Energy Economics 111 (2022) 106059

Contents lists available at ScienceDirect

Energy Economics

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

Forecasting crude oil volatility with exogenous predictors: As good as it GETS?

Jean-Baptiste Bonnier *

CRESE, Université Bourgogne Franche-Comté, France
LEMNA, Université de Nantes, France

ARTICLE INFO

Keywords:
Crude oil
Volatility forecasting
General-to-specific
Dynamic model averaging

ABSTRACT

This paper aims to investigate the usefulness of exogenous predictors to forecast crude oil volatility. We use the recent expansion of the general-to-specific (GETS) procedure to conditionally heteroskedastic models to estimate a parsimonious predictive model of crude oil volatility from a large set of predictors. Our results show that the GETS algorithm achieves good predictive accuracy compared to its competitors at the 1-day horizon. However, this accuracy deteriorates for more distant forecast horizons. We argue that it may be due to the fact that the GETS procedure is based on tests that are key in assessing explanatory power as opposed to reducing expected prediction error. Among its competitors, DMA achieves good predictive power in almost all situations. Still, our analysis provides interesting insights on the variables best suited to forecast crude oil volatility. In particular, forecasters might benefit from better exploiting the predictive content of exchange rates.

1. Introduction

Crude oil is a central input in global economic activity, and high oil price uncertainty is known to have depressing effects on investment and macroeconomic aggregates (Elder and Serletis, 2010). Volatility is also of crucial importance for risk management and hedging. Policy-makers, producers, consumers, and financial-market participants are therefore all interested in crude oil volatility forecasts. Forecasting is, however, a complicated task, not least because of the numerous factors that are thought to play a role in the determination of crude oil prices and volatility.

The literature on crude oil volatility forecasting is particularly abundant in recent years. Two main strands stand out. The first produces forecasts using GARCH-type models (Sadorsky, 2006; Sadorsky and McKenzie, 2008; Kang and Yoon, 2013; Charles and Darné, 2017; Pan et al., 2017; Wei et al., 2017; Herrera et al., 2018; Lin et al., 2020), while the other builds on the seminal work of Andersen et al. (2001) and relies on intraday data in realized volatility (RV) models (Sévi, 2014; Prokopczuk et al., 2016; Gong and Lin, 2018; Lv, 2018; Luo et al., 2020). These two strands rely heavily on historical volatility and prices, but only a limited subset of papers consider exogenous variables related to macroeconomic information, fundamentals in the demand and supply of crude oil, speculative activity, or sentiment (Conrad et al., 2014; Pan et al., 2017; Ma et al., 2018; Meng and Liu, 2019). As a consequence, despite important benefits to using exogenous variables for forecasting crude oil volatility, there is no fixed or commonly accepted set of predictors.

In this paper, we propose to reassess the usefulness of exogenous variables to forecast crude oil volatility with GARCH-type models. We consider different approaches with a particular emphasis on a parsimonious AR-X-log-GARCH-X model resulting from Sucarrat and Escribano's (2012) general-to-specific (GETS) procedure. GETS modelling is a procedure that combines many elements from the model selection literature (backwards elimination, tests on coefficients, diagnostic tests, and fit measures) to determine a parsimonious model from a general unrestricted model (GUM). Up until now, because of the important computational difficulties associated with the estimation of models with heteroskedastic errors, GETS modelling was restricted to modelling the mean of homoskedastic processes (Granger and Timmermann, 1999; McAleer, 2005). However, recent renewed interest in log-GARCH models provided a tractable equation-by-equation estimation framework (Sucarrat et al., 2016; Francq and Sucarrat, 2017) that allows to overcome many of the estimation issues that arise with log-GARCH models. In particular, the mean and volatility specifications can be consistently estimated step-by-step by least squares using the ARMA representation of the volatility specification. Sucarrat and Escribano (2012) built on these developments to extend GETS modelling to AR-X models with log-GARCH-X innovations, and developed associated

* Correspondence to: UFR SJEPG, 45D Av. de l'Observatoire, 25030 Besançon, France.
E-mail address: Jean_Baptiste.Bonnier@univ-fcomte.fr.

https://doi.org/10.1016/j.eneco.2022.106059
Received 2 May 2021; Received in revised form 25 April 2022; Accepted 30 April 2022
Available online 16 May 2022

Table 1
Selected literature review.

Article	Dependent variables	Exogenous variables	Sample	Models	Combination strategies	Volatility proxies	Horizons	OOS loss functions	Tests
Manera et al. (2016)	WTI (futures)	WT, MS, NL	Jan. 2000 Apr. 2014 (W)	GARCH-X, GARCH-M-X, TARCH-X, EGARCH-X					
Pan et al. (2017)	WTI, Brent (spot)	PROD, KIL	Jan. 1986 Dec. 2015 (D)	RS-GARCH-MIDAS, GARCH-MIDAS	FC	Squared returns	1 day	MSPE, QLIKE	DM
Wei et al. (2017)	WTI (spot)	PROD, KIL, SI, EPUs	Jan. 1996 Apr. 2016 (D)	GARCH-MIDAS	DMA	Squared returns	1 day	MSPE, MAFE	MCS
Bahloul and Gupta (2018)	WTI, Brent (futures)	SSI, SUI	Jan. 1991 May 2016 (D)	AR-X-GARCH-X					
Gong and Lin (2018)	WTI (futures)	OVX	May 2007 June 2016 (D)	HAR-RV models		5-min RV	1,5, 22 days	MAFE, HMAE	DM, MZ MCS
Lv (2018)	WTI (futures)	OVX	May 2007 Sep. 2015 (D)	HAR-RV models		5-min RV	1 day	MSPE, QLIKE	MCS,
Ma et al. (2018)	WTI (spot)	4 UMS, 25 MF, 16 TI	Jan. 1986 Dec. 2015 (M)	(Log-)RV regressions	FC (M), DMA, LASSO	RV	1 month	R^2_{OOS}	CW, CER
Meng and Liu (2019)	WTI, Brent (spot)	SI, 5 MF, 4 FP	Jan. 1990 Dec. 2015 (M)	(Log-)RV regressions	FC (M, TM, DMSPE)	RV	1 month	R^2_{OOS}, lnS	CW, SR, HLN, W
Yang et al. (2019)	WTI (futures)	FEARS, Leverage	July 2004 Dec. 2011 (D)	HAR-RV models		5-min RV	1, 5, 22 days	MAE, MSPE	DM, MZ
Zhang et al. (2019)	WTI (futures)		Jan. 2007 July 2016 (D)	HAR-RV models	FC LASSO, EN	5-min RV	1 day	MAE, MSPE, QLIKE, DoC	MCS, PM
Luo et al. (2020)	WTI, Brent (futures)	RV S&P 500, RV Brent/WTI	Jan. 2013 May 2018 (D)	IHM-HAR, IHMC-HAR		5-min RV	1, 5, 22 days	MSPE, QLIKE, P	MCS, EV

Notes: CER = Certainty Equivalent Return. CW = Clark–West test. DM = Diebold–Mariano test. DoC = Direction of Change. EN = Elastic Net. EPUs = Economic Policy Uncertainty indexes. EV = Economic Value. FC = Forecast Combinations. FP = Futures Prices. KIL = Kilian's activity index. HLN = Harvey–Leybourne–Newbold test. lnS = Expected logarithmic Score. MCS = Model Confidence Set. MF = Macroeconomic Fundamentals. MS = Market Share of non-commercial traders. MZ = Mincer–Zarnowitz regression. NL = Net Long positions of non-commercial traders on total open interest. P = Blair–Poon–Taylor P statistic. PM = Pesaran–Timmermann test. PROD = Global oil production. SI = Speculation Index. SR = Success Ratio. SSI/SUI = Scotti Surprise/Uncertainty Index. TI = Technical Indicators. UMS = Uncertainty and Market Sentiment variables. W = West test. WT = Working T index.

algorithms that automate multi-path GETS modelling of both the mean and the volatility.

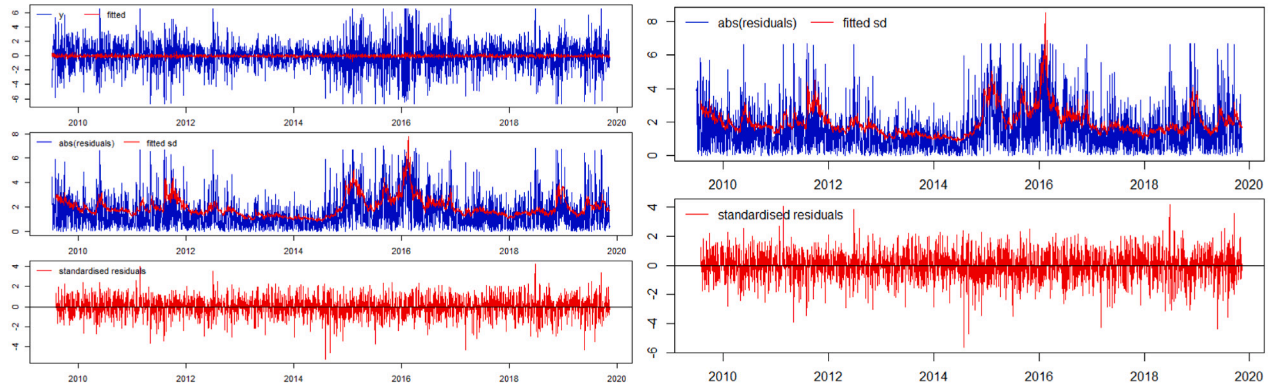
We first conduct an in-sample (IS) analysis in which we focus on the selected variables of the GETS terminal model. Then, in an out-of-sample (OOS) analysis, we compare the predictive power of this multivariate approach to different models and forecast combinations for several horizons (1 day, 5 days, 21 days, and 63 days). We consider standard models (GARCH(1,1), GJR-GARCH(1,1), EGARCH(1,1), log-GARCH(1,1), and log-GARCH(1,1)-X) and simple combination methods (mean, trimmed mean, and discounted mean squared prediction error), as well as more advanced techniques representing the state of the art. Specifically, we estimate individual GARCH-MIDAS models (Engle et al., 2013) that proved successful in exploiting low-frequency macroeconomic and sentiment information to forecast daily crude oil volatility (Pan et al., 2017; Wei et al., 2017), and we use the dynamic model averaging (DMA) technique proposed by Raftery et al. (2010) as a time-varying combination method (Liu et al., 2017; Wei et al., 2017). We evaluate the performance of candidate models with both Harvey et al.'s (1997) modified version of Diebold and Mariano's (1995) test of equal predictive ability, and Hansen et al.'s (2011) model confidence set (MCS) test that allows to determine a subset of best models from a large pool of candidates. In accordance with Patton (2011), these tests are reported under two loss functions that are the most robust to imperfect volatility proxies: the mean square predictive error (MSPE) and the Gaussian quasi-likelihood (QLIKE).

This paper is directly related to the literatures on the determinants of crude oil prices and volatility (Kilian, 2009; Kilian and Murphy, 2014; Baumeister and Kilian, 2016), and on the predictive ability of exogenous variables to improve crude oil volatility forecasts (Pan et al., 2017; Wei et al., 2017; Meng and Liu, 2019; Yang et al., 2019). At the most general level, oil price fluctuations can result from three types of shocks: (i) supply shocks to global crude oil production arising from political events in oil-producing countries, improvements in the technology of extracting crude oil, and the discovery of new fields; (ii) shocks to the demand for crude oil associated with unexpected changes in global economic activity; and (iii) shocks to the demand for above-ground crude oil inventories, reflecting shifts in expectations about future shortfalls (Kilian, 2009; Baumeister and Kilian, 2016). However, an abundant number of variables can prove helpful to capture changes in these supply and demand determinants. As an example, while production and inventories directly reflect current supply, political uncertainty can also hold information on future supply dynamics (Aloui et al., 2016; Wei et al., 2017). Similarly, several proxies of global aggregate demand for crude oil exist, but a demand component can also be connected to speculative activity and, hence, be captured by vastly different indicators (Wei et al., 2017; Meng and Liu, 2019). In this paper, we consider an important number of potential predictors that are meant to express these diverse elements (see Section 4).

This paper also contributes to the debate on single-predictor versus multiple-predictor models for forecasting. Do we produce more accurate forecasts by combining variables in a single model or by combining forecasts of several single-predictor models? Both approaches have their own advantages and drawbacks. Individual-predictor models can be underspecified and lead to poor out-of-sample performance, but forecast combinations help reduce the volatility of previsions (Timmermann, 2006). On the opposite, multiple-predictor models can lead to over-fitting and multicollinearity issues, but the GETS procedure is meant to solve these issues and produce a parsimonious model.[1]

Our results show that the predictive accuracy of the GETS algorithm at the 1-day horizon compares relatively well to standard benchmarks, single-variable models (including single-predictor GARCH-MIDAS models), and simple forecast combinations that are widespread in the literature. However, its accuracy deteriorates as the forecast horizon becomes more distant. This is partly due to the fact that models based on exogenous predictors require to produce forecasts of the regressors to be used as inputs, so that uncertainty accrues with each new step and each regressor in the model. Still, by comparison, DMA fares very well at all horizons and tends to outperform the GETS procedure when the pool of models is sufficiently large. It suggests that another

[1] Another multiple-predictor approach applies shrinkage methods, such as Tibshirani's (1996) lasso or Zou and Hastie's (2005) elastic net, to RV models (Zhang et al., 2019). We do not pursue this path here. Future research could investigate the respective predictive ability of these two multiple-predictor methodologies.



(a) GETS1 (b) GETS2

Fig. 1. Representation of fitted GETS models for crude oil volatility.

explanation may be at play. We argue that the statistical tools used in the GETS procedure might not be the most adequate to achieve the best forecasting model. Nevertheless, adding the GETS' terminal model as an input to the DMA results in further significant improvements of OOS predictions. It implies that forecast combinations need not be limited to simple models, and that we could benefit from applying time-varying combination methods to state-of-the-art models.

The flexible framework adopted in this paper also allows us to participate in the debate on the determinants of crude oil volatility. In line with other empirical papers, we find the OVX, a measure of implied volatility in crude oil futures markets based on options of the U.S. oil fund LP, to stand out as a preferred predictor. More surprisingly, out of our 17 other exogenous variables, several exchange rates remain in the terminal model when fundamentals about demand and supply of crude oil, macroeconomic factors, and speculation indicators, are all discarded. It reveals that the forward-looking nature of exchange rates highlighted by Chen et al. (2010) may not only be helpful to forecast commodity returns but also to forecast their volatility. It advocates for a larger exploration of their predictive ability in future work, as well as for a more widespread utilization in forecasting models.

Section 2 presents a selected literature review. Sections 3 and 4 show our methodology and data, respectively. We display our results in Section 5, and Section 6 concludes.

2. Literature review

A number of contributions have considered the usefulness of exogenous variables to forecast crude oil volatility (see Table 1). Yet, despite extensive work, no variable stands out as absolutely imperative in crude oil volatility models. First, several studies document the link between crude oil supply and demand fundamentals and its volatility (Nomikos and Pouliasis, 2011; Pan et al., 2017; Wei et al., 2017; Ma et al., 2018; Meng and Liu, 2019). According to the theory of storage, the basis is composed of a convenience yield net of storage costs and an opportunity cost of foregone interest (Fama and French, 1987; Gospodinov and Ng, 2013). It thus often serves as a measure of scarcity. Nomikos and Pouliasis (2011) show in a regime-switching context that the squared basis is a strongly significant predictor of crude oil volatility in a low-regime setting. They justify the difference in predictability between low- and high-volatility regimes by the fact that, in a high-volatility state, volatility fluctuations are mainly the consequences of short-lived random shocks that are very difficult to foresee. They also find the basis to be useful to forecast crude oil volatility out-of-sample (OOS).

Pan et al. (2017) consider the usefulness of the level and volatility of global crude oil production and demand to improve volatility forecasts of a RS-GARCH-MIDAS model. Even though including fundamentals does not always improve OOS performance, their model with the level of aggregate demand often does. Moreover, a forecast combination of single-predictor RS-GARCH-MIDAS models significantly outperforms their benchmark, thus highlighting the importance of fundamentals to forecast crude oil volatility. In the same vein, Meng and Liu (2019) show that forecast combinations of individual realized-volatility (RV) regressions including the volatility of supply and demand factors, and other macroeconomic and futures market indicators, could provide significantly more accurate forecasts than an autoregressive benchmark.

In a similar form of RV regressions, but with a more data-intensive approach, Ma et al. (2018) consider the predictive power of 4 uncertainty and market sentiment indexes, 25 macroeconomic variables, and 16 technical indicators. They find evidence that several variables in each category could prove helpful to forecast crude oil volatility, although evidence is thinner for technical indicators. They also examine three different methods to combine information from multiple variables: a mean forecast combination, Raftery et al.'s (2010) dynamic model averaging (DMA) approach as a time-varying combination strategy, and Tibshirani's (1996) lasso regression. While the mean forecast combination and the lasso usually outperform their benchmark, they find the DMA to generally produce the most accurate predictions. It does not mean, however, that there is a consensus around forecast combinations approaches. By contrast, Zhang et al. (2019) achieve better OOS performance with shrinkage methods (the lasso and Zou and Hastie's (2005) elastic net) than with forecast combinations (mean, median, trimmed mean, DMSPE) in a HAR-RV framework.

Ma et al. (2018) is but one of the recent papers that is interested in the predictive power of uncertainty and sentiment variables. Indeed, an important number of new uncertainty and sentiment measures have been developed in recent years (Baker and Wurgler, 2007; Hakkio and Keeton, 2009; Da et al., 2015; Jurado et al., 2015; Baker et al., 2016), and forecasters have logically considered their predictive ability. Wei et al. (2017), for instance, find in an OOS analysis that global and country-specific economic policy uncertainty (EPU) indicators compare favourably to several macroeconomic variables to forecast crude oil volatility. In the same direction, Bahloul and Gupta (2018) show with GARCH-X-type models that contemporaneous U.S. macroeconomic surprise and economic uncertainty in the U.S. and other economic areas (the U.K., the Euro Area, Japan, and Canada) can help explain movements in crude oil volatility. Yang et al. (2019) demonstrate how investor sentiment, as measured by Da et al.'s (2015) FEARS index, and leverage could help improve the predictive performance of several HAR-RV models. Relatedly, the OVX, also referred to as the investor fear gauge, has been fund to contain useful incremental information content for forecasting the volatility of crude oil futures in various HAR-RV models (Gong and Lin, 2018; Lv, 2018).

In addition to uncertainty and sentiment indices, and to supply and demand data for the physical commodity, there are important

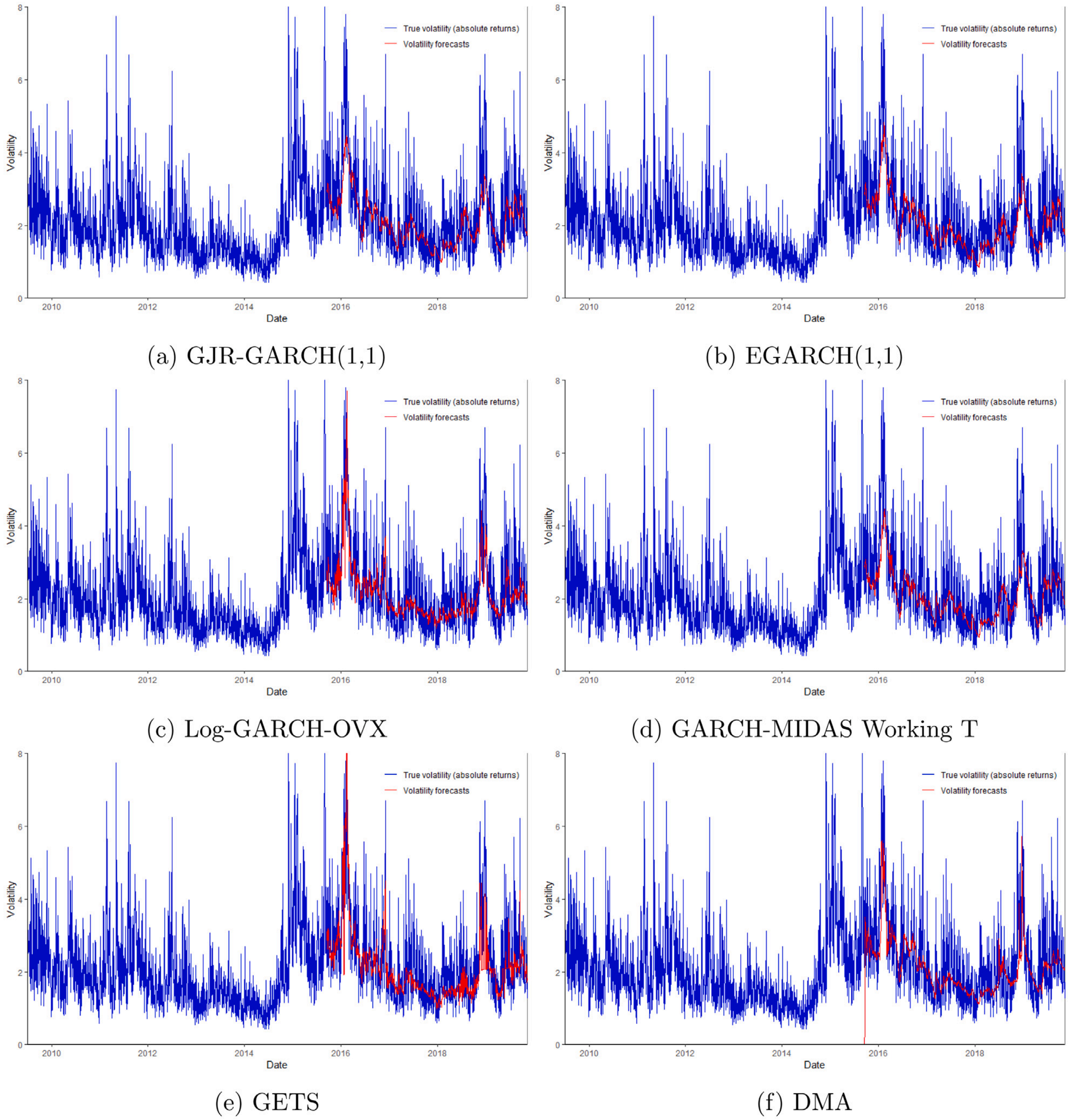
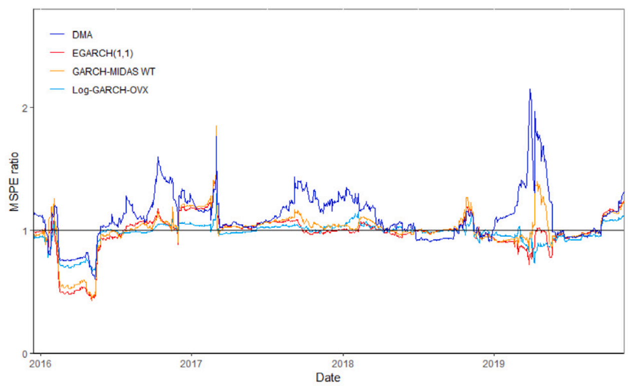


Fig. 2. Volatility forecasts of selected models.

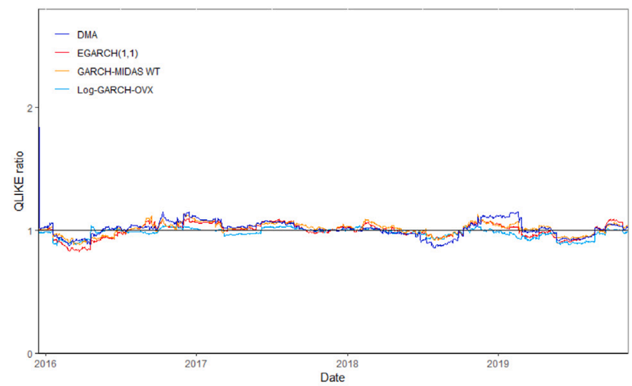
suspicions that speculation in futures markets may affect spot prices. This debate re-emerged in recent years because of the large inflows of capital from index traders starting in the first half of the 2000s (Cheng and Xiong, 2014; Sockin and Xiong, 2015). Measures of speculative activity could then hold fruitful information content to forecast crude oil volatility. Manera et al. (2016) confirm this intuition. They find three indexes of speculative activity to be contemporaneously negatively related to crude oil volatility in an AR-X-GARCH-X framework.

This financialization phenomenon fuelled a large literature. One of its most fruitful path has been concerned with how financialization affected the dependence structure in-between commodities and between commodities and other asset classes (Tang and Xiong, 2012; Delatte and Lopez, 2013; Ma et al., 2021). It is therefore possible that information from other markets could prove helpful to forecast crude oil volatility. Luo et al. (2020), in particular, find that realized volatility of the S&P 500 improves the performance of IHM-HAR models to forecast WTI futures volatility at short horizons.

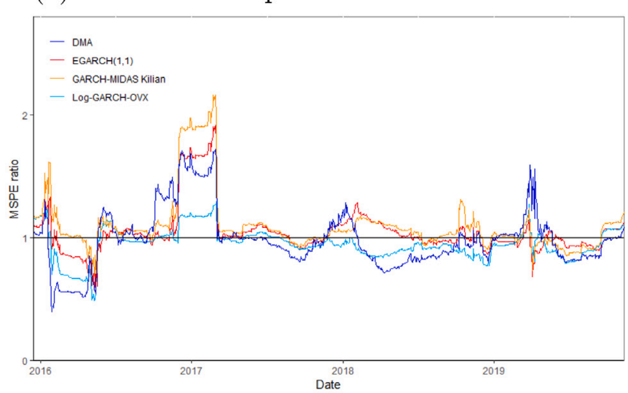
Finally, there is also important empirical evidence that the forward-looking nature of exchange rates makes them useful to forecast commodity prices (Chen et al., 2010; Groen and Pesenti, 2011) and that volatility spillovers exist between these markets (Wang and Wu, 2012). Global commodity price fluctuations substantially affect exports of commodity-exporting countries. As such, they constitute important terms of trades shocks to their currencies. Market participants anticipate commodity price shocks, so that they will be priced into current exchange rates through their anticipated impact on future export income and exchange rate values, while, by contrast, commodity prices will take longer to adjust as supply and demand for commodities are more inelastic (Chen et al., 2010).



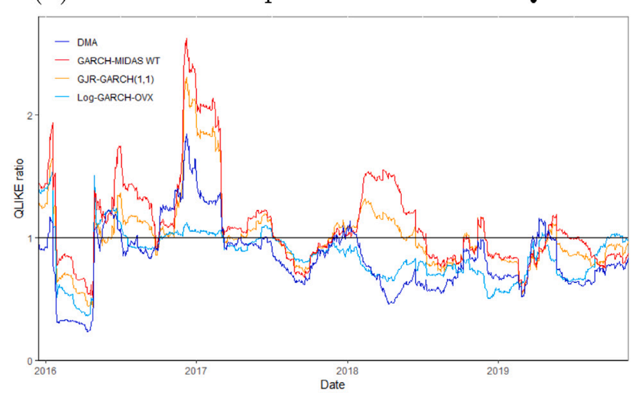
(a) Benchmark: squared return - LF: MSPE



(b) Benchmark: squared return - LF: QLIKE



(c) Benchmark: intraday range - LF: MSPE



(d) Benchmark: intraday range - LF: QLIKE

Fig. 3. Rolling window loss function ratio relative to the GETS.

3. Methodology

3.1. GARCH-type models

3.1.1. The GARCH model

The GARCH model proposed by Bollerslev (1986) is widely popular as it allows to model the clustering of large returns that are typical of financial time series.[2] A time series process, ϵ_t, described by a GARCH(1,1) model with mean zero takes the form:

$$\epsilon_t = \sigma_t z_t, \quad z_t \sim IID(0,1), \qquad (1)$$
$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2.$$

The non-negativity conditions are $\omega > 0$, $\alpha \geq 0$, and $\beta \geq 0$, and the stationarity of the process is guaranteed by the restriction $\alpha + \beta < 1$.

As noted by Nelson (1991), GARCH models have a number of important limitations. First, standard GARCH models do not allow for an asymmetric response to positive and negative returns, and we know since Black (1976) that financial returns tend to be negatively correlated with changes in volatility. The debate on asymmetric volatility is particularly vivid for commodities. Many papers on the topic for energy commodities find conflicting results, but, overall, evidence suggests that practitioners should consider models that cater for asymmetry (Kristoufek, 2014; Carnero and Pérez, 2019). Second, GARCH models face non-negativity constraints on the parameters that limit their flexibility and rule out the possibility of oscillatory fluctuations.

[2] See Bauwens et al. (2012) for a comprehensive survey of volatility models.

3.1.2. The GJR-GARCH model

The GJR-GARCH model of Glosten et al. (1993) respond to the first critic as it allows to capture asymmetric effects. The specification of the conditional variance for a GJR-GARCH(1,1) process is:

$$\sigma_t^2 = \omega + (\alpha + \gamma I_{t-1})\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \qquad (2)$$

where I_t is the indicator function that takes the value 1 if $\epsilon_t < 0$, and 0 otherwise. Volatility is positive if $\omega > 0$, and $\alpha, \gamma, \beta \geq 0$. The stationarity condition is satisfied for $\alpha+\beta+\gamma/2 < 1$. A positive γ means that negative shocks have a stronger effect on the volatility than positive shocks.

3.1.3. The EGARCH model

An entire subclass of GARCH models respond to the second critic. Exponential GARCH models specify the conditional variance in the form of the natural logarithm, and, hence, their fitted values are guaranteed to be non-negative in practice.[3] This specification enables richer autoregressive volatility dynamics.[4] Nelson's (1991) EGARCH is probably the most popular exponential GARCH model. The variance specification of an EGARCH(1,1) model is:

$$\ln(\sigma_t^2) = \omega + g(z_{t-1}) + \beta\ln(\sigma_{t-1}^2), \qquad (3)$$

where $g(z_t) = \theta z_t + \phi[|z_t| - E|z_t|]$. θz_t represents a sign effect. If θ is negative, a negative (positive) shock z_{t-1} will have a positive

[3] Exponential GARCH models most notably include log-GARCH models (Pantula, 1986; Geweke, 1986), Nelson's (1991) EGARCH model, and Harvey's (2013) Beta-t-EGARCH model.

[4] In more standard GARCH models, the more conditioning variables are added to the volatility equation, the more restrictions are needed to ensure positivity.

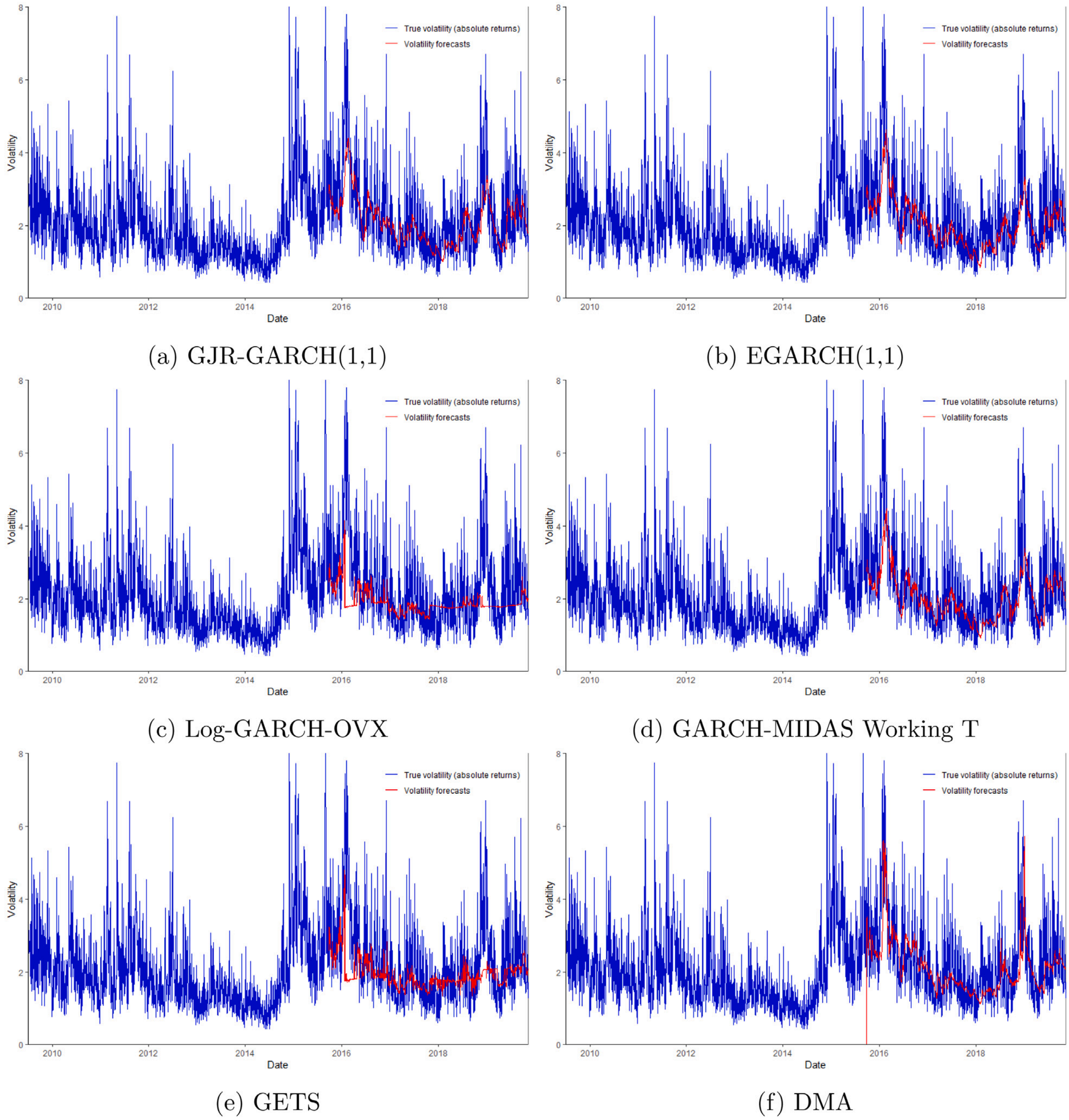


Fig. 4. 5-day volatility forecasts of selected models.

(negative) effect on the conditional variance. $\phi[|z_t| - E|z_t|]$ represents a magnitude effect. If ϕ is positive, a shock z_{t-1} will have a positive (negative) effect on the conditional variance if its absolute value is above (below) the mathematical expectation of its absolute value.[5] $g(.)$ is therefore a flexible specification that allows past disturbances to have an asymmetric effect on the conditional variance. Since the EGARCH(1,1) model does not impose positivity constraints on the coefficients, a sufficient stationarity condition is $|\beta| < 1$.

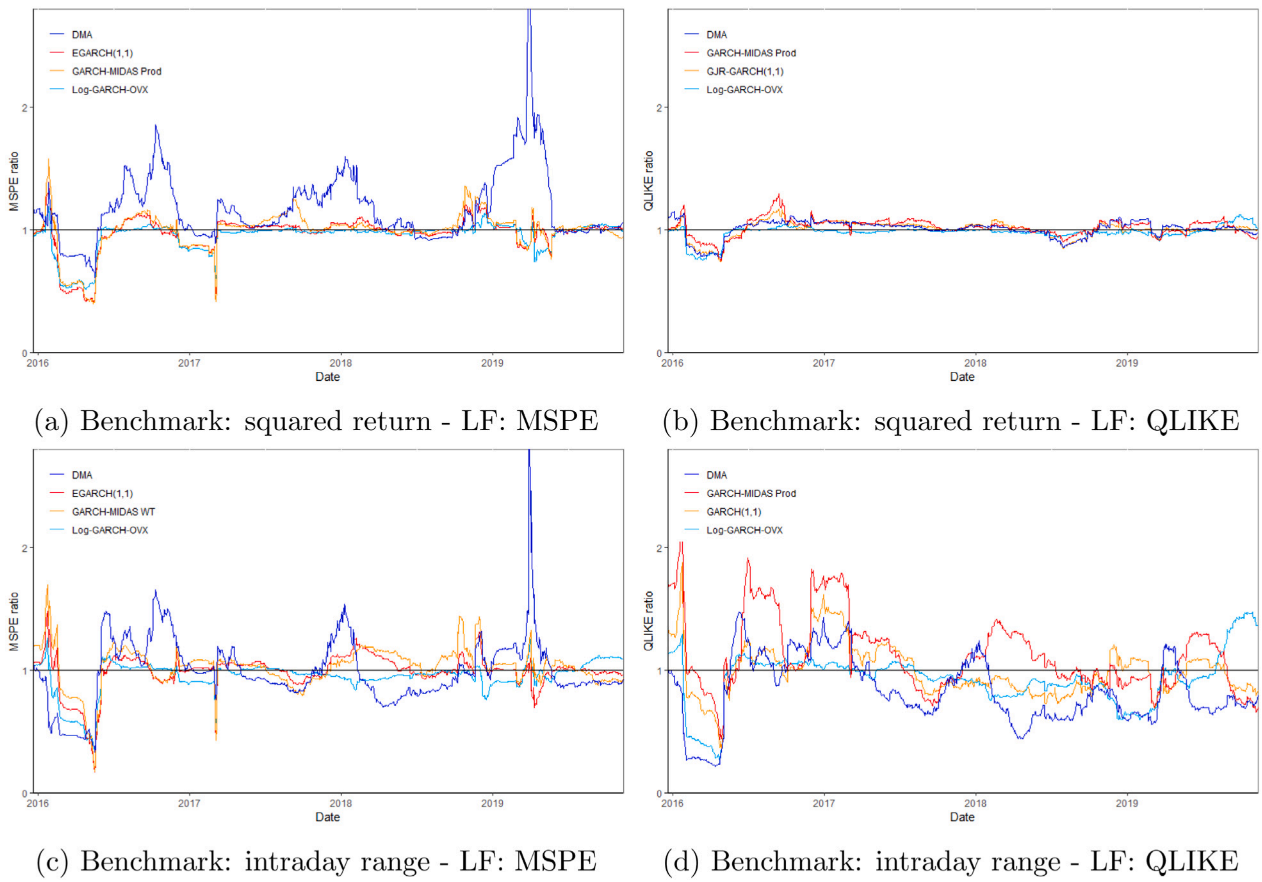
While the EGARCH representation brings interesting features to model the volatility, it also has its own shortcomings. First, the stability conditions are more restrictive than for GARCH models. It is notably not stable for Student t errors, which is the preferred fat-tailed distribution by practitioners. Second, there are very few theoretical results on consistent estimation and valid asymptotic inference for the EGARCH model. Only recently did Wintenberger (2013) prove the strong consistency and asymptotic normality of the quasi maximum likelihood estimator (QMLE) for the EGARCH(1,1) model under the additional condition of continuous invertibility.

3.1.4. The log-GARCH model

The log-GARCH model was independently introduced by Pantula (1986), Geweke (1986), and Milhøj (1987). A log-GARCH(P, Q) representation of the conditional variance of a time series takes the form[6]:

[5] $E|z_t|$ depends on the assumption made on the unconditional density of ϵ_t. For the Gaussian distribution, $E|z_t| = \sqrt{2/\pi}$.

[6] We present the form of the log-GARCH model with order P and Q because of its role in the GETS procedure (Section 3.2).



(a) Benchmark: squared return - LF: MSPE (b) Benchmark: squared return - LF: QLIKE

(c) Benchmark: intraday range - LF: MSPE (d) Benchmark: intraday range - LF: QLIKE

Fig. 5. 5-day rolling window loss function ratio relative to the GETS.

$$\ln(\sigma_t^2) = \omega + \sum_{p=1}^{P} \alpha_p \ln(\epsilon_{t-p}^2) + \sum_{q=1}^{Q} \beta_q \ln(\sigma_{t-q}^2). \tag{4}$$

$\text{Ln}(\sigma_t^2)$ is stable if $|E(\ln(z_t^2))| < \infty$, and if the modulus of all the roots of the lag polynomial $1 - \sum_{p=1}^{P^*}(\alpha_p + \beta_p)L^p$ are greater than 1, where $P^* = \max\{P, Q\}$ (Sucarrat et al., 2016). Moreover, if $\ln(\sigma_t^2)$ is stable, then σ_t^2 will generally be stable for distributions like the Generalized Error Distribution (GED) with shape parameter greater than 1, and the Student distribution with degrees of freedom greater than 2 (which is not the case of the EGARCH). Similarly to the GJR-GARCH, the specification of the log-GARCH model can be enriched with leverage effects to cater for asymmetry.

For a few years, renewed interest in log-GARCH models led to a stream of results that make them more tractable, and have more straightforward estimation properties than EGARCH models (Francq et al., 2013; Sucarrat et al., 2016; Francq and Sucarrat, 2017, 2018; Sucarrat and Escribano, 2018). In particular, Sucarrat et al. (2016) provide a result that enables consistent estimation and ordinary inference methods for log-GARCH models when the conditional density is unknown. To do so, they recourse to the moving average representation of the log-GARCH model. That is, if $|E(\ln(z_t^2))| < \infty$, then Eq. (4) admits the following ARMA(P, Q) representation:

$$\ln(\epsilon_t^2) = \phi_0 + \sum_{p=1}^{P} \phi_p \ln(\epsilon_{t-p}^2) + \sum_{q=1}^{Q} \theta_q u_{t-q} + u_t, \quad u_t = \ln(z_t^2) - E(\ln(z_t^2)), \tag{5}$$

where $\phi_0 = \omega + (1 - \sum_{q=1}^{Q} \beta_q) E(\ln(z_t^2))$, $\phi_p = \alpha_p + \beta_p$, and $\theta_p = -\beta_p$. It is then possible to use well-known ARMA estimation methods (see Brockwell and Davis, 2006) to obtain consistent and asymptotically normal estimates of all ARMA parameters.

Apart from the intercept, ω, log-GARCH parameters estimates can easily be deduced from the ARMA estimates. It is known that the estimation of ω is biased, and that this bias depends on the unknown density of the conditional standardized error. However, Sucarrat et al. (2016) derive a consistent estimate for $E(\ln(z_t^2))$ that takes the form: $-\ln[(1/T)\sum_{t=1}^{T} \exp(\hat{u}_t)]$. This new result means that all the coefficients of the log-GARCH model can be consistently estimated using the ARMA representation. Furthermore, Sucarrat et al. (2016) show that these results also hold for log-GARCH-X models, where the "X" denotes additional covariates. And, while there are many examples of papers that use covariates in GARCH models, few theoretical results on the consistency and asymptotic normality of models with covariates exist (Francq and Sucarrat, 2017). It makes Sucarrat et al.'s (2016) framework all the more interesting.

As a matter of fact, there is an important literature that supports the benefits of conditioning volatility forecasts on other variables than past returns. It is well-known however that including too many variables in a model may lead to over-fitting. It is therefore necessary to adopt a parsimonious framework that allows to carefully select relevant variables. Sucarrat and Escribano (2012) develop such a framework with the general-to-specific (GETS) procedure for the conditional volatility.

3.2. GETS modelling

General-to-specific (GETS) modelling is a procedure aimed at determining a parsimonious model from a general unrestricted model (GUM) using a series of tests to eliminate irrelevant variables and ensure that the model remains well specified.[7] So far, the important number of

[7] Important milestones in automated multi-path GETS search algorithms include Hoover and Perez (1999), the PcGets of Hendry and Krolzig (1999,

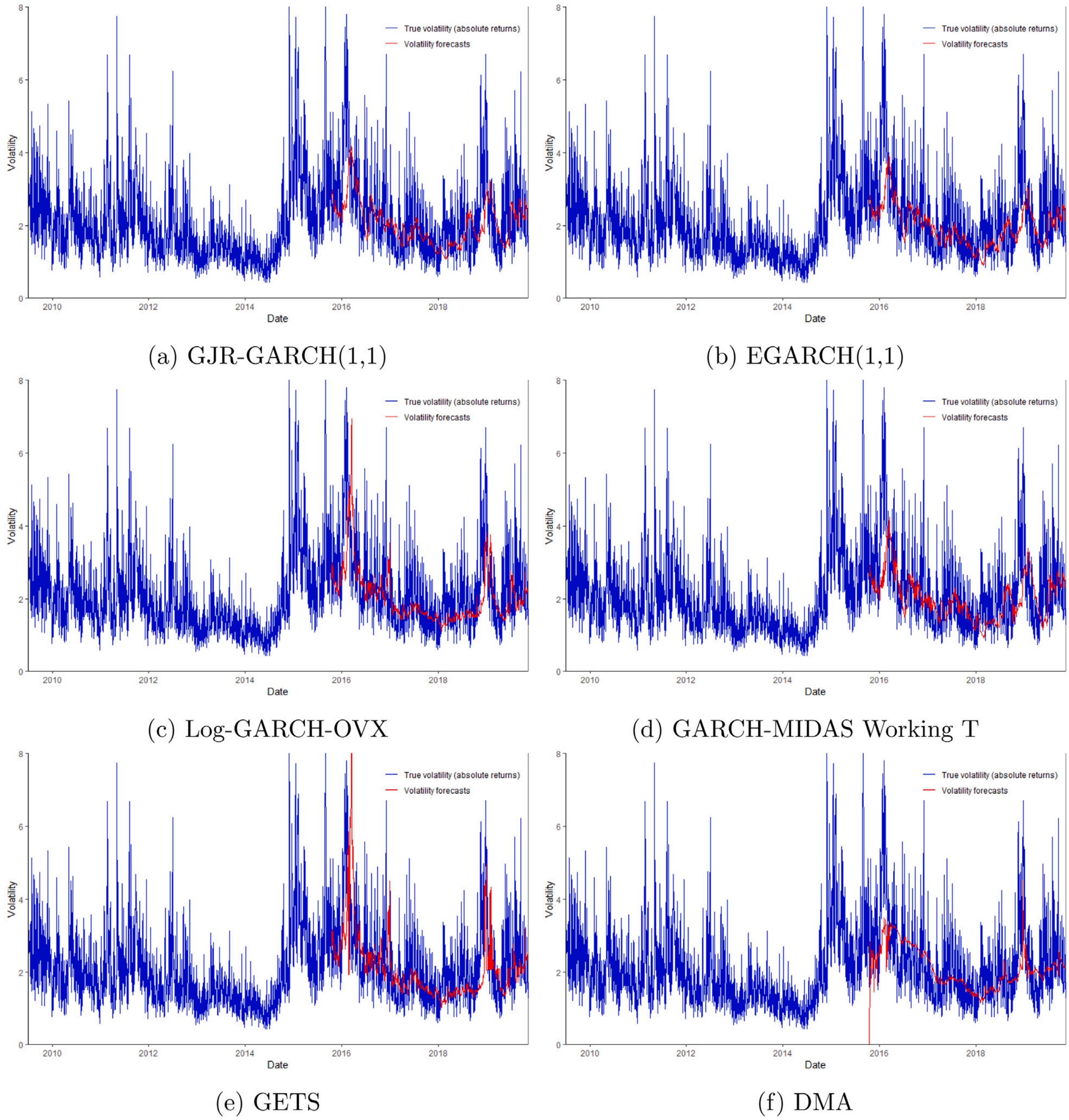


Fig. 6. 21-day volatility forecasts of selected models.

parameters and the computational difficulties usually associated with the estimation of conditional volatility models were considered to be important obstacles to the application of GETS modelling to financial times series (Granger and Timmermann, 1999; McAleer, 2005). However, the equation-by-equation estimation framework developed in Sucarrat et al. (2016) allows to overcome many of the inference and estimation issues – such as numerical approximation, multiple optima, convergence issues, negative variance, initial values, parameter constraints, or finite sample approximations – that arise with log-GARCH models. In particular, the mean and volatility specifications of a time series process can be consistently estimated step-by-step by ordinary least squares (OLS) using the ARMA representation of the volatility specification. Sucarrat and Escribano (2012) rely on this result to expand GETS modelling to conditional volatility models.

The initial specification of the GUM in Sucarrat and Escribano's (2012) GETS procedure (GETS, henceforth) must be contained in an AR-X model with log-ARCH-X errors:

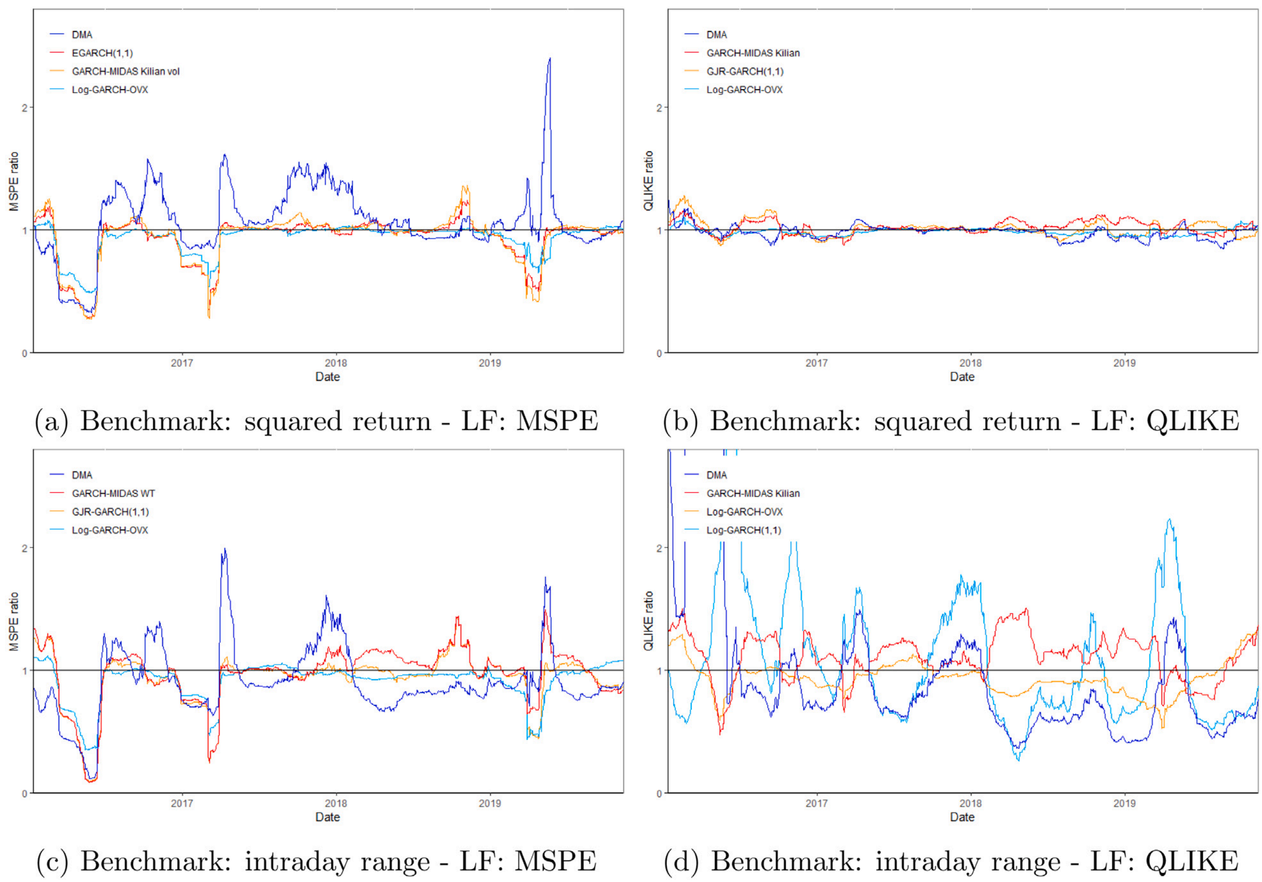
$$y_t = \lambda_0 + \sum_{r=1}^{R} \lambda_r y_{t-r} + \sum_{s=1}^{S} \delta_s x_s^m + \epsilon_t, \tag{6}$$

$$\epsilon_t = z_t \sigma_t, \quad z_t \sim IID(0,1),$$

$$\ln(\sigma_t^2) = \omega + \sum_{p=1}^{P} \alpha_p \ln(\epsilon_{t-p}^2) + \sum_{l=1}^{L} \gamma_l \ln(\epsilon_{t-l}^2) \mathbb{1}_{\{\epsilon_{t-l}<0\}}$$

$$+ \sum_{q \in Q} \beta_q \ln(\text{EqWMA}_{q,t-1}) + \sum_{k=1}^{K} \eta_k x_k^v. \tag{7}$$

2005), and the Autometrics algorithm of Doornik and Hendry (2013) and Doornik (2009).



(a) Benchmark: squared return - LF: MSPE (b) Benchmark: squared return - LF: QLIKE

(c) Benchmark: intraday range - LF: MSPE (d) Benchmark: intraday range - LF: QLIKE

Fig. 7. 21-day rolling window loss function ratio relative to the GETS.

Eq. (6) is the conditional mean equation that takes the form of an AR-X, an autoregressive model of order R with S exogenous variables. Eq. (7) is the conditional volatility equation that takes the form of a log-ARCH-X model of order P with L logarithmic asymmetry terms similar to the GJR-GARCH, realized volatility proxies defined as $\text{EqWMA}_{q,t-1} = (\epsilon^2_{t-1} + \cdots + \epsilon^2_{t-q})/q$, and K covariates.[8] The covariates in the mean and volatility equations need not be the same, and they may contain lagged variables. In this paper, we take $R, L = 5$, $P = 10$, $Q = \{5, 20, 60\}$ (which corresponds to volatility proxies in the last week, last month, and last quarter, respectively), and our lagged exogenous variables in the mean and the volatility are the same 19 indicators described in Section 4.[9]

The GETS follows three main steps. (i) The GUM must be defined both for the mean (the MGUM) and for the volatility (the VGUM), and validated against a set of chosen misspecification tests.[10] (ii) Simplification is undertaken along several paths by deleting a different insignificant regressor in each path. The validity of the simplified models are checked against the misspecification tests, and in a backtest (BaT) against the GUM. For each path in which the simplified model passes those tests, the simplification procedure pursues until there is no insignificant regressor, or until the simplified model does not pass the misspecification tests or the BaT against the GUM. Given that there are several deletion paths, multiple models can remain at the end of the procedure. (iii) The final step consists in implementing a selection procedure between those remaining models using an information criterion, which, by default, is the Bayesian information criterion (BIC). Steps (ii) and (iii) are implemented in two stages. In the first stage, the algorithm focuses on the multi-path specification search for the mean, while letting the VGUM unchanged. The parsimonious specification of the mean is then retained for the second stage, in which the multi-path specification search for the volatility is undertaken.

3.3. The GARCH-MIDAS model

The GARCH models presented above are limited to exploiting information available at the same frequency as the dependent variable. However, several interesting supply and production data relevant to the dynamics of crude oil volatility are only available at lower frequencies. Fortunately, Engle et al. (2013) proposed the GARCH-MIDAS class of models that can solve this frequency mismatch. These models decompose conditional volatility into a short-term and a long-term component. The short-term component follows a standard GARCH process, while the long-term component is represented by a mixed-frequency data sampling (MIDAS) regression with lower frequency variables (Ghysels et al., 2004):

$$\epsilon_{t,m} = \sqrt{\tau_m g_{t,m}} z_t, \quad z_t \sim i.i.d.N(0,1),$$

[8] EqWMA stands for equally weighted moving average, and the EqWMA variables are meant to proxy for log-GARCH terms. $\text{Ln}(\text{EqWMA}_{q,t-1})$ terms are used instead of $\ln(\sigma^2_{t-q})$ because their associated parameters are simpler to estimate, and because OLS inference of the coefficients β_q can be carried out. We choose $Q = \{5, 20, 60\}$ to account for the observation that volatility over long-time intervals has an effect on short-term volatility (Müller et al., 1997; Corsi, 2009).

[9] We start with $P = 10$ instead of 5 because the GUM would otherwise not pass the misspecification test for ARCH effects. After this adjustment, we find that the GUM is not misspecified against autocorrelated residuals and ARCH effects.

[10] In practice, we undertake the GETS procedure in R using Pretis et al.'s (2018) **gets** package. By default, misspecifications include autocorrelated standardized residuals and ARCH effects. While the user can choose to add other diagnostic tests, we keep the default specifications.

$$g_{t,m} = (1-\alpha-\beta) + \alpha\frac{\epsilon^2_{t-1,m}}{\tau_m} + \beta g_{t-1,m}, \tag{8}$$

$$\tau_m = \exp(\theta_0 + \theta\sum_{k=1}^{K}\phi_k x_{m-k}),$$

In (8), the short-term component, $g_{t,m}$ is assumed to be a simple GARCH(1,1).[11] The long-term component, τ_m, depends on the lags of a low-frequency variable, x_t. Given that it can take negative values, the expression is often given using the exponential form. The optimal lag order K is chosen so as to minimize the BIC. θ_0 is an intercept and θ is the slope of the weighted effects of the lagged exogenous variable. In line with the literature (Engle et al., 2013; Wei et al., 2017), we adopt the two-parameter Beta polynomials as weighting scheme:

$$\phi_k(\omega_1,\omega_2) = \frac{(k/K)^{\omega_1-1}(1-k/K)^{\omega_2-1}}{\sum_{j=1}^{K}(j/K)^{\omega_1-1}(1-j/K)^{\omega_2-1}}. \tag{9}$$

3.4. Forecast combinations and dynamic model averaging

3.4.1. Simple forecast combinations

It is well documented that forecast accuracy of single-predictor models can be very unstable, and that simple forecast combinations can lead to improved performance (Pan et al., 2017; Ma et al., 2018; Meng and Liu, 2019). Forecast combinations take the form:

$$\hat{\sigma}_{c,t+1} = \sum_{k=1}^{K}\omega_{k,t}\hat{\sigma}_{k,t+1}, \tag{10}$$

where $\hat{\sigma}_{c,t+1}$ and $\hat{\sigma}_{k,t+1}$ denote the combination forecast and model-k individual forecast, respectively. $\omega_{k,t}$ is the combining weight of model-k individual forecast. We consider the following 3 weighting schemes:

- Mean combination. All individual forecast are equally weighted.
- Trimmed mean combination. The weights of the 20% most extreme forecasts are shrinked to 0 (10% on each side), while the remaining forecasts are equally weighted.
- Discount mean squared predictive error (DMSPE). The weight of the lth individual forecast on day t depends on its past performance: $\omega_{l,t} = \phi^{-1}_{l,t}/\sum_{k=1}^{K}\phi^{-1}_{k,t}$, where $\phi_{k,t} = \sum_{s=1}^{t}(\tilde{\sigma}^2_s - \hat{\sigma}^2_{k,s})^2$, and $\tilde{\sigma}^2_s$ is a volatility proxy (see Section 3.5).

3.4.2. Dynamic model averaging

In a context where there are several time-varying competing models, and where the best predictive model is unknown, Raftery et al. (2010) draw from Bayesian model avering to propose the dynamic model averaging (DMA) methodology. Consider K competing state-space models of the form:

$$y_t = \omega'_{k,t}x_{k,t-1} + \epsilon_{k,t}, \quad \epsilon_{k,t} \sim i.i.d.N(0, V_{k,t}), \tag{11}$$

$$\omega_{k,t} = \omega_{k,t-1} + \eta_{k,t}, \quad \eta_{k,t} \sim i.i.d.N(0, W_{k,t}), \tag{12}$$

where $x_{k,t} \subseteq x_t$, for $k = 1,\dots,K$, denotes the set of predictors of model k at time t. If there are m predictors in x_t, there is a total of $K = 2^m$ possible combinations of these predictors. The one-day ahead DMA prediction is:

$$\hat{y}^{DMA}_t = \sum_{k=1}^{K}\pi_{t|t-1,k}\hat{\omega}'_{k,t-1}x_{k,t-1}, \tag{13}$$

where $\pi_{t|t-1,k} \equiv P[L_t = k|Y^{t-1}]$, with $Y^{t-1} = \{y_1,\dots,y_{t-1}\}$, is the conditional probability that the process is governed by model k. The model governing the process evolves according to a Markov chain. The DMA prediction is therefore a weighted average of the predictions of all models, where the weights are equal to the posterior predictive model probabilities for period t. To save space, we limit our description of the DMA methodology to this general presentation. The interested reader should refer to Raftery et al. (2010) for additional details relative to the recursive estimation of the state-space and Markov chain models.

In this paper, similarly to Wei et al. (2017) and Ma et al. (2018), we implement the DMA method as a time-varying combination approach. That is, x_t is the vector of predictions from our different models at time t, and the DMA dynamically combines all forecasts.[12]

3.5. Forecasting methodology and evaluation

Our primary focus is on out-of-sample (OOS) performance. Our daily sample starts on July 1st 2009, right after the end of the recession in the U.S., according to the NBER, and ends on November 12th 2019 (2608 observations). We choose a rolling window estimation scheme of 1565 observations.[13] It means that all conditional volatility models are estimated with a minimum of 1500 observations.[14] We re-estimate the parameters of the models with each new observation. Specifically, as a new observation becomes available, we (i) drop the first observation of the last window, (ii) add the new observation (so that the window size remains fixed), and (iii) re-estimate the parameters of the model to produce a new forecast. This procedure results in 1042 one-day ahead forecasts.[15]

Contrary to the mean, the volatility is not observable, even ex-post. We use two different volatility proxies that have their own advantages and drawbacks: squared return and the adjusted intraday range. In the hypothesis of zero mean return, since $E[y^2_t|I_{t-1}] = E[z^2_t\sigma^2_t|I_{t-1}] = \sigma^2_t$, squared return is a conditionally unbiased estimator of latent volatility. However, squared return can be very noisy as z^2_t usually exhibits a large degree of variation compared to σ^2_t (Andersen and Bollerslev, 1998). The intraday range is defined by: $rg_t = 100(\max_\tau \ln(P_\tau) - \min_\tau \ln(P_\tau))$, where $t-1 < \tau \le t$. Assuming that the natural logarithm of asset prices follows a Brownian motion, Parkinson (1980) shows that, while the intraday range is a biased estimator of volatility ($E[rg^2_t|I_{t-1}] = 4\ln(2)\sigma^2_t$), the adjusted intraday range ($rg^*_t = rg_t/(2\sqrt{\ln(2)})$) is unbiased and yields a better volatility proxy than squared return.[16] As pointed out by Patton (2011), however, a drawback of the adjusted intraday range is to rely on a particular data generating process (DGP).

[11] Because, in practice, we find the asymmetry term in the IS estimation of a GJR-GARCH(1,1) model to be strongly significant, we retain a slightly different specification of the GARCH-MIDAS model by adding an asymmetry term in its short-term component (similar to the GJR-GARCH model presented above).

[12] Given our important number of single-predictor models, we restrict ourselves to a single regressor in Eq. (11). It means that we have a total of $K = m$ state-space models instead of $K = 2^m$.

[13] Since we lose one observation when we lag our regressors, the first rolling window starts on July 2nd 2009 and ends on September 16th 2015. Therefore, we produce out-of-sample crude oil volatility forecasts from September 17th 2015 to November 12th 2019.

[14] The volatility specification in the GETS procedure is estimated on a minimum of 1500 observations whenever both the fifth autoregressive term in the mean specification, and EqWMA_{60} in the volatility specification, are selected.

[15] We also consider an alternative OOS specification where the size of the rolling window is extended to 2065 observations, which leads to 542 one-day ahead forecasts. The IS period then corresponds to a little bit more than 8 years, and roughly 80% of observations. The first in-sample window starts on July 2, 2009 and ends on September 12, 2017. Therefore, in this specification, we produce out-of-sample crude oil volatility forecasts from September 13, 2017 to November 12, 2019. Results of this robustness test can be found in Section 3 of the *Supplementary material* file for 1-day, 5-day, 21-day, and 63-day ahead forecasts. They are overall consistent with our baseline results.

[16] Since the price of the first nearby futures contract often serves as a proxy of the spot price, we build the intraday range using daily price minima and maxima of first nearby contracts. To avoid undesirable noise due to increased trading activity, we choose to follow Chantziara and Skiadopoulos (2008) and roll from the first to the second nearby contracts 5 days prior to expiration.

Besides the fact that it usually leads to an underestimation of the predictability of conditional volatility, using a proxy instead of the true latent volatility can also influence the comparison of competing models. Hansen and Lunde (2006) find that some criteria, such as the R^2 for Mincer and Zarnowitz's (1969) regression, do not satisfy the necessary conditions to ensure equivalence of rankings induced by the use of a volatility proxy. Likewise, Patton (2011) shows that only two out of nine commonly used loss functions (the MSPE and the QLIKE) satisfy the necessary and sufficient conditions to preserve the rankings of competing models when using noisy volatility proxies. Thereupon, we adopt these two loss functions that are also among the most popular in the literature. They are computed as follows:

$$\text{MSPE} = (1/S)\sum_{t=R+1}^{T}(\tilde{\sigma}_t^2 - \hat{\sigma}_t^2)^2,$$

$$\text{QLIKE} = (1/S)\sum_{t=R+1}^{T}((\tilde{\sigma}_t^2/\hat{\sigma}_t^2) - \ln(\tilde{\sigma}_t^2/\hat{\sigma}_t^2) - 1),$$

where T, R, and S denote our total sample size, the size of our rolling window, and our number of one-step ahead predictions, respectively. $\tilde{\sigma}_t^2$ is a variance proxy, and $\hat{\sigma}_t^2$ is our conditional variance forecast. These two loss functions have distinct features. First, the MSPE is symmetric, while the QLIKE is asymmetric. The latter penalizes much more strongly underestimations than overestimations. Second, while losses associated with the MSPE are usually higher than losses associated with the QLIKE, it is the opposite for large underestimations of the true variance.

To infer with some degree of confidence whether a model is better than another, we use Harvey et al.'s (1997) modified version of Diebold and Mariano's (1995) equal predictive accuracy test statistic, called MDM. Let $L(e_{0t})$ and $L(e_{it})$ denote the time-t loss functions associated with forecast errors from a benchmark and from an alternative model, respectively. Additionally, let $d_{0it} = L(e_{0t}) - L(e_{it})$ be the loss differential between these two forecasts. The null hypothesis for the MDM test is $H0 : E(d_{0it}) = 0$, and the associated statistic is:

$$\text{MDM}_{0i} = ((T-1)/T)^{1/2}\frac{\bar{d}_{0i}}{\hat{\sigma}_{\bar{d}_{0i}}},$$

where $\bar{d}_{0i} = \frac{1}{S}\sum_{t=R+1}^{T} d_{0it}$ is the sample mean loss differential, and $\hat{\sigma}_{\bar{d}_{0i}}$ is a consistent estimate of the standard deviation of $\bar{d}_{0i}$. Given the assumption that the loss differential is covariance stationary, MDM_{0i} follows a Student's t distribution with $(T-1)$ degrees of freedom.

In addition, we implement Hansen et al.'s (2011) model confidence set (MCS) test. The MCS test allows to determine a set of best models. Consider an initial set of I competing models, M_I. The relative performance of model i to model j is defined by: $d_{ijt} = L(e_{it}) - L(e_{jt})$. Model i is considered to be a better model than j in terms of expected loss if $E(d_{ijt}) < 0$. The best set of models can then be defined as: $M^* \equiv \{i \in M_I \mid E(d_{ijt}) \leq 0,\ \forall j \in M_I\}$. The determination of M^* is based on an iterative procedure. You first test the null hypothesis $H0_{M_I} : E(d_{ijt}) = 0,\ \forall i, j \in M_I$ for some confidence level $1-\alpha$. If $H0_{M_I}$ is rejected, you use some elimination rule, e_M, to withdraw the poorest models from M_I. You then have a smaller pool of competing models, $M_{I-1} \subset M_I$, for which you can reiterate the test until the null hypothesis is not rejected. This final set of models, $\widehat{M}^*_{1-\alpha}$, is the MCS.

The statistic associated with the hypothesis $H0_{M_{I-k}}$ is:

$$T^{MCS}_{M_{I-k}} = \max_{i \in M_{I-k}} t_{i.},$$

$$\text{with:} \quad t_{i.} = \frac{\bar{d}_{i.}}{\sqrt{\widehat{var}(\bar{d}_{i.})}}, \quad \text{for } i \in M_{I-k},$$

where $\bar{d}_{i.} = I^{-1}S^{-1}\sum_{j \in M_{I-k}}\sum_{t=R+1}^{T} d_{ijt}$, and $\widehat{var}(\bar{d}_{i.})$ denotes the estimate of $var(\bar{d}_{i.})$. The asymptotic distribution of $T^{MCS}_{M_{I-k}}$ is estimated with a bootstrap. If the null is rejected, a model is discarded from M_{I-k} according to the elimination rule: $e_M \equiv \arg\max_{i \in M} t_{i.}$.

Table 2
Descriptive statistics.

	Mean	SD	Skew	Kurt	JB p-val	ADF p-val
WTI	−0,017	2,022	−0,099	1,293	<0,01***	<0,01***
AOI	0,031	0,967	−0,234	1,086	<0,01***	<0,01***
AVOL	2,708	24,718	0,739	0,690	<0,01***	<0,01***
TB3M	0,050	1,559	0,195	2,916	<0,01***	<0,01***
LTGBY	−0,066	5,043	0,089	0,664	<0,01***	<0,01***
Basis	−0,011	54,269	−0,231	1,814	<0,01***	<0,01***
S&P 500	0,050	0,906	−0,345	2,023	<0,01***	<0,01***
MSCI EM	0,016	0,968	−0,187	1,099	<0,01***	<0,01***
GSCI	−0,002	1,196	−0,137	1,038	<0,01***	<0,01***
BCOM	−0,015	0,864	−0,064	0,969	<0,01***	<0,01***
USD–AUD	0,006	0,675	0,104	0,906	<0,01***	<0,01***
USD–CAD	0,004	0,525	0,049	0,830	<0,01***	<0,01***
USD–CLP	0,012	0,597	−0,006	1,020	<0,01***	<0,01***
USD–NZD	0,000	0,716	0,043	0,723	<0,01***	<0,01***
USD–ZAR	0,023	0,976	0,145	0,553	<0,01***	<0,01***
VIX	17,263	5,751	1,515	2,904	<0,01***	<0,01***
OVX	33,690	10,131	0,707	0,598	<0,01***	0,033**
BDI	−0,047	2,294	0,121	1,204	<0,01***	<0,01***
EPU	103,487	57,166	1,262	1,659	<0,01***	<0,01***
Prod	0,114	0,715	−0,139	0,241	<0,01***	<0,01***
Kilian	−25,724	50,244	0,223	−0,163	<0,01***	0,296
Prod vol	0,467	0,713	3,136	13,760	<0,01***	<0,01***
Kilian vol	332,060	473,940	2,112	4,421	<0,01***	<0,01***
Working T	1,053	0,026	1,597	2,454	<0,01***	0,026**

Notes: JB and ADF refer to the p-values of the Jarque Bera and augmented Dickey–Fuller tests, respectively. *, **, and *** denote rejections of the null hypothesis at the 10%, 5%, and 1% significance level, respectively.

4. Data

Table 2 displays key descriptive statistics of our variables. In accordance with recent results (Charles and Darné, 2017), we filter our daily series for outliers using the method proposed by Boudt et al. (2008). Our dependent variable is the daily spot price of the WTI retrieved from the website of the Energy Information Administration (EIA).[17] For stationary reasons, and in line with the literature, we build returns as the logarithmic difference of prices: $y_t = 100\ln(P_t/P_{t-1})$, where P_t is the spot price at time t. The mean of crude oil daily returns is almost null (−0.02%), while it has a much higher standard deviation (2.02%). The null hypotheses of normality and non-stationarity are rejected at the 1% level. Crude oil is skewed to the left and exhibits excess kurtosis. Hence the advantage of the GETS procedure to recourse to an estimation method that does not assume Gaussian standardized residuals. Even though GETS modelling allows to identify a parsimonious model, we still restrict our analysis to a handful of the predictors present in the literature to avoid overfitting. We consider a total of 18 daily variables: (1) the percent growth rate of aggregate open interest (AOI), (2) the percent growth rate of aggregate volume (AVOL), (3) the difference in the yield on the 3-month Treasury bill (TB3M), (4) the difference in the yield on the U.S. 10-year government bond (LTGBY), (5) the difference in the basis, (6) log-returns on the S&P 500, (7) log-returns on the MSCI Emerging markets (MSCI EM), (8) log-returns on the S&P GSCI, (9) log-returns on the Bloomberg Commodity Index (BCOM), log-returns on bilateral exchange rates (10) between the U.S. dollar and the Australian dollar (USD–AUD), (11) between the U.S. dollar and the Canadian dollar (USD–CAD), (12) between the U.S. dollar and the Chilean peso (USD–CLP), (13) between the U.S. dollar and the New-Zealand dollar (USD–NZD), and (14) between the U.S. dollar and the South African rand (USD–ZAR), (15) the VIX, (16) the OVX, (17) log-returns on the Baltic Dry Index (BDI), and (18) Baker et al.' (2016)

[17] In Section 5.3, we replicate the results of our paper with an alternative crude oil series, namely Brent crude oil. The source of this series is also the EIA website.

daily U.S. Economic Policy Uncertainty (EPU) index.[18] We construct the basis as the difference between the futures price and the spot price: $B_t = F_{t,T} - P_t$, where the futures price is the price of the first nearby contract. All variables are retrieved from Bloomberg, except for TB3M that comes from the Fred database, and EPU that comes from Baker et al.'s website.[19]

Our dataset for the long-term component of the GARCH-MIDAS models include 5 variables: the monthly level and volatility of (the log-growth of) global crude oil production and of Kilian's (2009) real global economic activity index, and the weekly Working speculative T index. Global supply and demand data are available on the website of the EIA and on Kilian's personal website, respectively. We follow the standard approach of Schwert (1989) and Engle et al. (2013) to construct the volatility of these variables:

$$x_t = \sum_{i=1}^{12} \alpha_i d_{it} + \sum_{i=1}^{12} \beta_i x_{t-i} + \epsilon_t, \tag{14}$$

where x_t is either the log-change in global crude oil production or Kilian's activity index, and d_{it} is a monthly dummy variable for month i (i.e., d_{it} takes the value 1 in month i and 0 every other month). The squared residuals, ϵ_t^2, are taken as the proxy of volatility of these macroeconomic variables (Pan et al., 2017; Meng and Liu, 2019). Working's (1960) T index is a standard measure of speculation in the literature. It reflects the level of speculation in excess of the minimum necessary to absorb hedging needs. It is computed using data from the Commitment of Traders reports of the Commodity Futures Trading Commission:

$$\begin{cases} T = 1 + SS/(HL + HS) & \text{if} \quad HS \geq HL \\ T = 1 + SL/(HL + HS) & \text{if} \quad HS < HL \end{cases} \tag{15}$$

where SL and SS denote the long and short positions of non-commercial traders, while HL and HS denote the long and short positions of commercial traders.

5. Results

5.1. In-sample

Table 3 summarizes results for two terminal models selected by the GETS algorithm. The first model, called GETS1, results from the default procedure. Whenever a variable is removed, the standardized residuals are checked for serial correlation and ARCH effects.[20] It explains why several variables in the mean equation are insignificant. It may be odd to keep so many variables in the mean equation, when the consensus in the literature is that daily oil return dynamics are dominated by variance (Wei et al., 2017; Pan et al., 2017). The second model, called GETS2, results from the same procedure but with the misspecification tests deactivated. Contrary to GETS1, no variable is retained in the mean specification. Selected variables in the volatility equation are very similar for both models. The only difference is that a second exchange rate remains in GETS2. The BIC is lower for GETS2 than for GETS1 suggesting that it might be preferable to keep a zero-mean specification.

Fig. 1 shows fitted standard deviation together with absolute residuals as well as standardized residuals for GETS1 (panel (a)) and GETS2 (panel (b)). For the first model, it also displays the fitted model for the mean. This last part shows very little fluctuations in the fitted mean specification, which is consistent with the widespread view that asset prices behave like a random walk. Given these results, we retain the GETS2 specification for the out-of-sample analysis. The later is more parsimonious as it will discard insignificant regressors whether standardized residuals exhibit serial correlation and ARCH effects or not.

Table 3
In-sample estimation results.

	GETS1		GETS2	
	Coef	t-stat	Coef	t-stat
Mean equation				
y_{t-1}	−0.062	−2.064**		
LTGBY	0.005	0.596		
USD–AUD	0.086	0.987		
USD–CAD	−0.265	−2.248**		
USD–CLP	−0.084	−1.129		
Volatility equation				
Volatility constant	−0.768	18.561***	−0.797	20.248***
ln(EqWMA$_{20}$)	0.281	2.724***	0.276	2.721***
USD–CAD			−0.179	−1.785*
USD–ZAR	0.116	2.596***	0.148	2.766***
OVX	0.050	6.491***	0.051	6.716***
Information criteria				
Log-likelihood	−5243.51		−5252.19	
BIC	4.083		4.076	
Diagnostic tests				
LB AR-X (lag)	0.36 (2)		0.56 (1)	
LB ARCH-X (lag)	0.31 (1)		0.30 (1)	

Notes: *, **, and *** denote statistical significance at the 10%, 5%, and 1% significance level, respectively.

Consistent with the literature, the OVX is by far the most significant exogenous variable retained in the volatility equation (Gong and Lin, 2018; Lv, 2018). Interestingly, two exchange rates are significant in GETS1: USD–CAD in the mean specification, and USD–ZAR in the volatility. Canada is one of the largest exporters of crude oil. Because of the forward-looking nature of exchange rates, expectation of future higher oil prices may reverberate preemptively on the CAD as investors anticipate an amelioration of Canada's terms of trades. A similar mechanism may be at play for the rand and volatility. South Africa is a net importer of crude oil. Expectations of a future rise in crude oil volatility increases uncertainty regarding the evolution of South Africa's terms of trade and the value of its currency. If volatility is priced, investors may discount the value of the ZAR.

An issue with these hypotheses is that the incorporation of one of those exchange rates in the terminal model does not seem independent from the incorporation of the other. Whenever we suppress either the USD–CAD or the USD–ZAR from the starting pool of variables, the other one also disappears from the terminal model. An alternative hypothesis is that these variables may capture an asymmetric relationship between returns and volatility. Indeed, even though no asymmetry term is selected by the GETS procedure, IS estimation of the GJR-GARCH(1,1) model (not reported) exhibits a strongly significant leverage effect. Moreover, the USD–CAD and USD–ZAR exchange rates appear with opposite signs, and are strongly positively correlated (0.56). This result is in contradiction with predictions of the theory of storage (Kaldor, 1939; Working, 1949), but is consistent with previous empirical findings (Kristoufek, 2014; Carnero and Pérez, 2019).[21] An

[18] For our robustness test with Brent crude oil, in Section 5.3, we made slight changes to our predictors. Specifically, we considered Brent futures contracts instead of WTI futures contracts for AOI, AVOL, and the basis.

[19] www.policyuncertainty.com.

[20] The Ljung–Box tests for serial correlation and ARCH effects use a number of lags equals to the AR and ARCH order of the model plus 1, respectively. The default p-value under which the diagnostics fail is 0.025. P-values are reported in Table 3 for the terminal models.

[21] The theory of storage (Kaldor, 1939; Working, 1949) predicts an inverse leverage effect for commodities. Two mechanisms can be highlighted. First, volatility has an effect on prices through inventories (Litzenberger and Rabinowitz, 1995; Pindyck, 2004). As uncertainty about the future price of a commodity rises, the marginal value of storage increases, supply for the current period falls, and spot prices increase. Second, in situation of scarcity, both short-term prices and volatility will be high (Fama and French, 1988; Ng and

appealing explanation calls for a distinction between consumption and investment commodities (Chiarella et al., 2016; Baur and Dimpfl, 2018). As they serve different purposes to their purchasers, market forces should impact them differently. An increase (decrease) in the price of a consumption commodity is "bad (good) news" for consumers and volatility should rise (fall). On the opposite, an increase (decrease) in the price of an investment commodity is "good (bad) news" for investors, and volatility should fall (rise). According to this narrative, the asymmetric return–volatility relation for crude oil is that of an investment commodity, similar to stocks.

It is unclear, however, why this asymmetric effect would pass through an exchange rate channel. We investigated whether it could be the consequence of a dollar effect by testing several different specifications in which we replace our selected exchange rates with a trade weighted U.S. dollar index, or in which we simply add the dollar index to the GETS procedure. It never remains in the terminal model. It is both an advantage and a drawback of the GETS procedure that we do not have to specify a precise model directly. On the one hand, the GETS aims at achieving the best forecasting model, and aside from the determination of the original set of variables, its approach is first and foremost a statistical one. On the other hand, while the GETS can have explanatory virtues as it helps uncover unexplored dynamics, some results can be hard to interpret. A careful study of the interactions between exchange rates and commodity prices is beyond the scope of this paper, and we let it for future research.

5.2. Out-of-sample

5.2.1. Single-predictor models

We now focus on OOS predictive accuracy of the GETS procedure. First, we compare its performance to several competing models: 4 benchmarks (a GARCH(1,1), a GJR-GARCH(1,1), an EGARCH(1,1), and a log-GARCH(1,1)), single-variable models based on the log-GARCH form (log-GARCH(1,1)-X), and single-variable models based on the GARCH-MIDAS.[22],[23]

Table 4 displays the MSPE and QLIKE of all models for our two volatility proxies as well as their ranks. We find the log-GARCH-X model with the OVX (henceforth, log-GARCH-OVX) to be the most accurate in 3 out of 4 situations. He ranks 9th only with squared return and the MSPE loss function. This result is all the more impressive that the log-GARCH(1,1) with no exogenous predictor ranks last in all cases. It strongly supports the role of the OVX as an indispensable predictor of crude oil volatility. Although, the OVX clearly stands out, and even if other log-GARCH-X models rarely rank among the best models, there are still sizable differences in the loss of these single-predictor models. It suggests that other exogenous variables may at times hold relevant information to forecast oil volatility.

Besides the log-GARCH(1,1) model, other benchmarks perform well. The GJR-GARCH(1,1) and the EGARCH(1,1), in particular, consistently rank in the first four places.[24] In a large majority of cases, GARCH-MIDAS models are outperformed by these benchmarks and the log-GARCH-OVX, but they outcompete other single-predictor log-GARCH-X models. How accurate is the GETS? It does better than all GARCH-MIDAS models except when using squared return as volatility proxy and the MSPE loss function. The GETS performs best relatively to other models when using the QLIKE with squared return and the MSPE with the intraday range. In the first situation, out of the 28 models, it achieves the 4th lowest QLIKE following the log-GARCH-OVX, the EGARCH(1,1), and the GJR-GARCH(1,1). Aside from these models, one-sided MDM tests show that the GETS has a significantly lower QLIKE than all models except for the GARCH(1,1) and GARCH-MIDAS models.[25] With the intraday range, the GETS procedure has the 3rd lowest MSPE right after the log-GARCH-OVX and the EGARCH(1,1). Again, MDM tests show that the GETS has a significantly lower MSPE than the log-GARCH(1,1) benchmark and all single-predictor models but the log-GARCH-OVX. Asterisks in Table 4 denote models included in the MCS for three different significance levels (10%, 25%, and 40%). The GETS always belong to the best performing models, but so are 3 of the 4 benchmarks.

Fig. 2 displays OOS forecasts of selected models with the true volatility as measured by the preferred intraday range proxy. Panel (a), (b), and (d) show that benchmarks and the best performing GARCH-MIDAS model exhibit relatively narrow fluctuations. Panel (c) and (e), on the contrary, reveals that the log-GARCH-OVX and the GETS exhibits stronger fluctuations. The latter, in particular, seems to capture more accurately large peaks in the volatility, especially in February and November 2016, and in November 2018.

5.2.2. Multiple-predictor models

Overall, the GETS compares well to single-variable models. Table 5 exhibits results from simple forecast combinations (mean, trimmed mean, and DMSPE) and from Raftery et al.'s (2010) DMA method for three pools of candidates: single-predictor log-GARCH-X models, GARCH-MIDAS models, and the sum of these two categories and the benchmarks.[26] We first note that, among these models, the GETS has the lowest QLIKE for squared return. Unsurprisingly, simple combinations of GARCH-MIDAS models yield lower MSPE than the GETS with squared return as it was already true for individual models. Aside from this case, more often than not, the GETS produces more accurate predictions than simple forecast combinations.

Pirrong, 1994; Geman and Ohana, 2009; Geman and Smith, 2013). Prices rise as purchasers raise their bids to secure supply. And volatility increases as news about short-term supply and demand reverberate strongly on prices.

[22] We also considered extensions of our other benchmarks (GARCH(1,1)-X, GJR-GARCH(1,1)-X, and EGARCH(1,1)-X), but we find them to be less interesting than the log-GARCH(1,1)-X. Adding covariates to the GARCH and GJR-GARCH models does not improve their accuracy. IS estimation shows that all coefficients are basically null, while OOS performance very slightly deteriorates. Results are different for the EGARCH and log-GARCH. IS results show that several covariates appear significant, while the BIC significantly changes depending on the exogenous variable. This is particularly true for the log-GARCH. Except for one variable for which the EGARCH-X fits the data particularly poorly, all models have a BIC between 4.07 and 4.09, while BIC of the log-GARCH-X models range from 4.05 to 4.11. Moreover, there are some estimation issues for the EGARCH-X out-of-sample. The default solver of the **rugarch** package often fails to converge, while other candidate solvers achieve diverging forecasts with errors that go to infinity. Hence our focus on the log-GARCH(1,1)-X specification for daily single-predictor models.

[23] Except for the GETS procedure, log-GARCH models are estimated using Francq and Sucarrat's (2018) Cex-χ^2 QMLE. In theory, Francq and Sucarrat's (2018) Cex-χ^2 QMLE combines the strength of several competing estimators: (i) maximum efficiency when standardized residuals are Gaussian, and (ii) zeros and missing values can be handled with Sucarrat and Escribano's (2018) recent algorithm. In addition, we find this estimator to have better IS fit than the least squares and Gaussian ARMA-QMLE estimators for the log-GARCH(1,1) and for all log-GARCH-X models.

[24] It is well-known that a misspecified standardized error density can result in a loss of efficiency (Bauwens et al., 2012). Standardized residuals of the GARCH(1,1), GJR-GARCH(1,1), and EGARCH(1,1) models are still negatively skewed, and exhibit excess kurtosis. As a robustness test, we considered two alternative distributions for the GARCH(1,1), GJR-GARCH(1,1), and EGARCH(1,1) models: (i) the Student's t distribution as it is the preferred fat-tailed distribution of practitioners, and (ii) the skewed Student's t distribution to cater for asymmetry. Our results were not significantly affected by these alternatives.

[25] Results of one-sided MDM tests are reported in Section 1 of the *Supplementary material* file.

[26] For obvious reasons, for the trimmed mean of GARCH-MIDAS models, we discard the lowest and highest predictions instead of the 20% most extreme forecasts as is indicated in Section 3.4.1.

Table 4
Out-of-sample results.

	Squared return				Intraday range			
	MSPE	Rank	QLIKE	Rank	MSPE	Rank	QLIKE	Rank
Benchmarks								
GARCH(1,1)	59,624***	8	1,387***	5	40,697***	5	0,361***	8
GJR-GARCH(1,1,1)	58,124***	2	1,376***	3	39,440***	4	0,352***	6
EGARCH(1,1)	**57,397*****	1	1,376***	2	38,101***	2	0,369***	14
log-GARCH(1,1)	419,904**	28	1,615	28	330,758**	28	0,427*	28
Single-predictor log-GARCH(1,1)-X								
AOI	148,645*	23	1,559	24	90,918*	23	0,378***	19
AVOL	146,344*	21	1,558	23	88,777*	21	0,376***	18
TB3M	139,361	19	1,560	25	87,130*	19	0,384***	23
LTGBY	170,113**	26	1,545	18	108,789**	26	0,371***	16
Basis	157,089**	24	1,552	21	96,083**	24	0,366***	10
S&P 500	184,963**	27	1,543	16	111,862**	27	0,360***	7
MSCI EM	117,876**	17	1,528	14	61,146**	15	0,352***	5
GSCI	123,160**	18	1,503	11	67,376**	18	0,329***	2
BCOM	109,534**	14	1,523	13	59,294**	14	0,350***	3
USD–AUD	102,337**	12	1,539	15	54,771**	12	0,365***	9
USD–CAD	87,306**	11	1,520	12	47,291**	11	0,350***	4
USD–CLP	104,987**	13	1,544	17	57,784**	13	0,367***	12
USD–NZD	117,549**	16	1,550	20	66,692***	17	0,375***	17
USD–ZAR	115,709**	15	1,549	19	63,067**	16	0,369***	13
VIX	146,647	22	1,558	22	87,984*	20	0,371***	15
OVX	61,867***	9	**1,350*****	1	**34,168*****	1	**0,299*****	1
BDI	144,738*	20	1,561	26	89,303*	22	0,380***	20
EPU	164,669*	25	1,563	27	106,981*	25	0,384***	22
GARCH-MIDAS								
Prod	58,924***	5	1,392***	7	42,891***	9	0,395***	24
Kilian	58,682***	3	1,395***	9	41,736***	7	0,403***	26
Prod vol	59,363***	7	1,393***	8	43,218***	10	0,398***	25
Kilian vol	58,985***	6	1,405***	10	42,822***	8	0,406***	27
Working T	58,914***	4	1,392***	6	41,447***	6	0,382***	21
AR-X-log-GARCH-X								
GETS	69,008***	10	1,378***	4	38,813***	3	0,366***	11

Notes: Bold characters signal models with the lowest information criteria. *, **, and *** mean that the model is in $\widehat{M}^*_{90\%}$, $\widehat{M}^*_{75\%}$, and $\widehat{M}^*_{60\%}$, respectively.

Our results also show that the DMA usually performs better than simple combinations when applied to a sufficiently large pool of models. With the intraday range, the DMA applied to the largest pool of models has the lowest MSPE and QLIKE of all models. Still, panel (f) of Fig. 2 shows that the DMA, although it does better than standard benchmarks or single-predictor models, falls short of capturing peaks in the volatility compared to the GETS and the log-GARCH-OVX. The DMA applied to the pool of single predictor log-GARCH-X models also has lower MSPE and QLIKE than the GETS with the intraday range. One-sided MDM tests show that these two DMA significantly outperform the GETS at the 10% level using the MSPE and at the 1% level using the QLIKE. Results are less clear with squared return as our volatility proxy. Both DMA procedures achieve a lower MSPE and a higher QLIKE than the GETS.

Despite the strong results of DMA, if we implement the MCS procedure after adding simple combination forecasts and the DMA as candidate models, we still find the GETS to always be in the final collection of best models. As a final test, to figure out whether the GETS can be useful in a forecast combination framework, we add it to the DMA procedure. With squared return, forecast accuracy of the DMA remains stable (MSPE = 60.645, QLIKE = 1.518). With the intraday range, however, accuracy improves further (MSPE = 33.228, QLIKE = 0.281). According to MDM tests, these improvements are significant at the 1% level.[27]

[27] Specifically, for the intraday range, p-values of the one-sided MDM tests yield 0.002 for the MSPE and 0.000 for the QLIKE (with the alternative hypothesis being that the DMA that includes the GETS produces better forecasts than the DMA that does not include it).

Table 5
Simple forecast combinations and dynamic model averaging.

	Squared return		Intraday range	
	MSPE	QLIKE	MSPE	QLIKE
Single-predictor log-GARCH-X models				
Mean	111,722	1,521	60,087	0,346
Trimmed mean	115,937	1,533	63,423	0,356
DMSPE	98,719	1,497	52,670	0,332
DMA	60,641	1,517	33,761	0,283
GARCH-MIDAS models				
Mean	**58,753**	1,389	42,202	0,388
Trimmed mean	58,809	1,390	42,319	0,391
DMSPE	59,215	1,394	42,814	0,395
DMA	60,739	1,527	36,805	1,402
Benchmarks + log-GARCH-X models + GARCH-MIDAS models				
Mean	91,705	1,469	47,547	0,318
Trimmed mean	88,034	1,469	45,173	0,318
DMSPE	69,474	1,413	35,628	0,296
DMA	60,471	1,518	**33,517**	**0,282**
AR-X-log-GARCH-X				
GETS	69,008	**1,378**	38,813	0,366

Notes: Bold characters signal models with the lowest information criteria.

5.2.3. Stability of the models' evaluation

We produce 1042 point forecasts over a four-year period. In the previous section, we evaluated our models over the entire OOS period. However, it is likely that the performance of the models changes over time. Some model might dominate over some period of time and be outperformed in another.

Table 6
5-day ahead out-of-sample results.

	Squared return				Intraday range			
	MSPE	Rank	QLIKE	Rank	MSPE	Rank	QLIKE	Rank
Benchmarks								
GARCH(1,1)	60,565***	8	1,402***	5	43,675***	5	0,386***	18
GJR-GARCH(1,1)	58,778***	2	1,388***	3	42,395***	4	0,389***	20
EGARCH(1,1)	**58,207*****	1	1,388***	4	41,797***	3	0,412***	22
log-GARCH(1,1)	363,221**	28	1,610	28	295,203***	28	0,430**	23
Single-predictor log-GARCH(1,1)-X								
AOI	115,712	20	1,539	23	73,263**	20	0,373***	13
AVOL	119,110	21	1,542	25	76,165**	22	0,376***	15
TB3M	111,526	18	1,547	26	72,994***	19	0,387***	19
LTGBY	126,971**	25	1,527	19	82,807***	25	0,368***	11
Basis	130,243**	26	1,535	22	85,463***	27	0,367***	10
S&P 500	134,898*	27	1,525	16	81,216***	24	0,355**	5
MSCI EM	96,285**	15	1,509*	13	53,140***	14	0,346***	3
GSCI	120,171**	22	1,498**	11	68,419***	18	**0,338*****	1
BCOM	97,680**	17	1,510*	14	56,227***	16	0,354***	4
USD–AUD	87,123**	12	1,519*	15	49,360***	12	0,360***	7
USD–CAD	79,880**	11	1,498*	12	41,601***	2	0,345***	2
USD–CLP	93,891**	14	1,526	18	55,631***	15	0,365***	9
USD–NZD	96,365**	16	1,531**	20	57,254***	17	0,370***	12
USD–ZAR	93,683**	13	1,526	17	52,089***	13	0,361***	8
VIX	122,152	23	1,560	27	77,179***	23	0,379***	17
OVX	59,217***	3	**1,364****	1	**38,921*****	1	0,360**	6
BDI	115,122	19	1,541	24	73,924***	21	0,376***	16
EPU	125,971*	24	1,532	21	83,853***	26	0,374***	14
GARCH-MIDAS								
Prod	60,022***	4	1,410***	6	46,022***	8	0,437***	24
Kilian	60,048***	5	1,417***	7	44,796***	7	0,446**	26
Prod vol	61,052***	9	1,422***	8	46,689***	10	0,450**	27
Kilian vol	60,297***	7	1,430***	10	46,380***	9	0,455**	28
Working T	60,212***	6	1,424***	9	44,733***	6	0,444***	25
AR-X-log-GARCH-X								
GETS	71,427***	10	1,387***	2	48,263***	11	0,399***	21

Notes: Bold characters signal models with the lowest information criteria. *, **, and *** mean that the model is in $\widehat{M}^*_{90\%}$, $\widehat{M}^*_{75\%}$, and $\widehat{M}^*_{60\%}$, respectively.

We investigate the stability of our models accuracy by computing the ratio between the loss functions of two competing models over 980 rolling sub-samples. We follow Herrera et al. (2018) and choose sub-samples of 63 observations. Mathematically, let $L(e_{it})$ and $L(e_{jt})$ denote the time-t loss functions of models i and j, we construct the loss ratio:

$$LR_t = \frac{\sum_{s=t-62}^{t} L(e_{is})}{\sum_{s=t-62}^{t} L(e_{js})}$$

where L is either the QLIKE or the MSPE, and where t goes from December 15th 2015 to November 12th 2019. Fig. 3 plots the ratios of the MSPE and QLIKE, for both measures of volatility, for four models relative to the GETS. In each case, we select the most accurate benchmark and GARCH-MIDAS model, the log-GARCH-OVX, and the DMA method applied to the largest pool of models. A ratio above 1 indicates a period during which the GETS is more accurate than the competing model, and, conversely, a ratio below 1 indicates a period during which the GETS is outperformed by the competing model.
Except for panel (b), the ratios vary considerably over the sample. If we focus on the intraday range, the GETS is outperformed by the selected models between roughly February and April 2016 as well as during the second half of 2017. The DMA continues to outperform the GETS over 2018. On the contrary, the GETS does better than all other models from May 2016 to July 2017.

Table 7
5-day ahead simple forecast combinations and dynamic model averaging.

	Squared return		Intraday range	
	MSPE	QLIKE	MSPE	QLIKE
Single-predictor log-GARCH-X models				
Mean	92,508	1,503	52,848	0,344
Trimmed mean	95,581	1,515	55,451	0,353
DMSPE	87,167	1,489	49,962	0,337
DMA	61,742	1,397	**38,086**	**0,301**
GARCH-MIDAS models				
Mean	**60,023**	1,412	45,421	0,435
Trimmed mean	60,126	1,414	45,650	0,439
DMSPE	60,495	1,425	46,065	0,447
DMA	61,544	1,409	41,219	1,422
Benchmarks + log-GARCH-X models + GARCH-MIDAS models				
Mean	80,974	1,461	46,263	0,326
Trimmed mean	77,865	1,460	44,600	0,326
DMSPE	69,833	1,429	40,038	0,319
DMA	61,778	1,397	38,133	0,302
AR-X-log-GARCH-X				
GETS	71,427	**1,387**	48,263	0,399

Notes: Bold characters signal models with the lowest information criteria.

5.2.4. *h*-step Ahead forecasts

We now focus on forecasts at more distant horizons: one week (5 days), one month (21 days), and three months (63 days).[28] The GETS and log-GARCH-X models require to produce forecasts of the regressors to be used as inputs. This is potentially an important drawback of models based on exogenous predictors as uncertainty accrues with each regressor and each new step. We choose a simple approach where forecasts of each regressor x_k are produced from an ARIMA model selected according to the BIC from a general ARIMA(63,2,63)

[28] Section 2 of the *Supplementary material* file provide details on our approach for h-step ahead forecasts.

Table 8
21-day ahead out-of-sample results.

	Squared return				Intraday range			
	MSPE	Rank	QLIKE	Rank	MSPE	Rank	QLIKE	Rank
Benchmarks								
GARCH(1,1)	66,683***	7	1,476***	4	54,583***	6	0,477***	20
GJR-GARCH(1,1)	64,897***	2	1,474***	3	**53,258*****	1	0,498***	22
EGARCH(1,1)	**64,217*****	1	1,480***	5	53,625***	3	0,532***	23
log-GARCH(1,1)	231,823***	27	1,616	27	206,973***	27	0,475***	19
Single-predictor log-GARCH(1,1)-X								
AOI	76,639***	13	1,517***	15	59,584***	15	0,416***	5
AVOL	82,668***	20	1,531***	21	64,034***	20	0,420***	10
TB3M	78,066***	18	1,547***	23	61,424***	18	0,446***	14
LTGBY	77,022***	15	1,512***	12	60,739***	17	0,418***	8
Basis	89,336***	22	1,527***	19	70,335***	23	0,412***	4
S&P 500	78,773***	19	1,504***	10	58,634***	14	0,396***	3
MSCI EM	72,827***	11	1,499***	8	53,467***	2	**0,393*****	1
GSCI	249,422***	28	1,620***	28	209,063***	28	0,465***	16
BCOM	126,460**	26	1,569**	24	96,646***	26	0,443***	13
USD–AUD	72,287***	10	1,514***	13	54,195***	5	0,418***	7
USD–CAD	101,304***	24	1,515**	14	70,172***	21	0,416***	6
USD–CLP	98,975***	23	1,579***	25	76,105***	24	0,472***	18
USD–NZD	73,272***	12	1,518***	16	56,534***	12	0,424***	11
USD–ZAR	77,807***	17	1,501***	9	56,000***	10	0,395***	2
VIX	104,157***	25	1,596**	26	79,538***	25	0,448***	15
OVX	69,158***	9	**1,440*****	1	55,642***	9	0,465***	17
BDI	77,469***	16	1,520***	18	60,507***	16	0,420***	9
EPU	76,998***	14	1,496***	7	61,572***	19	0,437***	12
GARCH-MIDAS								
Prod	65,164***	4	1,518***	17	55,585***	8	0,571*	26
Kilian	65,826***	6	1,493***	6	55,394***	7	0,559*	24
Prod vol	67,512***	8	1,544***	22	57,705***	13	0,612	28
Kilian vol	65,157***	3	1,509***	11	56,127***	11	0,580	27
Working T	65,383***	5	1,529***	20	54,185***	4	0,568*	25
AR-X-log-GARCH-X								
GETS	84,471***	21	1,461***	2	70,252***	22	0,495***	21

Notes: Bold characters signal models with the lowest information criteria. *, **, and *** mean that the model is in $\widehat{M}^*_{90\%}$, $\widehat{M}^*_{75\%}$, and $\widehat{M}^*_{60\%}$, respectively.

unrestricted model.[29,30] This selection procedure is carried out for each rolling window.

First, we note a few general remarks that resonate with results of 1-day ahead forecasts. Although the log-GARCH-OVX model is still arguably the best at the weekly horizon, it is no longer the case for monthly and quarterly horizons. GARCH(1,1), GJR-GARCH(1,1), and EGARCH(1,1) benchmarks still do very well even though their performance deteriorates comparatively to their competitors for 63-day volatility forecasts. The performance of GARCH-MIDAS models follows a similar pattern although with a significantly lower degree of precision.

There are still sizable differences between the accuracy of single-predictor models suggesting that they also hold relevant information for longer-horizon volatility forecasts. It is especially the case for quarterly forecasts for which several log-GARCH-X models perform very well. It is the case in particular of models that include the MSCI EM, the basis, the AVOL, the BDI and the LTGBY. Considering that our forecasts are built iteratively, the relative predictability of the regressors plays a fundamental role in the evolution of the rankings of log-GARCH-X models over the forecast horizon. The archetypal example being the log-GARCH-OVX whose performance deteriorates as the forecast horizon becomes more distant. Therefore, although different variables stand out at longer horizons, they are only more accurate predictors, empirically, insofar as their dynamics are less uncertain.

Table 9
21-day simple forecast combinations and dynamic model averaging.

	Squared return		Intraday range	
	MSPE	QLIKE	MSPE	QLIKE
Single-predictor log-GARCH-X models				
Mean	73,236	1,496	54,348	0,390
Trimmed mean	73,503	1,502	54,986	0,396
DMSPE	72,694	1,496	55,300	0,396
DMA	64,120	1,441	**45,168**	**0,283**
GARCH-MIDAS models				
Mean	65,166	1,505	55,157	0,557
Trimmed mean	65,253	1,516	55,496	0,573
DMSPE	65,014	1,507	55,433	0,559
DMA	64,206	1,440	54,012	1,568
Benchmarks + log-GARCH-X models + GARCH-MIDAS models				
Mean	72,459	1,479	54,397	0,390
Trimmed mean	69,739	1,477	53,351	0,396
DMSPE	68,828	1,469	53,044	0,396
DMA	**63,570**	**1,440**	45,289	0,665
AR-X-log-GARCH-X				
GETS	84,471	1,461	70,252	0,495

Notes: Bold characters signal models with the lowest information criteria.

Given the good performance of some log-GARCH-X models at the quarterly horizon, we could have expected, as a result, that the GETS would also produce relatively accurate forecasts. It is not the case. This is not specific to this horizon and the reader might feel disappointed with its overall performance. For instance, although several individual log-GARCH-X models rank in the first places at the 5-day, 21-day, and

[29] The general forecasting methodology of the regressors follows the procedure described in the first paragraph of Section 3.5

[30] Section 2.2 in the *Supplementary material* file, displays our results with a simpler alternative where the general ARIMA(63,2,63) model is replaced with an ARMA(5,5) specification.

Table 10
63-day ahead out-of-sample results.

	Squared return				Intraday range			
	MSPE	Rank	QLIKE	Rank	MSPE	Rank	QLIKE	Rank
Benchmarks								
GARCH(1,1)	71,389***	15	**1,520*****	1	60,806***	7	0,523***	4
GJR-GARCH(1,1)	69,490***	8	1,548***	11	61,306***	10	0,608***	12
EGARCH(1,1)	**68,224*****	1	1,543***	8	61,545***	11	0,636***	14
log-GARCH(1,1)	113,384***	22	1,576***	15	94,499***	22	0,501***	2
Single-predictor log-GARCH(1,1)-X								
AOI	68,860***	3	1,531***	5	60,185***	3	0,533***	7
AVOL	70,755***	12	1,547***	10	60,386***	5	0,525***	5
TB3M	72,374***	17	1,603***	16	63,546***	16	0,696***	17
LTGBY	69,058***	6	1,538***	7	60,636***	6	0,544***	9
Basis	70,609***	10	1,525***	3	59,663***	2	**0,495*****	1
S&P 500	69,699***	9	1,551***	12	60,926***	9	0,553***	10
MSCI EM	68,583***	2	1,524***	2	**59,536*****	1	0,517***	3
GSCI	2532,864*	28	4,347***	28	2464,094*	28	3,001***	28
BCOM	376,284**	26	2,475**	27	336,942***	26	1,384***	27
USD–AUD	74,463***	19	1,633***	19	62,987***	13	0,712***	19
USD–CAD	354,147***	25	1,774***	25	316,176***	25	0,940***	26
USD–CLP	388,616***	27	2,099***	26	384,027***	27	0,908***	25
USD–NZD	69,025***	5	1,544***	9	60,901***	8	0,555***	11
USD–ZAR	137,646***	23	1,529***	4	128,390***	24	0,529***	6
VIX	140,259***	24	1,697***	23	124,416***	23	0,614***	13
OVX	82,606***	20	1,573***	14	72,739***	20	0,669***	15
BDI	68,966***	4	1,533***	6	60,245***	4	0,535***	8
EPU	69,306***	7	1,675***	22	62,797***	12	0,817***	24
GARCH-MIDAS								
Prod	71,383***	14	1,657***	21	64,583***	18	0,775***	22
Kilian	71,192***	13	1,561***	13	63,131***	14	0,686***	16
Prod vol	72,805***	18	1,697***	24	65,914***	19	0,817***	23
Kilian vol	70,728***	11	1,619***	18	64,496***	17	0,751***	21
Working T	71,782***	16	1,643***	20	63,265***	15	0,724***	20
AR-X-log-GARCH-X								
GETS	107,439***	21	1,610***	17	93,382***	21	0,705***	18

Notes: Bold characters signal models with the lowest information criteria. *, **, and *** mean that the model is in $\widehat{M}^*_{90\%}$, $\widehat{M}^*_{75\%}$, and $\widehat{M}^*_{60\%}$, respectively.

63-day horizons with the intraday range, the GETS is far from the most accurate models. Several explanations can be put forward. An important point is that accuracy is dependent on the loss function and the proxy used for the true volatility. It is not rare that one model does well with one proxy or one loss function and not with others. It might explain to some extent why it is difficult for the GETS to rank high according to every measure. This is not completely satisfying, however, since results of DMA undermine this argument. Indeed, DMA applied to one of the two largest pools of models almost always rank first for every horizon, every volatility proxy, and every loss function.[31]

Arguably, the main explanation for the discrepancy between the performance of some single-predictor models and the GETS lies in the purpose of the statistical tools used within the procedure itself. These tools, and more crucially tests of statistical significance, have been identified in the literature to be key in assessing explanatory power, while measures of predictive power typically rely on out-of-sample metrics and their in-sample approximation (Shmueli, 2010). The fact that the GETS retains variables based on criteria that are primarily associated with the purpose of explanation rather than prediction can manifest in several forms. For instance, some variables that may at times hold relevant information for forecasting may not be statistically significant. A second example is the well-known bias–variance trade-off (Hastie et al., 2009, Ch. 7). As more variables are added to a model, its complexity rises. Predictive power first typically increases up to some point when it starts to deteriorate. And, although there are safeguards against over-fitting in the GETS procedure, it is not based on expected prediction error as should ideally be the case in a forecasting context.

[31] The only exceptions are for 5-day forecasts with squared return as our measure of volatility.

Table 11
63-day simple forecast combinations and dynamic model averaging.

	Squared return		Intraday range	
	MSPE	QLIKE	MSPE	QLIKE
Single-predictor log-GARCH-X models				
Mean	67,846	1,516	57,791	0,507
Trimmed mean	68,160	1,519	58,434	0,511
DMSPE	68,128	1,516	59,102	0,517
DMA	60,289	1,427	42,572	**0,423**
GARCH-MIDAS models				
Mean	70,771	1,611	63,471	0,711
Trimmed mean	70,748	1,625	63,700	0,735
DMSPE	70,492	1,611	63,444	0,713
DMA	60,235	1,427	45,605	1,680
Benchmarks + log-GARCH-X models + GARCH-MIDAS models				
Mean	75,435	1,497	61,866	0,480
Trimmed mean	68,586	1,512	58,936	0,515
DMSPE	68,022	1,496	58,363	0,498
DMA	**60,076**	**1,424**	**41,578**	2,564
AR-X-log-GARCH-X				
GETS	107,439	1,610	93,382	0,705

Notes: Bold characters signal models with the lowest information criteria.

In our baseline specification, we chose to select the forecasting models of our regressors within an ARIMA(63,2,63) specification. We made the choice of such a large specification mainly to improve forecasts of the OVX. Tables 2 to 7 in the *Supplementary material* file display our results when we use an alternative and more parsimonious ARMA(5,5) specification instead of the ARIMA(63,2,63). Indeed, we

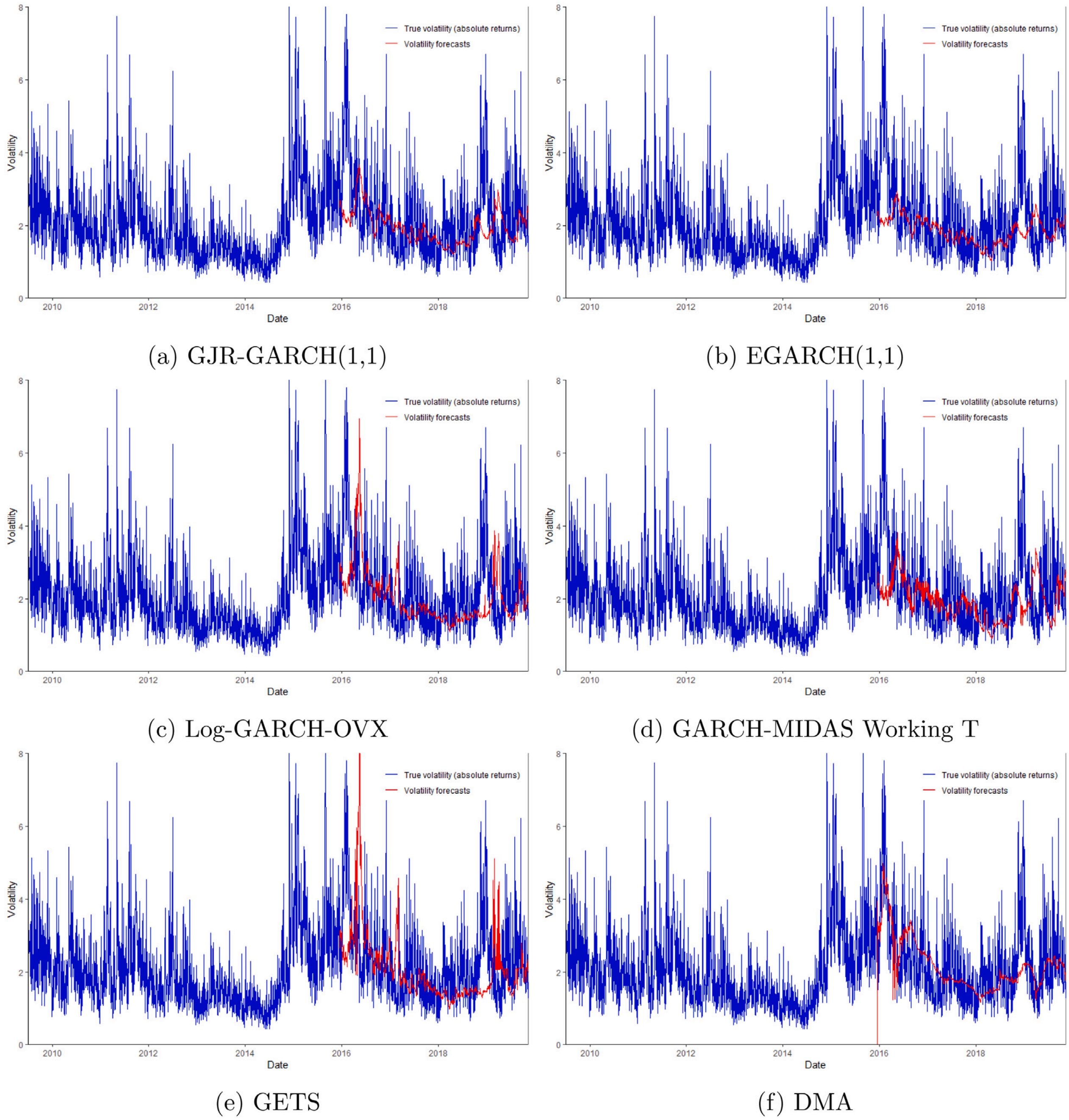
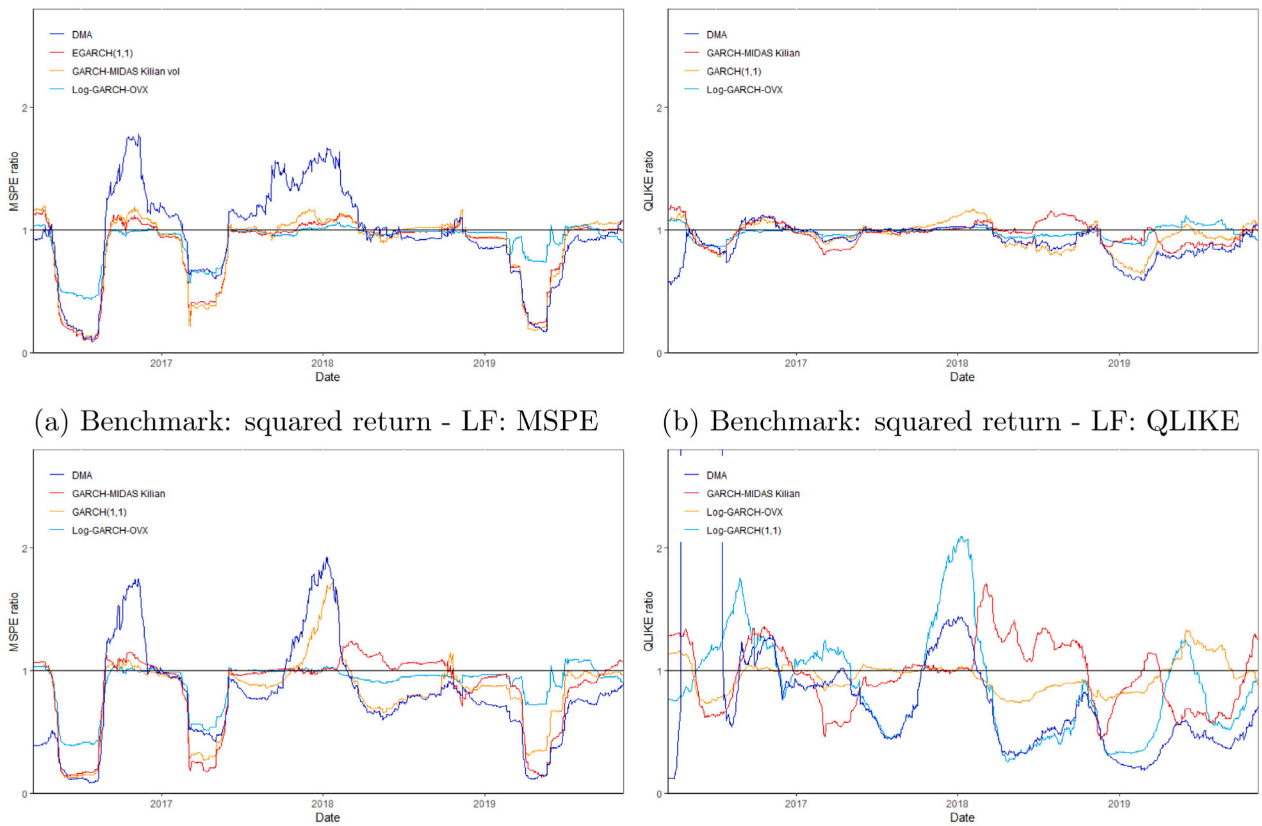


Fig. 8. 63-day volatility forecasts of selected models.

find that the log-GARCH-OVX greatly benefits from the flexibility of the ARIMA(63,2,63) at the 5-day horizon, and to a lower extent at the 21-day horizon. On the opposite, however, several other log-GARCH-X models produce more accurate forecasts with a more parsimonious specification of the regressors. This can be understood as a data-mining issue. The large range of possible models within an ARIMA(63,2,63) specification may lead to the selection of spurious models in-sample that do not translate into out-of-sample accuracy. It is not striking at the 5-day horizon where differences between the ARIMA(63,2,63) and the ARMA(5,5) are negligible, except for the GSCI and the BCOM, but it concerns an increasing number of models as the horizon increases and the downfall in accuracy of the baseline specification grows exponentially. In addition to the GSCI and the BCOM, it notably concerns the USD–CAD, the USD–CLP, the USD–ZAR, and the VIX at the monthly and quarterly horizons.

What are the consequences of the regressor forecast specifications for the GETS? At the weekly horizon, the improvement of OVX forecasts with the ARIMA(63,2,63) is sufficient to compensate for the deterioration of the quality of other forecasts. At the monthly horizon, it depends on the volatility proxy and on the loss function. While, finally, at the quarterly horizon, the parsimonious specification becomes more appealing.[32] These results suggest that there is room to improve prediction accuracy of the GETS by paying more attention to the optimal forecasting model of each regressor individually. Our general perception is that, even with such special care, the GETS will struggle to reach the performance of the DMA. By contrast, we find the effects

[32] With the ARMA(5,5) scheme, the GETS even ranks 1st with squared returns for both loss functions at the quarterly horizon.



(a) Benchmark: squared return - LF: MSPE (b) Benchmark: squared return - LF: QLIKE

(c) Benchmark: intraday range - LF: MSPE (d) Benchmark: intraday range - LF: QLIKE

Fig. 9. 63-day rolling window loss function ratio relative to the GETS.

of the regressor forecast specification to be less impactful on forecast combinations.

5.3. Alternative crude oil specification

As a robustness check, we replicate our forecast exercise for Brent crude oil. All results can be found in Section 4 of the *Supplementary material* file. They are overall consistent with our WTI baseline specification, although with a few notable differences. In-sample, the OVX is still the most significant variable in the volatility specification, and while it is not the same as for the WTI, the only other exogenous covariate is an exchange rate (USD–AUD).

Out-of-sample, at short horizons, the log-GARCH-OVX is still the most accurate log-GARCH-X models, but it is no longer the best model overall. This can easily be understood considering that the OVX is based on U.S. Oil Fund (USO), which aims to reflect daily changes in the near month WTI crude oil futures contract. Except for the log-GARCH(1,1), benchmarks often are the most accurate models with GARCH-MIDAS models just behind them.

At the 1-day ahead horizon, the GETS performs somewhat not as good as in the WTI specification. This is most likely directly linked to the fact that the OVX is less relevant for Brent crude oil than for the WTI. It still seems that both the GETS and the log-GARCH-OVX exhibit larger fluctuations than other models, and as such could be useful to capture peaks in crude oil volatility.

Finally, this robustness test corroborates all the main results of our baseline specification for longer horizon forecasts. Benchmarks and GARCH-MIDAS models perform relatively well at most horizons but their relative accuracy deteriorates as horizon increases. On the opposite, the performance of some single-predictor log-GARCH-X models drastically improve at the quarterly horizon. Furthermore, it overwhelmingly concerns the same predictors as in the WTI specification. Yet, again, this good performance does not translate into improved accuracy of the GETS. Lastly, DMA still shows impressive forecast accuracy at all horizons.

6. Conclusion

In this paper, we analysed the predictive power of a parsimonious AR-X-log-GARCH-X model, resulting from Sucarrat and Escribano's (2012) GETS procedure, for crude oil volatility. We compare the GETS to a number of standard benchmarks, single-predictor log-GARCH-X models, mixed-frequency GARCH-MIDAS models, as well as to forecast combinations using simple weighting schemes and Raftery et al.'s (2010) DMA method. Overall, the accuracy of the GETS is somewhat disappointing. While it fares relatively well for 1-day ahead forecast, predictability deteriorates at longer horizons. Two explanations can be proposed. First, the GETS algorithm is only useful as long as the pool of variables hold relevant information to forecast crude oil. Yet, for distant horizons, it is necessary to produce predictions of these explanatory variables which adds uncertainty to the model. Second, the GETS procedure is based on tests that are key in assessing explanatory power as opposed to reducing expected prediction error. A similar algorithm oriented toward reducing loss functions such as MSPE might result in better forecast accuracy. Still, the GETS can remain useful in a forecasting context as an input to a time-varying combination method such as DMA.

Furthermore, the combination of the GETS procedure and log-GARCH models provides a flexible environment for the analysis of the determinants of crude oil volatility. Our results confirm the role of implied volatility (OVX) as a somewhat preferred predictor. More interestingly, they also point to the information content of exchange rates. Because they adjust particularly fast to news, exchange rates of targeted countries may hold significant information not only for crude oil returns but also for its volatility (Chen et al., 2010). This finding calls for a more profound and targeted analysis of the information content of exchange rates to forecast commodity volatility in future research.

It is clear that the application of the GETS procedure to conditionally heteroskedastic models opens up new possibilities for research questions beyond forecasting. Related to this paper, a central question in finance is informational efficiency. According to its definition that is most relevant to economists, a market is efficient if "prices are right" (Barberis and Thaler, 2003). In the general stochastic discount factor (SDF) framework, the equilibrium expected return of an asset that requires no initial investment can be decomposed as the product of three terms: (i) the ratio between the conditional volatility of the SDF and its conditional mean, (ii) the conditional volatility of returns, and (iii) the conditional correlation between the SDF and returns (Cochrane, 2005). The GETS provides a parsimonious framework to investigate whether the same variables predict an asset returns and its volatility. We found in our IS investigation that no variable appears both in the mean and volatility specifications of the GETS' terminal model. Since first nearby futures contracts are expected to be closely related to spot, it is unlikely that the predictable component of crude oil return could be explained by the predictable component of its volatility. Given that the exposure of an asset to time-varying risk (i.e., the conditional correlation between the SDF and returns) is unlikely to be changing at the daily frequency, it confirms that one should focus on the variability of the SDF to test informational efficiency. This exercise could be expanded in future research to a larger pool of variables, as well as to different financial assets.

Declaration of competing interest

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

Appendix A

See Tables 6–11.
See Figs. 4–9.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eneco.2022.106059.

References

Aloui, R., Gupta, R., Miller, S.M., 2016. Uncertainty and crude oil returns. Energy Econ. 55, 92–100.

Andersen, T.G., Bollerslev, T., 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. Internat. Econom. Rev. 39 (4), 885–905.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H., 2001. The distribution of realized stock return volatility. J. Financ. Econ. 61 (1), 43–76.

Bahloul, W., Gupta, R., 2018. Impact of macroeconomic news surprises and uncertainty for major economies on returns and volatility of oil futures. Int. Econ. 156, 247–253.

Baker, S., Bloom, N., Davis, S., 2016. Measuring economic policy uncertainty. Q. J. Econ. 131 (4), 1593–1636.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. J. Econ. Perspect. 21 (2), 129–152.

Barberis, N., Thaler, R., 2003. A survey of behavioral finance. In: Constantinides, G., Harris, M., Stulz, R. (Eds.), Handbook of the Economics of Finance. Elsevier, pp. 1053–1128.

Baumeister, C., Kilian, L., 2016. Forty years of oil price fluctuations: Why the price of oil may still surprise us. J. Econ. Perspect. 30 (1), 139–160.

Baur, D.G., Dimpfl, T., 2018. The asymmetric return-volatility relationship of commodity prices. Energy Econ. 76, 378–387.

Bauwens, L., Hafner, C., Laurent, S., 2012. Volatility models. In: Handbook of Volatility Models and their Applications. John Wiley & Sons, Ltd, pp. 1–45.

Black, F., 1976. Studies of stock price volatility changes. In: Proceedings of the 1976 Meeting of the Business and Economic Statistics Section. American Statistical Association, Washington, pp. 177–181.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31 (3), 307–327.

Boudt, K., Peterson, B.G., Croux, C., 2008. Estimation and decomposition of downside risk for portfolios with non-normal returns. J. Risk 11 (2), 79–103.

Brockwell, P.J., Davis, R.A., 2006. Time Series: Theory and Methods, second ed. In: Springer Series in Statistics, Springer-Verlag, New York.

Carnero, M.A., Pérez, A., 2019. Leverage effect in energy futures revisited. Energy Econ. 82, 237–252.

Chantziara, T., Skiadopoulos, G., 2008. Can the dynamics of the term structure of petroleum futures be forecasted? Evidence from major markets. Energy Econ. 30 (3), 962–985.

Charles, A., Darné, O., 2017. Forecasting crude-oil market volatility: Further evidence with jumps. Energy Econ. 67 (C), 508–519.

Chen, Y.-c., Rogoff, K., Rossi, B., 2010. Can exchange rates forecast commodity prices? Q. J. Econ. 125 (3), 1145–1194.

Cheng, I.-H., Xiong, W., 2014. The financialization of commodity markets. Annu. Rev. Financ. Econ. 6, 419–441.

Chiarella, C., Kang, B., Nikitopoulos, C.S., Tô, T.-D., 2016. The return–volatility relation in commodity futures markets. J. Futures Mark. 36 (2), 127–152.

Cochrane, J.H., 2005. Asset Pricing. Princeton University Press.

Conrad, C., Loch, K., Rittler, D., 2014. On the macroeconomic determinants of long-term volatilities and correlations in U.S. stock and crude oil markets. J. Empir. Financ. 29 (C), 26–40.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. J. Financ. Econom. 7 (2), 174–196.

Da, Z., Engelberg, J., Gao, P., 2015. The sum of all FEARS investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32.

Delatte, A.-L., Lopez, C., 2013. Commodity and equity markets: Some stylized facts from a copula approach. J. Bank. Financ. 37 (12), 5346–5356.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13 (3), 253–263.

Doornik, J.A., 2009. Autometrics. In: The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry. Oxford University Press, pp. 88–121.

Doornik, J.A., Hendry, D.F., 2013. Empirical Econometric Modelling - PcGive 14 Vol. I. Timberlake Consultant Ltd, London.

Elder, J., Serletis, A., 2010. Oil price uncertainty. J. Money Credit Bank. 42 (6), 1137–1159.

Engle, R.F., Ghysels, E., Sohn, B., 2013. Stock market volatility and macroeconomic fundamentals. Rev. Econ. Stat. 95 (3), 776–797.

Fama, E.F., French, K.R., 1987. Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. J. Bus. 60 (1), 55–73.

Fama, E.F., French, K.R., 1988. Dividend yields and expected stock returns. J. Financ. Econ. 22 (1), 3–25.

Francq, C., Sucarrat, G., 2017. An equation-by-equation estimator of a multivariate log-GARCH-X model of financial returns. J. Multivariate Anal. 153, 16–32.

Francq, C., Sucarrat, G., 2018. An exponential chi-squared QMLE for log-GARCH models via the ARMA representation. J. Financ. Econom. 16 (1), 129–154.

Francq, C., Wintenberger, O., Zakoïan, J.-M., 2013. GARCH models without positivity constraints: Exponential or log GARCH? J. Econometrics 177 (1), 34–46.

Geman, H., Ohana, S., 2009. Forward curves, scarcity and price volatility in oil and natural gas markets. Energy Econ. 31 (4), 576–585.

Geman, H., Smith, W.O., 2013. Theory of storage, inventory and volatility in the LME base metals. Resour. Policy 38 (1), 18–28.

Geweke, J., 1986. Modelling the persistence of conditional variance: A comment. Econometric Rev. 5 (1), 57–61.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS Touch: Mixed Data Sampling Regression Models. CIRANO Working Paper, CIRANO.

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

Gong, X., Lin, B., 2018. The incremental information content of investor fear gauge for volatility forecasting in the crude oil futures market. Energy Econ. 74, 370–386.

Gospodinov, N., Ng, S., 2013. Commodity prices, convenience yields, and inflation. Rev. Econ. Stat. 95 (1), 206–219.

Granger, C., Timmermann, A., 1999. Data mining with local model specification uncertainty: A discussion of Hoover and Perez. Econom. J. 2 (2), 220–225.

Groen, J.J.J., Pesenti, P.A., 2011. Commodity prices, commodity currencies, and global economic developments. In: NBER Chapters. National Bureau of Economic Research, Inc, pp. 15–42.

Hakkio, C.S., Keeton, W.R., 2009. Financial stress: What is it, how can it be measured, and why does it matter? Econ. Rev. 94, 5–50.

Hansen, P.R., Lunde, A., 2006. Consistent ranking of volatility models. J. Econometrics 131 (1), 97–121.

Hansen, P.R., Lunde, A., Nason, J.M., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Harvey, A.C., 2013. Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series. Cambridge University Press.

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. Int. J. Forecast. 13 (2), 281–291.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer.
Hendry, D.F., Krolzig, H.-M., 1999. Improving on 'data mining reconsidered' by K.D. Hoover and S.J. Perez. Econom. J. 2 (2), 202–219.
Hendry, D.F., Krolzig, H.-M., 2005. The properties of automatic GETS modelling. Econ. J. 115 (502).
Herrera, A.M., Hu, L., Pastor, D., 2018. Forecasting crude oil price volatility. Int. J. Forecast. 34 (4), 622–635.
Hoover, K., Perez, S., 1999. Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. Econom. J. 2 (2), 167–191.
Jurado, K., Ludvigson, S.C., Ng, S., 2015. Measuring uncertainty. Amer. Econ. Rev. 105 (3), 1177–1216.
Kaldor, N., 1939. Speculation and economic stability. Rev. Econom. Stud. 7 (1), 1–27.
Kang, S.H., Yoon, S.-M., 2013. Modeling and forecasting the volatility of petroleum futures prices. Energy Econ. 36, 354–362.
Kilian, 2009. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. Amer. Econ. Rev. 99 (3), 1053–1069.
Kilian, L., Murphy, D.P., 2014. The role of inventories and speculative trading in the global market for crude oil. J. Appl. Econometrics 29 (3), 454–478.
Kristoufek, L., 2014. Leverage effect in energy futures. Energy Econ. 45, 1–9.
Lin, Y., Xiao, Y., Li, F., 2020. Forecasting crude oil price volatility via a HM-EGARCH model. Energy Econ. 87 (C).
Litzenberger, R.H., Rabinowitz, N., 1995. Backwardation in oil futures markets: Theory and empirical evidence. J. Finance 50 (5), 1517–1545.
Liu, J., Wei, Y., Ma, F., Wahab, M.I.M., 2017. Forecasting the realized range-based volatility using dynamic model averaging approach. Econ. Model. 61, 12–26.
Luo, J., Ji, Q., Klein, T., Todorova, N., Zhang, D., 2020. On realized volatility of crude oil futures markets: Forecasting with exogenous predictors under structural breaks. Energy Econ. 89 (C).
Lv, W., 2018. Does the OVX matter for volatility forecasting? Evidence from the crude oil market. Physica A 492, 916–922.
Ma, Y.-R., Ji, Q., Wu, F., Pan, J., 2021. Financialization, idiosyncratic information and commodity co-movements. Energy Econ. 94, 105083.
Ma, F., Liu, J., Wahab, M.I.M., Zhang, Y., 2018. Forecasting the aggregate oil price volatility in a data-rich environment. Econ. Model. 72, 320–332.
Manera, M., Nicolini, M., Vignati, I., 2016. Modelling futures price volatility in energy markets: Is there a role for financial speculation? Energy Econ. 53, 220–229.
McAleer, M., 2005. Automated inference and learning in modeling financial volatility. Econom. Theory 21 (1), 232–261.
Meng, F., Liu, L., 2019. Analyzing the economic sources of oil price volatility: An out-of-sample perspective. Energy 177, 476–486.
Milhøj, A., 1987. A Multiplicative Parameterization of ARCH Models. Research Report 101, University of Copenhagen: Institute of Statistics.
Mincer, J., Zarnowitz, V., 1969. The evaluation of economic forecasts. In: Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance. National Bureau of Economic Research, New York, pp. 3–46.
Müller, U.A., Dacorogna, M.M., Davé, R.D., Olsen, R.B., Pictet, O.V., von Weizsäcker, J.E., 1997. Volatilities of different time resolutions — Analyzing the dynamics of market components. J. Empir. Financ. 4 (2), 213–239.
Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59 (2), 347–370.
Ng, V.K., Pirrong, S.C., 1994. Fundamentals and volatility: Storage, spreads, and the dynamics of metals prices. J. Bus. 67 (2), 203–230.
Nomikos, N.K., Pouliasis, P.K., 2011. Forecasting petroleum futures markets volatility: The role of regimes and market conditions. Energy Econ. 33 (2), 321–337.
Pan, Z., Wang, Y., Wu, C., Yin, L., 2017. Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model. J. Empir. Financ. 43 (C), 130–142.
Pantula, S.G., 1986. Modelling the persistence of conditional variance: A comment. Econometric Rev. 5 (1), 71–74.
Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. J. Bus. 53 (1), 61–65.
Patton, A., 2011. Volatility forecast comparison using imperfect volatility proxies. J. Econometrics 160 (1), 246–256.
Pindyck, R.S., 2004. Volatility and commodity price dynamics. J. Futures Mark. 24 (11), 1029–1047.
Pretis, F., Reade, J.J., Sucarrat, G., 2018. Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. J. Stat. Softw. 86 (1), 1–44.
Prokopczuk, M., Symeonidis, L., Simen, C.W., 2016. Do jumps matter for volatility forecasting? Evidence from energy markets. J. Futures Mark. 36 (8), 758–792.
Raftery, A.E., Kárný, M., Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. Technometrics 52 (1), 52–66.
Sadorsky, P., 2006. Modeling and forecasting petroleum futures volatility. Energy Econ. 28 (4), 467–488.
Sadorsky, P., McKenzie, M.D., 2008. Power transformation models and volatility forecasting. J. Forecast. 27 (7), 587–606.
Schwert, G.W., 1989. Why does stock market volatility change over time?. J. Finance 44 (5), 1115–1153.
Sévi, B., 2014. Forecasting the volatility of crude oil futures using intraday data. European J. Oper. Res. 235 (3), 643–659.
Shmueli, G., 2010. To explain or to predict?. Statistical Science 25 (3), 289–310.
Sockin, M., Xiong, W., 2015. Informational frictions and commodity markets. J. Finance 70 (5), 2063–2098.
Sucarrat, G., Escribano, A., 2012. Automated model selection in finance: General-to-specific modelling of the mean and volatility specifications. Oxford Bull. Econ. Statist. 74 (5), 716–735.
Sucarrat, G., Escribano, A., 2018. Estimation of log-GARCH models in the presence of zero returns. Eur. J. Finance 24 (10), 809–827.
Sucarrat, G., Grønneberg, S., Escribano, A., 2016. Estimation and inference in univariate and multivariate log-GARCH-X models when the conditional density is unknown. Comput. Statist. Data Anal. 100, 582–594.
Tang, K., Xiong, W., 2012. Index investment and the financialization of commodities. Financ. Anal. J. 68 (6), 54–74.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier, pp. 135–196.
Wang, Y., Wu, C., 2012. Energy prices and exchange rates of the U.S. dollar: Further evidence from linear and nonlinear causality analysis. Econ. Model. 29 (6), 2289–2297.
Wei, Y., Liu, J., Lai, X., Hu, Y., 2017. Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? Energy Econ. 68 (C), 141–150.
Wintenberger, O., 2013. Continuous invertibility and stable QML estimation of the EGARCH(1,1) model. Scand. J. Stat. 40 (4), 846–867.
Working, H., 1949. The theory of price of storage. Amer. Econ. Rev. 39 (6), 1254–1262.
Working, H., 1960. Speculation on hedging markets. Food Res. Inst. Stud. 01 (2).
Yang, C., Gong, X., Zhang, H., 2019. Volatility forecasting of crude oil futures: The role of investor sentiment and leverage effect. Resour. Policy 61, 548–563.
Zhang, Y., Wei, Y., Zhang, Y., Jin, D., 2019. Forecasting oil price volatility: Forecast combination versus shrinkage method. Energy Econ. 80, 423–433.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.