Title: Forecasting stock market volatility using Realized GARCH model: International Evidence

Author: Prateek Sharma Vipul

PII: S1062-9769(15)00080-0
DOI: http://dx.doi.org/doi:10.1016/j.qref.2015.07.005
Reference: QUAECO 867

To appear in: *The Quarterly Review of Economics and Finance*

Received date: 29-11-2014
Revised date: 15-7-2015
Accepted date: 23-7-2015

Please cite this article as: Sharma, P.,Forecasting stock market volatility using Realized GARCH model: International Evidence, *Quarterly Review of Economics and Finance* (2015), http://dx.doi.org/10.1016/j.qref.2015.07.005

- We examine the volatility forecasting performance of the Realized GARCH (RG) model
- Inclusion of realized measures improves the forecasting performance of GARCH models
- The relative performance of the RG and EGARCH models depend on the loss criterion
- The RG model is unable to beat the other realized volatility based models
- The EWMA model, with realized measures, provides the best volatility forecasts

Title of paper: Forecasting stock market volatility using Realized GARCH model: International evidence

Authors: Prateek Sharma and Vipul

First Author Details:
Full Name: (Mr) Prateek Sharma
Affiliation: Assistant Professor
Institute of Management Technology, Ghaziabad
Mailing Address: Pocket- F , 180 D, GTB Enclave,
Nand Nagri,
Delhi - 110093, India
Telephone Number: +91-9794305142
Email Addresses: prateeksharma1985@gmail.com (Primary)
prateek@iiml.ac.in (secondary)

Second Author Details (corresponding author):
Full Name: (Dr) Vipul
Affiliation: Professor of Finance
Indian Institute of Management, Lucknow
Mailing Address: 215, Faculty Block
Indian Institute of Management, Lucknow
Prabandh Nagar, Near Sitapur Road,
Lucknow 226013, India
Telephone Number: +91 (522) 2736607/6696607
Email Address: vipul@iiml.ac.in
Fax: 0091 (522) 2734025

Forecasting stock market volatility using Realized GARCH model: International Evidence

ABSTRACT

This article compares the forecasting ability of the recently proposed Realized GARCH model with that of the standard GARCH models that use only the daily returns, and the other time series models based on the realized measures of volatility. Each model is used for forecasting the conditional variance of 16 international stock indices, for a sample period of about 14 years. We find that the relative forecasting performance of the Realized GARCH and EGARCH models is sensitive to the choice of the loss criterion. With the realized measures, the exponentially weighted moving average model generally outperforms the Realized GARCH model in out-of-sample forecasts. This result is robust across different volatility regimes and loss criteria.

Keywords: Realized GARCH; conditional variance; forecast; stock indices; volatility.

JEL Classification: G10, G15, G17

1. Introduction

Volatility forecasting of financial assets has important implications for option pricing, portfolio selection, risk-management and volatility trading strategies. More recently, the availability of reliable high-frequency data has catalyzed the evolution of realized measures of volatility. According to the theory of quadratic variation, the realized volatility can yield very precise estimates of the latent integrated volatility, by increasing the sampling frequency sufficiently. However, this approach is constrained by the microstructure noise and price jumps. There is compelling evidence that the realized measures lead to significant statistical and economic benefits in forecasting, when incorporated in different volatility models (Andersen & Bollerslev, 1998; Andersen, Bollerslev, Diebold, & Ebens, 2001; Andersen, Bollerslev, Diebold, & Labys, 2001; Barndorff-Nielsen, 2002; Barndorff-Nielsen & Shephard, 2002; Christoffersen, Feunou, Jacobs, & Meddahi, 2014; Fleming, Kirby, & Ostdiek, 2003; Koopman, Jungbacker, & Hol, 2005; Pong, Shackleton, Taylor, & Xu, 2004; Vortelinos, 2013; Vortelinos & Thomakos, 2012). Standard GARCH models use only the daily returns for the estimation and forecasting of conditional volatility. Since the daily returns contain far less information about the current level of volatility than the numerous intraday returns, the information set of standard GARCH models is limited. Additionally, as the GARCH models rely on the moving averages with gradually decaying weights, they are slow to adapt to the changing levels of volatility (Andersen, Bollerslev, Diebold, & Labys, 2003). Therefore, there is a motivation to include the realized measures of variance within the standard GARCH framework. Engle (2002) recommended including the realized measure as an exogenous variable in the GARCH equation (Forsberg & Bollerslev, 2002). However, this naive augmentation makes the model unsuitable for multi-period forecasts. Recently, Hansen, Huang, & Shek (2012) proposed the Realized GARCH model, which provides a unique framework for the joint modeling of a realized measure of volatility and the conditional variance. The Realized GARCH framework has several appealing properties. It is simple to estimate, using the method of maximum likelihood. It allows an ARMA structure for the realized measure and the conditional variance, which is a distinctive feature of the standard GARCH framework. The realized measure is modeled under a general specification. It models the leverage effects in the realized measure through a quadratic leverage function, and accommodates any bias in the realized measure caused by microstructure noise, price jumps and non-trading hours.

Despite its attractive features, it is not readily obvious why the Realized GARCH model should provide a superior out-of-sample forecast. The principle of parsimony suggests that the simpler models usually provide better forecasts than the more complex models, if all of them incorporate all the relevant information. In the case of Realized GARCH models, even the most basic AR(1)-Realized GARCH(1,1) model requires an estimation of nine parameters, which can hardly be considered parsimonious (refer equations 4, 8, and 9). The estimation of so many parameters requires a long history of data, and can result in large estimation errors, rendering the model unsuitable for forecasting applications. The simpler models with less numbers of parameters, which incorporate the same information directly through realized measures, may outperform the sophisticated Realized GARCH model. For instance, Garg & Vipul (2014) found that the random walk and exponentially weighted moving average models provide better forecasts than the more sophisticated time series models, for the two-scale realized volatility measure (Zhang, Mykland, & Ait-Sahalia, 2005).

We compare the daily forecasts of conditional variance for sixteen international stock indices, using the standard GARCH models based on daily returns, the Realized GARCH model, and the other conventional time series models based on the realized measures of volatility. This article contributes to the existing literature in a number of ways. First, it is one of the earliest applications of the recently proposed Realized GARCH model for forecasting the volatility of a wide cross-section of international stock indices. The Realized GARCH model is still relatively nascent, with few empirical applications in the equity markets. Other notable applications include Hansen et al. (2012), Louzis, Xanthopoulos-Sisinis, & Refenes (2013), and Watanabe (2012), all of which are based solely on the U.S. equity market. Second, our study uses a data set with a relatively long sample period, as compared to most of the existing studies on realized volatility forecasting. Third, we compare the forecasting performance of the Realized GARCH models with the conventional time series models based on the realized measures of volatility, and the standard GARCH. Note that the Realized GARCH model has a more general specification that nests the alternative models used in this study. However, for the reasons discussed earlier, the simpler time series models may hold an advantage over the Realized GARCH model. Fourth, by using a number of standard and robust measures of realized volatility, we allow for a two-dimensional comparison between the forecasting models and the realized measures. However, the relative forecasting performance of the GARCH models can be sensitive to the performance evaluation criterion (Balaban, 2004; Brooks & Burke, 1998; Lee, 1991). In view of this, we use multiple evaluation methods, including robust loss functions and the Diebold-Mariano test of equal predictive ability.

The rest of this article is organized as follows. Section 2 presents the econometric methodology. Section 3 describes the sample data. Section 4 presents the empirical results and discussion. Section 5 provides a brief summary and concludes the paper.

2. METHODOLOGY

2.1 Realized Measures

We use five realized measures of volatility in this study. The first measure is the realized variance (RV) proposed by Andersen & Bollerslev (1998). The RV measure for day *t* is

$$\mathrm{RV}_t = \sum_{j=1}^{M} (r_{t,j})^2 \tag{1}$$

where $r_{t,j}$ is the *j*th intraday return of the day and *M* is the total number of intraday returns for the day. Under the ideal conditions of no microstructure noise and no price jumps, the RV provides an asymptotically consistent measure of the latent integrated variance. In practice, however, these assumptions may not be valid. At very high sampling frequencies, the price observations are contaminated by the microstructure noise caused by bid-ask bounce and price rounding, among other reasons. It induces a bias in the RV measure, which progressively increases, as the sampling frequency is increased (Bandi & Russell, 2006, 2008; Zhou, 1996). To alleviate this problem, the general approach is to sample the prices at lower frequencies, typically at 5 to 30 minutes. In this study, we use the RV measure based on 5-minute intraday returns. The problem with this approach is that the sampling at a lower frequency discards a vast amount of data. Thus, the reduction in microstructure noise comes at a significant loss of efficiency. Zhang et al. (2005) recommended a subsampling approach, which makes a more efficient use of the available data. They sample the prices at a given frequency, using a variety of non-overlapping sub-grids. A set of values of the realized measure is obtained using these subsamples, which are then averaged to yield the subsampled realized measure. We use the subsampled RV measure RVS, as the second realized measure in our study. The RVS measure is calculated using 5-

minute returns with 1-minute subsampling, in the following manner. Suppose the first value of RV is computed, using the prices sampled at the time points 9:30, 9:35, 9:40 ..., etc. Now, another value of RV is computed, using the prices at the time points 9:31, 9:36, 9:41 ..., etc. In this manner, five values of RV are computed, using five non-overlapping subsamples for each day. Since the start and end times of the subsamples may not coincide with those of the trading session, these RV values may omit a small number of observations of the trading day. To adjust for these omissions, the RV values are proportionally inflated. The RVS is then calculated as the average of these five RV values.

Another problem with the RV estimate is the bias induced by price jumps. In the presence of price jumps, the RV measures the integrated volatility plus the cumulative squared jumps (Andersen, Dobrev, & Schaumburg, 2012; Barndorff-Nielsen & Shephard, 2004a, 2004b). Andersen, Bollerslev, & Diebold (2007) found that filtering the jump component from the realized variance estimates leads to a significant improvement in the volatility forecasts. To neutralize the effect of price jumps, we use the realized bipower variation measure (BV) of Barndorff-Nielsen & Shephard (2004b). The BV for day *t* is

$$\mathrm{BV}_t = \frac{\pi}{2}\sum_{j=2}^{M}|r_{t,j}||r_{t,j-1}| \tag{2}$$

Barndorff-Nielsen & Shephard (2004b) show that the BV is robust to rare jumps, and converges to the integrated volatility. In a similar manner as for the RV measure, we include the BV calculated with 5-minute returns, and its subsampled version BVS, which uses 5-minute returns with 1-minute subsampling. Finally, we include the realized kernel estimator RK of Barndorff-Nielsen, Hansen, Lunde, & Shephard (2008) as our fifth realized measure. The RK is robust to microstructure noise, and makes use of all the intraday data. The RK for day *t* is

$$\mathrm{RK}_t = \sum_{h=-H}^{H} k\left(\frac{h}{H+1}\right)\gamma_h\,, \quad where, \quad \gamma_h = \sum_{j=|h|+1}^{M} r_{t,j}\, r_{t,j-|h|} \tag{3}$$

Here, $k(x)$ is a kernel weight function. The Parzen kernel function is used as the kernel weight function[1]. The optimal bandwidth parameter *H* is calculated using the procedure of Barndorff-Nielsen, Hansen, Lunde, & Shephard (2009).

2.2 Forecasting Models

We compare the daily conditional variance forecasts, using six models: the GARCH model (Bollerslev, 1986), the EGARCH model (Nelson, 1991), the Realized GARCH model (RG), the random walk model (RW), the moving average model (MA), and the exponentially weighted moving average model (EW). The first two models are based on daily returns. The next four models are implemented using the five realized measures described in the previous section. Overall, we use 22 forecasting models for each stock index. Due to a richer information set of intraday returns, the Realized model is expected to outperform the standard GARCH models that rely solely on the daily returns.

Following Louzis et al.(2013), we use an AR(1) specification for modeling the conditional mean of the GARCH models[2].The conditional mean equation is

$$r_t = c + \phi_1 r_{t-1} + \varepsilon_t, \qquad \varepsilon_t = \sigma_t z_t,\; z_t \sim i.i.d.\; N(0,1) \tag{4}$$

Here, r_t is the close-to-close logarithmic return for the period [*t-1*, *t*], and σ_t^2 is the estimate of the latent conditional variance of the daily returns (described later). The conditional variance equations for the various GARCH models are specified as[3]

[1] Barndorff-Nielsen, Hansen, Lunde, & Shephard (2011) show that the Parzen kernel ensures a positive estimate, while allowing for dependence or endogeneity in the microstructure noise process.

[2] Awartani & Corradi (2005) compared the forecasting performance of GARCH-type models under six different specifications of the conditional mean equation. They found that the relative forecasting performance remains consistent under different specifications of the conditional mean equation.

$$\text{GARCH(1,1)}: \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{5}$$

$$\text{EGARCH(1,1)}: \log\sigma_t^2 = \omega + \beta\log\sigma_{t-1}^2 + \tau_1 z_{t-1} + \tau_2(|z_{t-1}| - E|z_{t-1}|) \tag{6}$$

$$\text{Realized GARCH(1,1)}: \log\sigma_t^2 = \omega + \beta\log\sigma_{t-1}^2 + \gamma\log x_{t-1} \tag{7}$$

$$\log x_t = \xi + \varphi\log\sigma_t^2 + \delta(z_t) + u_t,\ where,\ u_t \sim i.i.d.\ N(0,\sigma_u^2) \tag{8}$$

Here, α, ω, β, τ_1, τ_2, γ ,ξ, φ, δ_1, δ_2 are the model parameters. Equation (8) is the measurement equation used for modeling the realized measure x_t. The variable x_t denotes the concerned realized measure, i.e., $x_t = \text{RV}_t, \text{RVS}_t, \text{BV}_t, \text{BVS}_t$ or RK_t. $\delta(\cdot)$ is the leverage function given by $\delta(z_t) = \delta_1 z_t + \delta_2(z_t^2 - 1)$. The leverage function captures the asymmetric effect of negative return shocks on the volatility process. All models are estimated using the method of maximum likelihood. Following Hansen et al. (2012) and Frommel, Han, & Kratochvil (2014), we assume Gaussian specification for the log-likelihood functions. We use a rolling window of the most recent 2000 daily observations for the estimation of GARCH models. The estimate of the variance at the end of day t, σ_{t+1}^2, is used as the one-step-ahead variance forecast, $\hat{\sigma}_{t+1}^2$, for day *t*+1.

Next, we describe the forecasting models based on the realized measures. As the realized measures estimate the variance for the open-to-close period, we scale them to obtain the estimate of close-to-close variance. The close-to-close variance for day *t* is estimated as $\sigma_t^2 = \eta x_t$, where η is the scaling factor and $x_t = \text{RV}_t, \text{RVS}_t, \text{BV}_t, \text{BVS}_t$ or RK_t.The scaling factor η is calculated as

$$\eta = \frac{T^{-1}\sum_{t=1}^{T}(r_t - \mu_{cc})^2}{T^{-1}\sum_{t=1}^{T}\left(r_{oc,t} - \mu_{oc}\right)^2} \tag{9}$$

[3] In a comparison of 330 ARCH-type models, Hansen & Lunde (2005) found that a model with higher lags for the ARCH or GARCH terms, rarely performs better than the same model with fewer lags (out-of-sample). Based on their findings, we restrict our GARCH models to the simplest lag specifications.

where *T* is the total number of days in the sample period, $r_{oc,t}$ is the open-to-close logarithmic return for day t, $\mu_{cc} = T^{-1}\sum_{t=1}^{T} r_t$, and $\mu_{oc} = T^{-1}\sum_{t=1}^{T} r_{oc,t}$. Martens (2002), Koopman et al. (2005), Jacob & Vipul (2008) and Garg & Vipul (2014) apply similar scaling factors to obtain the variance for the whole day (close-to-close). The RW, EW and MA models are specified as

$$\text{RW: } \hat{\sigma}_{t+1}^2 = \sigma_t^2 \tag{10}$$

$$\text{EW: } \hat{\sigma}_{t+1}^2 = \lambda\sigma_t^2 + (1-\lambda)\hat{\sigma}_t^2 \tag{11}$$

$$\text{MA: } \hat{\sigma}_{t+1}^2 = p^{-1}\sum_{i=1}^{p} \sigma_{t+1-i}^2 \tag{12}$$

The EW model is estimated using a rolling window of the most recent 2000 daily observations, as for the GARCH models. The optimal decay parameter λ is estimated by minimizing the sum of the squared forecast errors. The approach follows the recommendation of Gardner (1985), and is widely employed for estimating the decay rate parameter for the EW model (Brailsford & Faff, 1996; Dimson & Marsh, 1990; Garg & Vipul, 2014). The MA forecasts are generated with a moving average of *p* days, where *p* = 5, 10, 15 and 20. As the MA model gives equal weightage to all the past data, selecting higher values of *p* is undesirable. This is because the volatility of financial returns exhibits clustering, and therefore, most recent data are more informative for the future volatility. This intuition is confirmed by a comparison across various MA models. In terms of out-of-sample performance, we find that the MA model with five lags (*p* = 5) generally outperforms the MA models with higher lags (*p* = 10, 15 and 20), across all the realized measures (detailed results available on request). For this reason, we include only the MA model with five lags in the subsequent analysis.

2.3 Forecast Evaluation

We use the scaled RK measure as a proxy for the latent conditional variance. Therefore, the true (benchmark) conditional variance for day *t* is $\sigma_t^2 = \eta \text{RK}_t$. The choice is based on the theoretical strengths of the RK measure; it is robust to microstructure noise and makes efficient use of the intraday data. Moreover, in a comparison of nineteen realized variance measures, Gatheral & Oomen (2010) found that the RK is one of the best estimators in terms of efficiency and robustness to time varying parameters.

The standard approach for evaluating the accuracy of volatility forecasts is to utilize statistical loss functions. As the true volatility is latent, the estimation error in the volatility proxy may distort the ranking of competing volatility forecasts. Patton (2011) examined a class of loss functions for their robustness to the estimation error in the volatility proxy. Comparing nine widely used loss functions, he demonstrated that only the mean squared error (MSE) and quasi-likelihood (QLIKE) loss functions are robust to an imperfection in the volatility proxy. We utilize these two loss criteria in this study. MSE is a loss criterion that penalizes the forecasting errors in a symmetrical manner. If the forecasts are unbiased, the ranking of the forecasting models, based on MSE, is the same as that based on the $\boldsymbol{R^2}$ of the Mincer-Zarnowitz regressions. QLIKE is an asymmetric loss function that penalizes the under-prediction more heavily than the over-prediction. It is more suitable for the applications like risk management and VaR forecasting, where an under-prediction of volatility can be more costly than an over-prediction. QLIKE is the implicit loss function in Gaussian likelihood, and therefore, is consistent with the likelihood based inference drawn by Hansen et al. (2012). MSE and QLIKE are defined as

$$\text{MSE} \equiv E\left(L_{1,k,t}\right),\ \ where,\ L_{1,k,t} = (\sigma_t^2 - \hat{\sigma}_t^2) \tag{13}$$

$$\text{QLIKE} \equiv E\left(L_{2,k,t}\right),\ \ where,\ L_{2,k,t} = (\log(\hat{\sigma}_t^2) + \sigma_t^2\,\hat{\sigma}_t^{-2}) \tag{14}$$

Here, $L_{1,k,t}$ and $L_{2,k,t}$ are the losses for the forecasting model *k*, with the MSE and QLIKE loss functions, respectively.

Next, we use the Diebold-Mariano test (Diebold & Mariano, 1995) for comparing the predictive accuracy of forecasting models. Suppose the losses for the forecasting models *i* and *j* are given by $L_{w,i,t}$ and $L_{w,j,t}$ (where *w* = 1 or 2). The Diebold-Mariano test verifies the null hypothesis that $E(L_{w,i,t}) = E(L_{w,j,t})$. The test statistic is based on the loss differential $d_{w,t} = L_{w,i,t} - L_{w,j,t}$. The null hypothesis of equal predictive accuracy is then

$$H_0: E(d_{w,t}) = 0 \tag{15}$$

The Diebold-Mariano test statistic is

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{Var(\bar{d})}}$$

where $\bar{d} = N^{-1}\sum_{j=1}^{N} d_{w,t+j}$, and N is the total number of forecasts. The variance of $\bar{d}$, $Var(\bar{d})$, is estimated by the heteroskedasticity and autocorrelation consistent (HAC) estimator of Newey & West (1987). Diebold & Mariano (1995) show that under the null of equal predictive accuracy $\mathrm{DM} \sim N(0,1)$.

3. DATA

We use the daily data on returns and realized variance for 16 stock indices. The sample period extends from 1 January 2000 to 12 September 2014. The data are sourced from the Oxford–Man Institute's Realized Library (Heber, Lunde, Shephard, & Sheppard, 2009). Table 1 provides the list of sample indices and their descriptive statistics. As the number of trading days varies across different exchanges, the total number of daily observations *T*, in Table 1, differs across the sample indices. For each index, we generate *N* variance forecasts, where *N* = *T*- 2000. Hereafter, the sample indices are referred to as I_1 to I_{16}.

4. EMPIRICAL RESULTS

Table 2 provides a comparison of the out-of-sample performance of forecasting models, based on the QLIKE criterion. Each row corresponds to a particular model, and the scores show the percentage of models (out of 22) that performed worse than that model. Thus, the worst, the median, and the best models score 0, 50, and 100, respectively. The last column reports the mean of these scores. Among the standard GARCH models, the EGARCH model (Mean score = 35.7) performs better than the GARCH model (Mean score = 27.7). This result is similar to that of the earlier studies, which found that the nonlinear models, which accommodate the leverage effect, provide better forecasts than the GARCH model (Bali & Demirtas, 2008; Peter R. Hansen & Lunde, 2005; Pagan & Schwert, 1990). Generally, the models based on realized measures

Table 1 Descriptive statistics

#	Ticker	Index	Country	T	r_t		r_t^2	
					μ (%)	σ (%)	μ (%)	σ (%)
I_1	FTSE	FTSE 100	United Kingdom	3687	0.000	1.193	1.422	4.175
I_2	N225	Nikkei 225	Japan	3558	-0.005	1.567	2.456	7.044
I_3	GDAXI	DAX	Germany	3721	0.010	1.538	2.364	6.635
I_4	AORD	All Ordinaries	Australia	3666	0.016	0.952	0.906	2.609
I_5	DJI	DJIA	United States	3669	0.011	1.198	1.434	4.476
I_6	FCHI	CAC 40	Canada	3740	-0.008	1.481	2.194	5.655
I_7	KS11	KOSPI	South Korea	3615	0.018	1.658	2.749	7.694
I_8	AEX	AEX	Netherlands	3739	-0.013	1.469	2.158	6.167
I_9	SSMI	SMI	Switzerland	3676	0.005	1.217	1.482	4.463
I_{10}	IBEX	IBEX 35	Spain	3705	-0.002	1.500	2.250	5.813
I_{11}	NSEI	S&P CNX Nifty	India	3109	0.052	1.745	3.046	15.756
I_{12}	MXX	IPC	Mexico	3673	0.051	1.389	1.933	5.058
I_{13}	BVSP	Bovespa	Brazil	3588	0.035	1.858	3.451	8.964
I_{14}	STOXX50E	Euro STOXX 50	Eurozone	3717	-0.011	1.510	2.280	5.870
I_{15}	FTSTI	FT Straits Times	Singapore	3626	0.007	1.203	1.448	6.159
I_{16}	FTMIB	FTSE MIB	Italy	3702	-0.018	1.533	2.349	6.148

Notes: This table provides the descriptive statistics for the daily returns series (r_t) and the squared daily returns series (r_t^2), for the period 1 January 2000 to 12 September 2014. T is the number of daily observations for an index. The returns are measured in percent per day.

outperform the standard GARCH models. This was expected, as the standard GARCH models have a limited information set that only includes the daily returns. Hansen et al. (2012) find that the RG model improves the out-of-sample likelihood function, when benchmarked to a standard GARCH model. Our results also indicate that the RG model outperforms the standard GARCH models on the QLIKE loss criterion. Since QLIKE is the implicit loss function in Gaussian likelihood, these results are consistent with the likelihood-based inference drawn by Hansen et al. (2012). However, the RG model does not provide the best variance forecasts. Regardless of the choice of the realized measure, the EW model outperforms the RG model. In fact, the EW models generally provide the most accurate forecasts, with EW-RK, EW-RV and EW-RVS ranking as the best models, with the mean scores of 94.6, 87.2 and 86.3, respectively. Overall, the EW ranks as the best model for the 15 out of 16 sample indices on the QLIKE criterion. The RG models dominate the RW models for all the realized measures, but they do not consistently outperform the MA models. With the RV, RVS and RK realized measures, the MA models perform better than the RG models. Therefore, the RG may not be the best model for forecasting, as it fails to outperform simpler time series models like the EW and MA models. Finally, the RW model provides the worst variance forecasts. This is not surprising, because the RW model has the most limited information set among all the models; it excludes all the historical data beyond a single day. A number of other studies also report that the RW model has poor out-of-sample performance, in the context of volatility forecasting (Brailsford & Faff, 1996; Brooks & Burke, 1998; Yu, 2002).

Table 3 provides a comparison of the out-of-sample performance of competing forecasting models, based on the MSE criterion. All RG models outperform the GARCH model; but the EGARCH model is better than the RG models. Therefore, the relative forecasting performance of the RG and EGARCH models is sensitive to the choice of loss criterion. With the asymmetric QLIKE loss function, the RG models are preferable to the EGARCH model. This suggests that the RG models are less prone to under-prediction of volatility than the EGARCH model. Nevertheless, with the symmetric MSE loss function, the RG model does not

Table 2 Relative performance ranking based on the QLIKE criterion

Model	1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	Mean
GARCH	71.4	9.5	9.5	19.0	23.8	42.9	23.8	4.8	23.8	0.0	0.0	95.2	28.6	47.6	38.1	4.8	27.7
EGARCH	**100.0**	4.8	14.3	47.6	33.3	52.4	28.6	0.0	38.1	4.8	38.1	81.0	23.8	38.1	66.7	0.0	35.7
RG-RV	28.6	38.1	57.1	81.0	28.6	76.2	81.0	66.7	71.4	76.2	71.4	23.8	4.8	52.4	47.6	28.6	52.1
RG-RVS	23.8	47.6	66.7	85.7	42.9	90.5	85.7	90.5	81.0	81.0	81.0	33.3	14.3	57.1	33.3	23.8	58.6
RG-BV	42.9	33.3	61.9	95.2	38.1	95.2	66.7	76.2	66.7	61.9	76.2	38.1	9.5	66.7	28.6	14.3	54.5
RG-BVS	38.1	42.9	71.4	**100.0**	47.6	**100.0**	71.4	71.4	76.2	85.7	85.7	47.6	19.0	76.2	23.8	9.5	60.4
RG-RK	33.3	52.4	52.4	90.5	52.4	85.7	76.2	81.0	85.7	47.6	66.7	28.6	0.0	61.9	42.9	19.0	54.8
EW-RV	81.0	81.0	95.2	71.4	95.2	71.4	**100.0**	**100.0**	90.5	90.5	**100.0**	85.7	57.1	95.2	90.5	90.5	87.2
EW-RVS	76.2	90.5	90.5	61.9	90.5	57.1	90.5	95.2	95.2	95.2	90.5	71.4	95.2	85.7	**100.0**	95.2	86.3
EW-BV	95.2	85.7	47.6	33.3	81.0	38.1	57.1	61.9	57.1	66.7	61.9	52.4	76.2	42.9	71.4	76.2	62.8
EW-BVS	90.5	95.2	33.3	23.8	61.9	23.8	61.9	57.1	61.9	71.4	57.1	61.9	61.9	23.8	61.9	71.4	57.4
EW-RK	85.7	**100.0**	**100.0**	76.2	**100.0**	81.0	95.2	85.7	**100.0**	**100.0**	95.2	**100.0**	**100.0**	**100.0**	95.2	**100.0**	94.6
MA-RV	52.4	66.7	81.0	52.4	76.2	61.9	52.4	52.4	42.9	57.1	52.4	76.2	66.7	81.0	81.0	66.7	63.7
MA-RVS	47.6	71.4	76.2	57.1	71.4	47.6	42.9	47.6	47.6	42.9	42.9	66.7	90.5	71.4	76.2	81.0	61.3
MA-BV	66.7	57.1	28.6	38.1	66.7	28.6	38.1	23.8	28.6	33.3	33.3	42.9	81.0	33.3	57.1	57.1	44.6
MA-BVS	61.9	61.9	19.0	42.9	57.1	9.5	33.3	19.0	33.3	38.1	28.6	57.1	71.4	9.5	52.4	61.9	41.1
MA-RK	57.1	76.2	85.7	66.7	85.7	66.7	47.6	38.1	52.4	52.4	47.6	90.5	85.7	90.5	85.7	85.7	69.6
RW-RV	4.8	23.8	42.9	9.5	9.5	19.0	9.5	42.9	9.5	19.0	23.8	9.5	42.9	19.0	19.0	42.9	21.7
RW-RVS	0.0	28.6	23.8	14.3	19.0	14.3	19.0	33.3	19.0	23.8	19.0	14.3	47.6	14.3	14.3	52.4	22.3
RW-BV	19.0	0.0	0.0	0.0	0.0	4.8	0.0	9.5	0.0	9.5	4.8	0.0	33.3	4.8	4.8	33.3	7.7
RW-BVS	14.3	14.3	4.8	4.8	4.8	0.0	4.8	14.3	4.8	14.3	9.5	4.8	38.1	0.0	0.0	38.1	10.7
RW-RK	9.5	19.0	38.1	28.6	14.3	33.3	14.3	28.6	14.3	28.6	14.3	19.0	52.4	28.6	9.5	47.6	25.0

Notes: This table provides the relative performance ranking of the forecasting models, based on the QLIKE criterion. Each row corresponds to a particular model, and the scores show the percentage of models (out of 22) that performed worse than that particular model. Thus, the worst, the median, and the best models score 0, 50, and 100. The last column is the average of these 16 scores.
For each index, the best forecasting model is shown in bold text.

Table 3 Relative performance ranking based on the MSE criterion

Model	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	Mean
GARCH	0.0	0.0	9.5	9.5	19.0	23.8	71.4	23.8	0.0	0.0	95.2	9.5	19.0	33.3	0.0	0.0	19.6
EGARCH	81.0	71.4	38.1	81.0	66.7	95.2	76.2	**100.0**	**100.0**	52.4	**100.0**	38.1	9.5	95.2	14.3	66.7	67.9
RG-RV	52.4	19.0	42.9	19.0	61.9	57.1	52.4	38.1	19.0	61.9	38.1	14.3	14.3	57.1	19.0	19.0	36.6
RG-RVS	66.7	42.9	66.7	14.3	85.7	76.2	33.3	14.3	52.4	57.1	47.6	19.0	38.1	66.7	28.6	28.6	46.1
RG-BV	57.1	47.6	57.1	42.9	**100.0**	52.4	61.9	42.9	23.8	71.4	66.7	28.6	33.3	61.9	9.5	33.3	49.4
RG-BVS	61.9	61.9	71.4	28.6	95.2	81.0	23.8	19.0	47.6	76.2	61.9	23.8	52.4	85.7	23.8	38.1	53.3
RG-RK	71.4	38.1	47.6	23.8	28.6	66.7	66.7	47.6	14.3	66.7	42.9	0.0	23.8	71.4	4.8	76.2	43.2
EW-RV	85.7	95.2	52.4	85.7	57.1	61.9	90.5	85.7	76.2	81.0	71.4	42.9	90.5	47.6	90.5	81.0	74.7
EW-RVS	90.5	**100.0**	95.2	**100.0**	81.0	**100.0**	95.2	95.2	90.5	90.5	14.3	90.5	**100.0**	90.5	95.2	85.7	88.4
EW-BV	76.2	85.7	90.5	90.5	90.5	71.4	85.7	76.2	81.0	95.2	90.5	95.2	66.7	76.2	81.0	90.5	83.9
EW-BVS	95.2	81.0	**100.0**	71.4	71.4	90.5	81.0	90.5	85.7	**100.0**	19.0	**100.0**	76.2	**100.0**	85.7	95.2	83.9
EW-RK	**100.0**	90.5	61.9	95.2	76.2	85.7	**100.0**	81.0	95.2	85.7	81.0	71.4	95.2	81.0	**100.0**	**100.0**	87.5
MA-RV	28.6	28.6	4.8	57.1	23.8	28.6	57.1	61.9	38.1	28.6	52.4	33.3	71.4	14.3	61.9	23.8	38.4
MA-RVS	33.3	23.8	23.8	76.2	42.9	42.9	47.6	71.4	66.7	33.3	9.5	66.7	85.7	42.9	71.4	42.9	48.8
MA-BV	42.9	14.3	14.3	66.7	52.4	38.1	38.1	52.4	57.1	38.1	85.7	85.7	61.9	28.6	38.1	52.4	47.9
MA-BVS	47.6	4.8	19.0	52.4	47.6	47.6	14.3	66.7	61.9	42.9	23.8	81.0	57.1	52.4	47.6	57.1	45.2
MA-RK	38.1	9.5	0.0	61.9	38.1	33.3	42.9	57.1	71.4	23.8	76.2	57.1	81.0	23.8	57.1	71.4	46.4
RW-RV	4.8	66.7	28.6	0.0	0.0	0.0	28.6	4.8	4.8	4.8	28.6	4.8	4.8	0.0	76.2	4.8	16.4
RW-RVS	14.3	76.2	85.7	38.1	9.5	4.8	9.5	28.6	33.3	14.3	0.0	52.4	28.6	9.5	42.9	9.5	28.6
RW-BV	19.0	52.4	76.2	33.3	33.3	14.3	19.0	33.3	9.5	19.0	57.1	61.9	42.9	19.0	52.4	47.6	36.9
RW-BVS	23.8	57.1	81.0	47.6	14.3	19.0	0.0	9.5	28.6	47.6	4.8	76.2	47.6	38.1	66.7	61.9	39.0
RW-RK	9.5	33.3	33.3	4.8	4.8	9.5	4.8	0.0	42.9	9.5	33.3	47.6	0.0	4.8	33.3	14.3	17.9

Notes: This table provides the relative performance ranking of the forecasting models, based on the MSE criterion. Each row corresponds to a particular model, and the scores show the percentage of models (out of 22) that performed worse than that particular model. Thus, the worst, the median, and the best models score 0, 50, and 100, respectively. The last column is the average of the 16 scores. For each index, the best forecasting model is shown in bold text.

improve upon the EGARCH model. More importantly, the EW model outperforms the RG model, for all the realized measures. Even with MSE loss function, the EW models provide the best forecasts, dominating the standard GARCH models, and the other models based on realized measures. The EW-RVS and EW-RK models are the best performing models, with the mean scores of 88.4 and 87.5, respectively. Overall, the EW ranks as the best model for the 12 out of 16 sample indices on the MSE criterion. With any realized measure, the RG model provides better forecasts than the RW model. However, the RG model underperforms the MA model with the RV, RVS and RK realized measures. This confirms our earlier observation that the RG model does not provide better out-of-sample forecasts than the EW and MA models.

It is widely documented that the volatility process exhibits clustering, with periods of high and low volatility (Mandelbrot, 1963). For each index, we compare the out-of-sample performance of the forecasting models for three different volatility levels: the high, the moderate, and the low. For each stock index, we divide the out-of-sample period (*N* daily observations) into five contiguous subperiods of equal size[4]. These subperiods are sorted in ascending order, based on the volatility of the index in that particular subperiod. The first and fifth subperiods are classified as the 'low' and 'high' volatility periods. The third subperiod is classified as the 'moderate' volatility period. The remaining two subperiods are ignored. Table 4 provides a comparison of different models for the entire out-of-sample period, and the three subperiods with different levels of volatility. For each period, the forecasting models are ranked from one (the best model) to twenty two (the worst model), for the sixteen stock indices separately. Table 4 provides the mean of these ranks (for the 16 indices) for each model. Overall, we find that the EW models provide the best forecasts. Regardless of the choice of the realized measure, the EW

[4] For example, the FTSE 100 index has 1687 observations in the out-of-sample period. We divide them into five contiguous subperiods, with 337 trading days in each subperiod. The remaining two days are ignored.

Table 4 Forecasting comparison across subperiods with different levels of volatility

Model	Loss Function: MSE				Loss Function: QLIKE			
	Full Period	High Vol. Period	Moderate Vol. Period	Low Vol. Period	Full Period	High Vol. Period	Moderate Vol. Period	Low Vol. Period
GARCH	17.875	17.125	18.438	16.813	17.125	17.500	16.688	12.625
EGARCH	7.750	6.813	14.750	17.938	15.625	10.938	16.063	13.000
RG-RV	14.313	14.625	11.813	15.688	10.750	11.188	9.938	10.938
RG-RVS	12.313	12.000	12.313	12.438	9.375	8.938	9.938	9.188
RG-BV	11.625	11.063	10.500	15.438	10.688	8.813	9.375	10.938
RG-BVS	10.813	10.000	11.000	13.563	9.250	7.375	9.125	9.813
RG-RK	12.938	12.813	11.813	13.563	10.125	10.938	9.688	9.250
EW-RV	6.313	7.563	6.188	7.313	3.563	4.750	4.000	6.313
EW-RVS	**3.438**	3.500	4.125	**3.875**	3.625	3.813	4.063	7.250
EW-BV	4.375	5.438	**4.000**	7.500	9.000	9.125	9.375	11.688
EW-BVS	4.375	**3.438**	5.438	4.250	10.000	8.875	10.125	11.813
EW-RK	3.625	4.063	6.063	**3.875**	**1.938**	**2.500**	**3.250**	**6.313**
MA-RV	13.938	13.625	14.063	10.688	8.438	11.188	9.313	8.000
MA-RVS	11.750	11.688	11.875	6.625	8.813	11.938	9.563	8.500
MA-BV	11.938	12.688	10.813	8.625	13.000	15.875	14.000	12.625
MA-BVS	12.500	12.063	11.188	7.438	13.438	16.875	14.688	12.438
MA-RK	12.250	13.000	13.625	7.125	7.125	9.875	8.188	8.313
RW-RV	18.563	18.938	17.875	19.000	17.250	15.000	16.188	15.375
RW-RVS	16.000	15.625	15.625	15.063	16.938	14.500	15.750	15.875
RW-BV	14.250	15.000	11.938	16.750	20.563	20.125	19.500	18.750
RW-BVS	13.813	13.313	13.000	13.125	19.813	19.438	18.688	18.625
RW-RK	18.250	18.625	16.563	16.313	16.563	13.438	15.500	15.375

Notes: This table compares the out-of-sample performance of the forecasting models for the full out-of-sample period (Full period), as well as for the high volatility period (High Vol. Period), moderate volatility period (Moderate Vol. Period), and low volatility period (Low Vol. Period). In each period, the forecasting models are ranked from one (the best model) to twenty two (the worst model), for the sixteen stock indices. The scores provided in this table are the mean of these ranks, for each model. For each period, the best forecasting model is shown in bold.

models outperform the RG models in all the subperiods. However, the RG models provide better forecasts than those of the GARCH and RW models. The forecasting performance of the RG models and the MA models are quite similar across all the subperiods. Finally, the evidence regarding the relative performance of the RG and EGARCH models is inconclusive. With the QLIKE loss function, the RG models outperform the EGARCH models. With the MSE loss function, the EGARCH models outperform (underperform) the RG models for the full and high (moderate and low) volatility periods.

Next, we provide the results of the Diebold-Mariano tests. For brevity, we exclude the RV based models, since they are usually inferior to the RVS based models. Table 5 reports the statistics for the Diebold Mariano tests, where the RG model is benchmarked to the EGARCH model, as it provides better forecasts than the GARCH model. All the statistics that are significant at the 5% level are marked with an asterisk. The sign of the test statistic indicates the relative performance of the models; a negative sign indicates that the EGARCH model performs better than the competing RG model, whereas a positive sign indicates the opposite. The last two rows of the table summarize the results. The summary metric B>X (B<X) is the number of indices for which the benchmark model is significantly better (worse) than the competing model. With the MSE loss function, the EGARCH model is significantly better than the RG models for five indices, whereas the RG models never provide significantly better forecasts than the EGARCH model. With the QLIKE loss function, the EGARCH model is significantly better (worse) than the RG models for two (eleven) indices. Clearly, the relative forecasting performance of the EGARCH and RG models is dependent on the choice of the loss criterion.

Table 6 compares the performance of the realized-measure-based forecasts, using the QLIKE loss function. It reports the statistics for the Diebold Mariano tests, where the EW, MA and RW models are benchmarked to the RG model, using various realized measures. We compare the forecasting models, based on the summary metrics (B<X and B>X). With the RVS and RK measures, the EW model provides significantly better (worse) forecasts than the RG model for eight (two) and nine (zero) indices, respectively. The performance of the MA and RG models is quiet similar. However, the RG model usually performs better than the RW model.

With the BV and BVS measures, the RG model performs better than the MA and RW models, but the EW model is still marginally better than the RG model.

Table 5 Comparison of EGARCH and Realized GARCH forecasts

Index	Benchmark: EGARCH, Loss: MSE				Benchmark: EGARCH, Loss: QLIKE			
	RG-RVS	RG-BV	RG-BVS	RG-RK	RG-RVS	RG-BV	RG-BVS	RG-RK
I_{1}	-0.06	-0.40	-0.29	-0.07	10.60*	9.99*	10.23*	12.23*
I_{2}	-1.15	-1.14	-0.90	-1.24	6.80*	6.30*	6.85*	5.78*
I_{3}	0.70	0.56	0.75	0.30	3.30*	3.84*	3.78*	3.54*
I_{4}	-3.40*	-3.24*	-3.67*	-3.18*	1.19	0.22	1.27	1.68
I_{5}	0.49	1.04	0.80	-0.74	7.77*	6.80*	8.16*	5.81*
I_{6}	-0.43	-1.34	-0.49	-0.74	5.03*	4.83*	4.72*	5.00*
I_{7}	-0.51	-0.40	-0.59	-0.27	11.44*	11.06*	11.12*	11.08*
I_{8}	-2.78*	-3.31*	-2.65*	-2.85*	3.52*	2.99*	3.34*	3.71*
I_{9}	-2.79*	-2.84*	-2.90*	-3.44*	8.81*	8.35*	8.99*	7.66*
I_{10}	0.31	0.57	0.62	0.45	2.28	2.07	2.49	1.95
I_{11}	-5.21*	-4.22*	-4.44*	-5.72*	-4.50*	-3.56*	-3.16*	-3.91*
I_{12}	-8.95*	-7.30*	-7.36*	-10.28*	-14.46*	-14.94*	-13.11*	-21.16*
I_{13}	2.03	1.37	2.13	1.20	4.98*	5.75*	5.72*	5.66*
I_{14}	-0.65	-1.39	-0.38	-0.61	-0.97	-1.31	-1.22	-0.78
I_{15}	0.77	-0.83	0.15	-1.40	14.57*	15.24*	12.71*	16.67*
I_{16}	-0.73	-0.99	-0.56	0.17	4.30*	4.00*	4.07*	5.17*
B>X	5	5	5	5	2	2	2	2
B<X	0	0	0	0	11	11	11	11

Notes: This table reports the statistics for the Diebold Mariano tests, where the EGARCH model is used as the benchmark. A negative value of the statistic indicates that the benchmark model performs better than the competing model, whereas a positive value indicates the opposite. All the statistics that are significant at the 5% level are marked with an asterisk. The metric B>X (B<X) is the number of indices for which the benchmark model performs significantly better (worse) than the competing model.

Table 6 Comparison of realized measure based forecasts on the QLIKE criterion

Index	Benchmark: RG-RVS			Benchmark: RG-BV			Benchmark: RG-BVS			Benchmark: RG-RK		
	EW-RVS	MA-RVS	RW-RVS	EW-BV	MA-BV	RW-BV	EW-BVS	MA-BVS	RW-BVS	EW-RK	MA-RK	RW-RK
I_1	9.71*	4.34*	-1.86	6.05*	3.45*	-2.20	7.35*	2.93*	-2.33	5.86*	2.27	-1.98
I_2	1.59	1.67	-1.91	-1.06	-1.99	-4.94*	-1.51	-2.45	-4.77*	2.94*	2.49	-1.43
I_3	-2.86*	-4.00*	-5.38*	-4.58*	-6.80*	-6.93*	-4.89*	-6.74*	-6.88*	-1.31	-3.79*	-4.28*
I_4	5.40*	3.45*	-5.36*	5.13*	3.07*	-6.39*	1.23	0.26	-6.34*	8.51*	5.29*	-5.96*
I_5	-5.85*	-6.44*	-8.97*	-9.69*	-8.36*	-10.67*	-10.00*	-9.32*	-12.34*	-0.82	-3.09*	-6.89*
I_6	0.01	-4.27*	-3.79*	-2.44	-5.39*	-5.19*	-2.69*	-5.42*	-4.26*	0.40	-4.29*	-3.89*
I_7	0.66	-3.52*	-3.50*	-1.75	-4.34*	-6.19*	-2.15	-4.54*	-5.15*	0.18	-3.88*	-4.65*
I_8	1.78	-3.31*	-4.81*	-1.32	-4.48*	-5.75*	-1.09	-4.61*	-5.60*	1.98	-3.57*	-5.37*
I_9	3.20*	-1.71	-4.81*	0.22	-3.23*	-6.86*	-0.73	-3.62*	-6.49*	6.59*	0.14	-2.27
I_{10}	1.63	-3.02*	-2.84*	-0.87	-4.49*	-4.16*	-1.63	-4.75*	-4.21*	2.45	-2.24	-2.36
I_{11}	3.61*	3.22*	-0.58	1.16	0.88	-2.71*	2.66*	0.85	-1.71	8.05*	5.47*	-0.45
I_{12}	14.61*	14.32*	6.37*	11.75*	13.68*	4.43*	8.91*	10.81*	4.00*	20.37*	18.07*	9.77*
I_{13}	2.88*	0.71	-4.02*	-2.97*	-3.93*	-7.66*	-4.33*	-5.09*	-7.94*	6.80*	3.57*	-3.59*
I_{14}	1.73	1.12	-1.39	1.45	0.98	-1.35	1.25	0.93	-1.38	1.49	0.98	-1.19
I_{15}	14.38*	12.03*	5.40*	10.71*	8.77*	2.25	10.87*	9.27*	4.23*	16.49*	12.92*	5.94*
I_{16}	7.37*	1.94	-1.75	3.41*	0.10	-4.23*	3.97*	0.37	-3.05*	5.35*	0.73	-2.35
B>X	2	6	9	3	8	12	4	8	11	0	5	7
B<X	8	5	2	5	4	1	5	3	2	9	5	2

Notes: This table reports the statistics for the Diebold Mariano tests, where the RG models are used as the benchmark. A negative value of the statistic indicates that the benchmark model performs better than the competing model, whereas a positive value indicates the opposite. All the statistics that are significant at the 5% level are marked with an asterisk. The metric B>X (B<X) is the number of indices for which the benchmark model performs significantly better (worse) than the competing model.

Table 7 Comparison of realized measure based forecasts on the MSE criterion

Index	Benchmark: RG-RVS			Benchmark: RG-BV			Benchmark: RG-BVS			Benchmark: RG-RK		
	EW-RVS	MA-RVS	RW-RVS	EW-BV	MA-BV	RW-BV	EW-BVS	MA-BVS	RW-BVS	EW-RK	MA-RK	RW-RK
I_{1}	1.13	-1.53	-0.99	0.69	-1.29	-1.20	2.20	-1.13	-0.81	0.67	-0.98	-0.89
I_{2}	1.94	-0.37	0.80	1.58	-0.90	0.03	1.53	-1.58	-0.10	1.48	-0.47	-0.05
I_{3}	1.69	-3.03*	0.50	1.66	-3.37*	0.26	1.77	-3.24*	0.43	0.20	-2.72*	-0.53
I_{4}	3.99*	3.55*	0.42	3.57*	2.80*	-0.43	2.95*	2.29	1.04	4.90*	4.26*	-0.87
I_{5}	-0.23	-1.37	-1.18	-0.25	-1.61	-1.16	-0.91	-1.42	-1.45	1.97	0.38	-0.85
I_{6}	0.58	-2.16	-1.32	0.57	-1.91	-1.15	0.64	-2.22	-1.14	0.39	-2.27	-1.27
I_{7}	1.17	0.08	-0.21	0.94	-0.26	-0.47	1.19	-0.06	-0.43	1.26	-0.42	-0.67
I_{8}	2.82*	0.66	0.06	2.28	0.40	-0.02	2.68*	0.53	-0.09	2.90*	0.35	-0.33
I_{9}	2.93*	0.37	-0.19	2.79*	0.56	-0.66	2.56	0.28	-0.25	3.74*	1.45	0.80
I_{10}	2.07	-2.08	-1.24	1.35	-2.43	-1.08	2.19	-1.90	-0.76	1.35	-2.66*	-1.39
I_{11}	-1.86	-2.17	-1.03	3.23*	3.36*	-0.40	-1.37	-2.00	-1.00	2.41	3.85*	-0.61
I_{12}	13.05*	11.68*	12.06*	13.84*	12.38*	11.77*	12.56*	10.44*	12.06*	11.48*	10.79*	8.14*
I_{13}	2.00	1.52	-0.29	2.50	1.89	0.41	2.03	1.29	-0.05	1.83	1.39	-0.60
I_{14}	1.23	-1.05	-1.11	1.04	-1.52	-1.10	1.14	-1.44	-1.00	0.06	-1.87	-1.20
I_{15}	2.74*	1.45	0.83	2.54	2.13	2.16	2.36	1.68	1.89	3.10*	2.63*	1.96
I_{16}	4.20*	0.50	-1.07	4.48*	1.10	0.38	4.59*	0.62	0.47	2.58*	-0.01	-0.72
B>X	0	1	0	0	1	0	0	1	0	0	2	0
B<X	6	2	1	5	3	1	4	1	1	6	4	1

Notes: This table reports the statistics for the Diebold Mariano tests, where the RG models are used as the benchmark. A negative value of the statistic indicates that the benchmark model performs better than the competing model, whereas a positive value indicates the opposite. All the statistics that are significant at the 5% level are marked with an asterisk. The metric B>X (B<X) is the number of indices for which the benchmark model performs significantly better (worse) than the competing model.

Table 7 compares the performance of the realized-measure-based forecasts, using the MSE loss function. With the MSE criterion, the RG model fails to outperform the EW model for any index, irrespective of the choice of realized measure. The forecasting performance of the RG and RW models is quiet similar, as we fail to reject the null hypothesis of equal predictability for the fifteen out of sixteen indices. The MA model also provides marginally better forecasts than the RG model. Overall, the Diebold-Mariano tests indicate that the RG model does not provide better forecasting performance than the EW and MA models. In fact, in terms of the out-of-sample performance, the EW model usually dominates the RG model on both the MSE and QLIKE criteria.

5. CONCLUSION

The RG model provides a unique framework for the joint modeling of conditional variance and realized measures of volatility. As the RG model is still relatively nascent, its forecasting performance has not been sufficiently explored. This article attempts to bridge this gap by comparing the predictive ability of the RG model with that of the GARCH and EGARCH models based on daily returns, and the EW, RW and MA models based on realized measures of volatility. The RG, EW, RW and MA models are implemented using five realized measures. Overall, our model space includes twenty-two forecasting models. The data set for this study comprises sixteen international stock indices, with a sample period ranging from 1 January 2000 to 12 September 2014.

We find that the relative forecasting performance of the EGARCH and RG models is sensitive to the choice of the loss criteria. With the QLIKE loss function, the RG model is preferable to the EGARCH model, whereas with the MSE loss function, the EGARCH model is preferable. Moreover, the RG model does not provide better forecasts than the EW and MA models. Regardless of the choice of the realized measure, the EW model outperforms the RG model, and usually ranks as the best forecasting model on the QLIKE and MSE loss criteria, across all volatility regimes. It may seem surprising that the EW models outperform the RG models, since the EW specification is nested within the more general RG specification. However, the superior forecasting performance of the EW can be explained by three reasons. First, even the most basic RG model requires an estimation of nine parameters, whereas the EW model requires the estimation of a single parameter. The estimation of a large number of parameters can often lead to considerable estimation errors that may make the model ill-suited for forecasting applications. The principle of parsimony suggests, that among a set of models that incorporate all the relevant information, the simpler forecasting models usually perform better than the more complex models. Second, as the EW estimate is not conditioned on the long-run mean variance, it is quicker to respond to the changes in the variance process as compared to the GARCH estimate (Taylor, 1986). Several other studies have indicated that the EW model can provide more accurate volatility forecasts than the GARCH models (Tse, 1991; Tse & Tung, 1992; Boudoukh, Richardson, & Whitelaw, 1997; Vipul & Jacob, 2007). Third, the RG model may not be ideal for forecasting as it attempts to combine the squared daily returns with a realized measure of volatility. It is well established that the realized measures provide more precise estimates of the latent volatility, whereas, the squared daily returns provide a noisy estimate (Andersen and Bollerslev, 1998). Therefore, the inclusion of squared daily returns is likely to introduce noise in the model, and adversely affect its out-of-sample performance. There is no informational gain by the inclusion of daily returns in the model, as the realized measures subsume the information contained in the daily returns. Our results indicate that one can generate better volatility forecasts, using simpler time series models with the realized measures. From a practical standpoint also, the EW model is easier to implement than the RG model, as it does not require a large history of data for model estimation.

REFERENCES

Andersen, T. G., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review*, *39*(4), 885–905. http://doi.org/10.2307/2527343

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. *Review of Economics and Statistics*, *89*(4), 701–720. http://doi.org/10.1162/rest.89.4.701

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, *61*(1), 43–76. http://doi.org/10.1016/S0304-405X(01)00055-1

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The Distribution of Realized Exchange Rate Volatility. *Journal of the American Statistical Association*, *96*(453), 42–55. http://doi.org/10.1198/016214501750332965

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and Forecasting Realized Volatility. *Econometrica*, *71*(2), 579–625.

Andersen, T. G., Dobrev, D., & Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, *169*(1), 75–93. http://doi.org/10.1016/j.jeconom.2012.01.011

Awartani, B. M. A., & Corradi, V. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting*, *21*(1), 167–183. http://doi.org/10.1016/j.ijforecast.2004.08.003

Balaban, E. (2004). Comparative forecasting performance of symmetric and asymmetric conditional volatility models of an exchange rate. *Economics Letters*, *83*(1), 99–105. http://doi.org/10.1016/j.econlet.2003.09.028

Bali, T. G., & Demirtas, K. O. (2008). Testing mean reversion in financial market volatility: Evidence from S&P 500 index futures. *Journal of Futures Markets*, *28*(1), 1–33. http://doi.org/10.1002/fut.20273

Bandi, F. M., & Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, *79*(3), 655–692. http://doi.org/10.1016/j.jfineco.2005.01.005

Bandi, F. M., & Russell, J. R. (2008). Microstructure Noise, Realized Variance, and Optimal Sampling. *Review of Economic Studies*, *75*(2), 339–369. http://doi.org/10.1111/j.1467-937X.2008.00474.x

Barndorff-Nielsen, O. E. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(2), 253–280. http://doi.org/10.1111/1467-9868.00336

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise. *Econometrica*, *76*(6), 1481–1536. http://doi.org/10.3982/ECTA6495

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2009). Realized kernels in practice: trades and quotes. *Econometrics Journal*, *12*(3), C1–C32. http://doi.org/10.1111/j.1368-423X.2008.00275.x

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2011). Subsampling realised kernels. *Journal of Econometrics*, *160*(1), 204–219. http://doi.org/10.1016/j.jeconom.2010.03.031

Barndorff-Nielsen, O. E., & Shephard, N. (2002). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, *17*(5), 457–477. http://doi.org/10.1002/jae.691

Barndorff-Nielsen, O. E., & Shephard, N. (2004a). *Measuring the impact of jumps in multivariate price processes using bipower covariation*. Discussion paper, Nuffield College, Oxford University. Retrieved from https://www.cass.city.ac.uk/__data/assets/pdf_file/0004/64876/BarndorffNielsen-Shephard.pdf

Barndorff-Nielsen, O. E., & Shephard, N. (2004b). Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics*, *2*(1), 1–37.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. http://doi.org/10.1016/0304-4076(86)90063-1

Boudoukh, J., Richardson, M., & Whitelaw, R. F. (1997). Investigation of a Class of Volatility Estimators. *Journal of Derivatives*, *4*(3), 63–71. http://doi.org/10.3905/jod.1997.407973

Brailsford, T. J., & Faff, R. W. (1996). An evaluation of volatility forecasting techniques. *Journal of Banking & Finance*, *20*(3), 419–438. http://doi.org/10.1016/0378-4266(95)00015-1

Brooks, C., & Burke, S. P. (1998). Forecasting exchange rate volatility using conditional variance models selected by information criteria. *Economics Letters*, *61*(3), 273–278. http://doi.org/10.1016/S0165-1765(98)00178-5

Christoffersen, P., Feunou, B., Jacobs, K., & Meddahi, N. (2014). The Economic Value of Realized Volatility: Using High-Frequency Returns for Option Valuation. *Journal of Financial and Quantitative Analysis*, *49*(03), 663–697. http://doi.org/10.1017/S0022109014000428

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. http://doi.org/10.2307/1392185

Dimson, E., & Marsh, P. (1990). Volatility forecasting without data-snooping. *Journal of Banking & Finance*, *14*(2–3), 399–421. http://doi.org/10.1016/0378-4266(90)90056-8

Engle, R. (2002). New frontiers for arch models. *Journal of Applied Econometrics*, *17*(5), 425–446. http://doi.org/10.1002/jae.683

Fleming, J., Kirby, C., & Ostdiek, B. (2003). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, *67*(3), 473–509.

Forsberg, L., & Bollerslev, T. (2002). Bridging the gap between the distribution of realized (ECU) volatility and ARCH modelling (of the Euro): the GARCH-NIG model. *Journal of Applied Econometrics*, *17*(5), 535–548. http://doi.org/10.1002/jae.685

Frommel, M., Han, X., & Kratochvil, S. (2014). Modeling the Daily Electricity Price Volatility with Realized Measures. *Energy Economics*, *44*, 492–502.

Garg, S., & Vipul. (2014). Volatility forecasting performance of two-scale realized volatility. *Applied Financial Economics*, *24*(17), 1111–1121. http://doi.org/10.1080/09603107.2014.924293

Gatheral, J., & Oomen, R. C. A. (2010). Zero-intelligence realized variance estimation. *Finance and Stochastics*, *14*(2), 249–283. http://doi.org/10.1007/s00780-009-0120-1

Gardner, E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28. http://doi.org/10.1002/for.3980040103

Hansen, P. R., Huang, Z., & Shek, H. H. (2012). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, *27*(6), 877–906. http://doi.org/10.1002/jae.1234

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, *20*(7), 873–889. http://doi.org/10.1002/jae.800

Heber, G., Lunde, A., Shephard, N., & Sheppard, K. (2009). *Oxford-Man Institute's realized library, version 0.1*. Oxford-Man Institute, University of Oxford.

Jacob, J., & Vipul. (2008). Estimation and forecasting of stock volatility with range-based estimators. *Journal of Futures Markets*, *28*(6), 561–581. http://doi.org/10.1002/fut.20321

Koopman, S. J., Jungbacker, B., & Hol, E. (2005). Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance*, *12*(3), 445–475. http://doi.org/10.1016/j.jempfin.2004.04.009

Lee, K. Y. (1991). Are the GARCH models best in out-of-sample performance? *Economics Letters*, *37*(3), 305–308. http://doi.org/10.1016/0165-1765(91)90227-C

Louzis, D. P., Xanthopoulos-Sisinis, S., & Refenes, A. P. (2013). The Role of High-Frequency Intra-daily Data, Daily Range and Implied Volatility in Multi-period Value-at-Risk Forecasting. *Journal of Forecasting*, *32*(6), 561–576. http://doi.org/10.1002/for.2249

Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *The Journal of Business*, *36*(4), 394–419.

Martens, M. (2002). Measuring and forecasting S&P 500 index-futures volatility using high-frequency data. *Journal of Futures Markets*, *22*(6), 497–518. http://doi.org/10.1002/fut.10016

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, *59*(2), 347–370. http://doi.org/10.2307/2938260

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55*(3), 703. http://doi.org/10.2307/1913610

Pagan, A. R., & Schwert, G. W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, *45*(1–2), 267–290. http://doi.org/10.1016/0304-4076(90)90101-X

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256. http://doi.org/10.1016/j.jeconom.2010.03.034

Pong, S., Shackleton, M. B., Taylor, S. J., & Xu, X. (2004). Forecasting currency volatility: A comparison of implied volatilities and AR(FI)MA models. *Journal of Banking & Finance*, *28*(10), 2541–2563. http://doi.org/10.1016/j.jbankfin.2003.10.015

Taylor, S. (1986). *Modelling financial time series*. Chichester, UK: Wiley.

Tse, Y. K. (1991). Stock returns volatility in the Tokyo stock exchange. *Japan and the World Economy*, *3*(3), 285–298. http://doi.org/10.1016/0922-1425(91)90011-Z

Tse Y. K., & Tung S. H. (1992). Forecasting Volatility in the Singapore Stock Market. *Asia Pacific Journal of Management*, *9*(1), 1–13.

Vortelinos, D. I. (2013). Portfolio analysis of intraday covariance matrix in the Greek equity market. *Research in International Business and Finance*, *27*(1), 66–79. http://doi.org/10.1016/j.ribaf.2012.06.003

Vortelinos, D. I., & Thomakos, D. D. (2012). Realized volatility and jumps in the Athens Stock Exchange. *Applied Financial Economics*, *22*(2), 97–112. http://doi.org/10.1080/09603107.2011.605751

Vipul, & Jacob, J. (2007). Forecasting performance of extreme-value volatility estimators. *Journal of Futures Markets*, *27*(11), 1085-1105. http://doi.org/10.1002/fut.20283

Watanabe, T. (2012). Quantile Forecasts of Financial Returns Using Realized Garch Models. *Japanese Economic Review*, *63*(1), 68–80. http://doi.org/10.1111/j.1468-5876.2011.00548.x

Yu, J. (2002). Forecasting volatility in the New Zealand stock market. *Applied Financial Economics*, *12*(3), 193–202. http://doi.org/10.1080/09603100110090118

Zhang, L., Mykland, P. A., & Ait-Sahalia, Y. (2005). A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data. *Journal of the American Statistical Association*, *100*(472), 1394–1411.

Zhou, B. (1996). High-Frequency Data and Volatility in Foreign-Exchange Rates. *Journal of Business & Economic Statistics*, *14*(1), 45–52. http://doi.org/10.1080/07350015.1996.10524628