Scientific Research Publishing

Theoretical Economics Letters, 2018, 8, 1-27
http://www.scirp.org/journal/tel
ISSN Online: 2162-2086
ISSN Print: 2162-2078

Evaluating Volatility Forecasts with Ultra-High-Frequency Data—Evidence from the Australian Equity Market

Kai Zhang[*]**, Lurion De Mello, Mehdi Sadeghi**

Department of Applied Finance and Actuarial Studies, Macquarie University, Sydney, Australia
Email: *kai.zhang@mq.edu.au

How to cite this paper: Zhang, K., De Mello, L. and Sadeghi, M. (2018) Evaluating Volatility Forecasts with Ultra-High-Frequency Data—Evidence from the Australian Equity Market. *Theoretical Economics Letters*, **8**, 1-27.
https://doi.org/10.4236/tel.2018.81001

Received: November 10, 2017
Accepted: January 1, 2018
Published: January 4, 2018

Abstract

Due to the unobserved nature of the true return variation process, one of the most challenging problems in evaluation of volatility forecasts is to find an accurate benchmark proxy for *ex-post* volatility. This paper uses the Australian equity market ultra-high-frequency data to construct an unbiased *ex-post* volatility estimator and then use it as a benchmark to evaluate various practical volatility forecasting strategies (GARCH class model based). These forecasting strategies allow for the skewed distribution of innovations and use various estimation windows in addition to the standard GARCH volatility models. In out-of-sample tests, we find that forecasting errors across all model specifications are systematically reduced if using the unbiased *ex-post* volatility estimator compared with those using the realized volatility based on sparsely sampled intra-day data. In particular, we show that the three benchmark forecasting models outperform most of the modified strategies with different distribution of returns and estimation windows. Comparing the three standard GARCH class models, we find that the asymmetric power ARCH (APARCH) model exhibits the best forecasting power in both normal and financial turmoil periods, which indicates the ability of APARCH model to capture the leptokurtic returns and stylized features of volatility in the Australian stock market.

Keywords

High-Frequency Volatility, Volatility Forecasting, GARCH, Volatility Forecast Evaluation

1. Introduction

Volatility forecasting is a critical task in a variety of financial activities for different financial instruments and markets around the world. Important financial decisions such as portfolio optimisation, derivative pricing, risk management and financial regulation heavily depend on volatility forecasts. In derivative pricing, such as in the estimation of the Black-Scholes option pricing model, volatility is the only parameter that needs to be forecasted. The prediction of volatility is also crucial in development of Value at Risk (VaR) and a variety of systemic risk models, as well as in banking and finance regulations. For example, according to the Basel Accord II and III, it is compulsory for all financial institutions to predict the volatility of their financial assets to incorporate the risk exposure of their capital requirements.

The focus of our study is on the predictive ability of the popular AutoRegressive Conditional Heteroscedastic (ARCH) class of models that originated from a seminal Nobel Prize-wining article by [1] [2] generalized his framework to obtain the GARCH model. ARCH and GARCH models are popular and standard volatility forecasting models in econometrics and finance. Documented stylized features of variation such as the clustering and long memory effect can be captured by GARCH class models, and the model parameters are relatively easy to estimate. A comprehensive survey of the GARCH family of models can be found in [3]. The current study selects three popular GARCH class models from the literature, including the standard GARCH, the threshold GARCH (TGARCH) of [4] and [5], and the asymmetric power ARCH (APARCH) of [6]. In addition to the three standard models, we consider 12 corresponding forecasting strategies associated with them, which involves different estimation windows and errors distributions.

The predictive power of a volatility model is evaluated based on an out-of-sample test in which the predicted volatility generated from the model is compared with the *ex-post* volatility measurements. Superior volatility forecasting models are supposed to have small forecasting errors, measured as the difference between the predicted and actual volatility. However, unlike the return, the volatility process cannot be observed. Therefore, in out-of-sample evaluation of volatility forecasting models, the crucial task is to find an accurate proxy for the underlying unobserved volatility process. In the mid-1990s, a series of empirical studies noted that while GARCH-type models are for fitting the time series returns, they failed to explain much of the variability in *ex-post* volatility measured by the squared returns in out-of-sample tests. Hence, the practical usefulness of GARCH models was challenged. [7] responds to the critique of the model and argues that the unsatisfied empirical results are due to the noisy volatility proxies used in these studies, that is, squared returns or absolute returns. In out-of-sample forecast evaluation, a common approach for evaluating the practical performance of any model is to compare the fitted predictions derived from the model with the subsequent realizations. However, volatility is not directly observed and dealt with as a latent variable in financial modelling. Therefore, the squared innovation return is usually employed as a proxy for volatility realization. It is empirically documented that the correlation between the predicted and the realized value is weak, with low R^2 values.[1] However, this finding is not surprising and is considered evidence of the poor predicting ability of a volatility model. Although in statistical terms, the squared innovation return is an unbiased estimator of underlying variance, it is not an accurate estimator and displays a large degree of observation-by-observation variation. In summary, empirical findings of poor forecasting performance are due to unsatisfied volatility proxies rather than the predictive power of GARCH class models.

The recent availability of nearly continuous high-frequency transaction data provides a chance to explicitly compute the market volatility by using intra-day data, referred to as realized volatility in the literature. [7] shows that the basic GARCH(1,1) model performs rather well when it is evaluated with a volatility measurement constructed using high-frequency data. This stems from the fact that high-frequency volatility is a more accurate measure of daily volatility compared with those estimated using low-frequency data. Instead of only using the opening or closing price of a trading day in squared daily returns, high-frequency volatility exploits more information contained in intra-day trading data. In the literature, five-minute intra-day data are popularly used to construct realised volatility. However, in our study, we show that while contamination with microstructure noise can be reduced if realised volatility is constructed based on sparsely sampled high-frequency data, it is still a biased estimator.

Because of the crucial role of *ex-post* volatility measurement in evaluating forecasting performance, in addition to using five-minutes tick data we calculate the realized variance using ultra-high-frequency data and relying on the Two Time Scaled Realized Volatility (TSRV) estimator proposed in the study of [11] and [12], which is shown to be an unbiased *ex-post* variance estimator. Unlike an arbitrarily and subjectively selected sample frequency such as five minutes, the TSRV employs all available high-frequency intra-day data and therefore exploits full information about return variation contained in the ultra-high-frequency data. Our results show that forecasting errors relying on TSRV and ultra-high-frequency data are significantly lower than those based on sparsely sampled intra-day data.

This paper is an attempt to mimic volatility forecasting strategies using GARCH family models in practice, and examine natural questions arising from employing such strategies in the stock market in Australia. These questions include, Which model provides the smallest error? Should we allow for heavy-tailed distributed innovations? What is the best choice for the estimation window? Would it be a growing window, a longer rolling window or a shorter rolling estimation window? Moreover, the performance of these volatility forecasting strategies in periods of financial turmoil such as the Global Financial

[1] The empirical evidence can be found in [8] [9] and [10].

Crisis (GFC) of 2008 are of concern for financial practitioners. The current paper responds to these questions by examining the empirical predictive power of various strategies in pure out-of-sample tests, relying on a recently developed unbiased *ex-post* volatility proxy and using ultra-high-frequency trading data in the Australian equity market.

Our research is also motivated by the fact that no such study has previously been conducted on the Australian stock market. [13] model and evaluate the monthly volatility forecasting techniques using Australian low-frequency data. They identify several unique features that distinguishing the Australian equity stock market from other stock markets. For instance, the top 20 firms listed on the Australian Stock Exchange (ASX) own 80% of the total assets in the market, whereas in the United States, the top 30 companies on the Dow Jones Industrial Average (DJIA) index hold only 65% of the total assets. The Australian market is dominated by energy and resources firms ([14]), which are more volatile than industrial shares. However, regulatory guidelines developed by the Australian Security and Investment Commission (ASIC) require disclosure of more financial information for investors to make decisions than is required in some other countries. This may reduce the level of information asymmetry and expected volatility in the stock market. Overall, Australian institutional features and settings are quite unique compared with other markets, therefore, examining volatility forecasting issues in Australian provides particular evidence in relation to the related literature.

In our empirical analysis using ultra-high-frequency data from the Australian stock market, we evaluate the predictability of three commonly used GARCH-type models and four variations of each. The three models are GARCH, TGARCH and APARCH[2]. The benchmark forecasting models are each GARCH-type model estimated with a growing estimation window and Gaussian innovations, and the variations are the models estimated with rolling forecasting windows (one-year and three-years rolling windows) and skewed and heavy-tailed innovation distributions (Student-t innovations and skew Student-t innovations). We firstly compare the predictive abilities of the variations of each separate GARCH model to the benchmark GARCH class model, which is estimated with normally distributed innovations in the full sample period from January 2005 to December 2013. Then we proceed by directly comparing the forecasting accuracy of each benchmark GARCH model to determine which one provides the best predictive performance. The forecasting ability comparison is based on two commonly used loss functions in the literature: the Mean Square Error (MSE) and the Quasi-like Loss function (QL). The Diebold-Mariano test is also implemented to examine the statistical significance of the improved forecast accuracy. Our results show that for each individual GARCH-type model, the benchmark usually is not significantly outperformed by the modifications. The exception is that the skew Student t distribution improves the forecast accuracy

[2] The Asymmetric Power ARCH model at least nests seven other GARCH-type models, see [6].

of the standard GARCH model. In the direct comparison of the three benchmark GARCH class of models, the APARCH usually gives the best forecasts in the out-of sample period from 2005 to 2013[3]. We note that the volatility forecasting procedure presented in this study is not limited to GARCH class models. In practice, the out-of-sample predictive performance of any volatility model can be evaluated based on our procedure.

The reminder of the paper is organized as follows: Section 2 provides a brief discussion of the theory around measuring high-frequency volatility. Section 3 and Section 4 respectively illustrate forecasting models and detailed procedures to compare prediction performance. Section 5 describes the daily data and tick data used in this paper. The main empirical results are presented in Section 6 and conclusions are drawn in Section 7.

2. Measurement of *Ex-Post* Daily Volatility

2.1. Theoretical Set-Up

Unlike for raw return, the actual daily return volatility process usually cannot be directly observed because there is just one daily return per trading day. Conventionally, volatility is treated as a latent variable[4] in parametric models such as GARCH-type and Stochastic volatility (SV) models that are inferred from *ex-post* low-frequency return data. Volatility measurement by these models is based on specific distributional assumptions and usually involves complex procedures in estimating model parameters. [15] introduced the concept of realized volatility for the first time.[5] The realized volatility is a non-parametric estimator that does not rely on the distribution of parameters. In the standard form, realized volatility is the second-order sample moment, that is, the sum of squares of the high-frequency returns over a fixed period, say, one day.

$$RV_t = \sum_{i=1}^{n} r_{t,i}^2$$

where $r_{t,i}$ is the *ith* high-frequency return for day *t*.

In financial asset pricing models, the asset price is assumed to be driven by a continuous time diffusion process,

$$\mathrm{d}X(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t)$$

where $X(t)$ is log price, $W(t)$ is standard Brownian Motion, and $\mu(t)$ and

[3] We also examine the performance of our selected volatility forecasting strategies during the financial turmoil period. The same procedure above is applied to the period of GFC in 2008. We choose three months after the date that the Lehman Brothers Holdings filed for Chapter 11 of the United States bankruptcy protection code as the test period. We compare the forecast performance of GARCH models in the crisis period with three months in early 2008 which has much lower unconditional daily volatility. We find that during financial turmoil, the degree of forecast losses are significantly increased, however, the overall ranking of the forecast does not change. The APARCH provides the best forecast across all cases suggesting the importance of the role of negative returns in predicting future volatility when the market is down.

[4] Latent variables cannot be directly observed and are estimated by using other observed variables

[5] In literature, realized volatility and realized variance are often used interchangeably.

$\sigma(t)$ are drift and diffusion term respectively. In integral format, the process is:

$$X(t) = X(0) + \int_0^t \mu_s \mathrm{d}s + \int_0^t \sigma_s \mathrm{d}W_s$$

The underlying daily variation of return is then measured by the Quadratic Variation (QV):

$$QV_t = \int_{t-1}^t \sigma_s^2 \mathrm{d}s$$

where σ_s is the spot volatility process. Since σ_s is latent, GARCH and SV models treat QV_t as unobservable, and infer it from past realized values of daily return data, that is $r_{t-1}, r_{t-2}, \cdots$. [16] [17] [18] prove that if the underlying asset log price is a semi-martingale, the quadratic variation theory ensures that RV_t converges in probability to the Quadratic Variation (QV_t) of asset return, which is the actual underlying volatility we would like to measure in the continuous framework:[6]

$$\sum_{i=1}^n r_{t,i}^2 \to \int_{t-1}^t \sigma_s^2 \mathrm{d}s$$

Thus, non-parametric realized volatility provides an efficient measure of daily market volatility and allows us to treat realized volatility as an observable variable rather than a latent one.

However, high-frequency raw data are contaminated by microstructure noise reflecting market frictions such as bid-ask bounce and price discreteness. Mathematically, the observed log price can be decomposed into two parts:

$$Y_t = X_t + \varepsilon_t$$

where X_t is the latent price process and ε_t denotes microstructure noise and which is independent of X_t. It then can be shown that

$$RV(Y_t) = 2nE(\varepsilon^2) + \int_0^t \sigma_s^2 \mathrm{d}s + O_p(\sqrt{n}) \tag{1}$$

where n is the sampling frequency. Note in Equation (1), as sampling frequency increases, the integrated variation that measures the actual volatility of the true price process will be swamped by the error terms. Hence, it may be unwise to sample the data too often when estimating RV. One way to mitigate the contamination caused by microstructure noise is sparsely sampling high-frequency data, such as sampling every five minutes instead of every second. This will reduce the bias term, because $n^{(sparse)} < n$. In empirical work, a five-minute time interval is widely used as the sampling frequency. However, although the effect of the bias term can be mitigated, sparse sampling cannot completely remove it. Moreover, too much data are thrown away if high-frequency data are sampled every five minutes and this violates the statistical principle. In recent years, a few consistent estimators have been proposed that are designed to accurately calculate realized volatility by directly modelling microstructure noise. Theoretical and simulation studies show that they can improve the estimation to a large extent. (see [11] [12] [19] [20])

[6] To be accurate, the convergence is in probability, that is $\operatorname{plim}\sum_{i=1}^n r_{t,i}^2 = \int_{t-1}^t \sigma_s^2 \mathrm{d}s$.

2.2. An Unbiased *Ex-Post* Volatility Estimator Using Ultra-High-Frequency Data

As we discussed above, estimating daily volatility (RV_t) by using all the high-frequency observations will lead to a rather unreliable result. Sparsely sampling can mitigate the effect of microstructure noise, but at the cost of discarding a huge amount of data, which is not advisable. Moreover, sparsely sampled estimators are statistically biased. [11] proposes a method of utilizing the full data set, which provides a consistent estimator, the TSRV, which is estimated based on the assumption that noise is independently and identically distributed (i.i.d) and independent of the price process. It involves three steps:

1) Sub-sampling

Firstly, partition a full grid of observation $\Lambda = \{t_0, \cdots, t_n\}$ into M non-overlap sub-grids, $\Lambda^{(m)}, m = 1, \cdots, M$, such that $n/M \to \infty$ as $n \to \infty$.

2) Averaging

For each sub-sample, we calculate the $RV_t^{(m)}$ and then average these estimators.

$$RV_t^{(avg)} = \frac{1}{M} \sum_{m=1}^{M} RV_t^{(m)}$$

3) Bias-Correcting

Although $RV_t^{(avg)}$ remains biased, the bias is now $2\bar{n}E\left(\varepsilon^2\right)$, which is smaller than the original bias because $\bar{n}$ is the average size defined as n/M which is much smaller than the full size n.

To remove the bias $2\bar{n}E\left(\varepsilon^2\right)$ from $RV_t^{(avg)}$, we need to estimate $E\left(\varepsilon^2\right)$. According to [11], this can be estimated by $\frac{1}{2n} RV_t^{(all)}$[7]. Now the unbiased estimator is:

$$RV_t^{(tsrv)} = RV_t^{(avg)} - 2\bar{n}E\left(\varepsilon^2\right) = RV_t^{(avg)} - \frac{\bar{n}}{n} RV_t^{(all)} \tag{2}$$

To the best of our knowledge, TSRV is the first proposed consistent estimator of quadratic variation (QV_t). [21] further proposes a multi-scale RV (MSRV) that generalizes TSRV in Equation (2) by averaging more scales instead of just two ($RV^{(avg)}$ and $RV^{(all)}$). However, MSRV is difficult to implement in practice, and it does not significantly outperform TSRV. [12] examine the TSRV after relaxing the assumption of i.i.d microstructure noise. They find that TSRV works even in the situation where the noise is serially dependent.

In implementing the calculation of TSRV, firstly we can partition the full grid of data point, $\Lambda = \{t_0, \cdots, t_n\}$ into subgrids, $\Lambda^{(m)}, m = 1, \cdots, M, n \gg M$, such that:

$$\Lambda = \bigcup_{m=1}^{M} \Lambda^{(m)}, \quad \text{where } \Lambda^{(k)} \cap \Lambda^{(l)} = \varnothing, \quad \text{when } k \neq l.$$

For example, consider the full grid in tick time is $\Lambda = \{t_0, \cdots, t_{10}\}$, if we choose to sub-sample every three transactions, then the sub-grids are:

[7] RV_t^{all} is the realized volatility estimated by using all the data.

$$\Lambda^{(1)} = \{t_0, t_3, t_6, t_9\},$$
$$\Lambda^{(2)} = \{t_1, t_4, t_7, t_{10}\},$$
$$\Lambda^{(3)} = \{t_2, t_5, t_8\}.$$

For each sub-sample, we calculate the RV by summing the squared returns and then averaging RV estimators:

$$RV_t^{(avg)} = \frac{1}{M}\sum_{m=1}^{M} RV_t^{(m)} \tag{3}$$

$RV_t^{(avg)}$ is thus the average of the $RV_t^{(m)}$ for M sub-samples with average size $\bar{n} = m/M$. The bias for $RV_t^{(avg)}$ is now $2\bar{n}E(\varepsilon^2)$ which is smaller than the original noise $2nE(\varepsilon^2)$.

The next step is subtracting the noise term from our estimator to render it unbiased. The unbiased estimator is thus:

$$RV_t^{(tsrv)} = RV_t^{(avg)} - 2\bar{n}E(\varepsilon^2).$$

To perform the final step, we need to estimate the error term $E(\varepsilon^2)$ which is unknown. Zhang *et al.* (2005) shows that $E(\varepsilon^2)$ can be consistently approximated using $RV_t^{(all)}$ which is the realized variance among all data sets:

$$\widehat{E(\varepsilon^2)} = \frac{1}{2n} RV_t^{(all)} \tag{4}$$

Hence the final unbiased estimator is:

$$RV_t^{(tsrv)} = RV_t^{(avg)} - \frac{\bar{n}}{n} RV_t^{(all)}. \tag{5}$$

3. Forecasting Models

This paper examines the empirical accuracy of various approaches in predicting the conditional daily volatility σ_t. Let I_t denote the information set at time *t*. Then for *h*-step-ahead volatility forecasts we have

$$\sigma_{t+h|t}^2 = Var[r_{t+h} \mid I_t] = Var[a_{t+h} \mid I_t]$$
$$a_{t+h} = \varepsilon_{t+h}\sigma_{t+h}$$

where r_t is the log return series, a_t are innovations and ε_t follows a distribution with zero mean and unit variance. The predicted value of σ_{t+h} is obtained from an alternative GARCH-type model estimated by maximum likelihood estimation (MLE) using historical daily return in the sample.

The natural choice for the benchmark model is GARCH as proposed by [2]. The simple GARCH(1,1) has the form:

$$a_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1\sigma_{t-1}^2,$$

where a_t is the innovation at time *t*, ε_t is a sequence of identical independently distributed (iid) random variables with zero mean and unit variance. If the forecast starting point is *h*, for one-step-ahead forecast, we have

$$\sigma_h^2(1) = \alpha_0 + \alpha_1 a_h^2 + \beta_1\sigma_h^2$$

and for *L*-step-ahead forecast, we have

$$\sigma_h^2(L) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(L-1)$$

GARCH models largely improve the practical use by adequately describing the feature of volatility of asset return while limiting the number of parameters that need to be estimated compared with the ARCH model of [1]. It has been documented that the simple GARCH(1,1) model can capture most of the volatility process of financial asset return. A comprehensive study by [22] also show that the most improved versions of GARCH models do not significantly outperform GARCH(1,1) in forecasting volatility.

However, the GARCH model deals with the symmetric price increase and drop, and fails to capture asymmetric effects on volatility. Empirical evidence has shown that volatility usually responds differently to large positive return versus large negative return. To overcome the weakness of the GARCH and the asymmetric effect, the TGARCH model was proposed by [4] and [5]. TGARCH(1,1) assumes the form

$$\sigma_t^2 = \alpha_0 + (\alpha_1 + \gamma_1 N_{t-1})a_{t-1}^2 + \beta_1\sigma_{t-1}^2,$$

where N_{t-1} is the indicator for negative return innovation hence:

$$N_{t-1} = \begin{cases} 1 & \text{if } a_{t-1} < 0 \\ 0 & \text{if } a_{t-1} \geq 0 \end{cases}$$

The last model examined in this paper is the APARCH model of [6]. This model is commonly used in practice and [23] shows it nests many other GARCH models to capture the long memory feature in the volatility process. The APARCH has the following form:

$$a_t = \sigma_t\varepsilon_t, \quad \sigma_t^\delta = \omega + \alpha_1\left(|a_{t-1}| + \gamma_1 a_{t-1}\right)^\delta + \beta_1\sigma_{t-1}^\delta,$$

Here, ε follows a general distribution with zero mean and unit variance, and δ is a positive real number.

Table 1 summarises the in-sample performance of the GARCH, TGARCH and APARCH models in our sample period spanning 2002-2012. All of the estimated parameters of the models are significantly different from zero and mostly significant at the 1% level[8]. These volatility models are estimated using conditional maximum likelihood method.[9] Moreover, the skewness coefficient of our sample log return is −0.582633, which suggests the log return is negatively skewed[10]. Therefore, to deal with the skewness of daily return, the volatility

[8] The adequacy of the fitted models can be checked by the property of standardized residual series $\tilde{a}_t = a_t/\sigma_t$. The $\tilde{a}_t$ process should be serially uncorrelated if the volatility model adequately captures the variations in asset return, our result indicates that after they are scaled by the fitted conditional volatility, the residuals are uncorrelated.

[9] The Quasi Maximum Likelihood estimation is also used, and it yields similar results.

[10] We performed formal test for the skewness. In the case of the GARCH model, the skew parameter is 0.8487 with a standard error of 0.0248. The t ratio is then $(0.8487-1)/0.0248 = -6.1$ with p-value < 0.01. Consequently, the null hypothesis of no skew is rejected. The same rejection decisions are for TGARCH and APARCH models. TGARCH has skew parameter 0.8382 with se 0.2475 and APARCH has skew parameter 0.8357 with se 0.2473.

Table 1. Estimated parameters of GARCH, TARCH and APARCH. Table reports estimated parameters of GARCH, TARCH and APARCH for sample daily returns from 2002 to 2012. The maximum of likelihood estimation (MLE) method are used and we allow for three types of MLE functions to handle the skewness of innovation returns: Gaussian, Student t and skew Student t. The level of statistical significance is denoted by: *10%, **5%, ***1%.

	GARCH				
Parameters	**constant**	**alpha**	**beta**		
Gaussian	7.376E−07**	0.081***	0.916***		
Student t	3.87E−07*	0.061***	0.937***		
Skew Student t	4.70E−07**	0.066***	0.931***		
	TGARCH				
Parameters	**constant**	**alpha**	**beta**	**gamma**	
Gaussian	1.56E−06***	0.078***	0.893***	0.391***	
Student t	6.13E−07**	0.053***	0.932***	0.349***	
Skew Student t	7.23E−07***	0.056***	0.927***	0.341***	
	APARCH				
Parameters	**constant**	**alpha**	**beta**	**gamma**	**delta**
Gaussian	8.35E−04***	0.093***	0.908***	0.701***	0.724***
Student t	3.46E−04***	0.081***	0.924***	0.626***	0.801***
Skew Student t	3.5E−04***	0.081***	0.923***	0.646***	0.820***

models are also fitted by maximising the Student-t and skewed Student-t likelihood functions.

4. Comparison Procedure

For comparing the predictive performance of various volatility forecasting models, the benchmark strategy involves a growing estimation window that employs all available data on and before each trading day in the forecasting sample and then fits the model by maximising the Gaussian likelihood function. In addition, instead of using a growing estimation window in the benchmark, forecasting strategies with a medium rolling window of three years and a short rolling window of one year are tested. Further, to handle the skewness of daily return, we also re-estimate and evaluate the models with Student-t and skew Student-t innovations. Overall, we compare three GARCH-type models with five prediction strategies.

We adopt the recursive forecasting method to obtain forecasted daily volatility. In the implementation, for each trading day t in the out-of-sample forecasting period we estimate GARCH class volatility models by using all the available daily data before date t. Then, for each GARCH-type model and each forecasting strategy, the fitted model is used to generate multiple forecasting horizons: one day, one week, two weeks and one month[11]. In this way, a series of overlapping

[11] Usually, one week has 5 trading days and one month has 22 trading days.

forecast paths is generated for each strategy. As the objective of the paper is to evaluate the empirical accuracy of each strategy, we compare the predicted daily volatility in the forecasting path with the high-frequency volatility, which is treated as a proxy for the true underlying return variation process. The high-frequency daily volatility is calculated by using both five-minute trading data (5-min RV) and the TSRV introduced in Section 2. As 5-min RV is noisy and TSRV is a robust and unbiased estimator of high-frequency volatility, our results mainly rely on the latter.

The comparison of empirical forecasting accuracy are based on loss functions, which examine the magnitude of the difference between the predicted volatility and the realized *ex-post* volatility proxy. The model with smaller forecasting loss is considered superior in predictability. We take the two loss functions suggested by [24] and [25]: Mean Squared Error (MSE) and Quasi-Like loss function (QL) which are defined as:

$$\begin{aligned}\text{MSE}\left(RV_{t+h},\theta^2_{t+h|t}\right)&=\left(RV_{t+h}-\theta^2_{t+h|t}\right)^2\\ \text{QL}\left(RV_{t+h},\theta^2_{t+h|t}\right)&=\frac{RV_{t+h}}{\theta^2_{t+h|t}}-\log\frac{RV_{t+h}}{\theta^2_{t+h|t}}-1\end{aligned}\tag{6}$$

where RV_{t+h} is the *ex-post* realized volatility at time $t+h$ and $\theta^2_{t+h|t}$ is the predicted corresponding value. Empirical studies of [22] [24] and [26] suggest that QL is a more robust function to compare forecasting loss, specifically for accuracy of comparisons across periods with different volatility levels. This is because the loss function of QL only depends on the scaled residual whereas MSE is determined by the additive errors. As a robustness test, we apply the Diebold-Mariano test[12] to statistically examine the performance of the loss functions (MSE and QL) for each forecasting strategy. The null hypothesis is the case that the two forecasts have the same accuracy and the alternative hypothesis is that one forecast has a higher level of accuracy.

5. Data Description

Daily Data and Tick Data

The total sample period is June 2002 - September 2013, and the out-of-sample period covers 2005-2013. The GARCH-type volatility models are estimated using the index daily data, and as a proxy for the true high-frequency daily volatility of ASX200 index, we use tick data for an exchange-traded index tracking fund (YSTW.AX) to calculate the 5-min RV and TSRV. Both the daily data and tick data are constructed from a database maintained by the Securities Industry Research Centre of Asia-Pacific, which has access to Thomson Reuters Tick History (TRTH). TRTH preserves Australian equities tick history from 1995.

The ASX uses a call auction procedure at opening and closing of each trading day. Therefore, to avoid abnormal trading patterns around the start and the end

[12] For details see [27].

of the trading day, trading data for the opening and closing call auctions are discarded. Only trading data collected during normal trading hours (10 am to 4 pm) are considered. We also exclude the overnight return. Due to the recording errors in tick data, the intra-day data for some trading dates are missing. We reconcile the tick data and daily data to match the volatility for each date. In the final form, our sample has 2791 trading days with 1,211,122 trading records.

6. Empirical Results

In this section, forecasting results for each separate GARCH-type model with modified strategies are reported. We also directly compare and test the forecasting accuracy of all three GARCH-type models. We examine whether the ranking of forecasting models changes during the GFC in 2008, and the results are shown in the Appendix.

Comparing Forecasting Accuracy for Various Volatility Forecasting Strategies

We begin comparing the out-of-sample prediction performance of various strategies with respect to the benchmark for the full period from 2005 to 2013. The benchmark is the GARCH-type model with the Gaussian innovation distribution estimated using all available data from 2002, and competing strategies are the modifications of the benchmark with a different forecasting window and innovations distributions. Both the 5-min RV and the more robust TSRV are calculated using intra-day trading data according to the model specifications in Section 2. They are treated as proxies for the underlying volatility process and used to yield prediction losses through comparison with the predicted value from each volatility forecasting model specification.

The forecasting performance of GARCH, TGARCH and APARCH volatility models, which are recursively estimated, are reported in Table 2 to Table 4 respectively. The prediction losses in the tables are based on two loss functions (QL and MSE) and two realized volatility measures (5-min RV and TSRV). The First column of each table under "forecasting strategies" shows the out-of-sample average losses of the benchmark forecasting strategy at multiple horizons (1, 5, 10 and 22 trading days). The second column through to the last column report relative forecasting performance, which are the average percentage gains or losses achieved by the modified prediction strategies with 1) Student t innovations, 2) skew Student-t innovations, 3) one-year rolling estimation windows and 4) three-year rolling estimation windows. For each positive percentage gain, we apply the Diebold-Mariano test to investigate whether the modification can significantly improve the forecasting accuracy with respect to the benchmark. As TSRV is less noisy than the 5-min RV and QL is a more robust loss function, our analyses are mainly concentrated on the results based on TSRV and QL. However, we report all of the results, allowing to examine how the conclusions are affected by using traditional MSE and 5-min RV.

Table 2. The Out-of-sample test for GARCH models. Table shows the evaluation of GARCH volatility forecast strategies for the period 2005-2013 . The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days) . From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

			Forecasting strategies				
Volatility proxy	Loss function	Forecasting horizons	Benchmark	Student *t*	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.815	−0.498	0.833**	−23.197	−23.001
		1 week	0.828	−0.262	1.27***	−21.286	−20.538
		2 weeks	0.844	0.030	1.631***	−19.272	−17.842
		1 month	0.862	0.134	2.144***	−17.292	−15.359
	MSE	1 day	1.005	−0.252	0.433	−9.550	−9.566
		1 week	1.010	−0.236	0.749	−9.751	−9.835
		2 weeks	1.017	−0.262	1.062	−10.107	−10.281
		1 month	1.031	−0.544	1.709	−10.776	−11.116
5 min-RV	QL	1 day	1.486	0.044	−0.379	−21.602	−24.467
		1 week	1.482	−0.105	−0.231	−23.443	−24.939
		2 weeks	1.494	0.229	−0.045	−24.009	−24.178
		1 month	1.511	0.985**	0.156	−25.729	−24.082
	MSE	1 day	9.851	0.002	−0.037	−0.302	−0.371
		1 week	9.851	0.002	−0.014	−0.530	−0.567
		2 weeks	9.854	0.006	0.011	−0.774	−0.781
		1 month	9.938	−0.022	0.053	−0.582	−0.534

It is apparent from Tables 2-4 that 5-min RV yields systematically higher losses than those of TSRV, as expected. This suggests that TSRV is a more precise proxy for the *ex-post* volatility than the sparsely sampled 5-min RV. This is because using sparsely sampled five-minute data discards too many high-frequency observations, along with the information they contain. Thus, we interpret the results relying on the out-of-sample forecasting losses yield from the TSRV. We note further from Tables 2-4 that except the standard GARCH model under skew Student t distribution, there are few systematic improvements in forecasting accuracy when the benchmark strategy is modified. In other words, the daily volatility forecasts obtained from the GARCH-type models estimated

Table 3. The out-of-sample test for TGARCH models. Table shows the evaluation of TGARCH volatility forecast strategies for the period 2005-2013. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation, 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.798	−0.858	0.013	−28.956	−19.365
		1 week	0.804	−0.744	−0.629	−25.492	−17.804
		2 weeks	0.798	−0.844	−1.688	−22.805	−18.231
		1 month	0.799	−0.396	−3.177	−20.229	−18.446
	MSE	1 day	1.037	1.848	2.193	−13.219	−13.349
		1 week	1.010	0.028	−0.3	−10.786	−10.882
		2 weeks	0.983	−1.494	−2.601	−8.197	−8.207
		1 month	0.933	−4.302	−6.844	2.767	−2.711
5 min-RV	QL	1 day	1.577	2.874	2.751**	−83.341	−60.638
		1 week	1.563	1.755	2.424	−78.156	−59.202
		2 weeks	1.573	1.046	2.098	−66.311	−54.694
		1 month	1.599	1.333	3.091	−54.484	−45.779
	MSE	1 day	9.921	0.506	0.48	−0.237	−0.124
		1 week	9.895	0.277	0.235	−0.659	−0.558
		2 weeks	9.875	0.104	0.057	−1.046	−0.983
		1 month	9.899	−0.177	−0.223	−1.342	−1.282

via Gaussian maximum likelihood using all available data set are not outperformed by the models which have modified innovation distributions and estimation windows. Particularly, when the models are estimated using Student-t likelihood, we do not observe significant improvements in forecasting performance. This may suggest that in the sample period, the distribution of the return does not dramatically violate the assumptions in the models. Although the results are not statistically significant, Student-t distribution seems to positively improve the forecasting accuracy for longer horizons. For GARCH with Student-t innovation in Table 2, shorter forecasting horizons (1 day and 1 week) have higher prediction losses than the benchmark, while the losses for longer horizons are

Table 4. The out-of-sample test for APARCH models. Table shows the evaluation of APARCH volatility forecast strategies for the period 2005-2013. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation, 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.7845	−0.479	0.07	−31.762	−24.622
		1 week	0.7934	0.065	0.296	−27.737	−20.698
		2 weeks	0.7919	1.076	0.949	−24.481	−19.202
		1 month	0.8134	3.159	2.861	−18.84	−14.713
	MSE	1 day	1.0104	−1.551	−0.644	−10.879	−11.09
		1 week	0.9711	−2.262	−1.353	−7.15	−7.377
		2 weeks	0.9404	−2.196	−1.378	−3.978	−4.177
		1 month	0.9117	−1.091	−0.599	0.422	0.606
5 min-RV	QL	1 day	1.5773	2.874**	2.751**	−83.341	−60.638
		1 week	1.5629	1.755	2.424	−78.156	−59.202
		2 weeks	1.5731	1.046	2.098	−66.311	−54.694
		1 month	1.5985	1.333	3.091	−54.484	−45.779
	MSE	1 day	9.9207	0.506	0.48	−0.237	−0.124
		1 week	9.8953	0.277	0.235	−0.659	−0.558
		2 weeks	9.8754	0.104	0.057	−1.046	−0.983
		1 month	9.8993	−0.177	−0.223	−1.342	−1.282

reduced. The applications of TGARCH and APARCH provide similar outcomes. The MSE loss provides mixed results, but they are not significant. Recall that MSE is rather noisy when the out-of-sample forecasting period is long and covers different volatility regimes. Overall, although Student-t distribution of innovations does not significantly improve the forecasting power of the three GARCH-type models, the prediction losses are alleviated when the forecasting horizon is longer. This is because the heavy tailed property is not generally disturbed by short run distribution of stock returns in the Australian equity market.

Our next step is to move to the models with skewed distribution of innovations. Table 2 shows that the GARCH model with skew Student-t likelihood significantly decreases QL losses when TSRV is used. The improvement is about 1% for the one day forecast horizon and 2% for one month ahead. This is consistent with the findings in Section 3 that shows the sample skewness is significantly negative. The skew distribution of innovations successfully captures this property and hence provides improved forecasts. We also note that the improvements are larger for longer forecasting horizons because of the increase in the skewness of returns over a longer period. Similar to the case of Student-t distribution, 5-min RV provides insignificant mixed results but they do not contradict the results of TSRV. The APARCH model provides similar but not statistically significant outcomes according to the Diebold-Mariano test (see Table 4). In Table 3, the TGARCH with skew innovation distribution appears to reduce forecasting accuracy for QL loss with TSRV proxy, compared with the benchmark. The possible reason is that the effect of negative skewness is incorporated in the standard TGARCH, because the TGARCH is designed to capture the heavier influence of past negative returns.

Using rolling estimation windows dramatically decreases forecasting performance. In the case of QL loss with TSRV proxy, the accuracy decreased by more than 20% and for other cases the maximum reduction is as much as 80%. The one-year rolling estimation window has the largest forecasting losses, but the medium three-year rolling window is better, however, it is still much worse than the full sample growing estimation window. This may suggest that the information contained in the short estimation period is not sufficient to capture the dynamic of the volatility process.

Finally we move to the direct comparison of the three benchmark GARCH models. Table 5 displays the forecast performance of each GARCH-type model in the sample period 2005-2013. As discussed earlier, the benchmark strategy for each model works well so only the loss functions for the benchmark and for each forecast horizon are shown in Table 5. The best performance which has the smallest prediction loss is highlighted. We also use the Diebold-Mariano test to evaluate the significance of the improved forecast accuracy. From the results, we see that both the TGARCH and APARCH have significantly better prediction performance than GARCH. Further, it is usually observed that APARCH gives the best forecast. In other words the asymmetric specification captures the asymmetry in return volatility in the Australian equity market. The MSE results favor the GARCH for one-day ahead forecast, but this should not be considered a discrepancy because the result is not significant and the losses are very similar.

We also examine the performance of our selected volatility forecasting strategies during the recent financial turmoil period. The same procedure above is applied to the period of the GFC in 2008. We find that during financial turmoil, the forecast loss of each model is significantly increased, however, the overall ranking of the models does not change. The details and results are provided in Appendix.

Table 5. Comparison of three standard GARCH class models. Table shows the direct comparison of GARCH models in predicting ASX200 daily volatility for the period 2005-2013. The numbers in the table are the forecasting losses for each loss function Asterisks represent the significance of the Diebold-Mariano test. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Loss function	Forecast horizons	Volatility model		
		GARCH	TGARCH	APARCH
QL	1 day	0.8148	0.7979*	**0.7845**
	1 week	0.8279	0.8036**	**0.7934***
	2 weeks	0.8436	0.7982***	**0.7919****
	1 month	0.8616	**0.7982*****	0.8134
MSE	1 day	**1.0053**	1.0374	1.0104
	1 week	1.0097	1.0104	**0.9711***
	2 weeks	1.0166	0.9833*	**0.9404****
	1 month	1.0313	0.9333*	**0.9117***

7. Conclusion

Volatility models are constructed to predict volatility, which is an essential input in financial asset pricing models and risk management practice. In this paper, we empirically investigate the predictive ability of various volatility forecasting strategies employing GARCH class models in the Australian equity market that has several distinguishing features. We specifically examine which specific features of GARCH models provide the best forecast; whether we should allow for heavy-tailed and skew distribution of innovations in our estimation; and whether we should use growing estimation or rolling estimation windows. Because of the crucial role of *ex-post* volatility estimators in out-of-sample tests for predictive ability of volatility models—along with the realized volatility constructed using five-minute intra-day data that are arbitrary, subjective and biased, our analyses rely more on an unbiased volatility estimator—the TSRV, which utilizes ultra-high-frequency intra-day data.

In the pure out-of-sample test, we evaluate the predictive abilities of volatility forecasting strategies involving three popularly used GARCH-type models and their modifications. Each strategy is compared with the corresponding benchmark standard GARCH class model. The forecasting accuracy is based on two commonly used loss functions: the Mean Square Error (MSE) and the Quasi-like Loss function (QL). For robustness test purposes, we use Diebold-Mariano statistic to examine the significance of forecasting accuracy. Our results suggest that, in the Australian stock market, modifications to the GARCH class models usually do not significantly outperform the associated standard model with normally distributed innovations and a growing estimation window. However, Student-t and skew Student-t distribution seems to alleviate the forecasting losses for all three GARCH-type models when forecasting horizon is increasing. With regard to the estimation window, the results show that growing estimation windows significantly outperform rolling estimation windows, which suggests that the information incorporated in a relative short estimation window is not sufficient to capture the time-varying volatility process.

Further, in the direct comparisons to the benchmark models, the APARCH model typically provides the best forecast in the out-of sample period of 2005-2013. By applying the same evaluation procedure to a period of financial turmoil, the 2008 GFC, we find that while the predictive abilities of volatility models are reduced, their ranking remains the same. Overall, the APARCH model provides the best forecast across all cases, which demonstrates the ability of the APARCH model in capturing the leptokurtic returns and other stylized facts of volatility in the Australian stock market.

References

[1] Engle, R.F. (1982) Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*: *Journal of the Econometric Society*, **50**, 987-1007.

[2] Bollerslev, T. (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

[3] Bollerslev, T. (2009) Glossary to Arch (Garch). In: Bollerslev, T., Russell, J. and Watson, M., Eds., *Volatility and Time Series Econometrics*, Oxford University Press, Oxford.

[4] 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. *The Journal of Finance*, **48**, 1779-1801.

[5] Zakoian, J.-M. (1994) Threshold Heteroskedastic Models. *Journal of Economic Dynamics and Control*, **18**, 931-955. https://doi.org/10.1016/0165-1889(94)90039-6

[6] Ding, Z., Granger, C.W.J. and Engle, R.F. (1993) A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance*, **1**, 83-106. https://doi.org/10.1016/0927-5398(93)90006-D

[7] Andersen, T.G. and Bollerslev, T. (1998) Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review*, **39**, 885-905.

[8] Cumby, R., Figlewski, S. and Hasbrouck, J. (1993) Forecasting Volatilities and Correlations with Egarch Models. *The Journal of Derivatives*, **1**, 51-63. https://doi.org/10.3905/jod.1993.407877

[9] Figlewski, S. (1997) Forecasting Volatility. *Financial Markets*, *Institutions & Instruments*, **6**, 1-88.

[10] Jorion, P. (1995) Predicting Volatility in the Foreign Exchange Market. *The Journal of Finance*, **50**, 507-528.

[11] Zhang, L., Mykland, P.A. and Ait-Sahalia, Y. (2005) A Tale of Two Time Scales. *Journal of the American Statistical Association*, **100**, 1394-1411.

[12] Aıt-Sahalia, Y., Mykland, P.A. and Zhang, L. (2011) Ultra High Frequency Volatility Estimation with Dependent Microstructure Noise. *Journal of Econometrics*, **160**,

160-175. https://doi.org/10.1016/j.jeconom.2010.03.028

[13] Brailsford, T.J. and Faff, R.W. (1996) An Evaluation of Volatility Forecasting Techniques. *Journal of Banking & Finance*, **20**, 419-438. https://doi.org/10.1016/0378-4266(95)00015-1

[14] Carvajal, M., Coulton, J.J. and Jackson, A.B. (2017) Earnings Benchmark Hierarchy. *Accounting & Finance*, **57**, 87-111.

[15] Merton, R.C. (1980) On Estimating the Expected Return on the Market: An Exploratory Investigation. *Journal of Financial Economics*, **8**, 323-361. https://doi.org/10.1016/0304-405X(80)90007-0

[16] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. (2001) The Distribution of Realized Stock Return Volatility. *Journal of Financial Economics*, **61**, 43-76.

[17] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2001) The Distribution of Realized Exchange Rate Volatility. *Journal of the American Statistical Association*, **96**, 42-55.

[18] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003) Modeling and Forecasting Realized Volatility. *Econometrica*, **71**, 579-625.

[19] Barndorff-Nielsen, O.E, Hansen, P.R., Lunde, A. and Shephard, N. (2008) Designing Realized Kernels to Measure the ex Post Variation of Equity Prices in the Presence of Noise. *Econometrica*, **76**, 1481-1536.

[20] Barndorff-Nielsen, O.E, Hansen, P.R., Lunde, A. and Shephard, N. (2011) Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading. *Journal of Econometrics*, **162**, 149-169. https://doi.org/10.1016/j.jeconom.2010.07.009

[21] Zhang, L., *et al.* (2006) Efficient Estimation of Stochastic Volatility using Noisy Observations: A Multi-Scale Approach. *Bernoulli*, **12**, 1019-1043. https://doi.org/10.3150/bj/1165269149

[22] Hansen, P.R. and Lunde, A. (2005) A Forecast Comparison of Volatility Models: Does Anything Beat a Garch (1, 1)? *Journal of Applied Econometrics*, **20**, 873-889.

[23] Hentschel, L. (1995) All in the Family Nesting Symmetric and Asymmetric Garch Models. *Journal of Financial Economics*, **39**, 71-104. https://doi.org/10.1016/0304-405X(94)00821-H

[24] Patton, A.J. and Sheppard, K. (2009) Optimal Combinations of Realised Volatility Estimators. *International Journal of Forecasting*, **25**, 218-238. https://doi.org/10.1016/j.ijforecast.2009.01.011

[25] Patton, A.J. (2011) Data-Based Ranking of Realised Volatility Estimators. *Journal of Econometrics*, **161**, 284-303. https://doi.org/10.1016/j.jeconom.2010.12.010

[26] Brownlees, C., Engle, R. and Kelly, B. (2012) A Practical Guide to Volatility Forecasting through Calm and Storm. *Journal of Risk*, **14**, 3-22. https://doi.org/10.21314/JOR.2012.237

[27] Diebold, F.X. and Mariano, R.S. (2002) Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, **20**, 339-350.

Appendix

We examine the forecast ability of GARCH models during the recent GFC with a high volatility level. Figure A1 plots the time series of the daily return, as expected, the variation in asset return is dramatically high during the period of high turmoil around autumn of 2008. We choose the test period of three months beginning 15 September, 2008 when Lehman Brothers Holdings filed for bankruptcy protection. We selected the three-months period from April to July 2008 as a control period. A similar procedure to the full sample period is also applied to these periods. The annualized unconditional daily variance for the GFC period is 0.224, which is four times as large as the variance in the control period with daily volatility of 0.055.

Tables A1-A3 report the forecast losses for the control period April- July 2008. We repeat the procedure for the turmoil period September - December 2008, and the results are presented in Tables A4-A6. From the "benchmark" column in each table, we can see that the forecast losses are dramatically increased during the GFC. For instance, the TGARCH benchmark strategy has QL losses of 0.4 for the second quarter of 2008, which rises to 0.6 in the autumn. However, the results show a similar tendency as in the full sample period (2005-2013). While the benchmark strategies across all three GARCH models are seldom beaten by their corresponding variation, skewed Student-t likelihood can significantly improve the forecast accuracy of each GARCH model. We also observe that during the GFC, the 1-year rolling estimation window improves the forecast for longer horizons. This is because historical data from earlier periods have less prediction power during financial turmoil. However, when the higher effect of negative returns is considered in the TGARCH and APARCH models, improved performances disappear.

Table A7 provides the result of the direct comparison of GARCH class models during the normal and turmoil periods. Similar to the full sample period, the TGARCH and APARCH models perform better than the GARCH model. The APARCH model usually provides the best forecast accuracy, particularly for the longer forecast horizons during the normal period in 2008. However, the model wins across all cases for the GFC period. This reflects the important role of higher negative returns in predicting future volatility during times of financial turmoil. Overall, although the forecast accuracies are reduced during the GFC, the ranking of the forecasting models is unchanged.

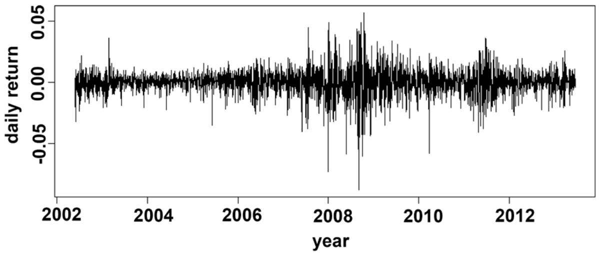

Figure A1. Plot of daily log return od ASX200.

Table A1. GARCH for April-July, 2008. Table shows the evaluation of GARCH volatility forecast strategies for the period April-July 2008. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days) . From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.5068	−5.323	0.451**	−30.251	−21.495
		1 week	0.5870	−5.618	1.218*	−24.998	−18.034
		2 weeks	0.5712	−6.213	1.933**	−24.533	−20.055
		1 month	0.6293	−8.754	3.35*	−17.483	−18.717
	MSE	1 day	1.5643	−9.459	1.235	−0.483	10.92
		1 week	1.6720	−9.95	3.204	2.23	11.267
		2 weeks	1.6895	−11.138	4.95	4.052	9.82
		1 month	1.6975	−15.955	8.134	11.706*	−4.969
5 min-RV	QL	1 day	0.9963	1.049	0.732**	−16.728	−19.726
		1 week	0.9315	0.478	0.665*	−39.133	−42.923
		2 weeks	0.8144	−0.481	0.334	−90.508	−93.362
		1 month	0.6215	−2.186	−0.606	−6.255	−5.975
	MSE	1 day	1.5030	−0.427	0.2	−26.786	−26.065
		1 week	1.5514	−0.359	0.291	−27.867	−27.176
		2 weeks	1.6057	−0.225	0.239	−29.666	−29.857
		1 month	1.7982	0.205	−0.341	−33.833	−33.809

Table A2. TGARCH for April-July, 2008. Table shows the evaluation of TGARCH volatility forecast strategies for the period April-July 2008. The 1st column under the “forecasting strategies” reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.4421	−15.667	−12.719	−65.756	−81.032
		1 week	0.4717	−19.015	−18.624	−66.674	−80.018
		2 weeks	0.4840	−24.036	−27.477	−65.626	−76.272
		1 month	0.4917	−48.699	−66.562	−44.2	−56.165
	MSE	1 day	1.8865	2.492	5.357	96.78	−97.908
		1 week	1.8353	−3.228	−4.973	−96.338	−97.496
		2 weeks	1.7685	−10.768	−19.692	−95.575	−96.68
		1 month	1.5344	−50.798	−89.796	−87.009	−89.398
5 min-RV	QL	1 day	1.4674	20.057	20.292	−13.929	−12.667
		1 week	1.5337	24.06 *	24.85 *	−9.451	−15.967
		2 weeks	1.6574	20.839	21.197	−35.704	−5.862
		1 month	1.8378	19.671	18.629	−120.186	−87.524
	MSE	1 day	1.6055	2.111	2.376	10.813	11.67
		1 week	1.6543	2.027	1.921	8.585	9.486
		2 weeks	1.7113	1.991	1.502	4.921	5.833
		1 month	1.9250	3.15	2.921	−3.534	−2.857

Table A3. APARCH for April-July, 2008. Table shows the evaluation of APARCH volatility forecast strategies for the period April-July 2008. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.4432	−8.571	−6.876	−65.999	−81.579
		1 week	0.4457	−10.853	−9.564	−64.731	−79.167
		2 weeks	0.4622	−13.432	−12.965	−60.703	−72.99
		1 month	0.4833	−18.974	−22.061	−25.437	−41.58
	MSE	1 day	1.6543	−2.295	−0.194	−96.629	−97.851
		1 week	1.7242	−10.322	−8.475	−95.281	−96.808
		2 weeks	1.5608	−18.783	−17.57	−92.897	−94.67
		1 month	1.4191	−39.502	−44.297	−62.861	−69.566
5 min-RV	QL	1 day	1.5197	11.494	11.018	−10.007	−19.045
		1 week	1.5282	9.83	9.551	−9.844	−20.109
		2 weeks	1.3615	10.428	10.384	−25.327	8.744
		1 month	9.9484	13.84	12.694	−85.425	−46.436
	MSE	1 day	15.8919	0.205	0.388	9.895	10.91
		1 week	16.1981	−0.021	0.076	6.641	7.776
		2 weeks	16.7667	0.067	0.105	2.958	4.183
		1 month	19.3395	1.149	1.04	−3.054	−2.033

Table A4. GARCH for Sep-Dec, 2008. Table shows the evaluation of GARCH volatility forecast strategies for the period September-December 2008. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.6493	−0.623	1.325***	−2.841	−9.433
		1 week	0.6663	−0.631	1.851*	−1.889	−1.887
		2 weeks	0.6955	−0.764	1.537	12.977**	−3.717
		1 month	0.6840	−1.598	2.848	9.392*	4.228
	MSE	1 day	2.9401	−1.867	1.957 ***	2.62	1.637
		1 week	3.1413	−1.983	3.457 *	2.695	1.847
		2 weeks	3.4482	−2.174	4.871 *	2.739	2.041
		1 month	3.5297	−3.586	9.396	5.238	−4.887
5 min-RV	QL	1 day	0.6850	−0.547	1.251***	7.774	−2.067
		1 week	0.7117	−0.603	1.651**	6.831**	0.412
		2 weeks	0.7367	−0.764	1.464**	4.909*	1.715
		1 month	0.7123	−1.633	2.856*	3.399	2.011
	MSE	1 day	3.0498	−1.827	1.92**	2.725	1.774
		1 week	3.2620	−1.971	3.36***	2.972	2.199
		2 weeks	3.6500	−2.196	4.739*	3.439	2.819
		1 month	3.5922	−3.606	9.288*	6.131**	5.817*

Table A5. TGARCH for Sep-Dec, 2008. Table shows the evaluation of TGARCH volatility forecast strategies for the period September-December 2008. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation, 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.6189	−5.39	−4.476	−28.355	−14.436
		1 week	0.6291	−7.044	−7.98	−29.191	−15.4
		2 weeks	0.6163	−8.719	−11.793	−28.563	−11.206
		1 month	0.6579	−18.668	−27.933	−23.379	−15.526
	MSE	1 day	3.2443	9.824*	1.564	−9.618	9.517
		1 week	3.3296	0.017	−2.526	−8.274	9.236
		2 weeks	2.9432	−11.195	−20.364	2.376	8.404
		1 month	3.4377	−4.989	−84.539	5.731	6.871
5 min-RV	QL	1 day	0.6543	−5.176	−4.358	−28.812	−15.194
		1 week	0.6733	−6.6	−7.489	−25.792	−13.192
		2 weeks	0.6572	−8.208	−11.165	−23.668	−13.286
		1 month	0.7182	−18.297	−27.722	−14.162	−14.506
	MSE	1 day	3.0599	9.526	11.244***	9.816	9.675
		1 week	3.4573	0.116	−2.374	9.83	9.722
		2 weeks	3.8544	−10.773	−19.762	8.645	8.577
		1 month	3.7833	−48.859	−83.002	8.129	9.15

Table A6. APGARCH for Sep-Dec, 2008. Table shows the evaluation of APARCH volatility forecast strategies for the period September-December 2008. The 1st column under the "forecasting strategies" reports the out-of-sample losses (QL and MSE) of the benchmark forecasting strategy at multi-step ahead (1, 5, 10 and 22 trading days). From the 2nd through the last column report the percentage gains or losses achieved by modifying the estimation strategy with 1) Student t innovations, 2) skew student t innovation, 3) 1 year rolling estimation window, 4) 3 year rolling estimation window. Asterisks after the percentage gains represent the significance of the Diebold-Mariano test for whether the modification can improve the forecasting accuracy. The level of statistical significance is denoted by: *10%, **5%, ***1%.

Volatility proxy	Loss function	Forecasting horizons	Forecasting strategies				
			Benchmark	Student t	Skew student	1-year	3-year
				innovation	t innovation	rolling window	rolling window
TSRV	QL	1 day	0.528	−11.349	−8.842	−3.229	−16.295
		1 week	0.5373	−15.271	−12.04	−37.634	−202.4
		2 weeks	0.5759	−20.065	−15.754	−44.734	−39.273
		1 month	0.5894	−40.855	−30.557	−46.463	−31.089
	MSE	1 day	2.1989	−14.535	−7.448	9.018	90.127
		1 week	2.2755	−30.592	−20.008	4.304	85.811
		2 weeks	2.5016	−51.161	−35.457	5.254	77.474
		1 month	2.8451	−21.477	−83.595	4.29	49.634
5 min-RV	QL	1 day	0.5134	−10.94	−8.592	−25.511	−15.824
		1 week	0.5815	−14.272	−11.271	−15.608	−16.175
		2 weeks	0.6121	−19.205	−15.066	−33.142	−11.552
		1 month	0.3018	−40.637	−30.276	−37.827	−24.413
	MSE	1 day	3.3016	−14.434	−7.444	9.263	9.324
		1 week	3.3737	−29.977	−19.611	5.186	6.58
		2 weeks	3.5766	−49.899	−34.552	7.999	7.986
		1 month	3.6146	−19.794	−82.378	4.126	5.856

Table A7. Comparing forecast ability of GARCH models before and during the 2008 financial crisis. Table shows the direct comparison of GARCH models in predicting ASX200 daily volatility before and during the 2008 financial crisis. The numbers in the table are the forecasting losses for each loss function Asterisks represent the significance of the Diebold-Mariano test. The level of statistical significance is denoted by: *10%, **5%, ***1%.

			April-July, 2008		
Proxy	Loss function	Forecast horizons	Volatility model		
			GARCH	TGARCH	APARCH
TSRV	QL	1 day	0.5068	0.4421**	0.4432**
		1 week	0.5870	0.4717	**0.4457**
		2 weeks	0.5712	0.4840	**0.4622**
		1 month	0.6293	0.4917	0.4833 *
	MSE	1 day	**1.5643**	1.8865	1.6543
		1 week	**1.6720**	1.8353	1.7242
		2 weeks	1.6895	1.7685	**1.5608**
		1 month	1.6975	1.5344	1.4191 *
			September-December, 2008		
			GARCH	TGARCH	APARCH
TSRV	QL	1 day	0.6493	0.6189*	0.528***
		1 week	0.6663	0.6291	0.5373**
		2 weeks	0.6955	0.6163	0.5759**
		1 month	0.6840	0.6579 **	0.5894**
	MSE	1 day	2.9401	3.2443	**2.1989**
		1 week	3.1413	3.3296	2.2755**
		2 weeks	3.4482	2.9432	2.5016***
		1 month	3.5297	3.1377 **	2.8451**