Forecasting Crude Oil Price Volatility

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

November 25, 2014

Abstract

We provide an extensive and systematic evaluation of the relative forecasting performance of several models for the volatility of daily spot crude oil prices. Empirical research over the past decades has uncovered significant gains in forecasting performance of Markov Switching GARCH models over GARCH models for the volatility of financial assets and crude oil futures. We find that, for spot oil price returns, non-switching models perform better in the short run, whereas switching models tend to do better at longer horizons.

Keywords: Crude oil price volatility, GARCH, Markov switching, forecast.
JEL codes: C22, C53, Q47

1 Introduction

Large variations in the price of crude oil have been observed during the past decade. The WTI price reached a maximum of $145.31 on July 3, 2008, possibly a consequence of geopolitical tensions over Iranian missile tests. It then fell sharply to $91.49 on September 16, 2008 in the midst of the financial crisis, and fluctuated around $40 by the end of the year. These recent large swings in the crude oil price, in conjunction with a widespread consensus that large fluctuations in the price of oil are detrimental for economic activity, has bolstered a line of research into how to improve oil price forecasts.[1] 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 the real and the nominal price of oil are predictable.

[*] 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

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

Despite this rich and growing literature, the number of studies on forecasting oil price volatility was rather limited until the 2000's. Yet, the increase in crude oil price volatility observed around the period of the global financial crisis (see Figure 1) has created new interest into how to improve volatility forecasts. Reliable forecasts of oil price volatility are of interest for various economic agents, first and most obviously, for those firms whose business greatly depends on oil prices. Examples include oil companies that need to decide whether or not to drill a new well, airline companies who use oil price forecasts to set airfares, and the automobile industry. Second, 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. Finally, oil price volatility also plays a role in households' decisions regarding purchases of durable goods, such as automobiles or heating systems.

The aim of this paper is to provide a comprehensive and systematic examination of the conditional volatility (hereafter volatility) of daily crude oil spot prices. Traditionally, oil price volatility has been modeled as a time-invariant GARCH process.[2] Nonlinear GARCH models such as EGARCH (Nelson 1991) and GJR-GARCH (Glosten, Jagannathan and Runkle 1993) are well suited for this task as they are capable of capturing features such as volatility clustering, fat tails, and possible asymmetric effects. Furthermore, these models 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). Nevertheless, 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. To the best of our knowledge, only two studies have addressed the possibility of changes in regime in oil price volatility: Fong and See (2002), and Nomikos and Pouliasis (2011). Both studies estimate Markov Switching GARCH (hereafter MS-GARCH) models to study the volatility of daily returns on oil futures, whereas the latter also estimates Mix-GARCH models.[3] Fong and See (2002) follow Gray's (1996) suggestion and 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. The evidence found in favor of switching models is mixed. Fong and See's (2002) results suggest that GARCH-t[4] and MS-GARCH-t models are very close competitors when forecasting the one-step-ahead volatility of daily GSCI oil futures. Instead, Nomikos and Pouliasis (2011) find that, for the one-step-ahead horizon, a Mix-GARCH-X[5] produces more accurate forecasts of the volatility in the returns of the NYMEX WTI oil futures.

In this paper, we model and forecast the volatility of the daily WTI spot prices instead.

[2] See Xu and Ouennich (2012) and references therein.

[3] The regime shifts are driven by i.i.d. mixture distributions, rather than by a Markov chain.

[4] t stands for Student's t distribution of the innovation.

[5] The GARCH-X model adds the squared lagged basis of futures prices to the GARCH specification of the conditional variance.

One advantage of using this price to evaluate volatility forecasts is that it is available with no delay and it is not subject to revisions. This eliminates concerns regarding differences between real-time forecasts and forecasts produced with information that only becomes available after the forecast is generated. For instance, a researcher interested in forecasting the monthly volatility using the refiners acquisition cost (RAC) would have to deal with the issue that this price is released by the Energy Information Agency with a delay and that values for the previous months tend to be revised. In contrast, the forecast we produce using only the information contained in the history of daily WTI prices is the real-time forecast. Moreover, whereas financial investors might be more interested in volatility in crude oil futures, models that investigate the role of oil price volatility in economic activity and investment decisions focus more on spot oil prices.

This paper contributes to the literature in four important dimensions. First, we evaluate the role of regime switches in the volatility of daily returns on spot oil prices. To the best of our knowledge, such a research question has only been explored by Vo (2009), who uses weekly spot prices of WTI crude oil prices to estimate a Markov switching Stochastic Volatility (SV) model and finds that incorporating regime switching into a SV model enhances forecasting power. Given that spot oil prices exhibit sudden jumps and that MS-GARCH models are well suited to capture changes in regimes triggered by sudden shocks, evaluating their relative forecasting ability is of particular interest.

Second, in contrast with previous studies on crude oil price volatility, we formally test for regime switches using a testing procedure proposed by Carrasco, Hu, and Ploberger (2014). Testing for regime switching in GARCH models is especially important since it has been noted in the literature that the commonly found high persistence in the unconditional variance in financial series 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)[6] 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.

Third, instead of following the estimation method of Gray (1996) or Haas et al. (2004), we use the technique developed by Klaassen (2002). This methodology makes efficient use of the conditional information when integrating out regimes to get rid of the path dependence. Furthermore, it has two advantages over Gray (1996): greater flexibility in capturing persistence of volatility shocks, and multi-step-ahead volatility forecasts that can be recursively calculated.[7] Meanwhile, a close look at Haas et al. (2004) reveals that their model has a simplified switching mechanism, where the regime switch occurs only in the GARCH effects. Our model, however, allows the conditional variance to switch to a different regime as well. For example, big shocks may be followed by a volatile period not only because of larger GARCH effects but also because of a possible switch to the

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

[7] By making multi-period ahead forecasts a convenient recursive procedure, Klaassen (2002) shows that MS-GARCH forecasts are better than single regime GARCH forecasts.

higher variance regime. As a result, our model allows for more flexibility in modeling the volatility and persistence levels.

Last, but not least, we assess the out-of-sample forecasting performance of the different models using a battery of tests. We first follow Hansen and Lunde (2005) in considering several statistical loss functions (e.g., mean square error, MSE, mean absolute deviation, MAD, quasi maximum likelihood, $QLike$) to evaluate out-of-sample forecasting performance, as no single criterion exists to select the best model when comparing volatility forecasts (Bollerslev et al. 1994, Lopez 2001). Then, we compute the Success Ratio (SR) and implement the Directional Accuracy (DA) tests from 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. Finally, we inquire into the stability of the forecasting accuracy for the preferred models over the evaluation period.

Our results suggest that EGARCH models yield more accurate out-of-sample forecasts at short horizons of 1 day and 5 days, whereas we generally favor MS-GARCH models at longer 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.[8] All in all, our results suggest that at longer horizons Markov switching models have superior predictive ability and yield more accurate forecasts than more restricted GARCH models where the parameters are time-invariant. Moreover, we uncover clear gains from using the MS-GARCH-t model for forecasting crude oil price volatility towards the end of the evaluation period at all horizons when comparing the mean squared prediction error ($MSPE$) of the preferred 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 Econometric Methodology

This paper focuses on the out-of-sample forecasting performance of a variety of models for predicting oil price volatility. The models considered here belong to the conventional GARCH family or are Markov Switching GARCH models. This section describes these models.

2.1 Conventional GARCH Models

The ARCH model by Engle (1982) and the GARCH model by Bollerslev (1986) have been widely employed for modeling volatility in financial assets and oil prices. Thus, the first

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

model we estimate is a standard GARCH$(1,1)$ regression model:

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

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. α_0, α_1 and γ_1 are all positive and $\alpha_1 + \gamma_1 \leq 1$.[9]

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 then given by

$$f_t(y_t) = f_t(y_t; \psi) = \frac{1}{\sqrt{h_t(\theta)}} f\left(\frac{\varepsilon_t(\theta)}{\sqrt{h_t(\theta)}}; \nu \right). \tag{2}$$

The most commonly used distributions for η_t include the standard normal, the Student's t, and the Generalized Error Distribution (GED). Both Student's t and GED are able to capture extra leptokurtosis –which is commonly observed in financial returns and oil price returns–, yet they require one additional nuisance parameter ν to be estimated, e.g., the degrees of freedom in the Student's t and the shape parameter in the GED. Namely, if we assume η_t is standard normal, there are no additional nuisance parameters for the probability density function (pdf) and it is simply

$$f(\eta_t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\eta_t^2}{2} \right). \tag{3}$$

Alternatively, if we assume η_t is distributed according to the Student's t distribution with ν degrees of freedom, the pdf of η_t is then given by

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

where $\Gamma(\cdot)$ is the Gamma function and ν is constrained to be greater than 2 so that the second moment exists and equals 1. If we assume a GED distribution, the pdf of η_t is modeled

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

with

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

[9] When $\alpha_1 + \gamma_1 = 1$, ε_t becomes an integrated GARCH (IGARCH) 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).

where $\Gamma(\cdot)$ is again the Gamma function; ν is 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. Once the distribution for η_t is specified, the parameter vector ψ can be estimated jointly using Maximum Likelihood Estimation (MLE).

A well-documented feature of financial data is the asymmetrical effects different types of shocks can have on volatility. For instance, political disruptions in the Middle East 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. Meanwhile, negative shocks seem to have a more pronounced effect on financial returns than positive shocks. The negative correlation between current returns and future volatility is known as the leverage effect. In order to allow negative and positive shocks to have a different effect on the conditional variance of oil prices we estimate the GJR-GARCH developed by Glosten, Jagannathan, and Runkle (1993). The conditional variance 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 positive ξ. ML estimation of GJR-GARCH can be conducted similarly under different distributional specifications.

Finally, a potential drawback of the standard GARCH model is the requirement that all of the parameters be positive. Nelson (1991) introduced the Exponential GARCH (EGARCH) model, which eliminated the non-negativity requirement. The logarithm of the conditional variance is described by

$$\log(h_t) = \alpha_0 + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| - E \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| \right) + \xi \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \log(h_{t-1}). \tag{6}$$

There are several interesting features of the EGARCH model. First, the equation for conditional variance is in log-linear form. Thus, the implied value of h_t can never be negative, permitting estimated coefficients to be negative. Second, 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. As Nelson (1991) argues, this allows for a more natural interpretation of the size and persistence of shocks since the standardized value of ε_{t-1} is unit-less. Finally, the EGARCH model also 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 (4),

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

Under the GED distribution specified in (5),

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

We simply plug these values in (6) and maximize the likelihood function across all parameters ψ in estimating EGARCH models.

2.2 MS-GARCH Models

As we mentioned in the introduction, a small number of studies have estimated MS-GARCH models to study the volatility of returns on oil price futures (see, e.g. Fong and See 2002, Nomikos and Pouliasis 2011). In fact, MS-GARCH models are of particular interest in the study of oil price volatility as the GARCH parameters are permitted to switch between regimes (e.g., periods that are perceived as of major political unrest versus periods of calm), thus providing flexibility over the standard GARCH models. For instance, a MS-GARCH model may better capture volatility persistence by allowing shocks to have a more persistent effect – through different GARCH parameters – during the high volatility regime and lower persistence during the low volatility regime. Meanwhile, MS-GARCH models can also capture the pressure-relieving effects of some large shocks, which may occur when large shocks that are not persistent are followed by relatively tranquil periods rather than by a switch to a higher volatility regime. Thus, a regime-switching model is flexible enough to accommodate volatility clustering and different levels of volatility persistence (Klaassen 2002).

We consider the MS-GARCH(1, 1) model given by

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

where we allow both the conditional mean $\mu^{S_t}$and the conditional variance h_t to be subject to a hidden Markov chain, S_t. In this paper, we focus on a two-state first-order Markov chain. That is, the transition probability of the current state only depends on the most adjacent past state:

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

where $\mathcal{I}_{t-2}$ denotes the information set up to $t-2$. We use p_{ij} to denote the transition probability that state i is followed by state j. We assume the Markov chain is geometric ergodic. More precisely, if S_t takes two values 1 and 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 (7) is computationally intractable, because the conditional variance h_t depends on the state-dependent h_{t-1}, consequently on all past states. Maximizing the likelihood function would require integrating out all possible unobserved regime paths, which grow exponentially with sample size T. Gray (1996) suggests integrating out the unobserved regime path $\tilde{S}_{t-1} = (S_{t-1}, S_{t-2}, ...)$ to avoid the path dependence, namely, replacing the path-dependent h_{t-1} by

$$\begin{aligned} h_{t-1} &= E_{t-2}\left[h_{t-1}^{(i)}\right] = p_{1,t-1}\left[\left(\mu_{t-1}^{(1)}\right)^2 + h_{t-1}^{(1)}\right] + p_{2,t-1}\left[\left(\mu_{t-1}^{(2)}\right)^2 + h_{t-1}^{(2)}\right] \\ &\quad - \left[p_{1,t-1}\mu_{t-1}^{(1)} + p_{2,t-1}\mu_{t-1}^{(2)}\right]^2, \end{aligned}$$

where $h_{t-1}^{(i)}$ and $\mu_{t-1}^{(i)}$ represent the the conditional variance and mean at time $t-1$ in state i, respectively, and $p_{1,t-1} = P(S_{t-1} = 1 \mid \mathcal{I}_{t-2})$ and $p_{2,t-1} = P(S_{t-1} = 2 \mid \mathcal{I}_{t-2})$ are the *ex-ante* probabilities. This specification avoids the path dependence issue and makes estimation very straightforward. But the disadvantage is that multi-step-ahead forecasting is very complicated.

In this paper we follow Klaassen (2002) and Marcucci (2005) 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

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

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

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

Similar to Gray (1996), this specification circumvents the path dependence by integrating out the path-dependent h_{t-1}. However, it uses the information set at time $t-1$ plus the current state S_t, which embodies Gray's $\mathcal{I}_{t-2}$ information set. Given that regimes are often observed to be highly persistent, S_t contains lots of information about S_{t-1}. Klaassen (2002) discovers that an empirical advantage of this specification over Gray's is the efficient use of all information available to the researcher. It also has the theoretical advantage of entailing a straightforward computation of the m-step-ahead volatility forecasts at time T as follows[10]:

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

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

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

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

Parameter estimates 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

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

The estimation method used here differs from other studies of oil price volatility that estimate MS-GARCH models. In particular, Fong and See (2002) follow Gray's (1996) suggestion and integrate out the unobserved regime paths. Nomikos and Pouliasis (2011) use the estimation method proposed by Haas et al. (2004) instead, where they simplify the regime shift mechanism to make the estimation computationally tractable. Our estimation method can be applied to the general MS-GARCH model, meanwhile making efficient use of the conditional information when integrating out regimes.

Since oil price returns exhibit similar characteristics to common financial returns, 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.

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 January 2, 1986 to April 5, 2013. Thus we have 6877 observations. Over this period of time, the average price for a barrel of crude oil was \$39.26, the median value equaled \$24.48, and the standard deviation was \$28.82. A maximum price of \$145.31 was observed on July 3, 2008; this record high was possibly a consequence of geopolitical tensions over Iranian missile tests. To model the returns in the oil price and its volatility, we calculate daily oil returns by taking 100 times the difference in the logarithm of consecutive days' closing prices. Table 1 shows the descriptive statistics for WTI rates of return. The mean rate of return is about 0.0187 with a standard deviation of 2.56. Note also that WTI returns are negatively skewed. Kurtosis is extremely high at the value of 17.70, compared with 3 for a normal distribution. These findings are consistent with previous studies by, e.g., Abosedra and Laopodis (1997), Morana (2001), Bina and Vo (2007), among others. Figure 1 plots the returns of the WTI spot prices and the squared deviations over the sample period. Large variations are observed during the period of the crude oil price collapse in 1986, the Iraq-Kuwait war in late 1990 and early 1991, the crude oil price crisis of 1998, as well as in the midst of the financial crisis in late 2008. 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.

In this paper we are interested in forecasting volatility. Two questions are of relevance here: (i) How do we measure volatility? (ii) How do we evaluate the relative forecasting performance of alternative models? We focus on the first question in this section and deal with the second question in Section 5. The main issue is that the true volatility σ_t^2 is not observable. Therefore, we need to compute some proxy for the true volatility. It seems natural to use the squared daily returns as a proxy. However, it has been noted in the literature (e.g. Andersen and Bollerslev 1998) that this is a noisy estimate of the true volatility. To see this, we can rewrite model (1) as $\varepsilon_t^2 = h_t \cdot \eta_t^2 = h_t + (\eta_t^2 - 1)h_t$. Leptokurtosis in the data suggests that the idiosyncratic component η_t would contribute a large amount of noise relative to the underlying volatility. In fact, this noise can lead to improper conclusions about the ability of the GARCH models to forecast volatility. Fortunately, the availability of high frequency futures data helps to solve this issue. We follow Andersen and Bollerslev (1998) and compute realized volatility using the sum of intraday squared futures returns at 5 minutes as our proxy of true volatility instead. Anderson et al. (2003) establish the theoretical justification for the realized volatility as an accurate measure of the underlying volatility. Furthermore, Andersen and Bollerslev (1998) test a variety of sampling frequencies for foreign exchange data to show that the realized volatility measure converges to the true volatility as the frequency of observation increases. However, they also find increasing the sampling frequency to less than 5 minutes has practical limitations due to market microstructure noise and discrete price observations. They determine 5-minute intraday returns are the best frequency for calculating their realized volatility measure. Liu, Patton, and Sheppard (2012) among others, also find that the 5-minute sampling frequency outperforms most other realized volatility measures across multiple asset classes including equities and interest rates.

Therefore, to compute our measure of realized volatility for the out-of-sample evaluation we obtained the oil futures[11] series from TickData.com. We downloaded 5-minute prices of 1-month futures contracts during the NYSE trading hours (9:30am to 4:00pm EST, Monday through Friday), excluding market holidays from January 5, 1987 (when this futures contract started trading) to April 5, 2013. Following Blair, Poon, and Taylor (2001), we constructed the daily realized volatility RV_t by summing the squared 5-minute returns during trading hours and then adding the square of the previous "overnight" return.[12]

[11] NYMEX Light Sweet Crude Oil, symbol CL.

[12] 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 not suitable here since the value for $\widehat{c}$ varies with sub-samples for our data series. For instance, prior to 7/1/2003, oil futures were traded from 10:00am until 2:30pm and $\widehat{c} = 1.19$. After 7/1/2003, trading hours were expanded to the entire day, with the exception of a 45-minute period from 5:15pm to 6:00pm when trading is halted. For the sub-sample of 7/1/2003 to 4/5/2013, $\widehat{c} = 1.85$. If instead we focus on the sample period 1/2/1992 to 1/31/1997 from Fong and See (2002), $\widehat{c} = 1.03$, whereas if we use the sample period 1/23/1991 to 12/31/1997 from Nomikos and Pouliasis (2011), $\widehat{c} = 1.33$. Finally, for our out-of-sample period 1/3/2012

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.[13]

We then evaluate the forecasting performance of various GARCH and MS-GARCH models with the realized volatility as reference. Since the forecasts will be utilized by agents who have differing investment horizons, we evaluate relative forecasting performance of the different models at various horizons. For example, central bankers typically need a monthly forecast. Oil exploration and production firms might be interested in longer horizons and this horizon might vary across regions. For instance, while the time to complete oil wells averages 20 days in Texas, it averages 90 days in Alaska. Therefore, we focus on 4 forecasting horizons at $m = 1$, 5, 22, and 66 days, corresponding to 1 day, 1 week, 1 month and 3 months respectively. Then, to calculate m-step-ahead realized volatility at time T, we simply sum the daily realized volatility over m days, denoted by:

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

We divide the whole sample into two parts: the first 6560 observations (corresponding to a period of January 2, 1986 to December 30, 2011) are used for in-sample estimation, while the remaining observations are used for out-of-sample forecast evaluation in the year 2012.[14]

4 Estimation Results

We regress the daily returns on a constant, and test the residuals for autocorrelations and ARCH effects. The Breusch-Godfrey test cannot reject the null of no serial autocorrelation.[15] However, the LM test for ARCH effects strongly rejects the null of no ARCH effect in all lag orders from 1 to 20.[16] So we start estimating our model with the conditional mean $r_t = \mu + \varepsilon_t$.

4.1 GARCH

The ML estimates for GARCH$(1,1)$, EGARCH$(1,1)$, and GJR-GARCH$(1,1)$ models are collected in Table 2. For each model, we report the results with Normal, Student's t, and GED innovations. Asymptotic standard errors are reported in parentheses.[17]

to 4/5/2013, $\widehat{c} = 2.33$.

[13] Anderson et al. (2003) have similar findings for the realized volatility on exchange rates.

[14] Our observations extend to April 5, 2013 to accommdate the m-step-ahead forecast at $m = 66$.

[15] p-value is 0.413.

[16] $p-$values are all at 0.

[17] The Maximum Likelihood estimates are obtained using the MATLAB's numerical optimization routine FMINCON. We use the nonlinear Sequential Quadratic Programming (SQP) method with FMIN-

In using a normal innovation, the conditional mean in all three models is insignificant at the 5% level. When a t or GED distribution is used, the conditional mean is significantly positive at around 0.06. Recall that the kurtosis of this return series is 17.86 from Table 1. Moreover, the degrees of freedom for the t distribution are estimated at around 6 in all three GARCH models[18] and the estimated shape parameter for GED distribution is around 1.32[19], which is consistent with the common finding in the literature that the normal error might not be able to account for all the mass in the tails in the distributions of daily returns.

In EGARCH-t and EGARCH-GED, the asymmetric effect (ξ) is significantly negative at 5%, suggesting that a negative shock would increase the future conditional variance more than a positive shock of the same magnitude. However, this asymmetric effect is not significant across all GJR specifications.

The estimates of the variance parameters reveal high persistence levels (indicated by $\alpha_1 + \gamma_1$ close to 1) throughout the GARCH specifications. In GJR and EGARCH models, the persistence levels are measured by $\alpha_1 + \gamma_1 + 0.5\xi$ and γ_1 instead. The estimates are also very close to 1, suggesting high persistence in all cases.

4.2 MS-GARCH

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 Chi-squared 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, CHP test thereafter). The advantage of this test is that it 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

CON to jointly estimate the conditional mean and conditional variance by maximizing the log-likelihood function. SQP closely mimics Newton's method for constrained optimization. For each iteration the Hessian of the Lagrangian function is updated using the BFGS quasi-Newton method.

[18] This suggests that the conditional moments exist up to the 6th order. Morever, since the conditional kurtosis for the t distribution is calculated by $3(\nu-2)/(\nu-4)$, $\nu = 6$ implies much fatter tails than normal distributions.

[19] 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 $\nu = 1.32$, the kurtosis is at 4.27, again confirming fat tails.

any subset of these parameters. We describe in detail how to conduct their test for regime switching in mean and variances. Specifically, the model under the null hypothesis (H_0) is (1) where $\mu_t = \mu$ and the alternative (H_1) model is (7).

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

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

We first obtain the MLE for the parameters $\hat{\theta}$ under H_0, where $\theta = (\mu, \alpha_0, \alpha_1, \gamma_1)'$. Then, we calculate the first and second derivatives of the log likelihood (8) with respect to θ evaluated at $\hat{\theta}$. The nuisance parameters specifying the Markov switching are not identified under H_0. Nevertheless, we denote the parameters as $\zeta = (h, \rho : \|h\| = 1, -1 < \underline{\rho} < \rho < \bar{\rho} < 1)$, where h is normalized and characterizes the direction of the alternative and ρ specifies the autocorrelation of the Markov chain. Given ζ, the first key component of the CHP test is $\Gamma_T^* = \sum \mu_{2,t}\left(\zeta, \hat{\theta}\right)/\sqrt{T}$, where

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

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

$$\text{supTS} = \sup_{\left\{h,\rho:\|h\|=1,\underline{\rho}<\rho<\bar{\rho}\right\}} \frac{1}{2}\left(\max\left(0, \frac{\Gamma_T^*}{\sqrt{\widehat{\epsilon}^{*\prime}\widehat{\epsilon}^*}}\right)\right)^2. \tag{9}$$

Alternatively, the exp test is:

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

where

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

That is, the unidentified nuisance parameters ζ are integrated out with respect to some priors in the supremum or exponential form to deliver an optimal test in the Bayesian sense.

We generate the 4×1 vector h uniformly over the unit sphere 60 times, corresponding to the switching mean and the three GARCH parameters.[20] The supTS is maximized over h and a grid search of ρ on the interval $[-0.95, 0.95]$ with the step length of 0.05.

[20] To test for switching in the variance equation only, we can simply set the first element of h to be 0 and generate the remaining 3×1 vector uniformly over the unit sphere.

Meanwhile, expTS is the average of $\Psi(h, \rho)$ above computed over those h and $\rho's$. For our data, the sup and exp test statistics are calculated to be 0.00522 and 0.675, respectively. Then we simulate the critical values by bootstrapping using $1,000$ iterations. We reject the null of constant parameters in favor of regime switching in both the mean and variance equations with p-values of 0 for both supTS and expTS. These results show 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.

Again with normal innovations, the results are not robust. For example, $\alpha_1^{(2)}$ is insignificantly different from 0, casting doubt upon the validity of the GARCH specification in this regime. Thus we focus on MS-GARCH-t and MS-GARCH-GED instead. The results are quite similar. In both models, regime 1 corresponds to a significantly positive mean, while the conditional mean in regime 2 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 1 occurs more often. About 62% of the observations are in regime 1, with the remaining 38% in regime 2. Moreover, regime 1 has a lower standard deviation than regime 2. In summary, we could call regime 1 –where most of the observations are located– the "good regime", with positive expected returns and lower volatility. In contrast, regime 2 is a "bad regime", where zero expected return is accompanied by higher volatility.

5 Forecast Evaluation

5.1 A Description of the Forecast Evaluation Methods

We compute 251 out-of-sample volatility forecasts (corresponding to the year 2012) for the 1-, 5-, 22-, and 66-step horizons using a rolling sample period. That is, we use the first 6560 daily observations spanning the period between January 2, 1986 and December 30, 2011 to estimate the volatility models; these estimates are then used to compute the forecasts at all horizons for the first out-of-sample period, January 3, 2012. We move to the next window by adding an observation at the end of the estimation period and drop an observation at the beginning, re-estimate our parameters, and compute a new forecast. We first present a description of the tests we employ and then an evaluation of the forecasts.

5.1.1 Statistical Loss Functions

Given that there is no unique criterion to select the best model when comparing volatility forecasts (see Bollerslev et al. 1994 and Lopez 2001), we follow Hansen and Lunde (2005) in computing six different loss functions for forecast evaluation. The use of various statistical functions has the advantage of allowing for a more systematic and complete comparison of the alternative forecast models. Given the volatility σ_t^2 and its model forecast $\hat{h}_t$, the first two criteria are the usual mean squared error (MSE) functions given by

$$MSE_1 = n^{-1} \sum_{t=1}^{n} \left(\sigma_t - \hat{h}_t^{1/2} \right)^2 \tag{10}$$

and

$$MSE_2 = n^{-1} \sum_{t=1}^{n} \left(\sigma_t^2 - \hat{h}_t \right)^2 . \tag{11}$$

We also compute two Mean Absolute Deviation (MAD) functions, as these criteria are more robust to outliers than the MSE functions. These are given by

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

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

Two disadvantages of the MAD are that they treat positive and negative errors symmetrically and they are not invariant to scale transformations.

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

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

$$QLIKE = n^{-1} \sum_{t=1}^{n} \left(\log \hat{h}_t + \sigma_t^2 \hat{h}_t^{-1} \right) . \tag{15}$$

Equation (14) represents the logarithmic loss function of Pagan and Schwert (1990). It 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. The $QLIKE$ is equivalent to the loss implied by a Gaussian likelihood.

5.1.2 Success Ratio and Directional Accuracy

We employ several methods to evaluate the relative forecasting performance of the different models. First, to evaluate the ability of the models to predict the direction of the change in the volatility, we calculate the Success Ratio (SR), which is the percentage of times the volatility forecasts move in the same direction as the actual volatility. Evaluating the proportion of times the direction of change in the volatility forecasts are correctly predicted is key for consumers of oil forecasts[21] because increases and decreases in volatility might have asymmetric effects on economic activity. Furthermore, we apply

[21] See Alquist and Kilian (2010).

the Directional Accuracy (DA) test of Pesaran and Timmermann (1992), which is constructed as a standardized statistic for SR and is asymptotically distributed as standard normal.

Using RV_t as a proxy for σ_t^2, the percentage of times the volatility forecasts move in the same direction as realized volatility is given by:

$$SR = n^{-1}\sum_{t=1}^{n}\mathcal{I}_{\left\{\overline{RV}_t\cdot\overline{h}_t>0\right\}}, \tag{16}$$

where $\overline{RV}_t$ is the demeaned realized volatility proxy at t, and $\overline{h}_t$ is the demeaned volatility forecast at t. If the realized 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 $\overline{RV}_t$ is positive and $\widehat{P}$ is the fraction of times that $\overline{h}_t$ is positive. The DA test is then calculated as:

$$DA = \frac{SR - SRI}{\sqrt{Var(SR) - Var(SRI)}}, \tag{17}$$

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 content for the underlying volatility RV_t.

5.1.3 Test of Equal Predictive Ability

It seems natural to rank the models according to the statistical loss functions, which would provide information about the relative forecasting ability of the different models. However, a common finding in the literature is that no unique model dominates the rest for all of the loss functions. A more rigorous comparison is obtained by evaluating the relative predictive accuracy with Diebold and Mariano's (1995) test of equal predictive ability (EPA). The EPA test is a pairwise comparison of two models, where the null hypothesis is that there is no difference in the predictive accuracy of the two forecasts. Building on the Diebold and Mariano framework, West (1996) develops the asymptotic theory for the EPA test; meanwhile Giacomini and White (2006) investigate the finite sample properties of the EPA test.

Suppose $\{\widehat{r}_{i,t}\}_{t=1}^{n}$and $\{\widehat{r}_{j,t}\}_{t=1}^{n}$ are two sequences of forecasts of the series $\{\widehat{r}_t\}$ generated by two competing models, i and j. Let $\{\widehat{e}_{i,t}\}_{t=1}^{n}$and $\{\widehat{e}_{j,t}\}_{t=1}^{n}$be the corresponding forecast errors. Consider the loss function $g(.)$ and define the difference between the two forecasts as $d_t \equiv [g(\widehat{e}_{i,t}) - g(\widehat{e}_{j,t})]$, where $g(\widehat{e}_{i,t})$ denotes the loss function for the benchmark model i and $g(\widehat{e}_{j,t})$ is the loss function for the alternative model j. Giacomini and White (2006) show that if the parameter estimates are constructed using a rolling scheme with a finite observation window, the asymptotic distribution of the sample mean loss differential $\overline{d} = \frac{1}{n}\sum_{t=1}^{n} d_t$ is asymptotically normal as long as $\{d_t\}_{t=1}^{n}$ is covariance stationary with a short memory. So the DM statistic for testing the null hypothesis of equal forecast accuracy between models i and j is simply $DM = \overline{d}/\sqrt{\widehat{V}(\overline{d})}$, where the asymptotic variance $\widehat{V}(\overline{d})$ can be estimated by Newey-West's HAC estimator.[22] 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.

5.1.4 Test of Superior Predictive Ability

In our case when the objective is to evaluate the relative performance of more than two models it is useful to consider White's (2000) Reality Check (RC) test for out-of-sample forecast evaluation. The RC evaluates whether a benchmark forecasting model is significantly outperformed by a set of alternative models given a particular loss function.

Consider comparing $l+1$ forecasting models where model 0 is defined as the benchmark model and $k = 1, ..., l$ denote the l alternative models. Let $L_{t,k} \equiv L(RV_t, \widehat{h}_{t,k})$ denote the loss if a forecast $\widehat{h}_{t,k}$ is made with the k-th model when the realized volatility equals RV_t. Similarly, the loss function of the forecasts from the benchmark model is denoted by $L_{t,0}$. The performance of the k-th forecast model relative to the benchmark is given by

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

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

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

The alternative is that the best model has a smaller loss function relative to the benchmark. If the null is rejected, then there is evidence that at least one of the competing models has a significantly smaller loss function than the benchmark. As a result, White's RC test is constructed from the test statistic

$$T_n^{RC} \equiv \max_{k=1,...,l} n^{\frac{1}{2}} \overline{f}_{n,k},$$

where $\overline{f}_{n,k} = n^{-1} \sum_{t=1}^{n} f_{t,k}$. T_n^{RC}'s asymptotic null distribution is also normal with mean 0 and some long-run variance Ω.

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

Note that 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:

$$T_n^{SPA} = \max_k \left(\frac{n^{\frac{1}{2}}\overline{f}_{n,k}}{\sqrt{\widehat{var}(n^{\frac{1}{2}}\overline{f}_{n,k})}}, 0 \right),$$

where $\widehat{var}(n^{\frac{1}{2}}\overline{f}_{n,k})$ is a consistent estimator of the variance of $n^{\frac{1}{2}}\overline{f}_{n,k}$ obtained via bootstrap. The distribution under the null is $N(\hat{\mu}, \Omega)$, where $\hat{\mu}$ is a chosen estimator for μ that conforms with H_0. Since different choices of $\hat{\mu}$ would result in 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_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. SPA_u uses the upper bound of $\hat{\mu}$ and it is a conservative test instead. It has the same asymptotic distribution as the RC test and is sensitive to the inclusion of poor models.

On a final note, the distinction between Hansen's SPA test and Diebold and Mariano's EPA test simply lies in the null hypothesis. H_0 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.

5.2 Evaluating Relative Out-of-Sample Performance

The volatility forecasts obtained from the EGARCH-GED, EGARCH-t and MS-GARCH-t models for the 1-, 5-, 22-, and 66-day horizons are collected in Figure 3.[23] 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 22-day horizon, both models are also able to forecast the major upward and downward movements in the realized volatility of oil futures. Only when we increase the forecast horizon to 66 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 a one-day forecast horizon, five out of the six loss functions rank the EGARCH-GED first. Only the MAD_2 ranks the EGARCH-GED second after the EGARCH-N. Instead, when we consider a 5-day (one-week) horizon the EGARCH-GED is uniformly ranked second, whereas the EGARCH-t is ranked first according to four of the criteria, and

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

the MS-GARCH-t is ranked first by the R^2LOG and the MAD_2. At longer horizons such as 22 and 66 days (one and three months, respectively), evidence in favor of a switching model is overwhelming: the MS-GARCH-t is ranked first by all loss functions.

The SR averages over 50% for all models and forecast horizons, indicating that all models forecast the direction of the change correctly in more that 50% of the sample. For the 5- and 22-day forecast horizons, the SR exceeds 60% for all models (averages 63% and 64%, respectively), whereas for the 1-day forecast horizon the SR ranges between 57% and 59% for four of the competing models, and equals or exceeds 60% for the remaining eight models. In addition, at a longer 66-day horizon the SR averages 55% across all models, suggesting the direction of the change is more difficult to predict for this longer 3-month horizon. 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 a majority of the models at 5- and 22-day forecast horizons. At the shorter 1-day horizon seven of the models exhibit a statistically significant DA at the 5% level. In contrast, for the 3-month forecast horizon only the three EGARCH models have a DA statistic that is significant at a 5% level.

Table 5 reports the selected DM test statistics with GARCH-t, EGARCH-GED, EGARCH-t, and MS-GARCH-t as benchmark models.[24] These test results are in line with the rankings reported in Table 4. Consider first the one-day-ahead forecast where the EGARCH-GED was ranked higher by most of the loss functions. As Table 5b shows, we reject the null of equal predictive ability at a 5% level for three of the eleven competing models under all six loss functions and three extra models under five loss functions, favoring the benchmark model. In addition, relative to the EGARCH-t and GJR-t models, the EGARCH-GED yields a more accurate forecast according to some of the loss functions. Note also that when an alternative model is used as the benchmark (see Tables 5a, 5c and 5d), we also find statistical evidence that the EGARCH-GED generates the smaller forecast errors in the majority of the cases.

Regarding the 5-day horizon (1 week), there is little statistical difference in the forecast accuracy comparison between the benchmark models and any of the non-switching models. The MS-GARCH-t is found to have equal accuracy as the benchmark models as well. In contrast, as the forecast horizon increases to 22 and 66 days (1 and 3 months), statistical evidence that the forecast accuracy differences are negative, in favor of switching models, especially for the MS-GARCH-t, is prevalent.

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 is that there are no other models that outperform the benchmark. The model in each row is the benchmark model under consideration. To economize space, the table reports the p-values only. The RC, SPA_c, and SPA_l correspond to the Reality Check p-value, Hansen's (2005) consistent, and lower p-values, respectively.[25] For the 1- and 5-day horizons, all three EGARCH models fail to

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

[25] 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.

reject the null regardless of the loss function, implying the EGARCH models outperform the other models. Meanwhile, the MS-GARCH-t also outperforms other models at the 5-day horizon, but not at the 1-day horizon (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 some of the EGARCH models, but also for the MS-GARCH-t.

It is interesting to consider here how our results differ from Fong and See (2002) and Nomikos and Pouliasis (2011), who find some evidence that MS-GARCH models are preferred over GARCH models for forecasting the volatility of oil futures. Recall that both studies use an estimation methodology that does not allow for a straightforward calculation of multi-step forecasts. Hence, they only compute one-step-ahead forecasts. Fong and See (2002) use three loss functions (MSE, MAE, which correspond to MSE_2 and MAD_2 in our paper, together with R^2) to evaluate the out-of-sample performance of a MS-GARCH-t and a GARCH-t model in forecasting one-day-ahead volatility. They find that the MS-GARCH-t yields a lower loss when the MSE_2 or MAD_2 are used, however, the ranking is reversed when the R^2 is used. Thus, it is not clear that the switching model performs unanimously better than the non-switching model for a short forecast horizon. In contrast, we evaluate volatility in spot oil prices at the 1-day horizon, where the EGARCH-GED is ranked above the switching models for five out of six loss functions. Yet, the GARCH-t and the MS-GARCH-t are closely tied: the MSE_2 ranks the GARCH-t higher but the MS-GARCH-t is preferred if we use the MAD_2. In other words, had we restricted ourselves to the models and loss functions used by Fong and See (2002), we would have reached similar conclusions. However, our consideration of a larger set of models and loss functions leads us to conclude that an EGARCH-GED performs substantially better than the alternative models.

Nomikos and Pouliasis (2011), on the other hand, consider a wider range of models and forecast evaluation methods than Fong and See (2002) but do not estimate EGARCH models. They instead focus on GARCH, MS-GARCH, and Mix-GARCH and also compute the one-step-ahead forecasts. Overall, they find evidence that the Mix-GARCH-X model yields smaller forecast errors and more accurate forecasts for NYMEX WTI futures. This result is also consistent with our finding that at the 1-day horizon MS-GARCH models are less favorable.

5.3 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 the year 2012, the EGARCH-GED and the

EGARCH-t exhibit lower $MSPE$ –as measured by the MSE_1 in (10)– for the 1- and 5-day forecast horizons respectively, whereas the MS-GARCH-t results in smaller $MSPE$ for the longer 22- and 66-day horizons. To investigate the stability of the forecast accuracy, we compute the $MSPE$ over 185 rolling sub-samples in the evaluation period, where the first sub-sample consists of the first 66 forecasts (three months) in the evaluation period, the second sub-sample is created by dropping the first forecast and adding the 67^{th} forecast at the end, and so on. In brief, these $MSPE$s are now computed as the average MSE_1 over a rolling window of size $n = 66$. Figure 4 plots the ratio of the $MSPE$ for three out of the four models relative to the "best model" at each of the four horizons, where the best model is selected based on the results reported in Table 4, i.e., for the whole evaluation period. Note that, because the last window used to compute the $MSPE$ spans the period between September 26, 2012 and December 30, 2012, the last $MSPE$ ratio is reported at September 25, 2012.

Panel A in Figure 4 illustrates that at the 1-day horizon the EGARCH-GED almost always has higher predictive accuracy than the GARCH-t as evidenced by the $MSPE$ ratio exceeding 1 over almost all of the evaluation period. In contrast, the predictive accuracy of the EGARCH-t and the EGARCH-GED is very similar. Consistent with being ranked lower in Table 4a, the MS-GARCH-t has lower predictive accuracy than EGARCH-GED for most of the evaluation sample; the only exception being the days in September 2012. Regarding the 5-day horizon (Panel B of Figure 4), the conclusions we draw from the $MSPE$ ratios are very similar to those for the 1-day horizon. Relative to the EGARCH-t, the forecast accuracy of the GARCH-t is considerably worse with the exception of the first quarter of 2012 where the ratio fluctuates around 0.8. That of the EGARCH-GED is comparable, and the accuracy of the MS-GARCH-t is worse during the first month but it is more accurate from July 2012 onwards. As for the longer 1-month and 3-month horizons, the MS-GARCH-t –which exhibits the lowest MSE_1 in Table 4 for the out-of-sample period under consideration– has been more accurate than any of the three closest competitors (Panels C and D of Figure 4) during the last two quarters of the evaluation period, especially during the last two quarters of the evaluation period.

We conclude that there are clear gains from using the MS-GARCH-t model for forecasting crude oil return volatility at longer horizons. Whereas these gains are not evident for the 1- and 5-day horizons over the one-year evaluation period (see Table 4), some gains become clear when we plot the ratio of the rolling window $MSPE$s of a sub-period of three months, especially towards the end of the evaluation period.

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 four 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 or a GED distribution are favored over those where a normal distribution is presumed. Second, for the one day horizon, nonlinear GARCH models, e.g., the EGARCH-GED and EGARCH-t, are often ranked higher in terms of loss functions and tend to yield more accurate forecasts than other GARCH or MS-GARCH models. Third, as the length of the forecast horizon increases, the MS-GARCH-t model outperforms non-switching GARCH models and other regime switching specifications. Lastly, when we analyzed the stability of the forecasting accuracy over different evaluation periods, we found clear gains from using the MS-GARCH-t model at the longer 1- and 3-months horizons as well as higher predictive accuracy for the 1-day and 5-day horizons towards the end of the evaluation period. All in all, our analysis suggested that the MS-GARCH-t model yields more accurate long-term forecasts of spot WTI return volatility.

Two 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., Anderson et al. 2006, among others). For practitioners who prefer unbiased forecasts, caution must be taken when using EGARCH models. Second, long horizon volatility forecasts that might be of interest to oil companies, such as the 1- and 3-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 22 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] 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.

[3] 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.

[4] 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.

[5] Andersen, T. G., T. Bollerslev, P. F. Christoffersen and F. X. Diebold (2006), “Volatility and Correlation Forecasting,” Handbook of Economic Forecasting, Amsterdam: North-Holland, 778-878.

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

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

[8] 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.

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

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

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

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

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

[14] Ferderer, J. P. (1996), “Oil price volatility and the macroeconomy,”*Journal of Macroeconomics,* 18:1, 1-26.

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

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

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

[24] 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.

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

[26] 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.

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

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

[29] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[44] 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
0.0187	2.56	-40.64	19.15	6.57	-0.76	17.70

$RV^{1/2}$

Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
0.021	0.014	0.0035	0.41	0.00021	6.70	112.35

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

Mean	Std. Dev	Min	Max	Variance	Skewness	Kurtosis
-3.98	0.48	-5.65	-0.90	0.23	0.58	4.50

: Note: WTI returns are over the sample period of January 2, 1986 to April 5, 2013 for 6876 observations. $RV^{1/2}$, and the natural logarithm of $RV^{1/2}$ series are from January 5,1987 to April 5, 2013 for 6580 observations.

Table 2: MLE Estimates of Standard GARCH Models

	GARCH			EGARCH			GJR		
	N	t	GED	N	t	GED	N	t	GED
μ	0.0325 (0.0217)	0.0634** (0.0230)	0.0596** (0.0219)	0.0323 (0.0227)	0.0542** (0.0227)	0.0535** (0.0219)	0.0366 (0.0239)	0.0582** (0.0230)	0.0575** (0.0222)
α_0	0.0895** (0.0095)	0.0754** (0.0120)	0.0799** (0.0134)	0.0311** (0.0031)	0.0177** (0.0041)	0.0187** (0.0043)	0.0723** (0.0093)	0.0684** (0.0134)	0.0679** (0.0133)
α_1	0.0938** (0.0039)	0.0645** (0.0060)	0.0739** (0.0063)	0.1976** (0.0072)	0.1413** (0.0117)	0.1610** (0.0118)	0.0990** (0.0045)	0.0584** (0.0083)	0.0719** (0.0078)
γ_1	0.8952** (0.0046)	0.9225** (0.0070)	0.9130** (0.0070)	0.9869** (0.0017)	0.9886** (0.0024)	0.9880** (0.0025)	0.8996** (0.0046)	0.9230** (0.0068)	0.9165** (0.0067)
ξ	-	-	-	-0.0036 (0.0040)	-0.0129* (0.0075)	-0.0138* (0.0069)	-0.0082 (0.0061)	0.0164 (0.0107)	0.0040 (0.0099)
ν	-	6.0212** (0.3930)	1.3238** (0.0229)	-	5.9813** (0.3845)	1.3213** (0.0229)	-	5.9370** (0.3896)	1.3224** (0.0239)
$Log(L)$	-14617.958	-14394.203	-14430.459	-14607.225	-14373.253	-14415.257	-14615.477	-14391.249	-14427.854

: 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 1/2/1986 to 12/31/11. 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
$\mu^{(1)}$	0.0729**	0.1148**	0.1150**
	(0.0247)	(0.0353)	(0.0330)
$\mu^{(2)}$	-0.3103**	0.0158	0.0093
	(0.1200)	(0.0337)	(0.0323)
$\sigma^{(1)}$	1.1239**	2.3263**	2.3630**
	(0.0351)	(0.0234)	(0.0254)
$\sigma^{(2)}$	42.7112**	2.5890**	2.9275**
	(0.3814)	(0.0212)	(0.0225)
$\alpha_1^{(1)}$	0.0420**	0.0214**	0.0265**
	(0.0149)	(0.0069)	(0.0072)
$\alpha_1^{(2)}$	4.27 E-09	0.1366**	0.1562**
	(0.0116)	(0.0212)	(0.0207)
$\gamma_1^{(1)}$	0.8053**	0.9616**	0.9550**
	(0.0131)	(0.0089)	(0.0091)
$\gamma_1^{(2)}$	0.9983**	0.8486**	0.8324**
	(0.0131)	(0.0206)	(0.0203)
p_{11}	0.9459**	0.9970**	0.9974**
	(0.0058)	(0.0016)	(0.0013)
p_{22}	0.7158**	0.9951**	0.9959**
	(0.0338)	(0.0024)	(0.0019)
ν	-	6.2133**	1.3506**
		(0.4160)	(0.0256)
$Log(L)$	-14520.17	-14373.95	-14410.13
$N.of\ Par.$	10	11	11
π_1	0.8401	0.6203	0.6119
π_2	0.1599	0.3797	0.3881
$\alpha_1^{(1)} + \gamma_1^{(1)}$	0.8473	0.9830	0.9815
$\alpha_1^{(2)} + \gamma_1^{(2)}$	0.9983	0.9852	0.9886

: 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 1/2/1986 to 12/31/11. 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	0.460	10	7.018	10	1.911	10	0.738	9	1.843	10	0.548	9	0.59	0.739
GARCH-t	0.450	8	6.708	7	1.904	8	0.749	10	1.843	9	0.553	10	0.57	0.446
GARCH-GED	0.449	7	6.762	8	1.905	9	0.737	8	1.831	7	0.548	8	0.57	0.253
EGARCH-N	0.405	3	5.754	5	1.867	2	0.691	2	1.735	2	**0.519**	**1**	0.63	2.280*
EGARCH-t	0.401	2	5.694	2	1.871	5	0.707	3	1.741	3	0.527	3	0.61	2.379**
EGARCH-GED	**0.396**	**1**	**5.645**	**1**	**1.867**	**1**	**0.691**	**1**	**1.721**	**1**	0.519	2	0.59	1.389
GJR-N	0.423	6	5.981	6	1.868	4	0.709	5	1.797	5	0.534	5	0.63	2.569**
GJR-t	0.416	5	5.749	4	1.873	6	0.729	6	1.798	6	0.542	6	0.61	2.142*
GJR-GED	0.410	4	5.720	3	1.868	3	0.708	4	1.775	4	0.533	4	0.61	1.973*
MS-GARCH-N	0.927	12	27.592	12	1.967	12	1.072	12	2.794	12	0.717	12	0.61	1.438
MS-GARCH-t	0.453	9	6.881	9	1.897	7	0.733	7	1.836	8	0.545	7	0.61	1.872*
MS-GARCH-GED	0.516	11	7.840	11	1.918	11	0.817	11	2.009	11	0.587	11	0.60	1.673*

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	1.184	8	85.883	9	3.521	8	0.329	6	6.386	6	0.829	6	0.63	2.588**
GARCH-t	1.103	4	74.115	5	3.517	7	0.323	5	6.162	5	0.814	5	0.61	2.084*
GARCH-GED	1.108	5	76.392	6	3.517	6	0.319	4	6.153	4	0.809	4	0.61	2.084*
EGARCH-N	1.198	9	83.609	8	3.524	10	0.340	9	6.554	9	0.851	9	0.63	2.830**
EGARCH-t	**1.020**	**1**	**63.927**	**1**	**3.512**	**1**	0.314	3	**6.014**	**1**	0.805	3	0.65	3.742**
EGARCH-GED	1.040	2	67.459	2	3.513	2	0.312	2	6.034	2	0.802	2	0.62	2.471**
GJR-N	1.302	10	98.960	10	3.523	9	0.349	10	6.928	10	0.881	10	0.65	3.337**
GJR-t	1.121	6	73.059	4	3.516	5	0.333	8	6.453	7	0.851	8	0.64	3.309**
GJR-GED	1.144	7	77.663	7	3.516	4	0.329	7	6.461	8	0.844	7	0.63	2.897**
MS-GARCH-N	4.131	12	610.011	12	3.641	12	0.748	12	13.823	12	1.467	12	0.62	2.231*
MS-GARCH-t	1.066	3	72.419	3	3.514	3	**0.309**	**1**	6.065	3	**0.801**	**1**	0.63	2.614**
MS-GARCH-GED	1.472	11	109.942	11	3.543	11	0.398	11	7.588	11	0.967	11	0.63	2.856**

: Note: The volatility proxy is given by the realized volatility calculated with five-minute returns aggregated with the overnight returns. * and ** denote 5% and 1% significance levels for the DA statistic, respectively.

Table 4b: Out-of-sample evaluation of the 22- and 66-step-ahead volatility forecasts

22-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	5.541	8	1757.638	8	5.027	8	0.324	8	32.255	8	1.945	8	0.64	2.088*
GARCH-t	4.602	4	1326.653	4	5.014	4	0.285	4	29.551	4	1.822	5	0.62	1.543
GARCH-GED	4.710	5	1397.245	5	5.015	5	0.287	5	29.647	5	1.818	4	0.62	1.508
EGARCH-N	5.905	9	1915.389	9	5.031	9	0.337	9	32.985	9	1.972	9	0.64	2.445**
EGARCH-t	3.699	2	1028.852	2	4.995	2	0.235	2	25.486	2	1.594	2	0.65	3.662**
EGARCH-GED	3.858	3	1115.967	3	4.997	3	0.239	3	25.973	3	1.611	3	0.65	3.000**
GJR-N	6.708	11	2371.000	11	5.041	10	0.365	10	35.648	10	2.089	10	0.67	3.330**
GJR-t	4.948	6	1462.159	6	5.019	6	0.301	6	30.688	6	1.875	6	0.65	2.885**
GJR-GED	5.268	7	1626.635	7	5.022	7	0.311	7	31.504	7	1.905	7	0.65	2.653**
MS-GARCH-N	24.116	12	13214.732	12	5.228	12	0.964	12	79.388	12	3.972	12	0.64	2.088*
MS-GARCH-t	**3.152**	**1**	**827.289**	**1**	**4.988**	**1**	**0.207**	**1**	**23.620**	**1**	**1.501**	**1**	0.64	2.208*
MS-GARCH-GED	6.381	10	1969.727	10	5.047	11	0.378	11	35.722	11	2.152	11	0.64	2.679**

66-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	32.161	9	29263.018	9	6.203	9	0.625	9	146.342	9	4.986	9	0.54	-0.460
GARCH-t	22.466	4	18344.934	4	6.153	4	0.475	4	114.508	4	4.071	4	0.50	-1.614
GARCH-GED	23.880	5	19971.742	5	6.160	6	0.496	6	118.963	5	4.197	5	0.51	-1.519
EGARCH-N	35.213	10	33789.444	10	6.214	10	0.656	10	156.743	10	5.237	10	0.61	2.229*
EGARCH-t	14.429	2	11311.522	2	6.098	2	0.319	2	86.211	2	3.156	2	0.60	2.294*
EGARCH-GED	14.692	3	11772.157	3	6.098	3	0.320	3	86.672	3	3.160	3	0.59	1.886*
GJR-N	38.503	11	38807.487	11	6.227	11	0.699	11	164.276	11	5.435	11	0.56	0.191
GJR-t	23.963	6	20310.300	6	6.159	5	0.493	5	120.781	6	4.243	6	0.53	-0.411
GJR-GED	26.506	7	23220.206	7	6.172	7	0.532	7	127.120	7	4.416	7	0.52	-0.969
MS-GARCH-N	104.275	12	138339.793	12	6.469	12	1.500	12	327.635	12	9.476	12	0.55	-0.257
MS-GARCH-t	**10.422**	**1**	**7604.255**	**1**	**6.072**	**1**	**0.243**	**1**	**72.123**	**1**	**2.720**	**1**	0.53	-0.493
MS-GARCH-GED	28.3567	8	24032.2875	8	6.1900	8	0.5774	8	139.2564	8	4.8280	8	0.55	0.2461

: Note: The volatility proxy is given by the realized volatility calculated with five-minute returns aggregated with the overnight returns. * and ** denote 5% and 1% significance levels for the DA statistic, respectively.

Table 5a: Diebold and Mariano test - GARCH-t Benchmark

Panel A: One day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-0.74	-1.32	-0.95	0.70	0.65	-0.01
GARCH-GED	0.29	-0.77	-0.31	1.89	1.72	1.10
EGARCH-N	2.44+	1.78	3.18++	2.38+	3.32++	2.94++
EGARCH-t	2.74++	2.51+	3.28++	2.10+	2.84++	2.84++
EGARCH-GED	3.23++	2.43+	3.65++	2.97++	3.97++	3.90++
GJR-N	1.29	1.07	2.93++	1.97+	1.99+	1.05
GJR-t	2.07+	1.82	2.61++	1.51	1.85	1.81
GJR-GED	2.51+	1.74	3.16++	3.19++	3.50++	2.81++
MS-GARCH-N	-2.69**	-1.81	-2.89**	-3.93**	-4.02**	-3.19**
MS-GARCH-t	-0.18	-0.73	0.83	0.60	0.60	0.15
MS-GARCH-GED	-2.65**	-2.09*	-1.33	-2.35*	-2.49*	-2.82**

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-1.16	-1.31	-0.94	-0.49	-0.71	-1.01
GARCH-GED	-0.25	-0.89	0.08	0.78	0.68	0.12
EGARCH-N	-1.16	-1.19	-0.62	-0.81	-1.12	-1.41
EGARCH-t	1.12	1.34	0.54	0.55	0.34	0.60
EGARCH-GED	1.00	1.18	0.38	0.63	0.49	0.61
GJR-N	-1.72	-1.55	-0.96	-1.49	-2.19*	-2.16*
GJR-t	-0.40	0.21	0.11	-1.10	-2.06*	-1.86
GJR-GED	-0.91	-0.77	0.19	-0.68	-1.7	-1.87
MS-GARCH-N	-3.32**	-2.25*	-4.64**	-4.54**	-4.80**	-3.90**
MS-GARCH-t	0.54	0.38	0.30	0.73	0.35	0.32
MS-GARCH-GED	-3.19**	-2.16*	-3.24**	-3.61**	-4.01**	-3.70**

Panel C: Twenty-two day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-2.45*	-1.86	-3.43**	-3.39**	-2.76**	-2.44*
GARCH-GED	-1.04	-1.24	-0.35	-0.49	0.26	-0.29
EGARCH-N	-2.39*	-2.11*	-2.10*	-2.31*	-1.70	-1.91
EGARCH-t	3.62++	2.85++	3.69++	3.59++	4.25++	4.32++
EGARCH-GED	3.48++	3.44++	3.03++	3.07++	3.61++	3.90++
GJR-N	-2.46*	-1.90	-3.15**	-3.20**	-2.83**	-2.53*
GJR-t	-2.13*	-2.19*	-1.47	-1.89	-1.47	-1.70
GJR-GED	-2.45*	-2.16*	-2.16*	-2.47*	-1.89	-2.06*
MS-GARCH-N	-4.18**	-2.94**	-6.59**	-5.87**	-6.73**	-5.25**
MS-GARCH-t	2.20+	1.93	2.03+	2.36+	2.20+	2.21+
MS-GARCH-GED	-4.19**	-2.83**	-4.11**	-4.66**	-3.51**	-3.64**

Panel D: Sixty-six day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-5.20**	-3.60**	-8.42**	-7.20**	-11.50**	-8.47**
GARCH-GED	-3.43**	-2.64**	-4.80**	-4.43**	-5.51**	-4.56**
EGARCH-N	-4.36**	-3.45**	-5.67**	-5.29**	-6.07**	-5.59**
EGARCH-t	5.93++	5.60++	5.81++	5.53++	5.93++	6.28++
EGARCH-GED	5.72++	5.95++	5.42++	5.22++	5.42++	5.76++
GJR-N	-3.93**	-2.81**	-6.38**	-5.59**	-7.29**	-5.65**
GJR-t	-1.95	-2.09*	-1.75	-1.54	-2.62**	-2.66**
GJR-GED	-3.28**	-2.72**	-3.97**	-3.81**	-4.32**	-3.88**
MS-GARCH-N	-7.28**	-4.92**	-13.02**	-10.03**	-18.95**	-12.17**
MS-GARCH-t	3.68++	3.21++	3.88++	3.86++	3.94++	3.89++
MS-GARCH-GED	-6.74**	-6.23**	-6.64**	-6.61**	-7.56**	-7.77**

: 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 - EGARCH-GED Benchmark

Panel A: One day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-2.60**	-2.35*	-3.17**	-1.95	-2.62**	-2.61**
GARCH-t	-3.23**	-2.43*	-3.65**	-2.97**	-3.97**	-3.90**
GARCH-GED	-2.83**	-2.38*	-3.36**	-2.30*	-3.30**	-3.23**
EGARCH-N	-0.85	-0.47	-0.06	-0.06	0.09	-0.50
EGARCH-t	-0.83	-0.41	-1.54	-2.20*	-2.13*	-1.36
GJR-N	-1.19	-0.70	-0.13	-0.74	-1.28	-1.45
GJR-t	-1.73	-0.65	-0.83	-2.00*	-2.66**	-2.65**
GJR-GED	-1.10	-0.34	-0.13	-0.94	-1.71	-1.80
MS-GARCH-N	-2.84**	-1.85	-4.44**	-4.22**	-4.39**	-3.40**
MS-GARCH-t	-2.36*	-2.09*	-2.88**	-1.69	-2.08*	-2.28*
MS-GARCH-GED	-3.56**	-2.49*	-4.57**	-4.42**	-4.51**	-4.06**

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-1.38	-1.41	-0.91	-0.92	-0.81	-1.02
GARCH-t	-1.00	-1.18	-0.38	-0.63	-0.49	-0.61
GARCH-GED	-0.97	-1.21	-0.38	-0.42	-0.27	-0.49
EGARCH-N	-2.28*	-1.82	-2.94**	-2.68**	-2.72**	-2.44*
EGARCH-t	0.67	1.10	0.36	-0.25	-0.34	0.20
GJR-N	-1.85	-1.64	-1.16	-1.65	-2.14*	-2.08*
GJR-t	-1.55	-1.96	-0.34	-1.19	-2.05*	-2.29*
GJR-GED	-1.68	-1.92	-0.33	-1.02	-1.69	-1.98*
MS-GARCH-N	-3.28**	-2.24*	-4.21**	-4.32**	-4.59**	-3.80**
MS-GARCH-t	-0.31	-0.67	-0.14	0.20	0.00	-0.10
MS-GARCH-GED	-3.08**	-2.06*	-3.85**	-4.34**	-4.25**	-3.45**

Panel C: Twenty-two day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-3.49**	-2.36*	-4.37**	-4.37**	-4.73**	-4.33**
GARCH-t	-3.48**	-3.44**	-3.03**	-3.07**	-3.61**	-3.90**
GARCH-GED	-3.35**	-2.69**	-3.10**	-3.16**	-3.56**	-3.80**
EGARCH-N	-3.74**	-2.63**	-5.22**	-5.21**	-5.13**	-4.37**
EGARCH-t	1.30	1.44	0.97	0.83	0.69	0.98
GJR-N	-3.12**	-2.15*	-4.63**	-4.60**	-4.61**	-3.80**
GJR-t	-4.50**	-3.73**	-3.97**	-4.41**	-4.23**	-4.53**
GJR-GED	-3.90**	-2.91**	-4.00**	-4.25**	-4.26**	-4.27**
MS-GARCH-N	-4.28**	-2.97**	-6.77**	-6.06**	-7.15**	-5.54**
MS-GARCH-t	1.18	1.25	0.80	1.12	0.81	0.94
MS-GARCH-GED	-5.17**	-3.24**	-5.55**	-6.04**	-5.21**	-5.30**

Panel D: Sixty-six day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-6.48**	-4.74**	-7.61**	-6.95**	-9.99**	-9.86**
GARCH-t	-5.72**	-5.95**	-5.42**	-5.22**	-5.42**	-5.76**
GARCH-GED	-6.00**	-5.52**	-5.86**	-5.59**	-6.24**	-6.65**
EGARCH-N	-6.31**	-4.49**	-8.90**	-7.87**	-12.35**	-10.26**
EGARCH-t	0.84	1.15	0.19	0.28	0.14	0.46
GJR-N	-5.35**	-3.51**	-8.39**	-7.41**	-10.71**	-8.48**
GJR-t	-7.09**	-5.87**	-7.24**	-6.97**	-7.74**	-8.24**
GJR-GED	-6.11**	-4.79**	-6.77**	-6.36**	-7.77**	-7.81**
MS-GARCH-N	-7.61**	-5.09**	-12.28**	-9.85**	-20.74**	-13.66**
MS-GARCH-t	1.64	1.53	1.67	1.76	1.58	1.58
MS-GARCH-GED	-8.00**	-8.03**	-7.46**	-7.13**	-9.09**	-10.08**

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

Panel A: One day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-2.13*	-2.26*	-2.70**	-1.11	-1.59	-1.81
GARCH-t	-2.74**	-2.51*	-3.28**	-2.10*	-2.84**	-2.84**
GARCH-GED	-2.31*	-2.38*	-2.90**	-1.36	-2.07*	-2.18*
EGARCH-N	-0.27	-0.17	0.62	0.82	0.86	0.15
EGARCH-GED	0.83	0.41	1.54	2.20+	2.13+	1.36
GJR-N	-0.81	-0.48	0.35	-0.07	-0.50	-0.88
GJR-t	-1.20	-0.22	-0.35	-1.22	-1.81	-1.88
GJR-GED	-0.56	-0.08	0.46	-0.07	-0.68	-0.92
MS-GARCH-N	-2.79**	-1.85	-4.26**	-3.93**	-4.10**	-3.27**
MS-GARCH-t	-2.01*	-2.11*	-2.35*	-0.91	-1.33	-1.72
MS-GARCH-GED	-3.15**	-2.46*	-3.81**	-3.27**	-3.51**	-3.38**

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-1.28	-1.38	-0.98	-0.71	-0.60	-0.89
GARCH-t	-1.12	-1.34	-0.54	-0.55	-0.34	-0.60
GARCH-GED	-1.00	-1.27	-0.52	-0.30	-0.13	-0.47
EGARCH-N	-1.81	-1.63	-1.89	-1.63	-1.68	-1.75
EGARCH-GED	-0.67	-1.10	-0.36	0.25	0.34	-0.2
GJR-N	-1.69	-1.57	-1.26	-1.38	-1.73	-1.79
GJR-t	-1.75	-1.91	-0.57	-1.21	-1.96	-2.18*
GJR-GED	-1.55	-1.69	-0.52	-0.88	-1.38	-1.65
MS-GARCH-N	-3.24**	-2.24*	-4.28**	-4.26**	-4.46**	-3.70**
MS-GARCH-t	-0.52	-0.99	-0.26	0.26	0.09	-0.15
MS-GARCH-GED	-2.76**	-1.97*	-3.76**	-3.65**	-3.70**	-3.06**

Panel C: Twenty-two day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-3.19**	-2.22*	-4.54**	-4.34**	-4.73**	-3.96**
GARCH-t	-3.62**	-2.85**	-3.69**	-3.59**	-4.25**	-4.32**
GARCH-GED	-3.12**	-2.34*	-3.52**	-3.44**	-3.86**	-3.72**
EGARCH-N	-3.41**	-2.46*	-5.10**	-4.85**	-4.92**	-3.98**
EGARCH-GED	-1.30	-1.44	-0.97	-0.83	-0.69	-0.98
GJR-N	-2.96**	-2.10*	-4.58**	-4.39**	-4.47**	-3.56**
GJR-t	-4.23**	-3.07**	-4.83**	-4.93**	-5.06**	-4.83**
GJR-GED	-3.55**	-2.60**	-4.35**	-4.34**	-4.49**	-4.05**
MS-GARCH-N	-4.25**	-2.96**	-6.83**	-6.05**	-7.15**	-5.47**
MS-GARCH-t	1.02	1.07	0.66	1.03	0.70	0.78
MS-GARCH-GED	-4.88**	-3.00**	-6.17**	-6.37**	-5.88**	-5.56**

Panel D: Sixty-six day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-6.19**	-4.45**	-7.71**	-6.97**	-10.26**	-9.60**
GARCH-t	-5.93**	-5.60**	-5.81**	-5.53**	-5.93**	-6.28**
GARCH-GED	-5.91**	-4.97**	-6.16**	-5.80**	-6.70**	-6.97**
EGARCH-N	-5.99**	-4.27**	-8.71**	-7.66**	-11.96**	-9.62**
EGARCH-GED	-0.84	-1.15	-0.19	-0.28	-0.14	-0.46
GJR-N	-5.13**	-3.40**	-8.22**	-7.21**	-10.45**	-8.06**
GJR-t	-6.69**	-5.13**	-7.59**	-7.17**	-8.39**	-8.51**
GJR-GED	-5.79**	-4.39**	-6.89**	-6.39**	-8.09**	-7.73**
MS-GARCH-N	-7.52**	-5.05**	-12.28**	-9.82**	-20.35**	-13.30**
MS-GARCH-t	1.67	1.54	1.72	1.81	1.62	1.62
MS-GARCH-GED	-8.17**	-7.48**	-7.89**	-7.48**	-9.82**	-10.85**

: 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 5d: Diebold and Mariano test - MS-GARCH-t Benchmark

Panel A: One day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-0.38	-0.60	-1.39	-0.20	-0.21	-0.15
GARCH-t	0.18	0.73	-0.83	-0.60	-0.6	-0.15
GARCH-GED	0.26	0.55	-0.94	-0.17	-0.22	0.12
EGARCH-N	2.07+	1.71	2.77++	1.78	2.09+	2.00+
EGARCH-t	2.01+	2.11+	2.35+	0.91	1.33	1.72
EGARCH-GED	2.36+	2.09+	2.88++	1.69	2.08+	2.28+
GJR-N	1.20	1.18	2.21+	0.89	0.75	0.64
GJR-t	1.39	1.65	1.67	0.12	0.20	0.65
GJR-GED	1.70	1.62	2.17+	0.88	0.82	1.08
MS-GARCH-N	-2.83**	-1.82	-3.26**	-4.20**	-4.28**	-3.40**
MS-GARCH-GED	-3.28**	-2.46*	-3.12**	-3.92**	-3.59**	-3.39**

Panel B: Five day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-1.13	-1.26	-0.85	-0.92	-0.65	-0.84
GARCH-t	-0.54	-0.38	-0.30	-0.73	-0.35	-0.32
GARCH-GED	-0.56	-0.72	-0.29	-0.53	-0.20	-0.28
EGARCH-N	-1.24	-1.09	-0.92	-1.36	-1.16	-1.24
EGARCH-t	0.52	0.99	0.26	-0.26	-0.09	0.15
EGARCH-GED	0.31	0.67	0.14	-0.20	0.00	0.10
GJR-N	-1.63	-1.51	-0.90	-1.57	-1.61	-1.75
GJR-t	-0.63	-0.09	-0.17	-1.07	-1.14	-1.06
GJR-GED	-0.86	-0.70	-0.15	-0.93	-0.98	-1.06
MS-GARCH-N	-3.38**	-2.27*	-4.37**	-4.58**	-4.60**	-3.91**
MS-GARCH-GED	-3.17**	-2.18*	-3.77**	-4.18**	-4.40**	-3.82**

Panel C: Twenty-two day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-2.44*	-1.97*	-2.53*	-2.84**	-2.53*	-2.46*
GARCH-t	-2.20*	-1.93	-2.03*	-2.36*	-2.20*	-2.21*
GARCH-GED	-2.12*	-1.85	-1.99*	-2.29*	-2.07*	-2.09*
EGARCH-N	-2.54*	-2.10*	-2.61**	-2.98**	-2.51*	-2.45*
EGARCH-t	-1.02	-1.07	-0.66	-1.03	-0.70	-0.78
EGARCH-GED	-1.18	-1.25	-0.80	-1.12	-0.81	-0.94
GJR-N	-2.55*	-1.99*	-2.86**	-3.17**	-2.82**	-2.65**
GJR-t	-2.41*	-2.10*	-2.19*	-2.65**	-2.31*	-2.34*
GJR-GED	-2.42*	-2.09*	-2.26*	-2.65**	-2.32*	-2.34*
MS-GARCH-N	-4.13**	-2.95**	-5.83**	-5.60**	-6.02**	-5.02**
MS-GARCH-GED	-3.65**	-2.61**	-3.84**	-4.49**	-3.77**	-3.65**

Panel D: Sixty-six day Horizon

Model	MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
GARCH-N	-4.45**	-3.52**	-5.14**	-4.94**	-5.80**	-5.47**
GARCH-t	-3.68**	-3.21**	-3.88**	-3.86**	-3.94**	-3.89**
GARCH-GED	-3.73**	-3.18**	-4.01**	-3.97**	-4.15**	-4.06**
EGARCH-N	-4.52**	-3.56**	-5.52**	-5.26**	-6.29**	-5.72**
EGARCH-t	-1.67	-1.54	-1.72	-1.81	-1.62	-1.62
EGARCH-GED	-1.64	-1.53	-1.67	-1.76	-1.58	-1.58
GJR-N	-4.22**	-3.09**	-5.55**	-5.25**	-6.16**	-5.42**
GJR-t	-3.87**	-3.24**	-4.28**	-4.23**	-4.46**	-4.30**
GJR-GED	-3.85**	-3.20**	-4.26**	-4.20**	-4.46**	-4.28**
MS-GARCH-N	-7.00**	-4.91**	-9.96**	-8.58**	-13.77**	-10.78**
MS-GARCH-GED	-5.46**	-4.67**	-5.75**	-5.56**	-6.70**	-6.69**

: 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
	SPAl	0.005	0.047	0.011	0.018	0.005	0.006		*SPAl*	0.021	0.016	0.056	0.115	0.218	0.145
GARCH-N	*SPAc*	0.378	0.511	0.011	0.018	0.005	0.006	GARCH-N	*SPAc*	0.423	0.498	0.056	0.115	0.247	0.163
	RC	0.378	0.511	0.019	0.237	0.164	0.325		RC	0.423	0.498	0.138	0.397	0.511	0.520
	SPAl	0.018	0.093	0.007	0.001	0.000	0.006		*SPAl*	0.045	0.073	0.072	0.054	0.127	0.147
GARCH-t	*SPAc*	0.399	0.574	0.007	0.001	0 .000	0.006	GARCH-t	*SPAc*	0.472	0.555	0.072	0.054	0.127	0.154
	RC	0.399	0.574	0.021	0.162	0.121	0.339		RC	0.482	0.586	0.197	0.352	0.389	0.535
	SPAl	0.015	0.094	0.011	0.010	0.002	0.007		*SPAl*	0.047	0.050	0.087	0.150	0.228	0.189
GARCH-GED	*SPAc*	0.412	0.56	0.011	0.010	0.002	0.007	GARCH-GED	*SPAc*	0.471	0.536	0.087	0.154	0.281	0.224
	RC	0.412	0.560	0.024	0.232	0.169	0.356		RC	0.478	0.558	0.199	0.456	0.552	0.611
	SPAl	0.238	0.522	0.68	0.503	0.501	0.249		*SPAl*	0.104	0.077	0.325	0.316	0.342	0.296
EGARCH-N	*SPAc*	0.650	0.730	0.719	0.613	0.826	0.568	EGARCH-N	*SPAc*	0.586	0.576	0.325	0.466	0.640	0.55
	RC	0.785	0.898	0.933	0.927	0.949	0.860		RC	0.621	0.609	0.660	0.701	0.828	0.798
	SPAl	0.244	0.411	0.381	0.185	0.103	0.234		*SPAl*	0.44	0.493	0.577	0.317	0.340	0.306
EGARCH-t	*SPAc*	0.807	0.814	0.417	0.283	0.303	0.491	EGARCH-t	*SPAc*	0.872	0.957	0.800	0.544	0.53	0.639
	RC	0.836	0.913	0.832	0.606	0.612	0.745		RC	0.892	0.961	0.885	0.733	0.756	0.838
	SPAl	0.806	0.933	0.756	0.544	0.45	0.764		*SPAl*	0.517	0.321	0.691	0.916	0.882	0.822
EGARCH-GED	*SPAc*	1.000	1.000	0.980	0.991	0.911	1.000	EGARCH-GED	*SPAc*	0.995	0.944	1.000	1.000	1.000	1.000
	RC	1.000	1.000	0.999	0.999	0.977	1.000		RC	0.997	0.958	1.000	1.000	1.000	1.000
	SPAl	0.085	0.279	0.656	0.172	0.041	0.038		*SPAl*	0.040	0.047	0.378	0.162	0.213	0.152
GJR-N	*SPAc*	0.547	0.670	0.697	0.270	0.083	0.058	GJR-N	*SPAc*	0.458	0.522	0.385	0.185	0.232	0.152
	RC	0.576	0.690	0.918	0.652	0.413	0.465		RC	0.458	0.522	0.795	0.531	0.506	0.501
	SPAl	0.044	0.618	0.236	0.027	0.007	0.008		*SPAl*	0.064	0.087	0.396	0.102	0.106	0.111
GJR-t	*SPAc*	0.517	0.861	0.236	0.046	0.017	0.008	GJR-t	*SPAc*	0.561	0.569	0.430	0.108	0.106	0.162
	RC	0.612	0.903	0.676	0.313	0.238	0.433		RC	0.619	0.787	0.808	0.422	0.385	0.546
	SPAl	0.123	0.652	0.637	0.130	0.016	0.014		*SPAl*	0.081	0.094	0.527	0.274	0.231	0.187
GJR-GED	*SPAc*	0.609	0.911	0.645	0.130	0.046	0.020	GJR-GED	*SPAc*	0.585	0.619	0.830	0.307	0.356	0.391
	RC	0.723	0.939	0.964	0.624	0.409	0.510		RC	0.653	0.705	0.975	0.696	0.636	0.689
	SPAl	0.002	0.040	0.000	0.000	0.000	0.000		*SPAl*	0.001	0.019	0.000	0.000	0.000	0.000
MS-GARCH-N	*SPAc*	0.002	0.040	0.000	0.000	0.000	0.000	MS-GARCH-N	*SPAc*	0.001	0.019	0.000	0.000	0.000	0.000
	RC	0.002	0.040	0.000	0.000	0.000	0.000		RC	0.001	0.019	0.000	0.000	0.000	0.000
	SPAl	0.019	0.074	0.020	0.039	0.015	0.014		*SPAl*	0.060	0.046	0.144	0.264	0.201	0.161
MS-GARCH-t	*SPAc*	0.019	0.074	0.020	0.041	0.016	0.014	MS-GARCH-t	*SPAc*	0.088	0.077	0.159	0.345	0.22	0.186
	RC	0.394	0.558	0.053	0.275	0.196	0.341		RC	0.511	0.559	0.309	0.606	0.453	0.514
	SPAl	0.000	0.014	0.000	0.000	0.000	0.000		*SPAl*	0.000	0.000	0.000	0.000	0.000	0.000
MS-GARCH-GED	*SPAc*	0.000	0.014	0.000	0.000	0.000	0.000	MS-GARCH-GED	*SPAc*	0.000	0.000	0.000	0.000	0.000	0.000
	RC	0.216	0.493	0.001	0.007	0.003	0.085		RC	0.247	0.433	0.007	0.030	0.013	0.13

: 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-two days

		Loss Function					
Benchmark		MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
	SPAl	0.016	0.014	0.127	0.167	0.118	0.064
GARCH-N	*SPAc*	0.016	0.512	0.127	0.184	0.125	0.064
	RC	0.437	0.513	0.301	0.507	0.412	0.435
	SPAl	0.075	0.055	0.296	0.213	0.192	0.224
GARCH-t	*SPAc*	0.115	0.569	0.328	0.267	0.258	0.34
	RC	0.564	0.607	0.529	0.59	0.554	0.637
	SPAl	0.07	0.027	0.286	0.355	0.289	0.239
GARCH-GED	*SPAc*	0.100	0.535	0.319	0.451	0.382	0.336
	RC	0.552	0.555	0.514	0.700	0.639	0.662
	SPAl	0.025	0.025	0.206	0.132	0.126	0.068
EGARCH-N	*SPAc*	0.026	0.524	0.216	0.135	0.131	0.072
	RC	0.475	0.525	0.385	0.445	0.426	0.452
	SPAl	0.585	0.541	0.804	0.559	0.337	0.415
EGARCH-t	*SPAc*	0.935	0.989	0.948	0.781	0.663	0.730
	RC	0.961	0.991	0.970	0.853	0.785	0.874
	SPAl	0.281	0.071	0.570	0.693	0.741	0.637
EGARCH-GED	*SPAc*	0.805	0.728	0.918	0.997	1.000	0.999
	RC	0.902	0.821	0.951	0.997	1.000	0.999
	SPAl	0.006	0.009	0.218	0.032	0.014	0.010
GJR-N	*SPAc*	0.006	0.471	0.238	0.032	0.014	0.010
	RC	0.391	0.471	0.500	0.337	0.212	0.278
	SPAl	0.068	0.019	0.499	0.142	0.057	0.057
GJR-t	*SPAc*	0.154	0.615	0.650	0.161	0.059	0.096
	RC	0.564	0.654	0.835	0.503	0.340	0.453
	SPAl	0.046	0.024	0.551	0.217	0.104	0.078
GJR-GED	*SPAc*	0.053	0.541	0.765	0.276	0.106	0.093
	RC	0.523	0.552	0.904	0.612	0.435	0.482
	SPAl	0.000	0.014	0.000	0.000	0.000	0.000
MS-GARCH-N	*SPAc*	0.000	0.014	0.000	0.000	0.000	0.000
	RC	0.000	0.014	0.000	0.000	0.000	0.000
	SPAl	0.143	0.079	0.398	0.543	0.342	0.263
MS-GARCH-t	*SPAc*	0.331	0.110	0.540	0.797	0.539	0.469
	RC	0.671	0.625	0.667	0.880	0.694	0.695
	SPAl	0.000	0.002	0.000	0.000	0.000	0.000
MS-GARCH-GED	*SPAc*	0.000	0.002	0.000	0.000	0.000	0.000
	RC	0.252	0.442	0.020	0.059	0.006	0.079

Horizon: Sixty-six days

		Loss Function					
Benchmark		MSE1	MSE2	QLIKE	R2LOG	MAD1	MAD2
	SPAl	0.018	0.010	0.183	0.118	0.230	0.096
GARCH-N	*SPAc*	0.018	0.507	0.210	0.146	0.283	0.117
	RC	0.448	0.507	0.421	0.480	0.568	0.493
	SPAl	0.081	0.044	0.404	0.206	0.450	0.336
GARCH-t	*SPAc*	0.150	0.580	0.461	0.292	0.596	0.500
	RC	0.562	0.609	0.667	0.604	0.776	0.737
	SPAl	0.085	0.024	0.418	0.334	0.553	0.363
GARCH-GED	*SPAc*	0.142	0.557	0.463	0.442	0.725	0.504
	RC	0.580	0.576	0.661	0.715	0.848	0.770
	SPAl	0.010	0.010	0.128	0.037	0.059	0.027
EGARCH-N	*SPAc*	0.010	0.500	0.128	0.037	0.059	0.027
	RC	0.439	0.500	0.323	0.345	0.327	0.373
	SPAl	0.663	0.589	0.827	0.428	0.678	0.702
EGARCH-t	*SPAc*	0.950	0.985	0.974	0.675	0.819	0.892
	RC	0.968	0.990	0.982	0.788	0.876	0.917
	SPAl	0.311	0.085	0.564	0.468	0.709	0.655
EGARCH-GED	*SPAc*	0.668	0.673	0.875	0.800	0.979	0.863
	RC	0.870	0.775	0.908	0.872	0.982	0.954
	SPAl	0.005	0.007	0.092	0.008	0.003	0.003
GJR-N	*SPAc*	0.005	0.471	0.092	0.008	0.003	0.003
	RC	0.370	0.471	0.332	0.268	0.141	0.233
	SPAl	0.035	0.028	0.462	0.056	0.033	0.031
GJR-t	*SPAc*	0.091	0.572	0.515	0.073	0.033	0.036
	RC	0.530	0.591	0.766	0.396	0.304	0.412
	SPAl	0.023	0.011	0.506	0.085	0.058	0.030
GJR-GED	*SPAc*	0.027	0.528	0.581	0.085	0.060	0.031
	RC	0.472	0.535	0.835	0.462	0.365	0.414
	SPAl	0.000	0.005	0.000	0.000	0.000	0.000
MS-GARCH-N	*SPAc*	0.000	0.005	0.000	0.000	0.000	0.000
	RC	0.000	0.005	0.000	0.000	0.000	0.000
	SPAl	0.231	0.102	0.521	0.577	0.595	0.496
MS-GARCH-t	*SPAc*	0.395	0.194	0.691	0.839	0.810	0.643
	RC	0.748	0.663	0.779	0.918	0.862	0.842
	SPAl	0.000	0.002	0.000	0.000	0.000	0.000
MS-GARCH-GED	*SPAc*	0.000	0.002	0.000	0.000	0.000	0.000
	RC	0.250	0.453	0.020	0.048	0.011	0.076

: 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.

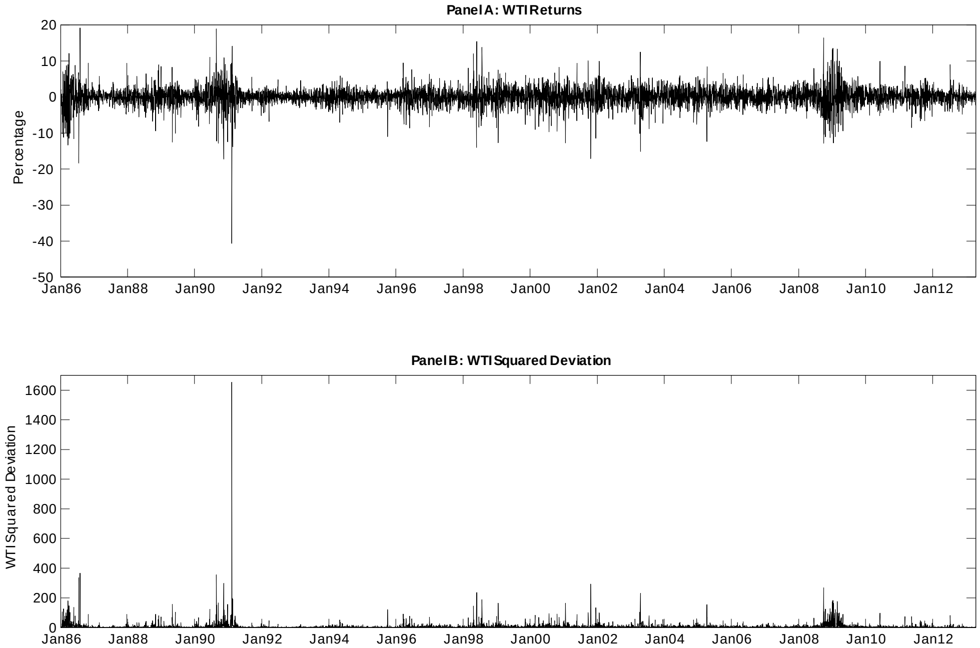

Figure 1. Daily WTI Crude Oil Returns and Squared Deviations. The sample period extends from January 2, 1987 through April 5, 2013.

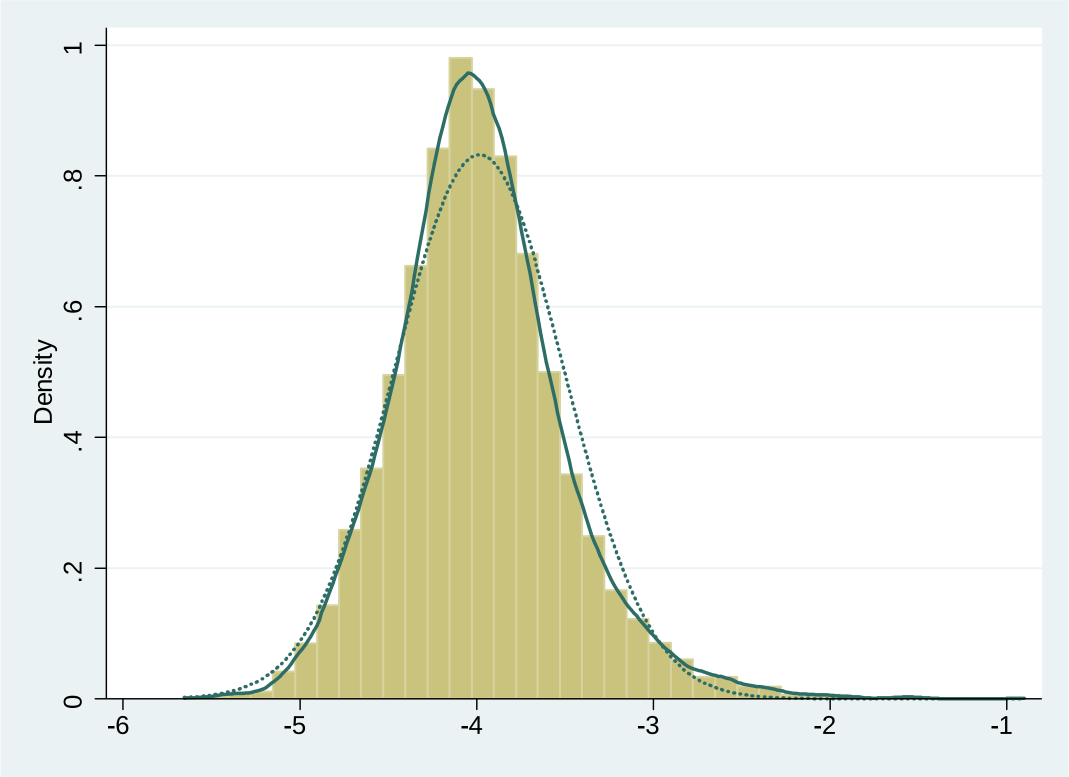

Figure 2. $\ln(RV^{1/2})$ distributions. The sample period extends from January 5, 1987 through April 5, 2013. The solid line is the kernel density. The dotted line is a normal density scaled to have the same mean and standard deviation of the data.

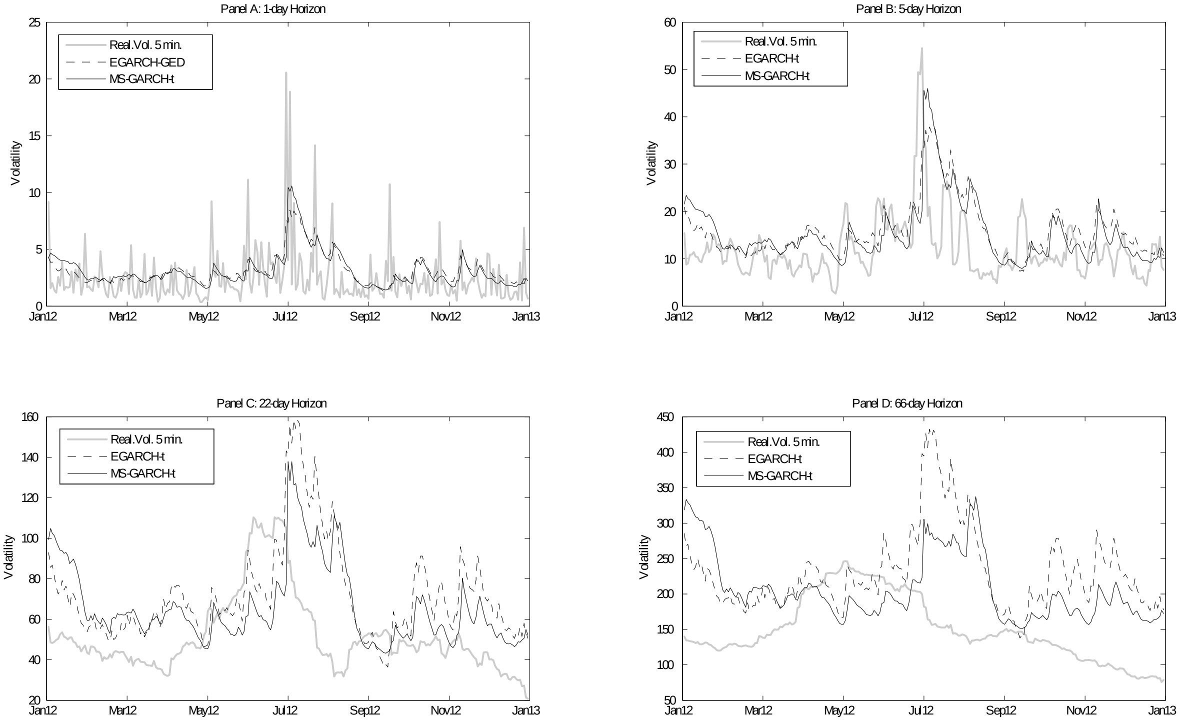

Figure 3. Volatility Forecast Comparisons for Select Models. The out-of-sample period extends from January 3, 2012 through Dec 30, 2013.

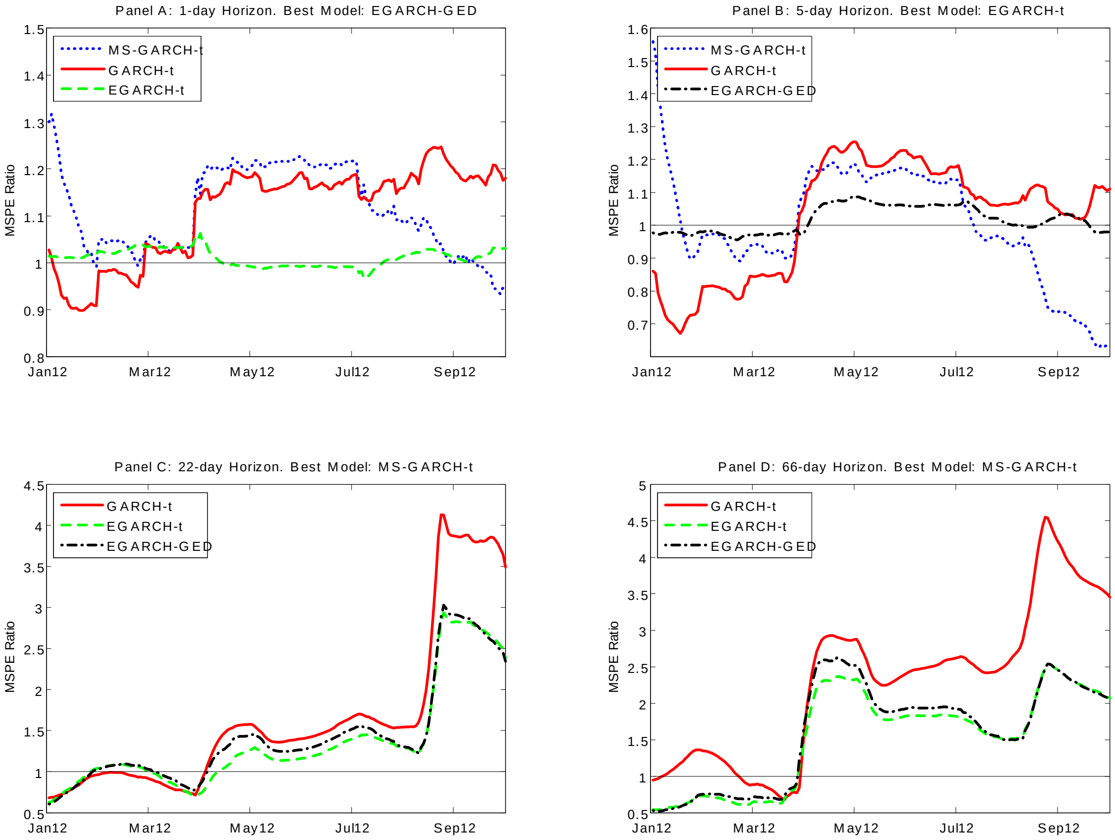

Figure 4. Rolling Window MSPE Ratio Relative to Best Models.