Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Forecasting 21 (2005) 167–183
international journal of forecasting
www.elsevier.com/locate/ijforecast

Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries

Basel M. A. Awartani, Valentina Corradi*

Queen Mary, University of London, Department of Economics, Mile End, London E14NS, United Kingdom

Abstract

In this paper, we examine the relative out of sample predictive ability of different GARCH models, with particular emphasis on the predictive content of the asymmetric component. First, we perform pairwise comparisons of various models against the GARCH(1,1) model. For the case of nonnested models, this is accomplished by constructing the [Diebold, F.X., & Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263 test statistic]. For the case of nested models, this is accomplished via the out of sample encompassing tests of [Clark, T.E., & McCracken, M.W., 2001. Tests for equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*, 85–110]. Finally, a joint comparison of all models against the GARCH(1,1) model is performed along the lines of the reality check of [White, H., 2000, A reality check for data snooping. *Econometrica*, *68*, 1097–1126]. Our findings can be summarized as follows: for the case of one-step ahead pairwise comparison, the GARCH(1,1) is beaten by the asymmetric GARCH models. The same finding applies to different longer forecast horizons, although the predictive superiority of asymmetric models is not as striking as in the one-step ahead case. In the multiple comparison case, the GARCH(1,1) model is beaten when compared against the class of asymmetric GARCH, while it is not beaten when compared against other GARCH models that do not allow for asymmetries.

Keywords: Asymmetric; Bootstrap *P*-values; Forecast evaluation; GARCH; Volatility

1. Introduction

The negative correlation between stock returns and volatility is a well established fact in empirical finance (see, e.g., Bekaert & Wu, 2000; Engle & Ng, 1993; Glosten, Jagannathan, & Runkle, 1993; Nelson, 1991; Wu, 2001; Zakoian, 1994 and references therein). Such a phenomenon is known as asymmetric volatility. One explanation for this empirical fact, first emphasized by Black (1976), is the so-called leverage effect, according to which a drop in the value of a stock (negative return) increases the financial leverage; this makes the stock riskier and thus increases its volatility. While often asymmetric volatility is meant as a synonym of leverage effect, another explanation can be in terms of volatility feedback or time-varying

* Corresponding author. Tel.: +44 20 7882 5087.
E-mail addresses: b.awartani@qmul.ac.uk (B.M.A. Awartani), V.Corradi@qmul.ac.uk (V. Corradi).

0169-2070/$ - see front matter © 2004 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2004.08.003

risk premium (see, e.g., Bekaert & Wu, 2000; Campbell & Hentschel, 1992; French, Schwert, & Stambaugh, 1987; Pindyck, 1984; Wu, 2001). If volatility is priced, an anticipated increase in volatility raises the required return on equity. Hence, the leverage and the volatility feedback explanation lead to a different causal nexus; in fact, the former prescribes a causal nexus from return to conditional volatility, while the latter, from conditional volatility to returns.

The early generation of GARCH models, such as the seminal ARCH(p) model of Engle (1982), the GARCH(p,q) of Bollerslev (1986), and their in-mean generalization (Engle, Lilien, & Robins, 1987) have the ability of reproducing another very important stylized fact, which is volatility clustering; that is, big shocks are followed by big shocks. However, only the magnitude of the shock, but not the sign, affects conditional volatility. Therefore, the first generation of GARCH models cannot capture the stylized fact that bad (good) news increase (decrease) volatility. This limitation has been overcome by the introduction of more flexible volatility specifications which allow positive and negative shocks to have a different impact on volatility. This more recent class of GARCH models includes the Exponential GARCH (EGARCH) model of Nelson (1991), the Asymmetric GARCH (AGARCH) of Engle and Ng (1993), the threshold GARCH by Glosten et al. (1993) (GJR-GARCH), the threshold GARCH of Zakoian (1994) (TGARCH), and the quadratic GARCH (QGARCH) of Sentana (1995). Finally, a new class of GARCH models which jointly capture leverage effects and contemporaneous asymmetry, as well as time varying skewness and kurtosis, has been recently introduced by El Babsiri and Zakoian (2001). In a recent paper, Patton (2004) also analyzes the use of asymmetric dependence among stocks; that is, the fact that stocks are more highly correlated during market downturns.

In this paper, we examine the relative out of sample predictive ability of different GARCH models, with particular emphasis on the predictive content of the asymmetric component. The main problem in evaluating the predictive ability of volatility models is that the "true" underlying volatility process is not observed. In the sequel, as a proxy for the unobservable volatility process we use squared returns. As pointed out by Andersen and Bollerslev (1998), squared returns are an unbiased but very noisy measure of volatility. However, in the present context, we are just interested in comparing the relative predictive accuracy. Also, as shown in the next section, the use of squared returns as a proxy for volatility ensures a correct ranking of models in terms of a quadratic loss function. In a related paper, Hansen and Lunde (in press) provide a test for the null hypothesis that no competing model, within the GARCH universe, provides a more accurate out of sample prediction than the GARCH(1,1) model. In their paper, as a proxy for volatility, they use realized volatility, which is the sum of squared intradaily (e.g., 5 min) returns. The rationale is that in the case of continuous semimartingale processes, such as diffusion processes, daily realized volatility converges in probability to the true integrated (daily) volatility (see, e.g., Andersen, Bollerslev, Diebold, & Labys, 2001; Andersen, Bollerslev, Diebold, & Labys, 2003; Barndorff-Nielsen & Shephard, 2001, 2002; Meddahi, 2002). Therefore, as the time interval approaches zero, realized volatility is a model free measure of integrated volatility.

Nevertheless, GARCH processes are discrete time processes and so are not continuous semimartingales; in this case, realized volatility is no longer a consistent estimator of daily volatility, even if, as shown by Hansen and Lunde, it is an unbiased estimator. Thus, in the case of discrete time processes, both squared returns and realized volatility are unbiased for the true underlying volatility, and so, in the quadratic loss function case, both of them ensure a correct ranking of the models. Finally, another way of overcoming the problem of unobservability of the volatility process is the use of economic loss function, based on option pricing, value at risk, or utility function. This is the approach followed by Gonzalez-Rivera, Lee, and Mishra (in press) who also compared the predictive ability of various GARCH models.

It should be pointed out that several empirical studies have already examined the impact of asymmetries on the forecast performance of GARCH models. The recent survey by Poon and Granger (2003) provides, among other things, a interesting and extensive synopsis of them. Indeed, different conclusions have been drawn from these studies.

In fact, some studies find evidence in favor of asymmetric models, such as EGARCH, for the case of

exchange rates and stock returns predictions. Examples include Cao and Tsay (1992), Heynen and Kat (1994), Lee (1991), Loudon, Watt, and Yadav (2000), and Pagan and Schwert (1990). Other studies find evidence in favor of the GJR-GARCH model. See Brailsford and Faff (1996) and Taylor (2001) for the case of stock returns volatility, and Bali (2000) for interest rate volatility. Nevertheless, other authors, including Bluhm and Yu (2000), Brooks (1998), and Franses and Van Dijk (1996) came to the conclusion that the role of asymmetry in forecasting volatility is rather weak. Along the same lines, Doidge and Wei (1998), Ederington and Guan (2000), and McMillan, Speigh, and Gwilym (2000) found that EGARCH does not outperform the simple GARCH model in forecasting volatility indices. A similar finding is obtained by Day and Lewis (1992, 1993) for the case of S&P-100 OEX options and crude oil futures.

In our empirical analysis, we shall proceed in three steps. First, we get a grasp of the predictive ability of the various models by computing their out of sample mean squared errors (MSE). Then, we proceed by computing pairwise comparisons of the various models against the GARCH(1,1). For the case of nonnested models, this is accomplished by constructing the Diebold and Mariano (1995) test. As is well known, the numerator of the DM statistic vanishes in probability when models are nested and the null hypothesis of equal predictive ability is true. Thus, for the case of nested models, pairwise comparisons are performed via the out of sample encompassing tests of Clark and McCracken (2001). Finally, a joint comparison of all models is performed along the lines of the reality check of White (2000). In particular, the benchmark model is the GARCH(1,1) and the null hypothesis is that none of the competing models provides a more accurate prediction than the benchmark does. We consider different classes of competing models, in particular, distinguishing between GARCH models that allow or do not allow for volatility asymmetry. We use daily observations, from January 1990 to September 2001, on the S&P-500 Composite Price Index, adjusted for dividends. Our findings can be summarized as follows: for the case of one-step ahead pairwise comparison, the GARCH(1,1) is beaten by asymmetric GARCH models. The same is true for longer forecast horizons, although the predictive improvement of asymmetric models is not as striking as in the one-step ahead comparison. In the multiple comparison case, the GARCH(1,1) model is beaten when compared against the class of asymmetric GARCH, while it is not beaten when compared against other GARCH models which do not allow for asymmetries. Such a finding is rather robust to the choice of the forecast horizon. Finally, the RiskMetrics exponential smoothing seems to be the worst model in terms of predictive ability.

The rest of this paper is organized as follows. Section 2 discusses the choice of the volatility proxy and shows that, in the case of quadratic loss functions, the use of squared returns as proxy for volatility ensures a correct ranking of models. Section 3 outlines the adopted methodology. The empirical findings are reported in Section 4. Finally, concluding remarks are given in Section 5.

2. Measuring volatility

As volatility is unobservable, we need a good proxy for it. If the conditional mean is zero, then squared returns provide an unbiased estimator of the true underlying volatility process.[1] However, in a very stimulating paper, Andersen and Bollerslev (1998) pointed out that squared returns are a very noisy measure. Hereafter, let $r_t = \ln S_t - \ln S_{t-1}$, where S_t is the stock price, and so r_t is the continuously compounded return of the underlying asset. AB have shown that the R^2 from the regression of r_t^2 over σ_t^2 and a constant (where σ_t^2 is the conditional variance under a given model) cannot exceed 1/3, even if σ_t^2 is the true conditional variance. Hence, AB concluded that low R^2 from such regression cannot be interpreted as a signal of low predictive ability of a given GARCH model, for example. However, in the present context, we are just interested in comparing the relative predictive accuracy of various models. In this case, if the loss function is quadratic, the use of squared returns ensures that we actually obtain the correct ranking of models. On the other hand, this is

[1] If the conditional mean is not zero, then we should use the squared residuals from the regression of r_t on say, a constant and r_{t-1} and/or other regressors. Of course, if we misspecify the conditional mean, such squared residuals are no longer unbiased estimators of the conditional variance.

not true if we use any generic (for example asymmetric) loss function. Let I_{t-1} be the sigma field containing all relevant information up to time $t-1$, and suppose that $E(r_t|I_{t-1})=0$ and $E(r_t^2|I_{t-1})=\sigma_t^{2\dagger}$, so that $\sigma_t^{2\dagger}$ is the true conditional variance process. Also, let $\sigma_{A,t}^2$ and $\sigma_{B,t}^2$ be two candidate GARCH models, of which we want to evaluate the relative predictive accuracy. In the sequel, we shall assume that $E((\sigma_t^{2\dagger})^2)$, $E((\sigma_{A,t}^2)^2)$, $E((\sigma_{B,t}^2)^2)$ are constant and finite (i.e., the volatility process is covariance stationary) and that the unconditional fourth moment (i.e., $E((r_t-E(r_t|I_{t-1}))^4)$ is finite. Sufficient conditions to ensure covariance stationarity, as well as β-mixing, are provided in Carrasco and Chen (2002, Table 1 and p. 27), for "almost" all GARCH processes. Also, from Carrasco and Chen (2002, p. 24) it follows that $E((r_t-E(r_t|I_{t-1}))^4)=E((\sigma_t^2)^2)E(\eta_t^4)$, where $\eta_t=(r_t-E(r_t|I_{t-1}))/\sigma_t$ is the innovation processes. Thus, covariance stationarity and the finite fourth moment of the innovation process ensure that the unconditional fourth moment is finite. In the empirical section, we check the conditions for covariance stationarity, evaluated at the estimated parameters.

If we ignore the issue of parameter estimation error; that is, we suppose that the parameters of $\sigma_{A,t}^2$ and $\sigma_{B,t}^2$ are known, then in the ergodic and geometrically mixing case because of the law of large numbers,

$$\frac{1}{T}\sum_{t=1}^{T}\left(\left(\sigma_t^{2\dagger}-\sigma_{A,t}^2\right)^2-\left(\sigma_t^{2\dagger}-\sigma_{B,t}^2\right)^2\right)$$
$$\xrightarrow{pr} E\left(\left(\sigma_t^{2\dagger}-\sigma_{A,t}^2\right)^2-\left(\sigma_t^{2\dagger}-\sigma_{B,t}^2\right)^2\right),$$

where $E((\sigma_t^{2\dagger}-\sigma_{A,t}^2)^2-(\sigma_t^{2\dagger}-\sigma_{B,t}^2)^2)=E((\sigma_1^{2\dagger}-\sigma_{A,1}^2)^2-(\sigma_1^{2\dagger}-\sigma_{B,1}^2)^2)$ for all t, given covariance stationarity. Now, suppose that $E((\sigma_t^{2\dagger}-\sigma_{A,t}^2)^2)<E((\sigma_t^{2\dagger}-\sigma_{B,t}^2)^2)$, that is, model A provides a more accurate prediction of the true underlying volatility process than model B does, at least in terms of a quadratic loss function. If we replace the unobservable process $\sigma_t^{2\dagger}$ by r_t^2, then any comparison based on quadratic loss function will ensure the correct ranking of models. In fact,

$$E\left(\left(r_t^2-\sigma_{A,t}^2\right)^2\right)=E\left(\left(\left(r_t^2-\sigma_t^{2\dagger}\right)-\left(\sigma_{A,t}^2-\sigma_t^{2\dagger}\right)\right)^2\right)$$
$$=E\left(\left(r_t^2-\sigma_t^{2\dagger}\right)^2\right)+E\left(\left(\sigma_{A,t}^2-\sigma_t^{2\dagger}\right)^2\right),$$

as the cross product is zero, given that $(r_t^2-\sigma_t^{2\dagger})$ is uncorrelated with any I_{t-1}-measurable function. Thus,

$$\frac{1}{T}\sum_{t=1}^{T}\left(\left(r_t^2-\sigma_{A,t}^2\right)^2-\left(r_t^2-\sigma_{B,t}^2\right)^2\right)$$
$$\xrightarrow{pr} E\left(\left(\sigma_{A,t}^2-\sigma_t^{2\dagger}\right)^2\right)-E\left(\left(\sigma_{B,t}^2-\sigma_t^{2\dagger}\right)^2\right),$$

where the right-hand side above is negative, if and only if $E((\sigma_{A,t}^2-\sigma_t^{2\dagger})^2)<E((\sigma_{B,t}^2-\sigma_t^{2\dagger})^2)$. Therefore, the correct ranking of models is preserved when we use the squared returns as a proxy for volatility.

In practice, we do not know the underlying parameters and we need to estimate them. In real time forecasting, a common practice is to split the sample T as $T=R+P$, where the first R observations are used for initial estimation and the last P observations are used for out of sample prediction. If the estimated parameters are $\sqrt{R}$ consistent and $P/R\to\pi<\infty$, then

$$\frac{1}{P}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\hat{\sigma}_{A,t+1}^2\right)^2-\left(r_{t+1}^2-\hat{\sigma}_{B,t+1}^2\right)^2\right)$$
$$=\frac{1}{P}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\sigma_{A,t+1}^2\right)^2-\left(r_{t+1}^2-\sigma_{B,t+1}^2\right)^2\right)$$
$$+O_P(R^{-1/2}),$$

and so the correct ranking is still preserved.

As $\frac{1}{P^{1/2}}\sum_{t=R}^{T-1}((r_{t+1}^2-\hat{\sigma}_{A,t+1}^2)^2-(r_{t+1}^2-\hat{\sigma}_{B,t+1}^2)^2)$ is the basic ingredient for both Diebold Mariano and (White)-Reality Check type tests, it follows that, if $P/R\to 0$, then the use of r_t^2 as a proxy of the underlying volatility, would lead to the choice of the "right" model, as it does not alter the correct comparison of models, at least in terms of a quadratic loss function. Of course, if our objective were the explanation of the fraction of volatility explained by a given GARCH model, then the noisiness of r_t^2, as a proxy of $\sigma_t^{2\dagger}$, would certainly constitute a problem. In fact, as pointed out by Andersen and Bollerslev (1998), even if, say $\sigma_{A,t}^2$ were the "true" model, it could explain no more than 1/3 of the true volatility, whenever r_t^2 is used as

volatility proxy. However, as pointed out above, our interest is in the relative predictive ability of different GARCH models, and so we are just interested in a correct comparison.

Recently, a new proxy for volatility, termed realized volatility has been introduced (see, among others, Andersen et al., 2001, 2003; Barndorff-Nielsen & Shephard, 2001, 2002; Meddahi, 2002). Let ln S_t, $t=1,\ldots,T$ be the series of daily stock prices, and let ln $S_{t+k\xi}$, $k=1,\ldots,m$ and $\xi=1/m$ denote the series of intraday observations.[2] The series of daily realized volatility constructed as $RV_{t,m}=\sum_{k=0}^{m-1}(\ln S_{t+(k+1)\xi}-\ln S_{t+(k)\xi})^2$. If ln S_t is a continuous semimartingale process, and $\sigma^{\dagger 2}(t)$ is the instantaneous volatility of that process, then as $m\to\infty(\xi\to 0)$, $RV_{t,m}\xrightarrow{pr}\int_{t-1}^{t}\sigma^{\dagger 2}(s)\mathrm{d}s$, for $t=1,2,\ldots$ (see the references above), where the right-hand side is called (daily) integrated volatility. Therefore, as the time interval approaches zero, realized volatility provides a measure free, consistent estimator of the true underlying volatility process. The fact that realized volatility is not only an unbiased but also a consistent estimator of the integrated volatility process allows comparison of models in terms of a wide range of loss functions, and not only in terms of quadratic loss functions. However, consistency holds if and only if the underlying (log) stock price is a continuous semimartingale process. GARCH processes are instead discrete time processes, and so realized volatility is not a consistent estimator. Nevertheless, as shown in a recent paper by Hansen and Lunde (in press), even in the case of discrete time data generating processes, realized volatility is an unbiased estimator of the true underlying volatility. Thus, when interested in model comparisons, in principle, there is not much to choose between the use of squared returns and realized volatility as proxy of the true unobservable volatility. In fact, both of them are unbiased estimators.

3. Methodology

We compare the relative predictive ability of the following models: GARCH, of Bollerslev (1986) (including the Integrated GARCH (IGARCH)), ABGARCH (absolute value GARCH) of Taylor (1986) and Schwert (1990), EGARCH (exponential GARCH) of Nelson (1991), TGARCH (Threshold GARCH) of Zakoian (1994), GJR-GARCH of Glosten et al. (1993), AGARCH (Asymmetric GARCH) of Engle and Ng (1993),[3] QGARCH (Quadratic GARCH) of Sentana (1995), and finally, the RiskMetrics Model (J.P. Morgan, 1997). A synopsis of all the models considered is given in Table 2. Note that EGARCH, TGARCH, GJR-GARCH, AGARCH, and QGARCH allow for asymmetry in the volatility process, in the sense that bad and good news are allowed to have a different effect. Also, note that EGARCH, ABGARCH, and TGARCH do not nest (neither are nested in) the GARCH model, while GJR-GARCH, AGARCH, and QGARCH nest the GARCH model (and RiskMetrics is nested in GARCH), at least for the same number of lags. For all models, the lag order has been chosen using both AIC and BIC criteria. As for estimation, in the sequel, we only present the results for the recursive estimation scheme and for one sample size; results for the fixed and the rolling sampling schemes as well as for longer samples are available from the authors. For the relative properties of fixed, recursive, and rolling sampling schemes, see West (1996) and West and McCracken (1998). Hereafter, let $T=R+P$. We use R observations to get a first estimator and to form the first h-step ahead prediction error, then we compute a second estimator using $R+1$ observations and get a second h- step ahead prediction error, and so on until we have a sequence of P h-step ahead prediction errors. The rolling sampling scheme works in an analogous manner, but for the fact that we use a rolling window of R observations;

[2] For example, if we are interested in 5-min observations, then we choose m to be 288.

[3] A more flexible generalization of this model—called Asymmetric Power ARCH—is proposed by Engle and Ng in the same paper. It is expressed as

$$\sigma_t^{\delta}=\omega+\sum_{i=1}^{q}\alpha_i(|u_{t-i}|-\gamma_i u_{t-i})^{\delta}+\sum_{j=1}^{p}\beta_j\sigma_{t-\mathrm{j}}^{\delta},$$

where $d>0,-1<\gamma_i<1(i=1,\ldots,q)$.

The APARCH includes the AGARCH, GJR-GARCH, GARCH, TGARCH, ABGARCH, and others as special cases. Its flexibility leads to some interesting results as evidenced by Giot and Laurent (2001) and by Peters (2001).

that is, at the first step we use observations from 1 to R, in the second from 2 to $R+1$, and finally from R to $R+P-1$. Parameters are estimated by Quasi Maximum Likelihood, using a Gaussian Likelihood. As for the conditional mean, we have tried six specifications which are presented at Table 1. All parameters, but for the intercept, are not significantly different from zero at 1%, some are significant at 5%, and most of them are significant at 10%. Therefore, in the empirical section, we will consider different models for the conditional mean. However, for ease of exposition, in the rest of this section, and without any loss of generality, we shall proceed as if $E(r_t|I_{t-1})=0$.

3.1. Pairwise comparison of nonnested models

We begin by analyzing pairwise comparison of the eight competing models against the GARCH. The null hypothesis is that of equal predictive accuracy, that is

$$H_0: E\left(\left(r_{t+1}^2-\sigma_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\sigma_{k,t+1}^2\right)^2\right)=0,$$

versus

$$H_A: E\left(\left(r_{t+1}^2-\sigma_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\sigma_{k,t+1}^2\right)^2\right)\neq 0,$$

where $\sigma_{0,t+1}^2$ denotes the conditional variance generated by model 0, while $\sigma_{k,t}^2$ denotes the $k-th$ (nonnested) competitor model. For the three models which are nonnested with GARCH (EGARCH, TGARCH, and ABGARCH), we compute Diebold Mariano statistics (Diebold and Mariano, 1995). In fact, for the case of nested models, the DM statistic does no longer have a well defined limiting distribution (see, e.g., McCracken, 2004). The DM statistic is computed as:

$$\mathrm{DM}_P=\frac{1}{\sqrt{P}}\frac{1}{\hat{\mathrm{S}}_P}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\hat{\sigma}_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\hat{\sigma}_{k,t+1}^2\right)^2\right), \tag{1}$$

where $\hat{\mathrm{S}}_P^2$ denotes a heteroskedasticity and autocorrelation robust covariance (HAC) estimator, i.e.,

$$\begin{aligned}\hat{\mathrm{S}}_P^2=&\frac{1}{P}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\hat{\sigma}_{0,t+1}^2\right)^2-\left(r_{t+h}^2-\hat{\sigma}_{k,t+1}^2\right)^2\right)^2\\&+\frac{2}{P}\sum_{\tau=1}^{l_P}w_\tau\sum_{t=R+\tau}^{T-1}\left(\left(\left(r_{t+1}^2-\hat{\sigma}_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\hat{\sigma}_{k,t+1}^2\right)^2\right)\left(\left(r_{t+1-\tau}^2-\hat{\sigma}_{0,t+1-\tau}^2\right)^2-\left(r_{t+1-\tau}^2-\hat{\sigma}_{k,t+1-\tau}^2\right)^2\right)\right)\end{aligned} \tag{2}$$

where $w_t=1-\tau/(l_P-1)$, and $l_P=o(P^{-1/4})$. If the prediction period P grows at a slower rate than the estimation period R, then the effect of parameter estimation error vanishes and the DM_P statistic converges in distribution to a standard normal. On the other hand, if P and R grow at the same rate and parameter estimation error does not vanish, then we need to use a different HAC estimator, which is able to capture the contribution of parameter uncertainty (see West, 1996).[4]

3.2. Pairwise comparison of nested models

In the nested case, one is interested in testing the hypothesis that the larger model does not beat the

Table 1
Alternative GARCH specifications

The conditional mean	
cm(1)	$E(r_t\|F_{t-1})=a_0$
cm(2)	$E(r_t\|F_{t-1})=a_0+a_1\sigma_t^2$
cm(3)	$E(r_t\|F_{t-1})=a_0+a_1(\sigma_t^2)^{0.5}$
cm(4)	$E(r_t\|F_{t-1})=a_0+a_1(\sigma_t^2)^{2/3}$
cm(5)	$E(r_t\|F_{t-1})=a_0+a_1R_{t-1}$
cm(6)	$E(r_t\|F_{t-1})=0$

[4] In the case of quadratic loss function and parameters estimated via OLS, the contribution of parameter estimation error vanishes regardless of the relative rate of growth of P and R. However, in the present context, parameters are estimated by QMLE and, given the nonlinear nature of GARCH processes, QMLE does not "coincide" with OLS.

smaller one. Thus, we state the hypotheses of interest as:

$$H_0' : E\left(\left(r_{t+1}^2 - \sigma_{0,t+1}^2\right)^2 - \left(r_{t+1}^2 - \sigma_{k,t+1}^2\right)^2\right) = 0,$$

versus

$$H_A' : E\left(\left(r_{t+1}^2 - \sigma_{0,t+1}^2\right)^2 - \left(r_{t+1}^2 - \sigma_{k,t+1}^2\right)^2\right) > 0,$$

where $\sigma_{0,t}^2$ denotes the conditional variance generated by a GARCH process, while $\sigma_{k,t}^2$ denotes the $k-th$ competitor model nesting the GARCH. All competitors nest the GARCH, except for RiskMetrics which is nested in the GARCH.[5] In this case, we perform two of the tests described in Clark and McCracken (2001) for comparing the predictive ability of nested models. Let

$$C_{t+1} = \left(\left(r_{t+1}^2 - \hat{\sigma}_{0,t+1}^2\right) - \left(r_{t+1}^2 - \hat{\sigma}_{k,t+1}^2\right)\right) \times \left(r_{t+1}^2 - \hat{\sigma}_{0,t+1}^2\right),$$

and $\bar{C} = \frac{1}{P}\sum_{t=R}^{T-1} C_{t+1}$, then

$$\text{ENC-T} = (P-1)^{1/2}\frac{\bar{C}}{\sqrt{P^{-1}\frac{1}{P}\sum_{t=R}^{T-1}(C_{t+1}-\bar{C})^2}}, \quad (3)$$

the ECN-T test was originally suggested by Harvey, Leybourne, and Newbold, (1998). Also, define the ENC-REG test (originally suggested by Ericsson (1992)), as

$$\text{ENC-REG} = (P-1)^{1/2}\frac{P^{-1}\sum_{t=R}^{T-1}\left(r_{t+1}^2 - \hat{\sigma}_{0,t+1}^2\right)\left(\left(r_{t+1}^2 - \hat{\sigma}_{0,t+1}^2\right) - \left(r_{t+1}^2 - \hat{\sigma}_{k,t+1}^2\right)\right)}{\sqrt{P^{-1}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2 - \hat{\sigma}_{0,t+1}^2\right) - \left(r_{t+1}^2 - \hat{\sigma}_{k,t+1}^2\right)\right)^2\left(P^{-1}\sum_{t=R}^{T-1}\left(r_{t+1}^2 - \hat{\sigma}_{0,t+1}^2\right)^2\right) - \bar{C}}} \quad (4)$$

If the prediction period P grows at a slower rate than the estimation period R, then both ENC-T and ENC-REG are asymptotically standard normal, under the null. This is true regardless of the sampling scheme we employ. For the case in which P and R grow at the same rate, Clark and McCracken (2001) show that ENC-T and ENC-REG have a nonstandard limiting distribution which is a functional of a Brownian motion; they also provide critical values for various π, where $\pi = \lim_{P,\, R\to\infty} P/R$, and for the three different sampling schemes. Such critical values have been computed for the case in which parameters have been estimated by OLS; therefore, they are not necessarily valid in the case of QML estimation of GARCH processes. For this reason, in the empirical section, we shall limit our attention to the case in which P grows slower than R. In any case, the results from these tests have to be interpreted with caution, especially for longer horizons. In fact, the tests of Clark and McCracken (2001) deal with the one-step ahead evaluation of conditional mean models, and so are based on the difference between the series and its (estimated) conditional mean under a given model. Instead, in the present context, the tests are based on the difference between the square of the series, r_t^2, and its (estimated) conditional variance.[6]

The tests above consider pairwise comparison of two given nested volatility models. If instead, interest lies in testing the null hypothesis that there are no alternative, generic, volatility models producing a more accurate prediction, then one can proceed along the lines of Corradi and Swanson (2002, 2004a). This is left for future research.

[5] In the RiskMetrics case, the formulation of the hypotheses and the statistics should be switched so that GARCH is treated as the larger model.

[6] The case of multistep ahead comparisons of nested model is considered in Clark and McCracken (2004) who showed that the limiting distribution is not free of nuisance parameters and that critical values have to be bootstrapped.

3.3. Multiple comparison of competing models

Finally, along the lines of the reality check test of White (2000), we fix one model (GARCH) as the benchmark and we test the null hypothesis that no competing model, within a given class, can produce a more accurate forecast than the benchmark model does. The alternative hypothesis is simply the negation of the null. The hypotheses of interest are stated as:

$$H_0'': \max_{k=1,\ldots,m} E\left(\left(r_{t+1}^2-\sigma_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\sigma_{k,t+1}^2\right)^2\right)\le 0$$

versus

$$H_A'': \max_{k=1,\ldots,m} E\left(\left(r_{t+1}^2-\sigma_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\sigma_{k,t+1}^2\right)^2\right)>0.$$

The associated statistics is:

$$S_P=\max_{k=1,\ldots,m} S_P(k), \tag{5}$$

where

$$S_P(k)=\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\hat{\sigma}_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\hat{\sigma}_{k,t+1}^2\right)^2\right). \tag{6}$$

From Proposition 2.2 in White (2000), we know that

$$\max_{k=1,\ldots,m}\left(S_P(k)-\sqrt{P}\left(E\left(\left(r_{t+1}^2-\sigma_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\sigma_{k,t+1}^2\right)^2\right)\right)\right)\xrightarrow{d}\max_{k=1,\ldots,m} Z_P(k),$$

where $Z_P=(Z_P(1),\ldots,Z_P(m))$ is a zero mean m-dimensional normal with covariance $\Omega=[\omega_{i,j}]$, $i,j=1,\ldots,m$ and $i,j-th$ element given by

$$\lim_{P\to\infty} E\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\hat{\sigma}_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\hat{\sigma}_{i,t+1}^2\right)^2\right)\times\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}\left(\left(r_{t+1}^2-\hat{\sigma}_{0,t+1}^2\right)^2-\left(r_{t+1}^2-\hat{\sigma}_{j,t+1}^2\right)^2\right)\right).$$

Thus, when $E((r_{t+1}^2-\sigma_{0,t+1}^2)^2-(r_{t+1}^2-\sigma_{k,t+1}^2)^2)=0$ for all k; that is, when all competitors have the same predictive ability as the benchmark, the limiting distribution of S_P is the maximum of a m-dimensional zero mean Gaussian process. In this case, the critical values of $\max_{k=1,\ldots,m} Z_P(k)$ provide asymptotically correct critical values for S_P. On the other hand, when some model is strictly dominated by the benchmark, that is, $E((r_{t+1}^2-\sigma_{0,t+1}^2)^2-(r_{t+1}^2-\sigma_{k,t+1}^2)^2)<0$ for some (but not all) k, then the critical values of $\max_{k=1,\ldots,m} Z_P(k)$ provide upper bounds for the critical values of S_P, thus leading to conservative inference. Finally, when all models are outperformed by the benchmark, that is, $E((r_{t+1}^2-\sigma_{0,t+1}^2)^2-(r_{t+1}^2-\sigma_{k,t+1}^2)^2)<0$ for all k, then S_P approaches minus infinity, and comparing the statistic S_P with the critical values of $\max_{k=1,\ldots,m} Z_P(k)$ leads to conservative inference, characterized by asymptotic size equal to zero. Summarizing, the reality check test is characterized by a composite null hypothesis, and the "exact" limiting distribution is obtained only for the least favorable case under the null. In the other cases, we draw conservative inference. White (2000) has suggested two ways of obtaining valid critical values for the distribution of $\max_{k=1,\ldots,m} Z_P(k)$. One is based on a Monte Carlo approach. We construct an estimator of Ω, say $\hat{\Omega}$, then we draw an m-dimensional standard normal, say η, and we take the largest element of $\hat{\Omega}^{1/2}\eta$. This provides one realization from the distribution of $\max_{k=1,\ldots,m} Z_P(k)$. We repeat this procedure B times, with B sufficiently large, so that we have B draws and can use their empirical distribution to obtain asymptotically valid critical values for $\max_{k=1,\ldots,m} Z_P(k)$. The problem with this approach is that often, especially when the number of models is high and the sample size is moderate. $\hat{\Omega}$ is a somewhat poor estimator for $\hat{\Omega}$ Another approach is based on the bootstrap. One way is to use the stationary bootstrap of Politis and Romano (1994); however, in our empirical study, we shall instead use the block bootstrap of Künsch (1989).[7]

[7] The main difference between the block bootstrap and the stationary bootstrap of Politis and Romano (1994) is that the former uses a deterministic block length, which may be either overlapping as in Künsch (1989) or nonoverlapping as in Carlstein (1986), while the latter resamples using blocks of random length. One important feature of the PR bootstrap is that the resampled series, conditional on the sample, is stationary, while a series resampled from the (overlapping or nonoverlapping) block bootstrap is nonstationary, even if the original sample is strictly stationary. However, Lahiri (1999) shows that all block bootstrap methods, regardless of whether the block length is deterministic or random, have a first-order bias of the same magnitude, but the bootstrap with deterministic block length has a smaller first-order variance. In addition, the overlapping block bootstrap is more efficient than the nonoverlapping block bootstrap.

Let $\hat{f}_{k,t+1}=(r_{t+1}^2)-\hat{\sigma}_{0,t+1}^2)^2-(r_{t+1}^2-\hat{\sigma}_{k,t+1}^2)^2$, $f_{k,t+1}=(r_{t+1}^2-\sigma_{0,t+1}^2)^2-(r_{t+1}^2-\sigma_{k,t+1}^2)^2$ for $k=1,\ldots,m$, and let $\hat{f}_{t+1}=(\hat{f}_{2,t+1},\ \ldots\ ,\hat{f}_{m,t+1})$. At each replication, draw b blocks (with replacement) of length l from the sample $\hat{f}_{k,t+1}$, $t=R,\ \ldots\ ,T-1$, $k=1,\ \ldots\ ,m$, where $P=lb$. Thus, for $k=1,\ \ldots\ ,m$ each of the b blocks is equal to $\hat{f}_{k,i+1},\ \ldots\ ,\hat{f}_{k,i+l}$, for some $i=R,\ \ldots\ ,R+P-l$, with probability $1/(\mathrm{P}-l+1)$.

More formally, let I_i, $i=1,\ \ldots\ ,b$ be *iid* discrete uniform random variables on $[R,\ \ldots\ ,R+P-1]$, and let $P=bl$. Then, the resampled series, $\hat{f}_{t+1}^*=(\hat{f}_{2,t+1}^*,\ \ldots\ ,\hat{f}_{m,t+1}^*)$, is such that $\hat{f}_{R+1}^*,\ \hat{f}_{R+2}^*,\ \ldots\ ,\hat{f}_{R+l}^*,\ \hat{f}_{R+l+1}^*,\ \ldots\ ,\hat{f}_{R+T}^*=\hat{f}_{I_1+1},\hat{f}_{I_1+2},\ \ldots\ ,\hat{f}_{I_1+l},\hat{f}_{I_2},\ \ldots\ ,\hat{f}_{I_b+l}$, and so a resampled series consists of b blocks that are discrete *iid* uniform random variables, conditionally on the sample. If the effect of parameter estimation error vanishes, that is, if P grows at a slower rate than R, then $S_P^*=\max_{k=1,\ldots,m}\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat{f}_{k,t+1}^*-\hat{f}_{k,t+1})$ has the same limiting distribution as $\max_{k=1,\ldots,m}\frac{1}{\sqrt{P}}\sum_{t=R}^{T-1}(\hat{f}_{k,t+1}-E(f_{k,t+1}))$, conditionally on the sample and for all samples but a set of measure zero, provided that at least one competitor is neither nested within the benchmark nor is nesting the benchmark. In fact, if all competitors are nested with the benchmark, then both S_P and S_P^* vanish in probability. Therefore, by repeating the resampling process B times, with B large enough, we can construct B bootstrap statistics, say $S_P^{*(i)}$, $i=1,\ \ldots\ ,B$, and use them in order to obtain the empirical distribution of S_P^*.[8] The bootstrap P-values are then constructed using the formula $B^{-1}\sum_{i=1}^{B}1\{S_P\le S_P^{*(i)}\}$. As usual, a low P-value (large) provides indication against the null (alternative).[9]

In the case in which all competitors are as accurate as the benchmark, then the bootstrap P-values are asymptotically correct. On the other hand, when some competitor is strictly dominated by the benchmark model, then the use of bootstrap P-values leads to conservative inference. In a recent paper, Hansen (2004) explores the point made by White (2000) that the reality check test can have level going to zero, at the same time that power goes to unity, and suggests a mean correction for the statistic in order to address this feature of the test.

4. Empirical results

The data set consists of 3065 daily observations on the S&P-500 Composite Price Index, adjusted for dividends, covering the period from January 1, 1990 to September 28, 2001. Recall that $T=R+P$, where R denotes the length of the estimation period and P the length of the prediction period. In the present context, we set $P=330$, so that $P/R=0.12$.[10] We consider six different prediction horizons, i.e., $h=1$, 5, 10, 15, 20, and 30. The parameters of all models have been recursively estimated by Quasi Maximum Likelihood (QML), using a Gaussian likelihood.[11,12] As for the conditional mean we tried six different specifications: constant mean, zero mean, AR(1), and three GARCH-M specifications, all presented in Table 1. In all cases, the mean parameters, except for the intercept, are not significant at 1%. The lag order has been selected via the AIC and BIC criterion. In almost all cases, the two criteria agree and lead to the choice $p=q=1$.

Hereafter, let $r_t=\ln S_t-\ln S_{t-1}$, where S_t is the stock price index, and so r_t is the continuously compounded return. As proxy of volatility, we use squared returns after filtering returns from the (estimated) conditional mean, for example, if the conditional mean is constant, we use $(r_t-(1/T)\sum_{j=1}^{T}r_j)^2$ as a proxy. Table 2 provides a synopsis of the various GARCH models we consider. Table 3

[8] Note that the validity of the bootstrap critical values is based on an infinite number of bootstrap replications, although in practice we need to choose B. Andrews and Buchinsky (2000) suggest an adaptive rule for choosing B; Davidson and McKinnon (2000) suggest a pretesting procedure, ensuring that there is a "small probability" of drawing different conclusions from the ideal bootstrap and from the bootstrap with B replications, for a test with a given level. In the empirical applications below, we set $B=1000$.

[9] Corradi and Swanson (2004b) provide valid bootstrap critical values for the reality check test in the case of nonvanishing parameter estimation error.

[10] The results for $P=660$, 1034 and so $P/R=0.27$, 05 are available from the authors.

[11] As all models are potentially misspecified, the QML estimators will be in general consistent to some pseudotrue parameters, which, in general, will differ from the true conditional variance parameters. The (quasi) likelihood function has been maximized using the Berndt, Hall, Hall, and Hausman (1974) algorithm.

[12] The results for fixed and rolling estimation schemes are available from the authors.

Table 2
Alternative GARCH specifications

The conditional variance		
(1)	GARCH	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j u_{t-j}^2$
(2)	EGARCH	$\log\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \log\sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j \times \left\{ \gamma \frac{u_{t-j}}{\sqrt{\sigma_{t-j}^2}} + \left[\frac{\lvert u_{t-j} \rvert}{\sqrt{\sigma_{t-j}^2}} - \sqrt{\frac{2}{\pi}} \right] \right\}$
(3)	GJR-GARCH	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} [\beta_j + \gamma_j I_{u_{t-j}<0}] u_{t-j}^2$
(4)	QGARCH	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \gamma_j u_{t-j} + \sum_{j=1}^{q} \gamma_{jj} u_{t-j}^2 + \sum_{i<j}^{q} \gamma_{ij} u_{t-i} u_{t-j}$
(5)	TGARCH	$\sigma_t = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i} + \sum_{j=1}^{q} [\beta_j \max(u_{t-j}, 0) - \gamma_j \min(u_{t-j}, 0)]$
(6)	AGARCH	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j [u_{t-j} - \gamma_j]^2$
(7)	IGARCH	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j u_{t-j}^2$ $1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j = 0$
(9)	RiskMetrics	$\sigma_t^2 = \alpha \sigma_{t-1}^2 + (1-\alpha) u_{t-1}^2$
(10)	ABGARCH	$\sigma_t = \omega + \sum_{i=1}^{p} \alpha_i \sigma_{t-i} + \sum_{j=1}^{q} \beta_j \lvert u_{t-j} \rvert$

reports the conditions for covariance stationarity and β-mixing (Carrasco & Chen, 2002), checked using estimated parameters and replacing expectations with sample mean of residuals, for the case of constant conditional mean. Such conditions are "borderline" satisfied. Thus, recalling the discussion in Section 2, the conditions for the finiteness of the fourth unconditional moments are "borderline" satisfied, provided the innovation process has finite fourth moments. Table 4 displays the in sample and out of sample mean squared errors (MSE) for the various models, for all six conditional mean models considered and for different forecast horizons. As pointed out by Inoue and Kilian (in press), out of sample MSE comparison does not give particularly meaningful information about the relative predictive ability of the various competitors, although it gives an initial overview of model performance. An examination of Table 4 reveals that asymmetric models exhibit a lower MSE. For one-step ahead prediction, the superiority of asymmetric GARCH holds across all conditional means. In particular, EGARCH exhibits the smallest MSE, then followed by other asymmetric models, such as TGARCH, LGARCH, AGARCH, and GJR-GARCH. The worst model is the RiskMetrics exponential smoothing, which is dominated by all other models across all horizons. In the multistep ahead case, asymmetric models still tend to outperform symmetric models, although their superiority is less striking than in the one-step ahead case.

As extreme return observations might have a confounding effect on parameter estimates, and hence, on volatility forecast performance and ranking of models, we recompute the MSEs after removing outliers, see Table 5. Note that MSEs have dropped sharply for all models, but the ranking remained unchanged for all horizons.

We now proceed by performing pairwise comparison of nonnested models, via the Diebold and Mariano statistic defined in (1). More precisely, we compare the GARCH(1,1) model to each of the nonnested competitors. The long run covariance matrix estimator is computed as in (2). This means that we do not take into account the contribution of parameter estimation error to the covariance matrix of the limiting distribution. However, as P/R=0.12, we expect that parameter estimation error should not matter. When performing the DM test, we confine our attention to the comparison of nonnested models. From Table 2, we see that the models, which are nonnested with GARCH, are TGARCH, EGARCH, and ABGARCH. The findings from the DM test are reported in Table 6, for all conditional mean models and across all forecast horizons. The null hypothesis is that of equal predictive accuracy of the two models; a significantly

Table 3
Conditions for covariance stationarity

Model	Condition	Sample value
(1)	$\alpha+\beta<1$	0.997
(2)	$\lvert\alpha\rvert<1$	0.982
(3)	$\alpha+\beta+\gamma*E(\max(0,u/\sigma)^2)<1$	0.985
(5)	$(\alpha+\gamma*\lvert u/\sigma\rvert+\beta*E([\max(0,-u/\sigma)])^2<1$	0.989
(6)	$\alpha+\beta(1+\gamma^2)<1$	0.985
(10)	$(\alpha+\beta*E(\lvert u/\sigma\rvert))^2<1$	0.993

The table presents conditions for stationarity and β-mixing (Carrasco & Chen, 2002) based on sample estimates. Note that the EGARCH is strictly stationary, but not necessarily covariance stationary.

Table 4
Out of sample mean squared error

Model	cm(1)	cm(2)	cm(3)	cm(4)	cm(5)	cm(6)
h=1						
(1)	0.826	0.807	0.812	0.809	0.826	0.829
(2)	0.771	0.758	0.761	0.760	0.772	0.775
(3)	0.778	0.760	0.762	0.761	0.772	0.778
(4)	0.797	0.774	0.773	0.772	0.820	0.793
(5)	0.775	0.763	0.764	0.764	0.775	0.778
(6)	0.788	0.773	0.774	0.774	0.794	0.792
(7)	0.830	0.809	0.811	0.820	0.825	0.835
(8)	0.824	0.825	0.831	0.830	0.845	0.858
(9)	0.823	0.810	0.814	0.813	0.824	0.830
h=5						
(1)	0.840	0.828	0.828	0.829	0.842	0.844
(2)	0.793	0.782	0.784	0.784	0.792	0.796
(3)	0.804	0.789	0.791	0.790	0.800	0.808
(4)	0.813	0.798	0.797	0.797	0.822	0.804
(6)	0.811	0.797	0.799	0.799	0.815	0.815
(7)	0.847	0.830	0.832	0.879	0.846	0.854
(8)	0.843	0.847	0.853	0.851	0.866	0.877
(9)	0.835	0.825	0.828	0.827	0.837	0.842
h=10						
(1)	0.855	0.836	0.840	0.838	0.859	0.853
(2)	0.785	0.774	0.777	0.776	0.786	0.788
(3)	0.803	0.780	0.782	0.770	0.793	0.801
(4)	0.817	0.791	0.792	0.792	0.832	0.820
(6)	0.809	0.793	0.794	0.795	0.813	0.813
(7)	0.859	0.843	0.844	0.876	0.853	0.869
(8)	0.857	0.861	0.867	0.865	0.881	0.893
(9)	0.852	0.839	0.843	0.842	0.853	0.859

Numbers in the table are of order 10^{-7}.

positive (negative) t-statistic indicates that the GARCH(1,1) model is dominated by (dominates) the competitor model. In the one-step ahead case, GARCH is outperformed by EGARCH and TGARCH, while ABGARCH seems to perform as well as GARCH. In the multistep ahead, EGARCH strongly outperforms GARCH across all horizons and conditional means, while ABGARCH does not outperform GARCH in almost all cases and for all horizons.

We now move to pairwise comparison of nested models using the statistics defined in (3) and in (4). The GARCH model is compared to LGARCH, QGARCH, and AGARCH respectively. Findings are reported in Tables 7 and 8. The null hypothesis is that of equal predictive accuracy, the alternative hypothesis is that the competitor (which is the "larger" model) provides a more accurate prediction. Overall, across different horizons and P/R ratios, GARCH

Table 5
Outlier analysis

MSE/cm(1)						
Model	h=1	h=5	h=10	h=15	h=20	h=30
(1)	0.275	0.279	0.283	0.285	0.289	0.291
(2)	0.271	0.275	0.277	0.277	0.283	0.290
(3)	0.270	0.277	0.280	0.281	0.288	0.296
(4)	0.400	0.391	0.370	0.344	0.343	0.336
(5)	0.274					
(6)	0.269	0.273	0.277	0.278	0.285	0.293
(7)	0.278	0.283	0.286	0.289	0.295	0.298
(8)	0.485	0.482	0.474	0.477	0.497	0.527
(9)	0.279	0.282	0.287	0.287	0.293	0.297

Table shows out of sample MSE of models after the removal of aberrant return observations. We removed return observations that were more than three standard deviations from the sample mean of the return series.

is beaten by the competing model. On the other hand, when GARCH is compared against RiskMetrics, it systematically beats it; this clearly results from Table 9. RiskMetrics is a constrained IGARCH (which is a constrained GARCH), and hence, this is not surprising.

Table 6
Diebold Mariano test results

Model	cm(1)	cm(2)	cm(3)	cm(4)	cm(5)	cm(6)
h=1						
(5)	2.861	−0.548	−0.375	−0.681	0.190	−0.078
(2)	3.345	3.031	3.079	3.050	3.110	3.307
(9)	0.468	2.567	2.646	2.608	2.780	2.858
h=5						
(2)	2.826	2.524	2.389	2.485	2.768	2.915
(9)	0.680	0.448	0.052	0.234	0.717	0.253
h=10						
(2)	3.643	2.832	2.813	2.803	3.192	3.486
(9)	0.401	−0.328	−0.332	−0.433	0.685	−0.556
h=15						
(2)	2.650	2.098	2.055	2.056	2.415	2.946
(9)	−1.602	−0.918	−1.192	−1.138	−0.860	−1.301
h=20						
(2)	2.348	1.727	1.577	1.606	2.002	2.367
(9)	0.605	−0.970	−1.496	−1.330	−1.213	−1.465
h=30						
(2)	2.078	1.245	1.293	1.330	1.771	2.465
(9)	−1.425	−1.309	−1.220	−1.158	−1.014	−1.299

Data represents the t statistics of the D&M (1995) test of equal predictive ability against the GARCH (1,1) model.

Finally, we move to the joint comparison of multiple models along the lines of the reality check test of White (2000). The reality check test is performed using the statistic defined in (6), where GARCH(1,1) is the benchmark model, i.e., model 0. Bootstrap P-values are constructed via the empirical distribution of S_P*, as described in the previous section. Bootstrap P-values are computed using 1000 bootstrap replications. We consider three cases:

(i) the benchmark (GARCH) is compared against all models;
(ii) the benchmark is compared against all asymmetric competitors;
(iii) the benchmark is evaluated against all symmetric competitors.

Table 7
Clark McCracken ENC-REG test results

Model	cm(1)	cm(2)	cm(3)	cm(4)	cm(5)	cm(6)
h=1						
(3)	5.482	4.891	5.035	4.778	5.326	5.039
(4)	4.858	4.728	5.061	4.953	1.567	4.704
(6)	5.198	4.612	4.925	4.653	4.375	4.958
h=5						
(3)	4.337	4.311	4.122	4.145	4.420	3.931
(4)	4.504	4.434	4.104	4.323	3.701	5.049
(6)	4.137	4.192	3.927	4.014	3.826	4.080
h=10						
(3)	5.963	5.761	5.815	6.364	6.342	5.211
(4)	6.028	5.903	5.507	5.504	4.600	4.292
(6)	6.040	5.181	5.341	5.106	5.686	5.441
h=15						
(3)	4.291	4.543	4.494	4.106	4.666	3.632
(4)	4.738	5.245	4.820	4.881	3.423	4.101
(6)	4.677	4.721	4.674	4.538	4.768	4.549
h=20						
(3)	3.188	3.329	3.100	3.179	3.258	1.878
(4)	3.807	4.939	4.251	4.336	4.104	3.357
(6)	4.429	4.419	4.187	4.093	4.151	3.891
h=30						
(3)	2.967	2.768	2.908	2.596	3.044	1.990
(4)	3.316	3.869	3.538	3.540	3.690	2.149
(6)	3.655	3.300	3.465	3.314	3.420	3.846

ENC-REG Clark & McCracken (2001) test results for models nesting the GARCH (1,1) model.

Table 8
Clark McCracken ENC-T test results

Model	cm(1)	cm(2)	cm(3)	cm(4)	cm(5)	cm(6)
h=1						
(3)	3.933	3.748	3.641	3.468	3.899	3.340
(4)	4.415	4.492	4.625	4.741	1.856	3.603
(6)	4.714	4.658	4.584	4.655	4.448	3.962
h=5						
(3)	4.283	3.523	3.569	3.107	3.695	2.936
(4)	4.736	3.987	3.931	4.056	3.538	3.545
(6)	4.713	4.055	3.852	3.943	3.838	3.493
h=10						
(3)	3.946	3.450	3.513	3.005	3.354	2.546
(4)	4.478	5.509	5.690	5.681	4.104	3.399
(6)	5.532	5.423	5.715	5.539	5.203	3.905
h=15						
(3)	3.527	3.227	3.170	3.250	3.176	2.141
(4)	5.553	5.209	4.811	4.930	3.462	3.723
(6)	5.808	4.850	4.688	4.638	5.371	4.825
h=20						
(3)	3.194	2.854	2.722	3.205	2.921	1.487
(4)	3.749	4.944	4.065	4.210	4.988	3.650
(6)	5.314	4.382	4.027	4.014	4.352	4.214
h=30						
(3)	4.036	2.891	3.211	3.143	3.623	2.044
(4)	4.251	4.209	3.827	3.965	4.029	2.672
(6)	5.047	3.367	3.755	3.675	3.910	4.451

ENC-T Clark and McCracken (2001) test results for models nesting the GARCH (1,1) model.

Table 9
Clark McCracken test results for RiskMetrics

Horizon	cm(1)	cm(2)	cm(3)	cm(4)	cm(5)	cm(6)
ENCREG						
h=1	−0.387	3.229	3.267	3.452	3.081	3.652
h=5	1.765	3.221	3.804	3.564	3.468	4.057
h=10	0.710	3.857	3.972	4.072	3.207	4.539
h=15	3.723	4.897	5.315	5.258	5.121	6.142
h=20	3.289	4.621	5.347	5.217	5.113	6.192
h=30	3.954	5.014	4.928	4.928	4.745	5.667
ENCT						
h=1	−0.295	2.889	2.823	3.063	2.547	2.774
h=5	1.767	2.772	3.273	3.092	2.832	3.033
h=10	0.436	3.084	3.123	3.249	2.370	3.306
h=15	4.174	4.221	4.490	4.497	4.114	4.697
h=20	3.099	4.129	4.803	4.696	4.355	5.096
h=30	4.510	5.002	4.839	4.813	4.433	5.043

Table 10
Reality check test results

Horizon	RC	p1	p2	p3
cm(1)/Against all				
h=1	0.987	0.003	0.008	0.001
h=5	0.850	0.000	0.004	0.008
h=10	1.272	0.000	0.001	0.004
h=15	1.001	0.000	0.005	0.013
h=20	0.925	0.006	0.015	0.018
h=30	1.007	0.005	0.008	0.030
cm(1)/Against asymmetric				
h=1	0.987	0.007	0.004	0.008
h=5	0.850	0.002	0.009	0.008
h=10	1.272	0.000	0.001	0.003
h=15	1.001	0.000	0.007	0.020
h=20	0.925	0.005	0.007	0.019
h=30	1.007	0.001	0.021	0.030
cm(1)/Against nonasymmetric				
h=1	0.089	0.274	0.263	0.328
h=5	0.073	0.372	0.413	0.425
h=10	0.088	0.366	0.391	0.430
h=15	−0.018	0.948	0.948	0.955
h=20	−0.013	0.926	0.918	0.926
h=30	−0.036	0.964	0.962	0.972
cm(2)/Against all				
h=1	0.884	0.009	0.006	0.006
h=5	0.845	0.009	0.011	0.003
h=10	1.138	0.003	0.006	0.002
h=15	1.055	0.013	0.022	0.003
h=20	0.971	0.019	0.039	0.007
h=30	0.883	0.051	0.09	0.021
cm(2)/Against asymmetric				
h=1	0.884	0.006	0.004	0.011
h=5	0.845	0.007	0.008	0.000
h=10	1.138	0.002	0.006	0.000
h=15	1.055	0.007	0.015	0.001
h=20	0.971	0.012	0.024	0.007
h=30	0.883	0.052	0.059	0.021
cm(2)/Against nonasymmetric				
h=1	−0.004	0.976	0.971	0.974
h=5	0.063	0.631	0.643	0.518
h=10	−0.052	0.904	0.918	0.940
h=15	−0.082	0.995	0.958	0.952
h=20	−0.073	0.931	0.929	0.931
h=30	−0.279	0.999	0.997	0.999
cm(3)/Against all				
h=1	0.920	0.011	0.008	0.012
h=5	0.791	0.010	0.015	0.000
h=10	1.143	0.001	0.007	0.000
h=15	1.018	0.009	0.019	0.007
h=20	0.878	0.028	0.048	0.020
h=30	0.901	0.041	0.071	0.020

Table 10 (*continued*)

Horizon	RC	p1	p2	p3
cm(3)/Against asymmetric				
h=1	0.920	0.008	0.006	0.009
h=5	0.791	0.008	0.023	0.001
h=10	1.143	0.004	0.006	0.001
h=15	1.018	0.010	0.025	0.012
h=20	0.878	0.035	0.040	0.009
h=30	0.901	0.048	0.071	0.019
cm(3)/Against nonasymmetric				
h=1	0.000	0.887	0.885	0.841
h=5	0.000	0.844	0.854	0.811
h=10	−0.050	0.876	0.899	0.889
h=15	−0.196	0.988	0.986	0.987
h=20	−0.294	0.993	0.992	0.997
h=30	−0.319	0.987	0.981	0.993
cm(4)/Against all				
h=1	0.883	0.016	0.007	0.016
h=5	0.817	0.057	0.097	0.023
h=10	1.231	0.090	0.075	0.063
h=15	1.019	0.263	0.276	0.171
h=20	0.892	0.343	0.369	0.283
h=30	0.891	0.425	0.426	0.381
cm(4)/Against asymmetric				
h=1	0.883	0.011	0.007	0.010
h=5	0.817	0.010	0.013	0.006
h=10	1.231	0.010	0.007	0.012
h=15	1.019	0.009	0.060	0.004
h=20	0.892	0.025	0.033	0.011
h=30	0.891	0.039	0.076	0.018
cm(4)/Against nonasymmetric				
h=1	−0.007	0.991	0.994	0.993
h=5	0.033	0.876	0.893	0.823
h=10	−0.071	0.931	0.937	0.919
h=15	−0.195	0.988	0.990	0.993
h=20	−0.281	0.989	0.991	0.994
h=30	−0.327	0.989	0.970	0.995
cm(5)/Against all				
h=1	0.984	0.009	0.003	0.005
h=5	0.888	0.003	0.005	0.002
h=10	1.329	0.002	0.001	0.000
h=15	1.073	0.005	0.012	0.003
h=20	0.942	0.013	0.023	0.004
h=30	1.024	0.013	0.026	0.002
cm(5)/Against asymmetric				
h=1	0.984	0.010	0.005	0.007
h=5	0.888	0.007	0.004	0.002
h=10	1.329	0.001	0.000	0.001
h=15	1.073	0.005	0.012	0.002
h=20	0.942	0.011	0.024	0.005
h=30	1.024	0.009	0.018	0.000

(continued on next page)

Table 10 (*continued*)

Horizon	RC	p1	p2	p3
cm(5)/Against nonasymmetric				
h=1	0.002	0.847	0.853	0.804
h=5	0.102	0.580	0.637	0.499
h=10	0.110	0.522	0.628	0.420
h=15	−0.065	0.952	0.955	0.930
h=20	0.225	0.997	0.995	0.997
h=30	−0.173	0.952	0.951	0.979
cm(6)/Against all				
h=1	0.996	0.014	0.003	0.016
h=5	0.859	0.003	0.003	0.002
h=10	1.176	0.001	0.002	0.003
h=15	0.963	0.015	0.036	0.004
h=20	0.805	0.049	0.097	0.005
h=30	0.986	0.022	0.042	0.002
cm(6)/Against asymmetric				
h=1	0.996	0.011	0.009	0.012
h=5	0.859	0.001	0.003	0.003
h=10	1.176	0.002	0.002	0.003
h=15	0.963	0.004	0.010	0.004
h=20	0.805	0.004	0.005	0.001
h=30	0.986	0.000	0.004	0.000
cm(6)/Against nonasymmetric				
h=1	0.000	0.878	0.881	0.841
h=5	0.039	0.753	0.789	0.680
h=10	−0.106	0.864	0.883	0.901
h=15	−0.288	0.974	0.982	0.991
h=20	−0.427	0.990	0.988	0.997
h=30	−0.465	0.983	0.969	0.996

The table above presents the p values (p1, p2, p3) of White's RC test for three different block sizes: 10, 15, and 5, respectively.

Table 10 reports the findings of the reality check test for different horizons, different conditional mean models and for different block length parameters, i.e., l=10, 15, 20. We recall that the null hypothesis is that no competitor provides a more accurate prediction than the benchmark, while the alternative is that at least one competitor provides a more accurate volatility forecast than the benchmark does.

When GARCH is compared to all models, the evidence is that at least one model in the sample is superior for all conditional mean cases and all horizons. However, this does not provide enough information about which models outperform GARCH. For this reason, in cases (ii) and (iii), we compare GARCH to the universe of asymmetric and symmetric models respectively. It is immediately obvious to see that when GARCH is compared with asymmetric models, the null is rejected (in fact all P-values are below 0.05) across different block lengths, forecast horizons, and conditional mean models. On the other hand, none of the symmetric models outperforms the benchmark. Therefore, there is clear evidence that asymmetries play a crucial role in volatility prediction. In particular, GARCH models that allow for asymmetries in the volatility process produce more accurate volatility predictions, across different forecast horizons and conditional mean models. Finally, for sake of completeness, Table 11 reports the findings for the statistics computed in sample, for the case of

Table 11
Summary for in-sample evaluation

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
MSE	0.547	0.525	0.529	0.533	0.526	0.535	0.548	0.550	0.548
Diebold Mariano		3.808			3.713				−0.301
ENC-REG			10.351	10.373		9.382			
ENC-T			3.794	6.343		5.298			
ENC-REG (RiskMetrics)									4.15
ENC-T (RiskMetrics)									3.99

White test results

	RC	p value1	p value2	p value3
Against all	1.227	0.001	0.000	0.000
Asymmetric	1.227	0.003	0.000	0.001
Nonasymmetric	−0.004	0.995	0.997	0.997

Summary of tests in an in-sample fashion. Diebold and Mariano (1995), Clark McCracken (2001), and White's (2000) Multiple testing results are presented in the table. GARCH (1,1) was taken as the benchmark model. The p values of the RC test (p1, p2, p3) in the table correspond to block sizes: 10, 15, and 20, respectively.

constant mean. The in-sample findings are qualitatively similar to the out of sample ones.

As already mentioned, in a related paper, Hansen and Lunde (in press) compare the GARCH(1,1) model against a universe of GARCH models, using both White's reality check and the predictive ability test of Hansen (2004). For the case of exchange rate data, Hansen and Lunde found that no competitor beats the GARCH(1,1). However, in the case of IBM returns, the GARCH(1,1) model is beaten. Their findings are in line with ours. In fact, the leverage effect is not present in exchange rate data and therefore the GARCH(1,1) model produces predictions which cannot be outperformed. On the other hand, stock returns data incorporate leverage effects, and thus by taking into account the asymmetric behaviour of volatility, one can obtain more accurate predictions.

5. Conclusions

In this paper, we have evaluated the relative out of sample predictive ability of different GARCH models, at various horizons. Particular emphasis has been given to the predictive content of the asymmetric component. The main problem in evaluating the predictive ability of volatility models is that the "true" underlying volatility process is not observed. In this paper, as a proxy for the unobservable volatility process, we use squared returns. As pointed out by Andersen and Bollerslev (1998), squared returns are an unbiased but very noisy measure of volatility. We show that the use of squared returns as a proxy for volatility ensures a correct ranking of models in terms of a quadratic loss function. In our context, this suffices, as we are just interested in comparing the relative predictive accuracy. Our data set consists of daily observations, from January 1990 to September 2001, on the S&P-500 Composite Price Index, adjusted for dividends. First, we compute pairwise comparisons of the various models against the GARCH(1,1). For the case of nonnested models, this is accomplished by constructing the Diebold and Mariano (1995) tests. For the case of nested models, pairwise comparison is performed via the out of sample encompassing tests of Clark and McCracken (2001). Finally, a joint comparison of all models is performed along the lines of the reality check of White (2000). Our findings can be summarized as follows: for the case of one-step ahead pairwise comparison, GARCH(1,1) is beaten by the asymmetric GARCH models. The same finding applies to different longer forecast horizons, although the predictive superiority of asymmetric models is not as striking as in the one-step ahead comparison. In the multiple comparison case, the GARCH model is beaten when compared against the class of asymmetric GARCH, while it is not beaten when compared against other GARCH model which do not allow for asymmetries. Such a finding is rather robust to the choice of the forecast horizon. Finally, the RiskMetrics exponential smoothing model seems to be the model with the lowest predictive ability.

Acknowledgements

We wish to thank an Editor, Mike Clements, two anonymous referees, as well as George Bulkley, Cherif Guermat and the seminar participants at the 2003 UK Study Group in Bristol for helpful comments and suggestions. We gratefully acknowledge ESRC grant RES-000-23-0006.

References

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

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, *96*, 42–55.

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

Andrews, D. W. K., & Buchinsky, M. (2000). A three step method for choosing the number of bootstrap replications. *Econometrica*, *68*, 23–52.

Bali, T. G. (2000). Testing the empirical performance of stochastic volatility models of the short-term interest rate. *Journal of Financial and Quantitative Analysis*, *35*(2), 191–215.

Barndorff-Nielsen, O. E., & Shephard, N. (2001). Non Gaussian OU based models and some of their use in financial economics. *Journal of the Royal Statistical Society. B*, *63*, 167–207.

Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society. B*, *64*, 253–280.

Bekaert, G., & Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies*, *13*, 1–42.

Berndt, E. K., Hall, B. H., Hall, R. E., & Hausman, J. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, *4*, 653–665.

Black, F. (1976). Studies of stock prices volatility changes. *Proceedings of the 976 Meeting of the American Statistical Association, Business and Economic Statistics Section* (pp. 177–181).

Bluhm, H. H. W., & Yu, J. (2000). *Forecasting volatility: Evidence from the German stock market.* Working Paper, University of Auckland.

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

Brailsford, T. J., & Faff, R. W. (1996). An evaluation of volatility forecasting techniques. *Journal of Banking and Finance*, *20*(3), 419–438.

Brooks, C. (1998). Predicting stock market volatility: Can market volume help? *Journal of Forecasting*, *17*(1), 59–80.

Campbell, J. Y., & Hentschel, L. (1992). No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics*, *31*, 281–318.

Cao, C. Q., & Tsay, R. S. (1992, December). Nonlinear time-series analysis of stock volatilities. *Journal of Applied Econometrics*, *Suppl. 1*(S), 165–185.

Carlstein, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics*, *14*, 1171–1179.

Carrasco, M., & Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, *18*, 17–39.

Clark, T. E., & McCracken, M. W. (2001). Tests for equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*, 85–110.

Clark, T. E., & McCracken, M. W. (2004). *Evaluating long-horizon forecasts*. Working Paper, University of Missouri-Columbia.

Corradi, V., & Swanson, N. R. (2002). A consistent test for out of sample nonlinear predictive accuracy. *Journal of Econometrics*, *110*, 353–381.

Corradi, V., & Swanson, N. R. (2004a). Some recent developments in predictive accuracy testing with nested models and nonlinear (generic) alternatives. *International Journal of Forecasting*, *20*, 185–199.

Corradi, V., & Swanson, N. R. (2004b). *Bootstrap procedures for recursive estimation schemes with applications to forecast model selection*. Queen Mary-University of London and Rutgers University.

Davidson, R., & Mackinnon, J. G. (2000). Bootstrap tests: How many bootstraps. *Econometric Reviews*, *19*, 55–68.

Day, T. E., & Lewis, C. M. (1992). Stock market volatility and the information content of stock index options. *Journal of Econometrics*, *52*, 267–287.

Day, T. E., & Lewis, C. M. (1993). Forecasting futures market volatility. *Journal of Derivatives*, *1*, 33–50.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Doidge, C., & Wei, J. Z. (1998). Volatility forecasting and the efficiency of the Toronto 35 index option market. *Canadian Journal of the Administrative Sciences*, *15*(1), 28–38.

Ederington, L. H. & Guan, W. (2000). *Forecasting volatility.* Working Paper, University of Oklahoma.

El Babsiri, M., & Zakoian, J. M. (2001). Contemporaneous asymmetries in GARCH processes. *Journal of Econometrics*, *101*, 257–294.

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

Engle, R. F., Lilien, D. V., & Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, *55*, 391–407.

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

Ericsson, N. R. (1992). Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extension and illustration. *Journal of Policy Modeling*, *14*, 465–495.

Franses, P. H., & Van Dijk, D. (1996). Forecasting stock market volatility using (nonlinear) GARCH models. *Journal of Forecasting*, *15*(3), 229–235.

French, K. R., Schwert, G. W., & Stambaugh, R. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, *19*, 3–29.

Giot, P., & Laurent, S. (2001). *Modelling daily value-at-risk using realized volatility and ARCH type models.* Maastricht University METEOR RM/01/026.

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

Gonzalez-Rivera, G., Lee, T. H., & Mishra, S. (2003). Does volatility modelling really matter? A reality check based on option pricing, VaR, and utility function. *International Journal of Forecasting*, (in press).

Hansen, P. R. (2004). *A test for superior predictive ability.* Working Paper, Brown University.

Hansen, P. R., & Lunde, A. (2003). A forecast comparison of volatility models: Does anything beat a GARCH(1,1). *Journal of Applied Econometrics*, (in press).

Harvey, D. I., Leybourne, S. J., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, *16*, 254–259.

Heynen, R. C., & Kat, H. M. (1994). Volatility prediction: A comparison of stochastic volatility, GARCH(1,1) and EGARCH(1,1) Models. *Journal of Derivatives*, 50–65.

Inoue, A., & Kilian, L. (2003). In sample or out of sample tests for predictability: Which one should we use? *Econometric Reviews*, (in press).

J.P. Morgan. (1997). *RiskMetrics technical documents* (4th ed.). New York.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, *17*, 1217–1241.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, *27*, 386–404.

Lee, K. Y. (1991). Are the GARCH models best in out of sample performance? *Economics Letters*, *37*(3), 9–25.

Loudon, G. F., Watt, W. H., & Yadav, P. K. (2000). An empirical analysis of alternative parametric ARCH models. *Journal of Applied Econometrics*, *2*, 117–136.

McCracken, M. W. (2004). *Asymptotics for out of sample tests of causality.* Working Paper, University of Missouri-Columbia.

McMillan, D. G., Speigh, A. H., & Gwilym, O. A. P. (2000). Forecasting UK stock market volatility. *Journal of Applied Economics*, *10*, 435–448.

Meddahi, N. (2002). A theoretical comparison between integrated and realized volatilities. *Journal of Applied Econometrics*, *17*, 479–508.

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

Pagan, A. R., & Schwert, G. W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, *45*(1–2), 267–290.

Patton, A. (2004). On the out of sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, *2*, 130–168.

Peters, J. P. (2001). *Estimating and forecasting volatility of stock indices using asymmetric GARCH models and (Skewed) Student-t densities.* University of Leige Working Paper.

Pindyck, R. S. (1984). Risk, inflation and stock market. *American Economic Review*, *74*, 334–351.

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

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

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

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

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

Taylor, J. W. (2001). *Volatility Forecasting with smooth transition exponential smoothing.* Working Paper, Oxford University.

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

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

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

Wu, G. (2001). The determinants of asymmetric volatility. *Review of Financial*, 837–859.

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

Biographies: Basel AWARTANI is currently a PhD student in Economics at Queen Mary, University of London. His current research focuses on predictive evaluation, market microstructure effects and jump diffusion processes.

Valentina CORRADI is professor of Econometrics at Queen Mary, University of London. Previously, she held posts at the University of Pennsylvania and at the University of Exeter. Corradi completed her PhD at the University of California, San Diego, 1994. Her current research focuses on densities forecast evaluation, bootstrap techniques for recursive and rolling schemes, testing and modelling volatility processes. She has published in numerous scholarly journals, including Journal of Econometrics, Econometric Theory, Journal of Economic Theory, Econometrics Journal, International Journal of Forecasting, Journal of Time Series Analysis and Macroeconomic Dynamics.