Forecasting Crude Oil Price Volatility

Ana María Herrera* Liang Hu† Daniel Pastor‡

April 27, 2018

Abstract

We use high-frequency intra-day realized volatility to evaluate the relative forecasting performance of commonly used models for the volatility of crude oil daily spot returns at multiple horizons. The set of models includes RiskMetrics, GARCH, asymmetric GARCH, Fractional Integrated GARCH and Markov switching GARCH models. We first implement Carrasco, Hu, and Ploberger's (2014) test for regime switching in the mean and variance of the GARCH(1,1), finding 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 the $QLIKE$ 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 accuracy at short forecast horizons whereas EGARCH(1,1) yields the most accurate forecast at medium horizons; and (iii) Markov switching GARCH shows 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 totality of the 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 volatility during periods of turmoil.

Keywords: Crude oil price volatility, GARCH models, long memory, Markov switching, volatility forecast, realized volatility.

JEL codes: C22, C53, Q47

*Department of Economics, University of Kentucky, Gatton Business and Economics Building, Lexington, KY 40506-0034; e-mail: amherrera@uky.edu; phone: (859) 257-1119; fax: (859) 323-1920

†Corresponding author. Department of Economics, Wayne State University, 2119 Faculty Administration Building, 656 W. Kirby, Detroit, MI 48202; e-mail: lianghu@wayne.edu; phone: (313) 577-2846; fax: (313) 577-9564

‡Department of Economics and Finance, The University of Texas at El Paso, College of Business Administration, El Paso, TX 79968; e-mail: djpastor@utep.edu; phone: (915) 747-7472; fax: (915) 747-6282

1 Introduction

Throughout the past 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 put in evidence concerns voiced by the International Energy Agency regarding the return of high volatility in crude oil markets. This time around, apprehension regarding higher volatility seems to stem 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, and after the onset of the 2007-2008 financial crisis, and more recently since the fall in oil prices that started in July 2014 (see Figure 1). Clearly, periods of heightened volatility in crude oil markets are recurrent, and these headlines manifest the importance of evaluating whether the econometric tools available to practitioners are able to generate reliable forecasts of crude oil volatility.

"Spot oil price volatility reflects the volatility of current as well as future values of [oil] production, consumption and inventory demand" (Pindyck 2004), thus they are relevant for various economic agents. Accurate forecasts are key for those firms whose business greatly depends 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 refining and transportation infrastructure, airline companies who use oil price forecasts to set airfares, and the automobile industry. Second, oil price volatility also plays a role in households' decisions regarding purchases of durable goods (Kahn 1986, Davis and Kilian 2011). 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 forecaster (see, e.g., Elder and Serletis 2010, Jo 2014).

The aim of this paper is to evaluate the out-of-sample forecasting performance 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 and Bollerslev 1998). More specifically, we investigate the predictive ability of 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 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 and Lunde 2005) and crude oil price volatility (see Xu and Ouennich 2012 and references therein). Asymmetric GARCH models such as EGARCH (Nelson 1991) and GJR-GARCH (Glosten, Jagannathan, and Runkle 1993) have been shown to have good out-of-sample performance when forecasting oil price volatility one-step ahead (Mohammadi and Su 2010, and Hou and Suardi 2012). As for Markov switching models, they have been found to be better suited to model situations where changes in regimes are triggered by sudden shocks to the economy. Thus, they might have good predictive ability for spot crude oil returns, which are characterized by

sudden jumps due to, for instance, political disruptions in the Middle East or military interventions in oil exporting countries. However, regime switching and long memory are intimately related and it is hard to differentiate a Markov switching model from a long memory model (Nelson and Inoue 2001). Therefore, we add the FIGARCH to our pool of models for forecasting evaluation.

We provide a comprehensive study on the relative out-of-sample forecasting performance at multiple horizons. We start by formally testing for regime switches using the procedure proposed by Carrasco, Hu, and Ploberger (2014). Then, we evaluate directional accuracy using Pesaran and Timmerman'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 employ Hansen, Lunde, and Nason's (2011) Model Confidence Set procedure to determine the best set of model(s) from the pool. All 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 forecasting accuracy for the preferred models over the evaluation period (2013-2014).

Our findings are summarized as follows: (i) the Student's t distribution is generally favored in the parametric models due to 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) at the longer 63-day horizon, the MS-GARCH model yields more accurate forecasts; and (v) the MS-GARCH model has higher predictive ability during periods of turmoil.

We are not the first to consider Markov switching models in 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 and Fisher 2001) vis-à-vis a set of GARCH-class models. Alternatively, Arouri et al. (2012) discover that accounting for structural breaks and long memory in the GARCH specifications leads to gains in forecasting the conditional volatility of spot and futures oil prices. Our paper clearly benefits from this literature, but also differs in several aspects. Specifically, the MS-GARCH specification in this paper allows for great flexibility in modeling the persistence and regime switches. The adopted estimation method not only facilitates calculation of the multi-step-ahead forecast, but also makes more efficient use of the information contained in the data. We also employ an accurate proxy for the underlying volatility (the realized volatility instead of squared returns) and investigate forecasting stability over time.

This paper is organized as follows. Section 2 introduces the econometric models used in estimating and forecasting 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. The last section concludes.

2 Model Specifications

In this section, we briefly describe the parametric models widely used by practitioners in modeling and forecasting oil price volatility.

2.1 Standard GARCH Models

The conventional GARCH models considered in this paper comprise the GARCH (Bollerslev 1986), the EGARCH (Nelson 1991), and the GJR-GARCH (Glosten, Jagannathan, and Runkle 1993). 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 possibly given by $\boldsymbol{\beta}'\mathbf{x}_t$ with $\mathbf{x}_t$ being the $k \times 1$ vector of stochastic covariates and $\boldsymbol{\beta}$ a $k \times 1$ vector of parameters 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 the indicator function equal to one if $\varepsilon_{t-1} < 0$, and zero otherwise.

2.2 MS-GARCH

In using GARCH models to estimate 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 the result of 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, for instance, political disruptions in the Middle East, military interventions in oil exporting countries or depressed aggregate demand following the financial crisis. Consequently 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, only depends on the most adjacent past state, S_{t-1}:

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

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 $p_{11} = P\left(S_t = 1 \mid S_{t-1} = 1\right)$ and $p_{22} = P\left(S_t = 2 \mid S_{t-1} = 2\right)$. S_t is geometric ergodic if $0 < p_{11} < 1$ and $0 < p_{22} < 1$.

2.3 FIGARCH

As noted earlier, IGARCH behavior has been widely reported 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. This does not seem compatible with the persistence observed after large shocks such as the global financial crisis in Figure 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 Markov switching is easily confused 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 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, Bollerslev and Mikkelsen (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:

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

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

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 ranges from January 2, 2007 to April 2, 2015, a time period that comprises 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. To model crude oil returns and their volatility, we calculate daily returns by taking 100 times the difference in the logarithm of consecutive days' closing spot prices.

To evaluate the forecasting performance of different models, we need 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 proxy. More specifically, we obtain 5-minute prices of 1-month WTI oil futures contracts series from TickData.com spanning 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 the trading hours.[3] Then, to calculate m-step-ahead realized volatility at time T, we simply sum the daily realized volatility over m days, denoted by:

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

Table 1 reports the summary statistics for the WTI rates of return, the $RV_t^{1/2}$ and the logarithm of $RV_t^{1/2}$. The mean rate for the WTI returns is -0.010 with a standard deviation of 2.426. Note that WTI returns are slightly positively skewed. The kurtosis

[2] Andersen and Bollerslev (1998) note that squared daily returns are a noisy proxy of the true volatility and this noise can lead to improper conclusions about the forecasting ability of GARCH-type models. Anderson et al. (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 where futures are not traded around the clock, Blair, Poon, and Taylor (2001) suggest constructing the measure of daily realized volatility 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 to measure 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. Then they scale the realized volatility rv_t by the constant $\widehat{c}$. This measure is less noisy compared with Blair, Poon, and Taylor (2001). During our sample period, crude oil futures are traded almost continually during the day with the exception of the 45 minute gap between 5:15 and 6:00 p.m. EST. We have tried scaling and it turns out that our results are robust to scaling for the daily 45-minute interval when trading is halted.

equals 8.491 which is high compared to 3 for a normal distribution.[4] The $RV_t^{1/2}$ series is severely right-skewed and leptokurtic. However, the logarithmic series is less skewed with a kurtosis close to 3.

Figure 1 plots the returns of the WTI spot prices and the squared returns over the sample period. Two salient characteristics of WTI crude returns are apparent in the figure. First, crude oil returns are characterized by periods of low (high) volatility followed by low (high) volatility most of the time. GARCH models are intended to capture this volatility clustering. Second, exceptionally large variations in the WTI returns are observed 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 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 it is conducted via maximum likelihood. We thus restrict our discussion to the estimation of the MS-GARCH model in (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 integrating out all possible unobserved regime paths, which grow exponentially with the sample size T. Therefore, to estimate the MS-GARCH model we follow Klaassen (2002)[5] and replace h_{t-1} by its expectation conditional on the information set at $t-1$ and the current state variable, namely,

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

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

[5] The choice of estimation method made in this paper 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 $\tilde{S}_{t-1} = (S_{t-1}, S_{t-2}, ...)$ 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 –modified later in 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 to calculate the Maximum Likelihood Estimator (MCML). However, the multi-step-ahead volatility forecasts are less straightforward using these methods.

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 (2) circumvents the path dependence by integrating out h_{t-1}. Because the conditional variance depends only on the current state S_t, estimation and computation of the forecasts are straightforward.[6] 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 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.[7] It is easy to observe 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 great flexibility in modeling the conditional variance.

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

4.2 Estimation Results

The whole sample is divided into two parts: the first 1512 observations (corresponding to the period of 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 the simplest conditional mean equation $r_t = \mu + \varepsilon_t$ is appropriate, whereas testing the residuals from this specification reveals very small autocorrelations yet tremendous ARCH effects.

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

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

4.2.1 Non-switching GARCH models

The ML estimates and asymptotic standard errors (in parenthesis) 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 GARCH and FIGARCH are very close to each other, with the fractional differencing parameter d very close to 1.[8] 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 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 GED distribution is around 1.5.[9] 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 the mass in the tails of the distribution.[10] 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 persistence for all models. In the GARCH specification $\alpha_1 + \gamma_1$ are estimated close to 1. In the FIGARCH, d is estimated to be very close to 1, suggesting the process is very close to an IGARCH. In the EGARCH and GJR-GARCH models the persistence level measured by γ_1 and $\alpha_1 + \gamma_1 + 0.5\xi$, respectively, is also close to 1. As mentioned before, such persistence might be indicative of possible structural breaks or regime switches (Lamoureux and Lastrapes 1990, Mikosch and Starica 2004).

4.2.2 MS-GARCH Models

Before using the MS-GARCH models, one needs to 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, consequently the (log) likelihood functions are not calculable.[11] Second, standard tests suffer from the famous Davies problem, where the nuisance parameters characterizing the regime switching are not identified under the null hypothesis of

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

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

[10] Our findings differ from Marcucci (2005) where a normal innovation is favored in modeling financial returns.

[11] Markov switching tests by e.g., Hansen (1992) or Garcia (1998) are not applicable here since they both involve examining the distribution of the likelihood ratio statistic, which is not feasible for MS-GARCH.

parameter constancy. Therefore, standard tests like the Wald or LR test do not have the usual χ^2 distribution.

We apply the test developed by Carrasco, Hu, and Ploberger (2014). This test is similar to a LM test and only requires estimating the model under the null hypothesis of constant parameters, yet the test is still optimal. In addition, it has the flexibility to test for regime switching in both the mean and/or the variance or any subset of these parameters. We compute two test statistics, the supTS and the expTS[12]; they equal 0.007 and 0.680, respectively. Then, we simulate the critical values by bootstrapping using $3,000$ iterations. We reject the null of constant parameters 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, hence we estimate the MS-GARCH models with a two-state Markov chain as described in (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 are negative –but seldom significant– in regime 2; (ii) the transition probabilities p_{11} and p_{22} are close to one, implying that both regimes are highly persistent; (iii) the majority of observations belong in regime 2; (iv) the persistence of shocks to the system in regime 2 is very close to 1, 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, the MS-GARCH-t and the MS-GARCH-GED have a more prevalent regime 2 (70% and 84% of the observations, respectively), with a mean that is insignificantly different from 0. In the MS-GARCH-t, regime 1 is specified by a t distribution with 4.56 degrees of freedom, and regime 2 is closer to a normal distribution (the degrees of freedom equal 15.10). In the meantime, MS-GARCH-GED's regime 1 is closer to being normal with the shape parameter at 1.91 and regime 2 is characterized by higher kurtosis.

To summarize, regime 1 is a relatively good regime with positive expected returns, 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 regards to various starting values and different numerical algorithms. This result should probably 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 greater flexibility in modeling leptokurtosis. Yet, because the density of the GED involves a double exponential function of the absolute value of the residuals, numerical convergence tends to be more difficult to attain. The practitioner should be aware that poor performance of the MS-GARCH-GED in forecasting may stem

[12] A detailed description of their testing procedure is in the appendix.

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.[13] We compute the forecasts using a rolling scheme and evaluate forecasting performance based on 504 out-of-sample volatility forecasts (corresponding to the years 2013 and 2014) for the 1-, 5-, 21-, and 63-step horizons (corresponding to 1 day, 1 week, 1 month, and 3 months, respectively).[14] We choose a rolling window scheme because it is more robust to the presence of time-varying parameters than the recursive one. We also report the forecasts from the RiskMetrics given its popularity among practitioners.[15]

Figure 2 plots the volatility forecasts obtained from four competing models: RiskMetrics, GARCH-t, EGARCH-t, and MS-GARCH-t.[16] The corresponding realized volatility is also plotted for reference. At 1- and 5-day horizons, the four models yield very similar forecasts. They move closely with the realized volatility and are able to capture the large increase in the realized volatility in mid-2014. At a 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 (3 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 at forecasting the sharp increase in volatility from mid-2014.

We compare volatility forecasts (denoted as $\widehat{h}_t$) 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 $MSE = n^{-1}\sum_{t=1}^{n}\left(\sigma_t^2 - \widehat{h}_t\right)^2$. The second one, $QLIKE = n^{-1}\sum_{t=1}^{n}\left(\log \widehat{h}_t + \sigma_t^2/\widehat{h}_t\right)$, is equivalent to the loss function implied by a Gaussian likelihood. Our motivation to focus on these particular loss functions derives from Patton (2011) who shows that only the MSE and $QLIKE$ loss functions generate optimal forecasts 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 ranking are reported in Table 4.

For the sake of brevity, and because models where the innovations are assumed to follow a Student's t 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

[13] Our observations extend to April 2, 2015 to accommodate the m-step-ahead forecast at $m = 63$.

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

[15] 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 square residual is set to $1 - \lambda$.

[16] To economize space, plots for the remaining models are relegated to the online appendix.

ranks the FIGARCH-t second and the EGARCH-t third, whereas the ranking is reversed for the $QLIKE$. Similarly, at the 5-day horizon RiskMetrics is ranked first by both loss functions. Yet, under both MSE and $QLIKE$ the FIGARCH models drop to the bottom of the ranking 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 third of the rankings, however 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, the EGARCH models continue to rank highly, the GARCH models and RiskMetrics drop in the rankings, and the FIGARCH models remain at the bottom.

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 observed in the volatility of crude oil returns. Second, the huge losses 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.[17]

5.1 Success Ratio and Directional Accuracy

To evaluate the ability of the models to predict the direction of the change in the volatility, we calculate the Success Ratio (SR) and apply the Directional Accuracy (DA) test of Pesaran and Timmermann (1992). The results are reported in Table 4.

For the 1- and 5-day horizons, the SR exceeds 68% for all models. This is also the case at the 21-day horizon, with the exception of the FIGARCH-N for which the SR equals 64%. At a longer 63-day horizon the SR averages 70% across all models but there is greater variability. For instance, the SR ranges between 44% for the FIGARCH-N and 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.

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

[17] For 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 severe bias, even at long lags.

5.2 Tests of Equal Predictive Ability

To assess the relative predictive accuracy of the volatility models we implement the Diebold-Mariano-West (Diebold and Mariano 1995 and West 1996) test of Equal Predictive Ability (EPA).[18] The results are reported in Table 5. Notice that since we use the rolling scheme with a finite observation window, the EPA test statistic does not suffer from the nested-model bias (see Giacomini and White 2006) and it has a normal distribution.[19] For the sake of brevity, and because RiskMetrics and MS-GARCH-t are, respectively, ranked higher at short and long horizons, we discuss the results where these two models are taken as benchmarks.[20]

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 significantly higher predictive accuracy against all competing models under $QLIKE$, but insignificantly under MSE. Similar results are obtained at the 5-day horizon, with the exception that RiskMetrics has significantly higher predictive accuracy than the FIGARCH family and MS-GARCH-GED not only under $QLIKE$ but also MSE. As we move from short forecast horizons to the medium (21-day) horizon, evidence that RiskMetrics has higher predictive accuracy than the competing models becomes less prevalent. In particular, RiskMetrics significantly dominates the FIGARCH family, the GJR-N and the MS-GARCH-GED under both loss functions, and the GARCH-N and GJR-GED under $QLIKE$. At the longer 63-day horizon, the EGARCH-t and the MS-GARCH-t beat RiskMetrics under MSE. RiskMetrics continues to have significantly higher predictive ability than the FIGARCH models and the MS-GARCH-GED; it 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 competing models across short horizons. The exceptions are 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 1- and 5-day horizons and in favor of the GARCH-t at the 5-day horizon. Nevertheless, at the 63-day horizon the MS-GARCH-t has significantly higher predictive accuracy than all competing models under MSE and twelve out of fifteen models under $QLIKE$.[21]

[18] White's (2000) Reality Check (RC) test, and Hansen's (2005) Superior Predictive Ability (SPA) test and test results are also reported in an online appendix.

[19] When two nested models are compared, the smaller model has an unfair advantage relative to the larger one because the larger model estimates extra parameters, thus introducing estimation error. Therefore, the larger model's sample loss function, e.g., MSE is expected to be greater. One may therefore erroneously conclude 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.

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

[21] Results for the superior predictive ability test and the reality check, reported in the online appendix, are in line with these findings.

5.3 Model Confidence Set

This section discusses the Model Confidence Set (MCS) computed according to the procedure developed by Hansen, Lunde, and Nason (2011). An 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 is sufficiently informative regarding which model is 'the best', then the MCS will contain only one (or a small set) of the competing models.

To determine the MCS we follow Hansen, Lunde, and Nason's (2011) suggestion to focus on the $T_{R,\mathcal{M}}$ statistic and report the p-values in Table 6.[22] The $T_{R,\mathcal{M}}$ test is computed with confidence level of 0.25 over 3000 bootstrap iterations. We denote the resulting confidence sets by $\widehat{\mathcal{M}}^*_{.75}$. The $\widehat{\mathcal{M}}^*_{.75}$ is reduced to a singleton with RiskMetrics at the 1-day horizon and the MS-GARCH-t at the 63-day horizons. At the 5- and 21-day horizon MSE produces more conservative sets than $QLIKE$ and, thus, the resulting MCS set contains more models. For instance, at a 5-day horizon, $\widehat{\mathcal{M}}^*_{.75}$ contains only RiskMetrics under $QLIKE$. In contrast, $\widehat{\mathcal{M}}^*_{.75}$ also contains GARCH-t, GARCH-GED, EGARCH-t and MS-GARCH-N under MSE. 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 and the GJR models are commonly 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 1- and 63-day forecast horizons, respectively. Instead, RiskMetrics, GARCH-t and EGARCH-t consistently appear in the MCS for 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 out-of-sample period used for 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, yet 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 indicates that for the entire evaluation period of 2013-2014, the RiskMetrics exhibits lower $MSPE$ –as measured by the loss functions (MSE, $QLIKE$)– for the 1- and 5-day forecast horizons, whereas the EGARCH-t results in smaller $MSPE$ for the 21-day horizon and the MS-GARCH-t for the 63-day forecast horizon.

To investigate the stability of the forecast accuracy, we compute the $MSPE$ ratio

[22] Hansen, Lunde, and Nason's (2011) proposed another statistic $T_{\max,\mathcal{M}}$ (see 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, Lunde, and Nason's (2011) paper.

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. In brief, these $MSPE$s are now computed as the average $QLIKE$ over a rolling window of size $n = 63$. Figure 3 plots the ratio of the $MSPE$ for 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 to compute the $MSPE$ spans the period between October 2, 2014 and December 31, 2014, the last $MSPE$ is reported at October 1, 2014.

Figure 3 illustrates that the $MSPE$ ratio contains a lot of time variation during the evaluation period. The GARCH-t tends to have low predictive accuracy at the beginning of the period. In contrast, RiskMetrics has higher predictive ability in the middle of the sample. Although, when considering the forecast period as a whole, we find that the EGARCH-t has 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 decrease 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 at predicting the increase in volatility even at short 1- and 5-day horizons. We conclude that there are clear gains from using the MS-GARCH-t model for forecasting crude oil return volatility, especially during periods of turmoil. Whereas 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 $MSPE$s over a sub-period of three months.

6 Conclusion

This paper offered an extensive empirical investigation of the relative forecasting performance of different models for the volatility of daily spot oil price returns at multiple horizons. Our finding is in favor of 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, to forecast crude oil volatility at short horizons. We also discover that the extremely high persistence level 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 literature's findings for other assets (see, e.g. Hansen and Lunde 2005). Because the GARCH(1,1) model implies a geometric decay of the autocorrelation of the squared returns, short-term volatility dynamics can be well captured 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 at forecasting volatility during periods of turmoil but

also yields more accurate long-term forecasts of the spot WTI return volatility.[23]

Two caveats are needed here. First, EGARCH models deliver an unbiased forecast for 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, for practitioners who prefer unbiased forecasts, caution must be taken 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 volatility and then use the estimates to directly predict the volatility over the next month. Alternatively, as we do here, she could compute an "iterated" forecast where a daily volatility forecasting model is first estimated and the monthly forecast is then computed by iterating over the daily forecasts for the 21 working days in the month. Ghysels, Rubia, and Valkanov (2009) find that iterated forecasts of stock market return volatility typically outperform the direct forecasts. Thus we opt for this forecasting scheme. Nevertheless, evaluating the relative performance of these two alternative methods and comparing it to the more recent mixed-data sampling (MIDAS) approach proposed by Ghysels, Santa-Clara, and Valkanov (2005, 2006) is the aim of our future research.

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

References

[1] Abosedra, S. S. and N. T. Laopodis (1997), "Stochastic behavior of crude oil prices: a GARCH investigation," Journal of Energy and Development 21(2), 283-291.

[2] Abramson, A. and I. Cohen (2007), "On the stationarity of Markov-switching GARCH processes," Econometric Theory 23, 485-500.

[3] Andersen, T. G. and T. Bollerslev (1998), "Answering the Critics: Yes ARCH Models DO Provide Good Volatility Forecasts," *International Economic Review* 39(4), 885-905.

[4] Andersen, T. G., T. Bollerslev, P. F. Christoffersen and F. X. Diebold (2006), "Volatility and Correlation Forecasting," In: Elliott G., Granger C., Timmermann A. (Eds.): *Handbook of Economic Forecasting*, North Holland, Amsterdam.

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

[6] Augustyniak, M. (2014), "Maximum likelihood estimation of the Markov-switching GARCH model," *Computational Statistics & Data Analysis* 76(0), 61-75, CFEnetwork: The Annals of Computational and Financial Econometrics 2nd Issue.

[7] Baillie, R. T., T. Bollerslev and H. L. Mikkelsen (1996), "Fractionally integrated generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics* 74(1), 3-30.

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

[9] Bauwens, L., A. Preminger, and J.V.K. Rombouts (2010), "Theory and Inference for a Markov-switching GARCH Model," *Econometrics Journal* 13, 218-244.

[10] Bina, C., and M. Vo (2007), "OPEC in the epoch of globalization: an event study of global oil prices," *Global Economy Journal* 7(1).

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

[12] Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics* 31(3), 307-327.

[13] Calvet, L. and A. Fisher (2001), "Forecasting multifractal volatility," *Journal of Econometrics* 105(1), 27-58.

[14] Caporale, G., N. Pittis and N. Spagnolo (2003), "IGARCH models and structural breaks," *Applied Economics Letters* 10(12), 765-768.

[15] Carrasco, M., L. Hu and W. Ploberger (2014), "Optimal Test for Markov Switching Parameters," *Econometrica* 82(2), 765-784.

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

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

[18] Diebold, F. X. and A. Inoue (2001), "Long memory and regime switching," *Journal of Econometrics* 105(1), 131-159.

[19] Diebold, F. X. and R. S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13(3), 253-263.

[20] Elder, J. and A. Serletis (2010), "Oil Price Uncertainty," *Journal of Money, Credit and Banking* 42(6), 1137-1159.

[21] Fong, W. and K. See (2002), "A Markov switching model of the conditional volatility of crude oil futures prices," *Energy Economics* 24, 71-95.

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

[23] Garcia, R. (1998), "Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models," *International Economic Review* 39, 763-788.

[24] Ghysels, E., A. Rubia and R. Valkanov (2009), "Multi-Period Forecasts of Volatility: Direct, Iterated, and Mixed-Data Approaches," working paper, University of North Carolina.

[25] Ghysels, E., P. Santa-Clara and R. Valkanov (2005), "There is a Risk-Return Tradeoff After All," *Journal of Financial Economics* 76, 509–548.

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

[27] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

[28] Glosten, L., R. Jagannathan and D. Runkle (1993), "On the Relation Between Expected Value and the Volatility of Nominal Excess Returns on Stocks," *Journal of Finance* 48, 1779-1901.

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

[30] Gray, S. (1996), "Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process," *Journal of Financial Economics* 42, 27-62.

[31] 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.

[32] Hansen, P. R. (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23(4), 365-380.

[33] Hansen, P. R. and A. Lunde (2005), "A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?," *Journal of Applied Econometrics* 20, 873-889.

[34] Hansen, P. R., A. Lunde and J.M. Nason (2011), "The Model Confidence Set," *Econometrica* 79(2), 453-497.

[35] Hou, A. and S. Suardi (2012), "A nonparametric GARCH model of crude oil price return volatility," *Energy Economics* 34, 618-626.

[36] Hsu, C. C. (2001), "Change point estimation in regressions with I(d) variables," *Economics Letter* 70(2), 147-155

[37] Jo, S. (2014), "The Effects of Oil Price Uncertainty on Global Real Economic Activity," *Journal of Money, Credit and Banking* 46(6), 1113-1135.

[38] Kahn, J.A. (1986), "Gasoline prices and the used automobile market:a rational expectations asset price approach," *Quarterly Journal of Economics* 101, 323–340.

[39] Kellogg, R. (2014), "The Effect of Uncertainty on Investment: Evidence from Texas Oil Drilling," *American Economic Review* 104, 1698-1734.

[40] Klaassen, F. (2002), "Improving GARCH Volatility Forecasts," *Empirical Economics* 27(2), 363-94.

[41] Lamoureux, C. G. and W. D. Lastrapes (1990), "Persistence in variance, structural change, and the GARCH model," *Journal of Business and Economic Statistics* 8(2), 225-234.

[42] Liu, L, A. J. Patton and K. Sheppard (2012), "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes," working paper, Duke University.

[43] Marcucci, J. (2005), "Forecasting Stock Market Volatility with Regime-Switching GARCH Models," *Studies in Nonlinear Dynamics and Econometrics* 9(4), Article 6.

[44] Mikosch, T. and C. Starica (2004), "Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects," *The Review of Economics and Statistics* 86, 378-390.

[45] Mohammadi, H. and L. Su (2010), "International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models ," *Energy Economics* 32, 1001-1008.

[46] Morana, C. (2001), "A semi-parametric approach to short-term oil price forecasting," *Energy Economics* 23(3), 325-338.

[47] Nelson, D. B. (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica* 59(2), 347-370.

[48] Nomikos, N. and P. Pouliasis (2011), "Forecasting petroleum futures markets volatility: The role of regimes and market conditions," *Energy Economics* 33, 321-337.

[49] Patton, A (2011), "Volatility forecast comparison using imperfect volatility proxies," *Journal of Econometrics* 160, 246-256.

[50] Pesaran, M. H. and A. Timmermann (1992), "A Simple Nonparametric Test of Predictive Performance," *Journal of Business and Economic Statistics* 10(4), 461-465.

[51] Pindyck, R. S. (2004), "A Volatility in natural gas and oil markets," *The Journal of Energy and Development* 30(1), 1-19.

[52] Wang, Y., C. Wu and L. Yang (2016), "Forecasting crude oil market volatility: A Markov switching multifractal volatility approach," *International Journal of Forecasting* 32(1), 1-9.

[53] West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.

[54] White, H. (2000), "A Reality Check for Data Snooping," *Econometrica* 68(5), 1097-1126.

[55] Xu, B. and J. Ouenniche (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.

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 realized volatility computed from the 5-minute returns on oil futures. WTI returns, $RV^{1/2}$, and the natural logarithm of $RV^{1/2}$ series are from the sample period of January 3, 2007 to April 2, 2015 for 2079 observations.

Table 2: MLE Estimates of Standard GARCH Models

	GARCH			EGARCH			GJR			FIGARCH		
	N	t	GED	N	t	GED	N	t	GED	N	t	GED
μ	0.1065**	0.0953*	0.1107**	0.0430	0.0488	0.0579	0.0443	0.0558	0.0672	0.1065**	0.0947*	0.1103**
	(0.0490)	(0.0497)	(0.0489)	(0.0472)	(0.0477)	(0.0472)	(0.0490)	(0.0494)	(0.0487)	(0.0507)	(0.0495)	(0.0495)
α_0	0.1230**	0.0734*	0.0971*	0.0255**	0.0155*	0.0179*	0.1187**	0.0922**	0.1043**	0.1250 **	0.0796 **	0.1001 **
	(0.0344)	(0.0333)	(0.0409)	(0.0067)	(0.0069)	(0.0077)	(0.0322)	(0.0343)	(0.0385)	(0.0423)	(0.0358)	(0.0418)
α_1	0.0887**	0.0722**	0.0790**	0.1382**	0.1168**	0.1253**	0.0279**	0.0213	0.0244*	0.0857 **	0.0672 **	0.0750 **
	(0.0105)	(0.0144)	(0.0144)	(0.0171)	(0.0226)	(0.0229)	(0.0091)	(0.0113)	(0.0116)	(0.0177)	(0.0171)	(0.0186)
γ_1	0.8908**	0.9171**	0.9052**	0.9855**	0.9899**	0.9880**	0.8976**	0.9161**	0.9075**	0.8916 **	0.9177 **	0.9063 **
	(0.0147)	(0.0169)	(0.0185)	(0.0039)	(0.0041)	(0.0046)	(0.0143)	(0.0165)	(0.0178)	(0.0220)	(0.0207)	(0.0229)
ξ	-	-	-	-0.0821**	-0.0669**	-0.0741**	0.1091**	0.0925**	0.0987**	-	-	-
				(0.0131)	(0.0151)	(0.0161)	(0.0205)	(0.0229)	(0.0247)			
d	-	-	-	-	-	-	-	-	-	0.99997**	0.99998**	0.99999**
										(0.0005)	(0.0005)	(0.0004)
ν	-	8.3776**	1.4941**	-	9.6838**	1.5375**	-	9.4739**	1.5299**	-	8.8282**	1.4994**
		(1.5261)	(0.0643)		(1.8938)	(0.0652)		(1.8476)	(0.0645)		(1.7528)	(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: * and ** represent significance at 10% and 5% level respectively. Each model is estimated with 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 in parentheses.

Table 3: Maximum Likelihood Estimates of 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
$N.of\ Par.$	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: * and ** represent significance at 10% and 5% level respectively. Each MS-GARCH model is estimated using 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 i-th regime. Asymptotic standard errors are in the parentheses.

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**

	Twenty-one Days						Sixty-three 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. * and ** denote 5% and 1% significance levels for the DA statistic, respectively.

Table 5: Equal Predictive Ability Test

RiskMetrics Benchmark

	One Day		Five Days		Twenty-one Days		Sixty-three 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**	-1.5	-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		Twenty-one Days		Sixty-three 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 ** represent the Diebold-Mariano-West test statistic for which the null hypothesis of equal predictive accuracy can be rejected at 5% and 1% significance level respectively and the test statistic is negative. + and ++ represent the test statistic is statistically positive at 5% and 1% level, respectively.

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

	One Day		Five Days		Twenty-one Days		Sixty-three Days	
Model	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
GARCH-N	0.0000	0.0000	0.0500	0.0058	0.6696*	0.0466	0.0000	0.0000
GARCH-t	0.0160	0.0000	1.0000*	0.0230	1.0000*	0.3254*	0.0014	0.0004
GARCH-GED	0.0000	0.0006	1.0000*	0.0098	0.9954*	0.1338	0.0000	0.0000
EGARCH-N	0.0072	0.0004	0.0280	0.0272	1.0000*	0.3148*	0.0066	0.0004
EGARCH-t	0.0748	0.0000	0.2572*	0.0346	1.0000*	1.0000*	0.0724	0.0276
EGARCH-GED	0.0318	0.0006	0.1322	0.0260	1.0000*	1.0000*	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*	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*	0.0032	1.0000*	0.1062	0.0004	0.0006
MS-GARCH-t	0.0050	0.0002	0.0120	0.0164	0.2698*	0.4388*	1.0000*	1.0000*
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*	1.0000*	1.0000*	1.0000*	1.0000*	1.0000*	0.1286	0.0854

: Note: This table presents the $T_{R,\mathcal{M}}$ p-values from the MCS test. The models in $\widehat{\mathcal{M}}^*_{.75}$ are identified by *.

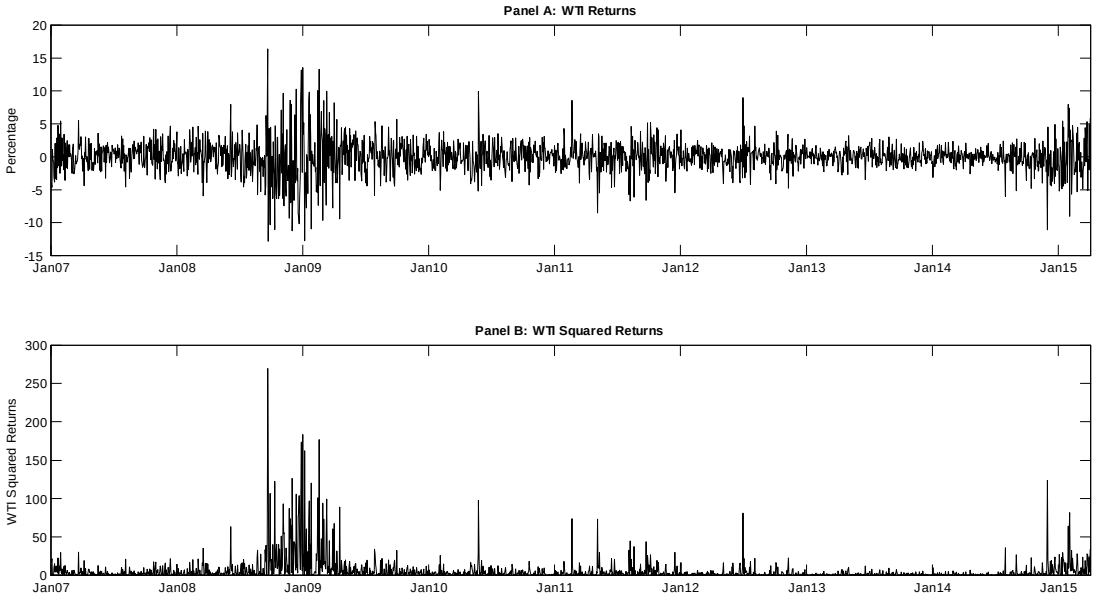

Figure 1: Daily WTI Crude Oil Returns and Squared Returns. The sample period extends from January 3, 2007 through April 2, 2015.

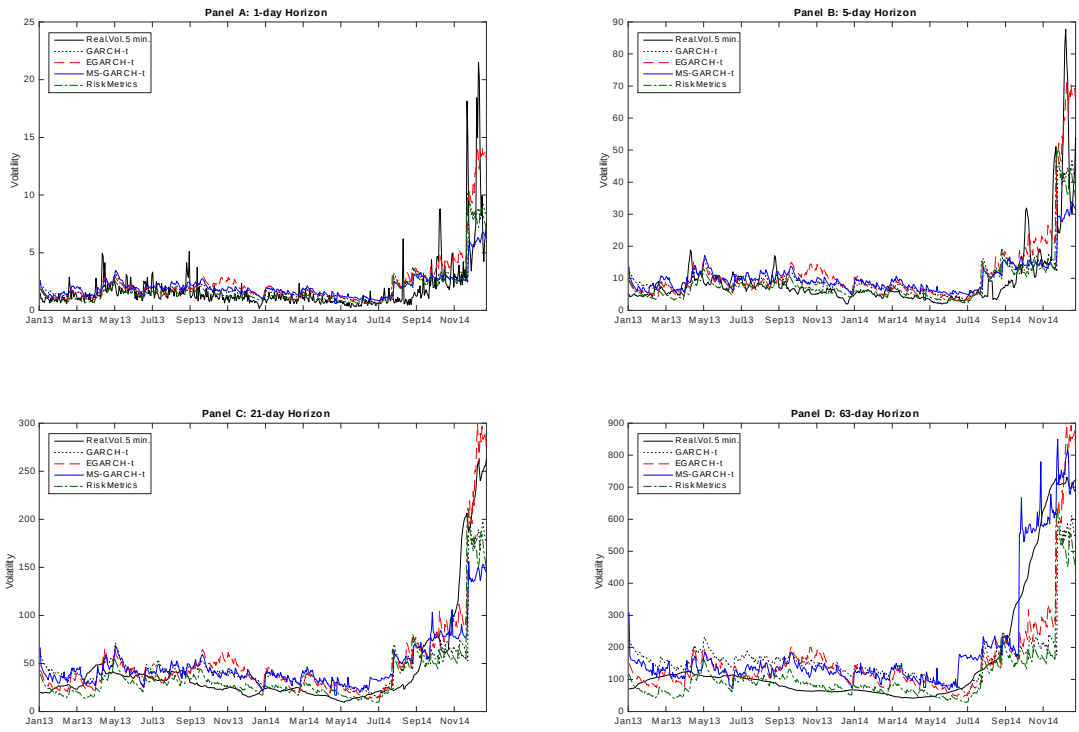

Figure 2: Volatility Forecast Comparisons for Select Models. The out-of-sample period extends from January 2, 2013 through Dec 31, 2014.

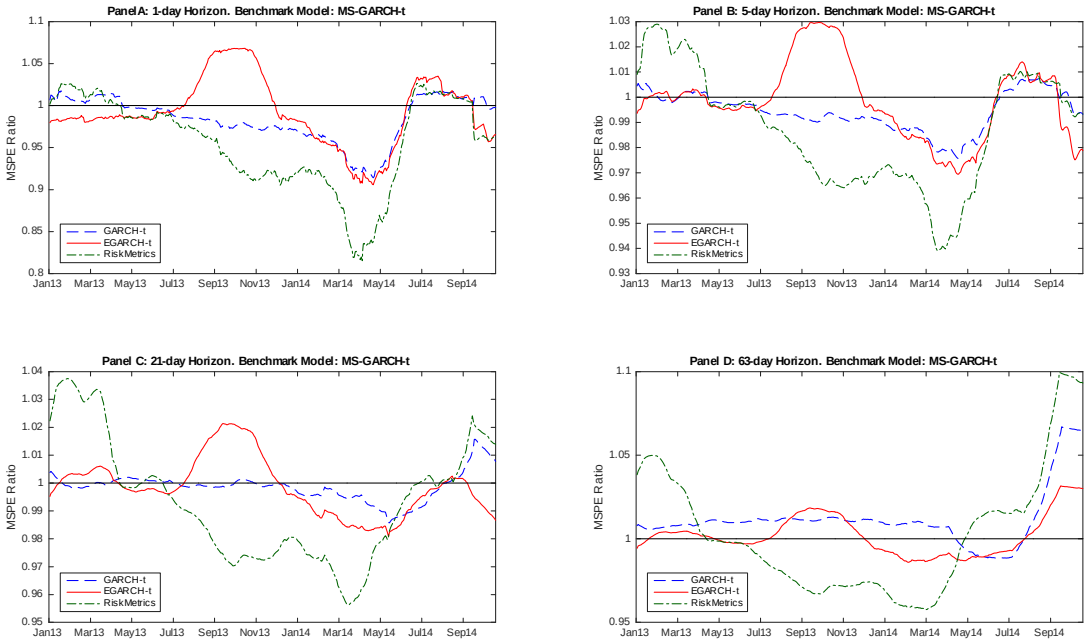

Figure 3: Rolling Window MSPE Ratio Relative to MS-GARCH-t model

7 Appendix

7.1 Model Specifications and Estimation Methods

We describe the parametric models used in this paper and the detailed estimation method for each model.

7.1.1 Conventional GARCH Models

The first model we estimate is the standard GARCH$(1,1)$, reproduced below:

$$\begin{cases} y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t, \\ \varepsilon_t = \sqrt{h_t}\cdot\eta_t,\ \eta_t \sim iid(0,1) \\ h_t = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \gamma_1 h_{t-1}. \end{cases} \tag{3}$$

Denote the parameters of interest as $\boldsymbol{\theta} = (\beta, \alpha_0, \alpha_1, \gamma_1)'$. Let $f(\eta_t;\nu)$ denote the density function for $\eta_t = \varepsilon_t(\boldsymbol{\theta})/\sqrt{h_t(\boldsymbol{\theta})}$ with mean 0, variance 1, and nuisance parameters $\nu \in \mathbb{R}^j$. The combined parameter vector is further denoted as $\boldsymbol{\psi} = (\boldsymbol{\theta}', \nu')'$. The likelihood function for the t-th observation is given by

$$f_t(y_t) = f_t(y_t;\boldsymbol{\psi}) = \frac{1}{\sqrt{h_t(\boldsymbol{\theta})}} f\left(\frac{\varepsilon_t(\boldsymbol{\theta})}{\sqrt{h_t(\boldsymbol{\theta})}};\nu\right).$$

When η_t is assumed to follow a standard normal, the probability density function (p.d.f.) is

$$f(\eta_t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\eta_t^2}{2}\right).$$

Alternatively, if η_t is assumed to be distributed according to the Student's t with ν degrees of freedom, the p.d.f. of η_t is then given by

$$f(\eta_t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{\eta_t^2}{\nu-2}\right)^{-\frac{(\nu+1)}{2}}, \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function and ν is constrained to be greater than 2 so that the second moment exists and equals 1. ν is a nuisance parameter that needs to be estimated.

Instead, if a GED distribution is assumed, the p.d.f. of η_t is

$$f(\eta_t;\nu) = \frac{\nu\exp\left[-\frac{1}{2}\left|\frac{\eta_t}{\lambda}\right|^{\nu}\right]}{\lambda 2^{\left(1+\frac{1}{\nu}\right)}\Gamma\left(\frac{1}{\nu}\right)}, \tag{5}$$

with

$$\lambda \equiv \left[\frac{\left(2^{-\frac{2}{\nu}}\Gamma\left(\frac{1}{\nu}\right)\right)}{\Gamma\left(\frac{3}{\nu}\right)}\right]^{\frac{1}{2}},$$

and ν defines the shape parameter indicating the thickness of the tails and satisfying $0 < \nu < \infty$. When $\nu = 2$, the GED distribution becomes a standard normal distribution. If $\nu < 2$, the tails are thicker than normal.

The Exponential GARCH (EGARCH) model is given by:

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

Note that the equation for the conditional variance takes a log-linear form. Thus, the implied value of h_t can never be negative, permitting the estimated coefficients to be negative. In addition, the level of the standardized value of ε_{t-1}, $\left|\varepsilon_{t-1}/\sqrt{h_{t-1}}\right|$, is used instead of ε_{t-1}^2.

Notice that in the EGARCH, $E\left|\varepsilon_{t-1}/\sqrt{h_{t-1}}\right|$ takes different values under different distribution specifications. When η_t is normal, $E\left|\varepsilon_{t-1}/\sqrt{h_{t-1}}\right|$ is the constant $\sqrt{\frac{2}{\pi}}$. Under the t distribution specified in (4),

$$\mathbb{E} \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| = \mathbb{E} \left| \eta_{t-1} \right| = \frac{2\sqrt{\nu-2}\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi} \cdot (\nu-1) \cdot \Gamma\left(\frac{\nu}{2}\right)}.$$

Under the GED distribution specified in (5),

$$\mathbb{E} \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| = \mathbb{E} \left| \eta_{t-1} \right| = \frac{\Gamma\left(\frac{2}{\nu}\right)}{\left[\Gamma\left(\frac{1}{\nu}\right)\Gamma\left(\frac{3}{\nu}\right)\right]^{1/2}}.$$

The conditional variance for the GJR-GARCH is given by

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

ML estimation of EGARCH and GJR-GARCH can be conducted similarly under different distributional specifications.

7.1.2 MS-GARCH

The MS-GARCH(1,1) specification is given as follows:

$$\begin{cases} y_t = \mu^{S_t} + \varepsilon^{S_t}, \\ \varepsilon^{S_t} = \sqrt{h_t^{S_t}} \cdot \eta_t, \ \eta_t \sim iid(0,1) \\ h_t^{S_t} = \alpha_0^{S_t} + \alpha_1^{S_t} \varepsilon_{t-1}^2 + \gamma_1^{S_t} h_{t-1}. \end{cases} \tag{6}$$

Parameter estimates for (6) can be obtained by maximizing the (conditional) log likelihood function

$$\begin{aligned} \mathcal{L} &= \sum_{t=1}^{T} \log f_{t-1}(y_t) \\ &= \sum_{t=1}^{T} \log \left[p_{t-1}(S_t = 1) f_{t-1}(y_t \mid S_t = 1) + p_{t-1}(S_t = 2) f_{t-1}(y_t \mid S_t = 2) \right], \end{aligned}$$

where $f_{t-1}(y_t \mid S_t = i)$ is the conditional density of y_t given regime i occurs at time t, and $p_{t-1}(S_t = i) = P(S_t = i \mid I_{t-1})$ are the ex-ante probabilities.

Recall the path dependent h_{t-1} is replaced by

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

Denote $p_{ji,t-1} = P\left(S_{t-1} = j \mid S_t = i, \mathcal{I}_{t-1}\right)$, which is calculated as

$$p_{ji,t-1} = \frac{p_{ji} P(S_{t-1} = j \mid \mathcal{I}_{t-1})}{P(S_t = i \mid \mathcal{I}_{t-1})} = \frac{p_{ji} p_{t-1}(S_{t-1} = j)}{p_{t-1}(S_t = i)},$$

where $p_{t-1}(S_{t-1} = j)$ can be computed as

$$\begin{aligned} p_{t-1}(S_{t-1} &= j) = P(S_{t-1} = j \mid y_{t-1}, \mathcal{I}_{t-2}) = \frac{P(S_{t-1} = j, y_{t-1} \mid \mathcal{I}_{t-2})}{f(y_{t-1} \mid \mathcal{I}_{t-2})} \\ &= \frac{f(y_{t-1} \mid S_{t-1} = j, \mathcal{I}_{t-2}) P(S_{t-1} = j \mid \mathcal{I}_{t-2})}{f_{t-2}(y_{t-1})} \\ &= \frac{f_{t-2}(y_{t-1} \mid S_{t-1} = j) p_{t-2}(S_{t-1} = j)}{f_{t-2}(y_{t-1})}. \end{aligned}$$

That is, $p_{t-1}(S_{t-1} = j)$ can be calculated recursively.

The ex-ante probability $p_{t-1}(S_t = i)$ in the log likelihood function follows immediately:

$$\begin{aligned} p_{t-1}(S_t &= i) = \sum_{j=1}^{2} P(S_t = i, S_{t-1} = j \mid \mathcal{I}_{t-1}) \\ &= \sum_{j=1}^{2} P(S_t = i \mid S_{t-1} = j, \mathcal{I}_{t-1}) P(S_{t-1} = j \mid \mathcal{I}_{t-1}) \\ &= \sum_{j=1}^{2} p_{ji} p_{t-1}(S_{t-1} = j). \end{aligned}$$

7.1.3 FIGARCH

The FIGARCH$(1, d, 1)$ model is reproduced here:

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

where $w_t = \varepsilon_t^2 - h_t$, and $\phi(L) = (1 - (\alpha_1 + \gamma_1)L)/(1 - L) \equiv (1 - \phi L)/(1 - L)$.

The conditional variance is as follows:

$$\begin{aligned} h_t &= \alpha_0 + \left(1 - \gamma_1 L - (1 - \alpha_1 L - \gamma_1 L)\left(1 - L\right)^{d-1}\right)\varepsilon_t^2 + \gamma_1 h_{t-1} \\ &= \alpha_0 + (\phi_1 L + \phi_2 L^2 + ...)\varepsilon_t^2 + \gamma_1 h_{t-1} \\ &= \alpha_0 + \phi_1 \varepsilon_{t-1}^2 + \phi_2 \varepsilon_{t-2}^2 + ... + \gamma_1 h_{t-1}. \end{aligned} \tag{7}$$

To solve for ϕ_j, we use the MacLaurin series:

$$\begin{aligned}(1-L)^{d-1} &= 1-\frac{(d-1)}{1!}L+\frac{(d-1)(d-2)}{2!}L^2-\frac{(d-1)(d-2)(d-3)}{3!}L^3+... \\ &= 1+\frac{(1-d)}{1!}L+\frac{(1-d)(2-d)}{2!}L^2+\frac{(1-d)(2-d)(3-d)}{3!}L^3+...\end{aligned}$$

Therefore, we can calculate the following sequences recursively:

$$\begin{array}{ll}\pi_1 = 1-d, & \phi_1 = -\pi_1+\alpha_1, \\ \pi_2 = \frac{(1-d)(2-d)}{2!}, & \phi_2 = -\pi_2+\pi_1(\alpha_1+\gamma_1), \\ \pi_3 = \frac{(1-d)(2-d)(3-d)}{3!}, & \phi_3 = -\pi_3+\pi_2(\alpha_1+\gamma_1), \\ \pi_4 = \frac{(1-d)(2-d)(3-d)(4-d)}{4!}, & \phi_4 = -\pi_4+\pi_3(\alpha_1+\gamma_1), \\ ... & ...\end{array}$$

The likelihood function is constructed conditional on initial values for ε_0^2, ε_{-1}^2, ... in (7) to be set at the unconditional sample variance. We choose the truncation lag at 1512, the in-sample window size.

7.2 Testing for Markov Switching

We follow Carrasco, Hu and Ploberger (2014) and illustrate how to test for regime switching in the mean and variance of the MS-GARCH model with a normal distribution. Specifically, under the null hypothesis (H_0) the model is given by (3) with a constant mean, whereas under the alternative (H_1) the model is given by (6).

The (conditional) log likelihood function under H_0 is

$$l_t = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1 h_{t-1}\right) - \frac{(y_t-\mu)^2}{2\left(\alpha_0+\alpha_1\varepsilon_{t-1}^2+\gamma_1 h_{t-1}\right)}. \tag{8}$$

We first obtain the MLE for the parameters $\hat{\theta}$ under H_0, where $\theta = (\mu, \alpha_0, \alpha_1, \gamma_1)'$. Then, we calculate the first and second derivatives of the log likelihood (8) with respect to θ evaluated at $\hat{\theta}$.

Note that the Markov chain S_t and the parameters driven by it $(\mu^{S_t}, \alpha_0^{S_t}, \alpha_1^{S_t}, \gamma_1^{S_t})'$ in the alternative model (6) are not present under H_0, therefore they cannot be consistently estimated and the test is nonstandard. Let ς denote the nuisance parameters specifying the alternative model, which are not identified under the null. In (6), ς consists of a constant c, which characterizes the amplitude of the alternative, and a vector $\zeta = (\eta, \rho : \|\boldsymbol{\eta}\| = 1, -1 < \underline{\rho} < \rho < \bar{\rho} < 1)$, where η is a normalized 4×1 vector that characterizes the direction of the alternative and ρ specifies the autocorrelation of the Markov chain. Given the nuisance parameters $\boldsymbol{\varsigma}$, Carrasco, Hu, and Ploberger (2014) first derive the test statistic process $\mu_{2,t}\left(\boldsymbol{\zeta}, \hat{\boldsymbol{\theta}}\right)$ by approximating the likelihood ratio, then they integrate out the process with respect to some prior distribution on $\boldsymbol{\zeta}$. Specifically, the first component

of their test is $\Gamma_T^* = \sum \mu_{2,t}\left(\boldsymbol{\zeta}, \hat{\boldsymbol{\theta}}\right)/\sqrt{T}$, and

$$\mu_{2,t}\left(\boldsymbol{\zeta}, \hat{\boldsymbol{\theta}}\right) = \frac{1}{2}\boldsymbol{\eta}'\left[\left(\frac{\partial^2 l_t}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \left(\frac{\partial l_t}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial l_t}{\partial\boldsymbol{\theta}}\right)'\right) + 2\sum_{s<t}\rho^{(t-s)}\left(\frac{\partial l_t}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial l_s}{\partial\boldsymbol{\theta}}\right)'\right]\boldsymbol{\eta}. \quad (9)$$

The second component, $\widehat{\epsilon}^*$, is the residual of the regression of $\mu_{2,t}\left(\boldsymbol{\zeta}, \hat{\boldsymbol{\theta}}\right)$ on $l_t^{(1)}\left(\hat{\boldsymbol{\theta}}\right)$. Then the sup test takes the form:

$$\text{supTS} = \sup_{\left\{\boldsymbol{\eta},\rho:\|\boldsymbol{\eta}\|=1,\underline{\rho}<\rho<\bar{\rho}\right\}} \frac{1}{2}\left(\max\left(0, \frac{\Gamma_T^*}{\sqrt{\widehat{\epsilon}^{*\prime}\widehat{\epsilon}^*}}\right)\right)^2. \quad (10)$$

Alternatively, the exp test is:

$$\text{expTS} = \underset{\left\{\boldsymbol{\eta},\rho:\|\boldsymbol{\eta}\|=1,\underline{\rho}<\rho<\bar{\rho}\right\}}{avg} \Psi\left(\boldsymbol{\eta}, \rho\right),$$

where

$$\Psi\left(\boldsymbol{\eta}, \rho\right) = \begin{cases} \sqrt{2\pi}\exp\left[\frac{1}{2}\left(\frac{\Gamma_T^*}{\sqrt{\widehat{\epsilon}^{*\prime}\widehat{\epsilon}^*}} - 1\right)^2\right]\Phi\left(\frac{\Gamma_T^*}{\sqrt{\widehat{\epsilon}^{*\prime}\widehat{\epsilon}^*}} - 1\right) & \text{if } \widehat{\epsilon}^{*\prime}\widehat{\epsilon}^* \neq 0, \\ 1 \text{ otherwise.} \end{cases}$$

That is, the unidentified nuisance parameters $\boldsymbol{\zeta}$ are integrated out with respect to some prior distributions to deliver an optimal test in the Bayesian sense.[24] The asymptotic distributions of the supTS and expTS tests are nonstandard; thus, the critical values are obtained by bootstrapping the empirical distributions.

To compute the test statistics, we use a uniform prior for η and generate the 4×1 vector uniformly over the unit sphere 100 times, corresponding to the switching mean and the three GARCH parameters.[25] The supTS is maximized with respect to $\boldsymbol{\eta}$ and ρ, where ρ takes on incremental values on the interval $[-0.95, 0.95]$ with the step length of 0.05. Meanwhile, expTS is the average of $\Psi\left(\boldsymbol{\eta}, \rho\right)$ above computed over those $\boldsymbol{\eta}$ and $\rho's$.

7.3 Forecast Evaluation Metrics

7.3.1 Statistical Loss Functions

We report the two loss functions, MSE and $QLIKE$ in the paper. Other commonly used loss functions include:

$$MSE_1 = n^{-1}\sum_{t=1}^{n}\left(\sigma_t - \widehat{h}_t^{1/2}\right)^2, \quad (11)$$

[24] The first part in (9) is the key component of the Information Matrix test commonly seen in testing for random coefficients, and the second part comes from the serial dependence of the time-varying coefficients. $\widehat{\epsilon}^{*\prime}\widehat{\epsilon}^*$ is the extra term to compensate for the difference in the likelihood ratio when we replace the true parameter $\boldsymbol{\theta}$ by its MLE $\hat{\boldsymbol{\theta}}$ under H_0. supTS is constructed from the supremum norm on c, $\boldsymbol{\eta}$ and ρ. expTS integrates out the exponential of test statistic process with an exponential prior on c^2 and uniform priors on $\boldsymbol{\eta}$ and ρ. Therefore, c seems to disappear from the formulae.

[25] To test for switching in the variance equation only, we can simply set the first element of $\boldsymbol{\eta}$ to be 0 and generate the remaining 3×1 vector uniformly over the unit sphere.

$$MAD_1 = n^{-1} \sum_{t=1}^{n} \left| \sigma_t - \widehat{h}_t^{1/2} \right|, \tag{12}$$

$$MAD_2 = n^{-1} \sum_{t=1}^{n} \left| \sigma_t^2 - \widehat{h}_t \right|, \tag{13}$$

$$R^2LOG = n^{-1} \sum_{t=1}^{n} \left[\log(\sigma_t^2 / \widehat{h}_t) \right]^2, \tag{14}$$

and

$$HMSE = n^{-1} \sum_{t=1}^{n} \left(\sigma_t^2 / \widehat{h}_t - 1 \right)^2. \tag{15}$$

Equations (12) and (13) are two Mean Absolute Deviation criteria. Equation (14) represents the logarithmic loss function of Pagan and Schwert (1990), whereas (15) is the heteroskedasticity-adjusted MSE proposed by Bollerslev and Ghysels (1996). Patton (2011) shows that only the MSE and $QLIKE$ loss functions generate optimal forecasts equal to the conditional variance σ_t^2, even when noisy volatility proxies are used in forecast evaluations. Nevertheless, for the proxy we choose, namely, the realized volatility constructed from the 5-minute returns over all the trading hours, the degree of distortion for other loss functions is also negligible. Test results for those loss functions are available from the authors upon request.

7.3.2 Success Ratio and Directional Accuracy

The percentage of times $\widehat{h}_t$ moves in the same direction as σ_t^2 is given by

$$SR = n^{-1} \sum_{t=1}^{n} \mathcal{I}_{\left\{\overline{\sigma_t^2} \cdot \overline{h_t} > 0\right\}},$$

where $\overline{\sigma_t^2}$ is the demeaned volatility at t, and $\overline{h_t}$ is the demeaned volatility forecast at t. If the volatility and the forecasted volatility move in the same direction, then $\mathcal{I}_{\{\omega>0\}}$ is equal to 1; 0 otherwise.

Having computed the SR, we calculate $SRI = P\widehat{P} + (1-P)(1-\widehat{P})$ where P is the fraction of times that $\overline{\sigma_t^2}$ is positive and $\widehat{P}$ is the fraction of times that $\overline{h_t}$ is positive. The DA test is given by

$$DA = \frac{SR - SRI}{\sqrt{\mathbb{V}(SR) - \mathbb{V}(SRI)}},$$

where $\mathbb{V}(SR) = n^{-1} SRI(1-SRI)$ and $\mathbb{V}(SRI) = n^{-1}(2\widehat{P}-1)^2 P(1-P) + n^{-1}(2P-1)^2 \widehat{P}(1-\widehat{P}) + 4n^{-2} P\widehat{P}(1-P)(1-\widehat{P})$. A significant DA statistic indicates the model forecast $\widehat{h}_t$ has predictive power for the direction of the movements in the underlying volatility σ_t^2.

7.3.3 Test of Equal Predictive Ability

Define the loss function $L(\widehat{h}_t, \sigma_t^2)$ where $\widehat{h}_t$ is the volatility forecast made when the underlying volatility is σ_t^2. Consider two sequences of forecasts generated by two competing models, i and j, $\left\{\widehat{h}_{i,t}\right\}_{t=1}^{n}$ and $\left\{\widehat{h}_{j,t}\right\}_{t=1}^{n}$. The loss differential between the two models is defined as $d_{ij,t} \equiv L_{i,t} - L_{j,t} = L(\widehat{h}_{i,t}, \sigma_t^2) - L(\widehat{h}_{j,t}, \sigma_t^2)$, where $L_{i,t} \equiv L(\widehat{h}_{i,t}, \sigma_t^2)$ denotes the loss function for the benchmark model i and $L_{j,t}$ is the loss function for the alternative model j. Giacomini and White (2006) show that if the parameters are estimated using a rolling scheme with a finite observation window, the asymptotic distribution of the sample mean loss differential $\overline{d} = n^{-1}\sum_{t=1}^{n} d_{ij,t}$ is asymptotically normal as long as $\{d_{ij,t}\}_{t=1}^{n}$ is covariance stationary with a short memory. So the Diebold-Mariano-West statistic for testing the null hypothesis of Equal Predictive Accuracy (EPA) between models i and j is $DMW = \overline{d}/\sqrt{\widehat{\mathbb{V}}(\overline{d})}$, where the asymptotic variance $\widehat{\mathbb{V}}(\overline{d})$ can be estimated by Newey-West's HAC estimator.[26] DMW has a standard normal distribution under H_0. If the test statistic DMW is significantly negative, the benchmark model is better since it has a smaller loss function; if DMW is significantly positive, then the benchmark model is outperformed.

7.3.4 Test of Superior Predictive Ability

Consider comparing $l+1$ forecasting models where model 0 is defined as the benchmark model and $k = 1, ..., l$ represent the l alternative models. Let $L_{k,t}$ and $L_{0,t}$ denote the loss when the k-th and the benchmark models are used to forecast the underlying volatility σ_t^2, respectively. The performance of the k-th forecast model relative to the benchmark is given by the loss differential

$$d_{0k,t} = L_{0,t} - L_{k,t}, \qquad k = 1, ..., l; \qquad t = 1, ..., T.$$

Under the assumption that $d_{0k,t}$ is stationary, the expected performance of model k relative to the benchmark can be defined as $\mu_k = \mathbb{E}\left[d_{0k,t}\right]$ for $k = 1, ..., l$. The value of μ_k will be positive for any model k that outperforms the benchmark. Hence, the null hypothesis for testing whether any of the competing models significantly outperforms the benchmark is defined in terms of μ_k for $k = 1, ..., l$ as:

$$H_0 : \mu_{\max} \equiv \max_{k=1,...,l} \mu_k \leq 0.$$

The alternative is that the best model has a smaller loss function relative to the benchmark. If the null is rejected, then there is evidence that at least one of the competing models has a significantly smaller loss function than the benchmark.

[26] $\widehat{\mathbb{V}}(\overline{d}) = n^{-1}\left(\widehat{\gamma} + 2\sum_{k=1}^{q}\omega_k\widehat{\gamma}_k\right)$, where $q = h-1$, $\omega_k = 1 - \frac{k}{q+1}$ is the lag window and $\widehat{\gamma}_i$ is an estimate of the i-th order autocovariance of the series $\{d_t\}$, where $\widehat{\gamma}_k = \frac{1}{n}\sum_{t=k+1}^{n}\left(d_t - \overline{d}\right)\left(d_{t-k} - \overline{d}\right)$ for $k = 1, ..., q$.

White's RC test is defined as

$$T_n^{RC} \equiv \max_{k=1,\ldots,l} n^{\frac{1}{2}} \bar{d}_k,$$

where $\bar{d}_k = n^{-1} \sum_{t=1}^{n} d_{0k,t}$. T_n^{RC}'s asymptotic null distribution is normal with mean 0 and some long-run variance Ω.

Note that the T_n^{RC}'s asymptotic distribution relies on the assumption that $\mu_k = 0$ for all k, however, any negative values of μ_k would also conform with H_0. Hansen (2005) proposes an alternative Super Predictive Ability (SPA) test statistic*:*

$$T_n^{SPA} = \max_{k=1,\ldots,l} \frac{n^{\frac{1}{2}} \bar{d}_k}{\sqrt{\widehat{\mathbb{V}}(n^{\frac{1}{2}} \bar{d}_k)}},$$

where $\widehat{\mathbb{V}}(n^{\frac{1}{2}} \bar{d}_k)$ is a consistent estimator of the variance of $n^{\frac{1}{2}} \bar{d}_k$ obtained via bootstrap. The distribution under the null is $N(\hat{\mu}, \Omega)$, where $\hat{\mu}$ is a chosen estimator for μ that conforms with H_0. Since different choices of $\hat{\mu}$ would result in different p-values, Hansen proposes three estimators $\hat{\mu}^l \leq \hat{\mu}^c \leq \hat{\mu}^u$. We name the resulting tests SPA_l, SPA_c, and SPA_u, respectively. SPA_c would lead to a consistent estimate of the asymptotic distribution of the test statistic. SPA_l uses the lower bound of $\hat{\mu}$ and the p-value is asymptotically smaller than the correct p-value, making it a liberal test. In other words, it is insensitive to the inclusion of poor models. In contrast, SPA_u uses the upper bound of $\hat{\mu}$ and it is a conservative test. It has the same asymptotic distribution as the RC test and is sensitive to the inclusion of poor models.

7.3.5 Model Confidence Set

Given the loss differential $d_{ij,t} = L_{i,t} - L_{j,t}$ for $i, j \in \mathcal{M}_0$ and $\mu_{ij} = \mathbb{E}\left[d_{ij,t}\right]$, the set of superior objects is defined as

$$\mathcal{M}^* = \left\{ i \in \mathcal{M}_0 : \mu_{ij} \leq 0 \text{ for all } j \in \mathcal{M}_0 \right\}.$$

The EPA hypothesis for a given set of models $\mathcal{M}$ can be formulated in two ways:

$$\begin{aligned} H_{0,\mathcal{M}} &: \mu_{ij} = 0 \text{ for all } i, j \in \mathcal{M} \subset \mathcal{M}_0, \\ H_{A,\mathcal{M}} &: \mu_{ij} \neq 0 \text{ for some } i, j \in \mathcal{M} \subset \mathcal{M}_0, \end{aligned} \tag{17}$$

or

$$\begin{aligned} H_{0,\mathcal{M}} &: \mu_{i.} = 0 \text{ for all } i, j \in \mathcal{M} \subset \mathcal{M}_0, \\ H_{A,\mathcal{M}} &: \mu_{i.} \neq 0 \text{ for some } i, j \in \mathcal{M} \subset \mathcal{M}_0, \end{aligned} \tag{18}$$

where $\bar{d}_{ij} = n^{-1} \sum_{t=1}^{n} d_{ij,t}$, $\bar{d}_{i.} = m^{-1} \sum_{j \in \mathcal{M}} \bar{d}_{ij}$ and $\mu_{i.} = \mathbb{E}(d_{i.})$. According to Hansen, Lunde and Nason (2001), we construct the t-statistics as in the EPA test for testing the pair (17):

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\mathbb{V}}(\bar{d}_{ij})}}, i, j \in \mathcal{M}.$$

Similarly, to test (18), the t-statistics is

$$t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\mathbb{V}}(\bar{d}_{i\cdot})}}, i, j \in \mathcal{M},$$

where $\bar{d}_{i\cdot}$ is the sample loss of the i-th model relative to the average across models in $\mathcal{M}$, and $\widehat{\mathbb{V}}(\bar{d}_{i\cdot})$ is the estimate of $\mathbb{V}(\bar{d}_{i\cdot})$.

Then the null hypotheses in (17) and (18) map to the following two test statistics respectively:

$$T_{R,\mathcal{M}} = \max_{i,j\in\mathcal{M}} |t_{ij}| \text{ and } T_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} t_{i\cdot}.$$

The asymptotic distributions of $T_{R,\mathcal{M}}$ and $T_{\max,\mathcal{M}}$ are nonstandard and can be simulated through bootstrap. The elimination rules applied are

$$e_{R,\mathcal{M}} = \arg\max_{i\in\mathcal{M}} \left\{ \sup_{j\in\mathcal{M}} t_{ij} \right\} \text{ and } e_{\max,\mathcal{M}} = \arg\max_{i\in\mathcal{M}} \{t_{i\cdot}\}.$$

References

[1] Bollerslev, T, and E. Ghysels (1996), "Periodic autoregressive conditional heteroscedasticity," *Journal of Business and Economic Statistics* 14(2), 139-151.

[2] Pagan, A. R. and G. W. Schwert (1990), "Alternative models for conditional stock volatility," *Journal of Econometrics* 45(1), 267-290.

Table A.1: Reality Check and Superior Predictive Ability Tests

		One Day		Five Days		Twenty-one Days		Sixty-three Days	
Benchmark		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
	SPAl	0.535	0	0.650	0.856	0.693	0.882	0.755	0.869
GARCH-N	*SPAc*	0.177	0	0.208	0	0.254	0	0.312	0.001
	SPAu	0.211	0	0.281	0	0.358	0	0.462	0.001
	SPAl	0.631	0.007	0.774	0.850	0.845	0.880	0.883	0.883
GARCH-*t*	*SPAc*	0.258	0.005	0.324	0.004	0.381	0.003	0.426	0.005
	SPAu	0.294	0.007	0.384	0.005	0.466	0.009	0.562	0.013
	SPAl	0.593	0.004	0.739	0.853	0.791	0.891	0.847	0.878
GARCH-GED	*SPAc*	0.207	0.003	0.282	0.002	0.348	0.001	0.405	0.001
	SPAu	0.233	0.004	0.336	0.002	0.513	0.002	0.505	0.002
	SPAl	0.483	0	0.397	0.849	0.422	0.885	0.385	0.882
EGARCH-N	*SPAc*	0.188	0	0.182	0	0.204	0	0.149	0
	SPAu	0.188	0	0.182	0	0.204	0	0.149	0
	SPAl	0.680	0	0.673	0.868	0.721	0.882	0.735	0.878
EGARCH-*t*	*SPAc*	0.280	0	0.315	0	0.36	0	0.345	0.002
	SPAu	0.289	0	0.319	0	0.374	0	0.354	0.003
	SPAl	0.547	0	0.494	0.853	0.535	0.879	0.547	0.874
EGARCH-GED	*SPAc*	0.227	0	0.249	0	0.256	0	0.260	0
	SPAu	0.234	0	0.249	0	0.256	0	0.260	0
	SPAl	0.379	0	0.137	0.850	0.131	0.887	0.113	0.880
GJR-N	*SPAc*	0.102	0	0.028	0	0.028	0	0.026	0
	SPAu	0.102	0	0.028	0	0.028	0	0.026	0
	SPAl	0.444	0	0.289	0.846	0.291	0.886	0.305	0.886
GJR-*t*	*SPAc*	0.125	0	0.093	0	0.084	0	0.084	0
	SPAu	0.125	0	0.093	0	0.084	0	0.084	0
	SPAl	0.393	0	0.188	0.852	0.211	0.886	0.182	0.878
GJR-GED	*SPAc*	0.106	0	0.043	0	0.052	0	0.039	0
	SPAu	0.106	0	0.043	0	0.052	0	0.039	0

: Note: This table presents the p-values of White's (2000) Reality Check test, and Hansen's (2005) Superior Predictive Ability test. The *SPAl*, *SPAc*, *SPAu* are the lower, consistent, and upper p-values from Hansen (2005), respectively. Each row contains the benchmark model. The null hypothesis is that none of the alternative models outperform the benchmark. The p-values are calculated using 3000 bootstrap replications with a block length of 2.

Table A.2: Reality Check and Superior Predictive Ability Tests

		One Day		Five Days		Twenty-one Days		Sixty-three Days	
Benchmark		MSE	QLIKE	MSE	QLIKE	MSE	QLIKE	MSE	QLIKE
	SPAl	0.564	0	0.684	0.853	0.741	0.887	0.789	0.881
MS-GARCH-N	*SPAc*	0.166	0	0.237	0	0.290	0	0.373	0
	SPAu	0.188	0	0.271	0	0.290	0	0.373	0
	SPAl	0.534	0	0.572	0.858	0.624	0.885	0.644	0.887
MS-GARCH-*t*	*SPAc*	0.196	0	0.215	0	0.233	0	0.240	0
	SPAu	0.196	0	0.215	0	0.237	0	0.251	0.001
	SPAl	0.390	0	0.264	0.846	0.287	0.872	0.291	0.883
MS-GARCH-GED	*SPAc*	0.050	0	0.037	0	0.043	0	0.043	0
	SPAu	0.050	0	0.037	0	0.043	0	0.048	0
	SPAl	0.528	0	0.009	0.001	0.009	0	0.005	0.051
FIGARCH-N	*SPAc*	0.124	0	0.009	0.001	0.009	0	0.005	0
	SPAu	0.175	0	0.009	0.001	0.009	0	0.005	0
	SPAl	0.714	0	0.012	0.007	0.008	0.002	0.004	0.003
FIGARCH-*t*	*SPAc*	0.232	0	0.012	0	0.008	0.002	0.004	0
	SPAu	0.307	0	0.012	0	0.008	0.002	0.004	0
	SPAl	0.146	0	0.01	0	0.008	0.003	0.004	0.005
FIGARCH-GED	*SPAc*	0.146	0	0.010	0	0.008	0.001	0.004	0.005
	SPAu	0.146	0	0.01	0	0.008	0.003	0.004	0.005
	SPAl	0.984	1	0.977	1	0.979	1	0.980	1
RiskMetrics	*SPAc*	0.687	0.690	0.692	0.653	0.663	0.655	0.671	0.610
	SPAu	0.687	1	0.692	1	0.663	1	0.827	1

: Note: This table presents the p-values of White's (2000) Reality Check test, and Hansen's (2005) Superior Predictive Ability test. The *SPAl*, *SPAc*, *SPAu* are the lower, consistent, and upper p-values from Hansen (2005), respectively. Each row contains the benchmark model. The null hypothesis is that none of the alternative models outperform the benchmark. The p-values are calculated using 3000 bootstrap replications with a block length of 2.