A Practical Guide to Volatility Forecasting through Calm and Storm*

Christian Brownlees Robert Engle Bryan Kelly

Abstract

We present a volatility forecasting comparative study within the 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 re-estimation. Test assets include a range of domestic and international equity indices and exchange rates. We find that model rankings are insensitive to forecast horizon and suggestions for estimation best practices emerge. While our main sample spans 1990-2008, we take advantage of the near-record surge in volatility during the last half of 2008 to ask if 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 one-day ahead, and should have been within risk managers' 1% confidence intervals up to one month ahead.

Keywords: Volatility, ARCH, Forecasting, Forecast Evaluation.

*Brownlees is at the Department of Finance, NYU Stern, Henry Kaufman Management Center, 44 West Fourth Street, New York, NY 10012 Office: 9-171 , phone: 212-998-4034, e-mail: ctb@stern.nyu.edu. Engle is at the Department of Finance, NYU Stern, Henry Kaufman Management Center, 44 West Fourth Street, New York, NY 10012 Office: 9-62 , phone: 212-998-0710, e-mail: rengle@stern.nyu.edu. Kelly is at the Department of Finance, Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue Chicago, IL 60637 phone: 773-702-8359, e-mail: bryan.kelly@chicagobooth.edu. Detailed results on the volatility forecasting exercise performed in this paper can be found in the appendix contained in Brownlees, Engle & Kelly (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 comments.

1 Introduction

The crash of 2008 has led practitioners and academics alike to reassess the adequacy of our financial models. Soaring volatilities across asset classes have made it especially important to know how well our standard tools forecast volatility, especially 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 essentially 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 hanging on a precipice of inadequate capital.

In this work we explore the performance of volatility forecasting within the class of 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 if 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 one day advance notice 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 not be interpreted merely as a critique of volatility forecasting but more importantly as a critique of our most widely used risk measures, value-at-risk 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 should be 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 GARCH specification, the threshold GARCH model of Glosten, Jagannanthan & Runkle (1993), is most often the best forecaster. How much data to use in estimation becomes an important issue if parameters are unstable since 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 re-estimated 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, Bollerslev, Diebold & Labys (2003), Deo, Hurvich & Lu (2006), Engle & Gallo (2006), Ait-Sähalia & Mancini (2008), Hansen, Huang & Shek (2010), Corsi (2010), and Shephard & Sheppard (2010), and Ghysels, Harvey & Renault (1995), among others. 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 impact 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 re-estimation using long samples optimizes precision while mitigating the impact of parameter drift.

This paper is related to the vast literature on volatility forecasting. Andersen, Bollerslev, Christoffersen & Diebold (2006) provide a comprehensive theoretical overview on the topic. An extensive survey of the literature's main findings is provided in Poon & 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 re-estimated, 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 & Lunde (2005*b*) and Patton (2009). Conflicting evidence reported by previous studies is due in part to the use of non-robust losses, and our assessment addresses this shortcoming.

2 Volatility Forecasting Methodology

2.1 Recursive Forecast Procedure

A time series of continuously compounded returns (including dividends) is denoted $\{r_t\}_{t=1}^{T}$, and $\mathcal{F}_t$ denotes the information set available at t. The unobserved variance of returns conditional on $\mathcal{F}_t$ is $\sigma^2_{t+i|t} \equiv \mathsf{Var}[r_{t+i}|\mathcal{F}_t]$. Variance predictions are obtained from a set of volatility models

$\mathcal{M} \equiv \{m_1, m_2, ..., m_M\}$. Model m can generically be represented as

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

where $h_{t+1}^{(m)}$ is an $\mathcal{F}_t$-measurable function and ϵ_{t+1} is an *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 $h_{t+i|t}^{(m)}$.

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 re-estimation. We use the fitted model to then predict volatility at different horizons (one, five, ten, 15 and 22 days ahead), resulting in a daily volatility forecast path $\{h_{t+i|t}^{(m)}\}$. This procedure generates a sequence of overlapping forecast paths, 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 if alternative estimation strategies can improve forecasting performance. In particular, we consider using four- and eight-year rolling estimation windows, rather than a growing window that uses the full post-1990 sample. We also explore re-estimating 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 trade-offs faced in estimation design. Interested readers will find exhaustive comparisons in the appendix of this work in Brownlees, Engle & Kelly (2011).

2.2 Volatility Models

The five models we consider for $h_{t+1}^{(m)}$ in Equation 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); Bollerslev (1986)), is a natural starting point for model com-

parison due to its ubiquity and progenesis of alternative models. GARCH describes the volatility process 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 volatilites to increase more when past returns are negative. Threshold ARCH (Glosten et al. (1993)), or TARCH, appends a linear asymmetry adjustment,

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

where I 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, Engle & Granger (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 δ can capture volatility dynamics more flexibly than other specifications, while asymmetries are

incorporated via γ. 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 thus 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 losses traditionally used in the volatility forecast literature but two:

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

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 multplicative forecast error, $\frac{\hat{\sigma}_t^2}{h_{t|t-k}}$. 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 the appendix. 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. MSE, which depends on additive errors, scales with the square of variance, 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. MSE is therefore an *iid* process times 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 η_t is a measurement error with unit expected value and variance τ^2. The expected value of MSE is then

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

while the expected value of QL is

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

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. QL avoids this ambiguity, making it easier to compare losses across volatility regimes.

3 Empirical Volatility Forecasting Results

3.1 Data

Daily split- and dividend-adjusted log return data on the S&P 500 index from 1990 to 2008 is from Datastream. The expanded data set for our large scale forecasting comparison includes three balanced panels of assets listed in Table 1. We use ten exchange rates, nine domestic sectoral equity indices, and 18 international equity indices. The sector index data are returns on S&P 500 industry sector SPDR 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 (the introduction of the EURO).

To proxy for true S&P 500 variance, we use daily realized volatility for the S&P 500 SPDR exchange traded fund. We construct this series from NYSE-TAQ intra-daily mid-quotes (filtered with procedures described in Brownlees & Gallo (2006) and Barndorff-Nielsen, Hansen, Lunde & Shephard (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 $p_{t,i}$ $(i = 1, ..., I_t)$ 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, $\hat{\sigma}^2_{\mathsf{rv}\,t} = \sum_{i=2}^{I_t}(p_{t\,i} - p_{t\,i-1})^2$. 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 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. US equity volatility 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.

Asset Class	Assets	Begin Date
Exchange Rates	Australian Dollar, British Pound, Canadian Dollar, EURO, Indian Rupee, Honk Kong Dollar, Japanese Yen, South Korean Won, Swiss Franc, Thai Baht	1999-01-05
Equity Sectors	Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Materials, Technology, Utilities	1998-12-23
International Equities	Singapore, Netherlands, Japan, Australia, Belgium, Canada, Germany, Hong Kong, Italy, Switzerland, Sweden, Spain, Mexico, UK, World, Emerging Markets, BRIC	1996-03-19

Table 1: Asset list. For each asset class, the table reports the list of assets used in the forecasting application and the first date of the sample.

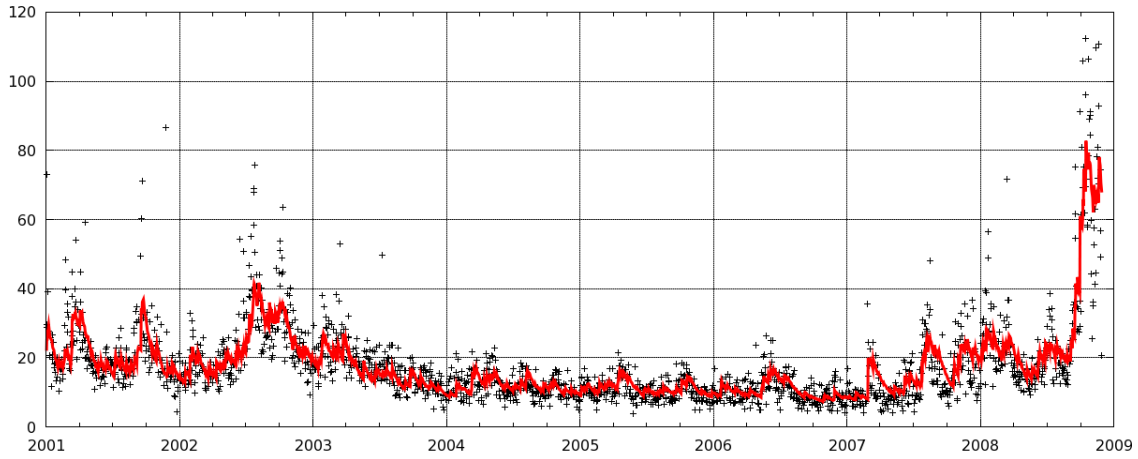

Figure 1: S&P 500 TARCH one step ahead volatility forecasts (solid line) and realized volatility (crosses). Volatilities are expressed in annualized terms.

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 summarizes the extensive analysis provided in our appendix 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 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 rv's improved efficiency. 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 and the ones obtained by the various modifications: if the mean of the loss differential is significantly different from zero, than the null of equal predictive ability is rejected.[1] The appendix presents evidence of model parameter instability, highlighting the relevance of choices for amount of data used in estimation and frequency of re-estimation. Our analysis suggests that the longest possible estimation window gives the best results, but suggest re-estimating 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.

	Vol. Proxy: r^2					Vol. Proxy: rv				
Model	1 d	1 w	2 w	3 w	1 m	1 d	1 w	2 w	3 w	1 m
GARCH	1.460	1.481	1.520	1.574	1.645	0.273	0.310	0.343	0.373	0.414
			UD		S					
TARCH	1.415	1.442	1.478	1.547	1.624	0.243	0.289	0.328	0.368	0.415
			S	S	S	WM	WM UD			
EGARCH	1.420	1.458	1.505	1.592	1.684	0.234	0.282	0.320	0.365	0.413
			S	S	S		UM	UM	UM	
APARCH	1.417	1.446	1.485	1.557	1.633	0.249	0.299	0.340	0.385	0.435
			S	S	S	WM S	WM UD	S WD UD	S WM	S
NGARCH	1.422	1.459	1.498	1.574	1.659	0.244	0.296	0.337	0.380	0.432
			S UD	S	S	UD		UD		

Table 2: Estimation Strategy Assessment. 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).

	2001-2008									
	Vol. Proxy: r^2					Vol. Proxy: rv				
Model	1 d	1 w	2 w	3 w	1 m	1 d	1 w	2 w	3 w	1 m
TARCH	1.415	1.442	1.478	1.547	1.624	0.243	0.289	0.328	0.368	0.415
	Fall 2008									
TARCH	1.461	1.560	1.985	2.311	2.875	0.304	0.353	0.590	0.672	1.380

Table 3: S&P 500 volatility prediction performance of the TARCH model from 2001 to 2008 and in Fall 2008. For each volatility proxy the table reports the out-of-sample QL loss at multiple horizons for the TARCH(1,1) model.

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. 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 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 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). This analysis uses the QL loss with squared returns as proxy. Volatility forecast losses are averaged across time and over

[1] The Superior Predictive Ability test (SPA) or Model Confidence Set (MCS) techniques could also be used to carry out this type of exercise (cf. Hansen (2005) and Hansen, Lunde & Nason (2003)).

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 & Lunde (2005*a*).

Interestingly, asymmetric models appear to improve exchange rate volatility forecasts during the 2008 crisis. This is consistent with a flight-to-quality during 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 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 magnitude 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 i) differences in value-at-risk (VaR) forecast errors and ii) option pricing errors. We show that the relative size of forecast errors across models provides an accurate description of relative magnitudes of both value-at-risk 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

$$\frac{.0146^2}{x^2} - \log\left(\frac{.0146^2}{x^2}\right) - 1 = \text{ average QL loss.}$$

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.[2] 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 $x_{GARCH,under} = 0.0105$ and $x_{GARCH,over} = 0.0222$, and $x_{TARCH,under} = 0.0107$ and $x_{TARCH,over} = 0.0217$. The reduction in volatility forecast errors achieved by moving from GARCH to TARCH is calculated as

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

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% value-at-risk return (calculated using the inverse cumulative distribution function of a Gaussian random variable, $\Phi^{-1}(0.01; \mu, \sigma)$). The value-at-risk error reduction from using TARCH rather than GARCH is calculated as

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

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

We next consider impled 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

[2]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.

of the underlying (and the strike price) are normalized to one. From above, each x value implies the price of an at-the-money call option according to the Black-Scholes model. Denoting the call option price as $BS(\sigma) = BS\left(\sigma, r_f = 1\%\ p.a., TTM = 1/365, S = 1, K = 1\right)$, we calculate the reduction in call option mispricing using TARCH relative to GARCH as

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

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[3] declared October to be "the most wild 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 destruction of over 25% of the US capital stock's value. The reaction by many policy makers, 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 (Jan. 2, 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 about risk management models during the crisis (Feb. 13, 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."

[3] *The New York Times*,"A Monthlong Walk on the Wildest Side of the Stock Market."

	2001-2008														
	Exchange Rates					Equity Sectors					International Equities				
Model	1 d	1 w	2 w	3 w	1 m	1 d	1 w	2 w	3 w	1 m	1 d	1 w	2 w	3 w	1 m
TARCH	1.976	2.038	2.093	2.118	2.138	2.236	2.266	2.313	2.356	2.412	2.253	2.289	2.337	2.389	2.464
	Fall 2008														
TARCH	1.925	2.081	2.454	2.636	3.244	2.217	2.120	2.523	2.834	3.870	2.153	2.054	2.574	3.349	4.250

Table 4: We report the out-of-sample loss for various asset classes at multiple horizons using the TARCH(1,1) model for the full 2001-2008 sample and fall 2008 subsample.

	1d	1w	2w	3w	1m
Jan. 1926 to Dec. 2008	1.5	1.6	1.7	1.7	1.8
Jan. 2003 to Aug. 2008	1.4	1.5	1.5	1.5	1.5
Sep. 2008 to Dec. 2008	1.4	1.6	2.1	2.8	4.1

Table 5: In-sample QL losses. 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.

In this section we attempt to put the forecasting results of the previous section in perspective and to ask the following question: Did volatility models, fundamental inputs for risk management tools, genuinely fail 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 obeys 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 presents average daily QL losses during the full sample, during the low volatility 2003-2007 subsample, and during the fall 2008 crisis 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 22-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 4 months period (that is, the length of our crisis sample). To do this, we divide our 1926 - 2008 sample in a sequence of overlapping 4 months 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 where 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 exeedence probabilities decrease quickly. At a 22-day horizon, losses at least as large as those observed during the crisis occurred only 1.3% of 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 82 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 5,000 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 exceedence probability drops to 2.0% at the 22-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 specification 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 (i) crisis forecasts deteriorated only at long horizons, and (ii) over one-day, errors were no larger than a typical day in the full 80 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. Over ally, 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 manage appropriately 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 2006-2007 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 re-estimated, 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

Ait-Sähalia, Y. & Mancini, L. (2008), 'Out of sample forecasts of Quadratic Variation', *Journal of Econometrics* **147**, 17–33.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. & Diebold, F. X. (2006), Volatility and correlation forecasting, *in* G. Elliott, C. W. J. Granger & A. Timmermann, eds, 'Handbook of Economic Forecasting', North Holland.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modelling and Forecasting Realized Volatility', *Econometrica* **71**(2), 579–625.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2009), 'Realised kernels in practice: trades and quotes', *Econometrics Journal* **12**.

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', *The Review of Economics and Statistics* **69**(3), 542–547.

Brownlees, C., Engle, R. & Kelly, B. (2011), Appendix to "A Practical Guide to Volatility Forecasting through Calm and Storm", Technical report. `http://pages.stern.nyu.edu/~cbrownle`.

Brownlees, C. & Gallo, G. M. (2006), 'Financial Econometric Analysis at Ultra High-Frequency: Data Handling Concerns', *Computational Statistics and Data Analysis* **51**(4), 2232–2245.

Chiriac, R. & Pohlmeier, W. (2010), How risky is the value at risk?, Technical report.

Corsi, F. (2010), 'A simple approximate long-memory model of realized volatility', *Journal of Financial Econometrics* **7**, 174–196.

Deo, R., Hurvich, C. & Lu, Y. (2006), 'Forecasting Realized Volatility Using a Long-Memory Stochastic Volatility Model: Estimation, Prediction and Seasonal Adjustment', *Journal of Econometrics* **131**(1-2), 29–58.

Ding, Z., Engle, R. & 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. (1990), 'Discussion: Stock Market Volatility and the Crash of '87', *Review of Financial Studies* **3**, 103–106.

Engle, R. F. (1982), 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica* **50**(4), 987–1007.

Engle, R. F. & Gallo, G. M. (2006), 'A Multiple Indicators Model for Volatility Using Intra-Daily Data', *Journal of Econometrics* **131**(1-2), 3–27.

Ghysels, E., Harvey, A. & Renault, E. (1995), Stochastic Volatility, *in* G. Maddala & C. Rao, eds, 'Handbook of Statistics 14, Statistical Methods in Finance,', North Holland, Amsterdam,, pp. 119–191.

Glosten, L. R., Jagannanthan, R. & Runkle, D. E. (1993), 'On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks', *The Journal of Finance* **48**(5), 1779–1801.

Hansen, P. R. (2005), 'A Test for Superior Predictive Ability', *Journal of Business & Economic Statistics* **23**(4), 365–380.

Hansen, P. R., Huang, Z. & Shek, H. H. (2010), Realized garch: A complete model of returns and realized measures of volatility, Technical report.

Hansen, P. R. & Lunde, A. (2005*a*), 'A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)', *Journal of Applied Econometrics* **20**(7), 873–889.

Hansen, P. R. & Lunde, A. (2005*b*), 'Consistent ranking of volatility models', *Journal of Econometrics* **131**(1–2), 97–121.

Hansen, P. R., Lunde, A. & Nason, J. M. (2003), 'Choosing the Best Volatility Models: The Model Confidence Set Approach', *Oxford Bulletin of Economics and Statistics* **65**, 839–861.

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.

Patton, A. (2009), Volatility Forecast Comparison using Imperfect Volatility Proxies, Technical report, University of Oxford.

Poon, S. & Granger, C. W. J. (2003), 'Forecasting Volatility in Financial Markets: A Review', *Journal of Economic Literature* **51**(2), 478–539.

Poon, S. & Granger, C. W. J. (2005), 'Practical Issues in Forecasting Volatility', *Financial Analysts Journal* **61**(1), 45–56.

Shephard, N. & Sheppard, K. (2010), 'Realising the future: forecasting with high frequency based volatility (heavy) models', *Journal of Applied Econometrics* **25**, 197–231.

	1d	1w	2w	3w	1m
Historical	54.5	38.2	12.3	3.6	1.3
Simulated	53.8	35.4	11.4	3.9	2.0

Table 6: QL loss exceedence probabilities September 2008 to December 2008. 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.