

## A practical guide to volatility forecasting through calm and storm

**Christian Brownlees**

Department of Finance, NYU Stern, Henry Kaufman Management Center,  
44 West Fourth Street, New York, NY 10012, USA; email: [ctb@stern.nyu.edu](mailto:ctb@stern.nyu.edu)

**Robert Engle**

Department of Finance, NYU Stern, Henry Kaufman Management Center,  
44 West Fourth Street, New York, NY 10012, USA; email: [rengle@stern.nyu.edu](mailto:rengle@stern.nyu.edu)

**Bryan Kelly**

Department of Finance, Booth School of Business, University of Chicago,  
5807 South Woodlawn Avenue, Chicago, IL 60637, USA;  
email: [bryan.kelly@chicagobooth.edu](mailto:bryan.kelly@chicagobooth.edu)

*We present a volatility forecasting comparative study within the autoregressive conditional heteroskedasticity (ARCH) class of models. Our goal is to identify successful predictive models over multiple horizons and to investigate how predictive ability is influenced by choices for estimation window length, innovation distribution, and frequency of parameter reestimation. Test assets include a range of domestic and international equity indices and exchange rates. We find that model rankings are insensitive to the forecast horizon and suggestions for best practices emerge. While our main sample spans from 1990 to 2008, we take advantage of the near-record surge in volatility during the last half of 2008 to ask whether forecasting models or best practices break down during periods of turmoil. Surprisingly, we find that volatility during the 2008 crisis was well approximated by predictions made one day ahead, and should have been within risk managers' 1% confidence intervals up to one month ahead.*

### 1 INTRODUCTION

The global financial crisis of 2008 led practitioners and academics alike to reassess the adequacy of our financial models. Soaring volatilities across asset classes have

---

Detailed results on the volatility forecasting exercise performed in this paper can be found in the appendix contained in Brownlees *et al* (2011). Data and analysis used in this study are available in part at <http://vlab.stern.nyu.edu>. We would like to thank Tim Bollerslev, Kevin Sheppard, David Veredas and participants at the NYU's "Volatilities and Correlations in Stressed Markets" conference for their comments.

made it especially important to know how well our standard tools forecast volatility, particularly amid episodes of turmoil that pervade all corners of the economy. Volatility prediction is a critical task in asset valuation and risk management for investors and financial intermediaries. The price of almost every derivative security is affected by swings in volatility. Risk management models used by financial institutions and required by regulators take time-varying volatility as a key input. Poor appraisal of the risks to come can leave investors excessively exposed to market fluctuations or institutions grounded on the edge of a precipice of inadequate capital.

In this paper we explore the performance of volatility forecasting within the class of autoregressive conditional heteroskedasticity (ARCH) models. The paper examines the design features that are involved in the implementation of a real-time volatility forecasting strategy: the type of model, the amount of data to use in estimation, the frequency of estimation update, and the relevance of heavy-tailed likelihoods for volatility forecasting. We perform the exercise on a wide range of domestic and international equity indices and exchange rates. Taking advantage of the near-record surge in volatility during the last half of 2008, we ask whether our conclusions regarding forecasting models or estimation strategies change during tumultuous periods. The surprising finding that we will report is that there was no deterioration in volatility-forecast accuracy during the financial crisis, even though forecasts are purely out-of-sample. However, this result is based on one-day-ahead forecasts. Most money managers will recognize that notice given one day in advance of increasing risk is insufficient for defensive action, particularly in illiquid asset classes. While longer-horizon forecasts exhibited some deterioration during the crisis period, we will argue that they remained within a 99% confidence interval. An interpretation of this observation is that there is always a risk that the risk will change. During the crisis, risks radically changed. When portfolios are formed in a low volatility environment, ignoring variability of risks leads institutions to take on excessive leverage.

This should be interpreted not merely as a critique of volatility forecasting but, more importantly, as a critique of our most widely used risk measures: value-at-risk (VaR) and expected shortfall. These measures inherently focus on short-run risk and yet are often used to measure the risk of long-horizon and illiquid assets. Thus, research to supplement these short-term risks with a term structure of risk is an important goal. We seek to understand how volatilities can change and how to formulate better long-run forecasts.

We find that, across asset classes and volatility regimes, the simplest asymmetric generalized autoregressive conditional heteroskedasticity (GARCH) specification, the threshold GARCH model of Glosten *et al* (1993), is most often the best forecaster. How much data to use in estimation becomes an important issue if parameters are unstable, as data from the distant past can bias estimates and pollute forecasts. While our estimates reveal slowly varying movements in model parameters, results show

that using the longest possible estimation window gives the best results. However, even when using long data histories, we find that models should be reestimated at least once per week to mitigate the effects of parameter drift. Finally, despite the documented prevalence of fat-tailed financial returns even after adjusting for heteroskedasticity (Bollerslev (1987)), we find no benefit to using the heavier-tailed Student  $t$  likelihood in place of the simple Gaussian specification. This is a statement about forecasting volatility, and does not imply that tail risks should be ignored in risk management.

Our study omits volatility prediction models based on high-frequency realized volatility measures and stochastic volatility models. Important contributions in these areas include Andersen *et al* (2003), Deo *et al* (2006), Engle and Gallo (2006), Aït-Sahalia and Mancini (2008), Hansen *et al* (2010), Corsi (2010), Shephard and Sheppard (2010) and Ghysels *et al* (1995). While realized volatility models often demonstrate excellent forecasting performance, there is still much debate concerning optimal approaches. As a result, a comprehensive comparison of alternative models would be vast in scope and beyond the bounds of this paper. Our goal is to document how estimation choices affect forecast performance, especially comparing high and low volatility regimes. By analyzing the ARCH class, we present estimation best practices for the most widely applied collection of volatility forecasting models. We suspect that our main conclusions extend to time-series models of realized volatility, including our finding that short-term volatility forecasts perform well during crisis periods, that asymmetric models are superior to symmetric ones and that frequent reestimation using long samples optimizes precision while mitigating the impact of parameter drift.

This paper is related to the vast literature on volatility forecasting. Andersen *et al* (2006) provide a comprehensive theoretical overview on the topic. An extensive survey of the literature's main findings is provided in Poon and Granger (2003, 2005). Volatility forecasting assessments are commonly structured to hold the test asset and estimation strategy fixed, focusing on model choice. We take a more pragmatic approach and consider how much data should be used for estimation, how frequently a model should be reestimated and what innovation distributions should be used. This is done for a range of models. Furthermore, we do not rely on a single asset or asset class to draw our conclusions. Volatility forecasting metastudies focus almost exclusively on one-day forecasts. Our work draws attention to the relevance of multi-step forecast performance for model evaluation, especially in crisis periods when volatility levels can escalate dramatically in a matter of days. Lastly, our forecast evaluation relies on recent contributions for robust forecast assessment developed in Hansen and Lunde (2005b) and Patton (2009). Conflicting evidence reported by previous studies is due in part to the use of nonrobust losses, and our assessment addresses this shortcoming.

The rest of the paper is structured as follows. Section 2 illustrates our forecasting methodology. We detail our empirical findings in Section 3. Section 4 gives an analysis of the volatility turmoil of the fall of 2008. Concluding remarks follow in Section 5.

## 2 VOLATILITY FORECASTING METHODOLOGY

### 2.1 Recursive forecast procedure

A time series of continuously compounded returns (including dividends) is denoted by  $\{r_t\}_{t=1}^T$ , and  $\mathcal{F}_t$  denotes the information set available at  $t$ . The unobserved variance of returns conditional on  $\mathcal{F}_t$  is  $\sigma_{t+i|t}^2 \equiv \text{var}[r_{t+i} | \mathcal{F}_t]$ . Variance predictions are obtained from a set of volatility models  $\mathcal{M} \equiv \{m_1, m_2, \dots, m_M\}$ . Model  $m$  can generically be represented as:

$$r_{t+1} = \epsilon_{t+1} \sqrt{h_{t+1}^{(m)}} \quad (2.1)$$

where  $h_{t+1}^{(m)}$  is an  $\mathcal{F}_t$ -measurable function and  $\epsilon_{t+1}$  is an independent and identically distributed (iid) zero mean/unit variance innovation. The specification of  $h_{t+1}^{(m)}$  determines the conditional variance evolution and is typically a function of the history of returns as well as a vector of unknown parameters to be estimated from the data. The  $i$ -step-ahead volatility forecast obtained by model  $m$  conditional on  $\mathcal{F}_t$  is denoted by  $h_{t+i|t}^{(m)}$ .

The real-time volatility forecasting procedure is implemented as follows. For each day  $t$  in the forecasting sample, we estimate model  $m$  using data ending at or before  $t$ , depending on the frequency of parameter reestimation. We then use the fitted model to predict volatility at different horizons (one, five, ten, fifteen and twenty-two days ahead), resulting in a daily volatility forecast path  $\{h_{t+i|t}^{(m)}\}$ . This procedure generates a sequence of overlapping forecast paths, with each path formulated from different conditioning information.

The baseline estimation strategy uses all available returns (beginning with 1990) and updates parameter estimates once per week by maximizing a Gaussian likelihood. We perturb this approach to determine whether alternative estimation strategies can improve forecasting performance. In particular, we consider using four-year and eight-year rolling estimation windows, rather than a growing window that uses the full post-1990 sample. We also explore reestimating parameters daily or monthly, in addition to weekly. Finally, maximum likelihood estimation is performed using both Gaussian and Student  $t$  likelihoods. We report a subset of these results that best highlight the tradeoffs faced in estimation design. Interested readers will find exhaustive comparisons in the online appendix of this work (Brownlees *et al* (2011)).

## 2.2 Volatility models

The five models we consider for  $h_{t+1}^{(m)}$  in Equation (2.1) are chosen from the vast literature on GARCH modeling for their simplicity and demonstrated ability to forecast volatility over alternatives. The first, GARCH(1,1) (Engle (1982) and Bollerslev (1986)), is a natural starting point for model comparison due to its ubiquity and progenesis of alternative models. The volatility process is described by GARCH models as:

$$h_{t+1} = \omega + \alpha r_t^2 + \beta h_t$$

Key features of this process are its mean reversion (imposed by the restriction  $\alpha + \beta < 1$ ) and its symmetry (the magnitude of past returns, and not their sign, influences future volatility).

We also include two asymmetric GARCH models, which are designed to capture the tendency for volatilities to increase more when past returns are negative. Threshold ARCH (TARCH) (Glosten *et al* (1993)) appends a linear asymmetry adjustment:

$$h_{t+1} = \omega + (\alpha + \gamma \mathbf{1}_{\{r_t < c\}}) r_t^2 + \beta h_t$$

where  $\mathbf{1}$  is an indicator equaling one when the previous period's return is below some threshold  $c$ . The inclination of equity volatilities to rise more when past returns are negative leads to  $\gamma > 0$ .

Exponential GARCH (Nelson (1991)), or EGARCH, models the log of variance:

$$\ln(h_{t+1}) = \omega + \alpha(|\epsilon_t| - E[|\epsilon_t|]) + \gamma \epsilon_t + \beta \ln(h_t)$$

where  $\epsilon_t = r_t / \sqrt{h_t}$ . The leverage effect is manifested in EGARCH as  $\gamma < 0$ .

The nonlinear GARCH (Engle (1990)), or NGARCH, models asymmetry in the spirit of previous specifications using a different functional device. When  $\gamma < 0$  the impact of negative news is amplified relative to positive news:

$$h_{t+1} = \omega + \alpha(r_t + \gamma)^2 + \beta h_t$$

Finally, asymmetric power ARCH (APARCH), devised by Ding *et al* (1993), evolves according to:

$$h_{t+1}^{\delta/2} = \omega + \alpha(|r_t| - \gamma r_t)^{\delta} + \beta h_t^{\delta/2}$$

Raising the left-hand side to  $2/\delta$  delivers the variance series. Ding *et al* (1993) show that serial correlation of absolute returns is stronger than squared returns. Hence, the free parameter  $\delta$  can capture volatility dynamics more flexibly than other specifications, while asymmetries are incorporated via  $\gamma$ . As noted by Hentschel (1995), APARCH nests at least seven other GARCH specifications.

### 2.3 Forecast evaluation

Our measure of predictive accuracy is based on the average forecast loss achieved by a model/strategy/proxy triplet. A model that provides a smaller average loss is more accurate and therefore preferred. Choices for loss functions are extensive, and their properties vary widely. Volatility forecast comparison can be tricky because forecasted values must be compared against an *ex post* proxy of volatility, rather than its true, latent value. Patton (2009) identifies a class of loss functions that is attractively robust in the sense that they asymptotically generate the same ranking of models regardless of the proxy being used. This rank preservation holds as long as the proxy is unbiased and minimal regularity conditions are met. It ensures that model rankings achieved with proxies like squared returns or realized volatility correspond to the ranking that would be achieved if forecasts were compared against the true volatility.

The Patton class is comprised of a continuum of loss functions indexed by a parameter on the real line. It rules out all but two losses traditionally used in the volatility forecasting literature:

$$\begin{aligned}\text{QL: } L(\hat{\sigma}_t^2, h_{t|t-k}) &= \frac{\hat{\sigma}_t^2}{h_{t|t-k}} - \log \frac{\hat{\sigma}_t^2}{h_{t|t-k}} - 1 \\ \text{MSE: } L(\hat{\sigma}_t^2, h_{t|t-k}) &= (\hat{\sigma}_t^2 - h_{t|t-k})^2\end{aligned}$$

where  $\hat{\sigma}_t^2$  is an unbiased *ex post* proxy of conditional variance (such as realized volatility or squared returns) and  $h_{t|t-k}$  is a volatility forecast based on  $t - k$  information ( $k > 0$ ). The quasi-likelihood (QL) loss, named for its close relation to the Gaussian likelihood, depends only on the multiplicative forecast error,  $\hat{\sigma}_t^2 / h_{t|t-k}$ . The mean squared error (MSE) loss depends solely on the additive forecast error,  $\hat{\sigma}_t^2 - h_{t|t-k}$ . Both QL and MSE are used in our extensive forecast evaluation reported in Brownlees *et al* (2011). However, the summary results that we report here focus on QL losses. There are a few reasons why we prefer QL for forecast comparison. First, as a result of the fact that QL depends on the multiplicative forecast error, the loss series is iid under the null hypothesis that the forecasting model is correctly specified. Mean-squared error, which depends on additive errors, scales with the square of variance, and thus contains high levels of serial dependence even under the null. To see this, divide MSE by  $\hat{\sigma}_t^4$  and note that the resulting quantity is iid under the null. Mean-squared error is therefore an iid process multiplied by the square of a highly serially correlated process. While loss functions are not required to be iid in order to identify successful forecasting models, this trait makes it easier to identify when a model fails to adequately capture predictable movements in volatility. Second, suppose that the volatility proxy  $\hat{\sigma}_t^2$  can be expressed as  $\hat{\sigma}_t^2 = h_{0,t} \eta_t$ , where  $h_{0,t}$  is the latent true variance and  $\eta_t$  is a measurement error with unit expected value and variance  $\tau^2$ . The

expected value of MSE is then:

$$\begin{aligned}
 E[\text{MSE}(\hat{\sigma}_t^2, h_{t|t-k})] &= E[(\hat{\sigma}_t^2 - h_{t|t-k})^2] \\
 &= E[(\hat{\sigma}_t^2 - h_{0t} + h_{0t} - h_{t|t-k})^2] \\
 &= E[(\eta_t - 1)h_{0t} + h_{0t} - h_{t|t-k})^2] \\
 &= \text{MSE}(h_{0t}, h_{t|t-k}) + \tau^2 h_{0t}^2
 \end{aligned}$$

while the expected value of QL is:

$$\begin{aligned}
 E[\text{QL}(\hat{\sigma}_t^2, h_{t|t-k})] &= E\left[\frac{\hat{\sigma}_t^2}{h_{t|t-k}} - \log \frac{\hat{\sigma}_t^2}{h_{t|t-k}} - 1\right] \\
 &= E\left[\frac{h_{0t}}{h_{t|t-k}} \eta_t - \log \frac{h_{0t}}{h_{t|t-k}} \eta_t - 1\right] \\
 &\approx \text{QL}(h_{0t}, h_{t|t-k}) + \frac{1}{2} \tau^2
 \end{aligned}$$

where the last line uses a standard Taylor expansion for moments of a random variable. MSE has a bias that is proportional to the square of the true variance, while the bias of QL is independent of the volatility level. Amid volatility turmoil, large MSE losses will be a consequence of high volatility without necessarily corresponding to deterioration of forecasting ability. The QL avoids this ambiguity, making it easier to compare losses across volatility regimes.

### 3 EMPIRICAL VOLATILITY FORECASTING RESULTS

#### 3.1 Data

We work with daily split-adjusted and dividend-adjusted log-return data on the S&P 500 index from 1990 to 2008 from Datastream. The expanded dataset for our large-scale forecasting comparison includes three balanced panels of assets listed in Table 1 on the next page. We use ten exchange rates, nine domestic sectoral equity indices and eighteen international equity indices. The sector index data are returns on S&P 500 industry sector exchange-traded funds. International index data are returns on iShares exchange-traded funds that track the MSCI country indices. Inception dates of the sector and country index exchange-traded funds are December 23, 1998 and March 19, 1996, respectively. The exchange rates dataset contains various exchanges versus the US dollar starting on January 5, 1999.

To proxy for true S&P 500 variance, we use daily realized volatility for the S&P 500 exchange-traded fund. We construct this series from NYSE-TAQ intra-daily mid-quotes (filtered with procedures described in Brownlees and Gallo (2006) and Barndorff-Nielsen *et al* (2009)). We sample every  $d_t$ th mid-quote (tick-time sampling), where  $d_t$  is chosen such that the average sampling duration is five minutes. Let

**TABLE 1** Asset list.

Asset class	Assets	Start date
Exchange rates	Australian dollar, British pound, Canadian dollar, euro, Indian rupee, Hong Kong dollar, Japanese yen, South Korean won, Swiss franc, Thai baht	01/05/1999
Equity sectors	Consumer discretionary, consumer staples, energy, financials, health care, industrials, materials, technology, utilities	12/23/1998
International equities	Singapore, Netherlands, Japan, Australia, Belgium, Canada, Germany, Hong Kong, Italy, Switzerland, Sweden, Spain, Mexico, UK, world, emerging markets, BRIC	03/19/1996

For each asset class, the table reports the list of assets used in the forecasting application and the first date of the sample. BRIC refers to the countries of Brazil, Russia, India and China.

$p_{t,i}$  ( $i = 1, \dots, I_t$ ) denote the series of log mid-quote prices on day  $t$ . Our realized volatility proxy is the “vanilla” (Andersen *et al* (2003)) estimator constructed using sums of intra-daily squared returns:

$$\hat{\sigma}_{RVt}^2 = \sum_{i=2}^{I_t} (p_{ti} - p_{ti-1})^2$$

The overnight return is omitted, as is often done in the literature.

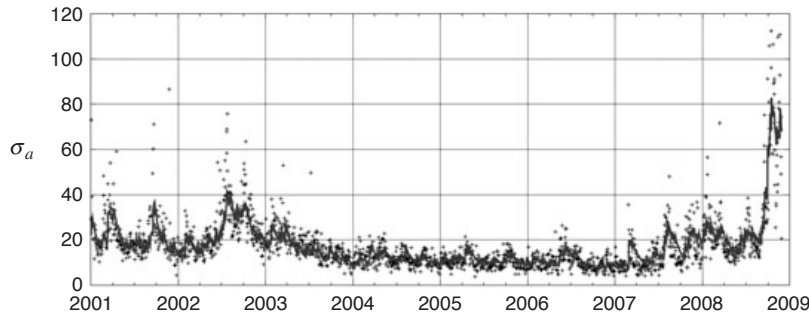
The out-of-sample forecast horizon covers 2001 to 2008 and contains periods of both very low volatility and severe distress. Figure 1 on the facing page shows the time series plot of daily realized volatility (in annualized terms) for the S&P 500 index alongside one-day-ahead predictions of a TARCH model. Equity volatility in the US reached its peak during the financial turmoil of fall 2008, with levels of realized volatilities exceeding 100%. This period is also characterized by high volatility of volatility. As of early September 2008, realized volatility was near 20%, and more than quadrupled in less than three months.

### 3.2 Forecasting S&P 500 volatility

We begin evaluating estimation strategies by assessing out-of-sample volatility forecast losses for the S&P 500 index. Table 2 on page 12 summarizes the extensive analysis provided in Brownlees *et al* (2011) and reports forecasting results for each of our five GARCH specifications using the QL loss function with squared return ( $r^2$ ) and realized volatility (RV) proxies over a range of forecast horizons. We present the



**FIGURE 1** S&P 500 TARCH one-step-ahead volatility forecasts (solid line) and realized volatility (crosses).



Volatilities are expressed in annualized terms.

base estimation strategy (using all available returns beginning in 1990 and updating parameter estimates once per week by maximizing a Gaussian likelihood). QL losses based on RV are substantially smaller than those based on  $r^2$  due to the improved efficiency of RV. However, results suggest that using a sufficiently long out-of-sample history leads to comparable findings despite the choice of the proxy. The labels beneath each loss indicate that the base strategy was significantly improved upon by modifying estimation with Student  $t$  innovations (S), medium estimation window (WM), long estimation window (WL), monthly estimation update (UM) or daily estimation update (UD). The significance of the improvement is assessed using a Diebold–Mariano predictive ability test. The test compares the forecast loss time series of the base strategy with the ones obtained by the various modifications: if the mean of the loss differential is significantly different from zero, then the null of equal predictive ability is rejected.<sup>1</sup> The appendix presents evidence of model parameter instability, highlighting the relevance of choices for amount of data used in estimation and frequency of reestimation. Our analysis suggests that the longest possible estimation window gives the best results, but we suggest reestimating at least once per week to counteract the effects of parameter drift. While there are exceptions (as expected from a comparison with a vast number of permutations), the comprehensive conclusion from this analysis is that there are no systematic large gains to be had by modifying the base procedure along the alternatives considered. A more detailed discussion of these and subsequent results is given in the appendix.

<sup>1</sup> The superior predictive ability (SPA) test or model confidence set (MCS) techniques could also be used to carry out this type of exercise (see Hansen (2005) and Hansen *et al* (2003)).

**TABLE 2** Estimation strategy assessment.

Model	Volatility proxy: $r^2$					Volatility proxy: RV				
	One day	One week	Two weeks	Three weeks	One month	One day	One week	Two weeks	Three weeks	One month
GARCH	1.460	1.481	1.520 UD	1.574	1.645 S	0.273	0.310	0.343	0.373	0.414
TARCH	1.415	1.442	1.478 S	1.547 S	1.624 S	0.243 WM	0.289 WM UD	0.328	0.368	0.415
EGARCH	1.420	1.458	1.505 S	1.592 S	1.684 S	0.234	0.282 UM	0.320 UM	0.365 UM	0.413
APARCH	1.417	1.446	1.485 S	1.557 S	1.633 S	0.249 WM S	0.299 WM UD	0.340 S WD UD	0.385 S WM	0.435 S
NGARCH	1.422	1.459	1.498 S UD	1.574 S	1.659 S	0.244 UD	0.296	0.337 UD	0.380	0.432

For each model and volatility proxy, the table reports out-of-sample QL losses at multiple horizons using the base estimation strategy. The labels underneath each loss mean that the base strategy was significantly improved upon by using Student  $t$  innovations (S), medium estimation window (WM), long estimation window (WL), monthly estimation update (UM) or daily estimation update (UD).

### 3.3 Direct comparison of GARCH models

Next, we directly compare GARCH model forecasts during the full sample and during the turmoil of fall 2008. Representative results regarding forecast accuracy across multiple horizons and volatility regimes are shown in Table 3 on the next page. Here we report out-of-sample QL losses for each volatility proxy using the TARCH(1,1) model. The appendix provides detailed results of this comparison across all models. For the full sample, asymmetric specifications provide lower out-of-sample losses, especially over one day and one week. At the one-month horizon, the difference between asymmetric and symmetric GARCH becomes insignificant as recent negative returns are less useful for predicting volatility several weeks ahead. When losses use squared return as proxy, results favor TARCH, while realized volatility selects EGARCH. The discrepancy should not be overstated, however, as the methods do not significantly outperform each other. Model rankings appear stable over various forecasting horizons.

Table 3 on the next page also shows that, during the extreme volatility interval from September 2008 through December 2008, forecast losses at all horizons are systematically larger than in the overall sample. Recall that QL is unaffected by changes in the level of volatility, so that changes in average losses purely represent differences in forecasting accuracy. One-step-ahead losses during fall 2008 are modestly higher, while, at one-month, QL losses are twice as large based on the squared return proxy and four times as large using realized volatility. The important finding from detailed cross-model comparisons in the appendix is that conclusions about model ranking remain largely unchanged during the crisis.

### 3.4 Volatility forecasting across asset classes

Table 4 on page 15 contains TARCH forecasting results for exchange rates, S&P 500 equity sector indices and international equity indices (see the detailed asset list in Table 1 on page 10). This analysis uses the QL loss with squared returns as proxy. Volatility forecast losses are averaged across time and over all assets in the same class over the full sample and the crisis subsample. Detailed cross-model comparisons from the crisis sample are provided in the appendix. We find strong evidence of volatility asymmetries in international and sectoral equity indices with TARCH as the universally dominant specification. The base GARCH model is a good descriptor of exchange rate volatility over the full sample, consistent with Hansen and Lunde (2005a).

Interestingly, asymmetric models appear to improve exchange rate volatility forecasts during the fall of 2008. This is consistent with a flight to quality during the peak of the crisis, leading to rapid appreciation of the US dollar amid accelerating exchange rate volatility. For all asset classes, one-day-ahead losses are virtually unchanged from

**TABLE 3** S&P 500 volatility prediction performance of the TARCh model from 2001 to 2008 and in fall 2008.

(a) 2001–8											
Model	Volatility proxy: $r^2$						Volatility proxy: RV				
	One day	One week	Two weeks	Three weeks	One month		One day	One week	Two weeks	Three weeks	One month
TARCh	1.415	1.442	1.478	1.547	1.624		0.243	0.289	0.328	0.368	0.415
(b) Fall 2008											
Model	Volatility proxy: $r^2$						Volatility proxy: RV				
	One day	One week	Two weeks	Three weeks	One month		One day	One week	Two weeks	Three weeks	One month
TARCh	1.461	1.560	1.985	2.311	2.875		0.304	0.353	0.590	0.672	1.380

For each volatility proxy the table reports the out-of-sample QL loss at multiple horizons for the TARCh(1,1) model.

**TABLE 4** Out-of-sample loss for various asset classes at multiple horizons using the TARCH(1,1) model for the full 2001–8 sample and fall 2008 subsample.

(a) 2001–8												
Model	Exchange rates			Equity sectors			International equities					
	One day	One week	Two weeks	Three weeks	One day	One week	One day	One week	Two weeks	Three weeks	One day	One month
TARCH	1.976	2.038	2.093	2.118	2.138	2.236	2.266	2.313	2.356	2.412	2.253	2.464
(b) Fall 2008												
Model	Exchange rates			Equity sectors			International equities					
	One day	One week	Two weeks	Three weeks	One day	One week	One day	One week	Two weeks	Three weeks	One day	One month
TARCH	1.925	2.081	2.454	2.636	3.244	2.217	2.120	2.523	2.834	3.870	2.153	4.250

those during the full sample, while one-month QL losses are magnified by a factor of nearly two. In general, results corroborate our findings for the S&P 500.

### 3.5 Interpreting forecast losses from an economic perspective

Statistically testing the differences in forecast error losses across models and methods is in itself economically meaningful because it captures how consistently one approach dominates another, which, in turn, is important in pricing and risk management. However, QL and MSE losses do not provide direct economic interpretations for the magnitudes of differences across approaches.

The relative magnitudes of forecast errors implied by the average losses from different models are useful for quantifying the economic importance of differences in forecast performance. To illustrate, consider two calibrated numerical examples. These examples translate differences in QL averages across forecasting models into:

- (1) differences in VaR forecast errors;
- (2) option-pricing errors.

We show that the relative size of forecast errors across models provides an accurate description of relative magnitudes of both VaR errors and option-pricing errors based on alternative models. We assume throughout the illustration that the true volatility of daily returns is 0.0146 (the daily volatility of the S&P 500 index over the 1990–2008 sample).

To calculate economic magnitudes, we begin by considering the typical forecast implied by our reported average forecast losses for each model. To do so, we solve the equation:

$$\frac{0.0146^2}{x^2} - \log\left(\frac{0.0146^2}{x^2}\right) - 1 = \text{average QL loss}$$

This equation is solved by two different volatility forecasts, one an underestimate and one an overestimate, that are positive and located asymmetrically around the true volatility of 0.0146.<sup>2</sup> The larger the average loss, the larger the absolute error  $|x - 0.0146|$ . Based on the reported average losses of 0.247 and 0.243 for the GARCH and TARCH models (using the RV proxy), we find  $x_{\text{GARCH, under}} = 0.0105$  and  $x_{\text{GARCH, over}} = 0.0222$ , and  $x_{\text{TARCH, under}} = 0.0107$  and  $x_{\text{TARCH, over}} = 0.0217$ . The

---

<sup>2</sup> Because QL is an asymmetric loss function, we consider the effect of volatility underestimates and overestimates separately. The two numbers reported in each comparison represent the effects of volatility underestimates and overestimates that each generate a loss equal to the appropriate average QL from Table 2.

reduction in volatility forecast errors achieved by moving from GARCH to TARCH is calculated as:

$$\text{Volatility error reduction}_i = 1 - \left| \frac{x_{\text{TARCH},i} - 0.0146}{x_{\text{GARCH},i} - 0.0146} \right|, \quad i \in \{\text{under, over}\}$$

Based on our estimated average losses, TARCH improves over GARCH 4.3% to 7.7% in volatility level forecasts.

Each of these  $x$  values implies a one-day-ahead 1% VaR return (calculated using the inverse cumulative distribution function of a Gaussian random variable,  $\Phi^{-1}(0.01; \mu, \sigma)$ ). The VaR error reduction from using TARCH rather than GARCH is calculated as:

$$\text{VaR error reduction}_i = 1 - \left| \frac{\Phi^{-1}(0.01; 0, x_{\text{TARCH},i}) - \Phi^{-1}(0.01; 0, 0.0146)}{\Phi^{-1}(0.01; 0, x_{\text{GARCH},i}) - \Phi^{-1}(0.01; 0, 0.0146)} \right|, \quad i \in \{\text{under, over}\}$$

The typical VaR forecast error reduction of TARCH relative to GARCH is 4.3% to 7.7%, which is equal to the volatility level forecast improvement to the nearest tenth of a percent.

Next we consider implied option-pricing errors from alternative models. Continuing from the previous example, we focus on one-day forecasts, and therefore on the value of an at-the-money call option with one day left until maturity. For simplicity, assume that the Black–Scholes model correctly prices options at this horizon, that the risk-free rate is 1% per annum, and that the value of the underlying (and the strike price) is normalized to one. From above, each  $x$  value implies the price of an at-the-money call option according to the Black–Scholes model. Defining the call option price as:

$$\text{BS}(\sigma) = \text{BS}(\sigma, r_f = 1\% \text{ p.a., TTM} = 1/365, S = 1, K = 1)$$

we calculate the reduction in call option mispricing using TARCH relative to GARCH as:

$$\text{BS error reduction}_i = 1 - \left| \frac{\text{BS}(x_{\text{TARCH},i}) - \text{BS}(0.0146)}{\text{BS}(x_{\text{GARCH},i}) - \text{BS}(0.0146)} \right|, \quad i \in \{\text{under, over}\}$$

The option-pricing error reduction of TARCH relative to GARCH is 4.3% to 7.7%, again equal to the volatility level forecast improvement to the nearest tenth of a percent.

#### 4 DID GARCH PREDICT THE CRISIS OF 2008?

On November 1, 2008, the *New York Times* (Norris (2008)) declared October of that year to be

the most volatile month in the 80-year history of the S&P 500.... In normal times, the market goes years without having even one [4% move]. There were none, for instance, from 2003 through 2007. There were three such days throughout the 1950s and two in the 1960s. In October, there were nine such days.

The economic fallout from this tumultuous period is now well understood, including the destruction of over 25% of the value of US capital stock. The reaction by many policy makers and academics and the popular press was to claim that economic models had been misused or were simply incorrect. Former Federal Reserve Chairman Alan Greenspan told one such story of misuse to the Committee of Government Oversight and Reform (Greenspan (2009)), concluding that risk models

collapsed in the summer of last year because the data inputted into the risk management models generally covered only the past two decades, a period of euphoria. Had instead the models been fitted more appropriately to historic periods of stress, capital requirements would have been much higher and the financial world would be in far better shape today.

Andrew Haldane, Executive Director for Financial Stability at the Bank of England, arrived at a starker conclusion regarding risk management models during the crisis (Haldane (2009)):

Risk management models have during this crisis proved themselves wrong in a more fundamental sense. These models were both very precise and very wrong.

In this section we attempt to put the forecasting results of the previous section in perspective and to ask whether volatility models, fundamental inputs for risk management tools, genuinely failed during the crisis. We first address this question with a simple thought experiment: how often would we observe forecast errors as large as those observed during the crisis if the world obeyed a GARCH model? Our answer takes a simple approach. First, we estimate a Gaussian TARCH model using the full sample of daily market returns from 1926 to 2008 and calculate multiple horizon in-sample forecast errors. Table 5 on the facing page presents average daily QL losses during the full sample, during the low volatility 2003–7 subsample, and during the fall 2008 sample. Average losses at all horizons in the full sample hover around 1.7, and are very similar to the losses experienced in the low volatility interval. As we turn to average crisis losses, we see that one-step-ahead losses are virtually the same as the rest of the sample. The severity of the crisis only becomes noticeable at longer forecast horizons. The twenty-two-day-ahead forecast loss appears to double during the crisis.

From the historical distribution of losses, we next calculate the probability of observing losses at least as large as those seen during the crisis over a four-month period (that is, the length of our crisis sample). To do this, we divide our 1926–2008



**TABLE 5** In-sample QL losses.

	One day	One week	Two weeks	Three weeks	One month
January 1926 to December 2008	1.5	1.6	1.7	1.7	1.8
January 2003 to August 2008	1.4	1.5	1.5	1.5	1.5
September 2008 to December 2008	1.4	1.6	2.1	2.8	4.1

The table reports average QL losses of the TARCH model with Student  $t$  innovations in the samples (i) January 1926 to December 2008, (ii) January 2003 to August 2008 and (iii) September 2008 to December 2008. The variance proxy is the squared return.

**TABLE 6** QL loss exceedance probabilities, September 2008 to December 2008.

	One day	One week	Two weeks	Three weeks	One month
Historical	54.5	38.2	12.3	3.6	1.3
Simulated	53.8	35.4	11.4	3.9	2.0

The table reports the historical and simulated probabilities of observing losses greater than or equal to those observed in the September 2008 to December 2008 sample.

sample into a sequence of overlapping four-month windows. In each the of these windows, we compute the forecast losses at the different forecasting horizons of interest. Finally, for each forecasting horizon we compute the proportion periods in which the losses were larger than the ones observed in the fall of 2008. These are reported in the first row of Table 6. The historical probability of observing a one-day loss at least as large as that observed during the crisis is 54.5%. That is, the average one-step crisis loss falls in the center of the empirical distribution. For longer horizons, the historical exceedance probabilities decrease quickly. At a twenty-two-day horizon, losses at least as large as those observed during the crisis occurred only 1.3% of the time between January 1926 and August 2008.

As a second approach to the question, we simulate data from the TARCH model using parameters estimated over the full sample using Student  $t$  innovations. In each simulation we generate eighty-two years of returns, then estimate the correctly specified model (estimation builds sampling error into Monte Carlo forecasts in analogy to our empirical procedure). Using estimated parameters, we construct in-sample forecasts at multiple horizons and calculate average losses. Next, we average the daily losses in the last four months of each simulated sample. Simulations are repeated 5000 times and produce a simulated distribution of average daily losses. Finally, we count the number of simulations in which losses meet or exceed crisis losses observed in the data. Results are reported in the last row of Table 6. Under the null model, the probability of observing one-step-ahead losses greater than the 1.4 value in the crisis

is 53.8%. This exceedance probability drops to 2.0% at the twenty-two-day horizon. The cumulative loss probabilities implied by the historical record and simulations under the null model tell the same story. In terms of one-step-ahead forecasts, the crisis sample was a typical season in a GARCH world. In contrast, one-month forecast losses were indeed aggravated during the crisis, but do not fall outside a 99% confidence interval. We have also performed this analysis using the other GARCH specifications used in the forecasting exercise. Interestingly, all the specifications that allow for asymmetric effects (that is, all models but the plain GARCH) deliver analogous findings.

The nature of volatility during the crisis seems to be captured by the facts that:

- (1) crisis forecasts deteriorated only at long horizons;
- (2) over one day, errors were no larger than a typical day in the full eighty-year sample.

On a given date during the crisis, conditioning on poor returns up until that day resulted in well-informed forecasts for the next day, and thus mild average one-day losses. However, this conditioning provided little help in predicting abnormally long strings of consecutive negative-return days that occurred during the crisis. Overall, the crisis does not lead us to reject standard time-series models used for volatility analysis as fundamentally flawed. However, it does indeed remind us that episodes of turmoil like the ones observed in the crisis do not have a negligible probability of occurring. We believe that the new challenge that has been raised is the development of effective ways to appropriately manage long-run risks. Most risk management practice is focused on short-run measures that are intrinsically myopic. Indeed, the extremely low levels of volatility observed in 2006–7 induced many institutions to take excessive risk, and this turned out to be a worsening factor during the crisis. Better long-run risk management would provide a more useful assessment of the actual level of downside exposure of an asset.

## 5 CONCLUSION

Volatility forecasting assessments are commonly structured to hold the test asset and estimation strategy fixed, focusing on model choice. We take a pragmatic approach and consider how much data should be used for estimation, how frequently a model should be reestimated, and what innovation distributions should be used. Our conclusions consider data from a range of asset classes, drawing attention to the relevance of multi-step-ahead forecast performance for model evaluation. We separately consider performance in crisis periods when volatility levels can escalate dramatically in a matter of days.

We find that asymmetric models, especially TARCH, perform well across methods, assets and subsamples. Models perform best using the longest available data series. Updating parameter estimates at least weekly counteracts the adverse effects of parameter drift. We find no evidence that the Student  $t$  likelihood improves forecasting ability, despite its potentially more realistic description of return tails. Preferred methods do not change when forecasting multiple periods ahead.

An exploration into the degree of extremity in volatility during the 2008 crisis reveals some interesting features. First and foremost, soaring volatility during that period was well described by short-horizon forecasts, as seen by mean forecast losses commensurate with historical losses and expected losses under the null. At longer horizons, observed losses have historical and simulated  $p$ -values of 1% to 2%. We conclude that while multi-step forecast losses are large and in the tail of the distribution, they cannot be interpreted as a rejection of GARCH models, and would have fallen within 99% predicted confidence intervals.

## REFERENCES

- Aït-Sahalia, Y., and Mancini, L. (2008). Out of sample forecasts of quadratic variation. *Journal of Econometrics* **147**, 17–33.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modelling and forecasting realized volatility. *Econometrica* **71**(2), 579–625.
- Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2006). Volatility and correlation forecasting. In *Handbook of Economic Forecasting*, Elliott, G., Granger, C. W. J., and Timmermann, A. (eds). North-Holland, Amsterdam.
- Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realised kernels in practice: trades and quotes. *Econometrics Journal* **12**, 1–32.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.
- Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics* **69**(3), 542–547.
- Brownlees, C., and Gallo, G. M. (2006). Financial econometric analysis at ultra high-frequency: data handling concerns. *Computational Statistics and Data Analysis* **51**(4), 2232–2245.
- Brownlees, C., Engle, R., and Kelly, B. (2011). A practical guide to volatility forecasting through calm and storm. (Appendix.) Technical Report. URL: <http://pages.stern.nyu.edu/~cbrownle>.
- Chiriac, R., and Pohlmeier, W. (2010). How risky is the value at risk? Technical Report, Rimini Centre for Economic Analysis.
- Corsi, F. (2010). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* **7**, 174–196.
- Deo, R., Hurvich, C., and Lu, Y. (2006). Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. *Journal of Econometrics* **131**(1), 29–58.

- Ding, Z., Engle, R., and Granger, C. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance* **1**(1), 83–106.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4), 987–1007.
- Engle, R. F. (1990). Discussion: stock market volatility and the crash of '87. *Review of Financial Studies* **3**, 103–106.
- Engle, R. F., and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics* **131**(1), 3–27.
- Ghysels, E., Harvey, A., and Renault, E. (1995). Stochastic volatility. In *Handbook of Statistics 14: Statistical Methods in Finance*, Maddala, G., and Rao, C. (eds), pp. 119–191. North-Holland, Amsterdam.
- Glosten, L. R., Jagannathan, R., and 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.
- Greenspan, A. (2009). Testimony to the Committee of Government Oversight and Reform (October 23). URL: <http://democrats.oversight.house.gov/images/stories/documents/20081023100438.pdf>.
- Haldane, A. G. (2009). Why banks failed the stress test. Speech presented at the Marcus Evans Conference on Stress-Testing (February 9–10).
- Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics* **23**(4), 365–380.
- Hansen, P. R., and Lunde, A. (2005a). A forecast comparison of volatility models: does anything beat a GARCH(1,1). *Journal of Applied Econometrics* **20**(7), 873–889.
- Hansen, P. R., and Lunde, A. (2005b). Consistent ranking of volatility models. *Journal of Econometrics* **131**(1), 97–121.
- Hansen, P. R., Lunde, A., and Nason, J. M. (2003). Choosing the best volatility models: the model confidence set approach. *Oxford Bulletin of Economics and Statistics* **65**, 839–861.
- Hansen, P. R., Huang, Z., and Shek, H. H. (2010). Realized GARCH: a complete model of returns and realized measures of volatility. Technical Report, School of Economics and Management, University of Aarhus.
- Hentschel, L. (1995). All in the family: nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics* **39**, 71–104.
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* **59**(2), 347–370.
- Norris, F. (2008). A monthlong walk on the wildest side of the stock market. *New York Times* (November 1).
- Patton, A. (2009). Volatility forecast comparison using imperfect volatility proxies. Technical Report, University of Oxford.
- Poon, S., and Granger, C. W. J. (2003). Forecasting volatility in financial markets: a review. *Journal of Economic Literature* **51**(2), 478–539.
- Poon, S., and Granger, C. W. J. (2005). Practical issues in forecasting volatility. *Financial Analysts Journal* **61**(1), 45–56.
- Shephard, N., and Sheppard, K. (2010). Realising the future: forecasting with high frequency based volatility (heavy) models. *Journal of Applied Econometrics* **25**, 197–231.