The Lahore Journal of Business
1:1 (Summer 2012): pp. 37–58

The Forecasting Ability of GARCH Models for the 2003–07 Crisis: Evidence from S&P500 Index Volatility

Mahreen Mahmud*

Abstract

This article studies the ability of the GARCH family of models to accurately forecast the volatility of S&P500 stock index returns across the financial crisis that affected markets in 2003–07. We find the GJR-GARCH (1,1) model to be superior in its ability to forecast the volatility of the initial crisis period (2003–06) compared to its realized volatility, which acts as a proxy for the actual. This model is then extended to make forecasts for the crisis period. We conclude that the model's ability to forecast volatility across the crisis is not substantially affected, thus supporting the use of the GARCH family of models in forecasting volatility.

Keywords: Forecasting, volatility clustering, financial crisis.

JEL classification: G01, G12, G17.

1. Introduction

Volatility modeling and forecasting has received enormous attention in the last two decades, driven by its importance to the financial sector. Many studies have tried to obtain accurate estimates of volatility, which is a key input into the pricing of options and assets and in hedging strategies.

There are several approaches to forecasting volatility. The options-based approach extracts a volatility estimate from the price of traded options. Another approach is to look at the past prices of financial securities, i.e., historical volatility. A third approach—which we employ in this study—makes use of the generalized autoregressive conditional heteroscedastic (GARCH) family of models. These have been specifically developed to model volatility in financial time series, and the basic model's extensions are able to take into account the asymmetric effects of good and bad news on volatility.

* The author is a research and teaching fellow at the Centre for Research in Economics and Business (CREB) based at the Lahore School of Economics. She can be contacted at mahreenm@gmail.com.

Our aim is twofold. The first is to determine which of the GARCH family models performs best at the out-of-sample forecasting of stock index returns volatility during the initial sample period (2003–06). The second aim relates to the recent subprime mortgage crisis that hit the US market. We will use the selected model to make forecasts for the crisis period and assess its performance for this period relative to before.

The trigger for the recent financial crisis was a shift in how mortgages were issued in the US. The crisis, the effects of which began to show in early 2007, had a major adverse impact on banks and financial markets in the country and around the world. The Standard and Poor's (S&P) 500 composite index, the leading indicator of the US economy's performance, went down 45 percent between 2007 and 2008. It is of interest to examine how well the GARCH models are able to forecast the returns volatility of the index during this turbulent time relative to the preceding tranquil time.

We need measures of true volatility in order to assess the quality of the forecasts made for both periods. Volatility is, however, a latent variable in that it is unobserved and develops stochastically over time. While squared returns are commonly used as a proxy for true volatility, they have proved a noisy estimator. Instead, we will use realized volatility (RV), a proxy that utilizes the extra information provided by intraday returns.

The following sections are organized as follows. The existing literature is outlined in Section 2, the dataset used is described in Section 3, and our proposed methodology explained in Section 4. Section 5 presents our empirical findings and Section 6 provides concluding remarks.

2. A Review of the Literature

2.1. Modeling and Forecasting Volatility

The history of the GARCH models originates in Engle's (1982) seminal study, followed by the more popular generalization proposed by Bollerslev (1986).[1] In one of the earliest studies on the topic, Akgiray (1989) found support for the GARCH (1,1) model's ability to better forecast monthly return variances (using CRSP value-weighted indices for 1983–86) than the ARCH model, historical volatility estimates, and exponentially weighted moving averages.

[1] Bollerslev, incidentally, used the same index as we do—the S&P500—to introduce the popular GARCH specification for modeling financial time series volatility.

Poon and Granger's (2003) extensive review of the literature on financial forecasting spans 93 published and working papers, providing a detailed analysis of the various techniques used in financial forecasting and of the quality of results obtained from each. They conclude that, from within the GARCH family, asymmetric models yield superior forecasts because they factor in the more pronounced effect of a negative shock to volatility than a positive one of same magnitude. In particular, for stock index data, Brailsford and Faff (1996) find evidence in favor of the Glosten-Jagannathan-Runkle (GJR)-GARCH (1,1) model when applied to Australian data, and Engle and Ng (1993) the same for Japanese daily stock index returns (the Japanese TOPIX index) during 1980–88.

Using late 19th- and early 20th-century US data, Pagan and Schwert (1990) propose the exponential GARCH (EGARCH) to be the best fit. Kim and Kon (1994) examine data on 30 individual stocks and three stock indices in the US over 1962–90, and find that the GJR-GARCH (1,3) model performs well for stocks while the EGARCH (1,3) best models stock index volatility. Thus, overall, there is mixed evidence on which specific model is superior to the other. All evidence, however, points toward the superiority of asymmetric GARCH models for stock index returns volatility relative to their symmetric counterparts.

Taylor (2004) employs eight different stock indices from across the world—including the S&P500 for New York—to concentrate on one-step-ahead forecasting. Using weekly data for the period 1987–1995, the author finds that the GJR-GARCH model performs best when the regression analysis uses RV as a proxy for actual unobserved volatility. The GJR-GARCH (1,1) model "estimated using daily returns outperforms all five GARCH models estimated using weekly returns. The extra information supplied by the higher frequency data is clearly beneficial for the GJR-GARCH model."

Corradi and Awartani (2005) use S&P500 index daily data for the period 1990–2001 to study the forecasting ability of several GARCH models. As mentioned in Section 1, a measure of true variance is required in order to evaluate the quality of the forecasts made. Since the true variance based on the population is latent, a proxy is used—in this case, the authors adopt the conventional approach of using squared returns as a proxy for unobservable volatility process since their aim is merely to rank the models. They find that the asymmetric GARCH models are better than the GARCH (1,1), although this dominance is smaller for forecasts of longer horizons.

2.2. RV as a Proxy for True Volatility

The burgeoning literature on time-varying financial market volatility abounds with empirical studies in which competing models are evaluated and compared on the basis of their forecast performance. The variable of interest (volatility) is not directly observable, rather being inherently latent. As a consequence, any assessment of forecast precision is plagued by problems associated with its measurement. Recognition of the importance of this issue led to a number of studies conducted in the late 1990s that advocated the use of so-called RV, constructed from the summation of squared high-frequency returns, as a method for improving the volatility measure (Anderson, Bollerslev, & Meddahi, 2005).

Groundbreaking work by Anderson and Bollerslev (1997) using two series of spot exchange rates (DM-$ and ¥-$ spot exchange rates from 1 October 1987 through 30 September 1992) shows that an alternate proxy, i.e., RV, helps the GARCH model explain more than half of true volatility. The basis of RV is found in continuous time whereby the extra information contained in intraday data reduces the sampling error, yielding better estimates of true unobserved volatility. Anderson and Bollerslev use five-minute frequency data to show improved out-of-sample forecasting as opposed to when squared returns are used.

Subsequent studies, such as that by Hansen and Lunde (2006), establish that the use of squared returns worsens the predictive ability of GARCH models out of sample even when they perform extremely well within the sample. McMillan and Speight (2004) lends further support to the use of RV, concluding that GARCH models can successfully model the conditional variance of financial time series but that their forecasting ability is adversely affected when compared with a fallacious estimate of volatility. Using RV as a proxy, they use data from 17 daily exchange rate series relative to the US dollar for the period 1990–1996 to prove that GARCH models do better at forecasting volatility than the smoothing and moving-average models that had earlier been thought superior (see Figlewski, 1997).

3. Description of Dataset

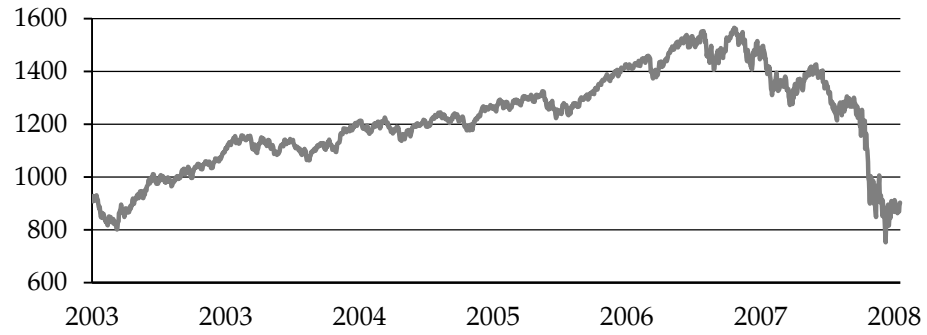
3.1. In-Sample Data

The first part of our empirical analysis is based on the S&P500 composite index for the period 2 January 2003 to 29 December 2006 (data obtained from Yahoo Finance). This in-sample period consists of 1,007 daily observations for the four-year trading period. The S&P500 is a

value-weighted index of the stocks of 500 leading industries traded on the US stock exchanges based on their market capitalization (Standard and Poor's Online) and a leading indicator of the US's equity market. The sample period is chosen as such to avoid the effects of the "dot.com" crash of the early 2000s and to end just before the latest crisis, the study of which is our objective.

The second part of the analysis extends this sample period from January 2003 to December 2007, by which point the effects of the crisis had begun to show (note the downward trend toward the end of the plotted graph in Figure 1), increasing number of observations to 1,257 for the five-year period.

Figure 1: S&P500 index (2003–07)



Source: Author's calculations.

The daily stock index value (P_t) series is nonstationary; it follows an upward trend and no mean reversion (Figure 1). This is formally confirmed by using the Dickey-Fuller test to test for the presence of a unit root, which is an indicator of the nonstationarity of the series:

$$\Delta y_t = \delta y_{t-1} + u_t \tag{1}$$

y_t is the P_t series and u_t the error term. The null hypothesis proposes that there is a unit root, $\delta = 0$. Regressing the first difference of the P_t series on its own lag yields a p-value of 0.8105 and so the null hypothesis cannot be rejected at a 5-percent confidence level, confirming that the series is non-stationary. Given that the P_t series is nonstationary, we use daily stock returns for analysis:

$$R_t = \ln (P_t/P_{t-1}) \tag{2}$$

Converting the index value series to a returns series results in a graph that reverts to its long-run mean instead of following an upward trend. The mean of the series is close to 0, and imposing a normal distribution line on its histogram shows that the R_t series is characterized by thicker tails than normal (Figure A1 in the Appendix).

The distribution is negatively skewed and, furthermore, more peaked than the normal curve, indicating excess kurtosis (Table 1). The joint skewness/kurtosis test for normality yields a p-value of 0.00, allowing us to reject the null of normality. The series also demonstrates another characteristic common to financial time series—volatility clustering, i.e., periods of well-defined high and low volatility.

Table 1: Summary statistics for R_t series

	Observations	Mean	Standard deviation	Skewness	Kurtosis	Min.	Max.
R_t	1,007	0.0004	0.0078	–0.11	4.86	–0.036	0.035

Source: Author's calculations.

3.2. Out-of-Sample Data

If true volatility is, as discussed above, latent, then in order to evaluate out-of-sample forecasts, it is important to find a proxy for it. Anderson, Bollerslev, Diebold, and Christoffersen (2006) define the R_t series as comprising an expected conditional mean return term (μ_t) and another term (ε_t) that comprises the standard deviation and an idiosyncratic error term (z_t) such that

$$R_t = \mu_{t|t-1} + \sigma_{t|t-1} z_t \tag{3}$$

The one-step-ahead volatility forecast can therefore be compared with squared returns:

$$R_t^2 = \sigma_{t|t-1}^2 z_t^2 \tag{4}$$

However the variance of z_t results in a great deal of noise when squared returns are used as the true underlying volatility. We therefore propose considering the R_t series as a continuous time process so that true volatility, referred to as integrated volatility (IV), is given by

$$IV(t) = \int_{t-1}^{t} \sigma^2(s) ds \tag{5}$$

If the returns series is sampled discretely and its variance taken for infinitesimally small periods, it will give an approximate measure of IV(t) (see Andersen, Bollerslev, Diebold, & Ebens, 2001; Andersen, Bollerslev, Diebold, & Labys, 2001, 2003). This approximate measure, RV, is not affected by the idiosyncratic error (z_t) as above, and is thus a far superior alternative that utilizes the additional information that intraday data has to offer. Hansen and Lunde's (2006) method is used to construct RV estimates as follows:

$$RV_t = \sum_{i=1}^{m} y_i^2 \tag{6}$$

y_i, $i = 1,\ldots, m$ are intraday returns, m being the number of returns in one trading day. The idea is that, with increased sampling frequency, this measure is a better approximation of true volatility ($t \to \infty$, $RV_t \to IV_t$). The important question that arises is the frequency at which the data should be sampled. At the highest frequencies, tick-by-tick returns violate the restrictions implied by the no-arbitrage assumptions in continuous-time asset pricing models. These same features also bias empirical RV measures constructed directly from ultra-high-frequency returns,[2] so in practice the measures are instead constructed from intraday returns sampled at an intermediate frequency (Anderson et al., 2005).

The S&P500 index is based on five-days-a-week trading starting at 0830 and ending at 1500. Market microstructure frictions can cause problems with very high-frequency data, so despite the availability of one-minute-interval data, we will use intermediate-frequency data sampled at five-minute intervals (our data source is TickData Online). This generates 78 observations for each day, which are then used to construct RV estimates for the two out-of-sample periods, each spanning six months beginning in January 2007 and January 2008, respectively.

4. Methodology

First, we model the volatility of the S&P500 index daily returns in order to predict their future values. Second, we compare the out-of-sample forecasting ability of the fitted models and select whichever model produces superior forecasts. The sample period is then extended up to December 2007, and the selected model re-estimated and used to produce crisis-period forecasts. These forecasts are then compared with the earlier

[2] This is due to market microstructure noise—bid-ask price spreads, jumps, and formation of patterns.

ones to establish the impact of the crisis period on their accuracy. We formally check for the presence of ARCH effects (conditional heteroscedasticity), using Engle's (1982) Lagrange Multiplier test.

The stylized facts concerning financial time series—persistence in volatility, mean-reverting behavior, and the asymmetric impact of negative- versus positive-return innovations—may significantly influence volatility. Among others, Engle and Patton (2001) illustrate these stylized facts and the GARCH models' ability—evaluated by their forecasting ability—to capture these characteristics. The sample employed in this study displays similar characteristics, thus the next logical step is to estimate the GARCH family of models:

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i} \tag{7}$$

This GARCH process does not differentiate between the impact of a positive and negative unexpected change in returns. It is therefore unable to capture the asymmetric effect of good or bad news on the volatility of the financial time series—a phenomenon termed the "leverage effect."

Anderson, Bollerslev, Diebold, and Ebens (2001) present two explanations for this so-called leverage effect. The first is that, when there is a negative shock, i.e., a negative return, it increases financial and operating leverage, which causes volatility to rise. The second is that, "if the market risk premium is an increasing function of volatility, large negative returns increase the future volatility by more than positive returns due to a volatility feedback effect." This means that the effect on volatility of unexpected bad news in the market would be higher than that of unexpected good news of the same magnitude. This renders the symmetry constraint imposed on the conditional variance equation in the GARCH process invalid. It is imperative to take account of this characteristic in order to make effective forecasts.

The presence of asymmetry in a financial time series necessitates the use of variants of the GARCH model that capture this phenomenon. Our review of the literature has established these models' goodness of fit as well as their forecasting ability. The mixed response concerning the superiority of a particular kind of model makes it difficult to allow a single choice and, so, we estimate the basic asymmetric GARCH (AGARCH) and the popular EGARCH and GJR-GARCH models.

Once the models have been fitted to the data, the next step is to generate forecasts using the estimated models. In order to select the best model, we evaluate the forecasts made by each based on the criteria detailed below. An alternative method would be to select the best model based on information criteria such as the Akaike or Bayesian, which would indicate the best in-sample fit. However, our aim is not to select the specification that best models the sample's volatility, but to select whichever makes the best forecasts out of sample. We therefore concentrate on the out-of-sample predictive capability of the models fitted and not on their in-sample fit.

4.1. Out-of-Sample Forecast Evaluation

For the purposes of forecast evaluation, we use each model to generate one-step-ahead forecasts of conditional volatility for the six-month period beginning in January 2007. Once each model has been estimated, in making each forecast we use the actual data available up to that point as an input into the equation estimated for conditional volatility by that model. The choice of the number of forecasts made is such that it ensures an adequate number of forecast observations for the analysis that is to be carried out. The quality of these forecasts is evaluated through the standard evaluation technique of employing loss-based functions and regression analysis. The chosen model is then re-estimated using the extended sample (now including 2007), and subsequently used to make predictions for the same horizon for the following year, i.e., one-step-ahead predictions for the six-month period beginning in January 2008 when the crisis hit the US equity market.

4.1.1. Regression-Based Evaluation

The first step in evaluating the quality of the forecasts made would be to regress the proxy for conditional volatility (RV_t) on the predicted volatility (h_t) from each model. This regression-based approach to evaluating out-of-sample forecasts—proposed by Mincer and Zarnowitz (1969)—has, however, been criticized in the literature. Pagan and Schwert (1990) note that, if the proxy RV_t contains large observations (outliers), problems arise when these regressions are run using ordinary least squares (OLS) because the OLS estimates are disproportionately affected by the larger values. Additionally, it "measures the level of variance errors rather than the more realistic proportional errors" thereby mainly assessing the performance of high values (Engle & Patton, 2001).

One solution to these two problems is to use the log of RV_t. Such log regressions are established as being less sensitive to the problems posed by larger observations. Thus, we run the following regression:

$$lnRV_t = \alpha + \beta * ln\hbar_t + u_t \tag{8}$$

If the forecasts (h_t) are perfect, the intercept (α) should equal 0 and the slope (β), 1. A model's superiority can be established by comparing the R^2 term—the higher the R^2 the better the forecasts explain the actual volatility

4.1.2. Loss Functions-Based Evaluation

An alternative to the regression analysis above is to assess how different the model's conditional variance predictions are from the proxy being used for the true variance. The simplest way of doing this is to calculate the mean forecast error (ME), which is

$$=\left(\frac{1}{m}\right)\sum_{t=1}^{m}(\widehat{y}_t - y_t) \tag{9}$$

The term m represents the number of forecasting observations, $\hat{y}_t$ is the predicted volatility, and y_t is the value of RV_t being used as a proxy for actual volatility. The lower the value, the better the forecast. Other, more sophisticated statistics that have been developed include a common forecast evaluation statistic, the mean squared error (MSE), which is

$$=\left(\frac{1}{m}\right)[\sum_{t=1}^{m}(\widehat{y}_t - y_t)^2] \tag{10}$$

The MSE squares the forecast errors ($\hat{y}_t + h - y_t + h$) and so penalizes larger errors more than smaller ones. Corradi and Awartani (2005) note that, since RV_t is measure-free and an unbiased estimator, it allows one to compare models in terms of loss functions other than quadratic. Thus, we can make use of the mean absolute percentage error (MAPE), which is

$$=(\frac{1}{m})[\sum_{t=1}^{m}\left|\frac{\widehat{y}_t - y_t}{y_t}\right|] \tag{11}$$

Unlike the MSE, the mean absolute error (MAE) does not penalize larger forecast errors more heavily than smaller ones. However, since it takes the absolute value, it does not allow the effect of under- and over-predictions of the same magnitude but carrying opposite signs to be cancelled out.

4.1.3. Diebold-Mariano Test

Diebold and Mariano's (DM) (1995) test allows one to compare the forecasting ability of two models. For one-step-ahead forecasts, let the forecast error ($\hat{y}_t + h - y_t + h$) be denoted by $g(e)$. The difference in loss in period i from using model 1 versus model 2 is defined as $d_i = g(e1_i - g(e2_i)$. The mean loss is given by

$$\overline{d} = 1/H \sum_{i=1}^{H} \left[g\left(e_{1i}\right) - g(e_{2i}) \right] \tag{12}$$

H is the number of forecast errors. The DM statistic is asymptotically standard normal when applied to non-nested forecasts, so that the t-test can be used to test the null hypothesis that any two fitted models have equal predictive abilities, which is when $đ = 0$.

$$DM = \overline{d} / \sqrt{\text{var}(\overline{d})} \tag{13}$$

This test is used to determine if there is any statistical difference between the forecasts generated by the chosen model before and during the crisis. Applying the test requires $var(đ)$. If the d_i series is uncorrelated, $var(đ)$ is given by $\gamma_0/(H - 1)$, else Enders' (2004) specification is followed where $var(đ) = (\gamma_0 + 2\gamma_1 + \ldots + 2\gamma_q)/H - 1$, and γ_i denotes the ith auto-covariance of d_i where the first q values of γ_i are significant.

5. Empirical Analysis

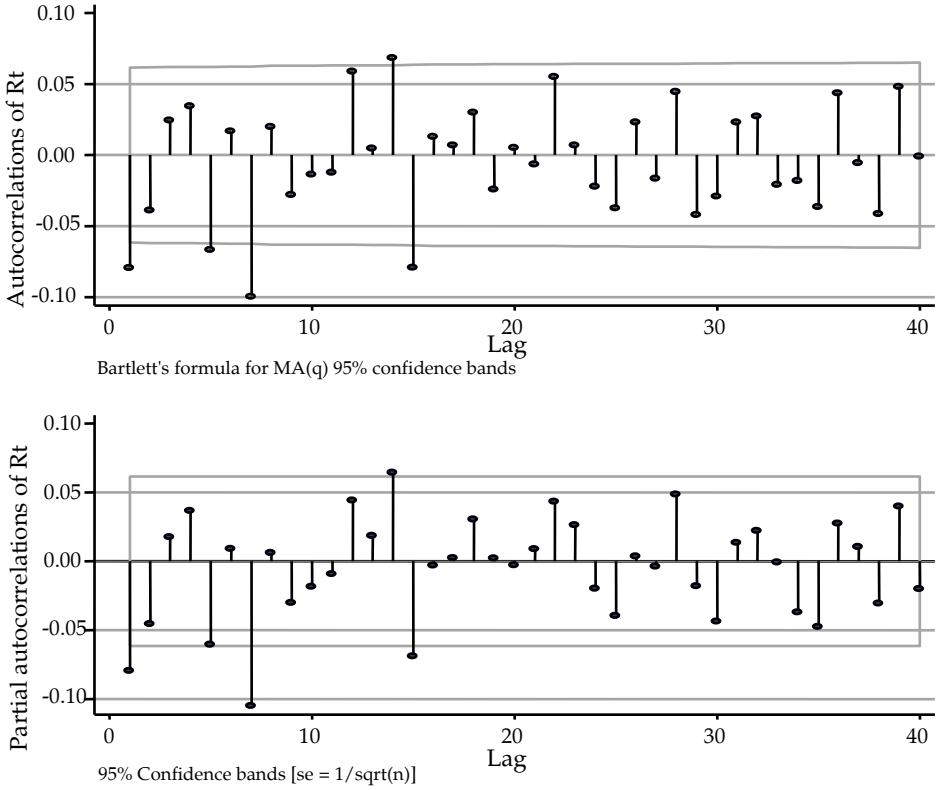
5.1. Modeling the Conditional Mean

The first step in our empirical analysis is to estimate the mean equation of returns. In order to identify the best-fitting model, the autocorrelogram (AC) and partial autocorrelogram (PAC) are plotted (Figure 2) to determine which lags are statistically significant at 5 percent ($\pm 1.96/\sqrt{T}$). Lags 1, 5, 7, 14, and 15 are found to be significant (Table A1 in the Appendix). Thus, the autoregressive moving average (ARMA) (15,15) is estimated using OLS.

Next, the estimated model's standardized residuals[3] are verified using the Portmanteau/Box-Pierce/Ljung-Box test to evaluate the adequacy of the fitted model. The null hypothesis for this test is that the errors are white noise. The p-value for the Q-statistic is 0.7126, thus the null hypothesis cannot be rejected at a 5-percent level of significance. The mean equation, therefore, fits the data.

Since it is important to check for second-order dependence in the residuals (conditional heteroscedasticity), we use Engle's (1982) Lagrange Multiplier test to check for serial correlation. This entails running OLS regression of the squares of the residuals on its own lags. The null hypothesis proposes that there is no ARCH effect, i.e., that the coefficients should all be jointly 0. The p-value yielded is 0.00, thus we can confidently reject the null hypothesis and conclude that there are, in fact, ARCH disturbances in the returns series.

Figure 2: Autocorrelogram and partial autocorrelogram of R_t series



Source: Author's calculations.

[3] Residuals divided by standard deviation.

A further check is to apply the Portmanteau test to the squared standardized residuals. The null hypothesis proposes that the errors are not serially correlated, i.e., that there are no ARCH effects. As above, a p-value of 0.000 allows us to confidently reject the null hypothesis and confirm the presence of ARCH disturbances in the data. Both tests prove that the variance is conditional on the past period, implying that we need to fit a model that can account for this effect.

5.2. Modeling the Conditional Volatility

The GARCH family of models adequately takes into account the presence of conditional heteroscedasticity, and these models are estimated in order to model volatility. Table 2 reports estimates of the maximum likelihood estimator parameters.

Table 2: Maximum likelihood estimation results

	GARCH (1,1)	EGARCH (2,1)	AGARCH (1,1)	GJR-GARCH (1,1)
α_0	7.74e-07	–0.0760587	8.48e-07	5.03e-07
	(4.39e-07)	(0.0370346)*	(3.19e-07)*	(2.52e-07)*
β_1	0.9376958	0.9922645	0.9487258	0.9562212
	(0.0179654)*	(0.0037748)*	(0.0148996)*	(0.0139044)*
α_1	0.0470485	–0.239826	0.0356606	0.0690447
	(0.0124951)*	(0.0430449)*	(0.0113298)*	(0.0150748)*
γ_1	-	–0.2274885	–0.0005019	0.0718498
		(0.0781236)*	(0.0001111)*	(0.0165285)*
α_2	-	0.1741572	-	-
		(0.0421315)*		
γ_2	-	0.2901534	-	-
		(0.079057)*		

Note: * = statistically significant at 1 percent. Standard errors are given in parentheses.
Source: Author's calculations.

We begin estimating the model using the most parsimonious specification, the GARCH (1,1). If the model is a good fit, it should be able to capture the serial correlation and no ARCH effects should remain. The p-value of the Q-statistic for the squared standardized residuals is 0.1284. Thus, we cannot reject the null hypothesis, implying that they are no remaining ARCH effects, and that the model of variance has been adequately fitted. A p-value of 0.3934 yielded by the Lagrange Multiplier test further confirms this since it allows us to accept the null hypothesis of

no remaining ARCH effects. The sum $\beta_1 + \alpha_1$ equals 0.96, which is less than 1 and satisfies the condition for stationarity. The results show that the coefficient of the lag of conditional variance, β_1 (0.94), is quite high, indicating the persistence of past effects.

As noted earlier, financial time series are characterized by the presence of leverage effects. Testing for this phenomenon entails regressing squared standardized residuals on the lags of standardized residuals, resulting in a p-value of 0.0007. This allows us to confidently reject the null hypothesis and conclude that the coefficients are not jointly equal to 0. It thus confirms the presence of leverage effects in the data. A further test involves the use of a dummy to signal any negative shocks that may have occurred in the previous period. When the squared standardized residuals are regressed on the dummy, its coefficient turns out to be significant (a p-value of 0.020). This is conclusive proof that negative shocks do, in fact, increase the conditional variance, as reported in the existing literature (see Corradi & Awartani, 2005; Taylor, 2004).

The next three models to be fitted formally account for this asymmetric effect in addition to the phenomena of volatility clustering and excess kurtosis. The first is the GJR-GARCH (1,1), which employs an indicator function that emerges when there is a negative shock in the past to account for the asymmetries. Table 2 shows that the coefficient of the indicator function γ_1 is significant and positive, implying that there are asymmetric effects. The p-value of the Q-statistic for the squared standardized residuals of this model is 0.9445. This signals that the null hypothesis cannot be rejected at a 5-percent level of significance. The GJR-GARCH (1,1) thus adequately models the second-order moment of the series.

The EGARCH (1,1) model is fitted next (Table A2 in the Appendix). The γ_1 coefficient appears to be insignificant, and the p-value yielded by the test that is applied to the model's squared standardized residuals is 0.000, implying that the null hypothesis can be rejected. Hence, the residuals are not white noise, and the model is not deemed an adequate fit. When a higher-order specification, the EGARCH (2,1), is fitted (Table 2), however, all the coefficients emerge as significant. The Portmanteau test confirms that the residuals are white noise (the p-value is 0.34). The parameter β_1 is equal to 0.99, i.e., less than 1, thus satisfying the condition for the process being stationary.

The AGARCH (1,1) model, which is fitted next, modifies the term that captures shocks that have occurred in previous periods. The estimates yielded confirm the presence of leverage effects since γ_1 is significant and negative. The standardized squared residuals are checked to ensure that no autocorrelation remains, and thus no additional lags are required. The p-value is 0.54, which does not allow us to reject the null hypothesis at a 5-percent level of significance, implying that a higher-order specification is not needed. All the coefficients are statistically significant with a β_1 that is close to 1, indicating persistent volatility.

5.3. Forecast Evaluation

Using five-minute-interval intraday return data, we construct RV_t estimates for six months that will act as a proxy for true (unobserved) volatility. This entails making one-step-ahead forecasts using each of the four models in order to evaluate the out-of-sample forecasts. The first stage of the evaluation process employs the loss function-based evaluation technique, with RV_t acting as a benchmark (Table 3). The lower the value of the criteria estimated using a particular model's forecasts, the better the forecasts. In the second stage, we carry out a regression analysis to verify the results obtained in the first stage.

Table 3: Loss function values for one-step-ahead predictions

Criterion	GARCH	EGARCH	AGARCH	GJR-GARCH
ME	–9.44887E-06	–2.106E-05	–1.15222E-05	–3.61821E-06
MSE	4.93968E-09	5.71367E-09	5.10746E-09	4.81257E-09
MAPE	0.868947666	0.626980967	0.795054713	0.983587439

Source: Author's calculations.

The value of the ME and MSE criteria is lowest in the case of the GJR-GARCH (1,1) model, with the GARCH (1,1) a close second in both. However, when the MAPE is taken into account, the EGARCH (2,1) emerges as the superior model. Thus, all three criteria indicate that the asymmetric models are superior. These results are in accordance with Corradi and Awartani (2005) who found that the asymmetric models dominated the GARCH (1,1) specification in making one-step-ahead forecasts.

After running log regressions of RV_t on the forecasted variance series[4] (see results in the Appendix), we check to see if the coefficients are statistically close to 1. The p-values of the GARCH (1,1) and AGARCH (1,1) coefficients are 0.02 and 0.00, respectively, which allows the null hypothesis to be rejected at a 5-percent level of significance, and implies that the coefficients are different from 1. However, the p-value of the GJR-GARCH (1,1) coefficient is 0.31 and that of the EGARCH (2,1) is 0.3559, which does not allow us to reject the null hypothesis, and suggests that these coefficients are not statistically different from 1. This is in synch with the results obtained from the loss function criteria, which also showed that the estimates of these two models are superior to those of the other two.

The adjusted R^2 values (Table 4) of the regression of log RV_t on the log of the forecasts show that the R^2 of the AGARCH (1,1) model is the lowest. The model's estimates appear to perform poorly on all criteria, confirming its poor predictive ability. The R^2 of both the GARCH (1,1) and GJR-GARCH (1,1) models is close to 42 percent, with the EGARCH (2,1) at 36 percent. This higher R^2 further supports the superiority of the GJR-GARCH (1,1) model over the EGARCH (2,1), which has a better MAPE measure. We can therefore proceed with the GJR-GARCH (1,1) model as the model with the best forecasting ability.

Table 4: Adjusted R^2 from regressions of log RV_t on log of predicted values

	AGARCH	GARCH	GJR-GARCH	EGARCH
R^2	0.3062	0.4255	0.4119	0.3603

Source: Author's calculations.

The next step is to extend the in-sample period from December 2006 to December 2007, and estimate the chosen model, the GJR-GRCH (1,1), based on this sample (Table A2 in the Appendix). The effect of positive news on conditional volatility is given by $\alpha_1 + \gamma_1$, the previous value of which was 0.0028 and is now 0.0057. All coefficients are still statistically significant. It is important to check if the model is an adequate fit. The p-value of the Q-statistic is 0.8276, verifying that the squared standardized residuals are white noise and that the model, therefore, fits the data.

[4] The volatility forecast series is tested for the presence of a unit root. If the series is nonstationary, the regression is spurious and yields meaningless coefficients. The null of nonstationarity is rejected at a 5-percent level of significance.

The selected model is then used to generate one-step-ahead forecasts for six months, which are compared with the out-of-sample forecasts that were made for the pre-crisis period. As before, RV_t is used as a benchmark against which to evaluate the forecasts. By January 2008, the impact of the subprime mortgage crisis had begun to show in the equity market, and the S&P 500 index had started to decline. As expected, Table 5 shows that the values of all three evaluation criteria for all horizons have increased relative to 2007.

Table 5: Loss function values for 2007 and 2008 forecasts from GJR-GARCH model

Year	ME	MSE	MAPE
2007	–3.61821E-06	4.81257E-09	0.983587
2008	–1.9694E-05	4.23284E-08	1.324632

Source: Author's calculations.

This signals worsened forecasting, which is not surprising given that crisis periods are more volatile than usual, which is why predicting becomes more difficult. However, a regression analysis of these forecasts yields an R^2 that has increased to 43 percent, lending support to the model's forecasting ability during the turbulent period.

The DM test is used to determine if there is any statistically significant difference between the model's forecasting ability in the two periods. The DM statistic yielded is 0.875, which is less than 1.96, implying that we cannot reject the null hypothesis at a 5-percent level of significance. Thus, there is no statistical difference between the model's forecasting ability in terms of period type—it is equally capable of predicting volatility for a crisis period and a normal period. The increased volatility that characterizes a crisis period is adequately accounted for. Moreover, the model fitted takes special account of leverage effects and thus effectively handles the downturn in index returns.

6. Conclusion

Based on an out-of-sample evaluation of the forecasts made by the four GARCH models, the model that best estimates daily returns volatility is the GJR-GARCH (1,1) model—in accordance with the findings of Corradi and Awartani (2005) and Taylor (2004)—when applied to the in-sample period from January 2003 to December 2008, which is characterized by relative tranquility.

Having selected the best model based on several evaluation criteria, the in-sample period is then extended up to December 2007 and the model is re-estimated. The one-step-ahead forecasts for six months obtained from this re-estimated model are compared to the prior forecasts obtained. This meets our second aim—to assess the model's ability to cope with the pronounced volatility characterizing the recent crisis that hit the US equity market. We have found that, while the model's predictive ability decreases, there is no substantial change. This supports the ability of the GJR-GARCH model in particular and of the asymmetric GARCH family of models in general to remain relatively robust across periods of pronounced volatility.

While we have used high-frequency data to construct an RV measure as a proxy for unobserved true volatility, to the presence of market microstructure noise meant that intraday returns were not aggregated at greater-than-five-minute intervals. This proxy could thus be further refined by increasing the sampling frequency and by explicitly accounting for the jumps and patterns that arise during the day when intraday data is used.

Other avenues of research could make use of the implied volatility that is extracted from options written on the index for further insight into how volatility forecasts are affected during crisis periods. Additionally, the stochastic volatility model, which has been found to be more flexible than ARCH-class models and to "fit financial market returns better and have residuals closer to standard normal" (Poon & Granger, 2003), has not been estimated here due to computational difficulties. Further research could use this model for detailed analyses analysis based on several modeling techniques.

References

Akgiray, V. (1989). Conditional heteroscedasticity in time series of stock returns: Evidence and forecasts. *Journal of Business*, *62*, 55–80.

Andersen, T. G., & Bollerslev, T. (1997). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885–905.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and correlation forecasting. In G. Elliot, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 778–787). Amsterdam: North-Holland.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of stock return volatility. *Journal of Financial Economics*, *61*, 43–76.

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*, 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*, 579–625.

Andersen, T. G., Bollerslev, T., & Meddahi, N. (2005). Correcting the errors volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica*, *73*, 279–296.

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

Brailsford, T. J., & Faff, R. W. (1996). An evaluation of volatility forecasting techniques. *Journal of Banking and Finance*, *20*, 419–438.

Corradi, V., & Awartani, B. M. A. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: The role of asymmetries. *International Journal of Forecasting*, *21*, 167–183.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–265.

Enders, W. (2004). *Applied econometrics time series* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, *48*(5), 1749–1778.

Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Journal of Banking and Finance*, *20*, 419–438.

Fair, R. C., & Shiller, R. J. (1990). Comparing information in forecasts from econometric models. *American Economic Review*, *80*(3), 375–380.

Figlewski, S. (1997). Forecasting volatility. *Financial Markets, Institutions and Instruments*, *6*(1), 1–88.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, *48*(5), 1779–1801.

Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, *131*(1–2), 97–121.

Kim, D., & Kon, S. J. (1994). Alternative models for the conditional heteroscedasticity of stock returns. *International Journal of Forecasting*, *67*(4), 563–598.

McMillan, D. G., & Speight, A. E. H. (2004). Daily volatility forecasts: Reassessing the performance of GARCH models. *Journal of Forecasting*, *23*, 449–460.

Mincer, J. A., & Zarnowitz, V. (1969). The evaluation of economic forecasts and expectations. In J. A. Mincer (Ed.), *Economic forecasts and expectations*. New York: National Bureau of Economic Research.

Nelson, D. B. (1991). Conditional heteroscedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370.

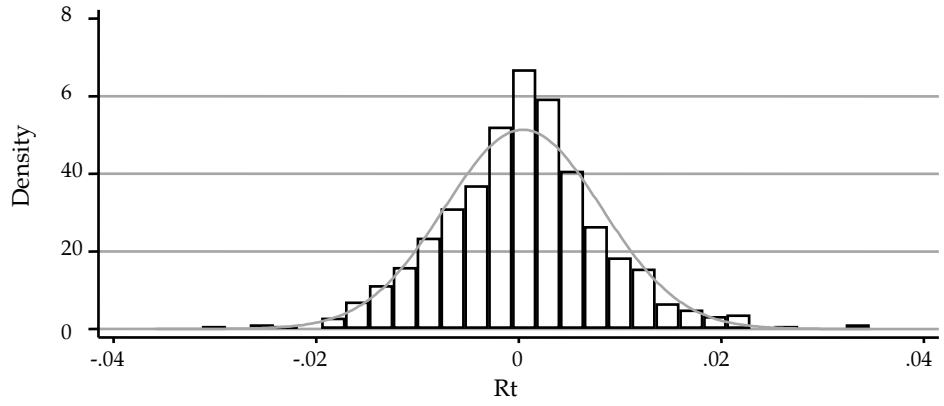
Pagan, A. R., & Schwert, G. W. (1990). Alternative models for conditional volatility. *Journal of Econometrics*, *45*, 267–290.

Poon, S.-H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, *41*(2), 478–539.

Taylor, J. W. (2004). Threshold volatility forecasting with smooth transition exponential smoothing. *International Journal of Forecasting*, *20*, 237–286.

Appendix

Figure A1: Histogram of R_t series with normal curve imposed



Source: Author's calculations.

Table A1: AC and PAC values for R_t series

LAG	AC	PAC
1	–0.0801	–0.0802
2	–0.0395	–0.0462
3	0.0241	0.0174
4	0.0349	0.0371
5	–0.0662	–0.0598
6	0.0171	0.0171
7	–0.0999	–0.1054
8	0.0200	0.0062
9	–0.0289	–0.0312
10	–0.0395	–0.0462
11	0.0241	0.0174
12	0.0600	0.0455
13	0.0059	0.0202
14	0.0683	0.0644
15	–0.0793	–0.0689
16	0.0129	–0.0034

Source: Author's calculations.

Table A2: Estimation results

Parameter	EGARCH (1,1)	GJR-GARCH (1,1)
α_0	–18.48361	1.05e-06
	(0.000)*	(0.000)*
β_1	0.9376958	0.9439015
	(0.000)*	(0.000)*
α_1	–0.0330587	0.0808454
	(0.122)	(0.000)*
γ_1	–0.0181544	–0.0865688
	(0.455)	(0.000)*

Note: * = significant at 1 percent; p-values are given in parentheses.
Source: Author's calculations.