Forecasting Crude Oil Price Volatility

Ana María Herrera[*] Liang Hu[†] Daniel Pastor[‡]

December 1, 2015

Abstract

We use high-frequency intra-day realized volatility to evaluate the relative forecasting performance of time invariant and Markov switching models for the volatility of crude oil daily spot returns. First, we implement Carasco, Hu and Ploberger's (2014) test for regime switching in the mean and variance of the GARCH(1,1), finding overwhelming support for a Markov switching model. We then perform a comprehensive out-of-sample forecasting performance evaluation using a battery of pairwise and groupwise test. We find that under the MSE_2 and the $QLIKE$ loss functions both types of tests favor the MS-GARCH-t. However, the EGARCH-t model emerges as a close competitior when other loss functions are used. This result is estabilished by computing the Equal Predictive Ability of Diebold and Mariano(1995), the Reality Check of White (2000), the test of Superior Predictive Ability of Hansen (2005) 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 computed using a rolling window suggests that MS-GARCH-t model is better at predicting volatility during periods of turmoil.

Keywords: Crude oil price volatility, GARCH, Markov switching, forecast.

JEL codes: C22, C53, Q47

1 Introduction

Crude oil price returns have fluctuated greatly during the last decade. In particular, the volatility of the daily West Texas Intermediate (WTI) spot returns surged during the financial crisis, then decreased for a few years and has increased again since the second semester of 2014 (see Figure 1). Although surges in the volatility of crude oil returns

[*] Department of Economics, University of Kentucky, 335Z 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, Wayne State University, 2103 Faculty Administration Building, 656 W. Kirby, Detroit, MI 48202; e-mail: daniel.pastor@wayne.edu; phone: (313) 971-3046; fax: (313) 577-9564

have been observed before, notably around the 1986 oil price collapse and the Gulf War, a natural question is whether the econometric tools that we possess nowadays allow us to generate reliable forecasts, especially during periods of turmoil.

A large number of studies on forecasting crude oil prices have focused on predicting the spot price. This is natural as it is often the price that constitutes an input used by economic analysts and policy makers in producing macroeconomic forecasts (Hamilton 2009, Edelstein and Kilian 2009). Indeed, this direction of research has provided important insights into the usefulness of macroeconomic aggregates, asset prices, and futures prices in forecasting the spot price of oil, as well as into the extent to which real and nominal oil prices are predictable.[1]

However, reliable forecasts of spot oil price volatility are also of interest for various economic agents as "[s]pot oil price volatility reflects the volatility of current as well as future values of [oil] production, consumption and inventory demand" (Pindyck 2004). First and most obviously, accurate forecasts are key for those firms whose business greatly depends on oil prices. Examples include 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, such as automobiles or heating systems (Kahn 1986, Davis and Kilian 2011). Last but not least, they are useful for those whose daily task is to produce forecasts of industry-level and aggregate economic activity, such as central bankers, business economists, and private sector forecasters. Indeed, reliable forecasts of oil price volatility are a crucial input for monetary authorities whose mandate is to stabilize inflation and promote output growth.

The aim of this paper is to evaluate the performance of different volatility models for the conditional variance (hereafter variance) of spot crude oil returns, where we replace the unobserved variance with the realized volatility of intra-day returns (Andersen and Bollerslev 1998). More specifically, we investigate the out-of-sample predictive ability of time-invariant and Markov switching GARCH (MS-GARCH) models. The motivation for focusing on this class of models is twofold. On one hand, time invariant GARCH(1,1) models have 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). In addition, nonlinear 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 at short horizons (Mohammadi and Su 2010, and Hou and Suardi 2012). On the other hand, oil prices are characterized by sudden jumps due to, for instance, political disruptions in the Middle East or military interventions in oil exporting countries. Markov switching models have been found to be better suited to model situations where changes in regimes are triggered by those sudden shocks to the economy. Yet, it remains an open question whether MS-GARCH models can beat the GARCH(1,1) or other time invariant GARCH models in

[1] See e.g. Alquist, Kilian and Vigfusson (2013) for a comprehensive study and a survey of the literature.

forecasting the volatility of spot crude oil returns. Moreover, how does predictive ability of the different models compare during periods of calm and periods of turmoil such as that experienced in the late 2014?

To answer these questions this paper provides a comprehensive analysis of forecast comparisons of various volatility models for crude oil returns. We start by formally testing for regime switches using the procedure proposed by Carrasco, Hu, and Ploberger (2014). Implementing such test in GARCH models is especially important since high persistence in the unconditional variance of crude oil returns may be the result of neglected structural breaks or regime changes (see, e.g. Lamoureux and Lastrapes 1990). In addition, Caporale, Pittis, and Spagnolo (2003) show via Monte Carlo studies that fitting (mis-specified) GARCH models to data generated by a MS-GARCH process tends to produce Integrated GARCH (IGARCH)[2] parameter estimates, leading to erroneous conclusions about the persistence levels. Indeed, we find overwhelming evidence in favor of a regime switching model for the daily crude oil price data.

We then provide a complete picture of the relative out-of-sample forecasting performance of the competing volatility models by reporting several statistical loss functions (e.g., mean square error, MSE, mean absolute deviation, MAD, quasi maximum likelihood, $QLIKE$) for a battery of tests. In particular, we compute the Success Ratio (SR) and implement the Directional Accuracy (DA) tests of Pesaran and Timmermann (1992), conduct pairwise comparisons between different candidate models with Diebold and Mariano's (1995) test of Equal Predictive Ability, and groupwise comparisons using White's (2000) Reality Check test and Hansen's (2005) test of Superior Predictive Ability. In addition, we employ Hansen, Lunde and Nason (2011)'s Model Confidence Set procedure to determine the best set of model(s) from a collection of time-invariant and time-varying models.

Lastly, we inquire into the stability of the forecasting accuracy for the preferred models over the evaluation period (2013-2014). As already mentioned, this period includes a large increase in oil price volatility, which could have had an effect on the risk exposure of producers and consumers of oil, on investment decisions of households and firms, as well as on the economic forecasts that shape the course of monetary policy.

Our primary finding is that, under the two most widely used losses in volatility forecasting (MSE_2 and $QLIKE$), both pairwise and groupwise tests favor the MS-GARCH-t model. In particular, the time-invariant GARCH models have lower predictive ability and yield less accurate forecasts, regardless of the horizon of interest (1, 5, 21 or 63 days). However, if we consider all six loss functions, the EGARCH-t emerges as a close competitor at the one- and five-day forecast horizons. We also find overwhelming evidence that a normal innovation is insufficient to account for the leptokurtosis in our data, thus Student's t or GED distributions are more appropriate.[3] Moreover, when comparing the Mean Squared Prediction Error ($MSPE$) of the preferred models, we uncover clear gains from using the MS-GARCH-t model for forecasting crude oil price volatility during peri-

[2] The conditional variance grows with time t and the unconditional variance becomes infinity.

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

ods of turmoil. Perhaps not surprisingly, this model does a better job at forecasting the transition into a period of higher volatility at both short and long horizons.

Our paper is close in spirit to the work of Fong and See (2002), Nomikos and Pouliasis (2011) and Wang, Wu and Yang (2016) that evaluate the out-of-sample forecasting performance of Markov switching volatility models for daily crude oil returns. While our paper clearly benefits from the insights of these papers –as well as other papers on volatility forecasting–, it differs in two important dimensions. Substantively, our object of interest is the volatility of spot returns at both short and long horizons. In contrast, Fong and See (2002) and Nomikos and Pouliasis (2011) focus exclusively on one-day ahead forecasts of the volatility of crude oil futures. Wang, Wu and Yang (2016), on the other hand, study the volatility of spot returns but use a rather noisy proxy for the unobserved volatility, squared daily returns. As a technical matter, Fong and See (2002) evaluate the performance MS-GARCH models by following Gray's (1996) suggestion to integrate out the unobserved regime paths. Nomikos and Pouliasis (2011) use the estimation method proposed by Haas et al. (2004), where they simplify the regime shifting mechanism to make the estimation computationally tractable. In contrast, Wang, Wu and Yang (2016) compare the out-of-sample performance of a Markov switching multifractal volatility model (Calvet and Fisher 2001) vis-à-vis set of GARCH-class models.

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. Estimation results are presented in Section 4. Section 5 discusses the out-of-sample forecast evaluation. Section 6 concludes.

2 Volatility Models

This section briefly reviews the volatility models used to compare the predictive ability in section 5. We focus on four different horizons (one, five, twenty one and sixty three days), which are of interest for different consumers of crude oil volatility forecasts.[4] We consider both conventional GARCH-class models and MS-GARCH models. The former group of models, especially the GARCH(1,1), has been shown to have good predictive ability for stock returns (Hansen and Lunde, 2005) and perform well in forecasting volatility of crude oil at short horizons (Mohammadi and Su 2010, and Hou and Suardi 2012). As for the MS-GARCH model, because the GARCH parameters are permitted to switch between regimes (e.g., periods that are perceived as of major political unrest versus periods of calm) they provide flexibility over the standard GARCH models, and might be better at capturing two features of volatility: (a) persistence by allowing shocks to have a longer lasting effects during the high volatility regime and (b) pressure-relieving effects of some large shocks when they are followed by relatively tranquil periods.

[4] Financial investors are likely to rely more on short term one and five-day forecasts. However, central bankers typically use monthly forecasts. For oil exploration and production firms longer horizons are of interest as the time spanned 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.

2.1 Conventional GARCH Models

The first model we employ in this paper is the standard GARCH$(1,1)$ proposed by Bollerslev (1986):

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

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 $\alpha_1 + \gamma_1 \leq 1$.[5]

Another model used in this paper is the Exponential GARCH (EGARCH) model introduced by Nelson (1991) where the logarithm of the conditional variance is defined as

$$\log(h_t) = \alpha_0 + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| - E \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| \right) + \xi \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \log(h_{t-1}). \tag{2}$$

Note that the equation for the conditional variance is in 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. The EGARCH model allows for an asymmetric effect, which is measured by a significant ξ. The effect of a positive standardized shock on the logarithmic conditional variance is $\alpha_1 + \xi$; the effect of a negative standardized shock would be $\alpha_1 - \xi$ instead.

Finally, we also consider the GJR-GARCH model developed by Glosten, Jagannathan, and Runkle (1993) given by

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

where $\mathcal{I}_{\{\omega\}}$ is the indicator function equal to one if ω is true, and zero otherwise. Here the asymmetric effect is characterized by a significant ξ.

All in all, an attractive feature of the EGARCH and GJR-GARCH models is that they allow for an asymmetric effect of positive and negative shocks on the conditional variance. In the case of crude oil prices, political disruptions in the Middle East or large decreases in global demand tend to increase volatility (see, e.g. Ferderer 1996, Wilson et al. 1996) whereas the effect of new oil field discoveries seems to have a more muted effect. (Note how Figure 1 reveals a large increase in the volatility of WTI crude oil returns around the global financial crisis.)

We estimate the above described models over three different distributional specifications for η_t, standard normal, Student's t, or Generalized Error Distribution (GED). The first distribution constitutes a natural benchmark whereas both Student's t and GED are able to capture extra leptokurtosis, which is observed in oil price returns (see Table 1).

[5] When $\alpha_1 + \gamma_1 = 1$, ε_t becomes an integrated GARCH process, where a shock to the variance will remain in the system. However, it is still possible for it to come from a strictly stationary process, see Nelson (1990).

2.2 MS-GARCH

The last volatility model considered here follows Klaassen's (2002) modification of Gray's (1996) MS-GARCH$(1,1)$ model and is given by

$$\begin{cases} y_t = \mu^{S_t} + \varepsilon_t, \\ \varepsilon_t = \sqrt{h_t} \cdot \eta_t,\ \eta_t \sim iid(0,1) \\ h_t = \alpha_0^{S_t} + \alpha_1^{S_t} \varepsilon_{t-1}^2 + \gamma_1^{S_t} h_{t-1}, \end{cases} \tag{3}$$

where both the conditional mean $\mu^{S_t}$and the conditional variance h_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 $0 < p_{22} < 1$.

Estimating the model in (3) is computationally intractable, because the conditional variance h_t depends on the state-dependent h_{t-1}, consequently on all past states. Indeed, maximizing the likelihood function is infeasible as it would require integrating out all possible unobserved regime paths, which grow exponentially with sample size T. To address this issue we follow Klaassen (2002)[6] and replace h_{t-1} by its expectation conditional on the information set at $t-1$ plus the current state variable, namely,

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

where

$$E_{t-1}\left[h_{t-1}^{(i)} \mid S_t\right] = \sum_{j=1}^{2} p_{ji,t-1}\left[\left(\mu_{t-1}^{(j)}\right)^2 + h_{t-1}^{(j)}\right] - \left[\sum_{j=1}^{2} p_{ji,t-1}\mu_{t-1}^{(j)}\right]^2,$$

$p_{ji,t-1} = P\left(S_{t-1} = j \mid S_t = i, \mathcal{I}_{t-2}\right),\ i,j = 1,2,$ and is calculated as

$$p_{ji,t-1} = \frac{p_{ji}\Pr(S_{t-1} = j \mid \mathcal{I}_{t-2})}{\Pr(S_t = i \mid \mathcal{I}_{t-2})} = \frac{p_{ji}p_{j,t-1}}{\sum_{j=1}^{2} p_{ji}p_{j,t-1}}.$$

This specification circumvents the path dependence by integrating out the path-dependent h_{t-1} and it uses the information set at time $t-1$ plus the current state S_t.

[6] The estimation method used here differs from Fong and See (2002) and Nomikos and Pouliasis (2011). The former follow Gray's (1996) suggestion and integrate out the unobserved regime paths. Instead, Nomikos and Pouliasis (2011) use the procedure proposed by Haas et al. (2004), where they simplify the regime shift mechanism such that the autoregressive behavior in each regime is subject to the conditional variance staying in the same regime. Consequently the path dependence is removed and the estimation is done using standard MLE.

Given that regimes are often observed to be highly persistent in the volatility of crude oil returns, S_t contains lots of information about S_{t-1}. The m-step-ahead volatility forecasts at time T can be computed as:[7]

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

where the τ-step-ahead volatility forecast in regime i made at time T can be calculated by

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

Note that this computation is done through a recursive procedure similar to the standard GARCH models. With this approach, Klaassen (2002) shows that the MS-GARCH model provides significantly better volatility forecasts than GARCH model in investigating exchange rates.

Our choice of estimation method is driven by our interest in forecast horizons that exceed a day. Clearly, there are alternative estimation methods for MS-GARCH models. These alternatives 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;[8] (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, none of these alternatives provide a straightforward way to compute multi-step-ahead volatility forecasts.[9]

Because oil price returns exhibit leptokurtosis, and to maintain comparability between the GARCH and MS-GARCH models, we also consider three different types of distributions for η_t: normal, Student's t, and GED distributions, which can be easily accommodated by Klaassen's method as well.

[7] The m-step-ahead volatility is the summation of the volatility at each step because of the absence of serial correlation in oil price returns.

[8] Computation of multi-step-ahead forecast with Gray's specification involves calculating $E_{T-1}\left[h^{(i)}_{T,T+\tau-1}\right]$ instead, which is very complicated.

[9] In addition, Klassen's method has some statistical advantages relative to the other methods. For instance, it makes a more efficient use of all information available to the researcher than Gray (1996). As for Bayesian estimation using the MCMC algorithms, the effectiveness of the algorithms relies heavily on the starting values. In addition, the label-switching problem can seriously complicate the Bayesian statistical inference. Namely, the MS models are generally not identifiable since the parameters from the two states may be exchanged without affecting the likelihood. When the Bayesian MCMC techniques are used to estimate the MS models, the posterior distribution of the parameters is invariant under a permutation of state indices when exchangeable priors are used (see Frühwirth-Schnatter 2006 for details).

3 Data Description

We use the daily spot price for the West Texas Intermediate (WTI) crude oil obtained from the U.S. Energy Information Administration. The sample period ranges from July 1, 2003 to April 2, 2015; the start of the sample coincides with the period when oil futures began to trade around the clock. Over this period of time the average price for a barrel of crude oil was $75.39, the median value equaled $76.08, and the standard deviation was $23.97. A maximum price of $145.31 was observed on July 3, 2008 and a minimum of $26.93 on September 19, 2003. 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. Table 1 shows the descriptive statistics for WTI rates of return. The mean rate of return was 0.0162 with a standard deviation of 2.34. Note also that WTI returns are slightly negatively skewed. The kurtosis equals 7.91, which is high compared to 3 for a normal distribution.[10] Figure 1 plots the returns of the WTI spot prices and the squared deviations over the sample period. Large variations are observed during the global financial crisis in late 2008 and since crude oil prices started decreasing in July 2014. Indeed, Figure 1 suggests crude oil returns are characterized by periods of low volatility followed by high volatility in the face of major political or financial unrest.

The object of interest here is the true volatility of crude oil returns, which is unobserved. Hence, to evaluate the out-of-sample performance of the various volatility models, we compute an estimated measure of the realized volatility using high-frequency intra-day returns on oil futures (see Andersen and Bollerslev, 1998). More specifically, we obtain 5-minute prices of 1-month WTI oil futures contracts series from TickData.com spanning the period between July 1, 2003 (when this futures contract started trading) and April 2, 2015. 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. Following Blair, Poon and Taylor (2001), we construct the daily realized volatility RV_t by summing the squared 5-minute returns over the trading hours.[11] Then, to calculate m-step-ahead realized volatility at time T, we simply sum the daily realized volatility over

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

[11] Hansen and Lunde (2005) suggest 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 directly adding the overnight returns. However, it is less suitable here for two reasons. 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). In addition, the value for $\widehat{c}$ varies over the sample period. For instance, $\widehat{c} = 1.12$ for the period between 7/1/2003 and 4/2/2013, whereas $\widehat{c} = 1.17$ for our out-of-sample period 1/3/2012 to 4/2/2015, . Nevertheless, 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.

m days, denoted by:

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

We list the summary statistics for both the $RV_t^{1/2}$ and the logarithm of $RV_t^{1/2}$ in Table 1. The $RV_t^{1/2}$ series is severely right-skewed and leptokurtic. However, the logarithmic series appears much closer to a normal distribution, which is further confirmed by comparing its kernel density estimates with the normal distribution in Figure 2.

4 Estimation Results

We estimate the models by setting the conditional mean to be $r_t = \mu + \varepsilon_t$. Testing the residuals from such a simple specification reveals very small autocorrelations yet tremendous ARCH effect.

4.1 Conventional GARCH models

The ML estimates and asymptotic standard errors (in parenthesis) for the GARCH$(1,1)$, EGARCH$(1,1)$, and GJR-GARCH$(1,1)$ models are reported in Table 2. The conditional mean in the GARCH models is significantly positive at around 0.1 regardless of the distribution. The estimated conditional mean is lower for the EGARCH and GJR-GARCH but still significantly positive when the t or GED distributions are used. Three features of the estimated models are worth noticing. First, the degrees of freedom for the t distribution are estimated at around 8.6 in all GARCH models[12] and the estimated shape parameter for GED distribution is around 1.48[13]. This is consistent with the high kurtosis of daily crude oil returns (7.90) and, in turn, with the inability of a normal error to account for all the mass in the tails in the distribution.

Second, the asymmetric effect (ξ) is significant in 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.

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; similarly 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.

[12] This suggests that the conditional moments exist up to the 8th order. Morever, since the conditional kurtosis for the t distribution is calculated by $3(\nu - 2)/(\nu - 4)$, $\nu = 8.6$ implies fatter tails than normal distributions.

[13] 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 $\nu = 1.48$, the kurtosis is at 3.81, again confirming fat tails.

4.2 MS-GARCH

Lamoureux and Lastrapes (1990) and Mikosch and Starica (2004) have shown that the high persistence observed in the variance of financial returns can be explained by time-varying GARCH parameters. It is natural to investigate whether the high persistence in the crude oil returns is also the consequence of neglected breaks or regime switches. Studies that estimate MS-GARCH models for oil price returns (e.g. Fong and See 2002, Vo 2009, and Nomikos and Pouliasis 2011) or a stock price index (e.g. Marcucci 2005), proceed to estimate the MS-GARCH models without testing for the existence of regime switching. In fact, testing for Markov switching in GARCH models is complicated mainly 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 (log) likelihood functions are not calculable. Second, standard tests suffer from the famous Davies problem, where the nuisance parameters characterizing the regime switching are not identified under the null. Therefore, standard tests like the Wald or LR test do not have the usual χ^2 distribution. Markov switching tests by e.g., Hansen (1992) or Garcia (1998) are not applicable here either since they both involve examining the distribution of the likelihood ratio statistic, which is not feasible for MS-GARCH. We adopt the testing procedure developed by Carrasco, Hu, and Ploberger (2014). The advantage of this test is that it 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 the sense that it is asymptotically equivalent to the LR test. In addition, it has the flexibility to test for regime switching in both the means and the variances or any subset of these parameters. We describe in detail how to conduct their test for regime switching in mean and variances with a normal distribution. Specifically, under the null hypothesis (H_0) the model is given by *(7)* with constant mean**;** under the alternative (H_1) the model is (3).

Given model *(7)*, the (conditional) log likelihood function under H_0 is

$$l_t = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \gamma_1 h_{t-1}\right) - \frac{(y_t-\mu)^2}{2\left(\alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \gamma_1 h_{t-1}\right)}. \tag{4}$$

We first obtain the MLE for the parameters $\hat{\boldsymbol{\theta}}$ under H_0, where $\boldsymbol{\theta} = (\mu, \alpha_0, \alpha_1, \gamma_1)'$. Then, we calculate the first and second derivatives of the log likelihood (4) with respect to $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$.

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 (3) are not present under H_0, therefore we cannot consistently estimate them. This problem is called the Davies problem and standard tests like the Wald or the LR test do not have the usual Chi-squared distribution. The test proposed by Carrasco, Hu and Ploberger (2014) is in essence a Bayesian test: given the nuisance parameters $\boldsymbol{\zeta}$ not identified under the null, they 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 $\boldsymbol{\zeta}$. More specifically, the nuisance parameters specifying the alternative model consist of a constant c, which characterizes the amplitude of the alternative, and a vector $\boldsymbol{\zeta} =$

$(\boldsymbol{\eta}, \rho : \|\boldsymbol{\eta}\| = 1, -1 < \underline{\rho} < \rho < \bar{\rho} < 1)$, where $\boldsymbol{\eta}$ is a normalized 4×1 vector[14] that characterizes the direction of the alternative and ρ specifies the autocorrelation of the Markov chain. Given $\boldsymbol{\zeta}$, the first component of Carrasco, Hu, and Ploberger (2014) test is $\Gamma_T^* = \sum \mu_{2,t}\left(\boldsymbol{\zeta}, \hat{\boldsymbol{\theta}}\right)/\sqrt{T}$, and

$$\mu_{2,t}\left(\boldsymbol{\zeta}, \hat{\boldsymbol{\theta}}\right) = \frac{1}{2}\boldsymbol{\eta}' \left[\left(\frac{\partial^2 l_t}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \left(\frac{\partial l_t}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial l_t}{\partial\boldsymbol{\theta}}\right)'\right) + 2\sum_{s<t}\rho^{(t-s)}\left(\frac{\partial l_t}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial l_s}{\partial\boldsymbol{\theta}}\right)'\right]\boldsymbol{\eta}. \tag{5}$$

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 simply takes the form:

$$\text{supTS} = \sup_{\left\{\boldsymbol{\eta},\rho:\|\boldsymbol{\eta}\|=1,\underline{\rho}<\rho<\bar{\rho}\right\}} \frac{1}{2}\left(\max\left(0, \frac{\Gamma_T^*}{\sqrt{\widehat{\epsilon}^{*\prime}\widehat{\epsilon}^*}}\right)\right)^2. \tag{6}$$

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[15]. The asymptotic distributions of the supTS and expTS tests are nonstandard thus the critical values are obtained by bootstrapping the empirical distribution.

To compute the test statistics, we generate the 4×1 vector $\boldsymbol{\eta}$ uniformly over the unit sphere 60 times[16], corresponding to the switching mean and the three GARCH parameters.[17] 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 $\rho's$. For our data, the sup and exp test statistics are calculated to be 0.0039 and 0.6743, respectively. Then we simulate the critical values by bootstrapping using 3,000 iterations. We reject the null of constant

[14] To guarantee identification.

[15] The first part in (5) 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.

[16] That is, we use a uniform prior for $\boldsymbol{\eta}$.

[17] 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.

parameters in favor of regime switching in both the mean and variance equations with p-values of 0.026 for supTS and 0.008 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.

Table 3 presents the parameter estimates for the three MS-GARCH models: MS-GARCH-N, MS-GARCH-t, and MS-GARCH-GED, respectively. MS-GARCH-t and MS-GARCH-GED estimates are very close to each other, but normal innovations lead to different results, where the ARCH parameter estimates in both regimes are insignificant. Thus we focus on MS-GARCH-t and MS-GARCH-GED. In both models, regime 2 corresponds to a significantly positive mean at around 0.1, while the conditional mean in regime 1 is insignificantly different from 0. The transition probabilities, p_{11} and p_{22}, are significant and close to one, implying that both regimes are highly persistent. However, the ergodic probabilities suggest that regime 2 occurs more often. About 70% of the observations are in regime 2, with the remaining 30% in regime 1. The standard deviations from both regimes are close, however, shocks are very persistent in regime 2 as $\alpha_1^{(2)} + \gamma_1^{(2)}$ is close to 1, but not in regime 1. In summary, regime 1 is a relatively bad regime with zero expected returns, and any shocks to the system do not persist for long and only 30% of the observations are in this regime. Majority of the observations are in regime 2, characterized by positive expected returns and persistent shock to the volatility, i.e., the shocks would remain in the system for a long time.

5 Forecast Evaluation

We divide the whole sample into two parts: the first 2388 observations (corresponding to a period of July 1, 2003 to December 31, 2012) are used for in-sample estimation, while the remaining observations are used for out-of-sample forecast evaluation (January 2, 2013 to December 31, 2014).[18] We compute the forecasts using a rolling scheme and evaluate forecasting performance using 504 out-of-sample volatility forecasts (corresponding to the years 2013 and 2014) for the 1-, 5-, 21-, and 63-step horizons.

5.1 Performance Metrics

To evaluate the relative predictive ability of the different volatility models we follow Hansen and Lunde (2005) in computing six different loss functions, where the realized volatility is substituted for the latent conditional variance. These functions are: the Mean Squared Error (MSE) functions written in terms of the standard deviation, MSE_1, and the variance, MSE_2; the Mean Absolute Deviation (MAD) functions, also in terms of the standard deviation, MAD_1, and the variance, MAD_2; the logarithmic loss function of Pagan and Schwert (1990), R^2LOG, which is similar to the R^2 from a regression of the squared first difference of the logged oil price on the conditional variance, and it penalizes volatility forecasts asymmetrically in low and high volatility regimes; and the $QLIKE$,

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

which is equivalent to the loss implied by a Gaussian likelihood. In addition, 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).[19] This battery of test allow us to provide a first ranking of the different volatility models.

To further assess the relative predictive accuracy of the volatility models we implement Diebold and Mariano's (1995) test of Equal Predictive Ability (EPA), White's (2000) Reality Check (RC) test for out-of-sample forecast evaluation, Hansen's (2005) Superior Predictive Ability (SPA) test and Hansen, Lunde and Nason (2011)'s Model Confidence Set (MCS).

The distinction between Hansen's SPA test and Diebold and Mariano's EPA test simply lies in the null hypothesis. The null hypothesis is a simple hypothesis in EPA whilst it is a composite hypothesis in SPA. In other words, EPA is a pairwise comparison, meanwhile SPA is a groupwise comparison. Moreover, the distribution of the SPA test under the null is $N(\hat{\mu}, \Omega)$, where $\hat{\mu}$ is a chosen estimator for μ. Since different choices of $\hat{\mu}$ would result in difference 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_u has the same asymptotic distribution as the RC test.

A disadvantage in doing a pairwise or groupwise forecast evaluation, as in the DM, RC or SPA tests, is that one has to specify a benchmark model for comparison. Hansen, Lunde and Nason (2011) proposed the alternative Model Confidence Set (MCS), which does not require a pre-specified benchmark model. Instead, it determines a set of "best" models M^* with respect to some loss functions given some specified level of confidence. Namely, rather than choosing a single model based on some model selection criteria, the MCS is a data-dependent set containing the best models. Given a collection of competing models, M_0 and a criterion, $L(.)$, MCS is a sequential testing procedure, constructed based on an equivalence test, $\delta_{\mathcal{M}}$ and an elimination rule, $e_{\mathcal{M}}$. First, the equivalence test is applied to the set of models $M = M_0$; if rejected, there is evidence that the models in M are not equally "good" and $e_{\mathcal{M}}$ is used to eliminate an object with poor sample performance from M. The procedure is repeated until $\delta_{\mathcal{M}}$ is accepted and the MCS now includes the set of surviving models and is referred to as the MCS.

5.2 Relative Out-of-Sample Performance

The volatility forecasts obtained from the EGARCH-t and MS-GARCH-t models for the 1-, 5-, 21-, and 63-day horizons are collected in Figure 3.[20] The corresponding realized volatility is also plotted for reference. At 1- and 5-day horizons, the forecasts the two models yield are very similar. They move closely with the realized volatility and are able to capture the huge spikes and dips in the realized volatility. Similarly, at a 21-day horizon, both models are also able to forecast the major upward and downward movements in the

[19] See the Appendix for a precise definition of the loss functions, the Success Ratio and the Directional Accuracy test.

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

realized volatility. Only when we increase the forecast horizon to 63 days, or 3 months, our forecasts contain less information about the aggregated realized volatility during the out-of-sample period, which is as expected.

The estimated loss functions of our out-of-sample forecasts, in addition to the Success Ratio (SR) and the Directional Accuracy (DA) test, are reported in Tables 4a and 4b. Recall that our volatility proxy is the realized volatility measure calculated from the 5-minute futures returns. At the 1- and 5-day forecast horizons, the EGARCH-t and MS-GARCH-t are tied with the MSE_1, MSE_2 and $QLIKE$ ranking the MS-GARCH-t higher and the EGARCH-t being ranked first by the three remaining loss functions. At longer horizons such as 21 and 63 days (one and three months, respectively), evidence in favor of a switching model is stronger: the MS-GARCH-t is ranked first by four loss functions.

The SR averages over 50% for most models and forecast horizons, indicating that most models forecast the direction of the change correctly in more that 50% of the sample. For the 1-, 5- and 21-day forecast horizons, the SR exceeds 60% for all models except EGARCH-N at 21-day horizon (averages 70%, 71% and 68% respectively). In addition, at a longer 63-day horizon the SR averages 60% across all models, suggesting the direction of the change is more difficult to predict for this longer 63-day horizon. Notice that at this horizon the SR is less than 50% for the three GARCH models and the EGARCH-N, yet all three MS-GARCH models have SR higher than 70%, which indicates that MS-GARCH models can do a much better job at predicting the direction of the change in volatility than the time-invariant models in the long run. The results of the DA test are consistent with this finding. Recall that a significant DA statistic indicates that the model forecasts have predictive content for the underlying volatility. In particular, the DA test is significant at the 1% level for majority of the models at 1- and 5-day forecast horizons. In contrast, for the longer 21- and 63-day forecast horizons the number of models that exhibit a significant DA decreases to six and four, respectively, and all MS-GARCH models are included.

Tables 5a and 5b reports selected DM test statistics with EGARCH-t and MS-GARCH-t as benchmark models.[21] These test results are in line with the rankings reported in Tables 4a and 4b. Consider first the 1-day-ahead forecast where the EGARCH-t was ranked higher by three loss functions R^2LOG, MAD_1 and MAD_2. As Table 5a shows, we reject the null of Equal Predictive Ability at 5% level for at least seven of the eleven competing models under the three loss functions, favoring the benchmark EGARCH-t model. Furthermore, all models but the EGARCH-t are shown to have equal or lower predictive accuracy than the MS-GARCH-t (see Table 5b). Regarding the 5-day horizon (1 week), there is some statistical difference in the forecast accuracy comparison between the benchmark model and the non-switching models. The EGARCH-t has higher predictive accuracy than 8 of the 11 competing models for $MAD_\mathbf{1}$ and $MAD_\mathbf{2}$. Yet, the MS-GARCH-t is found to have equal accuracy as the benchmark EGARCH-t. When the MS-GARCH-t is considered to be the benchmark, it has higher predictive accuracy than all the GARCH and GJR models for $QLIKE$, R^2LOG, $MAD_\mathbf{1}$ and MAD_2. In

[21] The complete list of all DM test statistics can be requested from the authors.

contrast, as the forecast horizon increases to 21 and 63 days (1 and 3 months), statistical evidence that the forecast accuracy differences are negative, in favor of switching models –especially the MS-GARCH-t– is prevalent. Indeed, the EGARCH-t has significantly better accuracy than majority of the non-switching models for $MAD_{\mathbf{1}}$ and MAD_2, but the MS-GARCH-t has higher predictive accuracy than the GARCH and GJR classes under all six loss functions. Moreover, it is worth noting that MS-GARCH-t is favored over the EGARCH-t for both MSE_2 and $QLIKE$ at 63-day horizon. In fact, at this longest horizon, the MS-GARCH-t has significantly higher predictive accuracy than all the 11 competing models for $QLIKE$.

The p-values for the RC and SPA tests are reported in Tables 6a and 6b, where each model is compared against all the others. Recall that the null hypothesis here is that no other models outperform the benchmark. The model in each row is the benchmark model under consideration. The RC, SPA_c, and SPA_l correspond to the Reality Check p-value, Hansen's (2005) consistent, and lower p-values, respectively.[22] For the 1- and 5-day horizons, all three EGARCH models fail to reject the null regardless of the loss function (except for EGARCH-GED with $QLIKE$ and MAD_1) at 5% level, implying no other models can outperform the EGARCH models. Meanwhile, the MS-GARCH-t also outperforms other models when the MSE_1, MSE_2 or $QLIKE$ is used, but not for the other loss functions (see Table 6a). Yet, consistent with the out-of-sample evaluation and the Diebold and Mariano's EPA test results, as the forecast horizon increases we fail to reject the null, not only for EGARCH models, but also for the MS-GARCH-t, with the exception of MAD_1 and MAD_2.

For the MCS, we compute both the $T_{\max,\mathcal{M}}$ and $T_{R,\mathcal{M}}$ tests with confidence level at 0.25 over 3000 bootstrap iterations. 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. We follow the authors' suggestion to focus on ($T_{R,\mathcal{M}}$, $e_{R,\mathcal{M}}$) and report the results in Table 7.[23] Regardless of the loss function, at *the* one-day forecast horizon the EGARCH-t is contained in the $\widehat{\mathcal{M}}^*_{.75}$. When the MAD_1 and MAD_2 are used, $\widehat{\mathcal{M}}^*_{.75}$ consists of two EGARCH models, EGARCH-N and EGARCH-t. Moreover, the EGARCH-t is the only model that ends up in $\widehat{\mathcal{M}}^*_{.75}$ according to the R^2LOG loss. However, under MSE_2 ($QLIKE$) loss, the MCS is larger as the $\widehat{\mathcal{M}}^*_{.75}$ contains all but one (all) specifications. At a five-day horizon, the test results are very similar except that for the R^2LOG loss all three EGARCH models are contained in the $\widehat{\mathcal{M}}^*_{.75}$. As the forecast horizon increases, the MCS becomes smaller under $QLIKE$, MAD_1, and MAD_2. At the 21-day horizon, the only specification that ends up in the $\widehat{\mathcal{M}}^*_{.75}$ is the MS-GARCH-t according to the $QLIKE$ and the EGARCH-t according to the MAD_1 and MAD_2. Similarly, at a 63-day horizon, under MSE_1 and $QLIKE$, the MS-GARCH-t is the only model in $\widehat{\mathcal{M}}^*_{.75}$, whereas under MAD_1 and MAD_2 losses the $\widehat{\mathcal{M}}^*_{75\%}$ only contains the EGARCH-t. At this longer horizon the GARCH models are always eliminated. In brief, the two MAD criteria favor the EGARCH-t across all

[22] The p-values are calculated using the stationary bootstrap from Politis and Romano (1994). The number of bootstrap re-samples B is 3000 and the block length q is 2.

[23] Results for the $T_{\max,\mathcal{M}}$ are available from the authors upon request.

forecast horizons, however, MS-GARCH-t is the "best" model according to $QLIKE$ at 21- and 63-day horizons. These results reinforce our previous findings.

To summarize, we find substantial evidence that the EGARCH-t and MS-GARCH-t forecasts of crude oil volatility are superior to other volatility forecast specifications. In particular, whereas the the time invariant specification is favored at shorter one- and five-day horizons, the regime switching model does a better job at longer one and three-month horizons.

5.3 Preferences for Loss Functions

Our empirical findings suggest that EGARCH-t and MS-GARCH-t are the two closest competitors in forecasting volatility. EGARCH-t is mostly favored for forecasting at short horizons whilst MS-GARCH-t generally does a better job at longer horizons. When we investigate further the combined results from the EPA, RC, SPA and MCS tests, we notice that different loss functions have persistent preferences over certain models. Specifically, the two MAD loss functions seem to favor the EGARCH-t across all horizons. However, the MSE criteria and especially the $QLIKE$ often rank the MS-GARCH-t higher. Now the question of interest is, are certain loss functions better than others in volatility forecast?

It is trivial to see that the MSE_2 and $QLIKE$ loss functions generate optimal forecast equal to the conditional variance σ_t^2. Patton (2011) shows that only these two among the six loss functions are robust to noise in the volatility proxy and all the rest suffer from bias distortion. Furthermore, Patton demonstrates that employing a measure of realized volatility –as we do here– to proxy for σ_t^2, alleviates the bias distortion relative to using other proxies such as the daily squared returns. Yet, as many authors have noted, the MSE_2 is sensitive to extreme observations and **to** the level of volatility of returns. Such episodes of extreme volatility are present in spot oil returns, which provides a motivation for using the $QLIKE$ in forecasting their volatility**,** especially when evaluating periods of turmoil. Not surprisingly, Patton (2011) argues that the moment conditions required under MSE_2 are also substantially stronger than those under $QLIKE$, which suggests employing the latter loss. Brownlees et al. (2011) also favor $QLIKE$ for two reasons: first, the $QLIKE$ only depends on the multiplicative forecast error, thus it is easier to identify when a model fails to adequately capture predictable movements in volatility; second, the MSE_2 has a bias that is proportional to the square of the true variance, suggesting that obtaining a large MSE_2 could be a consequence of high volatility without necessarily corresponding to deterioration of forecasting ability. In addition, Patton and Sheppard (2009) find that the power of the DM tests using $QLIKE$ loss is higher than using MSE_2 loss. In brief, there is ample motivation in the literature for using the $QLIKE$ loss rather than any of the other loss functions considered in this paper in forecasting volatility. In turn, using the $QLIKE$ loss favors the MS-GARCH-t model.

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 whole 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. As we have already mentioned, Table 4 indicates that for the evaluation period of 2013-2014, the MS-GARCH-t exhibits lower $MSPE$ –as measured by three loss functions (MSE_1, MSE_2, $QLIKE$)– for the 1- and 5-day forecast horizons, whereas the EGARCH-t results in smaller $MSPE$ when the remaining loss functions are used. In addition, under MSE_1, MSE_2, $QLIKE$ and R^2LOG losses, the switching model yields smaller $MSPE$ for the longer 21- and 63-day horizons.

To investigate the stability of the forecast accuracy, we compute the $MSPE$ from the preferred $QLIKE$ loss over 441 rolling sub-samples in the evaluation period, where the first sub-sample consists of the first 63 forecasts (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 4 plots the ratio of the $MSPE$ for the 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 $MSPE$ ratio is reported at October 1, 2014.

Figure 4 illustrates that for all forecast horizons, the MS-GARCH-t almost always has higher predictive accuracy than the GARCH-t. This is evidenced by the $MSPE$ ratio exceeding 1 over almost all of the evaluation period. This result confirms our finding that the MS-GARCH-t beats the GARCH(1,1) in forecasting spot crude oil price volatility. Regarding its closest competitor, the EGARCH-t, Figure 4 reveals that the MS-GARCH-t has lower predictive accuracy during the beginning of the evaluation period and from mid-2013 through mid-2014. In contrast, the MS-GARCH-t does a better job at at predicting the increase in volatility during the second half of 2014, when the WTI price fell considerably (44% between June and December of 2013) and returns became more volatile. 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 one- and five-day horizons over the two-year evaluation period (Table 4), they become clear when we plot the ratio of the rolling window $MSPE$s of 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. Our results suggest five key insights for practitioners interested in crude oil price volatility. First, given the extremely high kurtosis present in the data, models where the innovations are assumed to follow a Student's t distribution are favored over those where a normal distribution is presumed. Second, for the one day horizon the EGARCH-t is often ranked higher in terms of loss functions and tends to yield more accurate forecasts than other EGARCH and all GARCH models. Yet, predictive accuracy appears to be similar to that of the MS-GARCH-t. Third, as the length of the forecast horizon increases, the MS-GARCH-t model outperforms non-switching GARCH models and other regime switching specifications. Fourth, the $QLIKE$, being the most popular loss function for its good properties, favors the sole MS-GARCH-t model at longer horizons, which reinforced our findings. Lastly, when we analyzed the stability of the forecasting accuracy over different evaluation periods, we found MS-GARCH-t model has higher predictive accuracy for all horizons towards the end of the evaluation period when oil returns became considerably more volatility. All in all, our analysis suggested that the MS-GARCH-t model yields more accurate long-term forecasts of spot WTI return volatility and that it does a better job at forecasting during periods of turmoil.

Three caveats are needed here. First, as it is well known in the literature, EGARCH models deliver an unbiased forecast for the logarithm of the conditional variance, but the forecast of the conditional variance itself would be biased following Jensen's Inequality (e.g., Andersen et al. 2006, among others). For practitioners who prefer unbiased forecasts, caution must be taken when using EGARCH models. Second, 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 of the last semester of 2014. Lastly, long horizon volatility forecasts such as the one- and three-month horizons, may be computed in three different ways. For instance, if a researcher was 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 using 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. In this paper we use the "iterated" forecast to evaluate the relative out-of-sample performance of different models in the context of multi-period volatility forecast. 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.

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] Alquist, R., and L. Kilian (2010), "What do we learn from the price of crude oil futures?," *Journal of Applied Econometrics*, 25:4, 539-573.

[4] Alquist, R., L. Kilian and R. J. Vigfusson (2013), "Forecasting the Price of Oil," in: G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, 2, Amsterdam: North-Holland: 427-507.

[5] 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.

[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] 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.

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

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

[10] 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.

[11] Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, Vol 31, No 3, 307-327.

[12] Brownlees, C., R. Engle and B. Kelly (2011), "A practical guide to volatility forecasting through calm and storm," *Journal of Risk*, Vol 14, No 2, 3-22,.

[13] Caporale, G, N. Pittis and N. Spagnolo (2003), "IGARCH models and structural breaks," *Applied Economics Letters*, Vol 10, No 12, 765-768.

[14] Carrasco, M, L. Hu and W. Ploberger (2014), "Optimal Test for Markov Switching Parameters," *Econometrica*, Vol 82, No 2, 765-784.

[15] 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.

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

[17] Edelstein, P., and L. Kilian (2009), "How Sensitive are Consumer Expenditures to Retail Energy Prices?" *Journal of Monetary Economics*, 56, 766-779.

[18] Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of U.K. Inflation,"*Econometrica,* 50:4, 987-1008.

[19] Ferderer, J. P. (1996), "Oil price volatility and the macroeconomy," *Journal of Macroeconomics,* 18:1, 1-26.Fong, W., and K. See (2002), "A Markov switching model of the conditional volatility of crude oil futures prices," *Energy Economics*, 24, 71-95.

[20] 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.

[21] Frühwirth-Schnatter, S., 2006. *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, Springer, New York.

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

[23] 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.

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

[25] 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.

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

[27] 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.

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

[29] Haas, M., S. Mittnik and M. Paolella (2004), "A New Approach to Markov Switching GARCH Models," *Journal of Financial Econometrics,* 2, 493-530.

[30] Hamilton, J.D. (2009), "Causes and Consequences of the Oil Shock of 2007-08," *Brookings Papers on Economic Activity*, 1, Spring, 215-261.

[31] Hansen, B. (1992), "The Likelihood Ratio Test Under Non-Standard Conditions: Testing the Markov Switching Model of GNP,"*Journal of Applied Econometrics,* 7, S61-S82.

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

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

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

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

[40] Lopez, J. A. (2001),"Evaluating the Predictive Accuracy of Volatility Models,"*Journal of Forecasting*, 20:2, 87-109.

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

[42] 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.

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

[44] Morana, C. (2001), "A semi-parametric approach to short-term oil price forecasting," *Energy Economics*, Vol 23, No 3, 325-338.

[45] Nelson, D. B. (1990), "Stationarity and Persistence in the GARCH(1,1) Model,"*Econometric Theory*, 6:3, 318-334.

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

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

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

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

[50] Patton, A.J. and K. Sheppard (2009), "Evaluating volatility and correlation forecasts," *The Handbook of Financial Time Series*, Edited by T.G. Andersen, R.A. Davis, J.-P. Kreiss and T. Mikosch, Springer Verlag.

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

[52] Plante, M., & Traum, N. (2012). Time-varying oil price volatility and macroeconomic aggregates. Center for Applied Economics and Policy Research Working Paper, (2012-002).

[53] Politis, D. N. and J.P. Romano (1994), "The Stationary Bootstrap," *Journal of The American Statistical Association*, 89:428, 1303-1313.

[54] Vo, M. (2009), "Regime-switching stochastic volatility: Evidence from the crude oil market," *Energy Economics*, 31, 779-788.

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

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

[57] Wilson, B., R. Aggarwal and C. Inclan (1996), "Detecting volatility changes across the oil sector," *Journal of Futures Markets*, 47:1, 313-320.

[58] Xu, B. and J. Oueniche (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
0.0162	2.34	-12.83	16.41	5.45	-0.017	7.91

$RV^{1/2}$

Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
0.0197	0.0099	0.0040	0.187	0.00010	3.55	36.82

$\ln(RV^{1/2})$

Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
-4.02	0.41	-5.53	-1.67	0.17	0.39	4.10

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 July 1, 2003 to April 2, 2015 for 2955 observations

Table 2: MLE Estimates of Standard GARCH Models

	GARCH			EGARCH			GJR		
	N	t	GED	N	t	GED	N	t	GED
μ	0.0944**	0.1068**	0.1122**	0.0349	0.0700	0.0747	0.0603	0.0858*	0.0898*
	(0.0402)	(0.0397)	(0.0393)	(0.0424)	(0.0398)	(0.0398)	(0.0426)	(0.0402)	(0.0401)
α_0	0.2145**	0.1890**	0.2048**	0.0203**	0.0142*	0.015*	0.2026**	0.1670**	0.1847**
	(0.0336)	(0.0408)	(0.0451)	(0.0045)	(0.0059)	(0.006)	(0.0321)	(0.0391)	(0.0429)
α_1	0.0756**	0.0754**	0.0753**	0.0864**	0.0965**	0.0912**	0.0381**	0.0348**	0.0367**
	(0.0079)	(0.0126)	(0.0118)	(0.0089)	(0.0165)	(0.0146)	(0.0087)	(0.0126)	(0.0124)
γ_1	0.8809**	0.8854**	0.8826**	0.9887**	0.9461**	0.9895**	0.8857**	0.8931**	0.8896**
	(0.0124)	(0.0162)	(0.0168)	(0.0026)	(0.0036)	(0.0037)	(0.0127)	(0.0152)	(0.0164)
ξ	-	-	-	-0.0483**	-0.0539**	-0.0501**	0.0708**	0.0746**	0.0707**
				(0.0065)	(0.0111)	(0.0099)	(0.0168)	(0.0219)	(0.0221)
ν	-	8.6309**	1.4799**	-	8.4794**	1.4774**	-	8.8579**	1.4924**
		(1.1318)	(0.0441)		(1.0315)	(0.0417)		(1.1456)	(0.0436)
$Log(L)$	-5253.15	-5210.74	-5220.12	-5244.00	-5195.39	-5209.08	-5242.84	-5200.47	-5211.42

Note: * and ** represent significance at 5% and 1% level respectively. A one-sided test is conducted on ξ. Each model is estimated with Normal, Student's t, and GED distributions. The in-sample data consist of WTI returns from 7/1/03 to 12/30/12. The conditional mean is $r_t = \mu + \varepsilon_t$. The conditional variances are $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 h_{t-1}$, $\log(h_t) = \alpha_0 + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| - E \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| \right) + \xi \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \log(h_{t-1})$, and $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \xi \varepsilon_{t-1}^2 I_{\{\varepsilon_{t-1} < 0\}} + \gamma_1 h_{t-1}$ for GARCH, EGARCH, and GJR-GARCH respectively. Asymptotic standard errors are in parenthesis.

Table 3: Maximum Likelihood Estimates of MS-GARCH Models

	MS-GARCH-N	MS-GARCH-t	MS-GARCH-GED
$\mu^{(1)}$	-0.6921**	0.0932	0.1112
	(0.2004)	(0.0801)	(0.0739)
$\mu^{(2)}$	0.1727**	0.1141**	0.1010*
	(0.0452)	(0.0485)	(0.0493)
$\sigma^{(1)}$	9.3085**	2.045	2.253*
	(0.5011)	(1.2379)	(1.0062)
$\sigma^{(2)}$	1.6697**	2.3174**	1.9709**
	(0.3459)	(0.1902)	(0.1640)
$\alpha_1^{(1)}$	0.0142	0.0980	0.1628**
	(0.0145)	(0.0649)	(0.073)
$\alpha_1^{(2)}$	0.005	0.0625**	0.0484**
	(0.019)	(0.0130)	(0.0119)
$\gamma_1^{(1)}$	0.9750**	0.5697*	0.5004**
	(0.029)	(0.2755)	(0.0187)
$\gamma_1^{(2)}$	0.8234**	0.9221**	0.9369**
	(0.0348)	(0.0164)	(0.0146)
p_{11}	0.9003**	0.9936**	0.9856**
	(0.0027)	(0.0051)	(0.0081)
p_{22}	0.9787**	0.9973**	0.9941**
	(0.0068)	(0.0020)	(0.0032)
$\nu^{(1)}$	-	3.7976**	1.0240**
		(0.8557)	(0.0885)
$\nu^{(2)}$	-	17.8335*	1.9512**
		(7.3007)	(0.1431)
$Log(L)$	-5222.04	-5194.95	-5197.09
$N.of\ Par.$	10	12	12
π_1	0.1760	0.2967	0.2906
π_2	0.8240	0.7033	0.7094
$\alpha_1^{(1)} + \gamma_1^{(1)}$	0.9892	0.6677	0.6632
$\alpha_1^{(2)} + \gamma_1^{(2)}$	0.8284	0.9846	0.9853

Note: * and ** represent significance at 5% and 1% 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 7/1/03 to 12/30/12. The superscripts indicate the regime. The standard deviation conditional on the regime is reported: $\sigma^{(i)} = \left(\alpha_0^{(i)}/(1-\alpha_1^{(i)}-\gamma_1^{(i)})\right)^{1/2}$. 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 4a: Out-of-sample evaluation of the one- and five-step-ahead volatility forecasts

1-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	R^2LOG	Rank	MAD_1	Rank	MAD_2	Rank	SR	DA
GARCH-N	0.2106	5	2.5312	4	1.4239	5	0.5348	7	1.1161	7	0.3998	7	0.73	6.7663**
GARCH-t	0.1914	2	2.3649	2	1.4083	2	0.4890	5	1.0424	5	0.3751	5	0.73	6.7298**
GARCH-GED	0.2019	3	2.4422	3	1.4175	3	0.5160	6	1.0836	6	0.3895	6	0.73	6.6049**
EGARCH-N	0.2205	6	3.7800	10	1.4900	11	0.3785	2	0.8226	2	0.2855	2	0.64	0.9844
EGARCH-t	0.2043	4	3.5487	8	1.4773	10	**0.3537**	**1**	**0.7812**	**1**	**0.2717**	**1**	0.66	0.2615
EGARCH-GED	0.2257	7	3.8399	11	1.5042	12	0.3858	3	0.8267	3	0.2885	3	0.61	-1.4866
GJR-N	0.2756	11	3.9916	12	1.4404	8	0.5991	11	1.3186	12	0.4407	12	0.73	5.8730**
GJR-t	0.2478	8	3.5014	7	1.4235	4	0.5494	8	1.2207	8	0.4122	8	0.73	5.5640**
GJR-GED	0.2606	9	3.6928	9	1.4326	7	0.5758	10	1.2693	11	0.4274	11	0.73	5.7040**
MS-GARCH-N	0.2671	10	3.4140	5	1.4289	6	0.5746	9	1.2256	10	0.4131	9	0.7	5.2922**
MS-GARCH-t	**0.1899**	**1**	**2.3194**	**1**	**1.4012**	**1**	0.4708	4	0.9980	4	0.3566	4	0.68	5.2429**
MS-GARCH-GED	0.2829	12	3.5010	6	1.4501	9	0.6441	12	1.2234	9	0.4234	10	0.69	5.2802**

5-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	R^2LOG	Rank	MAD_1	Rank	MAD_2	Rank	SR	DA
GARCH-N	1.0821	5	55.7240	4	3.0784	7	0.5241	10	5.8666	8	0.9330	9	0.72	5.6439**
GARCH-t	0.9652	2	50.3921	2	3.0610	3	0.4739	5	5.4392	5	0.8715	5	0.72	5.7700**
GARCH-GED	1.0289	4	52.9307	3	3.0713	5	0.5035	8	5.6788	6	0.9076	8	0.72	5.9367**
EGARCH-N	1.1074	6	83.8685	11	3.1599	11	0.3628	2	4.0577	2	0.6245	2	0.63	0.1611
EGARCH-t	0.9816	3	71.7613	7	3.1348	10	**0.3389**	**1**	**3.8626**	**1**	**0.6017**	**1**	0.67	-0.8195
EGARCH-GED	1.1509	8	85.6236	12	3.1880	12	0.3833	3	4.1365	3	0.6458	3	0.61	-3.3194
GJR-N	1.2642	11	81.1383	10	3.0805	8	0.5469	11	6.4601	12	0.9766	12	0.76	7.6863**
GJR-t	1.1128	7	69.3462	6	3.0618	4	0.4934	6	5.9303	9	0.9073	7	0.76	7.4437**
GJR-GED	1.1803	10	73.8522	8	3.0715	6	0.5205	9	6.1823	10	0.9426	10	0.76	7.3697**
MS-GARCH-N	1.1690	9	67.3210	5	3.0591	2	0.4969	7	5.7440	7	0.8767	6	0.73	6.8979**
MS-GARCH-t	**0.8777**	**1**	**46.1996**	**1**	**3.0384**	**1**	0.4196	4	4.9220	4	0.7826	4	0.67	4.2898**
MS-GARCH-GED	1.4373	12	75.7412	9	3.1015	9	0.6319	12	6.3387	11	0.9750	11	0.7	5.5531**

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 4b: Out-of-sample evaluation of the 21- and 63-step-ahead volatility forecasts

21-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	R^2LOG	Rank	MAD_1	Rank	MAD_2	Rank	SR	DA
GARCH-N	6.5310	10	1282.5777	8	4.6439	9	0.6981	12	31.8949	12	2.3934	12	0.6	0.2237
GARCH-t	5.7824	4	1140.2466	5	4.6229	6	0.6327	8	29.5494	8	2.2426	9	0.65	1.4476
GARCH-GED	6.2000	6	1214.6674	6	4.6356	8	0.6715	11	30.8988	10	2.3332	11	0.63	0.9705
EGARCH-N	6.9819	11	2177.7107	11	4.8243	11	0.4940	3	21.7580	3	1.5319	2	0.54	-8.0191
EGARCH-t	6.3365	8	1997.8148	10	4.7864	10	0.4533	2	**20.1877**	**1**	**1.4342**	**1**	0.68	-4.6892
EGARCH-GED	7.0592	12	2179.0182	12	4.8473	12	0.5100	5	21.5664	2	1.5457	3	0.62	-7.8636
GJR-N	6.3119	7	1276.5772	7	4.6272	7	0.6652	10	31.0036	11	2.3073	10	0.75	6.4341**
GJR-t	5.4339	3	1058.1458	2	4.6041	3	0.5964	6	28.0513	6	2.1302	6	0.77	7.0618**
GJR-GED	5.7918	5	1138.4597	4	4.6146	4	0.6270	7	29.3222	7	2.2109	8	0.76	6.6794**
MS-GARCH-N	4.8156	2	1079.0748	3	4.5745	2	0.5078	4	25.0264	5	1.8810	5	0.73	6.2937**
MS-GARCH-t	**3.9713**	**1**	**850.7942**	**1**	**4.5488**	**1**	**0.4290**	**1**	22.1929	4	1.6899	4	0.7	5.7630**
MS-GARCH-GED	6.4389	9	1412.2395	9	4.6184	5	0.6489	9	29.9198	9	2.1904	7	0.71	5.5031**

63-step-ahead volatility forecasts

Model	MSE_1	Rank	MSE_2	Rank	$QLIKE$	Rank	R^2LOG	Rank	MAD_1	Rank	MAD_2	Rank	SR	DA
GARCH-N	35.7379	12	26523.4547	9	5.9959	9	1.0258	12	146.4412	12	5.6748	12	0.44	-6.5344
GARCH-t	32.6550	7	24671.2528	7	5.9764	7	0.9459	10	139.1392	10	5.4326	10	0.46	-6.1886
GARCH-GED	34.4449	9	25730.9848	8	5.9891	8	0.9939	11	143.5644	11	5.5856	11	0.45	-6.4869
EGARCH-N	35.5025	11	34950.0898	12	6.3686	11	0.7458	6	94.9033	4	3.5231	3	0.49	-13.1858
EGARCH-t	33.1379	8	33130.0637	10	6.3244	10	0.6996	4	**87.8113**	**1**	**3.2958**	**1**	0.67	-5.6943
EGARCH-GED	34.8127	10	34380.0924	11	6.3803	12	0.7401	5	90.2934	2	3.3933	2	0.61	-11.7889
GJR-N	30.7294	6	20893.3086	6	5.9514	6	0.9227	9	131.0116	9	5.2017	9	0.55	-0.6735
GJR-t	26.9771	4	18340.4726	2	5.9263	4	0.8336	7	121.9424	7	4.9126	7	0.69	3.2183**
GJR-GED	28.4416	5	19317.8082	3	5.9370	5	0.8698	8	125.5674	8	5.0319	8	0.63	1.4474
MS-GARCH-N	21.2259	2	20094.1955	5	5.8686	2	0.5779	2	101.8794	5	3.9724	5	0.76	7.9407**
MS-GARCH-t	**18.8717**	**1**	**16793.6501**	**1**	**5.8408**	**1**	**0.5278**	**1**	94.4618	3	3.6995	4	0.7	6.0831**
MS-GARCH-GED	23.6958	3	19438.8779	4	5.8857	3	0.6962	3	114.0805	6	4.4590	6	0.75	8.3764**

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 5a: Diebold and Mariano test - EGARCH-t Benchmark

Panel A: One day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-0.10	0.80	0.70	-2.18*	-3.62**	-3.23**
GARCH-t	0.21	0.93	0.91	-1.68	-3.01**	-2.58**
GARCH-GED	0.04	0.87	0.79	-1.98*	-3.37**	-2.94**
EGARCH-N	-1.16	-1.25	-0.63	-1.07	-1.24	-1.36
EGARCH-GED	-1.34	-1.32	-1.17	-1.40	-1.54	-1.43
GJR-N	-1.00	-0.24	0.47	-2.72**	-4.60**	-4.06**
GJR-t	-0.62	0.03	0.69	-2.21*	-3.89**	-3.51**
GJR-GED	-0.80	-0.08	0.57	-2.50*	-4.29**	-3.83**
MS-GARCH-N	-0.84	0.10	0.61	-2.14*	-3.39**	-3.34**
MS-GARCH-t	0.23	0.94	1.01	-1.43	-2.34*	-1.99*
MS-GARCH-GED	-1.07	0.04	0.34	-2.56*	-3.36**	-3.20**

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-0.31	0.58	0.68	-1.88	-3.30**	-3.07**
GARCH-t	0.05	0.78	0.90	-1.41	-2.76**	-2.48*
GARCH-GED	-0.14	0.69	0.77	-1.69	-3.08**	-2.81**
EGARCH-N	-1.06	-1.09	-0.81	-0.85	-0.72	-1.00
EGARCH-GED	-1.11	-1.04	-1.10	-1.26	-1.70	-1.57
GJR-N	-0.80	-0.25	0.64	-1.99*	-3.56**	-3.43**
GJR-t	-0.37	0.07	0.86	-1.51	-2.96**	-2.83**
GJR-GED	-0.57	-0.06	0.74	-1.76	-3.28**	-3.14**
MS-GARCH-N	-0.49	0.15	0.89	-1.41	-2.48*	-2.46*
MS-GARCH-t	0.31	0.92	1.17	-0.85	-1.86	-1.65
MRS-GARCH-GED	-1.16	-0.15	0.38	-2.25*	-3.09**	-3.01**

Panel C: Twenty-one day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-0.08	0.93	1.00	-1.61	-3.22**	-2.70**
GARCH-t	0.25	1.13	1.16	-1.22	-2.79**	-2.22*
GARCH-GED	0.06	1.02	1.07	-1.46	-3.05**	-2.50*
EGARCH-N	-1.66	-1.59	-1.25	-1.45	-1.65	-2.27*
EGARCH-GED	-1.69	-1.52	-1.58	-1.91	-2.33*	-2.32*
GJR-N	0.01	0.86	1.10	-1.37	-2.74**	-2.23*
GJR-t	0.37	1.11	1.26	-0.94	-2.22*	-1.63
GJR-GED	0.23	1.02	1.19	-1.14	-2.46*	-1.89
MS-GARCH-N	0.68	1.26	1.52	-0.38	-1.58	-1.19
MS-GARCH-t	1.06	1.45	1.72	0.18	-0.94	-0.49
MS-GARCH-GED	-0.04	0.79	1.19	-1.22	-2.53*	-2.26*

Panel D: Sixty-three day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-0.27	0.77	1.54	-1.48	-3.74**	-3.73**
GARCH-t	0.05	1.03	1.66	-1.16	-3.55**	-3.50**
GARCH-GED	-0.14	0.88	1.59	-1.36	-3.70**	-3.67**
EGARCH-N	-2.33*	-2.36*	-1.40	-1.77	-2.13*	-2.83**
EGARCH-GED	-1.53	-1.50	-1.46	-1.50	-1.41	-1.73
GJR-N	0.23	1.18	1.70	-0.99	-2.68**	-2.22*
GJR-t	0.61	1.45	1.84	-0.62	-2.38*	-1.84
GJR-GED	0.46	1.35	1.78	-0.77	-2.52*	-2.00*
MS-GARCH-N	1.60	2.06+	2.36+	0.69	-1.46	-1.36
MS-GARCH-t	1.69	1.99+	2.42+	0.95	-0.79	-0.53
MS-GARCH-GED	1.06	1.73	2.10+	0.02	-2.04*	-1.91

Note: * and ** represent the DM test statistic for which the null hypothesis of equal predictive accuracy can be rejected at 5% and 1%, respectively and the DM statistic is negative. + and ++ represent the 5% and 1% significance level when the DM test statistic is positive.

Table 5b: Diebold and Mariano test - MS-GARCH-t Benchmark

Panel A: One day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-1.69	-0.93	-2.81**	-2.67**	-3.45**	-2.60**
GARCH-t	-0.13	-0.23	-0.88	-0.75	-1.51	-1.03
GARCH-GED	-0.99	-0.60	-2.07*	-1.91	-2.70**	-1.96*
EGARCH-N	-0.42	-1.01	-1.05	1.06	1.96	1.53
EGARCH-t	-0.23	-0.94	-1.01	1.43	2.34+	1.99+
EGARCH-GED	-0.48	-1.04	-1.16	0.94	1.77	1.42
GJR-N	-2.28*	-1.30	-3.12**	-3.26**	-3.82**	-2.72**
GJR-t	-1.73	-1.15	-1.75	-1.95	-2.61**	-2.10*
GJR-GED	-2.04*	-1.22	-2.54*	-2.70**	-3.32**	-2.47*
MS-GARCH-N	-2.34*	-2.17*	-1.87	-2.06*	-2.66**	-2.73**
MS-GARCH-GED	-3.07**	-2.60**	-3.27**	-3.36**	-3.29**	-3.15**

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-3.18**	-1.45	-4.89**	-4.51**	-5.09**	-4.03**
GARCH-t	-1.40	-0.74	-2.92**	-2.45*	-3.21**	-2.35*
GARCH-GED	-2.44*	-1.14	-4.23**	-3.78**	-4.47**	-3.40**
EGARCH-N	-0.55	-1.00	-1.17	0.54	1.59	1.23
EGARCH-t	-0.31	-0.92	-1.17	0.85	1.86	1.65
EGARCH-GED	-0.61	-1.00	-1.26	0.32	1.32	1.09
GJR-N	-2.58**	-1.33	-3.61**	-3.65**	-4.32**	-3.08**
GJR-t	-1.73	-1.11	-2.02*	-2.07*	-2.93**	-2.23*
GJR-GED	-2.18*	-1.22	-2.92**	-2.94**	-3.75**	-2.71**
MS-GARCH-N	-1.92	-1.91	-1.51	-1.67	-2.12*	-2.20*
MS-GARCH-GED	-3.34**	-2.92**	-3.79**	-3.73**	-3.69**	-3.52**

Panel C: Twenty-one day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-8.96**	-6.39**	-9.55**	-8.98**	-9.98**	-9.73**
GARCH-t	-7.60**	-4.45**	-9.29**	-8.49**	-9.66**	-8.93**
GARCH-GED	-8.72**	-5.63**	-9.85**	-9.10**	-10.37**	-9.85**
EGARCH-N	-1.21	-1.51	-1.73	-0.44	0.58	0.10
EGARCH-t	-1.06	-1.45	-1.72	-0.18	0.94	0.49
EGARCH-GED	-1.22	-1.51	-1.79	-0.53	0.51	0.15
GJR-N	-5.93**	-2.47*	-6.34**	-7.10**	-6.36**	-4.95**
GJR-t	-3.90**	-1.44	-5.12**	-5.60**	-5.30**	-3.72**
GJR-GED	-4.95**	-1.91	-5.90**	-6.58**	-5.98**	-4.36**
MS-GARCH-N	-1.84	-1.58	-2.42*	-2.36*	-2.67**	-2.35*
MS-GARCH-GED	-3.77**	-3.16**	-4.49**	-4.38**	-4.83**	-4.46**

Panel D: Sixty-three day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-9.32**	-7.23**	-8.50**	-9.90**	-10.79**	-10.52**
GARCH-t	-10.35**	-7.60**	-9.85**	-10.61**	-12.35**	-12.02**
GARCH-GED	-10.04**	-7.68**	-9.35**	-10.43**	-11.84**	-11.53**
EGARCH-N	-1.86	-2.10*	-2.42*	-1.15	0.35	-0.04
EGARCH-t	-1.69	-1.99*	-2.42*	-0.95	0.79	0.53
EGARCH-GED	-1.76	-2.01*	-2.41*	-1.10	0.59	0.32
GJR-N	-4.91**	-1.68	-4.64**	-7.39**	-5.84**	-4.56**
GJR-t	-4.14**	-0.71	-4.30**	-7.32**	-5.73**	-4.00**
GJR-GED	-4.54**	-1.12	-4.55**	-7.48**	-5.89**	-4.31**
MS-GARCH-N	-1.57	-1.47	-2.40*	-2.07*	-2.22*	-1.81
MS-GARCH-GED	-2.73**	-1.42	-2.49*	-4.53**	-4.17**	-3.70**

Note: * and ** represent the DM test statistic for which the null hypothesis of equal predictive accuracy can be rejected at 5% and 1%, respectively and the DM statistic is negative. + and ++ represent the 5% and 1% significance level when the DM test statistic is positive.

Table 6a: Reality Check and Superior Predictive Ability Tests

Horizon: One day								Horizon: Five days							
		Loss Function								Loss Function					
Benchmark		MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2	Benchmark		MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	*SPAl*	0.624	0.870	0.433	0	0	0	GARCH-N	*SPAl*	0.438	0.824	0.406	0	0	0
	SPAc	0.339	0.168	0.053	0	0	0		*SPAc*	0.306	0.126	0.048	0	0	0
	RC	0.373	0.168	0.059	0	0	0		RC	0.322	0.126	0.048	0	0	0
GARCH-*t*	*SPAl*	0.979	0.979	0.760	0.002	0	0	GARCH-*t*	*SPAl*	0.870	0.955	0.575	0.004	0	0
	SPAc	0.665	0.409	0.136	0.002	0	0		*SPAc*	0.431	0.290	0.043	0.004	0	0
	RC	0.665	0.409	0.136	0.002	0	0		RC	0.431	0.290	0.078	0.004	0	0
GARCH-GED	*SPAl*	0.806	0.936	0.526	0	0	0	GARCH-GED	*SPAl*	0.626	0.901	0.470	0.001	0	0
	SPAc	0.422	0.264	0.046	0	0	0		*SPAc*	0.350	0.180	0.040	0.001	0	0
	RC	0.422	0.264	0.075	0	0	0		RC	0.393	0.180	0.041	0.001	0	0
EGARCH-N	*SPAl*	0.352	0.177	0.062	0.513	0.453	0.526	EGARCH-N	*SPAl*	0.338	0.172	0.062	0.640	0.625	0.650
	SPAc	0.258	0.169	0.062	0.122	0.061	0.062		*SPAc*	0.247	0.162	0.062	0.236	0.183	0.168
	RC	0.352	0.177	0.062	0.449	0.074	0.306		RC	0.338	0.172	0.062	0.563	0.247	0.454
EGARCH-*t*	*SPAl*	0.586	0.217	0.093	0.991	0.997	0.999	EGARCH-*t*	*SPAl*	0.490	0.191	0.084	0.966	0.982	0.990
	SPAc	0.369	0.197	0.093	0.599	0.595	0.646		*SPAc*	0.325	0.184	0.084	0.571	0.521	0.553
	RC	0.586	0.217	0.093	0.990	0.996	0.998		RC	0.490	0.191	0.084	0.965	0.982	0.987
EGARCH-GED	*SPAl*	0.315	0.155	0.061	0.415	0.387	0.512	EGARCH-GED	*SPAl*	0.286	0.149	0.047	0.441	0.418	0.517
	SPAc	0.227	0.143	0.061	0.077	0.034	0.051		*SPAc*	0.215	0.149	0.047	0.130	0.043	0.064
	RC	0.315	0.155	0.061	0.345	0.235	0.280		RC	0.286	0.149	0.047	0.375	0.257	0.313
GJR-N	*SPAl*	0.047	0.123	0.331	0	0	0	GJR-N	*SPAl*	0.065	0.157	0.348	0	0	0
	SPAc	0.046	0.123	0.060	0	0	0		*SPAc*	0.064	0.156	0.059	0	0	0
	RC	0.047	0.123	0.060	0	0	0		RC	0.064	0.157	0.059	0	0	0
GJR-*t*	*SPAl*	0.148	0.214	0.454	0	0	0	GJR-*t*	*SPAl*	0.212	0.250	0.488	0.001	0	0
	SPAc	0.143	0.179	0.064	0	0	0		*SPAc*	0.205	0.164	0.049	0.001	0	0
	RC	0.145	0.214	0.065	0	0	0		RC	0.206	0.213	0.049	0.001	0	0
GJR-GED	*SPAl*	0.089	0.177	0.389	0	0	0	GJR-GED	*SPAl*	0.142	0.220	0.417	0	0	0
	SPAc	0.088	0.165	0.065	0	0	0		*SPAc*	0.140	0.192	0.048	0	0	0
	RC	0.088	0.177	0.065	0	0	0		RC	0.140	0.208	0.048	0	0	0
MS-GARCH-N	*SPAl*	0.070	0.215	0.405	0	0	0	MS-GARCH-N	*SPAl*	0.090	0.234	0.447	0.001	0	0
	SPAc	0.070	0.111	0.058	0	0	0		*SPAc*	0.090	0.111	0.052	0.001	0	0
	RC	0.070	0.215	0.405	0	0	0		RC	0.090	0.234	0.447	0.001	0	0
MS-GARCH-*t*	*SPAl*	0.956	0.944	0.997	0.008	0	0	MS-GARCH-*t*	*SPAl*	0.992	0.974	1	0.031	0	0.005
	SPAc	0.667	0.556	0.532	0.008	0	0		*SPAc*	0.679	0.557	0.574	0.031	0	0.004
	RC	0.956	0.944	0.997	0.008	0	0		RC	0.992	0.974	1	0.031	0	0.005
MS-GARCH-GED	*SPAl*	0.022	0.158	0.223	0	0	0	MS-GARCH-GED	*SPAl*	0.014	0.158	0.207	0	0	0
	SPAc	0.022	0.103	0.056	0	0	0		*SPAc*	0.014	0.088	0.047	0	0	0
	RC	0.022	0.158	0.223	0	0	0		RC	0.014	0.158	0.207	0	0	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* and *SPAc* are the lower and consistent p-values from Hansen (2005), respectively. RC is the p-value from White's (2000) Reality Check test. 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 6b: Reality Check and Superior Predictive Ability Tests

		Horizon: Twenty-one days								Horizon: Sixty-three days					
		Loss Function								Loss Function					
Benchmark		MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2	Benchmark		MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
	SPAl	0.351	0.775	0.388	0	0	0		*SPAl*	0.308	0.703	0.360	0	0	0
GARCH-N	*SPAc*	0.270	0.121	0.036	0	0	0	GARCH-N	*SPAc*	0.238	0.101	0.023	0	0	0
	RC	0.272	0.121	0.264	0	0	0		RC	0.279	0.101	0.223	0	0	0
	SPAl	0.816	0.940	0.549	0.003	0	0		*SPAl*	0.745	0.92	0.505	0.006	0	0
GARCH-*t*	*SPAc*	0.422	0.254	0.043	0.003	0	0	GARCH-*t*	*SPAc*	0.359	0.218	0.027	0.006	0	0
	RC	0.469	0.254	0.054	0.003	0	0		RC	0.388	0.218	0.029	0.006	0	0
	SPAl	0.549	0.857	0.449	0.001	0	0		*SPAl*	0.471	0.830	0.421	0.001	0	0
GARCH-GED	*SPAc*	0.322	0.162	0.035	0.001	0	0	GARCH-GED	*SPAc*	0.272	0.138	0.025	0.001	0	0
	RC	0.359	0.162	0.037	0.001	0	0		RC	0.324	0.138	0.025	0.001	0	0
	SPAl	0.331	0.155	0.064	0.637	0.626	0.607		*SPAl*	0.311	0.144	0.044	0.590	0.619	0.581
EGARCH-N	*SPAc*	0.258	0.155	0.064	0.237	0.193	0.143	EGARCH-N	*SPAc*	0.243	0.144	0.044	0.203	0.175	0.128
	RC	0.331	0.155	0.064	0.561	0.251	0.404		RC	0.311	0.144	0.044	0.554	0.478	0.386
	SPAl	0.585	0.193	0.077	0.964	0.976	0.991		*SPAl*	0.594	0.189	0.063	0.980	0.979	0.993
EGARCH-*t*	*SPAc*	0.341	0.181	0.077	0.561	0.512	0.567	EGARCH-*t*	*SPAc*	0.326	0.172	0.063	0.580	0.528	0.533
	RC	0.585	0.193	0.077	0.964	0.976	0.990		RC	0.594	0.189	0.063	0.980	0.979	0.991
	SPAl	0.279	0.141	0.044	0.417	0.412	0.505		*SPAl*	0.239	0.137	0.035	0.353	0.369	0.455
EGARCH-GED	*SPAc*	0.231	0.141	0.044	0.120	0.044	0.064	EGARCH-GED	*SPAc*	0.202	0.137	0.035	0.104	0.025	0.035
	RC	0.279	0.141	0.044	0.382	0.242	0.291		RC	0.239	0.137	0.035	0.317	0.207	0.242
	SPAl	0.085	0.160	0.378	0	0	0		*SPAl*	0.080	0.151	0.355	0	0	0
GJR-N	*SPAc*	0.084	0.152	0.049	0	0	0	GJR-N	*SPAc*	0.079	0.135	0.030	0	0	0
	RC	0.084	0.160	0.049	0	0	0		RC	0.079	0.151	0.030	0	0	0
	SPAl	0.244	0.275	0.509	0.002	0	0		*SPAl*	0.240	0.266	0.500	0.004	0	0
GJR-t	*SPAc*	0.232	0.176	0.053	0.002	0	0	GJR-t	*SPAc*	0.216	0.170	0.023	0.004	0	0
	RC	0.234	0.227	0.053	0.002	0	0		RC	0.23	0.203	0.023	0.004	0	0
	SPAl	0.166	0.212	0.430	0	0	0		*SPAl*	0.156	0.218	0.404	0	0	0
GJR-GED	*SPAc*	0.161	0.172	0.042	0	0	0	GJR-GED	*SPAc*	0.151	0.167	0.025	0	0	0
	RC	0.161	0.172	0.042	0	0	0		RC	0.151	0.170	0.025	0	0	0
	SPAl	0.119	0.253	0.487	0.002	0	0		*SPAl*	0.138	0.266	0.511	0.005	0	0
MS-GARCH-N	*SPAc*	0.119	0.120	0.050	0.002	0	0	MS-GARCH-N	*SPAc*	0.138	0.113	0.040	0.005	0	0
	RC	0.119	0.253	0.487	0.002	0	0		RC	0.138	0.247	0.475	0.005	0	0
	SPAl	0.995	0.972	1	0.042	0	0.001		*SPAl*	0.998	0.979	1	0.068	0	0.002
MS-GARCH-*t*	*SPAc*	0.697	0.564	0.660	0.042	0	0.001	MS-GARCH-*t*	*SPAc*	0.718	0.566	0.742	0.067	0	0.002
	RC	0.995	0.957	1	0.042	0	0.001		RC	0.998	0.963	1	0.068	0	0.002
	SPAl	0.010	0.127	0.219	0	0	0		*SPAl*	0.010	0.114	0.197	0	0	0
MS-GARCH-GED	*SPAc*	0.010	0.079	0.039	0	0	0	MS-GARCH-GED	*SPAc*	0.010	0.057	0.028	0	0	0
	RC	0.010	0.127	0.219	0	0	0		RC	0.010	0.114	0.197	0	0	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* and *SPAc* are the lower and consistent *p*-values from Hansen (2005), respectively. RC is the *p*-value from White's (2000) Reality Check test. 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 7a: MCS $T_{R,\mathcal{M}}$ p-values

Panel A: One Day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	0.963*	1.000*	1.000*	0.021	0.002	0.001
GARCH-t	1.000*	1.000*	1.000*	0.039	0.007	0.006
GARCH-GED	1.000*	1.000*	1.000*	0.016	0.004	0.003
EGARCH-N	0.796*	0.663*	0.846*	0.201	0.360*	0.265*
EGARCH-t	1.000*	0.959*	0.981*	1.000*	1.000*	1.000*
EGARCH-GED	0.596*	0.487*	0.347*	0.074	0.245	0.229
GJR-N	0.003	0.021	1.000*	0.004	0.000	0.000
GJR-t	0.030	0.977*	1.000*	0.014	0.001	0.000
GJR-GED	0.011	0.821*	1.000*	0.019	0.000	0.000
MS-GARCH-N	0.010	0.994*	1.000*	0.024	0.001	0.001
MS-GARCH-t	1.000*	1.000*	1.000*	0.068	0.034	0.039
MS-GARCH-GED	0.004	0.975*	1.000*	0.014	0.005	0.002

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	1.000*	1.000*	1.000*	0.033	0.003	0.004
GARCH-t	1.000*	1.000*	1.000*	0.074	0.014	0.008
GARCH-GED	1.000*	1.000*	1.000*	0.054	0.008	0.005
EGARCH-N	0.999*	0.628*	0.930*	1.000*	0.581*	0.476*
EGARCH-t	1.000*	0.999*	0.998*	1.000*	1.000*	1.000*
EGARCH-GED	0.781*	0.527*	0.346*	0.309*	0.203	0.226
GJR-N	0.001	0.870*	1.000*	0.034	0.007	0.000
GJR-t	0.993*	1.000*	1.000*	0.068	0.007	0.001
GJR-GED	0.261*	0.994*	1.000*	0.048	0.005	0.000
MS-GARCH-N	0.428*	1.000*	1.000*	0.087	0.016	0.011
MS-GARCH-t	1.000*	1.000*	1.000*	0.158	0.058	0.063
MS-GARCH-GED	0.002	0.982*	1.000*	0.027	0.006	0.005

Panel C: Twenty-one Day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	0.907*	1.000*	0.000	0.004	0.009	0.014
GARCH-t	1.000*	1.000*	0.000	0.006	0.018	0.026
GARCH-GED	1.000*	1.000*	0.000	0.007	0.012	0.019
EGARCH-N	0.533*	0.368*	0.249	0.999*	0.189	0.056
EGARCH-t	1.000*	0.913*	0.205	1.000*	1.000*	1.000*
EGARCH-GED	0.431*	0.379*	0.227	0.617*	0.035	0.021
GJR-N	1.000*	1.000*	0.000	0.007	0.013	0.015
GJR-t	1.000*	1.000*	0.000	0.011	0.024	0.033
GJR-GED	1.000*	1.000*	0.000	0.007	0.021	0.024
MS-GARCH-N	1.000*	1.000*	0.000	0.561*	0.082	0.117
MS-GARCH-t	1.000*	1.000*	1.000*	1.000*	0.161	0.218
MS-GARCH-GED	1.000*	1.000*	0.000	0.008	0.030	0.031

Panel D: Sixty-three day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	0.162	0.000	0.000	0.000	0.007	0.006
GARCH-t	0.000	0.000	0.000	0.000	0.010	0.007
GARCH-GED	0.002	0.000	0.000	0.000	0.008	0.008
EGARCH-N	0.216	0.147	0.095	0.378*	0.004	0.129
EGARCH-t	0.239	0.164	0.096	0.897*	1.000*	1.000*
EGARCH-GED	0.231	0.169	0.086	0.463*	0.020	0.004
GJR-N	0.002	0.088	0.005	0.000	0.024	0.038
GJR-t	0.005	1.000*	0.006	0.000	0.034	0.058
GJR-GED	0.002	1.000*	0.004	0.000	0.030	0.040
MS-GARCH-N	0.022	0.263*	0.000	1.000*	0.099	0.108
MS-GARCH-t	1.000*	1.000*	1.000*	1.000*	0.206	0.238
MS-GARCH-GED	0.002	0.996*	0.003	0.904*	0.069	0.065

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

Table 7a: MCS $T_{\max,\mathcal{M}}$ *p*-values

Panel A: One Day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	0.128	1.000*	1.000*	1.000*	1.000*	1.000*
GARCH-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GARCH-GED	0.883*	1.000*	1.000*	1.000*	1.000*	1.000*
EGARCH-N	0.074	1.000*	0.000	1.000*	1.000*	1.000*
EGARCH-t	0.712*	1.000*	0.000	1.000*	1.000*	1.000*
EGARCH-GED	0.200	1.000*	0.000	1.000*	1.000*	1.000*
GJR-N	0.000	0.963*	0.392*	1.000*	1.000*	0.993*
GJR-t	0.000	1.000*	1.000*	1.000*	1.000*	1.000*
GJR-GED	0.000	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-N	0.000	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-GED	0.000	1.000*	0.000	0.296*	1.000*	1.000*

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GARCH-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GARCH-GED	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
EGARCH-N	1.000*	0.991*	0.000	1.000*	1.000*	1.000*
EGARCH-t	1.000*	1.000*	0.000	1.000*	1.000*	1.000*
EGARCH-GED	1.000*	0.939*	0.000	1.000*	1.000*	1.000*
GJR-N	0.810*	1.000*	1.000*	0.980*	1.000*	1.000*
GJR-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GJR-GED	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-N	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-GED	0.000	1.000*	0.000	0.000	1.000*	1.000*

Panel C: Twenty-one Day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	1.000*	1.000*	1.000*	1.000*	1.000*	0.929*
GARCH-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GARCH-GED	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
EGARCH-N	1.000*	0.000	0.000	1.000*	1.000*	1.000*
EGARCH-t	1.000*	0.000	0.000	1.000*	1.000*	1.000*
EGARCH-GED	1.000*	0.000	0.000	1.000*	1.000*	1.000*
GJR-N	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GJR-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
GJR-GED	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-N	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-t	1.000*	1.000*	1.000*	1.000*	1.000*	1.000*
MS-GARCH-GED	1.000*	0.999*	1.000*	1.000*	1.000*	1.000*

Panel D: Sixty-three day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	1.000*	0.000	1.000*	1.000*	0.109	0.000
GARCH-t	1.000*	0.000	1.000*	1.000*	0.000	0.000
GARCH-GED	1.000*	0.000	1.000*	1.000*	0.003	0.000
EGARCH-N	1.000*	0.000	0.000	1.000*	0.000	0.933*
EGARCH-t	1.000*	0.000	0.000	1.000*	1.000*	1.000*
EGARCH-GED	1.000*	0.000	0.000	1.000*	0.021	1.000*
GJR-N	1.000*	1.000*	1.000*	1.000*	0.000	0.000
GJR-t	1.000*	1.000*	1.000*	1.000*	0.000	0.000
GJR-GED	1.000*	1.000*	1.000*	1.000*	0.000	0.000
MS-GARCH-N	1.000*	1.000*	1.000*	1.000*	0.000	0.000
MS-GARCH-t	1.000*	1.000*	1.000*	1.000*	0.003	1.000*
MS-GARCH-GED	1.000*	1.000*	1.000*	1.000*	0.000	0.000

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

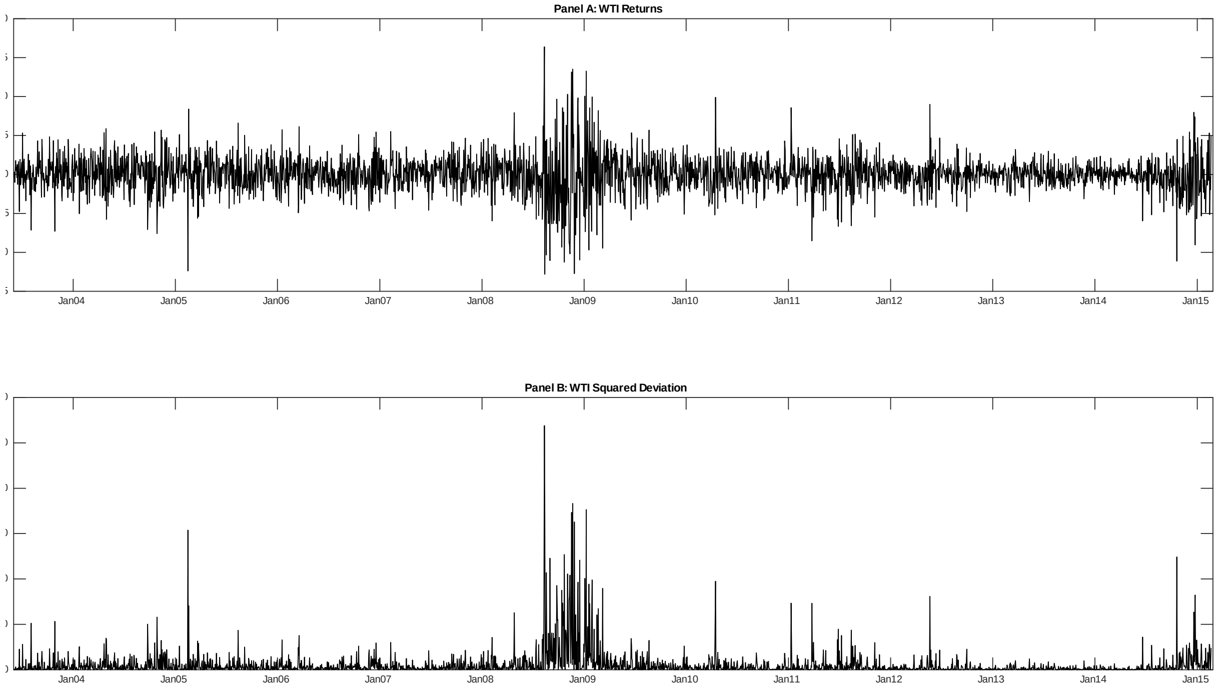

Figure 1: Daily WTI Crude Oil Returns and Squared Deviations. The sample period extends from July 1, 2003 through April 2, 2015.

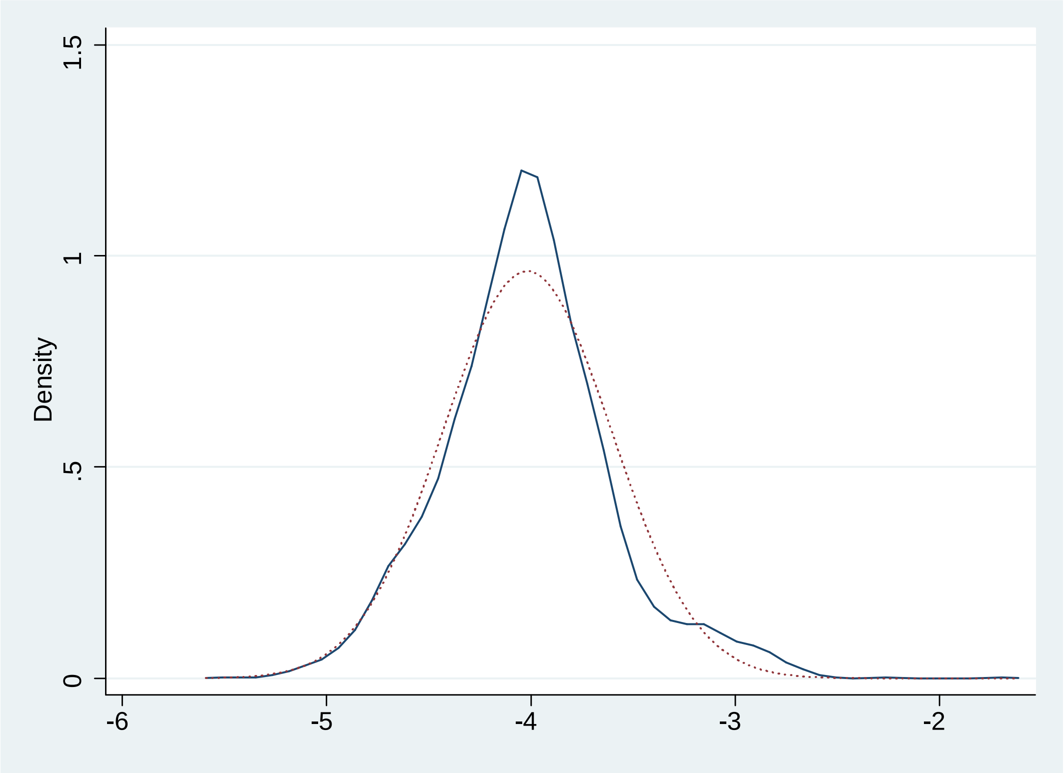

Figure 2: $\ln(RV^{1/2})$ distributions. The solid line is the kernel density. The dotted line is a normal density scaled to have the same mean and standard deviation of the data. The sample period extends from July 1, 2003 through April 2, 2015.

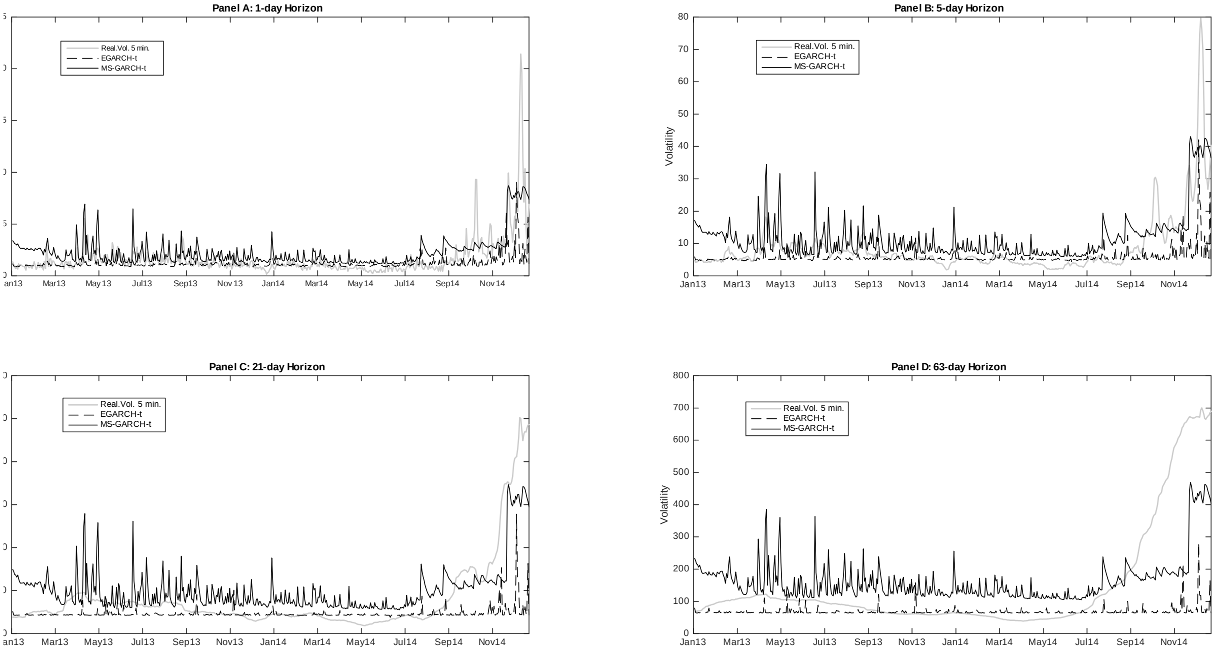

Figure 3: Volatility Forecast Comparisons for Select Models. The out-of-sample period extends from January 2, 2013 through Dec 31, 2014.

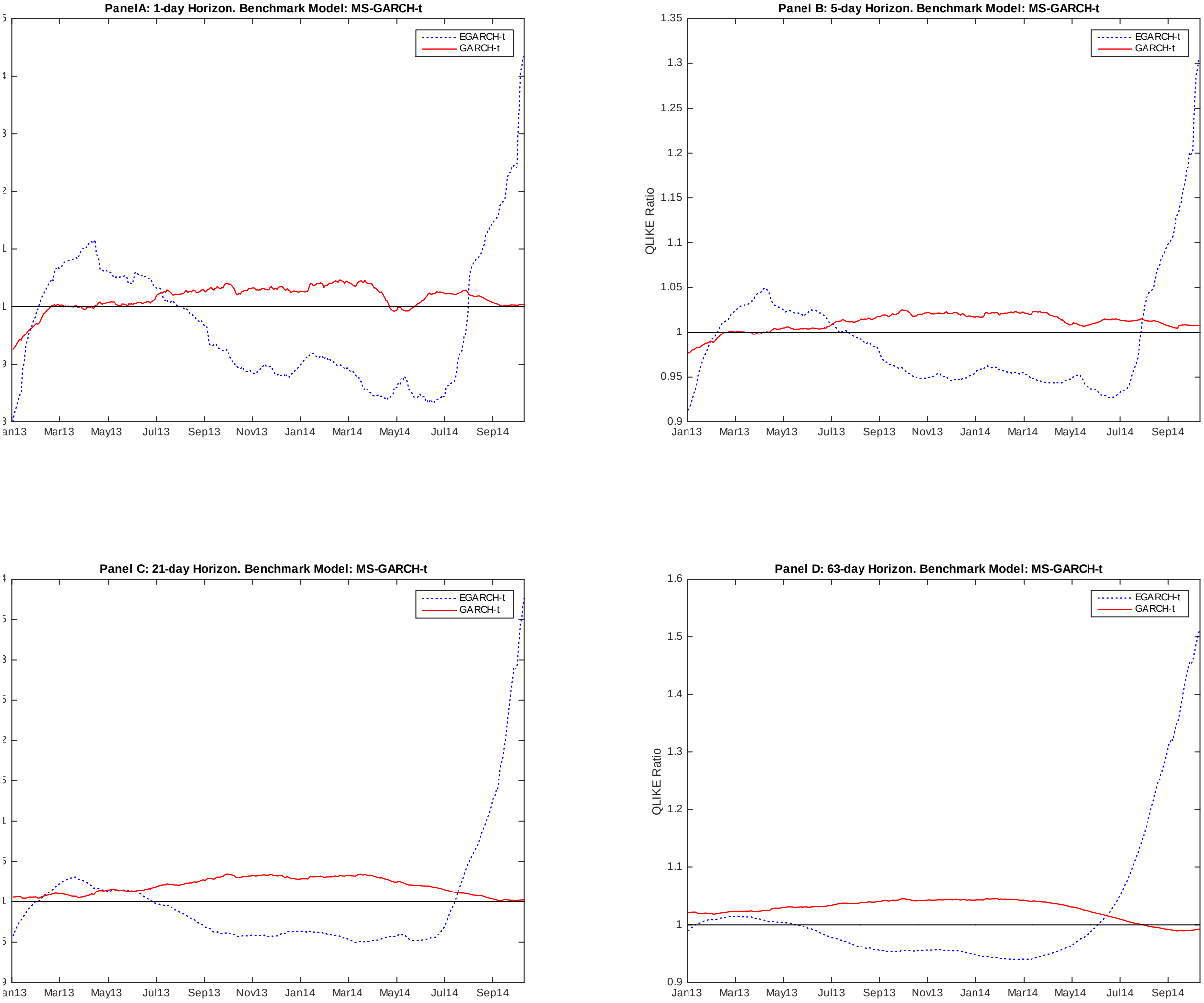

Figure 4: Rolling Window MSPE Ratio Relative to MS-GARCH-t model

7 Appendix

7.1 Conventional GARCH Models

The first model we estimate is the standard GARCH$(1, 1)$ proposed by Bollerslev (1986):

$$\begin{cases} y_t = \mu_t + \varepsilon_t, \\ \varepsilon_t = \sqrt{h_t} \cdot \eta_t, \ \eta_t \sim iid(0, 1) \\ h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 h_{t-1}, \end{cases} \tag{7}$$

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 $\alpha_1 + \gamma_1 \leq 1$.[24]

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

$$f_t(y_t) = f_t(y_t; \psi) = \frac{1}{\sqrt{h_t(\theta)}} f\left(\frac{\varepsilon_t(\theta)}{\sqrt{h_t(\theta)}}; \nu\right). \tag{8}$$

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

$$f(\eta_t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\eta_t^2}{2}\right). \tag{9}$$

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

$$f(\eta_t; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\eta_t^2}{\nu - 2}\right)^{-\frac{(\nu+1)}{2}}, \tag{10}$$

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. Then, 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 modeled as

$$f(\eta_t; \nu) = \frac{\nu \exp\left[-\frac{1}{2}\left|\frac{\eta_t}{\lambda}\right|^\nu\right]}{\lambda 2^{\left(1+\frac{1}{\nu}\right)}\Gamma\left(\frac{1}{\nu}\right)}, \tag{11}$$

with

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

[24] When $\alpha_1 + \gamma_1 = 1$, ε_t becomes an integrated GARCH process, where a shock to the variance will remain in the system. However, it is still possible for it to come from a strictly stationary process, see Nelson (1990).

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.

For the Exponential GARCH (EGARCH) model introduced by Nelson (1991) the logarithm of the conditional variance is defined as

$$\log(h_t) = \alpha_0 + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| - E \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| \right) + \xi \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \log(h_{t-1}). \tag{12}$$

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. The EGARCH model allows for an asymmetric effect, which is measured by a negative ξ. The effect of a positive standardized shock on the logarithmic conditional variance is $\alpha_1 + \xi$; the effect of a negative standardized shock would be $\alpha_1 - \xi$ instead.

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 (10),

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

Under the GED distribution specified in (11),

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

Finally, the conditional variance for the GJR-GARCH developed by Glosten, Jagannathan, and Runkle (1993) is modeled as

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

where $\mathcal{I}_{\{\omega\}}$ is the indicator function equal to one if ω is true, and zero otherwise. Then the asymmetric effect is characterized by a significant ξ. ML estimation of GJR-GARCH can be conducted similarly under different distributional specifications.

7.2 MS-GARCH

Parameter estimates for the $MS-GARCH$ model can be obtained by maximizing the log likelihood function

$$\mathcal{L} = \sum_{t=1}^{T} \log\left[p_{1,t} f_t(y_t \mid S_t = 1) + p_{2,t} f_t(y_t \mid S_t = 2)\right],$$

where $f_t(y_t \mid S_t = i)$ is the conditional density of y_t given regime i occurs at time t, and the ex-ante probabilities $p_{j,t}$ are calculated as

$$p_{j,t} = \Pr(S_t = j \mid \mathcal{I}_{t-1}) = \sum_{i=1}^{2} p_{ij} \frac{f_{t-1}(y_{t-1} \mid S_{t-1} = i) p_{i,t-1}}{\sum_{k=1}^{2} f_{t-1}(y_{t-1} \mid S_{t-1} = k) p_{k,t-1}}, j = 1, 2.$$

Define a 2×2 matrix A with $A_{ij} = \Pr(S_{t-1} = j \mid S_t = i)(\alpha_1^{(i)} + \gamma_1^{(i)})$, Klaassen (2002) derives the necessary conditions for second-order stationarity to be $A_{11}, A_{22} < 1$ and $\det(I_2 - A) > 0$. That is,

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

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

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. It is interesting 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 is small.

7.3 Forecast Evaluation Metrics

7.3.1 Statistical Loss Functions

The statistical loss functions used in this paper are defined as follows. Let the σ_t^2 denote the latent volatility, which is replaced by the 5-minute realized volatility, and $\hat{h}_t$ denote the model forecast. The first two metrics are the usual mean squared error (MSE) functions given by

$$MSE_1 = n^{-1} \sum_{t=1}^{n} \left(\sigma_t - \hat{h}_t^{1/2} \right)^2 \tag{13}$$

and

$$MSE_2 = n^{-1} \sum_{t=1}^{n} \left(\sigma_t^2 - \hat{h}_t \right)^2 . \tag{14}$$

We also compute two Mean Absolute Deviation (MAD) functions. These are given by

$$MAD_1 = n^{-1} \sum_{t=1}^{n} \left| \sigma_t - \hat{h}_t^{1/2} \right| , \tag{15}$$

$$MAD_2 = n^{-1} \sum_{t=1}^{n} \left| \sigma_t^2 - \hat{h}_t \right| . \tag{16}$$

The last two criteria are the R^2LOG and the $QLIKE$:

$$R^2LOG = n^{-1}\sum_{t=1}^{n}\left[\log(\sigma_t^2\hat{h}_t^{-1})\right]^2, \tag{17}$$

$$QLIKE = n^{-1}\sum_{t=1}^{n}\left(\log\hat{h}_t + \sigma_t^2\hat{h}_t^{-1}\right). \tag{18}$$

Equation (17) represents the logarithmic loss function of Pagan and Schwert (1990), whereas (18) is equivalent to the loss implied by a Gaussian likelihood.

7.3.2 Success Ratio and Directional Accuracy

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

$$SR = n^{-1}\sum_{t=1}^{n}\mathcal{I}_{\left\{\overline{\sigma_t^2}\cdot\overline{h_t}>0\right\}}, \tag{19}$$

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

$$DA = \frac{SR - SRI}{\sqrt{Var(SR) - Var(SRI)}}, \tag{20}$$

where $Var(SR) = n^{-1}SRI(1-SRI)$ and $Var(SRI) = n^{-1}(2\widehat{P}-1)^2P(1-P) + n^{-1}(2P-1)^2\widehat{P}(1-\widehat{P}) + 4n^{-2}P\widehat{P}(1-P)(1-\widehat{P})$. A significant DA statistic indicates the model forecast $\hat{h}_t$ 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 value 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 parameter 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 DM statistic for testing the null hypothesis of Equal Predictive Accuracy (EPA) between models i and j is

$DM = \overline{d}/\sqrt{\widehat{var}(\overline{d})}$, where the asymptotic variance $\widehat{var}(\overline{d})$ can be estimated by Newey-West's HAC estimator.[25] DM has a standard normal distribution under H_0. If the test statistic DM is significantly negative, the benchmark model is better since it has a smaller loss function; if DM 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

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

Under the assumption that $d_{0k,t}$ is stationary, the expected performance of model k relative to the benchmark can be defined as $\mu_k = E\left[d_{0k,t}\right]$ 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:

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

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.

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:

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

where $\widehat{var}(n^{\frac{1}{2}}\bar{d}_k)$ 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

[25] $\widehat{var}(\overline{d}) = n^{-1}\left(\widehat{\gamma} + 2\sum_{k=1}^{q}\omega_k\widehat{\gamma}_k\right)$, 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 $k = 1, ..., q$.

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 $\mu_{ij} = E\left[d_{ij,t}\right]$, 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:

$$\begin{aligned} H_{0,\mathcal{M}} &: \mu_{ij} = 0 \text{ for all } i, j \in \mathcal{M} \subset \mathcal{M}_0, \\ H_{A,\mathcal{M}} &: \mu_{ij} \neq 0 \text{ for some } i, j \in \mathcal{M} \subset \mathcal{M}_0, \end{aligned} \tag{21}$$

or

$$\begin{aligned} H_{0,\mathcal{M}} &: \mu_{i.} = 0 \text{ for all } i, j \in \mathcal{M} \subset \mathcal{M}_0, \\ H_{A,\mathcal{M}} &: \mu_{i.} \neq 0 \text{ for some } i, j \in \mathcal{M} \subset \mathcal{M}_0, \end{aligned} \tag{22}$$

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 $\mu_{i.} = E(d_{i.})$. According to Hansen, Lunde and Nason (2001), we construct the t-statistics as in the DM test for testing the pair (21):

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

Similarly, to test (22), the t-statistics is

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

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

Then the null hypotheses in (21) and (22) map to the two following test statistics respectively:

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

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

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