International Journal of Forecasting 34 (2018) 622–635

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting crude oil price volatility

Ana María Herrera [a], Liang Hu [b,*], Daniel Pastor [c]

[a] Department of Economics, University of Kentucky, United States
[b] Department of Economics, Wayne State University, United States
[c] Department of Economics and Finance, University of Texas at El Paso, United States

ARTICLE INFO

Keywords:
Crude oil price volatility
GARCH models
Long memory
Markov switching
Volatility forecast
Realized volatility

ABSTRACT

We use high-frequency intra-day realized volatility data to evaluate the relative forecasting performances of various models that are used commonly for forecasting the volatility of crude oil daily spot returns at multiple horizons. These models include the RiskMetrics, GARCH, asymmetric GARCH, fractional integrated GARCH and Markov switching GARCH models. We begin by implementing Carrasco, Hu, and Ploberger's (2014) test for regime switching in the mean and variance of the GARCH(1, 1), and find overwhelming support for regime switching. We then perform a comprehensive out-of-sample forecasting performance evaluation using a battery of tests. We find that, under the MSE and QLIKE loss functions: (i) models with a Student's t innovation are favored over those with a normal innovation; (ii) RiskMetrics and GARCH(1, 1) have good predictive accuracies at short forecast horizons, whereas EGARCH(1, 1) yields the most accurate forecasts at medium horizons; and (iii) the Markov switching GARCH shows a superior predictive accuracy at long horizons. These results are established by computing the equal predictive ability test of Diebold and Mariano (1995) and West (1996) and the model confidence set of Hansen, Lunde, and Nason (2011) over the entire evaluation sample. In addition, a comparison of the MSPE ratios computed using a rolling window suggests that the Markov switching GARCH model is better at predicting the volatility during periods of turmoil.

Published by Elsevier B.V. on behalf of International Institute of Forecasters.

1. Introduction

Over recent months, newspaper headlines such as "Oil prices will be much more volatile in 2017: IEA" (Reuters, January 15, 2017) and "IEA sees risk of volatile oil prices on weak upstream investment" (Bloomberg, September 17, 2017) have been evidence of the concerns voiced by the International Energy Agency regarding the return of high volatility in crude oil markets. This time around, the apprehension regarding a higher volatility seems to have stemmed from the slow pace of investment in new production. Nevertheless, surges in the volatility of the daily West Texas Intermediate (WTI) spot returns were observed around the 1986 oil price collapse, during the Gulf War, following the onset of the 2007–2008 financial crisis, and more recently since the fall in oil prices that started in July 2014 (see Fig. 1). Clearly, periods of heightened volatility in crude oil markets are recurrent, and these headlines manifest the importance of evaluating whether the econometric tools that are available to practitioners are able to generate reliable forecasts of the volatility of crude oil pricces.

Since the "spot oil price volatility reflects the volatility of current as well as future values of [oil] production, consumption and inventory demand" (Pindyck, 2004), it is relevant for various economic agents. Accurate forecasts are key for firms whose business depends heavily on oil prices; for instance, oil companies that need to decide whether to drill a new well (Kellogg, 2014) or to undertake long-term investments in their refining and transportation infrastructure, airline companies who use oil price forecasts to set airfares, and the automobile industry. Second,

* Corresponding author.
E-mail address: lianghu@wayne.edu (L. Hu).

https://doi.org/10.1016/j.ijforecast.2018.04.007
0169-2070/Published by Elsevier B.V. on behalf of International Institute of Forecasters.

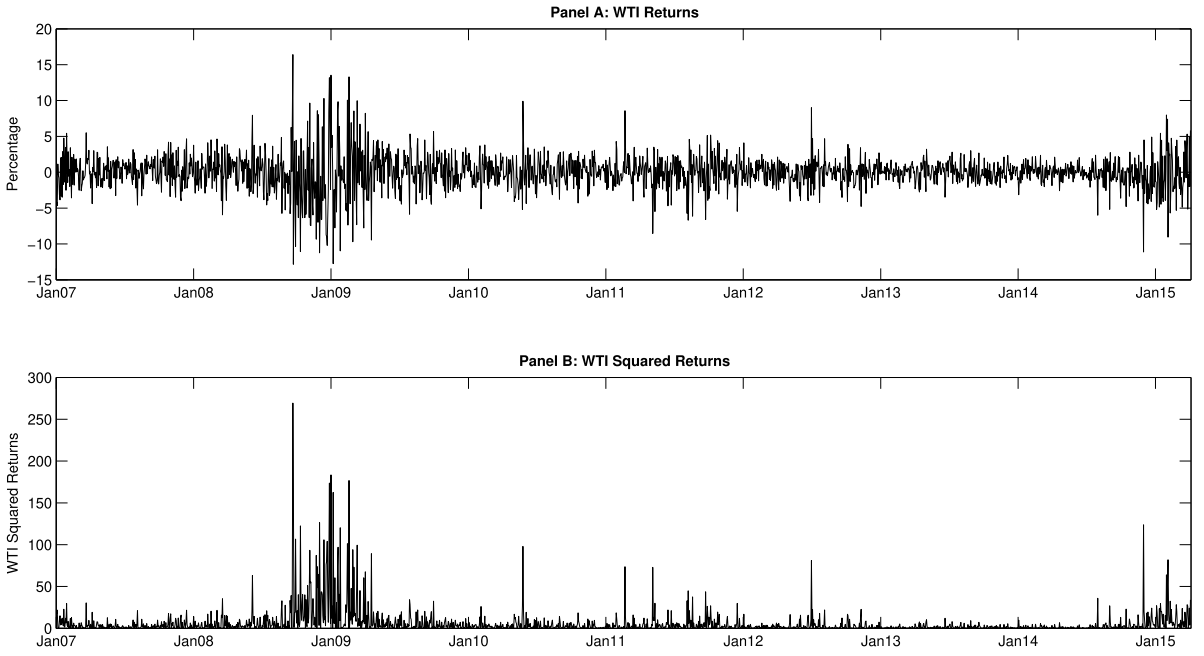

Fig. 1. Daily WTI crude oil returns and squared returns. The sample period extends from January 3, 2007, to April 2, 2015.

the oil price volatility also plays a role in households' decisions regarding their purchases of durable goods (Davis & Kilian, 2011; Kahn, 1986). Lastly, they are useful for agents whose daily task is to produce forecasts of industry-level and aggregate economic activities, such as policy makers, business economists, and private sector forecasters (see e.g. Elder & Serletis, 2010; Jo, 2014).

The aim of this paper is to evaluate the out-of-sample forecasting performances of different volatility models for the conditional variance (hereafter variance) of spot crude oil returns, where we proxy the unobserved variance with the realized volatility of intra-day returns (Andersen & Bollerslev, 1998). More specifically, we investigate the predictive abilities of the RiskMetrics, GARCH, asymmetric GARCH, Fractionally Integrated GARCH (FIGARCH) and Markov switching GARCH (MS-GARCH) models. The motivation for choosing these models is as follows. RiskMetrics remains a very popular empirical model among practitioners. Meanwhile, GARCH (Bollerslev, 1986) sets out the idea of modeling and forecasting the volatility as a time-varying function of currently available information. On the empirical side, the GARCH(1,1) model has also fared well in predicting the conditional volatility of financial assets (Hansen & Lunde, 2005) and the crude oil price volatility (see Xu & Ouenniche, 2012, and references therein). Asymmetric GARCH models such as EGARCH (Nelson, 1991) and GJR-GARCH (Glosten, Jagannathan, & Runkle, 1993) have been shown to have good out-of-sample performances when forecasting the oil price volatility one step ahead (Hou & Suardi, 2012; Mohammadi & Su, 2010). As for Markov switching models, they have been found to be more suitable for modeling situations in which changes in regimes are triggered by sudden shocks to the economy. Thus, they might have good predictive abilities for spot crude oil returns, which are characterized by sudden jumps, due to factors such as political disruptions in the Middle East or military interventions in oil-exporting countries, for instance. However, regime switching and long memory are related intimately, and it is hard to differentiate a Markov switching model from a long memory model (Diebold & Inoue, 2001). Therefore, we add the FIGARCH to our pool of models for forecast evaluation.

We provide a comprehensive study of the relative out-of-sample forecasting performances at multiple horizons. We start by testing formally for regime switches using the procedure proposed by Carrasco, Hu, and Ploberger (2014), then evaluate the directional accuracy using Pesaran and Timmermann's (1992) test. Furthermore, we conduct pairwise comparisons between different candidate models using Diebold and Mariano (1995) and West's (1996) test of equal predictive ability. In addition, we also employ Hansen, Lunde, and Nason's (2011) model confidence set procedure in order to determine the best (set of) model(s) from the pool. All of the tests are reported under two loss functions: the mean square error, MSE, and the quasi likelihood, QLIKE. We also inquire into the stability of the forecasting accuracy of the preferred models over the evaluation period (2013–2014).

Our findings can be summarized as follows: (i) the Student's t distribution is generally favored in the parametric models due to the extremely high kurtosis in the oil return volatility; (ii) the nonparametric model (RiskMetrics) and parsimonious models like GARCH(1,1) perform better at short (1- and 5-day) horizons; (iii) the EGARCH stands out at the 21-day horizon; (iv) the MS-GARCH model yields more accurate forecasts at the longer 63-day horizon; and (v) the MS-GARCH model has a higher predictive ability during periods of turmoil.

We are not the first to consider Markov switching models for forecasting the volatility of the crude oil market. For

example, Fong and See (2002) and Nomikos and Pouliasis (2011) both apply MS-GARCH to forecasting the volatility of crude oil futures and evaluate the out-of-sample forecasts at the one-day horizon. Wang, Wu, and Yang (2016) study the volatility of spot returns by comparing the forecasting performance of the Markov switching multifractal volatility model (Calvet & Fisher, 2001) vis-à-vis those of a set of GARCH-class models. Alternatively, Arouri, Lahiani, Lévy, and Nguyen (2012) discover that accounting for structural breaks and long memory in the GARCH specifications leads to gains in forecasting the conditional volatilities of spot and futures oil prices. Our paper clearly benefits from this literature, but also differs in several ways. Specifically, the MS-GARCH specification in this paper allows for a considerable degree of flexibility in modeling the persistence and regime switches. The adopted estimation method not only facilitates the calculation of the multi-step-ahead forecast, but also makes a more efficient use of the information that is contained in the data. We also employ an accurate proxy for the underlying volatility (the realized volatility instead of squared returns) and investigate the forecasting stability over time.

The remainder of the paper is organized as follows. Section 2 introduces the econometric models used for estimating and forecasting the oil price returns and volatility. Section 3 describes the data. The in-sample estimation results are reported in Section 4. The out-of-sample forecast evaluation follows; and the last section concludes.

2. Model specifications

This section briefly describes the parametric models that are used widely by practitioners for modeling and forecasting the oil price volatility.

2.1. Standard GARCH models

The conventional GARCH models considered in this paper comprise the GARCH (Bollerslev, 1986), EGARCH (Nelson, 1991), and GJR-GARCH (Glosten et al., 1993) models. The GARCH(1,1) is given by

$$\begin{cases} y_t = \mu_t + \varepsilon_t, \\ \varepsilon_t = \sqrt{h_t} \cdot \eta_t, \ \ \eta_t \sim iid(0,1) \\ h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 h_{t-1}, \end{cases}$$

where μ_t is the time-varying conditional mean that may be given by $\boldsymbol{\beta}'\mathbf{x}_t$, with $\mathbf{x}_t$ being the $k \times 1$ vector of stochastic covariates and $\boldsymbol{\beta}$ being a $k \times 1$ vector of parameters that is to be estimated. α_0, α_1, and γ_1 are all positive, and $\alpha_1 + \gamma_1 \leq 1$.

For the exponential GARCH (EGARCH) model, the logarithm of the conditional variance is defined as

$$\log(h_t) = \alpha_0 + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| - \mathbb{E} \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| \right) + \xi \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \log(h_{t-1}).$$

As for the GJR-GARCH, the conditional variance is given by

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \xi \varepsilon_{t-1}^2 \cdot I_{\{\varepsilon_{t-1} < 0\}} + \gamma_1 h_{t-1},$$

where $I_{\{\omega\}}$ is an indicator function that is equal to one if $\varepsilon_{t-1} < 0$, and zero otherwise.

2.2. MS-GARCH

When using GARCH models for estimating the conditional variance of economic or financial series, a common finding is that the persistence level is very high. Lamoureux and Lastrapes (1990) show that this may be due to neglected structural breaks or regime changes. In addition, Caporale, Pittis, and Spagnolo (2003) demonstrate via Monte Carlo studies that fitting (misspecified) GARCH models to data generated by a MS-GARCH process tends to produce integrated GARCH (IGARCH) parameter estimates, leading to erroneous conclusions about the persistence levels.

Oil prices are characterized by sudden changes in volatility due to factors such as political disruptions in the Middle East, military interventions in oil exporting countries or depressed aggregate demand following the financial crisis, for instance. As a consequence, standard GARCH models that ignore these sudden changes are likely to be misspecified. Therefore, we also consider the MS-GARCH(1,1), which is specified as follows:

$$\begin{cases} y_t = \mu^{S_t} + \varepsilon^{S_t}, \\ \varepsilon^{S_t} = \sqrt{h_t^{S_t}} \cdot \eta_t, \ \ \eta_t \sim iid(0,1) \\ h_t^{S_t} = \alpha_0^{S_t} + \alpha_1^{S_t} \varepsilon_{t-1}^2 + \gamma_1^{S_t} h_{t-1}, \end{cases} \tag{1}$$

where both the conditional mean μ^{S_t} and the conditional variance $h_t^{S_t}$ are subject to a hidden Markov chain, S_t. We assume a two-state first-order Markov chain so that the transition probability of the current state, S_t, depends only on the most adjacent past state, S_{t-1}:

$$P(S_t \mid S_{t-1}, \mathcal{I}_{t-2}) = P(S_t \mid S_{t-1}),$$

where $\mathcal{I}_{t-2}$ denotes the information set up to $t-2$. The transition probability that state i is followed by state j is denoted by p_{ij}. S_t takes on two values (1, 2) and has transition probabilities $p_{11} = P(S_t = 1 \mid S_{t-1} = 1)$ and $p_{22} = P(S_t = 2 \mid S_{t-1} = 2)$. S_t is geometric ergodic if $0 < p_{11} < 1$ and $0 < p_{22} < 1$.

2.3. FIGARCH

As was noted earlier, IGARCH behavior has been reported widely in the empirical literature on asset returns, commodity prices and exchange rates, especially at a daily frequency. The effect of any shock to the IGARCH volatility process will persist for an infinite horizon. However, this does not seem compatible with the persistence that is observed after large shocks, such as the global financial crisis in Fig. 1. Baillie, Bollerslev, and Mikkelsen (1996) argue that IGARCH may be a mathematical artifact of a mean-reverting long-memory FIGARCH instead. In fact, it is well documented in the literature that Markov switching and long memory are intimately related to each other. Diebold and Inoue (2001) point out that it is easy to confuse Markov switching with long memory, even asymptotically. Granger and Hyung (2004) show that occasional structural breaks also generate long memory, which is hard to distinguish from fractional integration. In addition, Hsu (2001) proves that the presence of long memory in time series may result

in the spurious detection of change-points.[1] Therefore, we also include FIGARCH in the set of volatility models to be evaluated.

We consider the FIGARCH(1, d, 1) of Baillie et al. (1996), which relies on the ARMA representation of ε_t^2 and takes the following form:

$$\phi(L)(1-L)^d\varepsilon_t^2 = \alpha_0 + (1-\gamma_1 L)w_t,$$

where $w_t = \varepsilon_t^2 - h_t$, $\phi(L) = (1-(\alpha_1+\gamma_1)L)/(1-L) \equiv (1-\phi L)/(1-L)$. d is the fractional differencing parameter and $0 < d \leq 1$. Hence, the conditional variance has the representation:

$$h_t = \alpha_0 + \gamma_1 h_{t-1} + [1-\gamma_1 L - (1-\phi L)(1-L)^d]\varepsilon_t^2.$$

3. Data description

Our measure of crude oil prices is the daily spot price for the West Texas Intermediate (WTI) crude oil, obtained from the U.S. Energy Information Administration. The sample covers the period from January 2, 2007, to April 2, 2015, which saw the rapid growth in oil production following the fracking revolution, the large upswing in oil prices during the economic expansion of the early 2000s, the downswing following the 2008–2009 global financial crisis, and the sharp decline since the second semester of 2014. We model crude oil returns and their volatility through the calculation of daily returns by taking 100 times the difference in the logarithm of consecutive days' closing spot prices.

The evaluation of the forecasting performances of different models requires a measure of the true underlying volatility. Since the true volatility of crude oil returns is unobserved, we use an estimated measure of the realized volatility as a proxy. More specifically, we obtain 5-minute prices of 1-month WTI oil futures contracts series from TickData.com over the period between January 2, 2007, and April 2, 2015.[2] These contracts are traded around the clock, with the exception of a 45-minute trading halt from 5:15pm to 6:00pm EST, Sunday through Friday, excluding market holidays. We construct the daily realized volatility RV_t by summing the squared 5-minute returns over all of the trading hours,[3] then calculate the m-step-ahead realized volatility at time T by simply summing the daily realized volatility over m days, denoted by:

$$\widehat{\mathrm{RV}}_{T,T+m} = \sum_{j=1}^{m} \widehat{\mathrm{RV}}_{T+j}.$$

Table 1 reports the summary statistics for the WTI rates of return, the $\mathrm{RV}_t^{1/2}$ and the logarithm of $\mathrm{RV}_t^{1/2}$. The mean rate for the WTI returns is -0.010, with a standard deviation of 2.426. Note that the WTI returns are slightly positively skewed. The kurtosis is equal to 8.491, which is high compared to 3 for a normal distribution.[4] The $\mathrm{RV}_t^{1/2}$ series is severely right-skewed and leptokurtic. However, the logarithmic series is less skewed, with a kurtosis close to 3.

Fig. 1 plots the returns and squared returns of the WTI spot prices over the sample period. Two salient characteristics of WTI crude returns are apparent in the figure. First, crude oil returns are characterized mostly by periods of low (high) volatility followed by low (high) volatility. GARCH models are designed to capture this volatility clustering. Second, we observe exceptionally large variations in the WTI returns during the global financial crisis in late 2008 and since crude oil prices started decreasing in July 2014. In other words, periods of low volatility may be followed by periods of elevated volatility in the face of major political or financial unrest. This behavior supports the use of MS-GARCH models, where the GARCH parameters are allowed to switch between two regimes according to a Markov chain.

4. In-sample estimation

This section describes the estimation methods used and discusses the in-sample estimation results for the parametric models.

4.1. Estimation methods

Estimation of the GARCH-family and FIGARCH models is standard, and is conducted via maximum likelihood.[5] Thus, we restrict our discussion here to the estimation of the MS-GARCH model in Eq. (1), which is computationally intractable because the conditional variance h_t depends on the state-dependent h_{t-1}, and consequently on all past states. In other words, computing the likelihood function is infeasible, as it requires us to integrate out all possible unobserved regime paths, the number of which grows exponentially with the sample size T. Therefore, we estimate

[1] We thank an anonymous referee for bringing this issue to our attention.

[2] Andersen and Bollerslev (1998) note that squared daily returns are a noisy proxy of the true volatility, and that this noise can lead to improper conclusions about the forecasting abilities of GARCH-type models. Andersen, Bollerslev, Christoffersen, and Diebold (2006) establish the theoretical justification for the realized volatility as an accurate measure of the underlying volatility. Liu, Patton, and Sheppard (2012) among others, also find that the 5-minute sampling frequency outperforms most other realized volatility measures across multiple asset classes.

[3] For markets in which futures are not traded around the clock, Blair, Poon, and Taylor (2001) suggest that the measure of the daily realized volatility be constructed by summing the 5-minute returns during the trading hours and then adding the square of the previous "overnight" return. Hansen and Lunde (2005) propose an alternative way of measuring the daily realized volatility. They first calculate the constant $\widehat{c} = [n^{-1}\sum_{t=1}^{n}(r_t-\widehat{\mu})^2]/[n^{-1}\sum_{t=1}^{n} rv_t]$, where r_t and $\widehat{\mu}$ are the close-to-close return of the daily prices and the mean respectively, and rv_t is the 5-minute realized volatility during the trading hours only. They then scale the realized volatility rv_t by the constant $\widehat{c}$. This measure is less noisy than that of Blair et al. (2001). During our sample period, crude oil futures are traded almost continually through the day, with the exception of the 45-minute gap between 5:15 and 6:00 p.m. EST. We tried scaling, and it turns out that our results are robust to scaling for the daily 45-minute interval when trading is halted.

[4] These numbers are consistent with those of previous studies by, e.g., Abosedra and Laopodis (1997), Morana (2001), and Bina and Vo (2007), among others.

[5] Details of the log likelihood functions and estimation can be found in Section A of the online appendix.

Table 1
Descriptive statistics.

WTI returns						
Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
−0.010	2.426	−12.827	16.414	5.887	0.055	8.491
$RV^{1/2}$						
Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
0.020	0.012	0.004	0.184	0.00014	3.207	26.494
$\ln(RV^{1/2})$						
Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
−4.027	0.469	−5.457	−1.692	0.220	0.553	3.608

Note: WTI returns denotes the log difference of the West Texas Intermediate daily spot closing price. *RV* denotes the realized volatility computed from the 5-minute returns on oil futures. The WTI returns, $RV^{1/2}$, and the natural logarithm of the $RV^{1/2}$ series are from the sample period of January 3, 2007, to April 2, 2015, with 2079 observations.

the MS-GARCH model by following Klaassen (2002)[6] and replacing h_{t-1} with its expectation, conditional on the information set at $t-1$ and the current state variable, namely

$$h_t^{(i)} = \alpha_0^{(i)} + \alpha_1^{(i)}\varepsilon_{t-1}^2 + \gamma_1^{(i)}\mathbb{E}_{t-1}\left[h_{t-1} \mid S_t = i\right], \tag{2}$$

where

$$\mathbb{E}_{t-1}\left[h_{t-1} \mid S_t = i\right] = \sum_{j=1}^{2} P\left(S_{t-1} = j \mid S_t = i, \mathcal{I}_{t-1}\right) h_{t-1}^{(j)},$$
$$i, j = 1, 2.$$

The specification in Eq. (2) circumvents the path dependence by integrating out h_{t-1}. Because the conditional variance depends only on the current state S_t, the estimation and computation of the forecasts are straightforward.[7] Indeed, the m-step-ahead volatility forecast at time T is calculated through a recursive procedure as follows:

$$\hat{h}_{T,T+m} = \sum_{\tau=1}^{m} \hat{h}_{T,T+\tau} = \sum_{\tau=1}^{m}\sum_{i=1}^{2} P(S_{T+\tau} = i \mid \mathcal{I}_T)\hat{h}_{T,T+\tau}^{(i)},$$

where the τ-step-ahead volatility forecast in regime i made at time T is given by

$$\hat{h}_{T,T+\tau}^{(i)} = \alpha_0^{(i)} + \left(\alpha_1^{(i)} + \gamma_1^{(i)}\right)\mathbb{E}_T\left[h_{T,T+\tau-1}^{(i)} \mid S_{T+\tau} = i\right].$$

Note that the necessary conditions for second-order stationarity, which follow from the work of Klaassen (2002), are:

$$p_{11}(\alpha_1^{(1)} + \gamma_1^{(1)}) < 1,\ p_{22}(\alpha_1^{(2)} + \gamma_1^{(2)}) < 1,$$

and

$$p_{11}(\alpha_1^{(1)} + \gamma_1^{(1)}) + p_{22}(\alpha_1^{(2)} + \gamma_1^{(2)}) + (1 - p_{11} - p_{22})(\alpha_1^{(1)} + \gamma_1^{(1)})(\alpha_1^{(2)} + \gamma_1^{(2)}) < 1.$$

Abramson and Cohen (2007) further show that these conditions are not only necessary, but also sufficient.[8] It is easy to see that these conditions do not require stationarity within each regime. For example, regime 1 could be nonstationary, or even slightly explosive (e.g. $\alpha_1^{(1)} + \gamma_1^{(1)} \geq 1$), as long as the probability of staying in regime 1 (p_{11}) is small. Thus, the MS-GARCH model allows for a great flexibility when modeling the conditional variance.

Finally, because oil price returns exhibit leptokurtosis, we consider three different types of distributions for η_t across all parametric models: the standard normal, Student's t, and GED distributions.

4.2. Estimation results

The whole sample is divided into two parts: the first 1512 observations (corresponding to the period January 3, 2007, to December 31, 2012) are used for in-sample estimation, and the rest are reserved for out-of-sample evaluation. Model specification tests suggest that the simplest conditional mean equation $r_t = \mu + \varepsilon_t$ is appropriate, whereas testing the residuals from this specification reveals very small autocorrelations but tremendous ARCH effects.

4.2.1. Non-switching GARCH models

The ML estimates and asymptotic standard errors (in parentheses) for the GARCH(1,1), EGARCH(1,1), GJR-GARCH(1,1) and FIGARCH(1, d, 1) models are reported in Table 2. Notice that the results from the GARCH and FIGARCH models are very similar to each other, with the fractional differencing parameter d being very close to

[6] This paper's choice of estimation method is driven by our interest in multi-step-ahead forecasts. Alternative estimation methods for MS-GARCH models include: (1) Gray's (1996) proposal to integrate out the unobserved regime path $\tilde{S}_{t-1} = (S_{t-1}, S_{t-2}, \ldots)$ in h_{t-1} in order to avoid the path dependence; (2) Francq and Zakoian's (2008) generalized method of moments (GMM) estimator using the autocovariances of the powers of the squared process; (3) Bauwens, Preminger, and Rombouts's (2010) Markov chain Monte Carlo (MCMC) algorithm — later modified by Bauwens, Dufays, and Rombouts (2014) — where the parameter space is enlarged to include the state variables and Bayesian estimation is done using Gibbs sampling; and (4) Augustyniak's (2014) combination of a Monte Carlo expectation–maximization (MCEM) algorithm and Bayesian importance sampling for calculating the maximum likelihood estimator (MCML). However, the multi-step-ahead volatility forecasts are less straightforward using these methods.

[7] Given that regimes are often observed to be highly persistent, S_t contains a lot of information about S_{t-1}. Thus, conditioning on S_t provides extra information that also leads to more efficient estimation.

[8] Francq and Zakoian (2008) also derived the conditions for weak stationarity and the existence of moments for MS-GARCH(p, q) processes.

Table 2
MLE estimates of the standard GARCH models.

	GARCH			EGARCH			GJR			FIGARCH		
	N	*t*	GED	*N*	*t*	GED	*N*	*t*	GED	*N*	*t*	GED
μ	0.1065** (0.0490)	0.0953* (0.0497)	0.1107** (0.0489)	0.0430 (0.0472)	0.0488 (0.0477)	0.0579 (0.0472)	0.0443 (0.0490)	0.0558 (0.0494)	0.0672 (0.0487)	0.1065** (0.0507)	0.0947* (0.0495)	0.1103** (0.0495)
α_0	0.1230** (0.0344)	0.0734* (0.0333)	0.0971* (0.0409)	0.0255** (0.0067)	0.0155* (0.0069)	0.0179* (0.0077)	0.1187** (0.0322)	0.0922** (0.0343)	0.1043** (0.0385)	0.1250** (0.0423)	0.0796** (0.0358)	0.1001** (0.0418)
α_1	0.0887** (0.0105)	0.0722** (0.0144)	0.0790** (0.0144)	0.1382** (0.0171)	0.1168** (0.0226)	0.1253** (0.0229)	0.0279** (0.0091)	0.0213 (0.0113)	0.0244* (0.0116)	0.0857** (0.0177)	0.0672** (0.0171)	0.0750** (0.0186)
γ_1	0.8908** (0.0147)	0.9171** (0.0169)	0.9052** (0.0185)	0.9855** (0.0039)	0.9899** (0.0041)	0.9880** (0.0046)	0.8976** (0.0143)	0.9161** (0.0165)	0.9075** (0.0178)	0.8916** (0.0220)	0.9177** (0.0207)	0.9063** (0.0229)
ξ	–	–	–	−0.0821** (0.0131)	−0.0669** (0.0151)	−0.0741** (0.0161)	0.1091** (0.0205)	0.0925** (0.0229)	0.0987** (0.0247)	–	-	–
d	–	–	–	–	–	–	–	–	–	0.99997** (0.0005)	0.99998** (0.0005)	0.99999** (0.0004)
ν	–	8.3776** (1.5261)	1.4941** (0.0643)	–	9.6838** (1.8938)	1.5375** (0.0652)	–	9.4739** (1.8476)	1.5299** (0.0645)	–	8.8282** (1.7528)	1.4994** (0.0744)
Log(*L*)	−3340.90	−3340.90	−3323.33	−3330.34	−3312.609	−3316.20	−3331.90	−3314.04	−3317.40	−3340.64	−3318.87	−3323.04

Note: Each model is estimated with the normal, Student's *t*, and GED distributions. The in-sample data consist of WTI returns from 1/3/07 to 12/31/12. The conditional mean is $r_t = \mu + \varepsilon_t$. The conditional variances are $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 h_{t-1}$, $\log(h_t) = \alpha_0 + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| - \mathbb{E} \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| \right) + \xi \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \log(h_{t-1})$, $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \xi \varepsilon_{t-1}^2 I_{\{\varepsilon_{t-1}<0\}} + \gamma_1 h_{t-1}$ and $h_t = \alpha_0 + \gamma_1 h_{t-1} + [1 - \gamma_1 L - (1 - (\alpha_1 + \gamma_1)L)(1 - L)^d] \varepsilon_t^2$ for GARCH, EGARCH, GJR-GARCH and FIGARCH respectively. Asymptotic standard errors are given in parentheses.

* Represent significance at the 10% level.

** Represent significance at the 5% level.

one.[9] The conditional mean in the GARCH/FIGARCH models is significantly positive, at around 0.1, regardless of the distribution. The estimated conditional mean is lower for the EGARCH and GJR-GARCH models than for the GARCH, and is insignificant across all distributions. Three features are worth noticing. First, the degrees of freedom for the t distribution are estimated to be greater than 8.37 in all three models, and the estimated shape parameter for the GED distribution is around 1.5.[10] This is consistent with the high sample kurtosis of daily crude oil returns (8.491), and, in turn, with the potential inability of a normal error to account for all of the mass in the tails of the distribution.[11]

Second, the asymmetric effect (ξ) is significant in the EGARCH and GJR-GARCH models across all distributions, suggesting that a negative shock would increase the future conditional variance more than a positive shock of the same magnitude. This result is consistent with political disruptions and large decreases in global demand leading to larger increases in volatility than, for instance, the fracking revolution. Third, the parameter estimates for the variance equation reveal high levels of persistence for all models. In the GARCH specification, $\alpha_1 + \gamma_1$ are estimated to be close to one. In the FIGARCH, d is estimated to be very close to one, suggesting that the process is very close to an IGARCH. In the EGARCH and GJR-GARCH models, the persistence levels measured by γ_1 and $\alpha_1 + \gamma_1 + 0.5\xi$, respectively, are also close to one. As was mentioned earlier, such a persistence might indicate possible structural breaks or regime switches (Lamoureux & Lastrapes, 1990; Mikosch & Starica, 2004).

4.2.2. MS-GARCH models

Before using the MS-GARCH models, one must test whether Markov switching exists in the data. Testing for Markov switching in GARCH models is complicated, for two reasons. First, the GARCH model itself is highly nonlinear. When the parameters are subject to regime switching, path dependence, together with nonlinearity, makes the estimation intractable, and as a consequence, the (log) likelihood functions cannot be calculated.[12] Second, standard tests suffer from the famous Davies problem, where the nuisance parameters that characterize the regime switching are not identified under the null hypothesis of parameter constancy, and therefore standard tests like the Wald and LR tests do not have the usual χ^2 distributions.

We apply the test developed by Carrasco et al. (2014), which is similar to a LM test, and only requires the model to

[9] This suggests that long memory might not be present in the in-sample estimation window. Nevertheless, we leave the FIGARCH in the pool for evaluation because we use a rolling-window scheme to calculate the out-of-sample forecasts.

[10] The conditional kurtosis for the t distribution is calculated by $3(\nu - 2)/(\nu - 4)$, with $\nu = 8.37$ implying a kurtosis of 4.37. The kurtosis for the GED distribution is given by $(\Gamma(1/\nu)\,\Gamma(5/\nu))/\Gamma^2(3/\nu)$. When $\nu = 1.5$, the kurtosis is 3.76.

[11] Our findings differ from those of Marcucci (2005), who found that a normal innovation is favored when modeling financial returns.

[12] The Markov switching tests proposed by Hansen (1992) or Garcia (1998), for example, are not applicable here, since they both involve an examination of the distribution of the likelihood ratio statistic, which is not feasible for MS-GARCH.

Table 3
Maximum likelihood estimates of the MS-GARCH models.

	MS-GARCH-N	MS-GARCH-t	MS-GARCH-GED
$\mu^{(1)}$	0.4181**	0.5367**	0.7125**
	(0.0938)	(0.1588)	(0.1761)
$\mu^{(2)}$	−0.2323**	−0.1570	−0.0730
	(0.1080)	(0.1411)	(0.0973)
$\alpha_0^{(1)}$	9.4156E−06	8.7386E−06	0.1952
	(0.0026)	(0.0024)	(0.2126)
$\alpha_0^{(2)}$	0.2541**	0.1266*	0.1643**
	(0.0887)	(0.0653)	(0.0642)
$\alpha_1^{(1)}$	1.0828E−07	0.0293	0.0045
	(6.0075E−05)	(0.0425)	(0.0225)
$\alpha_1^{(2)}$	0.0628**	0.0812**	0.0733**
	(0.0226)	(0.0282)	(0.0225)
$\gamma_1^{(1)}$	0.8673**	0.8689**	0.5918**
	(0.0432)	(0.0667)	(0.1622)
$\gamma_1^{(2)}$	0.9372**	0.9188**	0.9244**
	(0.0226)	(0.0282)	(0.0235)
p_{11}	0.8603**	0.8186**	0.7258**
	(0.0480)	(0.1037)	(0.1082)
p_{22}	0.9077**	0.9226**	0.9496**
	(0.0313)	(0.0393)	(0.0240)
$\nu^{(1)}$	–	4.5596*	1.9116**
		(2.4744)	(0.5866)
$\nu^{(2)}$	–	15.0977*	1.5313**
		(8.3849)	(0.0872)
$Log(L)$	−3325.7	−3312.5	−3316.4
No. of Par.	10	12	12
π_1	0.3977	0.2992	0.1554
π_2	0.6023	0.7008	0.8446
$\alpha_1^{(1)} + \gamma_1^{(1)}$	0.8673	0.8982	0.5963
$\alpha_1^{(2)} + \gamma_1^{(2)}$	0.99996	0.99997	0.9977

Note: Each MS-GARCH model is estimated using a different distribution, as described in the text. The in-sample data consist of WTI returns from 1/3/07 to 12/31/12. The superscripts indicate the regime. π_i is the ergodic probability of being in regime i; $\alpha_1^{(i)} + \gamma_1^{(i)}$ measures the persistence of shocks in the ith regime. Asymptotic standard errors are given in parentheses.
* Represent significance at the 10% level.
** Represent significance at the 5% level.

be estimated under the null hypothesis of constant parameters, but is still optimal. In addition, it has the flexibility to test for regime switching in the mean and/or variance or any subset of these parameters. We compute two test statistics, the supTS and the expTS,[13] which are equal to 0.007 and 0.680, respectively. We then simulate the critical values by bootstrapping using 3000 iterations. The null of constant parameters is rejected in favor of regime switching in both the mean and variance equations, with p-values of 0.028 for supTS and 0.018 for expTS. These results reveal overwhelming support for a Markov switching model, and hence, we estimate the MS-GARCH models with a two-state Markov chain, as described in Eq. (1).

Table 3 presents the parameter estimates for the three MS-GARCH models: MS-GARCH-N, MS-GARCH-t, and MS-GARCH-GED, respectively. In all three specifications, the common findings are: (i) regime 1 corresponds to significantly positive expected returns, whereas the expected returns in regime 2 are negative (but seldom significant); (ii) the transition probabilities p_{11} and p_{22} are close to one, implying that both regimes are highly persistent; (iii) the majority of the observations belong to regime 2; (iv) the persistence of shocks to the system in regime 2 is very close to one, suggesting a close-to-IGARCH behavior in this regime; and (v) shocks to the conditional variance are less persistent in regime 1. Specifically, the MS-GARCH-N has a significantly negative mean at −0.2323 in regime 2, and 60% of the observations lie in this regime. Meanwhile, regime 2 is more prevalent in the MS-GARCH-t and MS-GARCH-GED models (70% and 84% of observations, respectively), with a mean that is not significantly different from zero. In the MS-GARCH-t, regime 1 is specified by a t distribution with 4.56 degrees of freedom, while regime 2 is closer to a normal distribution (the degrees of freedom is equal to 15.10). In the meantime, MS-GARCH-GED's regime 1 is closer to being normal, with a shape parameter of 1.91, and regime 2 is characterized by a higher kurtosis.

To summarize, regime 1 is a relatively good regime, with positive expected returns and a much smaller dispersion, and any shocks to the conditional variance do not persist for long. The majority of observations lie in regime 2, which is characterized by either negative or zero expected returns, and the shocks to the conditional variance are highly persistent.

We conclude this section with a caveat. Of the three MS-GARCH models considered here, the MS-GARCH-t produces the most stable results with regard to various starting values and different numerical algorithms. This result probably should not come as a surprise to the reader, as the MS-GARCH-N is more restrictive and may not be able to accommodate the extra kurtosis that is present in the data. Alternatively, the MS-GARCH-GED allows for a greater flexibility in modeling leptokurtosis. However, numerical convergence tends to be more difficult to attain because the density of the GED involves a double exponential function of the absolute value of the residuals. The practitioner should be aware that a poor forecasting performance of the MS-GARCH-GED may stem from less accurate computation, rather than from the model itself.

5. Forecast evaluation

The out-of-sample forecast evaluation spans the period from January 2, 2013, to December 31, 2014.[14] We compute the forecasts using a rolling scheme and evaluate the forecasting performances based on 504 out-of-sample volatility forecasts (corresponding to the years 2013 and 2014) for horizons of 1, 5, 21, and 63 steps (corresponding to 1 day, 1 week, 1 month, and 3 months, respectively).[15] We choose a rolling window scheme because it is more robust to the presence of time-varying parameters than a

[13] A detailed description of the testing procedures can be found in Section B of the online appendix.

[14] Our observations extend to April 2, 2015, so as to accommodate the m-step-ahead forecast for $m = 63$.

[15] Financial investors are likely to rely more on short-term 1- and 5-day forecasts, whereas central bankers typically use monthly forecasts. For oil exploration and production firms, longer horizons are of interest because the time between pre-drilling activities and production easily exceeds one month and varies across regions. For instance, while the time to complete an oil well averages 20 days in Texas, it averages 90 days in Alaska.

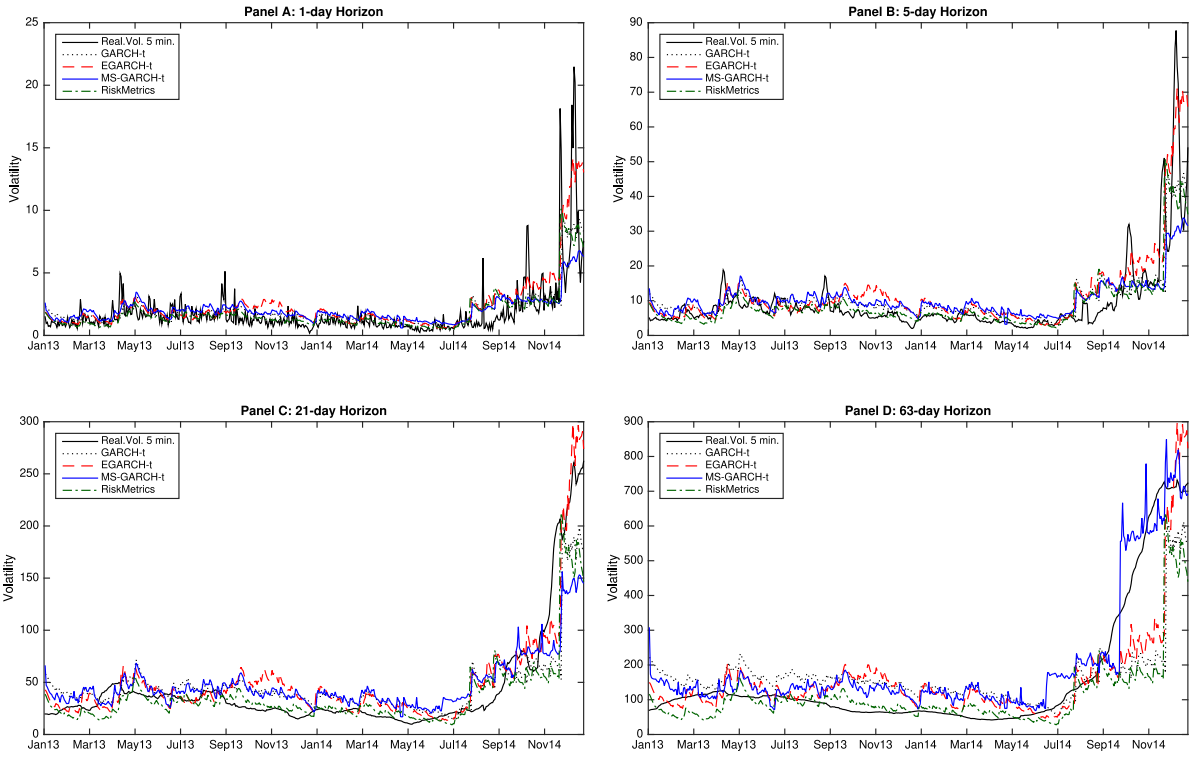

Fig. 2. Volatility forecast comparisons for select models. The out-of-sample period extends from January 2, 2013, to December 31, 2014.

recursive one. We also report the forecasts from RiskMetrics because of its popularity among practitioners.[16]

Fig. 2 plots the volatility forecasts obtained from four competing models: RiskMetrics, GARCH-t, EGARCH-t, and MS-GARCH-t.[17] The corresponding realized volatility is also plotted for reference. The four models yield very similar forecasts at the 1- and 5-day horizons. They move closely with the realized volatility and are able to capture the large increase in the realized volatility in mid-2014. At the 21-day horizon, all models are able to forecast the major upward and downward movements in the realized volatility, although the EGARCH-t seems to yield a more accurate forecast of the spike at the end of 2014. Only when we increase the forecast horizon to 63 days (three months) do our forecasts contain less information about the aggregated realized volatility during the out-of-sample period, which is as expected. However, the MS-GARCH-t does a good job of forecasting the sharp increase in volatility from mid-2014 onward.

We compare the volatility forecasts (denoted by $\widehat{h}_t$) based on two widely-used loss functions, where the realized volatility is substituted for the latent conditional variance (denoted as σ_t^2). The first one is the common mean square error, defined as $\text{MSE} = n^{-1}\sum_{t=1}^{n}\left(\sigma_t^2 - \widehat{h}_t\right)^2$. The second one, $\text{QLIKE} = n^{-1}\sum_{t=1}^{n}\left(\log \widehat{h}_t + \sigma_t^2/\widehat{h}_t\right)$, is equivalent to the loss function implied by a Gaussian likelihood. Our motivation for focusing on these particular loss functions is derived from the work of Patton (2011), who showed that only the MSE and QLIKE loss functions generate optimal forecasts that are equal to the conditional variance σ_t^2, even when noisy volatility proxies are used in forecast comparisons. The loss functions from all competing models and their rankings are reported in Table 4.

For the sake of brevity, and because models in which the innovations are assumed to follow a Student's t distribution fit the data better, we restrict our discussion to these models. At the 1-day forecast horizon, both the MSE and QLIKE rank RiskMetrics first. The MSE ranks the FIGARCH-t second and the EGARCH-t third, whereas this ranking is reversed for the QLIKE. Similarly, at the 5-day horizon, RiskMetrics is ranked first by both loss functions. However, the FIGARCH models drop to the bottom of the ranking under both MSE and QLIKE, and the GARCH-t emerges as the closest competitor to RiskMetrics. As the forecast horizon increases, the EGARCH models tend to rank higher than the GARCH models, with the EGARCH-t ranking first (second) at the 21-day horizon according to the MSE (QLIKE), and the GARCH-t ranking fifth. At this forecast horizon, RiskMetrics remains in the top three of the rankings, but the loss differential between RiskMetrics and GARCH-t (EGARCH-t) is smaller at the 21-day horizon than at the 1- or 5-day horizons. At the longer 63-day horizon, the MS-GARCH-t emerges as the winner under both loss functions, while the EGARCH models continue to rank highly, the GARCH models and RiskMetrics drop in the rankings, and the FIGARCH models remain at the bottom.

[16] RiskMetrics is equivalent to an IGARCH model (with normally distributed innovations) where the autoregressive parameter is set to $\lambda = 0.94$ and the coefficient on the squared residual is set to $1 - \lambda$.

[17] Plots for the remaining models are available from the authors upon request.

Table 4
Out-of-sample evaluation of the volatility forecasts.

	One day						Five days					
Model	MSE	Rank	QLIKE	Rank	SR	DA	MSE	Rank	QLIKE	Rank	SR	DA
GARCH-N	2.9916	7	1.4323	10	0.70	4.3826**	47.5283	6	3.0616	8	0.71	5.4174**
GARCH-t	2.7977	4	1.4198	4	0.69	4.2114**	42.7514	2	3.0460	2	0.72	6.0005**
GARCH-GED	2.8719	5	1.4249	6	0.70	4.5531**	44.5346	3	3.0522	4	0.72	5.8787**
EGARCH-N	3.2616	11	1.4257	7	0.70	3.4073**	60.6607	9	3.0595	6	0.72	4.5157**
EGARCH-t	2.7733	3	1.4174	2	0.69	3.5224**	46.5105	5	3.0489	3	0.70	3.9106**
EGARCH-GED	3.0544	9	1.4246	5	0.69	3.5224**	53.8590	8	3.0562	5	0.70	3.9703**
GJR-N	4.3695	15	1.4485	15	0.73	5.0201**	91.3195	13	3.0872	12	0.76	6.7207**
GJR-t	3.4927	12	1.4374	13	0.73	5.0201**	66.9322	11	3.0716	10	0.76	6.8970**
GJR-GED	3.9189	14	1.4439	14	0.73	5.0201**	78.4154	12	3.0795	11	0.76	6.7207**
MS-GARCH-N	2.9479	6	1.4323	9	0.68	4.7602**	46.4689	4	3.0631	9	0.71	6.5707**
MS-GARCH-t	3.1016	10	1.4321	8	0.68	4.9266**	53.2638	7	3.0607	7	0.70	6.0637**
MS-GARCH-GED	3.6191	13	1.4814	16	0.71	4.3033**	65.7369	10	3.1209	13	0.71	4.5082**
FIGARCH-N	3.0058	8	1.4350	12	0.72	5.6985**	130.8121	14	4.3071	14	0.73	6.4477**
FIGARCH-t	2.6977	2	1.4185	3	0.71	5.4694**	134.1039	15	4.5046	16	0.73	6.8015**
FIGARCH-GED	50.5641	16	1.4324	11	0.73	4.2957**	181.0000	16	4.4164	15	0.75	4.6106**
RiskMetrics	**2.2407**	**1**	**1.3812**	**1**	0.72	4.9563**	**40.8392**	**1**	**3.0268**	**1**	0.72	5.1754**
	21 days						63 days					
Model	MSE	Rank	QLIKE	Rank	SR	DA	MSE	Rank	QLIKE	Rank	SR	DA
GARCH-N	805.0850	9	4.5939	10	0.69	4.0120**	18559.0190	12	5.9356	12	0.65	2.3983**
GARCH-t	705.4356	5	4.5619	5	0.72	5.3630**	16354.4177	9	5.8806	8	0.73	6.6624**
GARCH-GED	739.6034	7	4.5745	7	0.71	4.6018**	17178.7713	10	5.9034	10	0.70	5.5209**
EGARCH-N	657.7634	4	4.5625	6	0.76	6.5598**	11710.8017	4	5.8359	4	0.79	9.1202**
EGARCH-t	**448.1821**	**1**	4.5457	2	0.75	6.2716**	10929.9002	2	5.8107	2	0.78	8.7478**
EGARCH-GED	527.3319	2	4.5522	3	0.75	6.4433**	10991.1017	3	5.8113	3	0.79	9.2570**
GJR-N	1203.2662	13	4.6121	12	0.78	7.0732**	18297.7123	11	5.9271	11	0.82	9.8728**
GJR-t	787.5492	8	4.5827	9	0.78	7.3624**	13803.3579	5	5.8767	7	0.80	9.2404**
GJR-GED	960.2393	11	4.5944	11	0.78	7.4936**	15199.0102	6	5.8918	9	0.81	9.5234**
MS-GARCH-N	716.5405	6	4.5758	8	0.74	6.9141**	15731.5260	8	5.8735	6	0.64	3.5959**
MS-GARCH-t	825.9422	10	4.5612	4	0.74	6.0481**	**4266.8562**	**1**	**5.7903**	**1**	0.84	10.1785**
MS-GARCH-GED	1199.0743	12	4.6765	13	0.70	3.0337**	27755.6880	13	6.0497	13	0.48	−6.1371
FIGARCH-N	3869.7152	14	14.0268	14	0.64	1.7795*	56387.1570	14	40.0194	14	0.44	−7.8655
FIGARCH-t	3899.8170	15	15.6592	16	0.68	4.4536**	56535.3748	16	45.7293	16	0.53	−2.3056
FIGARCH-GED	3905.4469	16	14.9680	15	0.74	2.0092*	56442.8738	15	43.5655	15	0.56	−7.1929
RiskMetrics	652.1611	3	**4.5425**	**1**	0.76	6.2046**	15418.4813	7	5.8562	5	0.81	10.0461**

Note: The volatility proxy is given by the realized volatility calculated with five-minute returns.
* Indicate significance of the DA statistic at the 5% level.
** Indicate significance of the DA statistic at the 1% level.

These results reveal important information. First, given that RiskMetrics can be considered as an IGARCH(1,1) with normal errors, the fact that it ranks highly suggests that the volatility exhibits IGARCH behavior. Either long memory or Markov switching could cause the extremely high persistence that is observed in the volatility of crude oil returns. Second, the huge losses that we see for the FIGARCH models imply that long memory can probably be ruled out (in favor of regime switching) as the reason for the high persistence in the volatility level.[18]

5.1. Success ratio and directional accuracy

We evaluate the models' abilities to predict the direction of the change in the volatility by calculating the success ratio (SR) and applying the directional accuracy (DA) test of Pesaran and Timmermann (1992).[19] The results are reported in Table 4.

For the 1- and 5-day horizons, the SR exceeds 68% for all models. Such is also the case at the 21-day horizon, with the exception of the FIGARCH-N, for which the SR equals 64%. At the longer 63-day horizon, the SR averages 70% across all models, but there is a greater variability. For instance, the SR ranges from 44% for the FIGARCH-N to 84% for the MS-GARCH-t. These results imply that, in the long run, the MS-GARCH-t does an exceptional job at predicting the direction of the change in volatility.

The results of Pesaran and Timmermann's DA test reinforce this finding. The test is significant at the 5% level for all models at most forecast horizons, which indicates that the forecast models have predictive power for the directional change in the underlying volatility. The exceptions are the FIGARCH models and the MS-GARCH-GED at a 63-day horizon.

[18] For the FIGARCH models, the estimation involves a truncation of the MacLaurin sequence of the polynomials. However, the long-run dependence implied by an IGARCH would be so highly persistent that any truncation would cause a severe bias, even at long lags.

[19] A detailed description of all of the forecast evaluation tests used in this paper is collected in Section C of the online appendix.

Table 5
Equal predictive ability test.

RiskMetrics benchmark

	One day		Five days		21 days		63 days	
Model	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
GARCH-N	−1.18	−3.47**	−1.03	−3.17**	−1.38	−2.71**	−1.82	−2.53*
GARCH-t	−1.01	−2.91**	−0.49	−2.03*	−0.60	−1.17	−0.69	−0.92
GARCH-GED	−1.07	−3.14**	−0.74	−2.51*	−0.90	−1.82	−1.20	−1.67
EGARCH-N	−1.06	−3.76**	−1.02	−2.41*	−0.04	−0.98	1.16	0.53
EGARCH-t	−0.66	−3.28**	−0.38	−1.79	1.38	−0.18	2.07+	1.48
EGARCH-GED	−0.90	−3.69**	−0.74	−2.24*	0.89	−0.52	1.93	1.40
GJR-N	−1.60	−4.89**	−1.77	−3.89**	−2.12*	−2.96**	−0.87	−1.72
GJR-t	−1.27	−4.55**	−1.34	−3.14**	−0.77	−1.86	0.61	−0.55
GJR-GED	−1.47	−4.79**	−1.60	−3.53**	−1.50	−2.32*	0.08	−0.92
MS-GARCH-N	−1.14	−3.94**	−0.79	−3.08**	−0.58	−1.88	−0.20	−0.66
MS-GARCH-t	−1.54	−3.71**	−1.57	−2.92**	−1.58	−1.09	2.24+	1.70
MS-GARCH-GED	−1.71	−6.42**	−2.01*	−5.74**	−3.47**	−5.15**	−4.88**	−4.46**
FIGARCH-N	−1.63	−4.80**	−2.52*	−10.20**	−3.16**	−11.57**	−3.82**	−8.07**
FIGARCH-t	−1.16	−3.60**	−2.55*	−11.40**	−3.18**	−13.48**	−3.83**	−9.17**
FIGARCH-GED	−1.02	−3.95**	−2.33*	−11.04**	−3.19**	−12.54**	−3.82**	−8.62**

MS-GARCH-t benchmark

	One day		Five days		21 days		63 days	
Model	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
GARCH-N	0.20	−0.03	0.45	−0.13	0.14	−4.18**	−3.53**	−7.39**
GARCH-t	0.68	1.77	1.03	2.33+	0.94	−0.10	−2.75**	−3.98**
GARCH-GED	0.47	0.99	0.77	1.29	0.62	−1.80	−3.01**	−5.21**
EGARCH-N	−0.14	0.47	−0.30	0.10	1.19	−0.14	−3.14**	−4.15**
EGARCH-t	0.32	1.16	0.32	1.04	2.10+	2.01+	−2.22*	−1.53
EGARCH-GED	0.04	0.56	−0.03	0.38	1.85	1.12	−2.30*	−1.66
GJR-N	−0.88	−1.07	−1.16	−1.94	−1.72	−4.80**	−5.03**	−11.48**
GJR-t	−0.33	−0.38	−0.54	−0.89	0.22	−2.49*	−3.53**	−7.67**
GJR-GED	−0.63	−0.81	−0.89	−1.44	−0.74	−3.48**	−4.05**	−8.96**
MS-GARCH-N	0.26	−0.03	0.49	−0.33	0.70	−2.09*	−2.76**	−3.98**
MS-GARCH-GED	−0.69	−5.54**	−0.70	−5.98**	−2.07*	−8.62**	−5.97**	−10.53**
FIGARCH-N	0.16	−0.30	−2.66**	−9.68**	−3.18**	−11.42**	−3.71**	−8.02**
FIGARCH-t	0.80	1.70	−2.69**	−10.80**	−3.20**	−13.28**	−3.72**	−9.12**
FIGARCH-GED	−1.00	−0.03	−2.26*	−10.45**	−3.20**	−12.37**	−3.71**	−8.57**
RiskMetrics	1.54	3.71++	1.57	2.92++	1.58	1.09	−2.24*	−1.70

Note: * and ** indicate rejections of the Diebold–Mariano-West test statistic with a null hypothesis of equal predictive accuracy at the 5% and 1% significance levels, respectively, where the test statistic is negative. + and ++ indicate that the test statistic is statistically positive at the 5% and 1% levels, respectively.

To summarize, we find that RiskMetrics and the conventional GARCH models do a good job of predicting the direction of the change in the volatility at short (1- and 5-day) and medium (21-day) horizons. However, the MS-GARCH-t model is more capable of directional prediction at longer horizons.

5.2. Tests of equal predictive ability

We assess the relative predictive accuracies of the volatility models by implementing the Diebold–Mariano-West (Diebold and Mariano, 1995, and West, 1996) test of equal predictive ability (EPA).[20] The results are reported in Table 5. Note that our use of the rolling scheme with a finite observation window means that the EPA test statistic does not suffer from the nested-model bias (see Giacomini & White, 2006), and has a normal distribution.[21] For the sake of brevity, and because RiskMetrics and MS-GARCH-t are ranked higher at short and long horizons respectively, we discuss only the results where these two models are taken as benchmarks.[22]

First, consider RiskMetrics, which is ranked highest by both MSE and QLIKE at the 1- and 5-day horizons. At the 1-day horizon, RiskMetrics has a significantly higher predictive accuracy than any of the competing models under QLIKE, but the differences in accuracy are not significant under MSE. Similar results are obtained at the 5-day horizon, with the exception that RiskMetrics has a significantly higher predictive accuracy than the FIGARCH family and MS-GARCH-GED under MSE as well as under QLIKE. As we move from short forecast horizons to a medium (21-day) horizon, the evidence of RiskMetrics having a higher

[20] The results of White's (2000) reality check (RC) test and Hansen's (2005) superior predictive ability (SPA) test are also reported, and can be found in Section C.4 of the online appendix.

[21] When two nested models are compared, the smaller model has an unfair advantage over the larger one because the larger model estimates additional parameters, thus introducing estimation error. Therefore, the larger model's sample loss function, e.g., MSE, is expected to be greater, which may lead one to conclude erroneously that the smaller one is better, resulting in size distortions where the larger model is rejected too often. In this case, one can use Clark and McCracken's ENC test, which corrects for the finite sample bias. See Clark and McCracken (2001) for details.

[22] The EPA test results for other benchmark models are available from the authors upon request.

Table 6
MCS $T_{R,\mathcal{M}}$ p-values.

	One day		Five days		21 days		63 days	
Model	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
GARCH-N	0.0000	0.0000	0.0500	0.0058	0.6696[a]	0.0466	0.0000	0.0000
GARCH-t	0.0160	0.0000	1.0000[a]	0.0230	1.0000[a]	0.3254[a]	0.0014	0.0004
GARCH-GED	0.0000	0.0006	1.0000[a]	0.0098	0.9954[a]	0.1338	0.0000	0.0000
EGARCH-N	0.0072	0.0004	0.0280	0.0272	1.0000[a]	0.3148[a]	0.0066	0.0004
EGARCH-t	0.0748	0.0000	0.2572[a]	0.0346	1.0000[a]	1.0000[a]	0.0724	0.0276
EGARCH-GED	0.0318	0.0006	0.1322	0.0260	1.0000[a]	1.0000[a]	0.0588	0.0586
GJR-N	0.0004	0.0000	0.0000	0.0008	0.0000	0.0376	0.0000	0.0000
GJR-t	0.0004	0.0004	0.0000	0.0070	0.8790[a]	0.1238	0.0000	0.0000
GJR-GED	0.0000	0.0000	0.0000	0.0024	0.0000	0.0824	0.0000	0.0000
MS-GARCH-N	0.0000	0.0000	0.2930[a]	0.0032	1.0000[a]	0.1062	0.0004	0.0006
MS-GARCH-t	0.0050	0.0002	0.0120	0.0164	0.2698[a]	0.4388[a]	1.0000[a]	1.0000[a]
MS-GARCH-GED	0.0000	0.0000	0.0000	0.0000	0.0000	0.0022	0.0000	0.0000
FIGARCH-N	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FIGARCH-t	0.0436	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FIGARCH-GED	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RiskMetrics	1.0000[a]	1.0000[a]	1.0000[a]	1.0000[a]	1.0000[a]	1.0000 [a]	0.1286	0.0854

Note: This table presents the $T_{R,\mathcal{M}}$ p-values from the MCS test.
[a] Indicates the models in $\widehat{\mathcal{M}}^{*}_{0.75}$.

predictive accuracy than the competing models becomes less prevalent. In particular, RiskMetrics dominates the FIGARCH family, the GJR-N and the MS-GARCH-GED significantly under both loss functions, and the GARCH-N and GJR-GED significantly under QLIKE. At the longer 63-day horizon, the EGARCH-t and the MS-GARCH-t both beat RiskMetrics under MSE. However, RiskMetrics continues to have a significantly greater predictive ability than the FIGARCH models and the MS-GARCH-GED, and is also found to be more accurate than the GARCH-N under QLIKE.

When the MS-GARCH-t is considered as the benchmark, the null of equal predictive ability cannot be rejected for the majority of the competing models across short horizons. The exceptions are the MS-GARCH-GED under QLIKE at the 1- and 5-day horizons and the FIGARCH models under both loss functions at the 5-day horizon. In addition, under QLIKE, we reject the null in favor of RiskMetrics at the 1- and 5-day horizons and in favor of the GARCH-t at the 5-day horizon. Nevertheless, the MS-GARCH-t has a significantly higher predictive accuracy at the 63-day horizon than all competing models under MSE, and than twelve of our fifteen models under QLIKE.[23]

5.3. Model confidence set

This section discusses the model confidence set (MCS), computed according to the procedure developed by Hansen et al. (2011). One advantage of the MCS over the EPA tests is that it does not require a pre-specified benchmark model; instead, it determines a set of "best" models M^{*} with respect to a loss function, given some specified level of confidence. Furthermore, if the data are sufficiently informative regarding which model is 'the best', the MCS will contain only one (or a small set) of the competing models.

We determine the MCS by following Hansen et al.'s (2011) suggestion to focus on the $T_{R,\mathcal{M}}$ statistic, and report the resulting p-values in Table 6.[24] The $T_{R,\mathcal{M}}$ test is computed with a confidence level of 0.25 over 3000 bootstrap iterations. We denote the resulting confidence sets by $\widehat{\mathcal{M}}^{*}_{0.75}$. The $\widehat{\mathcal{M}}^{*}_{0.75}$ is reduced to a singleton with RiskMetrics at the 1-day horizon and the MS-GARCH-t at the 63-day horizon. At the 5- and 21-day horizons, MSE produces more conservative sets than QLIKE, meaning that the resulting MCS sets contain more models. For instance, at a 5-day horizon, $\widehat{\mathcal{M}}^{*}_{0.75}$ under QLIKE contains only RiskMetrics. In contrast, under MSE, it also contains GARCH-t, GARCH-GED, EGARCH-t and MS-GARCH-N. Similarly, at the 21-day horizon, the MCS set contains six out of sixteen models under QLIKE and ten models under MSE. The FIGARCH models are all ruled out from the MCS, while the GJR models are mostly ruled out, except for the GJR- t at a 21-day horizon under MSE.

To summarize, RiskMetrics and the MS-GARCH-t emerge as the single best forecasting models at the 1- and 63-day forecast horizons, respectively. On the other hand, RiskMetrics, GARCH-t and EGARCH-t consistently appear in the MCS for the 5- and 21-day forecast horizons.

5.4. How stable is the forecasting accuracy of the preferred models?

One concern with using a single model to forecast over a long time period is that the predictive accuracy might depend on the specific out-of-sample period used for the forecast evaluation. In particular, a model might be chosen for its highest predictive accuracy when evaluating the loss functions over the entire out-of-sample period, but one of the competing models might exhibit a lower mean squared predictive error (MSPE) at a particular point (or points) in time during the evaluation period. For instance, Table 4

[23] The results for the superior predictive ability test and the reality check, reported in Tables A.1 and A.2 in the online appendix, are in line with these findings.

[24] Hansen et al. (2011) also proposed another statistic, $T_{\max,\mathcal{M}}$ (see the online appendix for details). Our results suggest that ($T_{\max,\mathcal{M}}$, $e_{\max,\mathcal{M}}$) are conservative and produce relatively large model confidence sets, which is consistent with the Corrigendum to Hansen et al.'s (2011) paper.

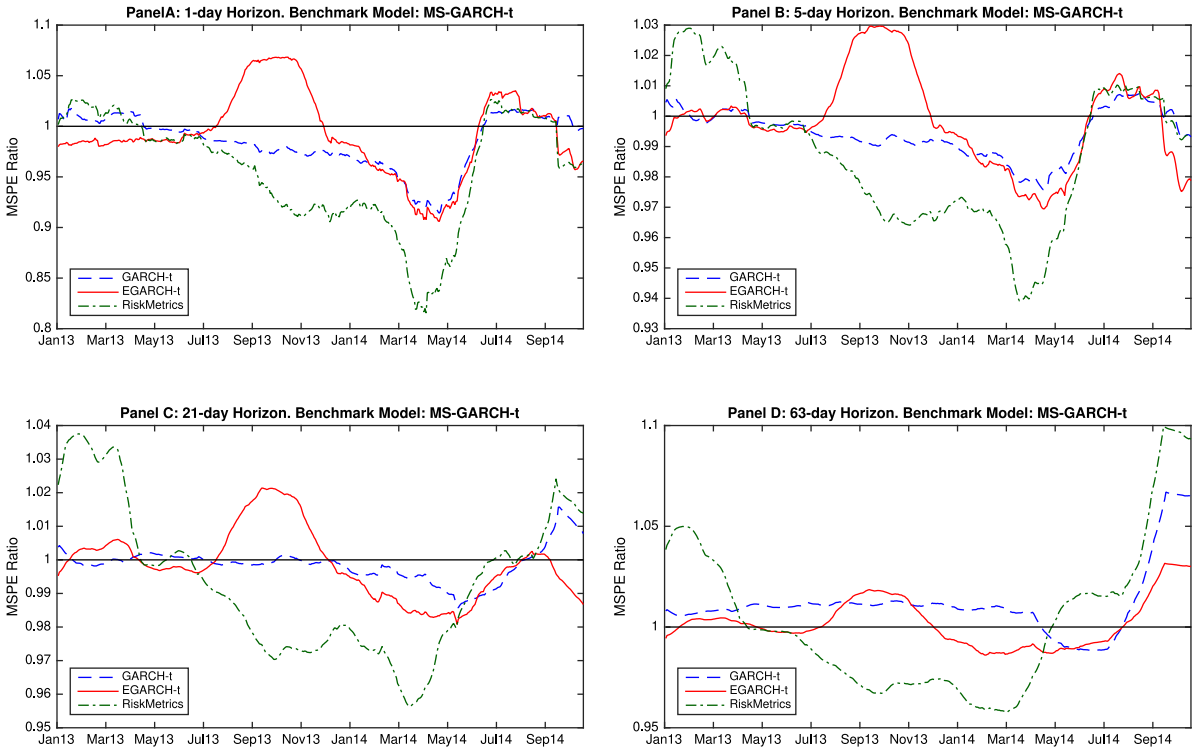

Fig. 3. Rolling window MSPE ratio relative to the MS-GARCH-t model.

indicates that over the entire evaluation period of 2013–2014, the RiskMetrics exhibits lower MSPEs — as measured by the loss functions MSE and QLIKE — for the 1- and 5-day forecast horizons, whereas the EGARCH-t and the MS-GARCH-t result in a smaller MSPEs for the 21- and 63-day forecast horizons, respectively.

We investigate the stability of the forecast accuracy by computing the MSPE ratio from the preferred QLIKE loss over 442 rolling sub-samples in the evaluation period. The first sub-sample consists of the first 63 forecasts (spanning three months) in the evaluation period, the second sub-sample is created by dropping the first forecast and adding the 64th forecast at the end, and so on. That is, these MSPEs are now computed as the average QLIKE over a rolling window of size $n = 63$. Fig. 3 plots the ratio of the MSPE for the RiskMetrics, GARCH-t and EGARCH-t models relative to the MS-GARCH-t at each of the four horizons. Note that, because the last window used for computing the MSPE spans the period between October 2, 2014, and December 31, 2014, the last MSPE reported is for October 1, 2014.

Fig. 3 illustrates that the MSPE ratio contains a lot of time variation over the evaluation period. The GARCH-t tends to have a low predictive accuracy at the beginning of the period, whereas RiskMetrics has a higher predictive ability in the middle of the sample. Although we find when considering the forecast period as a whole that the EGARCH-t has a good predictive ability at all horizons, it is outperformed by the MS-GARCH-t between September and December 2013. Recall that this was a period of consistent decreases in the WTI price. Similarly, during the second half of 2014 when the WTI price fell sharply (a 44% drop between June and December of 2014) and returns became more volatile, the MS-GARCH-t does a better job of predicting the increase in the volatility, even at the short 1- and 5-day horizons. We conclude that there are clear gains from using the MS-GARCH-t model for forecasting the crude oil return volatility, especially during periods of turmoil. While these gains are not as evident for the 1- and 5-day horizons over the two-year evaluation period (Table 4), they become clear when we plot the ratio of the rolling window MSPEs over a sub-period of three months.

6. Conclusion

This paper has offered an extensive empirical investigation of the relative forecasting performances of different models for the volatility of daily spot oil price returns at multiple horizons. Our finding is in favor of the RiskMetrics and GARCH models for short-horizon forecasts, EGARCH at medium horizons and MS-GARCH at long horizons. Thus, our results support the widespread use by practitioners of a naïve volatility model, RiskMetrics, for forecasting the crude oil volatility at short horizons. We also discover that the extremely high persistence that is observed in the volatility of crude oil prices is driven by Markov switching, rather than by long memory. The insights derived here are also in line with the findings in the literature for other assets (see e.g. Hansen & Lunde, 2005). Because the GARCH(1,1) model implies a geometric decay of the autocorrelation of the squared returns, the short-term volatility

dynamics can be captured well by such a parsimonious model. Alternatively, the MS-GARCH has the additional feature of incorporating abrupt changes in the parameters, and consequently allowing a more flexible functional form for the autocorrelation of the squared returns. Hence, it is not surprising that the MS-GARCH-t model not only does a better job of forecasting the volatility during periods of turmoil, but also yields more accurate long-term forecasts of the spot WTI return volatility.[25]

Two caveats are needed here. First, EGARCH models deliver an unbiased forecast of the logarithm of the conditional variance, but the forecast of the conditional variance itself will be biased following Jensen's inequality (see e.g. Andersen et al., 2006, among others). Hence, practitioners who prefer unbiased forecasts must be cautious when using EGARCH models. Second, long horizon volatility forecasts such as the one- and three-month horizons may be computed in various ways. For instance, if a researcher is interested in obtaining a one-month-ahead forecast, she could compute a "direct" forecast by first estimating the horizon-specific (e.g., monthly) GARCH model of the volatility and then using the estimates to predict the volatility over the next month directly. Alternatively, as we do here, she could compute an "iterated" forecast, where a daily volatility forecasting model is estimated first, and then the monthly forecast is computed by iterating over the daily forecasts for the 21 working days in the month. As Ghysels, Rubia, and Valkanov (2009) find that iterated forecasts of the stock market return volatility typically outperform the direct forecasts, we opt for this forecasting scheme here. Nevertheless, in our future research, we aim to evaluate the relative performances of these two alternative methods and compare them to the more recent mixed-data sampling (MIDAS) approach proposed by Ghysels, Santa-Clara, and Valkanov (2005, 2006).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2018.04.007.

References

Abosedra, S. S., & Laopodis, N. T. (1997). Stochastic behavior of crude oil prices: a GARCH investigation. *Journal of Energy and Development*, *21*(2), 283–291.

Abramson, A., & Cohen, I. (2007). On the stationarity of Markov-switching GARCH processes. *Econometric Theory*, *23*, 485–500.

Andersen, T. G., & Bollerslev, T. (1998). Answering the critics: Yes, ARCH models DO provide good volatility forecasts. *International Economic Review*, *39*(4), 885–905.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and correlation forecasting. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting*. Amsterdam: North Holland.

Arouri, M. E. H., Lahiani, A., Lévy, A., & Nguyen, D. K. (2012). Forecasting the conditional volatility of oil spot and futures prices with structural breaks and long memory models. *Energy Economics*, *34*, 283–293.

Augustyniak, M. (2014). Maximum likelihood estimation of the Markov-switching GARCH model. *Computational Statistics and Data Analysis*, *76*, 61–75.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. L. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*(1), 3–30.

Bauwens, L., Dufays, A., & Rombouts, J. V. K. (2014). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*, *178*, 508–522.

Bauwens, L., Preminger, A., & Rombouts, J. V. K. (2010). Theory and inference for a Markov-switching GARCH model. *Econometrics Journal*, *13*, 218–244.

Bina, C., & Vo, M. (2007). OPEC in the epoch of globalization: an event study of global oil prices. *Global Economy Journal*, *7*(1), 1524–5861.

Blair, B. J., Poon, S., & Taylor, S. (2001). Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. *Journal of Econometrics*, *105*, 5–26.

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

Calvet, L., & Fisher, A. (2001). Forecasting multifractal volatility. *Journal of Econometrics*, *105*(1), 27–58.

Caporale, G., Pittis, N., & Spagnolo, N. (2003). IGARCH models and structural breaks. *Applied Economics Letters*, *10*(12), 765–768.

Carrasco, M., Hu, L., & Ploberger, W. (2014). Optimal test for Markov switching parameters. *Econometrica*, *82*(2), 765–784.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85–110.

Davis, L. W., & Kilian, L. (2011). The allocative cost of price ceilings in the US residential market for natural gas. *Journal of Political Economy*, *119*, 212–241.

Diebold, F. X., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, *105*(1), 131–159.

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

Elder, J., & Serletis, A. (2010). Oil price uncertainty. *Journal of Money, Credit and Banking*, *42*(6), 1137–1159.

Fong, W., & See, K. (2002). A Markov switching model of the conditional volatility of crude oil futures prices. *Energy Economics*, *24*, 71–95.

Francq, C., & Zakoian, J. (2008). Deriving the autocovariances of powers of Markov-switching GARCH models, with applications to statistical inference. *Computational Statistics and Data Analysis*, *52*, 3027–3046.

Garcia, R. (1998). Asymptotic null distribution of the likelihood ratio test in Markov switching models. *International Economic Review*, *39*, 763–788.

Ghysels, E., Rubia, A., & Valkanov, R. (2009). *Multi-period forecasts of volatility: direct, iterated, and mixed-data approaches*. Working paper, University of North Carolina.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2005). There is a risk-return tradeoff after all. *Journal of Financial Economics*, *76*, 509–548.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, *131*, 59–95.

Giacomini, R., & White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, *74*, 1545–1578.

Glosten, L., Jagannathan, R., & Runkle, D. (1993). On the relation between expected value and the volatility of nominal excess returns on stocks. *Journal of Financ*, *48*, 1779–1901.

Granger, C. W. J., & Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Financ*, *11*, 399–421.

Gray, S. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, *42*, 27–62.

Hansen, B. (1992). The likelihood ratio test under non-standard conditions: testing the Markov switching model of GNP. *Journal of Applied Econometrics*, *7*, 61–82.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, *23*(4), 365–380.

Hansen, P. R., & Lunde, A. (2005). *A forecast comparison of volatility models: does anything beat a GARCH(1,1)? Journal of Applied Econometrics*, *20*, 873–889.

[25] For example, our finding that the MS-GARCH-t model is clearly preferred at long horizons is robust both to the use of a longer in-sample period ranging from January 2, 1986, to December 30, 2011, and to evaluating the forecasting ability on a shorter out-of-sample period (the year 2012), which excludes the large increase in volatility that was seen in the second half of 2014.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.
Hou, A., & Suardi, S. (2012). A nonparametric GARCH model of crude oil price return volatility. *Energy Economics*, *34*, 618–626.
Hsu, C. C. (2001). Change point estimation in regressions with $I(d)$ variables. *Economics Letters*, *70*(2), 147–155.
Jo, S. (2014). The effects of oil price uncertainty on global real economic activity. *Journal of Money, Credit and Banking*, *46*(6), 1113–1135.
Kahn, J. A. (1986). Gasoline prices and the used automobile market:a rational expectations asset price approach. *Quarterly Journal of Economics*, *101*, 323–340.
Kellogg, R. (2014). The effect of uncertainty on investment: evidence from Texas oil drilling. *American Economic Review*, *104*, 1698–1734.
Klaassen, F. (2002). Improving GARCH volatility forecasts. *Empirical Economics*, *27*(2), 363–394.
Lamoureux, C. G., & Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business and Economic Statistics*, *8*(2), 225–234.
Liu, L., Patton, A. J., & Sheppard, K. (2012). *Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes*. Working paper, Duke University.
Marcucci, J. (2005). Forecasting stock market volatility with regime-switching GARCH models. *Studies in Nonlinear Dynamics and Econometrics*, *9*(4), Article 6.
Mikosch, T., & Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *The Review of Economics and Statistics*, *86*, 378–390.
Mohammadi, H., & Su, L. (2010). International evidence on crude oil price dynamics: applications of ARIMA-GARCH models. *Energy Economics*, *32*, 1001–1008.
Morana, C. (2001). A semi-parametric approach to short-term oil price forecasting. *Energy Economics*, *23*(3), 325–338.
Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, *59*(2), 347–370.
Nomikos, N., & Pouliasis, P. (2011). Forecasting petroleum futures markets volatility: the role of regimes and market conditions. *Energy Economics*, *33*, 321–337.
Patton, A. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*, 246–256.
Pesaran, M. H., & Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics*, *10*(4), 461–465.
Pindyck, R. S. (2004). A volatility in natural gas and oil markets. *The Journal of Energy and Developmen*, *30*(1), 1–19.
Wang, Y., Wu, C., & Yang, L. (2016). Forecasting crude oil market volatility: a Markov switching multifractal volatility approach. *International Journal of Forecasting*, *32*(1), 1–9.
West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*, 1067–1084.
White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.
Xu, B., & Ouenniche, J. (2012). A data envelopment analysis-based framework for the relative performance evaluation of competing crude oil prices' volatility forecasting models. *Energy Economics*, *34*(2), 576–583.