Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Forecasting 20 (2004) 629–645
international journal of forecasting
www.elsevier.com/locate/ijforecast

Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood

Gloria González-Rivera[1], Tae-Hwy Lee*, Santosh Mishra[2]

Department of Economics, University of California, Riverside, Riverside, CA 92521-0427, USA

Abstract

We analyze the predictive performance of various volatility models for stock returns. To compare their performance, we choose loss functions for which volatility estimation is of paramount importance. We deal with two economic loss functions (an option pricing function and an utility function) and two statistical loss functions (a goodness-of-fit measure for a value-at-risk (VaR) calculation and a predictive likelihood function). We implement the tests for superior predictive ability of White [Econometrica 68 (5) (2000) 1097] and Hansen [Hansen, P. R. (2001). An unbiased and powerful test for superior predictive ability. Brown University]. We find that, for option pricing, simple models like the Riskmetrics exponentially weighted moving average (EWMA) or a simple moving average, which do not require estimation, perform as well as other more sophisticated specifications. For a utility-based loss function, an asymmetric quadratic GARCH seems to dominate, and this result is robust to different degrees of risk aversion. For a VaR-based loss function, a stochastic volatility model is preferred. Interestingly, the Riskmetrics EWMA model, proposed to calculate VaR, seems to be the worst performer. For the predictive likelihood-based loss function, modeling the conditional standard deviation instead of the variance seems to be a dominant modeling strategy.

Keywords: ARCH; Data snooping; Option pricing; Predictive likelihood; Reality check; Superior predictive ability; Utility function; VaR; Volatility

1. Introduction

During the last two decades, volatility has been one of the most active areas of research in time series econometrics. Volatility research has not been just limited to the area of time series econometrics dealing with issues of estimation, statistical inference, and model specification. More fundamentally, volatility research has contributed to the understanding of important issues in financial economics such as portfolio allocation, option pricing, and risk management. Volatility, as a measure of uncertainty, is of most interest to economists and, in particular, to those interested in decision making under uncertainty.

The development of volatility models has been a sequential exercise. Surveys as in Bera and Higgins (1993), Bollerslev, Chou, and Kroner (1992), Bol-

* Corresponding author. Tel.: +1-909-827-1509; fax: +1-909-787-5685.
E-mail addresses: gloria.gonzalez@ucr.edu (G. González-Rivera), tae.lee@ucr.edu (T.-H. Lee), santos_28@hotmail.com (S. Mishra).
[1] Tel.: +1-909-827-1590; fax: +1-909-787-5685.
[2] Tel.: +1-909-827-3266; fax: +1-909-787-5685.

0169-2070/$ - see front matter © 2004 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2003.10.003

lerslev, Engle, and Nelson (1994), and Poon and Granger (2003) attest to the variety of issues in volatility research. As a starting point, a volatility model should be able to pick up the stylized facts that we frequently encounter in financial data. The motivation for the introduction of the first generation of ARCH models (Engle, 1982) was to account for clusters of activity and fat-tail behavior of the data. Subsequent models accounted for more complex issues. Among others and without being exclusive, we should mention issues related to asymmetric responses of volatility to news, distribution of the standardized innovation, i.i.d. behavior of the standardized innovation, persistence of the volatility process, linkages with continuous time models, intraday data and unevenly spaced observations, seasonality and noise in intraday data. The consequence of this research agenda has been a vast array of specifications for the volatility process.

When the researcher and/or the practitioner faces so many models, the natural question becomes which one to choose. There is not a universal answer to this question. The best model depends upon the objectives of the researcher. Given an objective function, we look for best predictive ability while controlling for possible biases due to "data snooping" (Lo & MacKinlay, 1999).

The literature that compares the relative performance of volatility models is either centered around a statistical loss function or an economic loss function. The preferred statistical loss functions are based on moments of forecast errors (mean-error, mean-squared error, mean absolute error, etc.). The best model minimizes a function of the forecast errors. The volatility forecast is often compared to a measure of realized volatility. With financial data, the common practice has been to take squared returns as a measure of realized volatility. However, this practice is questionable. Andersen and Bollerslev (1998) argued that this measure is a noisy estimate and proposed the use of the intraday (at each 5 min interval) squared returns to calculate the daily realized volatility. This measure requires intraday data, which is subject to the variation introduced by the bid-ask spread and the irregular spacing of the price quotes.

Some authors have evaluated the performance of volatility models with criteria based on economic loss functions. For example, West, Edison, and Cho (1993) considered the problem of portfolio allocation based on models that maximize the utility function of the investor. Engle, Kane, and Noh (1997) and Noh, Engle, and Kane (1994) considered different volatility forecasts to maximize the trading profits in buying/selling options. Lopez (2001) considered probability scoring rules that were tailored to a forecast user's decision problem and confirmed that the choice of loss function directly affected the forecast evaluation of different models. Brooks and Persand (2003) evaluated volatility forecasting in a financial risk management setting in terms of value-at-risk (VaR). The common feature to these branches of the volatility literature is that none of these has controlled for forecast dependence across models and the inherent biases due to data-snooping. Our paper fills this void.

We consider 15 volatility models for the daily S&P500 index that are evaluated according to their out-of-sample forecasting ability. Our forecast evaluation is based on two economic loss functions, an option pricing formula and a utility function, and two statistical loss functions, a goodness-of-fit based on a VaR calculation and the predictive likelihood function. For option pricing, volatility is a key ingredient. Our loss function assess the difference between the actual price of a call option and the estimated price, which is a function of the estimated volatility of the stock. Our second economic loss function refers to the problem of wealth allocation. An investor wishes to maximize her utility allocating wealth between a risky asset and a risk-free asset. Our loss function assesses the performance of the volatility estimates according to the level of utility they generate. The statistical function based on the goodness-of-fit of a VaR calculation is important for risk management. The main objective of VaR is to calculate extreme losses within a given probability of occurrence and the estimation of the volatility is central to the VaR measure.

To control for the fact that as the number of models increases, so does the probability of finding superior predictive ability among the collection of models, we implement the "reality check" of White (2000). A problem associated with White's reality check is that the power of the test is sensitive to the inclusion of a poor model. The test is conservative in that the null hypothesis, which involves a benchmark model, is designed to be the least favorable to the alternative hypothesis. Hence, the inclusion of a bad model adversely affects the power of the reality check test. In this instance, the benchmark model may hardly be dominated. Hansen (2001) addressed this issue by suggesting a modification to the White's test. In our paper, we also implement Hansen's modification.

Concurrently and independently, Hansen and Lunde (2002) have also examined the predictive ability of volatility forecasts for the Deutsche Mark/US Dollar exchange rate and IBM stock prices with Whites reality check test. The main difference between their work and ours is the choice of loss functions and the data set. They have formed statistical loss functions where realized volatility is proxied by the mean of intraday squared returns as suggested in Andersen and Bollerslev (1998). None of their statistical loss functions include either a VaR goodness-of-fit or a predictive likelihood function. Our results are also very different. Hansen and Lunde claimed that the GARCH(1,1) model was not dominated by any other model. More recently, Awartani and Corradi (2003) have provided a comparison of the relative out-of-sample ability of various volatility models, with particular attention to the role of asymmetries. They show that while the true underlying volatility process is unobservable, using squared returns may be used as a valid proxy in assessing the relative predictive performance of various volatility models.

We claim that the preferred models depend very strongly upon the loss function chosen by the researcher. We find that, for option pricing, simple models such as the exponential weighted moving average (EWMA) proposed by Riskmetrics performed as well as any GARCH model. For an utility loss function, an asymmetric quadratic GARCH model is the most preferred. For VaR calculations, a stochastic volatility model dominates all other models. And, for a predictive likelihood function, modeling the conditional standard deviation instead of the variance results in a dominant model.

The organization of the paper is as follows. In Section 2, we present various volatility models. In Section 3, we discuss the White's reality check and the Hansen's modification. In Section 4, we present the loss functions. In Section 5, we explain our results and, in Section 6, we conclude.

2. Volatility models

In this section, we present various volatility models developed over the last two decades. To establish notation, suppose that the return series $\{y_t\}_{t=1}^{T+1}$ of a financial asset follows the stochastic process $y_{t+1} = \mu_{t+1} + \varepsilon_{t+1}$, where $E(y_{t+1}|\mathscr{F}_t) = \mu_{t+1}(\theta)$ and $E(\varepsilon^2_{t+1}|\mathscr{F}_t) = \sigma^2_{t+1}(\theta)$ given the information set $\mathscr{F}_t$ (σ-field) at time t. Let $z_{t+1} \equiv \varepsilon_{t+1}/\sigma_{t+1}$ have the conditional normal distribution with zero conditional mean and unit conditional variance. In Table 1, we summarize the models considered in this paper and introduce further notation.

These models can be classified in three categories: MA family, ARCH family, and stochastic volatility (SV) family.

First, the simplest method to forecast volatility is to calculate a historical moving average variance, denoted as MA(m), or an exponential weighted moving average (EWMA). In the empirical section where we deal with daily data, we set $m = 20$ and we follow Riskmetrics (1995) for the EWMA specification with $\lambda = 0.94$. For these two MA family models, there is no parameters to estimate.

Second, the ARCH family consists of the following models: ARCH(p) of Engle (1982); GARCH model of Bollerslev (1986); Integrated GARCH (I-GARCH) of Engle and Bollerslev (1986); Threshold GARCH (T-GARCH) of Glosten, Jaganathan, and Runkle (1993); Exponential GARCH (E-GARCH) of Nelson (1991); three variations of quadratic GARCH models (Q-GARCH), namely, Q-GARCH$_1$ of Sentana (1995), Q-GARCH$_2$ and Q-GARCH$_3$ of Engle and Ng (1993); Absolute GARCH (ABS-GARCH) of Taylor (1986) and Schwert (1990); Logarithmic GARCH (LOG-GARCH) of Geweke (1986) and Pantula (1986); Asymmetric GARCH (A-GARCH) of Zakonian (1994); and Smooth Tran-

Table 1
Volatility models

Name	Model
MA(m)	$\sigma_t^2 = 1/m \sum_{j=1}^{m} (y_{t-j} - \hat{\mu}_t^m)^2, \quad \hat{\mu}_t^m = 1/m \sum_{j=1}^{m} y_{t-j}$
EWMA	$\sigma_t^2 = (1-\lambda) \sum_{j=1}^{t-1} \lambda^{j-1} (y_{t-j} - \hat{\mu}_t)^2,$ $\hat{\mu}_t = 1/(t-1) \sum_{j=1}^{t-1} y_{t-j}, \quad \lambda = 0.94$
ARCH(p)	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2$
GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2$
I-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2,\ \alpha + \beta = 1$
T-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 \mathbf{1}\ (\varepsilon_{t-1} \geq 0)$
ST-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 F(\varepsilon_{t-1}, \delta),$ $F(\varepsilon_{t-1}, \delta) = [1 + \exp(\delta\varepsilon_{t-1})]^{-1} - 0.5$
E-GARCH	$\ln \sigma_t^2 = \omega + \beta \ln \sigma_{t-1}^2 + \alpha[\,\lvert z_{t-1}\rvert - cz_{t-1}]$
Q-GARCH$_1$	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha(\varepsilon_{t-1} + \gamma)^2$
Q-GARCH$_2$	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha(\varepsilon_{t-1} + \gamma\sigma_{t-1})^2$
Q-GARCH$_3$	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha(z_{t-1} + \gamma)^2$
ABS-GARCH	$\sigma_t = \omega + \beta\sigma_{t-1} + \alpha\,\lvert\varepsilon_{t-1}\rvert$
LOG-GARCH	$\ln\sigma_t = \omega + \beta \ln\sigma_{t-1} + \alpha\,\lvert\varepsilon_{t-1}\rvert$
A-GARCH	$\sigma_t = \omega + \beta\sigma_{t-1} + \alpha_1\varepsilon_{t-1}^{+} - \alpha_2\varepsilon_{t-1}^{-}$
SV	$\sigma_t^2 = \exp(0.5h_t),\ \ln(y_t^2) = -1.27 + h_t + \xi_t,$ $h_t = \gamma + \phi h_{t-1} + \eta_t$

(1) MA = moving average, EWMA = exponentially weighted MA, I-GARCH = integrated GARCH, T-GARCH = threshold GARCH, E-GARCH = exponential GARCH, Q-GARCH = quadratic GARCH, ABS-GARCH = absolute GARCH, LOG-GARCH = logarithmic GARCH, A-GARCH = asymmetric GARCH, ST-GARCH = smooth transition GARCH, SV = stochastic volatility; (2) $\mathbf{1}(\cdot)$ is an indicator function. For A-GARCH, $\alpha_1, \alpha_2 > 0$, $\varepsilon^+ = \max(\varepsilon, 0)$, and $\varepsilon^- = \min(\varepsilon, 0)$. For ST-GARCH, we fix $\delta = 3$ to ease the convergence in estimation. For SV, η_t is i.i.d. $N(0, \sigma_\eta^2)$ and ξ_t is i.i.d. $N(0, \pi^2/2)$.

sition GARCH (ST-GARCH) of González-Rivera (1998).

The EWMA specification can be viewed as an integrated GARCH model with $\omega = 0$, $\alpha = \lambda$, and $\beta = 1 - \lambda$. In the T-GARCH model, the parameter γ allows for possible asymmetric effects of positive and negative innovations. In Q-GARCH models, the parameter γ measures the extent of the asymmetry in the news impact curve. For the A-GARCH model, α_1, $\alpha_2 > 0$, $\varepsilon^+ = \max(\varepsilon, 0)$, and $\varepsilon^- = \min(\varepsilon, 0)$. For the ST-GARCH model, the parameter γ measures the asymmetric effect of positive and negative shocks, and the parameter $\delta > 0$ measures the smoothness of the transition between regimes, with a higher value of δ making ST-GARCH closer to T-GARCH. We fix $\delta = 3$ to ease the convergence in estimation.[3]

Third, for the SV family, we consider the stationary SV model of Taylor (1986) where η_t is i.i.d. $N(0, \sigma_\eta^2)$ and ξ_t is i.i.d. $N(0, \pi^2/2)$. This model is estimated by quasi-maximum likelihood (QML) method by treating ξ_t as though it were i.i.d. $N(0, \pi^2/2)$. The Kalman filter is used to obtain the Gaussian likelihood, which is numerically maximized. Ruiz (1994) showed that QML estimation within the Kalman filter algorithm works well.

3. Reality check

Consider various volatility models and choose one as a benchmark. For each model, we are interested in the out-of-sample one-step ahead forecast. This forecast will be fed into an objective function, for instance, a utility function or a loss function. Our interest is to compare the utility (loss) of each model to that of the benchmark model. We formulate a null hypothesis where the model with the largest utility (smallest loss) is not any better than the benchmark model. If we reject the null hypothesis, there is at least one model that produces more utility (less loss) than the benchmark.

Formally, the testing proceeds as follows. Let l be the number of competing volatility models ($k = 1, \ldots, l$) to compare with the benchmark volatility model (indexed as $k = 0$). For each volatility model k, one-step predictions are to be made for P periods from R through T, so that $T = R + P - 1$. As the sample size T increases, P and R may increase. For a given volatility model k and observations 1 to R, we estimate the parameters of the model $\hat{\theta}_R$ and compute the one-step volatility forecast $\sigma_{k,R+1}^2(\hat{\theta}_R)$. Next, using observations 2 to $R+1$, we estimate the model to obtain $\hat{\theta}_{R+1}$ and calculate the one-step volatility forecast $\sigma_{k,R+2}^2(\hat{\theta}_{R+1})$. We keep "rolling" our sample one observation at a time until we reach T, to obtain θ_T and the last one-step volatility forecast $\sigma_{k,T+1}^2(\hat{\theta}_T)$. Consider an objective function that depends on vola-

[3] It is a well-known fact that ST-GARCH models face convergence problem when smoothing parameter δ is estimated. We carried out a grid search for the δ in the interval [0, 20] and from the comparison of likelihood values we arrived at the value $\delta = 3$.

tility, for instance, a loss function $L(Z,\sigma^2(\theta))$ where Z typically will consist of dependent variables and predictor variables. $L(Z,\sigma^2(\theta))$ needs not be differentiable with respect to θ. The best forecasting model is the one that minimizes the expected loss. We test a hypothesis about an $l\times 1$ vector of moments, $E(\mathbf{f}^\dagger)$, where $\mathbf{f}^\dagger\equiv\mathbf{f}(Z,\theta^\dagger)$ is an $l\times 1$ vector with the kth element $f_k^\dagger=L(Z,\sigma_0^2(\theta^\dagger))-L(Z,\sigma_k^2(\theta^\dagger))$, for $\theta^\dagger=p\lim\hat{\theta}_T$. A test for a hypothesis on $E(\mathbf{f}^\dagger)$ may be based on the $l\times 1$ statistic $\bar{\mathbf{f}}\equiv P^{-1}\sum_{t=R}^{T}\hat{\mathbf{f}}_{t+1}$, where $\hat{\mathbf{f}}_{t+1}\equiv\mathbf{f}(Z_{t+1},\hat{\theta}_t)$.

Our interest is to compare all the models with a benchmark. An appropriate null hypothesis is that all the models are no better than a benchmark, i.e., H_0: $\max_{1\le k\le l}E(f_k^\dagger)\le 0$. This is a multiple hypothesis, the intersection of the one-sided individual hypotheses $E(f_k^\dagger)\le 0$, $k=1,\ldots,l$. The alternative is that H_0 is false, that is, the best model is superior to the benchmark. If the null hypothesis is rejected, there must be at least one model for which $E(f_k^\dagger)$ is positive. Suppose that $\sqrt{P}(\bar{\mathbf{f}}-E(\mathbf{f}^\dagger))\xrightarrow{d}N(0,\Omega)$ as $P(T)\to\infty$ when $T\to\infty$, for Ω positive semi-definite. White's (2000) test statistic for H_0 is formed as $\bar{V}\equiv\max_{1\le k\le l}\sqrt{P}\bar{f}_k$, which converges in distribution to $\max_{1\le k\le l}G_k$ under H_0, where the limit random vector $G=(G_1,\ldots,G_l)'$ is $N(0,\Omega)$. However, as the null limiting distribution of $\max_{1\le k\le l}G_k$ is unknown, White's Theorem 2.3 shows that the distribution of $\sqrt{P}(\bar{\mathbf{f}}^*-\bar{\mathbf{f}})$ converges to that of $\sqrt{P}(\bar{\mathbf{f}}-E(\mathbf{f}^\dagger))$, where $\bar{\mathbf{f}}^*$ is obtained from the stationary bootstrap of Politis and Romano (1994). By the continuous mapping theorem this result extends to the maximal element of the vector $\sqrt{P}(\bar{\mathbf{f}}^*-\bar{\mathbf{f}})$ so that the empirical distribution of

$$\bar{V}^*=\max_{1\le k\le l}\sqrt{P}(\bar{f}_k^*-\bar{f}_k), \tag{1}$$

may be used to compute the p-value of $\bar{V}$ (White, 2000, Corollary 2.4). This p-value is called the "reality check p-value".

3.1. Remarks

The following four remarks, each related to the issues of (i) differentiability of the loss function and the impact of parameter estimation error, (ii) nestedness of models under comparison, (iii) the forecasting schemes, and (iv) the power of the reality check test, are relevant for the present paper.

First, White's Theorem 2.3 is obtained under the assumption of the differentiability of the loss function (as in West, 1996, Assumption 1). Also, White's Theorem 2.3 is obtained under the assumption that either (a) the same loss function is used for estimation and prediction (i.e., $F\equiv E[(\partial/\partial\theta)f(Z,\theta^\dagger)]=0$) or (b) $(P/R)\log\log R\to 0$ as $T\to\infty$; so that the effect of parameter estimation vanishes (as in West, 1996, Theorem 4.1(a)). Thus, White's Theorem 2.3 does not immediately apply to the nonsmooth functions and the presence of estimated parameters. Nevertheless, White (2000, p. 1113) notes that the results analogous to Theorem 2.3 can be established under similar conditions used in deriving the asymptotic normality of the least absolute deviations estimator. When no parameter estimation is involved, White's procedure is applicable to nondifferentiable f. We expect that the approach of Randles (1982) and McCracken (2000, Assumption 4) may be useful here, where the condition $E[(\partial/\partial\theta)f(Z,\theta^\dagger)]=0$ is modified to $(\partial/\partial\theta)Ef(Z,\theta^\dagger)=0$ to exploit the fact that the expected loss function may be differentiable even when the loss function is not.[4] We conjecture that when parameter estimation is involved, White's procedure continues to hold either when $(\partial/\partial\theta)Ef(Z,\theta^\dagger)=0$ or when P grows at a suitably slower rate than R. This proof is much involved and has to be pursued in further work. Since we are using different criteria for in-sample estimation and forecast evaluation, there is no reason to expect that $(\partial/\partial\theta)Ef(Z,\theta^\dagger)=0$. Hence, it is important to have very large R compared to P. In our empirical section, for the option loss function, we have $R=7608/(\tau-t)$ and $P=429$, where the maturity τ of the option is $(\tau-t)$ ahead of the current date t. For the other three loss functions (utility function, VaR loss function,

[4] The condition $(\partial/\partial\theta)Ef(Z,\theta^\dagger)=0$ is indeed weaker than the condition $E[(\partial/\partial\theta)f(Z,\theta^\dagger)]=0$, because for example, for the loss function Q to be defined in the next section, $Ef(Z,\theta^\dagger)$ is differentiable while $f(Z,\theta^\dagger)$ is not differentiable. See McCracken (2000, p. 202) and Giacomini and Komunjer (2002, Proof of Proposition 2). See also Kim and Pollard (1990, p. 205) for a set of sufficient conditions for continuous differentiability of expectations of indicator functions. Randles (1982) provides the further conditions under which the parameter estimators are asymptotically normal when the condition $(\partial/\partial\theta)Ef(Z,\theta^\dagger)=0$ holds.

and predictive likelihood), we have $R=6648$ and $P=999$. Supporting evidence is provided by Monte Carlo experiments reported in Sullivan and White (1998), where, for the case of the indicator function and with parameter estimation, the stationary bootstrap reality check delivers quite good approximations to the desired limiting distribution (White, 2000, p. 1113).

Second, White (2000) does not require that Ω be positive definite (that is required in West, 1996), but that Ω be positive *semi*-definite (White, 2000, pp. 1105–1106). Hence, it is required that at least one of the competing models ($k=1,\ldots,l$) is nonnested with respect to the benchmark.

Third, White (2000, pp. 1107–1108) discussed that it would not be necessary to deal explicitly with the forecast schemes such as the "recursive", "rolling", and "fixed" forecasting schemes, defined in West and McCracken (1998, p. 819). West and McCracken (p. 823) and McCracken (2000, p. 203) showed how Ω may be differently affected by parameter estimation uncertainty depending on the choice of the forecasting schemes. When there is no parameter estimation involved, we may not need to deal explicitly with the forecasting schemes in using the bootstrap reality check. However, when parameters are to be estimated, we note that this may be a nontrivial issue due to the potential effect of the in-sample parameter estimation errors and that Corradi and Swanson (2003a, 2003b) have examined the validity of the block bootstrap in the presence of the parameter estimation error for the fixed forecasting scheme and for the recursive forecasting scheme. While the recursive scheme has the advantage of using more observations, we use the rolling forecasting scheme, as described in the beginning of the section, because it may be more robust to a possible parameter variation during the nearly 30-year sample period in the presence of potential structural breaks.

Finally, we note that the White's reality check may be quite conservative when a poor model is included in the set of l competing models. The inclusion of $\bar{f}_k$ in Eq. (1) guarantees that the statistic satisfies the null hypothesis $E(\bar{f}_k^* - \bar{f}_k)=0$ for all k. This setting makes the null hypothesis the least favorable to the alternative and consequently, it renders a very conservative test. When a poor model is introduced, the reality check p-value becomes very large and, depending on the variance of $\bar{f}_k$, it may remain large even after the inclusion of better models. Hence, the White's reality check p-value may be considered as an upper bound for the true p-value. Hansen (2001) considered different adjustments to Eq. (1) providing a lower bound for the p-value as well as intermediate values that depend on the variance of $\bar{f}_k$. In Hansen, the statistic (1) is modified as

$$\bar{V}^* = \max_{1\le k\le l} \sqrt{P}(\bar{f}_k^* - g(\bar{f}_k)). \tag{2}$$

Different $g(\cdot)$ functions will produce different bootstrap distributions that are compatible with the null hypothesis. If $g(\bar{f}_k)=\max(\bar{f}_k,0)$, the null hypothesis is the more favorable to the alternative, and the p-value associated with the test statistic under the null will be a lower bound for the true p-value. Hansen (2001) recommended setting $g(\cdot)$ as a function of the variance of $\bar{f}_k$, i.e.

$$g(\bar{f}_k) = \begin{cases} 0 & \text{if } \bar{f}_k \le -A_k \\ \bar{f}_k & \text{if } \bar{f}_k > -A_k \end{cases} \tag{3}$$

where $A_k=(1/4)P^{-1/4}\sqrt{\mathrm{var}(P^{1/2}\bar{f}_k)}$ with the variance estimated from the bootstrap resamples.

In our empirical section, we report three reality check p-values: the upper bound p-values with $g(\bar{f}_k)=\bar{f}_k$ as in Eq. (1) (denoted as White), lower bound p-values with $g(\bar{f}_k)=\max(\bar{f}_k,0)$ (denoted as Hansen$_L$), and intermediate p-values with $g(\bar{f}_k)$ determined from Eq. (3) (denoted as Hansen).

4. Loss functions

In this section, we present the four loss functions (to be denoted as O, U, Q, and W) through which we evaluate the predictive ability of the various volatility models. We deal with two economic loss functions where volatility is of paramount importance. The first function (O) is based on the Black–Scholes option pricing formula. The second function (U) deals with maximizing the utility of an agent who holds a portfolio of a risk-free asset and a risky asset. We also consider two statistical loss functions. The loss function (Q) is a goodness-of-fit measure for a value-at-risk calculation. As the loss Q is a nondifferentiable function, we also use a smooth approximation to Q, denoted as $\tilde{Q}$, which is differentiable. The second statistical loss function is based on the predictive log-likelihood function (W) under the assumption of conditional normality.[5]

4.1. Option pricing-based loss function

We consider an European call option written on a stock. A holder of a call option has the right to buy the stock at the expiration date of the option, at the strike price agreed in the contract. Black and Scholes (1973) and Merton (1973) derived the price of a call option under the assumption of no market imperfections, continuous trading, no borrowing constraints, no arbitrage opportunities, and geometric Brownian dynamics for the stock price. Under these assumptions, the price of a call option is given by

$$C_{t+1,t} = S_t \exp[-d_t(\tau - t)]\Phi(\delta_1) - X\exp[r_t(\tau - t)] \times \Phi(\delta_2), \quad (4)$$

where $C_{t+1,t}$ is the one-period ahead predicted price of the call option at time t that expires in $(\tau - t)$ periods; S_t is the price of the underlying stock at time t; $(\tau - t)$ is the option time to maturity; r_t is the risk-free interest rate at time t; d_t is the dividend yield on the underlying stock at time t; X is the strike stock price; $\Phi(\cdot)$ is the normal cumulative distribution function; $\delta_1 = [\ln(S_t/X) + (r_t - d_t + 0.5\sigma^2_{\tau,t})(\tau - t)] \div \sigma_{\tau,t}\sqrt{\tau - t}$; $\delta_2 = \delta_1 - \sigma_{\tau,t}\sqrt{\tau - t}$; and $\sigma^2_{\tau,t}$ is the volatility of the stock price at time t to remain constant till the expiration time τ.

For the derivation of the result and other option related issues, we refer to Hull (2000) and Merton (1992). In the call option formula, the only argument that is not observable is the volatility. For each volatility model, we can compute a volatility forecast that will be fed into the option formula to produce the predicted option price. Our volatility model evaluation is based on comparing the predicted option price with the actual option price.[6]

An important issue is the computation of the volatility forecast for $\tau - t$ periods. The question becomes on how to construct the volatility forecast in order to be faithful to the assumption of constant variance over the expiration period.

The first approach is due to Noh et al. (1994), whose estimator of volatility is an average of multistep forecasts of a GARCH model over the expiration period of the option. Aside the fact that this approach allows for time-varying variances during the expiration time of the option, we do not follow Noh et al. approach because of mainly two reasons. One reason is related to the properties of a multistep forecast. If the process is stationary, the multistep forecast of the conditional variance should converge to the unconditional variance of the process as the forecasting horizon increases. Since our purpose is to differentiate among variants of GARCH models, an average of multistep forecasts will not be helpful when the expiration time of the option is relatively long because the average will be dominated by the unconditional variance of the process and thus produce *under*-estimates of long-horizon volatility. Another reason is that multistep forecasts of GARCH processes are

[5] Strictly speaking, we do not need conditional normality because the QML estimators will be consistent. Also, the condition $(\partial/\partial\theta)Ef(Z,\theta^\dagger) = 0$ or $E[(\partial/\partial\theta)f(Z,\theta^\dagger)] = 0$ will be satisfied when we use the same loss function for the out-of-sample forecast evaluation (Gaussian predictive likelihood) as for the in-sample estimation.

[6] We understand that using the Black–Scholes formulation for option pricing is a strong simplification of the problem. It is conceivable that one separately derives the option pricing formula for each of the volatility models. Heston (1993) and Heston and Nandi (2000) provide the closed-form option pricing formula for stochastic volatility and GARCH volatility dynamics, respectively. But given the varied nature of the volatility models considered here, it is nearly impossible to get a closed-from option pricing formula for nonlinear volatility models. Even finding the ordinary differential equation (that needs to be solved numerically) is nontrivial for some models considered here. The only work that comes close to providing a solution is that of Duan (1997) (in the form of an augmented GARCH model), which provides a diffusion approximation to many symmetric and asymmetric GARCH. Unfortunately, it does not shed any light on the corresponding option pricing formulas. Thus, we take the Black–Scholes formula and, to account for the constancy of volatility over the expiration period, we do suitable aggregation as discussed shortly.

highly complicated mainly when the model includes nonlinear features.

The second approach, which is the popular industry practice (e.g., Riskmetrics, 1995) for computing multistep volatility forecasts, is to scale up the high-frequency volatility forecasts to get a low-frequency volatility measure (i.e., converting 1-day standard deviation to h-day standard deviation by scaling with $\sqrt{h}$). See Diebold, Hickman, Inoue, and Schuermann (1998) and Tsay (2002, p. 260). However, Christoffersen, Diebold, and Schuermann (1998), Diebold et al., and Tsay (p. 267) showed that this method will produce over-estimates of long-horizon volatility and hold only for the special case of Riskmetrics' EWMA model.

The third approach is based on temporal aggregation formulae as presented in Drost and Nijman (1993), who addressed the issue of temporal aggregation for linear ARCH models and showed that "weak GARCH" models can be temporally aggregated. As Christoffersen and Diebold (2000, p. 13) pointed out, this approach has some drawbacks; i.e., the aggregation formulae assume the fitted model as the true data generating process and there are no formulae yet available for nonlinear GARCH models.[7]

The fourth approach that we use in this paper is to work directly at the horizons of interest, thereby avoiding temporal aggregation entirely (Christoffersen & Diebold, 2000, p. 13). The approach consists of calculating one-step forecast of the variance of an aggregated process where the level of aggregation is dictated by the expiration time of the call option. If the option expires in m days, the stock price series is aggregated at m period intervals and we forecast one-step ahead (that is m days) conditional variance from the aggregated process. Effectively, from the current period through the expiration time of the option the conditional variance is constant.

Now, we define our option-based loss function, denoted as O. We consider call options on the S&P500 index with strike prices X ranging from 1200 through 1600 index points with intervals of 25 points, with a total of 17 different strike prices X_i ($i=1,\ldots,17$). The option data was collected for 11 months ($j=1,\ldots,11$), with expiration dates ranging from January 2000 through November 2000. Hence, we index the price of a call option expressed in Eq. (4) by using indices i and j, that is $C^{i,j}_{t+1,t}$. The maximum life for the traded options is rounded up to 39 days because we observe only significant trading over this time span. We denote the maximum life of the options by $\tau=39$.

Let $\hat{C}^{i,j}_{t+1,t}$ be the one-period ahead predicted call option price at time t using the formula in Eq. (4). Let $C^{i,j}_{t+1}$ be the actual price at time $t+1$ for the same call option and let $\omega^{i,j}_{t+1}$ be the volume share of the option with strike price X_i expiring in month j with respect to the total volume of the call option across all strike prices for month j. Define the volume-weighted sum of squared pricing errors (WSSE) (sum for the options with 17 different strike prices)

$$\mathrm{WSSE}^j_{t+1} \equiv \sum_{i=1}^{17} \omega^{i,j}_{t+1}(\hat{C}^{i,j}_{t+1,t} - C^{i,j}_{t+1})^2. \tag{5}$$

Then the option-based loss function for the option expiring in month j ($j=1,\ldots,11$) will be defined as

$$O^j \equiv \tau^{-1} \sum_{t=1}^{39} \mathrm{WSSE}^j_{t+1}. \tag{6}$$

Instead of evaluating models in terms of O^j for each month j, we take the average of O^j over the 11 months and define our first loss function O as

$$O \equiv J^{-1} \sum_{j=1}^{J=11} O^j = (J \times \tau)^{-1} \sum_{j=1}^{J=11} \sum_{t=1}^{\tau=39} \mathrm{WSSE}^j_{t+1}. \tag{7}$$

[7] The issue of aggregation is an open question in the realm of nonlinear GARCH models. Drost and Werker (1996) provides the result for the GARCH models and show the strong and semi-strong GARCH models are not robust to temporal aggregation. To the best of our knowledge, no such result is available for the host of GARCH models that we consider here. We do acknowledge that the ranking may depend on the extent of aggregation. As our result is based on averaging over $\tau=39$ levels of aggregation, we believe that any abnormal performance of a given model for a given level of aggregation will also be smoothed out. Alternatively, we may use simulation to find the relationship between parameters of different levels of aggregation. It is possible to use simulation if the data generating process is closed under aggregation. Otherwise, it is very difficult to locate the right model for the different level of aggregation. Thus, to find the actual relationship between the disaggregated and aggregated parameters might be very difficult.

The advantage of using O as a loss function instead of O^j is two-fold: one is to simplify the presentation of results and another is to increase the out-of-sample size for the reality check from $\tau = 39$ to $P \equiv J \times \tau = 11 \times 39 = 429$, which contributes to improve the power of the reality check tests.[8]

4.2. Utility-based loss function

In the exchange rate market, West et al. (1993) evaluated the performance of a GARCH model against ARCH, ABS-ARCH, and nonparametric models using a utility-based criterion. They considered an agent who optimizes the one period expected wealth when holding a portfolio of two assets: a foreign asset and a domestic asset. In this paper, we borrow their utility-based criterion to compare the predictive performance of many more volatility models controlling, at the same time, for potential data snooping problems. In our case, the agent maximizes her expected utility given that her wealth is allocated between a risky asset (S&P500 index) and a riskless asset (the 3-month treasury bill)

$$\max_{\alpha_t} E(U_{t+1} \mid \mathscr{F}_t) \equiv E(w_{t+1} - 0.5\gamma w_{t+1}^2 \mid \mathscr{F}_t),$$

$$\text{s.t.} \quad w_{t+1} = \alpha_t y_{t+1} + (1 - \alpha_t) r_{t+1} \tag{8}$$

where w_{t+1} is the return to the portfolio at time $t+1$, γ is a risk aversion parameter, α_t is the weight of the risky asset in the portfolio, y_{t+1} is the S&P500 return, and r_{t+1} is the risk-free rate, which is assumed known. In West et al. (1993) framework, it is assumed that all relevant moments of the return distribution are known except for the conditional variance. Solving (utility) gives the maximum expected utility

$$E(U^*_{t+1} \mid \mathscr{F}_t) = E(c_{t+1}(\gamma) + d_{t+1}(\gamma) x(e_{t+1}^2, \hat{\sigma}_{t+1}^2) \mid \mathscr{F}_t), \tag{9}$$

where $e_{t+1} \equiv y_{t+1} - r_{t+1}$ is the excess return to the risky asset, $\hat{\sigma}^2_{t+1}$ is the estimated conditional variance of e_{t+1}, and $\mu_{t+1} \equiv E(e_{t+1} | \mathscr{F}_t)$,

$$c_{t+1}(\gamma) \equiv r_{t+1} - 0.5\gamma r_{t+1}^2, \tag{10}$$

$$d_{t+1}(\gamma) \equiv \mu_{t+1}^2 \frac{(1 - \gamma r_{t+1})^2}{\gamma}, \tag{11}$$

and

$$x(e_{t+1}^2, \hat{\sigma}_{t+1}^2) \equiv \frac{1}{(\mu_{t+1}^2 + \hat{\sigma}_{t+1}^2)} - 0.5 \frac{(\mu_{t+1}^2 + e_{t+1}^2)}{(\mu_{t+1}^2 + \hat{\sigma}_{t+1}^2)^2}. \tag{12}$$

We should note that this utility function is asymmetric. Miscalculations of the conditional variance are paid in units of utility. A risk averse agent will have lower expected utility when the conditional variance is underestimated than when it is overestimated. Based on this criterion, our second economic loss function is

$$U \equiv -P^{-1} \sum_{t=R}^{T} \hat{U}^*_{t+1} = -P^{-1} \sum_{t=R}^{T} (c_{t+1}(\gamma) + \hat{d}_{t+1}(\gamma) \hat{x}(e_{t+1}^2, \hat{\sigma}_{t+1}^2)) \tag{13}$$

where $\hat{d}(\cdot)$ and $\hat{x}(\cdot,\cdot)$ are obtained from Eqs. (11) and (12) by replacing μ_{t+1} with the predicted excess

[8] Our effective sample size is 429 because we consider 11 different expiration months and 39 time period for each expiration months. It is possible that there are contemporaneous observations but there is no repetition of the observations, as two options trading in the same time but expiring at different months are not identically priced. Also, to make sure that the time series dependence (if any) across the options over the 11 different expiration months may not affect the bootstrap adversely, we have used various smoothing parameters q of the stationary bootstrap that is corresponding to the mean block length ($1/q$) of the stationary bootstrap. The results were robust to the various values of $q = 0.25$, 0.50, 0.75, and 1.00 ($q = 1$ corresponds to the mean block length 1).

return $\hat{\mu}_{t+1}$. In the empirical section, γ is set at 0.5 but we have experimented with different values of the risk aversion coefficient and our results remain unchanged. Note that U is to be minimized.[9]

4.3. VaR-based loss function

The conditional value-at-risk, denoted as $\mathrm{VaR}^{\alpha}_{t+1}$, can be defined as the conditional quantile

$$\Pr(y_{t+1} \le \mathrm{VaR}^{\alpha}_{t+1} \mid \mathcal{F}_t) = \alpha. \tag{14}$$

If the density of y belongs to the location-scale family (e.g., Lehmann, 1983, p. 20), it may be estimated from

$$\mathrm{VaR}^{\alpha}_{t+1} = \mu_{t+1}(\hat{\theta}_t) + \Phi^{-1}_{t+1}(\alpha)\sigma_{t+1}(\hat{\theta}_t), \tag{15}$$

where $\Phi_{t+1}(\cdot)$ is the forecast cumulative distribution (not necessarily standard normal) of the standardized return, $\mu_{t+1}(\theta) = E(y_{t+1}|\mathcal{F}_t)$ is the conditional mean forecast of the return, and $\sigma^2_{t+1}(\theta) = E(\varepsilon^2_{t+1}|\mathcal{F}_t)$ the conditional variance forecast based on the volatility models of Section 2, and $\hat{\theta}_t$ is the parameter vector estimated by using the information up to time t. We fit an AR(0) model with a constant term in the mean equation and the estimated values of the constant are very close to zero. We assume conditional normality of the standardized return.[10] We consider the quantile $\alpha = 0.05$ and thus $\Phi^{-1}_{t+1}(0.05) = -1.645$ for all t.

[9] It may be noted that $\hat{\sigma}^2_{t+1}$ is not the optimal forecast of the conditional variance under the asymmetry of the loss function. Christoffersen and Diebold (1996) provide some results for the GARCH(1,1) under the LinLin loss. It will be difficult to derive the optimal volatility forecast for all volatility models and for our loss functions. But we do acknowledge that the forecasts need not be optimal when the models are estimated by QML, while the forecasts are evaluated via asymmetric loss functions.

[10] We did carry out the analysis with Student's t distribution and qualitative nature of the result is same as what we obtained under conditional normality.

Our first statistical loss function Q is the loss function used in the quantile estimation (see, e.g., Koenker & Bassett, 1978), that is, for given α,

$$Q \equiv P^{-1} \sum_{t=R}^{T} (\alpha - d^{\alpha}_{t+1})(y_{t+1} - \mathrm{VaR}^{\alpha}_{t+1}), \tag{16}$$

where $d^{\alpha}_{t+1} \equiv \mathbf{1}(y_{t+1} < \mathrm{VaR}^{\alpha}_{t+1})$. This is an asymmetric loss function that penalizes more heavily with weight $(1-\alpha)$ the observations for which $y - \mathrm{VaR}^{\alpha} < 0$. Smaller Q indicates a better goodness of fit.

Note that the loss Q is not differentiable due to the indicator function. As discussed in Section 3.1, White's (2000) procedure may continue to be valid and applicable for nondifferentiable losses. We expect that when parameter estimation is involved, the impact of parameter estimation uncertainty is asymptotically negligible when P grows at a suitably slower rate than R. Thus, in our empirical work, we choose the prediction period ($P = 999$) that is much smaller than the estimation period ($R = 6648$).

Granger (1999, p. 165) notes that the problem of nondifferentiability may be just a technicality because there may exist a smooth function that is arbitrarily close to the nonsmooth function. Hence, we deal with the nondifferentiability of Q by running our experiments with a smoothed version of the loss Q where the indicator function is replaced with a continuous differentiable function. We denote this smoothed Q as $\tilde{Q}$ and define

$$\tilde{Q} \equiv P^{-1} \sum_{t=R}^{T} (\alpha - m_{\delta}(y_{t+1}, \mathrm{VaR}^{\alpha}_{t+1}))(y_{t+1} - \mathrm{VaR}^{\alpha}_{t+1}), \tag{17}$$

where $m_{\delta}(a,b) = [1 + \exp\{\delta(a-b)\}]^{-1}$. Note that $m_{\delta}(a,b) = 1 - m_{\delta}(b,a)$. The parameter $\delta > 0$ controls the smoothness. A higher value of δ makes $\tilde{Q}$ closer to Q. For $\tilde{Q}$, we consider many values of δ and we find that for values of $\delta > 10$ the loss values for both Q and $\tilde{Q}$ are very similar. We report the results for $\delta = 25$. The results with other values of δ are available and very similar to those reported here. The results of Q and $\tilde{Q}$ in Section 5 indicate the validity of the stationary bootstrap reality check with respect to the nondifferentiable loss.[11]

4.4. Predictive likelihood-based loss function

Our second statistical loss function is the predictive likelihood. The *negative* average predictive likelihood under the conditional normality assumption, denoted W, is given by

$$W \equiv -P^{-1} \sum_{t=R}^{T} \log l(Z_{t+1}, \hat{\theta}_t),$$

where

$$\log l(Z_{t+1}, \hat{\theta}_t) = -\log(\sqrt{2\pi}) - \frac{1}{2}\log\sigma^2_{t+1}(\hat{\theta}_t) - \frac{\varepsilon^2_{t+1}(\hat{\theta}_t)}{2\sigma^2_{t+1}(\hat{\theta}_t)},$$

$\varepsilon_{t+1}(\theta) = y_{t+1} - \mu_{t+1}(\theta)$ is a forecast error, $\mu_{t+1}(\theta) = E(y_{t+1}|\mathscr{F}_t)$, $\sigma^2_{t+1}(\theta) = E(\varepsilon^2_{t+1}|\mathscr{F}_t)$, and $\hat{\theta}_t$ is the parameter vector estimated by using the information up to time t. The loss W is to be minimized. See Bjørnstad (1990) for a review on predictive likelihood. Note that we evaluate the conditional models for $\mu_{t+1}(\theta)$ and $\sigma^2_{t+1}(\theta)$ in terms of the Gaussian predictive likelihood, which is different from a density forecast evaluation (e.g., Diebold, Gunther, & Tay, 1998).

5. Empirical results

In this section, we describe the data and explain the results presented in Tables 2 and 3.

5.1. Data

We consider closing prices of call options on the S&P500 index with strike prices ranging from 1200 through 1600 index points with intervals of 25 points, traded in the Chicago Board of Options Exchange (CBOE). We have omitted those options for which the trading volume is mostly zero. We consider mostly at-the-money options. The time period considered is 39 trading days before expiration since the number of days with nonzero volume is quite small. The option data was collected for 11 months, with expiration dates ranging from January 2000 through November 2000. The option data was purchased from Dialdata.com.

We consider 7647 daily observations of the S&P500 index from April 1, 1970 till November 17, 2000. The index was collected from finance.yahoo.com. The daily dividend data was collected from Datastream for the same period as that of the index. The risk-free rate is the secondary market 3-month treasury bill rate and it was retrieved from St. Louis Federal Reserve Bank.

For the option-based loss function we used the S&P500 percentage returns from April 1, 1970 until the date on which the option is traded to forecast one-step ahead conditional variance of the properly aggregated return series. This in turn was used to estimate the price of the call option.

For the utility-based loss function, VaR-based loss function, and predictive likelihood function, no ag-

[11] We do not have a theoretical proof on the consistency and asymptotic refinement of the stationary bootstrap with respect to the non-differentiable loss Q. As the Edgeworth expansion of the indicator function is very complicated as discussed by De Angelis, Hall, and Young (1993), we do not know whether the bootstrap can provide the asymptotic refinement for the non-smooth estimators (although we suspect so). Our experiment (with replacing the indicator function with a smooth function, thereby producing a modified objective function $\tilde{Q}$ whose derivatives are continuous) is to show (empirically) that bootstrap may work for the non-smooth loss function. In fact, the reality check results using the smoothed objective function $\tilde{Q}$ and the original non-smooth objective function Q are virtually identical. Hence, this confirms the theoretical results on the bootstrap consistency for the smoothed LAD estimator (Horowitz, 1998) and for the smoothed maximum score (MS) estimator (Horowitz, 2002), where a smooth kernel is used to replace the indicator function. It may be carried over to the other quantiles than the median. It may be shown that the smoothed and unsmoothed estimators are first-order asymptotically equivalent. We can also show the asymptotic normality of the quantile estimators (see Komunjer, 2003). Due to the first-order asymptotic equivalence of the smoothed and unsmoothed quantile estimators, due to the asymptotic normality of the quantile estimators, and due to the virtually identical empirical results we obtained for Q and $\tilde{Q}$, we conjecture that the bootstrap will work for the unsmoothed objective function Q. However, this is only a conjecture because the theoretical results of Horowitz (1998, 2002) and Hahn (1995) do not cover the dependent series and the theoretical result of Fitzenberger (1998) does not cover the parameter estimation error in the out-of-sample forecasting. The extension of Corradi and Swanson (2003a, 2003b) to the case non-smooth estimators (e.g., quantile estimator) would be an interesting future research topic.

Table 2
Reality check

Panel A. Based on economic loss functions

Benchmark	O	White	Hansen	Hansen_L	U	White	Hansen	Hansen_L
GARCH	58,441.8	0.969	0.515	0.456	− 0.024	0.286	0.000	0.000
Q-GARCH$_1$	58,378.7	0.971	0.524	0.468	− 0.027	1.000	0.518	0.495
E-GARCH	62,361.5	0.818	0.570	0.102	− 0.022	0.206	0.000	0.000
T-GARCH	65,906.5	0.714	0.366	0.133	− 0.023	0.219	0.000	0.000
ST-GARCH	60,364.3	0.862	0.487	0.144	− 0.023	0.251	0.000	0.000
I-GARCH	62,501.5	0.775	0.475	0.088	− 0.024	0.331	0.000	0.000
Q-GARCH$_2$	59,706.5	0.876	0.225	0.207	− 0.023	0.221	0.000	0.000
Q-GARCH$_3$	75,971.6	0.575	0.104	0.091	− 0.022	0.150	0.000	0.000
ARCH(5)	60,682.2	0.868	0.475	0.184	− 0.023	0.229	0.000	0.000
ABS-GARCH	57,828.0	0.999	0.963	0.867	− 0.024	0.302	0.000	0.000
A-GARCH	122,546.0	0.111	0.076	0.075	− 0.023	0.207	0.000	0.000
EWMA	58,030.4	0.976	0.654	0.466	− 0.023	0.246	0.000	0.000
MA(20)	58,528.9	0.965	0.548	0.431	− 0.022	0.168	0.000	0.000
LOG-GARCH	58,116.2	0.977	0.606	0.546	− 0.023	0.215	0.000	0.000
SV	233,962.0	0.000	0.000	0.000	0.010	0.000	0.000	0.000

Panel B. Based on statistical loss functions

Benchmark	Q	White	Hansen	Hansen_L	W	White	Hansen	Hansen_L
GARCH	1.807	0.000	0.000	0.000	1.602	0.532	0.040	0.015
Q-GARCH$_1$	1.807	0.000	0.000	0.000	1.602	0.521	0.034	0.007
E-GARCH	1.509	0.000	0.000	0.000	1.608	0.523	0.129	0.051
T-GARCH	1.796	0.000	0.000	0.000	1.592	0.719	0.189	0.018
ST-GARCH	1.771	0.000	0.000	0.000	1.587	0.843	0.072	0.059
I-GARCH	1.880	0.000	0.000	0.000	1.603	0.547	0.068	0.038
Q-GARCH$_2$	1.745	0.000	0.000	0.000	1.585	0.913	0.101	0.086
Q-GARCH$_3$	1.614	0.000	0.000	0.000	1.638	0.376	0.004	0.004
ARCH(5)	1.659	0.000	0.000	0.000	1.637	0.373	0.004	0.004
ABS-GARCH	1.760	0.000	0.000	0.000	1.603	0.536	0.018	0.004
A-GARCH	1.737	0.000	0.000	0.000	1.581	0.993	0.914	0.530
EWMA	1.830	0.000	0.000	0.000	1.607	0.469	0.024	0.019
MA(20)	1.818	0.000	0.000	0.000	1.639	0.384	0.024	0.023
LOG-GARCH	1.816	0.000	0.000	0.000	1.611	0.465	0.006	0.000
SV	1.041	1.000	0.516	0.495	2.632	0.000	0.000	0.000

(1) We compare each model as the benchmark model with all the remaining $l = 14$ models. (2) "White", "Hansen" and "Hansen_L" denote reality check p-values of the White's test, Hansen's intermediate test, and Hansen's liberal test, respectively. The bootstrap reality check p-values are computed with 1000 bootstrap resamples and smoothing parameter $q = 0.25$. See Politis and Romano (1994) or White (2000) for the details. The p-values for $q = 0.75$ and 0.50 are similar and are not reported. (3) The sample period of the data is from April 1, 1970 to November 17, 2000 with $T = 7647$ observations. (4) For the O loss function, $R = 7608/(\tau - t)$, where the maturity of the option is $(\tau - t)$ ahead of the current date. For the O loss function, the forecast horizon for every option is 39 periods but as we aggregate across months $P = \tau \times J = 39 \times 11 = 429$. (5) For the loss functions U, Q, and W, the models are estimated using $R = 6648$ observations and the forecast evaluation period is $P = 999$. (6) All the loss functions are to be minimized.

gregation of the data was needed. We divide the S&P500 data into two subsamples: the most recent 999 observations is the forecasting period ($P = 999$) and the rest is the estimation period ($R = 6648\%$). We choose large R to make $(P/R)\log\log R$ small to reduce the impact of the parameter estimation uncertainty (White, 2000, Theorem 2.3) while we also keep P reasonably large enough to maintain the power of the reality check (White, 2000, Proposition 2.5).

5.2. Results

We evaluate the out-of-sample predictive ability of the various volatility models described in Section 2,

Table 3
Reality check based on smoothed VaR loss function

Benchmark	$\tilde{Q}$	White	Hansen	Hansen_L
GARCH	1.808	0.000	0.000	0.000
Q-GARCH$_1$	1.808	0.000	0.000	0.000
E-GARCH	1.508	0.000	0.000	0.000
T-GARCH	1.797	0.000	0.000	0.000
ST-GARCH	1.771	0.000	0.000	0.000
I-GARCH	1.879	0.000	0.000	0.000
Q-GARCH$_2$	1.745	0.000	0.000	0.000
Q-GARCH$_3$	1.614	0.000	0.000	0.000
ARCH(5)	1.659	0.000	0.000	0.000
ABS-GARCH	1.760	0.000	0.000	0.000
A-GARCH	1.737	0.000	0.000	0.000
EWMA	1.830	0.000	0.000	0.000
MA(20)	1.818	0.000	0.000	0.000
LOG-GARCH	1.816	0.000	0.000	0.000
SV	1.041	1.000	0.516	0.496

For the loss function $\tilde{Q}$, the models are estimated using $R=6648$ observations and the forecast evaluation period is $P=999$. For $\tilde{Q}$, the smoothing parameter δ is set to be 25.

using the evaluation methods described in Section 3 and the objective functions of Section 4. We consider a total of 15 models.

In Table 2, we take into account the specification search and we present a multiple comparison of the benchmark model with all of the remaining 14 models. The p-values are computed using the stationary bootstrap of Politis and Romano (1994) generating 1000 bootstrap resamples with smoothing parameter $q=0.25$. The p-values for $q=0.75$ and 0.50 are similar (not reported), which is consistent with White (2000, p. 1116). The null hypothesis is that the best of the remaining 14 models is no better than the benchmark. For example, when GARCH is the benchmark White's p-value is 0.969, which indicates the null hypothesis may not be rejected. When SV is the benchmark White's p-value is 0.000 and so the null hypothesis is clearly rejected and there exists a better model than SV.

For the option loss function, we find that the White's reality check p-values for most of the benchmark models are very high. On the other hand, the Hansen's p-values seem to discriminate better among models. The stochastic volatility model is clearly dominated by the rest. The A-GARCH model comes next as the second least preferred model. In contrast, the ABS-GARCH seems to be the most preferred, it has the largest Hansen's p-value. Once again the simplest models such as EWMA and MA(20) are as good as any other specification. In general, there is not a highly preferred specification; none of the models that incorporate asymmetries seem to dominate the symmetric models, even under the most liberal Hansen's test. It seems that only three specifications—the stochastic volatility model, the A-GARCH model, and to a lesser extent the Q-GARCH$_3$ model—are clearly dominated models.

For the utility function, there is a most preferred model that clearly dominates all the rest, this is the Q-GARCH$_1$, which is an asymmetric model. We run the experiment for several values of the absolute rate of risk aversion to assess the robustness of our results. The values considered are 0.5, 0.6. 0.75, 0.8, 0.85, 0.9, and 0.95. Even though, the value of the loss function changes, the Q-GARCH$_1$ remains the preferred model. The worst seems to be the SV model. With the exception of the SV model, there are not very large differences across models.

For VaR-based loss functions Q, the SV model clearly dominates all the other models. It is interesting to note that the worst performers are IGARCH and EWMA, which are the popular models proposed by Riskmetrics (1995) for the VaR computation.

For the predictive likelihood, there seems to be a preference for asymmetric models and the preferred one is the A-GARCH, followed by the Q-GARCH$_2$ and the ST-GARCH. Modeling the conditional standard deviation (A-GARCH, ABS-GARCH, and LOG-GARCH), instead of the variance, seems to be a dominant modeling strategy.

In Table 3, we consider the smoothed version of the VaR loss function. As discussed in Section 3, White's Theorem 2.3 does not readily apply to nondifferentiable loss functions and the presence of estimated parameters, and thus the effect of parameter estimation might not vanish asymptotically (as in West, 1996, Theorem 4.1(b)). While the theoretical results for this nondifferentiable case are not yet available, we confirm the Monte Carlo results reported in Sullivan and White (1998), where it is shown that, for the case with the indicator function and with the parameter estimation, the stationary bootstrap reality check delivers quite good approximations to the desired limiting distribution. We note that the differences between the estimated loss function Q (Table 2) and its smoothed version $\tilde{Q}$ (Table 3) are negligible, implying that the differentiability of the loss function is not an issue for the implementation of the stationary bootstrap reality check. The bootstrap p-values for Q and $\tilde{Q}$ are also virtually the same.

The different p-values differ substantially for loss functions O, U, and W, when the SV model is not used as a benchmark, and for the Q loss function when the SV is used as the benchmark. This is due to fact that the inclusion of a bad model adversely affects the power of the reality check test. A problem in White's (2000) set-up may be that the null hypothesis is composite, H_0: $\max_{1 \leq k \leq l} E(f_k^\dagger) \leq 0$. When $E(f_k^\dagger) = 0$ for all $1 \leq k \leq l$, then the reality check p-value of White (2000) will provide an asymptotically correct size. However, when some models are strictly dominated by the benchmark model, i.e., $E(f_k^\dagger) < 0$ for some $1 \leq k \leq l$, i.e., when bad models are included in the set of the competing models, White's test tends to behave conservatively. Hansen's (2001) modification is basically to remove those (very) bad models in the comparison and to restore the test power. Note that Hansen's p-values are lower than White's p-values.

6. Summary and concluding remarks

In this paper, we have analyzed the predictive performance of multiple volatility models for stock returns. We have considered linear and nonlinear GARCH processes, some of the models are nested and some others are not, such as the stochastic volatility model. We have also included simple models that do not involve the parameter estimation such as MA and EWMA.

To evaluate the performance of this models, we have chosen both economic and statistical loss functions. Statistical functions that are based on some function of the forecast error are not the most appropriate to evaluate volatility models because volatility is not observable and any proxy to realized volatility is subject to estimation error. Our choice of loss functions spans the fields of finance, risk management, and economics. We have considered two statistical loss functions: the goodness-of-fit for a VaR calculation and the average predictive likelihood, where no assumption is required regarding the realized value of volatility.

For each loss function, the statistical framework in which the volatility forecast models are evaluated is that of White (2000). A pairwise comparison of models may result in data snooping biases because the tests are mutually dependent. Since we have multiple volatility models, it is important to take this dependence into account.[12]

As we were expecting there is not an unique model that is the best performer across the four loss functions considered. When we consider an option loss function, simple models like the Riskmetrics EWMA and MA(20) are as good performers as any of the more sophisticated specifications. This is interesting because either EWMA or MA(20) do not require statistical parameter estimation, and their implementation is almost costless. When we consider the VaR loss function the stochastic volatility model performs best. EWMA was proposed by Riskmetrics to calculate VaR but, in our analysis, this model is the worst performer in terms of the conditional quantile goodness-of-fit. When the utility loss function is considered, the Q-GARCH$_1$ model performs best, but, with the exception of the SV model, there are not large differences among the remaining models. We also find that different degrees of risk aversion do not affect the robustness of our results. Finally, for the predictive likelihood-based loss function, asymmetric models, based on the conditional standard deviation (A-GARCH, ABS-GARCH, and LOG-GARCH) instead of the conditional variance, are preferred, with the A-GARCH performing the best.

Different loss functions are relevant for different decision makers, as different types of forecast errors are penalized for different decisions. Our results of particular ranking of the models obtained across the different loss functions is in fact consistent with various important features of different models. For the option loss, the EWMA and a long distributed lag MA(20) models work well, reflecting high persistence in the implied volatility process. The utility loss function penalizes underforecasts more

[12] While the data snooping bias may be caused by the pair-wise tests, potential bias may also be caused from taking different models as benchmarks. It is probably not a big problem, but we acknowledge that this type of dependence is not being taken into account in our current testing framework.

than overforecasts. The asymmetric GARCH models may be more adequate for this particular loss. For the VaR loss, which has a focus on the tails of the density, the SV model can be more flexible than the ARCH class because the volatility equation—allowing for an extra innovation term—performs the best when it is evaluated in terms of the tail quantiles. The predictive likelihood, which deals with the whole distribution in contrast to the VaR loss, places much less emphasis on large values in the tails, so a standard deviation-based model is better than the variance-based models since the impact of large values is magnified in the variance-based models.[13]

Finally, we note that the validity of the stationary bootstrap reality check (White, 2000, Theorem 2.3) is proved under the absence of parameter estimation uncertainty; i.e. under the assumption that either the same loss function is used for estimation and prediction or the estimation sample is suitably larger than the prediction sample. However, in the present paper, we do not use the same loss function for estimation and prediction (except for the predictive likelihood for which we use the Gaussian likelihood for both estimation and prediction). While the volatility models are estimated using the Gaussian likelihood, the forecasts are compared by different loss functions. Recently, Christoffersen and Jacobs (2004), Patton and Timmermann (2003), and Skouras (2001) emphasize the importance of matching the in-sample estimation criterion to the forecast evaluation criterion. We leave this interesting issue for the future research.

[13] While we emphasize these different aspects of various loss functions, we note that our results (on ranking) may not be immediately generalizable to other data sets. Further studies in this line of research with different data sets would be warranted. That the out-of-sample loss function is different from the estimation loss function is one reason that this may not be generalized. The fact that the loss function plays a critical role in the evaluation of nonlinear models has previously been observed in a series of papers by Diebold and co-authors, among others. Chritoffersen and Jacobs (2003) presented results on a similar question using our option pricing loss function that there is a clear link between which loss function is used to estimate the model parameters and which loss function is use to evaluate forecasts. However, we note that our empirical findings and the particular ranking of the models obtained across the different loss functions are consistent with various important features of the loss functions and models, as summarized here.

References

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885–905.

Awartani, B. M. A., & Corradi, V. (2003). *Predicting the volatility of the S&P-500 index via GARCH models: The role of asymmetries*. University of Exeter.

Bera, A. K., & Higgins, M. L. (1993). ARCH models: Properties, estimation, and testing. *Journal of Economic Surveys*, *7*, 305–366.

Bjørnstad, J. F. (1990). Predictive likelihood: A review. *Statistical Science*, *5*(1), 242–265.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*, 637–654.

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

Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH models in finance. *Journal of Econometrics*, *52*, 5–59.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). *ARCH models. Handbook of Econometrics, vol. 4*.

Brooks, C., & Persand, G. (2003). Volatility forecasting for risk management. *Journal of Forecasting*, *22*(1), 1–22.

Christoffersen, P., & Diebold, F. X. (1996). Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, *11*, 561–571.

Christoffersen, P., & Diebold, F. X. (2000). How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics*, *82*, 1–11.

Christoffersen, P., Diebold, F. X., & Schuermann, T. (1998, October). Horizon problems and extreme events in financial risk management. *Economic Policy Review* (pp. 109–118). New York: Federal Reserve Bank of New York.

Christoffersen, P., & Jacobs, L. (2004). The importance of the loss function in option valuation. *Journal of Financial Economics*, (in press).

Corradi, V., & Swanson, N. R. (2003a). *Bootstrap conditional bootstrap tests in the presence of dynamic misspecification*. University of Exeter and Rutgers University.

Corradi, V., & Swanson, N. R. (2003b). *The block bootstrap for recursive m-estimators with applications to predictive evaluation*. University of Exeter and Rutgers University.

De Angelis, D., Hall, P., & Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in L^1 regression. *Journal of the American Statistical Association*, *88*, 1310–1316.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*, 863–883.

Diebold, F. X., Hickman, A., Inoue, A., & Schuermann, T. (1998). Converting 1-day volatility to h-day volatility: Scaling by $\sqrt{h}$ is worse than you think, Wharton Financial Institutions Center, Working Paper 97-34. Published in condensed form as "Scale Models". *Risk*, *11*, 104–107.

Drost, F. C., & Nijman, T. E. (1993). Temporal aggregation of GARCH processes. *Econometrica*, *61*, 909–927.

Drost, F.C., & Werker, B.J.M. (1996). Closing the GARCH Gap:

Continuous GARCH Modeling. *Journal of Econometrics*, *74*(1), 31–57.

Duan, J. C. (1997). Augmented GARCH (p,q) process and it's diffusion limit. *Journal of Econometrics*, *79*, 97–127.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica*, *50*, 987–1008.

Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, *5*, 1–50.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, *484*(5), 1749–1778.

Engle, R. F., Kane, A., & Noh, J. (1997). Index-option pricing with stochastic volatility and the value of accurate variance forecasts. *Review of Derivatives Research*, *1*, 139–157.

Fitzenberger, B. (1998). The moving block bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics*, *82*(2), 235–287.

Geweke, J. (1986). Comment-Modeling Persistence of Conditional Variances. *Econometric Review*, *6*, 57–61.

Giacomini, R., & Komunjer, I. (2002). *Evaluation and Combinations of Conditional Quantile Forecasts*. UCSD and Caltech.

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

González-Rivera, G. (1998). Smooth-transition GARCH models. *Studies in Nonlinear Dynamics and Econometrics*, *3*(2), 61–78.

Granger, C. W. J. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review*, *1*, 161–173.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, *11*, 105–121.

Hansen, P. R. (2001). *An unbiased and powerful test for superior predictive ability*. Brown University.

Hansen, P. R., & Lunde, A. (2002). *A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?* Brown University.

Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, *6*, 327–343.

Heston, S., & Nandi, S. (2000). A closed-form GARCH option valuation model. *Review of Financial Studies*, *13*, 585–625.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, *66*(6), 1327–1351.

Horowitz, J. L. (2002). Bootstrap critical values for tests based on the smoothed maximum score estimator. *Journal of Econometrics*, *111*, 141–167.

Hull, J. (2000). *Options, futures, and other derivatives* (4th ed.). New York: Prentice-Hall.

Kim, J., & Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, *18*, 191–219.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Komunjer, I. (2003). Conditional Quantile Estimation-A Quasi-Maximum Likelihood Approach. Caltech.

Lehmann, E. L. (1983). *Theory of point estimation*. New York: Wiley.

Lo, A. W., & MacKinlay, A. C. (1999). *A non-random walk down wall street*. Princeton: Princeton University Press.

Lopez, J. A. (2001). Evaluating the predictive accuracy of volatility models. *Journal of Forecasting*, *20*, 87–109.

McCracken, M. W. (2000). Robust out-of-sample inference. *Journal of Econometrics*, *99*, 195–223.

Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, *4*, 141–183.

Merton, R. C. (1992). *Continuous time finance*. Oxford, UK: Blackwell.

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

Noh, J., Engle, R. F., & Kane, A. (1994). Forecasting volatility and option prices of the S&P 500 index. *Journal of Derivatives*, 17–30.

Pantula, S. G. (1986). Modelling the persistence of conditional variances: A comment. *Econometric Reviews*, *5*, 71–74.

Patton, A. J., & Timmermann, A. (2003). *Properties of Optimal Forecasts*. LSE and UCSD.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*, 1303–1313.

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

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics*, *10*(2), 462–474.

Riskmetrics (1995). *Technical manual* (3rd ed.).

Ruiz, E. (1994). Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of Econometrics*, *63*(1), 289–306.

Schwert, G. W. (1990). Stock volatility and the crash of '87. *Review of Financial Studies*, *3*(1), 77–102.

Sentana, E. (1995). Quadratic ARCH models. *Review of Economic Studies*, *62*(4), 639–661.

Skouras, S. (2001). *Decisionmetrics: A decision-based approach to econometric modeling*. Santa Fe Institute, Working Paper No. 01-11-064.

Sullivan, R., & White, H. (1998). Finite Sample Properties of the Bootstrap Reality Check for Data-Snooping: A Monte Carlo Assessment, QRDA, LLC Technical Report, San Diego.

Taylor, S. J. (1986). *Modelling financial time series*. New York: Wiley.

Tsay, R. S. (2002). *Analysis of financial time series*. New York: Wiley.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*, 1067–1084.

West, K. D., Edison, H. J., & Cho, D. (1993). A utility based comparison of some models of exchange rate volatility. *Journal of International Economics*, *35*, 23–45.

West, K. D., & McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review*, *39*(4), 817–840.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.

Zakonian, J. M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, *18*, 931–955.

Biographies: Gloria GONZÁLEZ-RIVERA is Associate Professor and Chair of the Department of Economics at University of California, Riverside. She received her PhD in Economics from the University of California, San Diego in 1991 and she has held the position of Senior Economist at Freddie Mac in the Financial Research and Housing Economics Department during 2000. Her research interests include econometrics, time series analysis, nonlinear time series, forecasting theory, and applications to finance and macroeconomics, volatility modeling, financial risk management, and development econometrics.

Tae-Hwy LEE is Associate Professor at the Department of Economics, University of California, Riverside. He received his PhD in Economics from the University of California, San Diego in 1990 and he has held the position of the ASA/NSF/BLS Senior Research Fellow (1995–1996). His research area is time series econometrics and applications, with interests on the topics of cointegration, long memory, nonlinear time series, spectral analysis, aggregation, specification testing (nonlinearity, serial dependence, and martingale hypothesis), forecast theory, volatility and quantile models, skewed and heavy-tailed distributions, jump processes, copula, bootstrap, empirical likelihood, and nonparametric methods.

Santosh MISHRA is a doctoral candidate (expecting his PhD in June 2004) at the Department of Economics, University of California, Riverside. His research interests include econometrics, time series analysis, financial econometrics, forecasting, financial risk management, and nonparametric volatility modeling.