ELSEVIER

Journal of Econometrics 69 (1995) 367–391

JOURNAL OF
Econometrics

The predictive ability of several models of exchange rate volatility

Kenneth D. West*[,a], Dongchul Cho[b]

[a]Department of Economics, University of Wisconsin, Madison, WI 53706, USA
[b]Department of Economics, Texas A&M University, College Station, TX 77843, USA

(Received June 1993; final version received May 1994)

Abstract

We compare the out-of-sample forecasting performance of univariate homoskedastic, GARCH, autoregressive, and nonparametric models for conditional variances, using five bilateral weekly exchange rates for the dollar, 1973–1989. For a one-week horizon, GARCH models tend to make slightly more accurate forecasts. For longer horizons, it is difficult to find grounds for choosing between the various models. None of the models perform well in a conventional test of forecast efficiency.

Key words: Conditional heteroskedasticity; ARCH; Exchange rate; Prediction; Forecasting
JEL classification: C52; C53; F31

1. Introduction

This paper compares the forecasting performance of six models for a univariate conditional variance, using bilateral weekly data for the dollar versus the

*Corresponding author.

This research is an outgrowth of work conducted jointly with Hali J. Edison, to whom we are greatly indebted. We also thank Neil Ericcson, Blake Le Baron, Wen-Ling Lin, three anonymous referees, an associate editor of this journal, and participants in a seminar at the Federal Reserve Board of Governors for helpful comments and discussions. West thanks the National Science Foundation, the Sloan Foundation, and the University of Wisconsin Graduate School for financial support.

SSDI 0304407694016541

currencies of Canada, France, Germany, Japan, and the United Kingdom, 1973–1989. The six models include a homoskedastic one, two GARCH models (Bollerslev, 1986; Engle and Bollerslev, 1986), two autoregressions, and a nonparametric one. We compare the out-of-sample realization of the square of the weekly change in an exchange rate with the value predicted by a model of the conditional variance, for horizons of one, twelve, and twenty-four weeks. The measure of performance that we focus on is mean squared prediction error (*MSPE*).

For twelve- and twenty-four-week-ahead forecasts of the squared weekly change, it is difficult to find grounds to choose one model over another. But at a one-week horizon, we find that GARCH models have a slight edge over the other models. The GARCH mean squared prediction errors tend to be slightly smaller, and regressions of realized exchange rate squares on their estimated conditional variances tend to find somewhat more evidence of predictive power. But statistical tests typically cannot reject at conventional significance levels the null that the *MSPE* from the GARCH models is equal to those of other models, and standard regression tests for bias and efficiency strongly reject the null that the GARCH conditional variance differs from the realized exchange rate square by a white noise error. It appears that GARCH models leave something to be desired, even at the one-week horizon.

Other papers have compared univariate volatility models in related frameworks. Using monthly stock return data and a one-month-ahead *MSPE* criterion, Akigray (1989) found GARCH models preferable to naive and ARMA ones, and Pagan and Schwert (1990a) found GARCH and ARMA models preferable to nonparametric and Markov switching ones. While we, too, find that GARCH models perform well, our results complement and extend these earlier ones in three ways.

First, and least important, we use exchange rate instead of stock price data. One would obviously like to know if what works with one type of data works with another as well. Second, we formally test for equality of MSPEs across models, using a straightforward asymptotic technique that may be of general interest. Third, we consider not only one-period but multi-period horizons as well. Since, in the end, we could not reject the null of equality of *MSPE*s across models, and since we found no grounds for preferring one estimator over another at horizons of more than one period, our endorsement of GARCH models is more moderate than it would have been had we not performed these tests and examined these horizons.

Before turning to the analysis, a final remark seems advisable. The literature on conditional volatility has grown enormously in recent years (see Bollerslev et al., 1992, for an excellent survey), and it is simply not practical to simultaneously study every model that has been proposed. While we feel that we have chosen a representative set of models, we recognize that some readers might prefer a different set. We hope that such readers will nonetheless find it useful that our analysis leads us to speculate that successful models will allow for what standard tests suggest is movement in unconditional variances.[1]

Section 2 describes our data and models, and Sections 3 and 4 our empirical results. Section 5 concludes. An Appendix available on request from the authors contains some results omitted from the paper to save space.

2. Data, models, and estimation techniques

2.1. Data

Our exchange rates are measured as dollars per unit of foreign currency, between the U.S. and Canada, France, Germany, Japan, and the United Kingdom.[2] The data are Wednesday, New York noon bid rates, as published in *The Federal Reserve Bulletin.* When Wednesday was a holiday we used Thursday data; when Thursday was a holiday as well we used Tuesday data. After an initial observation was lost due to differencing (see below), the sample for each country included the 863 observations from March 14, 1973 to September 20, 1989. Figs. 1.A1 to 1.A5 plot the levels rather than differences of the series, with the vertical axis measured in cents per unit of foreign currency. Figs. 1.B1 to 1.B5 will be discussed below.

Prior to our formal analysis, we took logarithmic differences of the series, and then multiplied by 100. That is, our exchange rate series is

$$e_t \equiv 100 * \ln(\text{exchange rate in week } t/\text{exchange rate in week } t-1),$$

and thus has the interpretation of percentage change in the level of the exchange rate. With a slight abuse of terminology, we will sometimes refer to our data as 'exchange rates' rather than 'percentage changes in exchange rates'.

Table 1 contains some summary statistics on these data. Most standard errors and p-values in the remainder of the table also are robust to the possible presence of serial correlation and conditional heteroskedasticity, and are computed as described below. Table 1 is consistent with the results of many earlier studies (e.g., Baillie and Bollerslev, 1989; Diebold and Nerlove, 1989; Engle and Bollerslev, 1986). Exchange rate changes appear to have zero unconditional

[1] We find this to also be a message, perhaps implicit, in the studies using stock price data by Pagan and Schwert (1990a, b) and Chou et al. (1991), as well in Loretan and Phillips (1992).

[2] We also obtained Italian data. But in-sample statistics such as those reported in Table 1 suggest a nonzero unconditional mean. Fitted GARCH models tended to be explosive, with $\hat{\alpha} + \hat{\beta} > 1$ in the notation of Table 2; apparently this resulted in part from the nonzero sample mean since removing this mean lessened the tendency to get explosive estimates. We dropped Italy rather than fit means as well as variances.

Table 1
Summary statistics

	Canada	France	Germany	Japan	U.K.
Panel A: e_t					
1. Mean	− 0.020 (0.020)	− 0.044 (0.050)	0.042 (0.056)	0.068 (0.056)	− 0.052 (0.055)
2. Standard deviation	0.552 (0.029)	1.408 (0.053)	1.466 (0.064)	1.361 (0.063)	1.406 (0.080)
3. Skewness	− 0.423 (0.455)	0.103 (0.245)	0.480 (0.203)	0.385 (0.233)	0.261 (0.270)
4. Excess kurtosis	4.981 (2.636)	2.490 (0.846)	2.133 (0.900)	2.587 (0.715)	3.092 (0.857)
5. Modified L-B (10)	7.47 [0.681]	20.89 [0.022]	13.43 [0.200]	20.52 [0.025]	13.57 [0.194]
6. Modified L-B (50)	61.84 [0.122]	63.47 [0.096]	53.74 [0.333]	71.92 [0.023]	54.10 [0.321]
7. Modified L-B (90)	111.99 [0.058]	98.46 [0.254]	87.99 [0.540]	122.26 [0.005]	88.47 [0.526]
8. Minimum	− 4.164	− 6.825	− 4.488	− 6.587	− 5.691
9. Q1	− 0.313	− 0.851	− 0.850	− 0.641	− 0.773
10. Median	− 0.030	0.000	0.023	− 0.027	− 0.034
11. Q3	0.272	0.675	0.855	0.606	0.708
12. Maximum	2.550	7.741	8.113	6.546	7.397
Panel B: e_t^2					
13. Mean	0.305 (0.030)	1.983 (0.148)	2.147 (0.190)	1.854 (0.166)	1.978 (0.223)
14. Standard deviation	0.809 (0.213)	4.196 (0.598)	4.395 (0.763)	4.001 (0.513)	4.446 (0.768)
15. L-B (10)	34.27 [0.000]	37.82 [0.000]	56.72 [0.000]	51.92 [0.000]	98.12 [0.000]
16. L-B (50)	52.50 [0.377]	129.59 [0.000]	134.75 [0.000]	101.16 [0.000]	322.19 [0.000]
17. L-B (90)	65.41 [0.976]	178.42 [0.000]	166.25 [0.000]	138.44 [0.000]	337.07 [0.000]

1) The variable e_t is the percentage change in the weekly exchange rate. The sample includes 863 weekly observations from March 14, 1973 to September 20, 1989.

2) In rows 1–4, 13, and 14, heteroskedasticity and autocorrelation consistent asymptotic standard errors are in parentheses.

3) Rows 5 to 7 and 15 to 17 contain Ljung–Box statistics of order given in the header to the row. In rows 5 to 7, the statistics are computed as in Eq. (4) to allow for possible conditional heteroskedasticity in e_t. The p-values of the asymptotic chi-squared statistics are given in the lower halves of the rows.

means (line 1), and, with the possible exception of Japan, appear to be serially uncorrelated (lines 5 to 7).

Exchange rate changes are also very fat-tailed. This leptokurtosis may be seen in Figs 1.B1 to 1.B5, each of which plots both a normal density whose mean and variance match sample estimates and a histogram of the data. More formal evidence is in panel A of Table 1. The standard deviation of exchange rate changes is about 1% per week (line 2); the maximum and minimum changes in this sample of size 863 are generally five or more standard deviations away from the mean (lines 8 and 12), and the interquartile range is much less than two standard deviations (lines 9 and 11). Excess kurtosis is greater than two and is significantly different from zero at any conventional significance level for all countries except Canada (line 4). With the exception of Germany, there is no evidence of skewness (line 3).

Panel B of Table 1 contains some summary statistics on squared exchange rates. The means and standard deviations (lines 13 and 14) are presented for convenience of interpretation of our empirical results; they are redundant in the sense that the point estimates can be deduced from the appropriate entries in panel A. Rows 15 to 17 in panel B suggest that, in stark contrast to the levels, the squares of exchange rates are highly serially correlated. This, too, is a result consistent with many earlier studies.

2.2. Models and estimation techniques

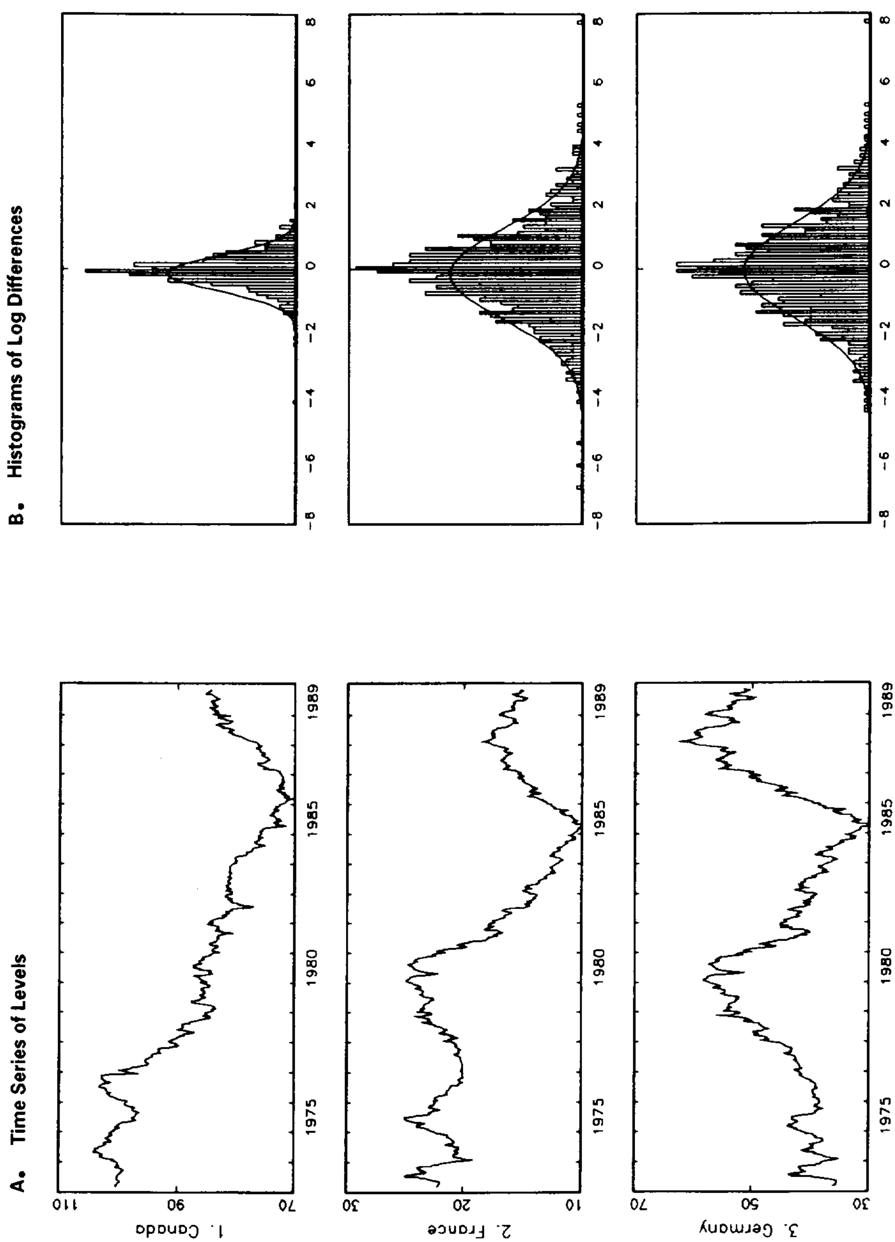
The in-sample evidence in Table 1 that e_t is linearly unpredictable is supported by the stronger results from other studies, some of which use out-of- as well as in-sample evidence, that there is not even any nonlinear dependence in the conditional mean of e_t. The most salient reference is Diebold and Nason (1990), who drew their data from exactly the same source as did we, but over the slightly shorter sample period 1973–1987. Despite much in-sample evidence of nonlinear dependence in the mean of e_t, they found little out-of-sample evidence of such dependence. Papers that come to similar conclusions using other data, sometimes with multivariate information sets, include Meese and Rogoff (1983) and Meese and Rose (1991). We therefore will limit ourselves to models in which the conditional mean of e_t is zero.

To define our models, some notation is needed. Let

$$h_{t,j} = \mathrm{var}_t(e_{t+j}) = \mathrm{E}_t e_{t+j}^2 = \text{(population) variance of } e_{t+j}, \text{ conditional on information generated by past } e_s,\ s \leqslant t; \tag{1a}$$

$$\hat{h}_{mt,j} = \text{fitted conditional variance of } e_{t+j}, \text{ according to model } m \text{ (e.g., model } m \text{ is GARCH(1,1) or homoskedastic), estimated using data on past } e_s,\ s \leqslant t; \tag{1b}$$

$$h_t \equiv h_{t,1}, \qquad \hat{h}_{mt} \equiv \hat{h}_{mt,1}; \tag{1c}$$


A. Time Series of Levels
B. Histograms of Log Differences
1. Canada
2. France
3. Germany

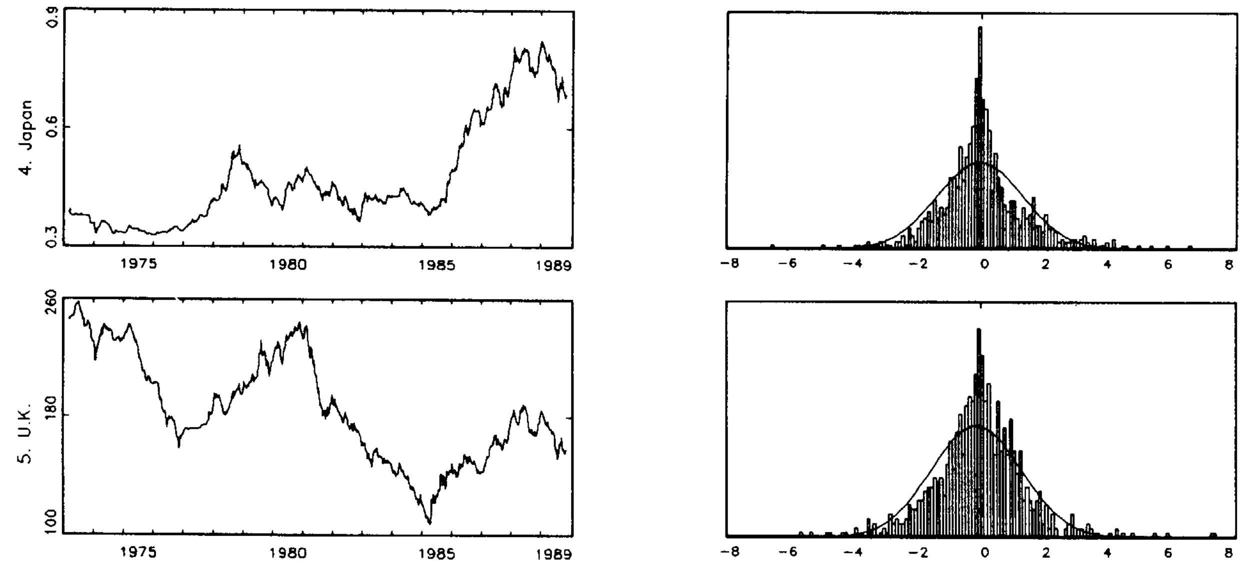


Fig. 1. Basic data – Time series of levels and histograms of log differences

In panel A, the vertical axis is measured in cents per unit of foreign currency. In panel B, the vertical axis of the histograms is the relative frequency of the data falling in every 0.1 interval from − 8 to 8; the corresponding densities were computed in the same way for each country using the normal density with the sample mean (row 1 in Table 1) and the sample variance (row 2) for each country.

R = endpoint of first sample used in estimation of regression parameters; (1d)

T = endpoint of last sample used in estimation. (1e)

Note our dating convention: what we denote h_t corresponds to what is often called h_{t+1} or σ_{t+1} (e.g., Engle, 1982). For concreteness in interpreting (1b) and (1c), it may help to note that in the tables below we report results for $j = 1, 12, 24$, corresponding to approximately weekly, quarterly, and semiannual horizons. To do so for a given horizon, we obtain for each model $T - R + 1$ fitted values $\hat{h}_{mt,j}$, $t = R, \ldots, T$, for models $m = 1, \ldots, M$, where the number of models M in the tables below is six. We then compute the root mean squared prediction error ($RMSPE$) for model m at horizon j as

$$\left[(T - R + 1)^{-1} \sum_{t=R}^{T} (e_{t+j}^2 - \hat{h}_{mt,j})^2 \right]^{1/2}. \tag{2}$$

We focus on $RMSPE$ because mathematical expectations have minimum $RMSPE$, so a good statistical model for the expected value of exchange rate squares will tend to have forecast errors whose average squared value is small.[3]

Column 1 of Table 2 lists the models we estimated, column 3 the acronyms used in some subsequent tables.[4] Column 2 gives the formula for the one-period-ahead conditional variance, except for the nonparametric estimator for which the formula for the arbitrary j-period-ahead forecast is given. Since all the other models are linear, multi-period forecasts can be obtained by the usual recursive prediction formulas. Consistent with the assumption that exchange rate changes have zero conditional mean, in such forecasts the changes were assumed to be conditionally uncorrelated at all nonzero lags (i.e., $E_{t-1} e_t e_{t+j} = 0$ for all $j > 0$).

The homoskedastic model (line 1) simply sets the conditional variance at all horizons equal to the sample mean of lagged e_t^2's.

Two GARCH models were used (lines 2 and 3). Both were estimated by maximum likelihood assuming conditional normality, using analytical derivatives, with presample values of h and e^2 set to sample means. Lee and Hansen (1991) and Lumsdaine (1989) show that the conditional normality assumption is not necessary for the consistency and asymptotic normality of the estimators.[5]

[3] In related work (West, Edison, and Cho, 1993) we consider an alternative measure of model quality, which also tends to favor GARCH.

[4] We also used these models in another paper (West, Edison, and Cho, 1993), and some of the prose in the remainder of this subsection also appears in that paper.

[5] For efficiency reasons one might nonetheless prefer to assume, say, a conditional t-distribution, if the conditional density is in fact t. Our reading of the in-sample evidence is that this is not essential; e.g., Baillie and Bollerslev (1989) found little support for the use of a t in weekly exchange rate data.

Table 2
Models, formula for the one-period-ahead conditional variance, and acronyms used in later tables

Column 1 Model	*Column 2* Formula for h_t	*Column 3* Acronym
Homoskedastic model		
1. Homoskedastic	$h_t = \omega$	homo
GARCH models		
2. GARCH (1,1)	$h_t = \omega + \alpha e_t^2 + \beta h_{t-1}$	(1,1)
3. IGARCH (1,1)	$h_t = \alpha e_t^2 + (1-\alpha) h_{t-1}$	ig
Autoregressive models		
4. AR (12) in e_t^2	$h_t = \omega + \sum_{i=1}^{12} \alpha_i e_{t-i+1}^2$	e2AR
5. AR (12) in $\|e_t\|$	$h_t = (\pi/2)(\mathrm{E}_t\|e_{t+1}\|)^2$, $\mathrm{E}_t\|e_{t+1}\| = \omega + \sum_{i=1}^{12} \alpha_i \|e_{t-i+1}\|$	|e|AR
Nonparametric model		
6. Gaussian kernel	$h_{t,j} = \mathrm{E}(e_{t+j}^2 \| e_t)$; $\hat{h}_{t,j} = \sum_{t=1}^{N-j} w_{tN,j} e_{t-j}^2$, $w_{tN,j} = c_{tN,j} / \sum_{s=1}^{N-j} c_{sN,j}$, $c_{tN,j} = \exp[-0.5(e_N - e_t)^2/b^2]$, b = bandwidth defined in text	nonp

We chose GARCH(1,1) and IGARCH from a larger set of possible GARCH models after (1) analysis of some in-sample diagnostics seemed to suggest GARCH(1,1) for Canada, Germany, and the U.K., and IGARCH for France and Japan, and (2) a little experimentation with ARCH(1), ARCH(2), GARCH(1,2), and GARCH(2,1) models suggested that *MSPE*s from these models are comparable or worse than the two we chose to study.

We also studied two autoregressive models, both of which were estimated by OLS. One autoregression used e_t^2 (line 4). It is included because GARCH models imply ARMA processes for e_t^2 (see Bollerslev, 1986); OLS estimation of such autoregressions therefore might perform comparably to more complicated GARCH estimation (although under the GARCH null, such OLS estimation is asymptotically inefficient). As in Schwert (1989a, b), whose work is based on that of Davidian and Carroll (1987), the other autoregression used $|e_t|$ (line 4). Schwert suggests the factor of $(\pi/2)$ because the variance of a zero mean normally distributed random variable is $(\pi/2)$ times the square of the expected value of its absolute value. For both autoregressions, the lag length of 12 was chosen because for all countries in-sample results indicated that such a lag length was more than sufficient to produce a Q-statistic that implied white noise residuals.

Finally, we also tried a nonparametric estimator (line 6). It can be interpreted as working off the basic definition

$$\mathrm{E}(e_{t+j}^2 \mid e_t) = \int_0^\infty e_{t+j}^2 f(e_{t+j}^2 \mid e_t)\, \mathrm{d}e_{t+j}^2,$$

where $f(e_{t+j}^2 \mid e_t)$ is the density of e_{t+j}^2 conditional on e_t. See Pagan and Ullah (1990a, b) for an excellent exposition. As in Pagan and Schwert (1990a) we used a Gaussian kernel, defined in column 2, with the bandwidth $b = \hat{\sigma}(R-j)^{-1/5}$, $\hat{\sigma}$ the sample standard deviation of e_t, $t = 1, \ldots, R-j$, $j = 1, 12, 24$. We did not try any other kernel. We did a little experimentation with some alternative fixed bandwidths and information sets, comparing *MSPE*s, but found that these yielded similar results.

There remain two questions before we can begin our model evaluation. The first is where to begin the out of sample exercise. We arbitrarily began our forecasts at the midpoint of the sample, and the first sample for which we fit any models included the 432 observations from March 14, 1973 to June 17, 1981. Because the final 24 weeks of the sample (April 12, 1989 to September 20, 1989) were used only for forecast evaluation, the last observation of our final estimation sample was April 5, 1989. [In the notation of (1d), $R = 432$ and $T = 839$.] For our one-week horizon, the predictions and realizations of e_t^2 spanned the 408-weeks from June 24, 1981 to April 12, 1989; the comparable 408-week period for the 12- and 24-week horizons may be obtained by shifting the one-week dates forward by 11 and 23 weeks, respectively.

The other question concerned what sample should be used for estimation as additional observations were added beyond the June 17, 1981 date at which our first sample ended. In our initial work, we estimated each of our models on both (1) rolling samples, in which the sample size used for estimation was fixed at 432 and what had been the initial observation as each additional observation was added, and (2) expanding samples, in which the sample size grew as additional observations were added. *RMSPE*s were quite similar for rolling and expanding samples, with those for rolling samples perhaps showing a slight tendency to be smaller (rolling *RMSPE*s were smaller in 63 of the 90 experiments [90 = 5 countries times 6 models times 3 horizons]). To keep the project manageable, we therefore decided to subject only the rolling estimators to detailed analysis.

2.3. Procedures for asymptotic inference

Most of our inference is based on asymptotic approximations described below. In addition to the usual reasons to be concerned about the finite sample accuracy of such approximations, there are grounds to be concerned about the applicability of regularity conditions typically underlying such approximations: exchange rate data may lack suitable higher-order moments (e.g., Loretan and Phillips, 1992); one of our models uses a nonparametric estimator; more generally, the previous paragraph's observation that forecast quality did not deteriorate when we used rolling rather than expanding samples suggests that the usual conditions may not hold. Nevertheless, we conduct most of our inference using such theory, for two reasons. First, a small Monte Carlo experiment to double-check one piece of our asymptotic analysis suggested that the asymptotic approximation is unlikely to be very misleading if some minimal conditions do hold, and second, the computational cost of using bootstrap methods throughout is enormous, given the nonlinear search required to estimate GARCH and IGARCH models.

To explain the asymptotic procedures that we used: Let P be the sample size. Under suitable regularity conditions, it is well-known that if g_t is a zero mean, covariance stationary random vector, $P^{-1/2}\sum_{t=1}^{P} g_t \overset{A}{\sim} \mathrm{N}(0,S)$, where $S \equiv \sum_{j=-\infty}^{\infty} \Gamma_j$, $\Gamma_j \equiv \mathrm{E}g_t g'_{t-j}$ (e.g., Hannan, 1973); White (1984) summarizes some parallel results that apply when data satisfy some mixing conditions but possibly are not stationary. Suppose that g_t is a function of an underlying vector of parameters of interest, say, θ, and that $\hat{\theta}$ is estimated by setting $P^{-1}\sum_{t=1}^{P} g_t(\hat{\theta}) = 0$. A straightforward Taylor series argument yields $P^{1/2}(\hat{\theta}-\theta) \overset{A}{\sim} \mathrm{N}(0,V)$, $V \equiv (\mathrm{E}\partial g_t/\partial\theta)^{-1} S (\mathrm{E}\partial g_t/\partial\theta)^{-1\prime}$; see Hansen (1982) for a formal argument in the stationary case, Gallant and White (1988) for the parallel argument, and more complicated formulas, under conditions that allow for the possibility that g_t is not stationary.

In our applications of this result, $\partial g_t/\partial\theta$ does not depend on θ and so $\mathrm{E}\partial g_t/\partial\theta$ is a matrix of known constants. The estimator of S that we used was that suggested by Newey and West (1987):

$$\hat{S} = \hat{\Gamma}_0 + \sum_{j=1}^{k} [1 - j/(k+1)](\hat{\Gamma}_j + \hat{\Gamma}'_j), \tag{3}$$

where $\hat{\Gamma}_j$ is the jth sample autocovariance of $\hat{g}_t$, $\hat{\Gamma}_0 \equiv P^{-1}\sum_{t=j+1}^{P} \hat{g}_t \hat{g}'_{t-j}$. The value of k in (3) was determined by a data-dependent automatic rule that has certain asymptotic optimality properties (Newey and West, 1994): Let n be the integer part of $4(P/100)^{2/9}$, so that $n = 6$ for estimates based on the 863 observations in the whole sample (e.g., Table 1), $n = 5$ for estimates based on 408 observations in the forecasting sample (e.g., Table 4). Also, let w be a vector of ones of the same dimension as g_t, $\hat{\sigma}_j \equiv w'\hat{\Gamma}_j w$, $\hat{s}^{(0)} \equiv \hat{\sigma}_0 + 2\sum_{j=1}^{n} \hat{\sigma}_j$, $\hat{s}^{(1)} \equiv 2\sum_{j=1}^{n} j\hat{\sigma}_j$. Then k was set to the integer part of $1.1447\{\hat{s}^{(1)}/\hat{s}^{(0)}\}^{2/3} \times \{\text{sample size}\}^{1/3}$. The resulting values for k in Table 1, for example, were Canada – 4, France – 1, Germany – 6, Japan – 7, and the U.K. – 11. The values for the remaining tables are available on request.

Some details may be helpful in understanding how we used this framework. In Table 1, lines 1–4, 13, and 14, begin by defining the (4×1) vector $X_t \equiv (e_t, e_t^2, e_t^3, e_t^4)'$. Let $\theta \equiv (\mathrm{E}e_t, \mathrm{E}e_t^2, \mathrm{E}e_t^3, \mathrm{E}e_t^4)'$, $\hat{\theta} \equiv (P^{-1}\sum_{t=1}^{P} e_t, P^{-1}\sum_{t=1}^{P} e_t^2, P^{-1}\sum_{t=1}^{P} e_t^3, P^{-1}\sum_{t=1}^{P} e_t^4)'$, $g_t \equiv X_t - \theta$, and $\hat{g}_t \equiv X_t - \hat{\theta}$. Then $P^{1/2}(\hat{\theta}-\theta) \overset{A}{\sim} \mathrm{N}(0,S)$, $S \equiv \sum_{j=-\infty}^{\infty} \Gamma_j, \Gamma_j \equiv \mathrm{E}g_t g'_{t-j}$. Given the estimate of S, standard errors on the relevant entries in Table 1 can be computed using the delta method. A similar method was used in panel A of Table 6 below.

In the modified Ljung–Box statistic, Table 1, lines 5–7, $\theta \equiv (\mathrm{E}e_t e_{t-1}, \ldots, \mathrm{E}e_t e_{t-r})'$, $r = 10, 50, 90$, $\hat{\theta}$ the corresponding sample moments, $X_t \equiv (e_t e_{t-1}, \ldots, e_t e_{t-r})'$, $g_t \equiv X_t - \theta$, $\hat{g}_t \equiv X_t - \hat{\theta}$. We assume that the conditional first moment of e_t is zero, which implies that g_t is serially uncorrelated, and that $\mathrm{E}e_t^2 e_{t-i} e_{t-j} = 0$ for $i \neq j$, which implies that $\mathrm{E}X_t X_t'$ is diagonal. With a little algebra, this validates the following: For $j = 0, \ldots, r$, let $\hat{\sigma}_j$ be the jth element of $\hat{\theta}$, $\hat{\sigma}_j \equiv P^{-1}\sum_{t=j+1}^{P} e_t e_{t-j}$, and let $\hat{\kappa}_j \equiv P^{-1}\sum_{t=j+1}^{P} e_t^2 e_{t-j}^2$, $\hat{\rho}_j \equiv \hat{\sigma}_j / \hat{\sigma}_0$. Then for any fixed r,

$$P(P+2)\hat{\sigma}_0^2 \sum_{j=1}^{r} (P-j)^{-1}(\hat{\rho}_j^2/\hat{\kappa}_j) \overset{\mathrm{A}}{\sim} \chi^2(r). \tag{4}$$

If the data are conditionally homoskedastic, so that $\mathrm{E}e_t^2 e_{t-j}^2 = \mathrm{E}e_t^2 e_{t-j}^2 \equiv \sigma_0^2$, $\hat{\kappa}_j \xrightarrow{\mathrm{P}} \sigma_0^2$ and this statistic is asymptotically equivalent to the standard Ljung–Box statistic.[6]

In Tables 4 and 5 below, which report inference about forecasts or forecast errors, the conceptual experiment that underlies our asymptotic approximation is one in which both the number of observations used in estimation (R in the notation of (1d)) and the number used for forecasting ($T - R + 1 \equiv P$) go to infinity, with $(T - R + 1)/R$ approaching a finite constant (possibly zero).

Consider, for example, the one-period-ahead *MSPE*. For notational simplicity, assume stationarity (rather than, say, just mixing). Let h_{mt} be model m's population prediction of e_{t+1}^2 at time t (i.e., the prediction it would make if an infinite-sized sample had been used in estimation). Let δ be the vector of the entire set of regression parameters, across all models (the constant for the homoskedastic model, the constant and coefficients on e_t^2 and h_{t-1} for the GARCH(1,1) model, ...). Let $u_{mt+1} = e_{t+1}^2 - h_{mt}$, $\sigma_m^2 = \mathrm{E}u_{mt+1}^2 = \mathrm{E}(e_{t+1}^2 - h_{mt})^2$, $\hat{u}_{mt+1} = e_{t+1}^2 - \hat{h}_{mt}$, $\hat{\sigma}_m^2 = (T-R+1)^{-1}\sum_{t=R}^{T}(e_{t+1}^2 - \hat{h}_{mt})^2$, $\hat{\theta} \equiv (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2)'$, $\theta \equiv (\sigma_1^2, \ldots, \sigma_M^2)'$, $g_{t+1}(\theta - \delta) = (u_{1t+1}^2 - \sigma_1^2, \ldots, u_{Mt+1}^2 - \sigma_M^2)'$, $g_{t+1}(\hat{\theta}, \hat{\delta}) = (\hat{u}_{1t+1}^2 - \hat{\sigma}_1^2, \ldots, \hat{u}_{Mt+1}^2 - \hat{\sigma}_M^2)'$. It may be shown that under suitable conditions, sampling error in $\hat{\delta}$ is irrelevant for asymptotic inference on θ (West, 1993), and we apply the logic above to $(T-R+1)^{-1} \sum_{t=R}^{T} g_{t+1}(\hat{\theta}, \hat{\delta}) \equiv P^{-1}\sum_{t=R}^{T}\hat{g}_{t+1}$. The implication is that $P^{1/2}(\hat{\theta} - \theta) \overset{\mathrm{A}}{\sim} \mathrm{N}(0, S)$, where the (i,q) element of the $M \times M$ matrix S is $\sum_{j=-\infty}^{\infty} \mathrm{E}(u_{1t}^2 - \sigma_i^2)(u_{qt-j}^2 - \sigma_q^2)$. A test statistic for the equality of the *MSPE*s across all M models is constructed as follows. Let B be the $(M-1) \times M$ matrix whose first column is $(-1, -1, \ldots, -1)'$ and whose $(M-1)$ other columns contain the identity

[6] Diebold and Mariano (1993) have independently suggested conducting inference on forecast errors using similar techniques, and Diebold (1988) suggested our modification of the Ljung–Box statistic in the specific case of a GARCH data-generating process.

matrix; the null is that $B\theta = 0$. Then for $\hat{S}$ constructed as in (3),

$$(T - R + 1)[\hat{\theta}' B'(B\hat{S}B')^{-1} B\hat{\theta}] \equiv P[\hat{\theta}' B'(B\hat{S}B')^{-1} B\hat{\theta}] \overset{A}{\sim} \chi^2(M - 1), \quad (5)$$

$$\hat{S} \equiv \hat{\Gamma}_0 + \sum_{j=1}^{k} [1 - j/(k+1)](\hat{\Gamma}_j + \hat{\Gamma}'_j), \qquad \hat{\Gamma}_j \equiv P^{-1} \sum_{t=R+j}^{T} \hat{g}_t \hat{g}'_{t+j}.$$

Note that since we select k as described and do not constrain k to be zero, we allow the forecast errors to be serially correlated. Similar formulas apply for the 12- and 24-period-ahead predictions, with, e.g., $u_{m,t+12} \equiv e^2_{t+12} - h_{m,t+12}$ and $\sigma^2_{m,12} = \mathrm{E}(e^2_{t+12} - h_{m,t+12})^2$.

3. Basic empirical results

To frame our discussion, Table 3a presents estimates of the GARCH(1,1) model for the first of our rolling samples. The Appendix available on request has parallel estimates for the other models;we present GARCH(1,1) here because of its simplicity and because, as we shall see, it worked relatively well in forecasting. For the benefit of those familiar with GARCH, we briefly note that the estimates suggest, as usual, considerable persistence, since $\alpha + \beta$ is estimated to be above 0.80 in all five countries, above 0.90 in France, Germany, and Japan; the null that $\alpha + \beta = 1$ could not be rejected at the 5% percent level for France and Japan (not reported in the table).

What is of particular interest to us is how such parameters translate into for *RMSPE*s at various horizons. Suppose e_t^2 is stationary ($\alpha + \beta < 1$, under a GARCH(1,1) parameterization). With the exception of IGARCH, all our estimators will then yield essentially the same predictions in population for a sufficiently long horizon, since all will predict that the e_t^2 will be near its unconditional mean. Accordingly, the *RMSPE*s will also be essentially the same. We use the GARCH(1,1) estimates in Table 3a to get an idea of how long a horizon is needed for this to occur.

Table 3b reports the ratio of the population *RMSPE*s of a homoskedastic model to that of a GARCH(1,1) model, for each of our three horizons and for each of the five sets of estimates of α and β reported in Table 3a.[7] According to columns 1 and 5 in Table 3b, the Table 3a estimates for Canada and the U.K.

[7] These population figures ignore the effects of sampling error in the estimation of model parameters. Reinsel (1980) and Ericcson and Marquez (1989), among others, have suggested a refinement to the computation of the *RMSPE* that accounts for sampling error in such estimation. But inspection of their formulae and simulation results indicates that the refinement has a noticeable effect only when the following ratio is much larger than in our application: (number of regressors)/(sample size). While neither of these papers considers data that are conditionally heteroskedastic, we take the message to be that such a refinement is unlikely to much affect the Table 3b figures.

Table 3a
GARCH (1.1) estimates, sample period: 3/14/73 to 6/17/81

	$\omega(\times 10^5)$	α	β
1. Canada	0.5	0.26	0.54
	(0.1)	(0.02)	(0.06)
2. France	1.3	0.35	0.61
	(0.2)	(0.05)	(0.04)
3. Germany	2.0	0.30	0.61
	(0.4)	(0.05)	(0.04)
4. Japan	0.07	0.05	0.94
	(0.03)	(0.01)	(0.01)
5. U.K.	1.9	0.11	0.73
	(0.7)	(0.05)	(0.10)

Table 3b
Population root mean square prediction errors, homoskedastic relative to GARCH (1, 1)

Horizon	*Column 1* $\alpha = 0.26$ $\beta = 0.54$	*Column 2* $\alpha = 0.35$ $\beta = 0.61$	*Column 3* $\alpha = 0.30$ $\beta = 0.61$	*Column 4* $\alpha = 0.05$ $\beta = 0.94$	*Column 5* $\alpha = 0.11$ $\beta = 0.73$
1	1.09	1.60	1.23	1.06	1.02
12	1.00	1.15	1.02	1.05	1.00
24	1.00	1.05	1.00	1.04	1.00

1) The numbers in parentheses in panel A are asymptotic standard errors.

2) Panel B presents the ratio of *RMSPE*s for the indicated horizons, computed assuming that the data are driven by a GARCH (1, 1) model with the indicated parameters, and abstracting from sampling error in estimation of the model parameters. The ratio is invariant to ω. The *RMSPE* for the homoskedastic model is constant for all horizons. The ratio asymptotes to 1 as the horizon approaches infinity, for each pair of α and β.

suggest sufficiently rapid mean reversion that our proposed comparisons of 12- and 24-week horizons are probably not of interest. On the other hand, columns 2 to 4 indicate that other Table 3a estimates imply as sharp a difference in *RMSPE*s at one or both of these longer horizons as occurs at a one-period horizon for the U.K. parameters in column 5.

We will *not* attempt to squeeze an interpretation of the results of our out-of-sample comparison into the Tables 3 figures. Even under a GARCH(1,1) null the Tables 3 figures will be misleading insofar as sampling error has affected the point estimates of α and β. Rather, we interpret Tables 3 as presenting in-sample evidence that it may be possible to distinguish different estimators at horizons of as long as 24 weeks.

Table 4 presents our attempts to do so, for forecasts of 1 and 12 weeks as well as 24 weeks ahead, in panels A, B, and C, respectively. In each panel, under each country are two columns. The second, labelled '*RMSPE*' gives the root mean squared prediction error, computed according to Eq. (2). The other column, labelled 'Rank', indicates the relative size of that model's *RMSPE*, 1 indicating the best (smallest) *RMSPE* and 6 the worst (largest). The rows labelled 'H_A', 'H_B', and 'H_C' (at the bottom of each panel) will be discussed below.

We begin with two general comments, before beginning a comparison of the models. First, as one would expect, given the noisiness of exchange rate data, these out-of-sample *RMSE*s generally are larger than the in-sample *RMSE*s reported in line 14 of Table 1. That is, the out-of-sample predictions using the estimated conditional variances are usually less accurate than an in-sample prediction using the in-sample unconditional variance. Second, and somewhat surprisingly, there does not appear to be a tendency for *RMSPE*s to increase at longer horizons; the median Table 4 values for the 1-, 12-, and 24-week horizons are 4.746, 4.791, and 4.503, for example. In the context of GARCH(1,1) models, the implication is that mean reversion occurs as rapidly as in, say, column 5 of Table 3b. The figures in columns 2 to 4 of that table suggest otherwise, so there is a clear conflict between the out-of-sample and in-sample evidence.

Turn now to comparing the models. At the one week horizon, panel A of Table 4 indicates that one of the two GARCH models had the smallest *RMSPE* for all five countries. The IGARCH model was probably the most consistent performer overall, being best in three countries (France, Germany, and U.K.), second and third best in the other two (Japan and Canada). At the 12-week horizon (panel B), the best model was either the homoskedastic (Canada, France, and Germany) or autoregression in absolute values (Japan and U.K.). At the 24-week horizon (panel C), depending on the country, one of four different models had the lowest *RMSPE*, GARCH(1,1) being the only model that was best in two countries (Germany and U.K.). But the most consistent performer at 24 weeks was probably the autoregression in exchange rate squares, which was second in four countries and first in one (Japan).

Which model performs best, then, varies from country to country and horizon to horizon; if there is an underlying pattern, it is difficult for us to discern, and, at least superficially, Table 1 suggests that it might be largely a matter of chance which model produces the smallest *RMSPE*.[8] That performance is quite similar across models is also suggested by casual inspection of the point estimates of the *RMSPE*s; even at a one-period horizon, in only one case is the worst model's

[8] Consistent with this statement, and with the literature surveyed in Clemen (1989), a prediction formed by averaging the six forecasts typically performs better than any of the individual forecasts, at least at the longer horizons. Of the 15 comparisons, the ranking of the average forecast was: 1 – 7 times (two of the seven occur for a one-period horizon; IGARCH performs roughly comparably here); 2 – 6 times; 3 – once; 4 – once. Details are in the Appendix that is available on request.

Table 4
Root mean squared prediction errors

	Canada		France		Germany		Japan		U.K.	
	Rank	*RMSPE*	Rank	*RMSPE*	Rank	*RMSPE*	Rank	*RMSPE*	Rank	*RMSPE*
Panel A: One-week horizon										
homo	5	0.714	2	5.167	2	4.704	3	4.380	4	5.745
(1, 1)	1	0.702	6	5.351	5	4.783	1	4.323	2	5.632
ig	3	0.706	1	5.161	1	4.695	2	4.343	1	5.563
e2AR	4	0.712	5	5.273	6	4.925	5	4.411	5	6.033
|e|AR	2	0.704	3	5.200	4	4.767	4	4.388	3	5.726
nonp	6	0.737	4	5.201	3	4.724	6	4.442	6	6.537
H_A	9.70	[0.084]	8.91	[0.113]	8.23	[0.144]	6.42	[0.268]	3.76	[0.584]
H_B	1.24	[0.265]	0.01	[0.918]	0.01	[0.912]	0.77	[0.380]	1.52	[0.217]
H_C	4.24	[0.237]	8.99	[0.029]	4.13	[0.247]	2.86	[0.413]	3.59	[0.310]
Panel B: Twelve-week horizon										
homo	1	0.695	1	5.219	1	4.754	3	4.435	5	5.794
(1, 1)	2	0.697	6	5.696	6	4.831	5	4.451	4	5.756
ig	6	0.731	5	5.268	5	4.817	6	4.454	2	5.692
e2AR	3	0.700	3	5.251	4	4.796	2	4.433	3	5.726
|e|AR	4	0.701	4	5.267	3	4.785	1	4.430	1	5.674
nonp	5	0.704	2	5.250	2	4.762	4	4.447	6	5.841
H_A	16.71	[0.005]	15.06	[0.010]	7.32	[0.198]	1.08	[0.956]	8.49	[0.131]
H_B	n.a.		n.a.		n.a.		0.01	[0.921]	1.29	[0.256]
H_C	5.96	[0.114]	13.60	[0.004]	5.83	[0.120]	0.39	[0.943]	5.66	[0.129]
Panel C: Twenty-four-week horizon										
homo	1	0.695	3	5.094	3	4.500	2	4.424	4	5.770
(1, 1)	5	0.703	6	5.694	1	4.490	6	4.498	1	5.708
ig	6	0.743	1	5.060	5	4.509	5	4.483	5	5.834
e2AR	2	0.695	2	5.087	2	4.498	1	4.422	2	5.721
|e|AR	3	0.697	4	5.109	4	4.505	4	4.441	3	5.729
nonp	4	0.702	5	5.131	6	4.535	3	4.436	6	5.943
H_A	18.95	[0.002]	18.08	[0.003]	3.80	[0.578]	6.05	[0.301]	5.82	[0.324]
H_B	n.a.		0.25	[0.619]	0.11	[0.741]	0.12	[0.728]	0.07	[0.789]
H_C	3.35	[0.340]	17.64	[0.001]	1.07	[0.785]	5.82	[0.121]	1.95	[0.583]

1) The '*RMSPE*' columns present the out of sample root mean squared error in predicting e_{t+j}^2 for horizon j ($j = 1, 12, 24$) and the indicated country and model. The 'Rank' columns index the relative size of the *RMSPE*s for a given country and horizon, 1 indicating the smallest *RMSPE* and 6 the largest.

2) The H_A, H_B, and H_C rows present χ^2 statistics (asymptotic p-values in brackets) for the following hypotheses: A – equality of *MSPE*s of all six models ($\chi^2(5)$); B – equality of *MSPE*s of best and homo models ($\chi^2(1)$); C – equality of *MSPE*s from homo, (1, 1), e2AR, and |e|AR models ($\chi^2(3)$). The statistics are computed as in Eq. (5).

RMSPE more than 5% larger than the best model's (U.K.); once again, such point estimates are surprising in light of Table 3b.

For an additional measure of similarity of *RMSPE*s, we turn to formal statistical testing of the hypothesis that these are the same across various models, for a given horizon. In Table 4, the 'H_A', 'H_B', and 'H_C' rows at the bottom of each panel give statistics and, in brackets, p-values assuming an asymptotic chi-squared distribution, for the following three hypotheses:

H_A: *MSPE*s for all six models are equal ($\chi^2(5)$).

H_B: *MSPE*s for the best model and the homoskedastic model are equal ($\chi^2(1)$). (6)

H_C: *MSPE*s for the homoskedastic, GARCH(1,1), and two autoregressive models are the same ($\chi^2(3)$).

Hypothesis A is an obvious one.[9] Tests of hypothesis B were performed because the homoskedastic model is the simplest one, and therefore probably the model of most appeal if, in fact, performance is similar across models. Tests of hypothesis C were performed because the formal asymptotic theory that underlies the test makes assumptions that rule out our nonparametric estimator and possibly the IGARCH estimator as well.

Table 4 indicates that the H_A test of the null of equal *RMSPE*s across all models is rejected at the 0.05 level in four of our fifteen experiments (Canada and France, 12- and 24-week horizons) and once at the 0.10 but not 0.05 level (Canada, 1-week horizon). This suggests that the seeming similarity of point estimates of *RMSPE*s might be misleading, at least for Canada and France. In no case, however, can one reject at conventional significance levels the null that the homoskedastic model's *RMSPE* is the same as that of the best model: the lowest of p-value for H_B is 0.217 (U.K., 1-week horizon). The H_C test of equal *RMSPE*s for the homoskedastic, GARCH(1,1), and two AR models rejects at the 0.05 level for France for all three horizons, again suggesting that the seeming similarity of point estimates of *RMSPE*s might be misleading for France.

These asymptotic tests may well be deceptive in finite samples, even if the asymptotic theory eventually yields a good approximation. One indication that this may be the case is that of the four rejections at the 0.05 level of equality of all six models, three occur in experiments in which the homoskedastic model is the best (Canada, 12- and 24-week; France, 24-week). If, indeed, a homoskedastic model were generating the data, at least four of the other five models would

[9] For computational convenience, we computed tests for equality of the *MSPE*s rather than the asymptotically equivalent tests for the *RMSPE*s; for expositional convenience, in all discussion apart from the statement of the tests in the preceding paragraph in the text, we refer to these as tests on the *RMSPE*s.

produce exactly the homoskedastic forecast in an infinitely large sample (the possible exception is IGARCH, whose asymptotic behavior under these conditions is unclear to us). But this suggests a tendency to reject too much, not too little, a result that we have found in related Monte Carlo studies using data generated by GARCH processes (Newey and West, 1994).

But to double-check the possibility that our asymptotic tests are instead rejecting too infrequently, we undertook two exercises. First, we examined a seventh model, which set $\hat{h}_t = e_t^2$ – the conditional variance in week t is equal to the realized square of the exchange rate. (Reminder to readers familiar with the GARCH literature: what we call h_t here is usually called h_{t-1}.) To our knowledge, this has not been seriously proposed as a model for exchange rate volatility, for the good reason that it is not an appealing one: the *RMSPE*s for the one-week horizon, for example, are: Canada – 0.933; France – 7.119; Germany – 6.574; Japan – 5.593; U.K. – 7.466. These are a good 25% above the panel A figures of Table 4 for the other models. We use it here to see if our asymptotic tests have enough power to recognize the substantial difference between this model and the others. And they do, as is indicated by the following summary of test results. Of fifteen $\chi^2(6)$ tests of the equality of *RMSPE*s across all seven models, eleven reject at the 0.05 level, thirteen at the 0.10 level. Of fifteen $\chi^2(1)$ tests of the equality of the *RMSPE* from this additional model and that of the worst of the six models reported in Table 4, thirteen reject at the 0.05 level, fourteen at the 0.10 level (the exception was U.K., one-week horizon, which rejects at the 0.15 level). It seems, then, that whatever the problems with our asymptotic tests, these tests do have enough power to reject an egregiously poor model at conventional significance levels.

The second exercise we undertook to check the validity of our asymptotic tests was a small Monte Carlo experiment. Because of space constraints, we limit ourselves to the succinct statement that the experiment suggested that if our asymptotic procedures have a small sample bias, that bias is towards rejecting too much, not too little. A detailed discussion of the experiment is available in the Appendix that is available on request.

4. Additional empirical results

It seems that to a first approximation all our models are equally good as predictors of exchange rates squares. To compare them from a slightly different perspective, we conducted a standard efficiency test (e.g., Pagan and Schwert, 1990a), estimating by OLS the regression

$$e_{t+1}^2 = b_0 + b_1 \hat{h}_{mt} + \varepsilon_{t+1}. \tag{7}$$

If, indeed, $\mathrm{E}_t e_{t+1}^2 = \hat{h}_{mt}$, one should get $b_0 = 0$, $b_1 = 1$. One should also find that ε_{t+1} is serially uncorrelated. But a quick look at the autocorrelations of the residuals suggested that this was rarely, if ever, the case. So we do not formally test for the absence of serial correlation, and instead correct the variance–covariance matrix of the estimated parameter vector for conditional heteroskedasticity as well as serial correlation, using the techniques described above.

Results are in Table 5. Asymptotic standard errors for b_0 and b_1 are given in parentheses beneath the point estimates. For all five countries, the $\chi^2(2)$ column gives the point estimate and asymptotic p-value for H_0: $b_0 = 0$, $b_1 = 1$. The ** and * after the estimates of $\hat{b}_1$ indicate significant differences from zero, not one.

We note first that the rankings by R^2 are quite similar to those by *RMSPE*. This indicates that models with relatively low *RMSPE*s also have *RMSPE*s whose variance component is relatively low, since R^2 reflects the variance but not bias-squared component of *MSPE*. Some new information is yielded by the other estimates. Of the thirty $\chi^2(2)$ tests of H_0: $b_0 = 0$, $b_1 = 1$, twenty-seven reject at the 0.10 level (the exceptions are GARCH(1,1) for Japan, IGARCH for Japan and U.K.), twenty-five at the 0.05 level (the additional exceptions are the two autoregressions for Canada). The standard errors on b_0 and b_1 yield compatible conclusions.

Perhaps unsurprisingly, then, none of the models pass this efficiency test: the Monte Carlo simulation indicates that this test has good power, being very likely to reject the null (not reported in the table). More encouraging is that seven of the estimates of b_1 are significantly different from zero at the 0.05 level, five of these being for GARCH models (see the ** entries). This shows that there is some predictive power in the estimated conditional variances. For future reference, note the marked tendency of the models to have predictive power for Canadian data.

We also performed the efficiency test in Table 5 for the 12- and 24-week horizons. For these horizons, the results did not help discriminate between models, and we therefore limit ourselves to a summary of the results. Of the sixty $\chi^2(2)$ tests, fifty-eight reject at the 0.10 level (the exceptions are GARCH(1,1) for Canada 12-week and U.K. 24-week), fifty-six at the 0.05 level (the additional exceptions are homoskedastic for Canada 12- and 24-week). More troubling is that while $\hat{b}_1$ was different from zero at the 0.05 level seven times, only two of those estimates were positive (GARCH(1,1) and IGARCH for U.K., 12-week).

Overall, then, it seems that at the one-period horizon there is some evidence favoring GARCH models: while Table 4 cannot reject the null that the *RMSPE*s are the same for all models, GARCH models do tend to produce lower *RMSPE*s, and Table 5 suggests that they have markedly more predictive power for next period's e_{t+1}^2. On the other hand, at longer horizons, we find little grounds for preferring one model over another.

This is a disappointing, and surprising, result. It seems that mean reversion in the conditional variance occurs rapidly enough that no model dominates the others at 12-week or longer horizons. This suggests that the in-sample fits

Table 5
Regression tests of efficiency, one week horizon

	b_0	b_1	R^2	$\chi^2(2)$	b_0	b_1	R^2	$\chi^2(2)$
	Canada				**France**			
homo	1.34	−2.99*	0.018	8.07	3.92	−0.71	0.004	12.15
	(0.53)	(1.62)		[0.018]	(1.13)	(0.50)		[0.002]
(1, 1)	0.15	0.60**	0.044	6.98	2.26	0.09	0.0009	45.45
	(0.06)	(0.17)		[0.031]	(0.40)	(0.14)		[0.000]
ig	0.17	0.55**	0.037	10.04	2.23	0.10	0.0004	13.53
	(0.05)	(0.17)		[0.007]	(0.67)	(0.24)		[0.001]
e2AR	0.20	0.48*	0.019	4.90	2.86	−0.15	0.001	82.32
	(0.09)	(0.28)		[0.086]	(0.41)	(0.13)		[0.000]
|e|AR	0.16	0.66**	0.031	5.44	2.38	0.05	0.0001	25.05
	(0.09)	(0.29)		[0.066]	(0.48)	(0.22)		[0.000]
nonp	0.28	0.27**	0.012	31.66	2.86	−0.16	0.0008	27.63
	(0.05)	(0.13)		[0.000]	(0.57)	(0.22)		[0.000]
	Germany				**Japan**			
homo	2.65	−0.06	0.00003	14.52	3.40	−0.59	0.001	7.46
	(0.74)	(0.35)		[0.001]	(1.37)	(0.70)		[0.024]
(1, 1)	1.95	0.23	0.005	19.85	0.89	0.60	0.023	1.25
	(0.44)	(0.21)		[0.000]	(0.79)	(0.37)		[0.537]
ig	1.56	0.37	0.008	7.93	1.02	0.60**	0.011	2.66
	(0.58)	(0.27)		[0.019]	(0.63)	(0.28)		[0.264]
e2AR	2.45	0.03	0.0001	92.85	1.56	0.32	0.008	11.09
	(0.28)	(0.11)		[0.000]	(0.47)	(0.22)		[0.004]
|e|AR	2.18	0.15	0.001	19.73	1.43	0.40*	0.012	11.39
	(0.54)	(0.27)		[0.000]	(0.42)	(0.21)		[0.003]
nonp	2.48	0.02	0.00001	17.94	2.51	−0.14	0.0006	12.44
	(0.59)	(0.25)		[0.000]	(0.73)	(0.32)		[0.002]

	U.K.			
homo	3.55 (0.81)	− 0.42 (0.36)	0.002	19.18 [0.000]
(1, 1)	1.25 (0.48)	0.53** (0.22)	0.045	7.20 [0.027]
ig	0.95 (0.61)	0.69** (0.30)	0.042	3.90 [0.142]
e2AR	2.40 (0.44)	0.12 (0.13)	0.003	45.07 [0.000]
|e|AR	1.84 (0.51)	0.36 (0.23)	0.010	13.08 [0.001]
nonp	2.78 (0.32)	− 0.03 (0.04)	0.0003	943.63 [0.000]

1) This reports results of the regression $e_{t+1}^2 = b_0 + b_1 \hat{h}_{mt} + \varepsilon_t$. For b_0 and b_1, heteroskedasticity and autocorrelation consistent standard errors are in parentheses. This $\chi^2(2)$ tests H_0: $b_0 = 0$, $b_1 = 1$, with asymptotic p-value in brackets.

2) For b_1, ** denotes significance at the 5% level, * at the 10% level.

overstate the conditional predictability of exchange rate squares. Lamoureux and Lastrapes (1990) have shown that occasional discrete shifts in the mean level of volatility cause substantial upward bias in estimates of the persistence of volatility. We close this section with some evidence that such shifts may have occurred here, and thus may help account for our inability to sharply distinguish one model from another.

Panel A of Table 6 reports split sample estimates of the standard deviation of e_t. As one can see, the point estimate is markedly higher in the second half of the sample for all countries except perhaps Canada.[10] In addition, line 3 of the table indicates that the null of equality is rejected at the 0.05 level for France, Japan, and the U.K., and at the 0.10 level for Germany.

Given that we began forecasting at the sample midpoint, the choice of the midpoint as a date to test for a shift is natural, but nonetheless still arbitrary. In panel B we report a Pagan and Schwert (1990b) test for the constancy of the unconditional variance of e_t that does not require *a priori* specification of a date. The details of the test are described in the notes to the table. As indicated in the table, the null of constancy is rejected at the 0.05 level for all countries but Canada, for which it is not rejected at even the 0.20 level. See row 1 of panel B.

Rows 2 to 4 of Table 6 report the results of applying this test on three subsamples for each country: the first half of our total sample (March 14, 1973 to June 17, 1981), the middle two-fourths (April 24, 1977 to July 31, 1985), and the last half (June 24, 1981 to September 20, 1989). Of the fifteen tests for constant variances (15 = 3 subsamples times 5 countries), only two tests rejected at the 0.05 level (Japan, beginning and middle subsamples).

Now, if the data were driven by a stationary model that allows time-varying conditional variances, it would not be surprising if tests such as those in Table 6 found evidence of shifts in variance at short but not long horizons. We, however, find the converse. And, as briefly noted in Section 2, forecast quality was no better for expanding than for rolling samples, which also seems to suggest a failure of the stationarity assumption.

In this study, we followed many others (e.g., Engle et al., 1990) and implicitly allowed for a failure of stationarity by using rolling samples. We conjecture that it will be productive to explore models that explicitly allow for seeming or actual

[10] This raises the question of whether our exercise would produce different results if applied to split samples, a question also raised by a referee who noted that in the mid-1980s central banks attempted to drive down the dollar. We computed one-week-ahead *RMSPE*s for samples running from (1) 6/17/81 to 9/18/85 (number of predictions = 223) and (2) 9/18/85 to 4/5/89 (number of predictions = 185); the split date was chosen because the Plaza Accord was announced on 9/22/85. The *RMSPEs* were generally higher in the later sample. But GARCH or IGARCH still fared relatively well: one or the other was best in all five comparisons in the early sample, in three of the five comparisons in the later sample. Details are in the Appendix that is available on request.

Table 6
Subsample statistics on e_t

	Canada	France	Germany	Japan	U.K.
Panel A: Standard deviation					
1. Standard deviation 3/14/73–6/17/81	0.499 (0.049)	1.185 (0.097)	1.309 (0.117)	1.174 (0.102)	1.093 (0.076)
2. Standard deviation 6/24/81–9/20/89	0.600 (0.050)	1.603 (0.153)	1.609 (0.149)	1.526 (0.146)	1.663 (0.159)
3. Row 2 − row 1	0.101 (0.070)	0.418 (0.181)	0.299 (0.190)	0.352 (0.178)	0.570 (0.177)
Panel B: Modified range scale tests for constancy of unconditional variance					
1. 3/14/73–9/20/89 No. of obs. = 863	1.428	2.238**	1.901**	1.874**	1.753**
2. 3/14/73–6/17/81 No. of obs. = 432	1.413	1.659*	1.540	1.892**	1.046
3. 4/24/77–7/31/85 No. of obs. = 432	1.242	1.725*	1.365	1.852**	1.317
4. 6/24/81–9/20/89 No. of obs. = 431	1.260	1.226	1.326	1.561	1.667

1) In panel A, heteroskedasticity and autocorrelation consistent asymptotic standard errors are in parentheses.

2) In panel B, let $x_t = (e_t - \bar{e})^2$, where e is the mean of e_t in the sample in question, and let $\bar{x}$ be the corresponding mean of x_t. Let $\psi(r) \equiv [\sum_{t=1}^{r}(x_t - \bar{x})]/(T\hat{s})^{1/2}$, where $1 \leqslant r \leqslant T$, $T =$ 431,432, 863 is the sample size, and $\hat{s}$ is an estimate of the asymptotic variance of $T^{-1/2}\sum_{t=1}^{T}[(e_t - \mathrm{E}e_t)^2 - \mathrm{E}(e_t - \mathrm{E}e_t)^2]$. The table reports the difference between the maximum and minimum of $\psi(r)$.

3) ** means significant at the 5% level, * at the 10% level, according to Table la in Haubrich and Lo (1989).

movement in the unconditional variance of e_t. The sort of movement that one wants to capture appears to be slow enough that it might not be detectable in samples that are eight years long, but rapid enough that it is marked in samples sixteen years long.

Canadian data were unusual in that Table 4's tests of equality of *RMSPE*s tended to find differences across models, and Table 5's efficiency tests were unusually likely to be able to find predictive power in the estimated conditional variances. Perhaps the distinctive results for Canada are no accident, but instead are linked to the stationary behavior of its exchange rates.

5. Conclusions

The in-sample evidence summarized in Tables 1 and 3 strongly suggests that a homoskedastic model should be dominated by the other models that we studied. This did not turn out to be the case. We speculate that models that allow for seeming or actual drift in unconditional moments may result in superior performance. Possibilities include processes that allow occasional discrete jumps (Jorion, 1988) and models with time-varying parameters (Chou et al., 1990).

References

Akgiray, Vedat, 1989, Conditional heteroskedasticity in time series of stock returns: Evidence and forecasts, Journal of Business 62, 55–80.

Baillie, Richard T. and Tim Bollerslev, 1989, The message in daily exchange rates: A conditional variance tale, Journal of Business and Economic Statistics 7, 297–305.

Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics 31, 307–327.

Bollerslev, Tim, Ray Y. Chou, and Kenneth F. Kroner, 1992, ARCH modelling in finance: A review of the theory and empirical evidence, Journal of Econometrics 52, 5–59.

Chou, Ray, Robert F. Engle, and Alex Kane, 1991, Measuring risk aversion from excess returns on a stock index, NBER working paper no. 3643.

Clemen, Robert T., 1989, Combining forecasts: A review and annotated bibliography, International Journal of Forecasting 5, 559–583.

Davidian, M. and R.J. Carroll, 1987, Variance function estimation, Journal of the American Statistical Association 82, 1079–1091.

Diebold, Francis X., 1988, Empirical modeling of exchange rate dynamics (Springer-Verlag, Berlin).

Diebold, Francis X. and Roberto S. Mariano, 1993, Comparing predictive accuracy, Manuscript (University of Pennsylvania, Philadelphia, PA).

Diebold, Francis X. and James Nason, 1990, Nonparametric exchange rate prediction?, Journal of International Economics 28, 315–332.

Diebold, Francis X. and Marc Nerlove, 1989, The dynamics of exchange rate volatility: A multivariate latent factor ARCH model, Journal of Applied Econometrics 4, 1–21.

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

Engle, Robert F. and Tim Bollerslev, 1986, Modelling the persistence of conditional variances, Econometric Reviews 5, 1–50.

Engle, Robert F., Che-Hsiung Hong, and Alex Kane, 1990, Valuation of variance forecasts with simulated options markets, Manuscript (University of California, San Diego, CA).

Ericcson, Neil R. and Jamie R. Marquez, 1989, Exact and approximate multi-period mean-square forecast errors for dynamic econometric models, International finance discussion paper no. 348 (Federal Reserve Board of Governors, Washington, DC).

Gallant, A. Ronald and Halbert White, 1988, A unified theory of estimation and inference for nonlinear dynamic models (Basil Blackwell, New York, NY).

Hannan, Edward J., 1973, Central limit theorems for time series regressions, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 26, 157–170.

Hansen, Lars Peter, 1982, Asymptotic properties of generalized method of moments estimators, Econometrica 50, 1029–1054.

Haubrich, Joseph G. and Andrew W. Lo, 1989, The sources and nature of long-term memory in the business cycle, NBER working paper no. 2951.

Jorion, Philippe, 1989, On jump processes in the foreign exchange and stock markets, Review of Financial Studies 1, 427–445.

Lamoureux, Christopher G. and William D. Lastrapes, 1990, Persistence in variance, structural change and the GARCH Model, Journal of Business and Economic Statistics, 225–234.

Lee, Sang-Won and Bruce E. Hansen, 1991, Asymptotic properties of the maximum likelihood estimator and test of the stability of parameters of the GARCH and IGARCH models, Manuscript.

Loretan, Mico S. and Peter C.B. Phillips, 1992, Testing the covariance stationarity of heavy-tailed time series: An overview of the theory with applications to several financial datasets, Journal of Empirical Finance, forthcoming.

Lumsdaine, Robin L., 1989, Asymptotic properties of the maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models, Manuscript (Harvard University, Cambridge, MA).

Meese, Richard A. and Kenneth S. Rogoff, 1983, Empirical exchange rate models of the seventies: Do they fit out of sample?, Journal of International Economics 14, 3–24.

Meese, Richard A. and Andrew K. Rose, 1991, An empirical assessment of non-linearities in models of exchange rate determination, Review of Economic Studies 58, 603–618.

Newey, Whitney K. and Kenneth D. West, 1987, A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica 55, 703–708.

Newey, Whitney K. and Kenneth D. West, 1994, Automatic lag selection in covariance matrix estimation, Review of Economic Studies 61, 631–654.

Pagan, Adrian R. and G. William Schwert, 1990a, Alternative models for conditional stock volatility, Journal of Econometrics 45, 267–290.

Pagan, Adrian R. and G. William Schwert, 1990b, Testing for covariance stationarity in stock market data, Economics Letters 33, 165–170.

Pagan, Adrian R. and Aman Ullah, 1990a, Chapter 3: Methods of density estimation, Manuscript (University of Rochester, Rochester, NY).

Pagan, Adrian R. and Aman Ullah, 1990b, Chapter 4: Non-parametric estimation of conditional moments, Manuscript (University of Rochester, Rochester, NY)

Reinsel, Greg, 1980, Asymptotic properties of prediction errors for the multivariate autoregressive model using estimated parameters, Journal of the Royal Statistical Society B, 328–333.

Schwert, G. William, 1989a, Business cycles, financial crises and stock volatility, in: K. Brunner and A.H. Meltzer, eds., IMF policy advice, market volatility, commodity price rules and other essays, Carnegie-Rochester series on public policy, Vol. 31 (North-Holland, Amsterdam) 83–126.

Schwert, G. William, 1989b, Why does stock market volatility change over time?, Journal of Finance 44, 1115–1154.

West, Kenneth D., 1993, Asymptotic inference about predictive ability, Manuscript (University of Wisconsin, Madison, WI).

West, Kenneth D., Hali J. Edison, and Dongchul Cho, 1993, A utility based comparison of some models of exchange rate volatility, Journal of International Economics 35, 23–45.

White, Halbert, 1984, Asymptotic theory for econometricians (Academic Press, New York, NY).