1.014	0.988	0.973	1.012	1.012
LHARCJ (OLS)	0.993	1.015	0.966	0.944	0.980	1.036
SHAR (OLS)	0.991	0.991	1.003	0.998	1.007	1.002
HAR (WLS)	1.117	0.992	0.996	0.993	0.976	0.992
HARCJ (WLS)	1.144	1.011	0.987	0.968	0.990	0.994
LHARCJ (WLS)	1.338	1.013	0.968	0.940	0.961	1.020
SHAR (WLS)	1.138	0.985	0.998	0.991	0.984	0.989
NEWS IV (RF)	**0.541**	**0.563**	**0.474**	**0.496**	**0.514**	**0.517**
HAR (RF)	1.060	1.097	0.971	1.020	1.075	1.004
HARCJ (RF)	0.893	0.994	0.844	0.938	1.007	1.042
LHARCJ (RF)	0.873	0.991	0.821	0.910	0.978	1.024
SHAR (RF)	1.061	1.086	0.955	1.018	1.085	0.997

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.20

MAE-SD RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	0.954	1.038	0.992	0.983	1.032	1.069
LHARCJ (OLS)	0.994	1.072	0.972	0.954	1.006	1.102
SHAR (OLS)	0.988	0.987	1.007	0.994	1.005	0.994
HAR (WLS)	0.994	1.003	1.000	1.003	1.017	1.021
HARCJ (WLS)	0.960	1.036	0.992	0.986	1.041	1.074
LHARCJ (WLS)	0.982	1.066	0.977	0.958	1.011	1.102
SHAR (WLS)	0.982	0.991	1.007	0.998	1.021	1.018
NEWS IV (RF)	**0.595**	**0.577**	**0.476**	**0.511**	**0.540**	**0.515**
HAR (RF)	1.105	1.157	0.975	1.075	1.106	1.091
HARCJ (RF)	0.903	1.054	0.828	1.012	1.028	1.174
LHARCJ (RF)	0.896	1.097	0.826	0.971	1.011	1.145
SHAR (RF)	1.106	1.132	0.954	1.077	1.127	1.060

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.

TABLE B.21

MAPE RESULTS FOR H=22

	BOVA11	**BBDC4**	**CSNA3**	**GGBR4**	**ITUB4**	**PETR4**
HAR (OLS)	1.000	1.000	1.000	1.000	1.000	1.000
HARCJ (OLS)	**0.904**	1.019	0.961	1.001	1.077	1.000
LHARCJ (OLS)	1.093	1.161	0.957	0.988	1.088	0.978
SHAR (OLS)	0.997	0.994	1.012	0.998	1.000	1.006
HAR (WLS)	0.922	0.949	0.967	0.927	1.032	0.894
HARCJ (WLS)	**0.858**	0.974	0.932	0.929	1.081	0.937
LHARCJ (WLS)	0.958	1.087	0.939	0.908	1.063	0.942
SHAR (WLS)	0.911	0.938	0.978	0.923	1.028	0.902
NEWS IV (RF)	**0.638**	**0.596**	**0.467**	**0.540**	**0.638**	**0.464**
HAR (RF)	**1.137**	1.172	0.976	1.062	1.220	0.983
HARCJ (RF)	**0.817**	1.095	0.799	1.029	1.152	1.224
LHARCJ (RF)	**0.827**	1.167	0.829	0.979	1.131	1.113
SHAR (RF)	1.132	1.108	0.950	1.068	1.237	0.905

Note: This table shows the loss ratios of various models compared to the HAR estimated through OLS. Bold numbers indicate models that are part of the MCS at a 90% confidence level.