Evaluating the Forecasting Performance of GARCH Models Using White's Reality Check*

Leonardo Souza**
Alvaro Veiga***
Marcelo C. Medeiros****

Abstract

The important issue of forecasting volatilities brings the difficult task of back-testing the forecasting performance. As volatility cannot be observed directly, one has to use an observable proxy for volatility or a utility function to assess the prediction quality. This kind of procedure can easily lead to poor assessment. The goal of this paper is to compare different volatility models and different performance measures using White's Reality Check. The Reality Check consists of a non-parametric test that checks if any of a number of concurrent methods yields forecasts significantly better than a given benchmark method. For this purpose, a Monte Carlo simulation is carried out with four different processes, one of them a Gaussian white noise and the others following GARCH specifications. Two benchmark methods are used: the naive (predicting the out-of-sample volatility by in-sample variance) and the Riskmetrics method.

Keywords: Time series, GARCH models, Bootstrap, Reality check, Volatility, Financial econometrics, Monte Carlo, Forecasting, Riskmetrics, Moving average.

JEL Codes: C45, C51, C52, C61, G12.

*Submited in July 2004. Revised in October 2004. The Reality Check is protected by US Patent 5,893,069, details of which can be obtained with Halbert White. E-mail halwhite@earthlink.net. The authors would like to thank the CNPq for the financial support and the Department of Economics, University of Warwick and the Department of Economic Statistics, Stockholm School of Economics for their kind hospitality. The comments from Jeremy Smith, Dick van Dijk, Timo Teräsvirta, and from an anonymous referee are gratefully acknowledged.

**Brazilian Ministry of Planning.
***Department of Electrical Engineering, Pontifícia Universidade Católica do Rio de Janeiro.
****Department of Economics, Pontifícia Universidade Católica do Rio de Janeiro.

1. Introduction

Modeling and forecasting the conditional variance, or the volatility, of financial time series has been one of the major topics in financial econometrics. Conditional variance forecasts are used, for example, in portfolio selection, derivative pricing and hedging, risk management, market timing, and market making. Among the solutions to tackle this problem, the ARCH (Autoregressive Conditional Heteroskedasticity) model proposed by Engle (1982) and the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) specification introduced by Bollerslev (1986) are certainly among the most widely used and are now fully incorporated into the econometric practice.

However, the important issue of forecasting volatilities brings the difficult task of back-testing the forecasting performance. As volatility cannot be observed directly, one has to use an observable proxy for the volatility or a utility function to assess the prediction quality. This kind of procedure can easily lead to poor assessment. Working with zero mean processes, the most common observable proxy for the volatility is the squared observation, as its expected value is the variance of the process. As pointed out by several authors, in spite of highly significant in-sample parameter estimates, standard volatility models explain very little of the out-of-sample variability of the squared returns (Cumby et al., 1993, Jorion, 1995, 1996, Figlewski, 1997). On the other hand, Andersen and Bollerslev (1998) showed that volatility models do produce strikingly accurate interdaily forecasts when intradaily variance is used as a proxy for volatility; see also Hansen and Lunde (2003). However, intradaily data are, in some cases, very difficult to obtain and the volatility proxy may not be the only explanation for the poor forecasting performance of GARCH models. Another possible source is model misspecification. For example, Teräsvirta (1996) and Malmsten and Teräsvirta (2004) pointed out that the GARCH(1,1) model fails to capture many of the stylized facts of financial time series; see also He and Teräsvirta (1999b) and He and Teräsvirta (1999a). In addition, several papers in the nonlinear time series literature have shown, using simulated data, that, in some cases, even when the correct model is estimated the forecast performance is not statistically different from the ones made by simpler linear models (Clements and Smith, 1997, Lundbergh and Teräsvirta, 2002, Van Dijk et al., 2002).

The goal of this paper is to evaluate the forecasting performance of GARCH models in comparison with simpler methods when different error measures and utility functions are used and when the true data generating process (DGP) is in fact a GARCH process. We check whether a practitioner can have a good assessment of the accuracy of volatility forecasts using the following measures: The root mean squared error (RMSE), the heteroskedasticity-adjusted mean squared error (HMSE), the logarithmic loss (LL), and the likelihood (LKHD). As suggested by a referee, in order to check the effect of the choice of a noisy variable as a proxy for true volatility, we also compare the estimated volatilities with true volatility

and we call this measure $RMSE_{true}$. A Monte Carlo simulation is carried out with four different DGPs: one of which is a Gaussian white noise, whereas the others follow first-order GARCH specifications. The main difference between this paper and others that have appeared recently in the literature [1] is that we use simulated data instead of real time series to check the forecasting performance of GARCH models. We proceed in that way in order to avoid any possible source of model misspecification. To verify if the forecasts are statistically different we use White's Reality Check White (2000).

The Reality Check consists of a non-parametric test that checks if any of a number of concurrent methods yields forecasts significantly better than a given benchmark method. In this paper, two benchmark methods are used: the naive (predicting the out-of-sample volatility by in-sample variance) and the Riskmetrics method (Morgan, 1996) with parameter $\lambda = 0.94$. This choice is based on the fact that the Riskmetrics method is often used as a benchmark in practical applications. The comparison is made by a statistic computed on the out-of-sample errors and respective volatilities. The null hypothesis to be tested is that no method is better than the benchmark.

The main findings of the paper are as follows. First, the choice of the comparison statistics affects the results to a great extent. We would recommend the RMSE and the likelihood for the purpose of comparing volatility forecasts, among the statistics tested here. Second, the forecasting performance of GARCH models increases with an increase in the DGP kurtosis, provided that the DGP is really a GARCH process. Third, the choice of the volatility proxy is also very important in comparing different models. When true volatility is used instead of the squared observations, the results have improved dramatically. This fact is not very surprising and has been discussed in several papers; See, for example, Hansen and Lunde (2003). Finally, beyond the initial motivation of the paper, we find that the Reality Check may not be suitable to compare volatility forecasts within a superior predictive ability framework, and we conjecture that this is due to assumptions made on the test statistic as reported in Hansen (2001). Hansen (2001) proved that the RC suffers from a nuisance parameter problem, causing the results to be sensitive to the inclusion of poor and irrelevant models in the comparison. The author also proposed a new test that compares favorable to White's Reality Check as the former is more powerful and unaffected by poor and irrelevant alternatives. In this paper we decided to keep the original Reality Check test to assess the empirical relevance of the inclusion of poor models in the comparison.

The paper is organized as follows. Section 2 briefly describes the Reality Check, while section 3 describes the experiment and shows some results. Finally, section 4 gives some concluding remarks.

[1] See, for example, Hansen and Lunde (2001).

2. The Reality Check

There are some specific kinds of time series for which there is a benchmark method of forecasting their future observations, in the absence of any overall better method. For instance, one can cite the naive method, behind which lies the random walk model, used as a benchmark in some financial time series. It is desirable to have a forecasting method better than the benchmark, and a comparison between methods is necessary to conclude that a method outperforms the benchmark in a specific series. The comparison is made by using a statistic that stands for the goodness of the predicted observations. Data Mining may compare many methods with the benchmark. However, a question arises: By comparing many methods, what is the probability of a model obtaining a good statistic just by chance? In other words, when the benchmark is the best method, what is the probability of considering another method better than the benchmark, just as a result of (bad) luck? The Reality Check tests for the significance of the best statistic obtained. White (2000) proves that, under some conditions, such as when the series is a stationary strong mixing sequence, the Reality Check converges asymptotically to a 100% power, even with an almost 0% size. However, for finite samples, neither theoretical results nor Monte Carlo realizations are offered.

The Reality Check is a non-parametric hypothesis test with its simplified version consisting of the following: Suppose one wants to predict a time series h-steps ahead over a period and a benchmark method is available. However, one wants to predict even better than the benchmark, and to do so, check many methods against it. Then, one splits the available time period into two parts, in-sample and out-of-sample. The in-sample observations are used to fit a model (whether there is a model behind the method) and the out-of-sample, by means of a measure statistic, to verify the forecast accuracy. If too many methods are tested, there is a chance of at least one method obtaining a statistic better than the benchmark, even when the benchmark method is known to be the best model. Consequently, a critical value for accepting the best statistic must be given. The Reality Check accounts for the increasing number of alternative models being tested, by increasing the critical value as more methods are added to the comparison. This occurs because the best statistic is a maximum, and the bootstrap procedure uses all methods being compared to compute bootstrap maxima, in order to obtain a non-parametric empirical distribution for the maximum (best) statistic under the null. The hypotheses are:

$$\mathrm{H}_0 : \text{No method is better than the benchmark.}$$
$$\mathrm{H}_1 : \text{At least one method is better than the benchmark.}$$

Let F_j be the statistic that accounts for the goodness of fit and f_j its observed value for the fitted model j and corresponding errors. So, f_0 is the statistic for the benchmark method, and $j = 1, \ldots, p$ are the indexes corresponding to the

p models being tested against the benchmark. Let us consider a statistic that increases with the goodness of fit, which means that the higher the statistic, the better is the adjustment (for example, the likelihood). If the statistic decreases with the goodness of fit, the problem is symmetric and one needs only to replace max by min and $<$ by $>$ in the following formulas to obtain the same results. Since the test is non-parametric, it does not require the chosen statistic to belong to a special probability density family. A new statistic V_j is defined as follows:

$$V_j = F_j - F_0 \tag{1}$$

which means that the statistic V_j has a positive expected value conditioned on the method j being better than the benchmark. Let V be the best statistic among the V_js, so that it is defined as follows:

$$V = \max_j V_j \tag{2}$$

The test is then focused on determining the significance of the observed value v of V, as the hypotheses can be written as:

$$\begin{aligned} &\mathrm{H}_0 : E[V] \leq 0 \\ &\mathrm{H}_1 : E[V] > 0 \end{aligned} \tag{3}$$

It is not an easy task to derive the theoretical distribution of V under the null. A non-parametric empirical distribution is computed for V under the null using the Stationary Bootstrap Politis and Romano (1994) applied on the out-of-sample residuals. The Stationary Bootstrap accounts for some dependence left in the residuals, by making the probability of picking contiguous observations conditional on a Bernoulli random variable.

For having a bootstrap distribution of V under the null, it is necessary to have B bootstrap replications v_i^*, $i = 1, \ldots, B$, of $v - E[V]$. In each bootstrap replication, a bootstrap version of the residuals (and the corresponding parameters in the model, e.g., the volatility associated with each point) is generated using the Stationary Bootstrap. This is done using the same bootstrap indexes for all methods. Then, f_{i0}^* and f_{ij}^*, the i^{th} bootstrap replications of f_0 and f_j, $j = 1, \ldots, p$, are computed from these residuals. In order to obtain v_i^*, one must generate all the v_{ij}^*, the i^{th} bootstrap replications of $v_j - E[V_j]$, by doing

$$v_{ij}^* = (f_{ij}^* - f_{i0}^*) - (f_j - f_0) \tag{4}$$

and then

$$v_i^* = \max_j v_{ij}^* \tag{5}$$

Many (B) instances of v_i^* form a bootstrap distribution for V under the null, attaching equal weights *to* each instance. Sorting all v_i^*, $i = 1, \ldots, B$, into $v_{[i]}^*$ and

picking k such that $v^*_{[k]} \leq v < v^*_{[k+1]}$ gives a p-value for v in the following way:

$$P_{RC} = 1 - \frac{k}{B} \tag{6}$$

Hence, one rejects the null hypothesis and considers v significant if P_{RC} is less than a threshold value (for instance, 0.05 for a 5% significance level).

3. The Experiment and Results

3.1 The models

In this paper two benchmark models are used. The first one consists of predicting the out-of-sample volatility (h_{out}) by the in-sample unconditional variance $\left(\sigma^2_{in}\right)$, herein called the naive method. The second one is the RiskMetrics method, defined by equation (8), with the parameter λ set to 0.94 as suggested in the RiskMetrics manual (Morgan, 1996).

As forecasting alternatives, we considered specifications of the following models/methods:

1. GARCH(p,q)

$$\begin{aligned} y_t &= h_t^{1/2}\varepsilon_t \\ h_t &= \alpha_0 + \sum_{i=1}^{q}\alpha_i\varepsilon^2_{t-i} + \sum_{j=1}^{p}\beta_j h_{t-j} \end{aligned} \tag{7}$$

   where $p > 0$, $\alpha_0 > 0$, $\alpha_i \geq 0$ $(i = 1, \ldots, q)$, $\beta_j \geq 0$ $(j = 1, \ldots, p)$, $\sum_{i=1}^{q}\alpha_i + \sum_{j=1}^{p}\beta_j < 1$, and $\varepsilon_t \sim \text{NID}(0, 1)$.

2. RiskMetrics [RM(λ)]

$$\begin{aligned} y_t &= h_t^{1/2}\varepsilon_t \\ h_t &= (1-\lambda)\varepsilon^2_{t-1} + \lambda h_{t-1} \end{aligned} \tag{8}$$

   where $1 > \lambda > 0$ and $\varepsilon_t \sim \text{NID}(0, 1)$.

3. Moving Average Windows [MA(N)]

$$\begin{aligned} y_t &= h_t^{1/2}\varepsilon_t \\ h_t &= \frac{1}{N}\sum_{i=0}^{N-1} y^2_{t-i} \end{aligned} \tag{9}$$

The following concurrent specifications are used: GARCH(1,1), RiskMetrics with $\lambda = 0.85, 0.97$, and 0.99, and moving averages with $N = 5, 10, 22, 43, 126$, and 252.

3.2 Forecasting performance measures

In order to check the forecasting performance of the concurrent models, we consider four goodness-of-fit measures. The first one is the out-of-sample logarithm of the normal likelihood (LKHD). The best predictor is considered the one with the highest value of the log-likelihood in the out-of-sample period defined as

$$LKHD = -\frac{1}{2}\sum_{t=t_0+1}^{T}\frac{y_t^2}{\hat{h}_{jt}} - \sum_{t=t_0+1}^{T}\ln\left(\hat{h}_{jt}^{1/2}\right) \tag{10}$$

where y_t is the observation at time t, $\hat{h}_{jt}$ is the estimated volatility at the time t by method j, t_0 is the total observations in the in-sample period and T is the total number of observations. The second measure used is the root mean squared error (RMSE) of the square of the out-of-sample observations. The best predictor is the one with the lowest RMSE of the squared out-of-sample observations given by

$$RMSE = \sqrt{\frac{1}{T-t_0}\sum_{t=t_0+1}^{T}\left(y_t^2 - \hat{h}_{jt}\right)^2} \tag{11}$$

As recommended by a referee, we also consider an RMSE measure using the true volatility h_t instead of y_t^2, defined as:

$$RMSE_{true} = \sqrt{\frac{1}{T-t_0}\sum_{t=t_0+1}^{T}\left(h_t - \hat{h}_{jt}\right)^2} \tag{12}$$

As suggested by Lopez (2001) and Bollerslev et al. (1994), we also use two asymmetric loss functions: The heteroskedasticity-adjusted mean squared error (HMSE) (Bollerslev and Ghysels, 1996) defined as

$$HMSE = \sqrt{\frac{1}{T-t_0}\sum_{t=t_0+1}^{T}\left(\frac{y_t^2}{\hat{h}_{jt}} - 1\right)^2} \tag{13}$$

and the Logarithmic Loss (LL) Pagan and Schwert (1990) given by

$$LL = \sqrt{\frac{1}{T-t_0}\sum_{t=t_0+1}^{T}\left(\log\left(y_t^2\right) - \log\left(\hat{h}_{jt}\right)\right)^2} \tag{14}$$

3.3 Data generating processes

The following DGPs are used in the simulation.

1. Model 1: Gaussian white noise with zero mean and unit variance.
2. Model 2: GARCH(1,1): $\alpha_0 = 0.5 \times 10^{-5}$, $\alpha_1 = 0.25$, $\beta_1 = 0.70$.
3. Model 3: GARCH(1,1): $\alpha_0 = 1.0 \times 10^{-5}$, $\alpha_1 = 0.05$, $\beta_1 = 0.90$.
4. Model 4: GARCH(1,1): $\alpha_0 = 1.0 \times 10^{-5}$, $\alpha_1 = 0.09$, $\beta_1 = 0.90$.

The first GARCH(1,1) specification (model 2) is very interesting because it does not have a well-defined theoretical kurtosis. The second specification (model 3) has kurtosis around three (3.16). Finally the last GARCH specification (model 4) has a high kurtosis (16.14).

In-sample and out-of-sample vary in length throughout the simulations. The in-sample sizes are 1000, 5000, and 15000, and the respective out-of-sample sizes are 200, 500 and 1000.

3.4 Parameter estimates

Brooks et al. (2001) pointed out that the GARCH parameter estimates are quite different depending on the software used to estimate them. To check the precision of the parameter estimates used in our experiment we conducted a Monte Carlo simulation to check the quality of the estimation algorithm implemented in Matlab. We simulated 1,000 replications of the GARCH(1,1) models defined above and estimated the parameters. Table 1 shows the mean and the standard deviation of the estimates over the Monte Carlo replications. As can be seen, the maximum likelihood estimation leads to very precise parameter estimates for the in-sample lengths used if the DGP is a GARCH(1,1). However, it is somewhat imprecise when a Gaussian white noise generates the data.

Table 1
Mean and standard deviation of the GARCH parameter estimates for models 1–4

	1000 observations			
	Model 1	Model 2	Model 3	Model 4
α_0	0.54 (0.31)	5.40×10^{-6} (1.51×10^{-6})	1.84×10^{-5} (2.62×10^{-5})	1.42×10^{-5} (7.58×10^{-6})
α_1	0.01 (0.02)	0.25 (0.04)	0.05 (0.02)	0.09 (0.02)
β_1	0.45 (0.31)	0.69 (0.04)	0.85 (0.14)	0.89 (0.02)
	5000 observations			
	Model 1	Model 2	Model3	Model 4
α_0	0.67 (0.31)	5.10×10^{-6} (6.07×10^{-7})	1.06×10^{-5} (3.01×10^{-6})	1.08×10^{-5} (2.44×10^{-6})
α_1	6.4×10^{-3} (8.6×10^{-3})	0.25 (0.02)	0.05 (0.01)	0.09 (0.01)
β_1	0.33 (0.31)	0.70 (0.02)	0.90 (0.02)	0.90 (0.01)
	15000 observations			
	Model 1	Model 2	Model3	Model 4
α_0	0.66 (0.30)	5.02×10^{-6} (3.63×10^{-7})	1.02×10^{-5} (1.53×10^{-6})	1.03×10^{-5} (1.35×10^{-6})
α_1	3.0×10^{-3} (4.6×10^{-3})	0.25 (0.01)	0.05 (4.5×10^{-3})	0.09 (4.5×10^{-3})
β_1	0.34 (0.29)	0.70 (0.01)	0.90 (0.01)	0.90 (0.01)

3.5 Forecasting results

Table 2 shows the number of times where each model is the best one according to the forecasting performance measures described in subsection 3.2. When the true DGP is a white noise (model 1), it is interesting to observe that, according to the RMSE and LKHD, the GARCH (1,1) model and the naive method have almost the same performance. When the LL statistic is used, the results are not conclusive and several alternatives have equivalent forecasting performances. When the HMSE is considered, the naive method has the best forecasting performance. However, when the true volatility is used instead of the squared observations, the naive method is, as expected, the best ranked one.

The results concerning a GARCH(1,1) process with no theoretical kurtosis (model 2) point the GARCH(1,1) model as the best forecaster when the RMSE, the $RMSE_{true}$, the LKHD, and the HMSE are used. Note that the likelihood and the $RMSE_{true}$ choose the GARCH(1,1) *in* a hundred percent of the cases. However, the LL points the MA(5) as the best forecasting alternative.

When a GARCH(1,1) process with kurtosis around three is used as DGP (model 3), the RMSE, the $RMSE_{true}$, and LKHD point the GARCH(1,1) model as having superior forecasting ability. When the HMSE is considered, the naive method wins the horse-race 337 times, having a similar performance as the GARCH (1,1) model (412 times). Again the LL leads to results that make no sense, being not suitable to compare volatility forecasts.

By analyzing the results concerning Model 4, one may observe that they are very similar to the previous case (model 3). The major difference is that, using the RMSE, the number of times where the GARCH(1,1) is chosen as the best model falls by approximately a quarter.

Table 2 depicts the winning percentages of each model for each statistic and DGP. However, it gives no idea about the significance of these wins. We then proceed by using the Reality Check with significance levels $0.01, 0.02, \ldots$, and 0.2. The RC experiment depicts the significance of the wins, but does not picture the winning method, being table 2 and the RC results complimentary to each other. Figures 1–4, panels d, e, and f show the percentage of cases where the null hypothesis is rejected for the four DGPs, using the naive method as the benchmark. Panels a, b, and c, in turn, are shown solely to illustrate how the inclusion of poor models affects the RC ability to detect forecasting quality, as they include the MA(5), the MA(10) and the RM(0.85) in the comparison. Hence the main results of the paper concern only panels d, e and f, while the remaining, panels a, b and c relate to the secondary result. Figure 1 shows the results for a white noise as the DGP. One would expect rejection percentages close to the 45^o line, since no method captures the volatility dynamics better than the benchmark. However, this behavior is observed only for the HMSE for the smallest sample size. As the sample size increases the HMSE tends to detect fewer cases where some model would forecast significantly better than the benchmark. The LL has proven

Table 2
Number of times where each model is the best model according to each statistic

Model	Model 1 – 15000 observations					Model 2 – 15000 observations				
	RMSE	$RMSE_{true}$	LKHD	LL	HMSE	RMSE	$RMSE_{true}$	LKHD	LL	HMSE
GARCH(1,1)	426	266	425	163	99	791	1000	1000	2	792
RM(0.85)	0	0	0	148	0	156	0	0	17	0
RM(0.94)	0	0	0	3	0	10	0	0	0	1
RM(0.97)	1	0	1	5	1	0	0	0	0	0
RM(0.99)	17	0	21	3	95	0	0	0	0	1
MA(5)	0	0	0	138	0	18	0	0	936	0
MA(10)	0	0	0	177	0	24	0	0	45	0
MA(22)	0	0	0	79	0	1	0	0	0	0
MA(43)	0	0	0	51	0	0	0	0	0	0
MA(126)	7	0	5	40	36	0	0	0	0	0
MA(252)	50	0	47	85	233	0	0	0	0	0
Naïve	499	734	501	108	536	0	0	0	0	206

Model	Model 3 – 15000 observations					Model 4 – 15000 observations				
	RMSE	$RMSE_{true}$	LKHD	LL	HMSE	RMSE	$RMSE_{true}$	LKHD	LL	HMSE
GARCH(1,1)	896	1000	942	1	412	675	1000	967	0	496
RM(0.85)	1	0	0	235	0	15	0	0	242	0
RM(0.94)	10	0	2	2	4	195	0	24	1	32
RM(0.97)	10	0	11	0	64	31	0	7	0	28
RM(0.99)	33	0	17	0	129	0	0	0	0	34
MA(5)	0	0	0	231	0	4	0	0	256	0
MA(10)	0	0	0	352	0	15	0	0	450	0
MA(22)	0	0	0	58	0	55	0	2	20	0
MA(43)	1	0	0	7	0	10	0	0	1	0
MA(126)	8	0	5	2	5	0	0	0	0	0
MA(252)	11	0	6	10	49	0	0	0	4	0
Naïve	30	0	17	102	337	0	0	0	26	410

unreliable in table 2, and rejects the null hypothesis far more than the significance level would tell. The RMSE, $RMSE_{true}$, and the LKHD barely rejected the null.

Figures 2–4 show the results for DGPs 2–4, all of them GARCH(1,1). Note that their respective kurtoses are not defined, 3.16 and 16.14. The percentage of null hypothesis rejections increases with the DGP kurtosis. Furthermore, an increase in the sample size seems to favor more the RMSE and the LKHD than the HMSE. The HMSE rejects the null at most 55% of the times, for the greatest sample size, for model 2, and a significance level of 0.2, whereas the RMSE attains 68% and the likelihood 97% for the same model and sample size but a significance level of only 0.01. The RMSE and the LKHD have fairly comparable performance, with the latter slightly beating the former. The low DGP kurtosis (figure 3) makes it hard to detect forecast performance superiority when a noisy variable is used as a proxy for true volatility. In fact, the statistics, apart from the LL and the $RMSE_{true}$, have rejection percentages around the 45^o line. The LL, in general and especially for the smallest sample size and confidence levels, rejects the null more often than any other statistic, but, as pointed out before, is not a reliable statistic for comparing volatility forecasts. When the high DGP kurtosis is considered (figure 4), the performance of the RMSE and the likelihood improved dramatically. As expected, the $RMSE_{true}$ rejects the null 100% of the time in almost all the cases considered.

Figures 5–8 show the same as figures 1–4, but with the RiskMetrics with parameter $\lambda = 0.94$ as the benchmark, instead of the naive method. Again, panels a, b, and c are secondary while c, d and f refer to the main results. Differences in the forecast performance in this case (RiskMetrics as the benchmark) tend to be smaller and consequently harder to detect than in the previous case (the naive method as the benchmark) since the RiskMetrics volatility dynamics, even using fixed parameter λ, is not too different from the DGP specifications. Moreover, this case is more realistic than the previous one since no one will use a white noise as benchmark if one suspects there is any dynamics in the volatility. Figure 5 refers to the case where a Gaussian white noise is the DGP. The number of times the RMSE and the LKHD reject the null increases with sample size, the RMSE being always better. This increase occurs with less intensity for the HMSE, while the number of rejections actually decreases for the LL. Figure 6 relates to model 2. It is the highest rejection proportion among figures 5–8, although less than the rejections shown in figure 2, which refers to the naive as the benchmark. In this case the LKHD outperforms the RMSE. The HMSE seems to be insensitive to changes in the sample size and the results concerning the LL statistic are not as strange as before. Figures 7–8 refer to models 3 and 4 as the DGPs. The RMSE and the likelihood fail to detect significant difference in the forecasting performance between any method and the RiskMetrics in a proportion higher than the RC significance level, particularly when model 3 is the DGP. The exception is the likelihood for model 4 as the DGP and significance level higher than 0.1. In these

cases the HMSE outperforms the RMSE and the likelihood, although without any improvements with sample size increases. The same statement would apply to the LL if someone would trust it as comparison statistics for volatility forecast. Again, as expected, the $RMSE_{true}$ rejects the null 100% of the time in almost all the cases considered. Remember that the RC results must be complemented by those shown in table 2, which do not demonstrate good performance of the LL.

When we include the MA(5), the MA(10), and the RM(0.85) (poor) methods, the change in results is dramatic and can be seen in panels a, b and c of Figures 1–8. The statistics, apart from the $RMSE_{true}$, cannot distinguish forecasting performance properly using the RC, unless the DGP has high kurtosis and the benchmark is as naive as the naive method. This result illustrates the statement that the inclusion of poor methods in the comparison negatively affects the RC as explored in Hansen (2001). Hansen (2001) shows that when poor methods, with errors with large expected values and standard deviations, such as the MA(5) and the RM(0.85), are included in the comparison, the Reality Check can be undersized and have little power. This is due to approximating the composite null hypothesis $E[V_j] \leq 0$ by the simple hypothesis $E[V_j] = 0$ to construct the statistic distribution under the null.

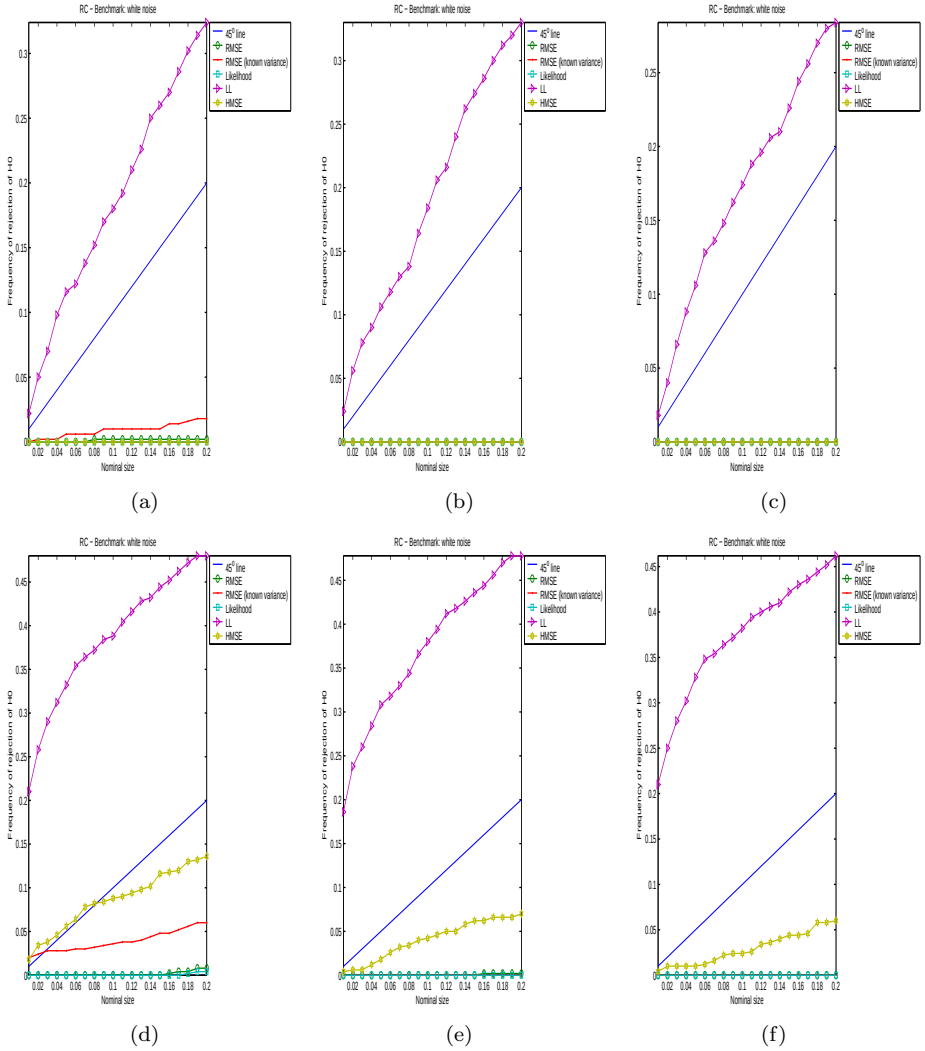

Figure 1
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 1. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

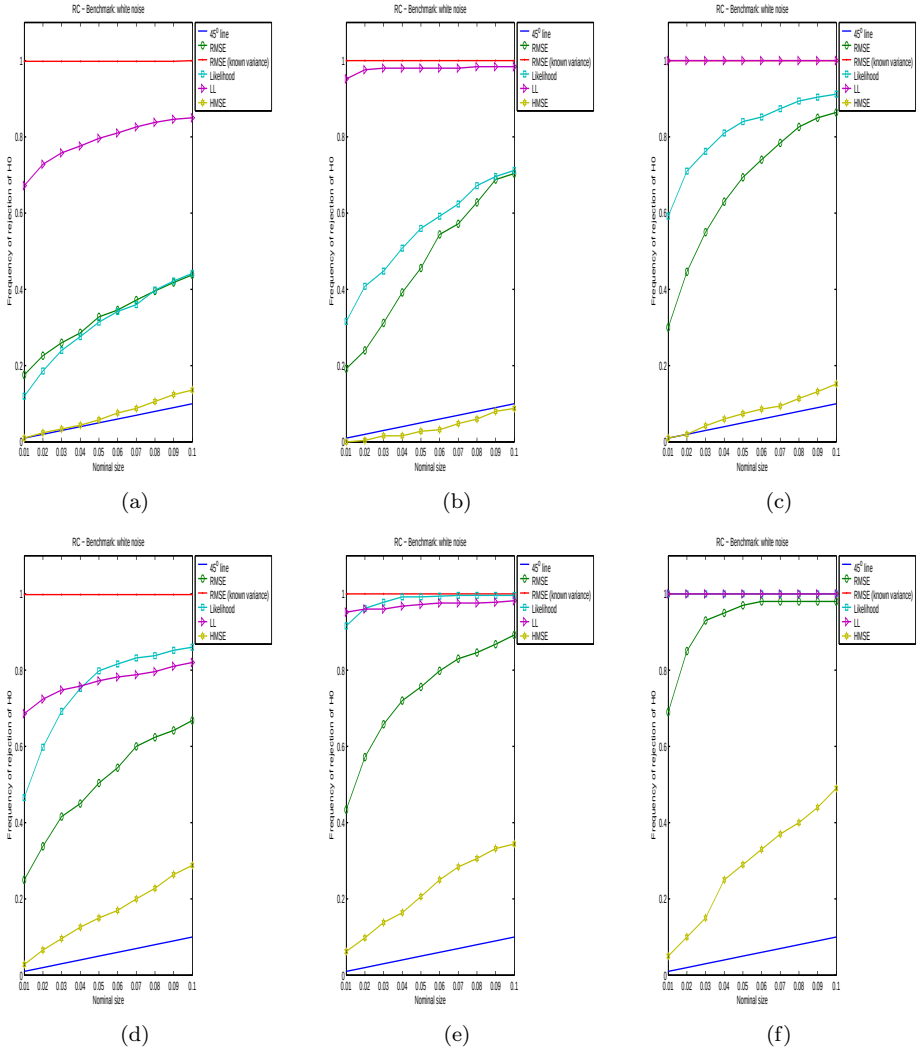

Figure 2
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 2. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

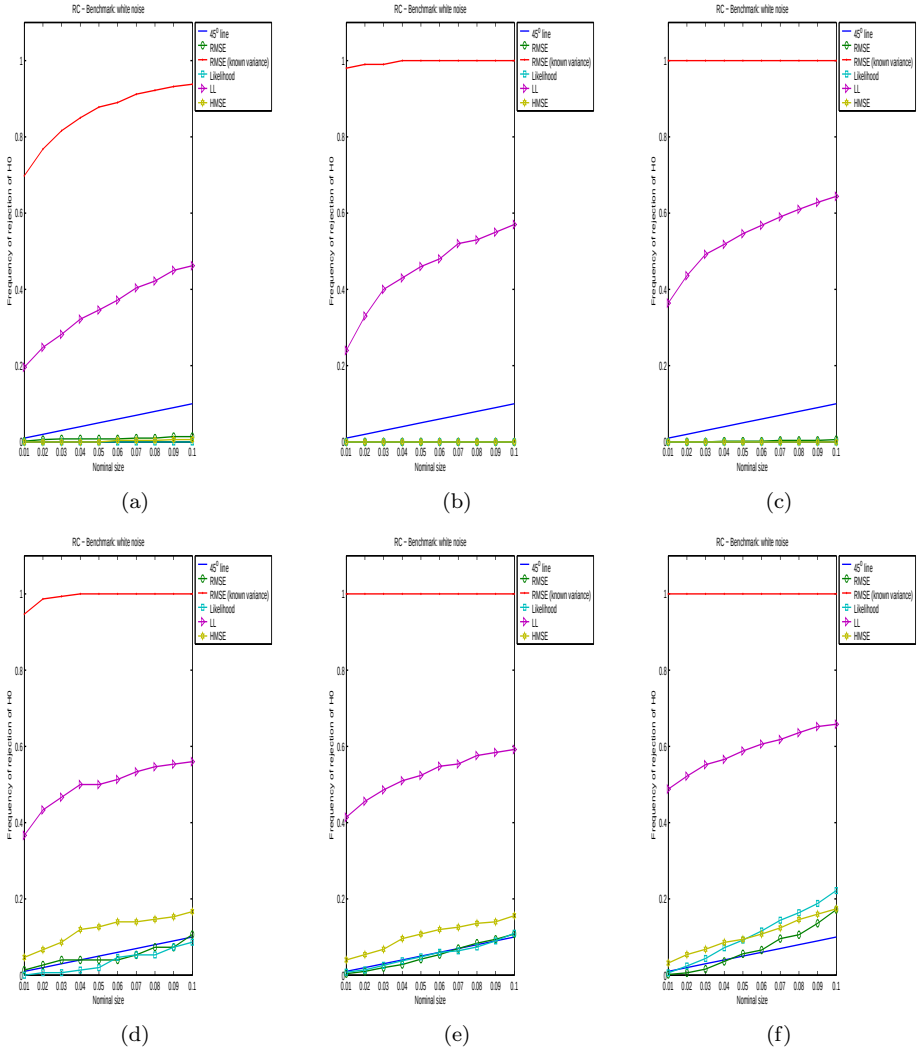

Figure 3
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 3. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

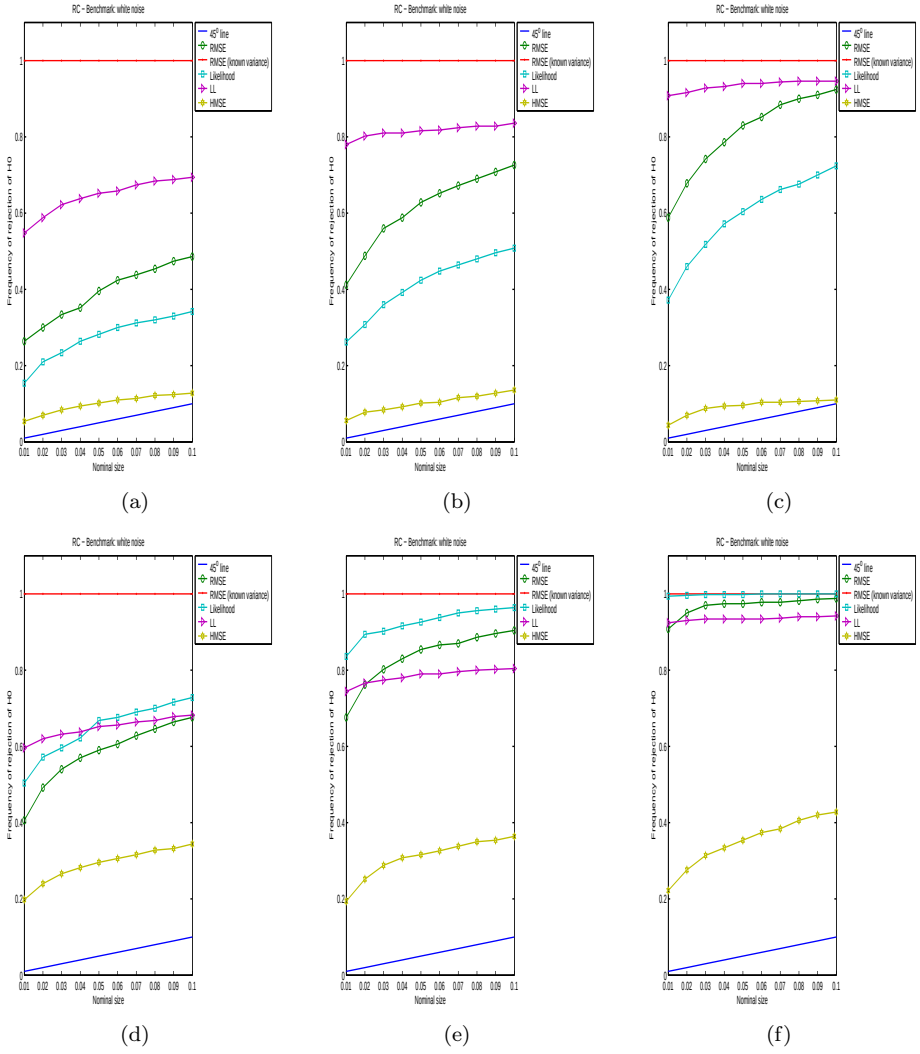

Figure 4
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 1. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

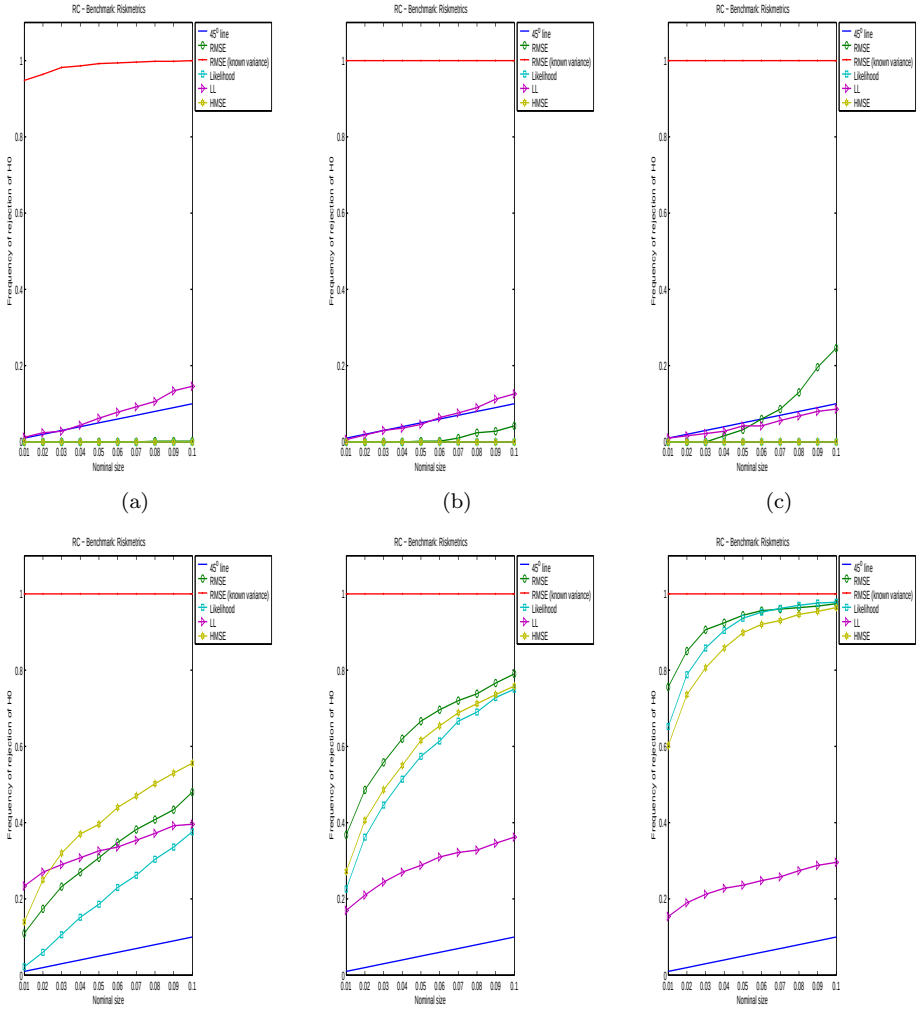

Figure 5
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 1. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

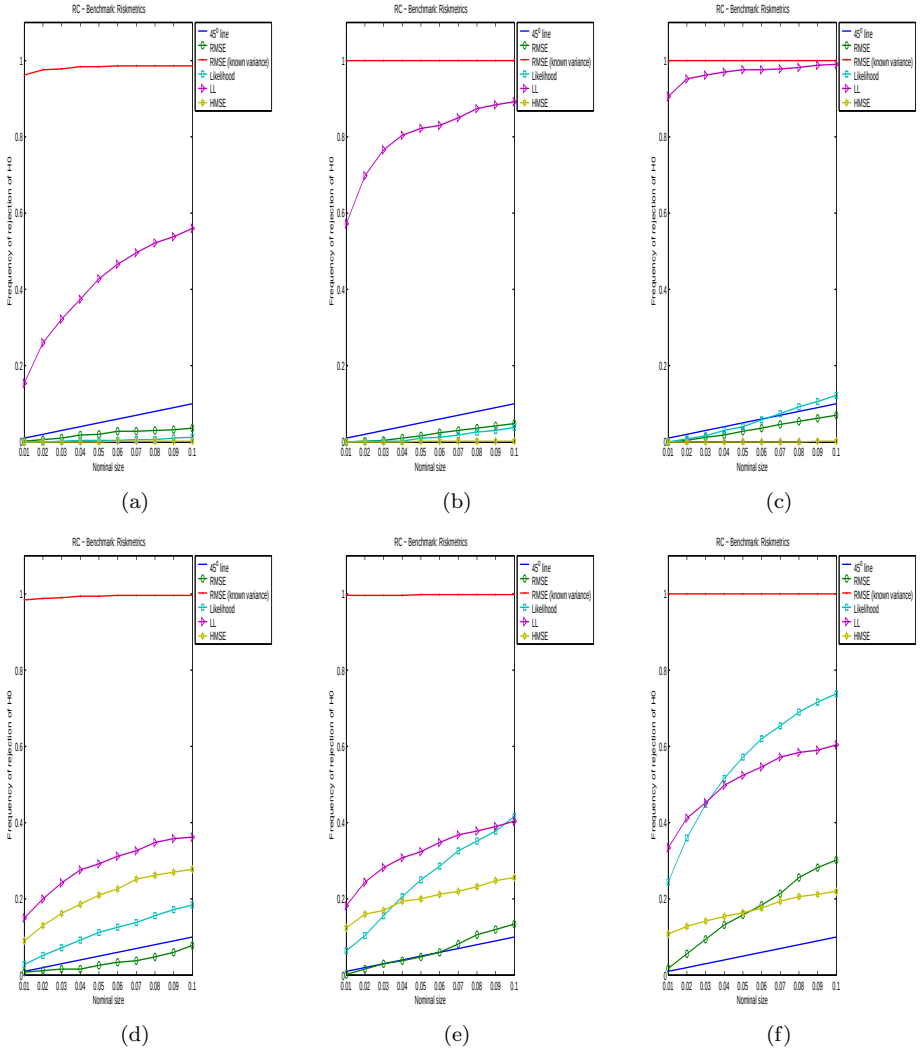

Figure 6
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 2. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

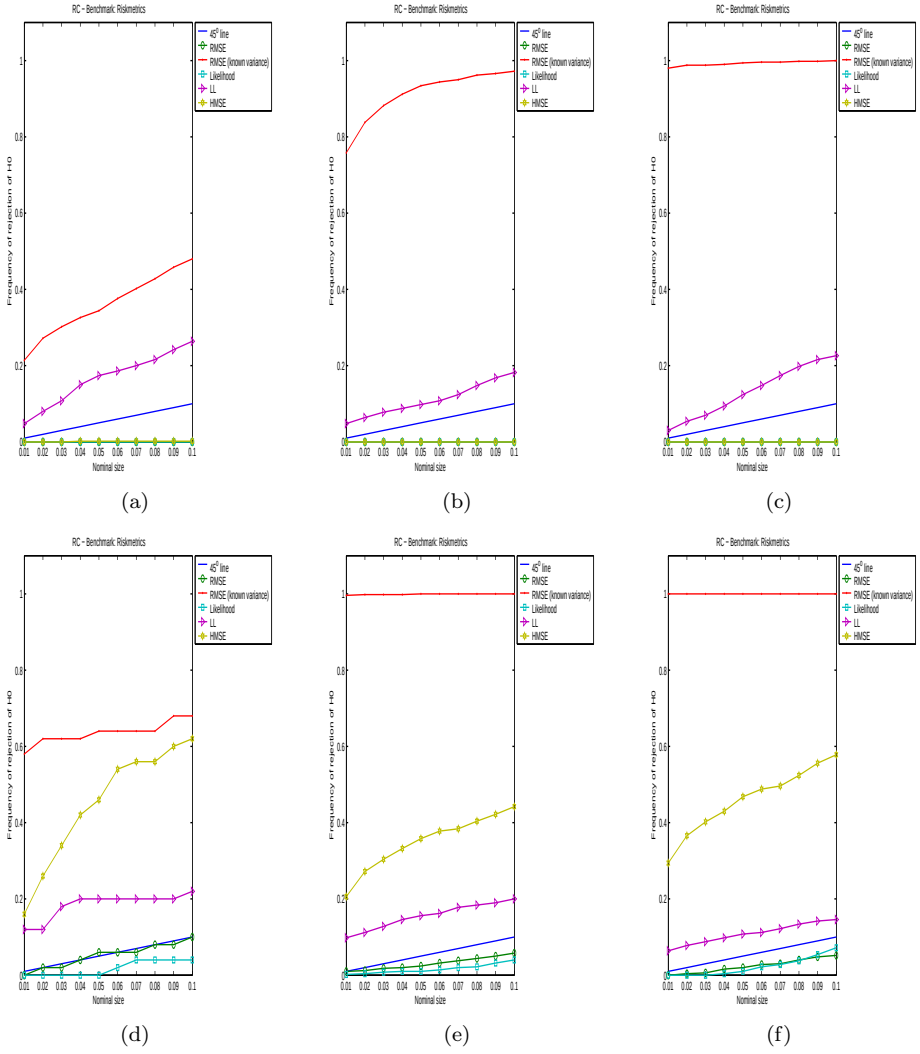

Figure 7
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 3. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

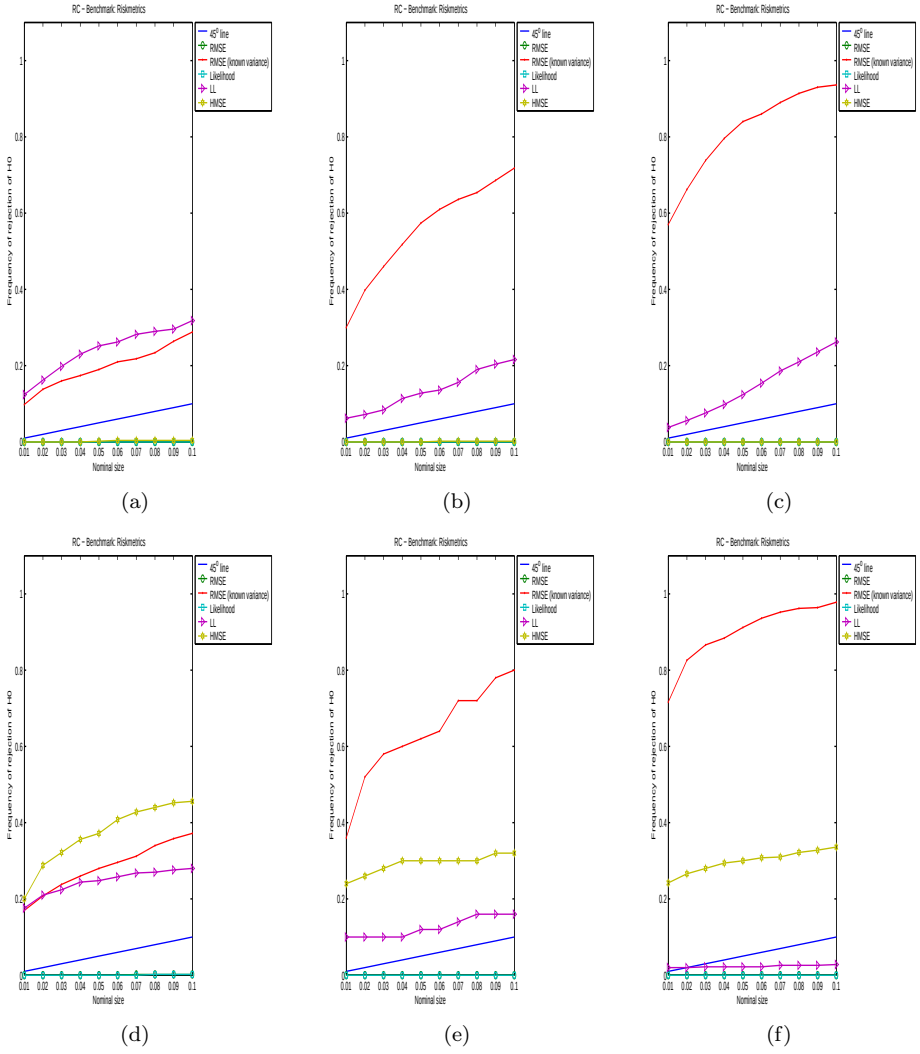

Figure 8
Frequencies of the cases where any of the concurrent models/methods are better than the benchmark for different significance levels of the Reality Check test when data are generated according Model 4. Panel (a) refers to 1000 observations. Panel (b) refers to 5000 observations. Panel (c) refers to 15000 observations. Panel (d) refers to 1000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (e) refers to 5000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation. Panel (f) refers to 15000 observations with RM(0.85), MA(5), and MA(10) removed from the simulation.

4. Conclusions

In this paper, we compared volatility forecasts using White's Reality Check White (2000), using five different measures. For this purpose, a Monte Carlo simulation was carried out with four different processes, one of them a Gaussian white noise and the others following GARCH specifications. As benchmark methods we used the naive (predicting the out-of-sample volatility by in-sample variance) and the Riskmetrics method with parameter $\lambda = 0.94$. The main conclusions are: The choice of the comparison statistics affects the results to a great extent and we would recommend the RMSE and the likelihood for the purpose of comparing volatility forecasts, among the statistics tested in the paper. Particularly, the LL does not prove suitable as a volatility error measure. Second, the ability to distinguish the goodness of volatility forecasts increases with DGP kurtosis. Third, the choice of the proxy for true volatility has a strong effect on the ranking of different models. Finally, the Reality Check may not be suitable to compare volatility forecasts within a superior predictive ability framework, and we relate this to assumptions made on the test statistic. According to the Monte Carlo evidence, we could regard the Reality Check as a very conservative test. Specifically, the test is constructed as if having a simple null hypothesis while it is in fact composite. Hansen (2001) depicts the consequences in detail, showing that the RC suffers from a nuisance parameter, causing the results to be sensitive to the inclusion of poor and irrelevant models in the comparison and producing inconsistent p -values. The author also proposed a new test for comparing different volatility models and we strongly recommend that the practitioner uses Hansen's test instead of White's Reality Check.

References

Andersen, T. & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–906.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 21:307–328.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). ARCH models. In Engle, R. F. & McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 4, pages 2959–3038. North Holland.

Bollerslev, T. & Ghysels, E. (1996). Periodic autoregressive conditional heteroskedasticity. *Journal of Business and Economic Statistics*, 14:139–157.

Brooks, C., Burke, S. P., & Persand, G. (2001). Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting*, 17:45–56.

Clements, M. & Smith, J. (1997). The performance of alternative forecasting methods for SETAR models. *International Journal of Forecasting*, 13:463–475.

Cumby, R., Figlewski, S., & Hasbrouck, J. (1993). Forecasting volatility and correlations with EGARCH models. *Journal of Derivatives*, Winter:51–63.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflations. *Econometrica*, 50:987–1007.

Figlewski, S. (1997). Forecasting volatility. *Financial Markets, Institutions, and Instruments*, 6:1–88.

Hansen, P. R. (2001). A test for superior predictive ability. Department of Economics Working Paper in Economics, 01–06, Brown University.

Hansen, P. R. & Lunde, A. (2001). A forecast comparison of volatility models: Does anything beat a GARCH(1,1) model? *Journal of Applied Econometrics*, forthcoming.

Hansen, P. R. & Lunde, A. (2003). Consistent ranking of volatility models. *Journal of Econometrics*.

He, C. & Teräsvirta, T. (1999a). Properties of moments of a family of GARCH processes. *Journal of Econometrics*, 92:173–192.

He, C. & Teräsvirta, T. (1999b). Properties of the autocorrelation function of squared observations for second order GARCH processes under two sets of parameter constraints. *Journal of Time Series Analysis*, 20:23–30.

Jorion, P. (1995). Predicting volatility in the foreign exchange market. *Journal of Finance*, 50:507–528.

Jorion, P. (1996). Risk and turnover in the foreign exchange market. In Frankel, J. A., Galli, G., & Giovanni, A., editors, *The Microstructure of Foreign Exchange Markets*, pages 19–37. University of Chicago Press.

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

Lundbergh, S. & Teräsvirta, T. (2002). Forecasting with smooth transition autoregressive models. In Clements, M. P. & Hendry, D. F., editors, *A Companion to Economic Forecasting*, pages 485–509. Oxford: Blackwell.

Malmsten, H. & Teräsvirta, T. (2004). Stylized facts of financial time series and three popular models of volatility. Working Paper Series in Economics and Finance 563, Stockholm School of Economics.

Morgan, J. P. (1996). *J. P. Morgan/Reuters Riskmetrics – Technical Document*. J. P. Morgan, New York. J. P. Morgan, New York.

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

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

Teräsvirta, T. (1996). Two stylized facts and the GARCH(1,1) model. Working Paper Series in Economics and Finance 96, Stockholm School of Economics.

Van Dijk, D., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models – a survey of recent developments. *Econometric Reviews*, 21:1–47.

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