EVALUATING THE PREDICTIVE ACCURACY OF VOLATILITY MODELS

by
Jose A. Lopez

Federal Reserve Bank of New York
Research Paper No. 9524

November 1995

Evaluating the Predictive Accuracy of Volatility Models

Jose A. Lopez

Research and Market Analysis Group
Federal Reserve Bank of New York
33 Liberty Street
New York, NY 10045
(212) 720-6633

Draft Date: October 25, 1995

Abstract: The volatility forecast evaluations most meaningful to forecast users are those conducted under economically relevant loss functions. Although several such loss functions are proposed in the literature, their implied economic costs are of interest only to specific types of volatility forecast users. A forecast evaluation framework that incorporates a more general class of economic loss functions is proposed. A user's loss function specifies the three key elements of the evaluation framework: the economic events to be forecast, the criterion with which to evaluate these forecasts, and the subsets of the forecasts of particular interest. Volatility forecasts are transformed into probability forecasts of the specified events, and the probability forecasts are evaluated using statistical criteria, such as probability scoring rules, tailored to the user's interests. An empirical example using exchange rates illustrates the procedure and confirms that the choice of loss function directly affects the forecast evaluation results.

Keywords: volatility, ARCH, probability forecasts, exchange rates
JEL Classification: C32, C52, C53, G12

Acknowledgements: The views expressed here are those of the author and not those of the Federal Reserve Bank of New York or the Federal Reserve System. I thank Frank Diebold for his comments on earlier drafts of this paper as well as for his overall guidance. I also thank Roberto Mariano, Lee Ohanian, Tim Bollerslev, Peter Christoffersen, Esther Ruiz and Ken West for their comments as well as Jeremy Berkowitz, Valentina Corradi, Pedro de Lima, Ken Singleton and seminar participants at the Derivatives Project Meeting of the National Bureau of Economic Research, the Econometric Society Seventh World Congress, the University of Pennsylvania, Arizona State University, Fordham University, the Federal Reserve Bank of Boston, the Federal Reserve Bank of Kansas City, the Federal Reserve Board of Governors and the Federal Reserve Bank of New York.

"Volatility forecasting is a little like predicting whether it will rain; you can be correct in predicting the probability of rain, but still have no rain."
- Engle (1993)

I. Introduction

Although dynamics in the variance of financial time series were observed at least as early as Mandelbrot (1963), efforts to empirically model these dynamics have only developed in the last fifteen years. GARCH models, pioneered by Engle (1982) and Bollerslev (1986), are the volatility models most commonly used in the economics literature, although numerous alternatives exist. Specifically, stochastic volatility models, which arose from the theoretical finance literature, are increasingly employed in empirical research.[1]

Volatility models and their forecasts are of interest to many types of economic agents. For example, options traders require asset volatilities to price options, and central banks or international investors forecasting exchange rates may require interval forecasts, which are readily derived from volatility forecasts. Given the vast number of models available, such agents must decide which forecasts to use as well as the evaluation criterion upon which to base that decision.

Forecast evaluation is typically conducted by minimizing a loss function, and mean squared error (MSE) is the one most commonly used.[2] Yet, the quadratic loss function implied by MSE is inappropriate for evaluating volatility forecasts because it penalizes positive and negative forecasts symmetrically. Furthermore, MSE is a purely statistical loss function with little economic meaning. As suggested by Bollerslev *et al.* (1994), economic loss functions that explicitly incorporate the costs faced by volatility forecast users provide the most meaningful forecast evaluations.[3] West *et al.* (1993) and Engle *et al.* (1993) propose such loss functions based on specific economic questions. However, these questions are arguably too

---

[1] For recent surveys of GARCH models, see Bollerslev *et al.* (1994) as well as Diebold and Lopez (1995a). Andersen (1994) presents a rigorous categorization of volatility models.

[2] Numerous studies have used MSE to evaluate the performance of volatility models; for example, see Taylor (1986, 1987), Friedman and Kuttner (1988), Pagan and Schwert (1990), Akgiray (1989), Kroner *et al.* (1993), West and Cho (1994), Donaldson *et al.* (1994), Lee (1991, 1994) and Bollerslev and Ghysels (1994).

[3] For further discussion of this issue with respect to exchange rates, see Stockman (1987).

specific; the loss functions are only relevant to small subsets of forecast users. For example, it is unclear whether central banks can use profit-based forecast evaluations to minimize their loss functions.

The forecast evaluation framework proposed in this paper incorporates a more general class of economic loss functions. Such loss functions, implicitly or explicitly based on volatility forecasts, are translated into statistical loss functions and hypothesis tests based on probability forecasts. Forecast evaluations are thus tailored to key elements of the relevant economic loss functions. The first such element is the economic events of interest to the user. Volatility forecasts are transformed into probability forecasts by integrating over a model's assumed distribution for the innovation terms, and the ranges of integration are chosen to correspond with the economic events specified.[4] In this way, economic loss functions are directly incorporated into the probability forecasts.

Once generated, the probability forecasts are evaluated using probability scoring rules and calibration tests; i.e., statistical criteria that further incorporate elements of the relevant loss functions into the forecast evaluations. Probability scoring rules measure the accuracy of probability forecasts with respect to whether or not the forecasted events occur. Forecast users can select the scoring rules best suited to their particular loss functions. The most common scoring rule (and the one used in this paper) is the quadratic probability score (QPS), the analog of MSE for probability forecasts. Calibration tests, as developed by Seillier-Moiseiwitsch and Dawid (1993), examine the degree of equivalence between an event's predicted and observed frequencies of occurrence within specified subsets of the forecasts. Thus, forecast users can evaluate subsets of the transformed volatility forecasts that are of particular interest under their loss functions. The tests of comparative predictive accuracy under a general loss function proposed in Diebold and Mariano (1995) are also used to evaluate the forecasts.

The paper is structured as follows. In Section II, the loss functions, both statistical and

---

[4] Although the distributions of the innovation terms will be assumed, the proposed evaluation framework can incorporate estimated distributions as in Engle and Gonzalez-Rivera (1991).

economic, previously used for volatility forecast evaluation are reviewed. Section III describes the three elements of the proposed evaluation framework that are specified by the user's loss function -- the events to be forecast, the scoring rule used to evaluate the forecasts, and the subsets of the forecasts of particular interest. Section IV presents an empirical example using daily foreign exchange rates to illustrate the evaluation procedure. Five economic loss functions implicitly based on volatility forecasts are used to specify the elements of the evaluation framework, and volatility forecasts from several models are evaluated. Section V concludes and suggests directions for further research.

II. Previous Evaluations of Volatility Forecasts

Volatility models are generally expressed as

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t = h_t^{1/2} v_t, \qquad v_t \sim D(0, 1),$$

where the conditional mean $\mu_t = E\left[y_t \mid \Omega_{t-1}\right]$, Ω_{t-1} is the information set available at time t-1, ε_t is the innovation term, h_t is its conditional variance and D(0,1) is a symmetric, standardized distribution, such as the normal or t-distribution; thus, $\varepsilon_t \mid \Omega_{t-1} \sim D\left(0, h_t\right)$. For example, the GARCH(p,q) model is characterized as $h_t = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$.

The parameters of a volatility model are estimated over a specified in-sample period t = 1, ..., T, and volatility forecasts based on these estimates are generated over the out-of-sample period t = T+1, ..., T+T*.[5]

Several in-sample procedures, such as maximization of the log-likelihood function or fitting the news impact curve proposed by Engle and Ng (1993), are used to compare volatility model specifications. However, out-of-sample forecast accuracy provides a more useful comparison; a model generating accurate forecasts provides a reasonable approximation of the

[5] In a slight abuse of terminology, in-sample, fitted conditional variances are often called in-sample forecasts. This convenient phrasing will be used in this paper.

underlying data generating process. Out-of-sample volatility forecasts have been evaluated using three types of loss functions -- mean-squared error and related statistical measures, utility-based loss functions and profit-based loss functions.

2a. Mean-Squared Error and Related Statistical Measures

The mean-squared error (MSE) criterion is commonly used to evaluate both in-sample and out-of-sample volatility forecasts. Focussing on out-of-sample forecasts, MSE is the average squared difference between h_{T+t}, the actual conditional variance, and the corresponding volatility forecast. However, since h_{T+t} is unobservable, the squared innovation ε^2_{T+t} is used as a proxy. Thus, MSE $= \frac{1}{T^*}\sum_{t=1}^{T^*}\left(\varepsilon^2_{T+t} - \hat{h}_{T+t}\right)^2$, where $\left\{\hat{h}_{T+t}\right\}_{t=1}^{T^*}$ are the one-step-ahead forecasts based on the in-sample parameter estimates and the relevant information sets.[6]

Volatility forecast evaluation based on MSE have two shortcomings. First, even though ε^2_{T+t} is an unbiased estimator of h_{T+t}, it is an imprecise or "noisy" estimator. For example, if $\nu_{T+t} \sim N(0, 1)$, $\varepsilon^2_{T+t} = h_{T+t}\nu^2_{T+t}$ has a conditional mean of h_{T+t} since $\nu^2_{T+t} \sim \chi^2_{(1)}$. However, since the median of a $\chi^2_{(1)}$ distribution is 0.455, $\varepsilon^2_{T+t} < \frac{1}{2}h_{T+t}$ more than 50% of the time. In fact,

$$\Pr\left(\varepsilon^2_{T+t} \in \left[\frac{1}{2}h_{T+t}, \frac{3}{2}h_{T+t}\right]\right) = \Pr\left(\nu^2_{T+t} \in \left[\frac{1}{2}, \frac{3}{2}\right]\right) = 0.2588;$$

that is, even if one is willing to accept a proxy that is up to 50% different than h_{T+t}, ε^2_{T+t} would fulfill this condition only 25% of the time. This proxy introduces a source of error into volatility forecast evaluations that diminishes as T^* increases.

---

[6] Although this paper focusses on one-step-ahead volatility forecasts, k-step-ahead forecasts can be generated and evaluated using statistical loss functions. However, generating the corresponding probability forecasts is more difficult since the conditional distribution of the k-step-ahead innovation term is not easily determined, as shown in Baillie and Bollerslev (1992) for GARCH models.

The second and more important difficulty with MSE is the quadratic loss function that it implies. As discussed by Bollerslev *et al.* (1994), such a loss function penalizes negative volatility forecasts (which are meaningless) and positive forecasts symmetrically. In addition, the in-sample MSE is minimized by forecasts based on a least squares regression of ε_t^2 on $\Omega_{t-1} = \{\varepsilon_{t-1}, ..., \varepsilon_1\}$. Thus, MSE is a purely statistical loss function that addresses neither the specific requirements of volatility forecast evaluation nor the economic issues underlying such evaluations.

Alternative statistical loss functions are proposed in the literature; for example, mean absolute error (MAE), $MAE = \frac{1}{T^*}\sum_{t=1}^{T^*}\left|\varepsilon_{T-t}^2 - \hat{h}_{T-t}\right|$, although the same criticisms apply.

Two loss functions that penalize volatility forecasts asymmetrically are the logarithmic loss (LL) function employed by Pagan and Schwert (1990),

$$LL = \frac{1}{T^*}\sum_{t=1}^{T^*}\left[\ln\left(\varepsilon_{T-t}^2\right) - \ln\left(\hat{h}_{T-t}\right)\right]^2,$$

and the heteroskedasticity-adjusted MSE (HMSE) of Bollerslev and Ghysels (1994),

$$HMSE = \frac{1}{T^*}\sum_{t=1}^{T^*}\left[\frac{\varepsilon_{T-t}^2}{\hat{h}_{T-t}} - 1\right]^2.$$

Bollerslev *et al.* (1994) suggest the loss function implicit in the Gaussian quasi-maximum likelihood function often used in estimating volatility models; that is,

$$GMLE = \frac{1}{T^*}\sum_{t=1}^{T^*}\left[\ln\left(\hat{h}_{T-t}\right) + \frac{\varepsilon_{T-t}^2}{\hat{h}_{T-t}}\right]$$

However, these alternative loss functions are also based on purely statistical considerations. Loss functions based on users' economic costs provide more meaningful forecast evaluations.

2b. Utility-Based Loss Functions

West *et al.* (1993) evaluate volatility forecasts using an economic loss function based

on expected utility; the criterion is how much an investor with a mean-variance utility function would pay to use the forecasts from a particular model. The utility function is quadratic in wealth, and in the investment framework specified, the first two moments of the asset return distributions are assumed to be finite. The average utility over the out-of-sample period, denoted as U, is

$$U = \frac{1}{T^*}\sum_{t=1}^{T^*} W\left(c_{T+t} + d_{T+t}\, f\left(\varepsilon^2_{T+t}, \hat{h}_{T+t}; \phi_{T+t}\right)\right).$$

The c_{T+t} and d_{T+t} terms are functions of the asset returns and the coefficient of relative risk aversion, and W is the wealth required to achieve U given the volatility forecasts, ε^2_{T+t} and ϕ_{T+t}, the interest rate differential. Note that

$$f\left(\varepsilon^2_{T+t}, \hat{h}_{T+t}; \phi_{T+t}\right) = \left(\phi^2_{T+t} + \hat{h}_{T+t}\right)^{-1} - \frac{1}{2}\left(\phi^2_{T+t} + \varepsilon^2_{T+t}\right)\left(\phi^2_{T+t} + \hat{h}_{T+t}\right)^{-2}$$

is asymmetric with respect to $\hat{h}_{T+t}$. The optimal model is defined as the one whose forecasts require the lowest wealth, say W^*, to achieve a specified U. The difference between W^* and W_j, the wealth required using model j's forecasts, is viewed as the fee an investor would pay to use the optimal forecasts; the smaller the fee, the closer the forecasts are to being optimal.

This evaluation framework is economically meaningful since it is motivated by the investment decisions of risk-averse utility maximizers, but it has two weaknesses. It relies on the ε^2_{T+t} proxy for the unobservable h_{T+t}, and the economic loss function is very specific and highly parameterized, especially with respect to the utility function and the moments of the asset return series. Although the evaluation results are robust across parameterizations, this utility-based criterion would not be useful to most forecast users.

2c. Profit-Based Loss Functions

Engle *et al.* (1993) propose an economic loss function based on trading profits earned

in a simulated options market.[7] The market consists of M traders, each using one of the volatility models to be examined. The traders price options on the chosen stock portfolio, such as the S&P 500, at the start of a trading period using the one-step-ahead forecasts from their models and the Black-Scholes option-pricing formula. The traders then examine the set of forecast-induced prices and trade according to a specified algorithm. At the end of the period, the realized value of the portfolio return is used to determine the traders' profits. The trader with the largest accumulated profit is said to have used the model generating the most appropriate volatility forecasts.

In addition to being economically relevant, this profit criterion avoids proxying for the unobservable conditional variances; instead, it relies on the observed value of the stock portfolio for evaluating the forecasts. However, this loss function addresses only options trading, which is irrelevant to many other users of volatility forecasts. Moreover, the procedure's complicated structure, which entails specifying a trading algorithm and selecting an option-pricing formula, diminishes its usefulness.

The forecast evaluation framework proposed in Section III retains the advantages of these two economic criteria and addresses their weaknesses. The framework incorporates a more general class of economic loss functions, is based solely on observable outcomes, and has a relatively simple structure.

III. Generating and Evaluating Probability Forecasts

The proposed forecast evaluation framework consists of two stages at which elements of economic loss functions are incorporated into the evaluation procedure. The first stage consists of transforming volatility forecasts into probability forecasts of events of interest to the user. In the second stage, the probability forecasts are evaluated using statistical tools tailored to the user's loss function. The first part of this section describes how the probability forecasts are generated, and the remainder describes the evaluation criteria employed -- probability scoring rules, the predictive accuracy tests of Diebold and Mariano (1995), and the calibration

---

[7] This line of research is continued in Engle *et al.* (1993) and Noh *et al.* (1994).

tests of Seillier-Moiseiwitsch and Dawid (1993).

3a. Transforming Volatility Forecasts into Probability Forecasts

Given the assumption that $\varepsilon_t \mid \Omega_{t-1} \sim D(0, h_t)$ volatility forecasts are transformed into probability forecasts by integrating over the standardized distribution $D(0, 1)$. The range of integration is chosen by the user in accordance with their loss function and its implied economic events of interest; thus, a general class of economic loss functions implicitly or explicitly based on volatility forecasts can be incorporated into the evaluation framework. To illustrate this, two categories of implicit economic loss functions and their associated economic events are discussed.

The probability forecast notation used is as follows. Volatility models are fit to the in-sample observations $t = 1,\ldots,T$, and P_t is the in-sample probability forecast for time t based on the estimated parameters and Ω_{t-1}. Out-of-sample, P_{T+t}, $t = 1,\ldots,T^*$, is the one-step-ahead probability forecast conditional on the parameter estimates and Ω_{T+t-1}. The subsequent discussion will focus on out-of-sample forecasts. The event space to be examined is created by partitioning the set of all possible outcomes into N mutually exclusive and collectively exhaustive subsets according to the forecast user's interests. If N=2, a binomial event is specified, and P_{T+t} and R_{T+t}, an indicator variable equalling one if the event occurs and zero otherwise, are scalar variables.[8]

An important category of volatility forecast users are those interested in the behavior of the innovation term ε_{T+t}. This category includes, for example, spot traders or options traders structuring their hedging strategies with respect to ε_{T+t}. For such users' economic loss functions, the event of interest is $\varepsilon_{T+t} \in [L_{\varepsilon,T+t}, U_{\varepsilon,T+t}]$ and the associated probability forecast

[8] If N>2, P_{T+t} is an (Nx1) vector probability forecast such that $P_{n,T+t} \geq 0$ for $n = 1,\ldots,N$ and $\sum_{n=1}^{N} P_{n,T+t} = 1$. This paper focusses exclusively on binomial events.

is $\Pr\left(L_{\varepsilon,T-t} \le \varepsilon_{T-t} \le U_{\varepsilon,T-t}\right)$. Given $\hat{h}_{T-t}$ and the assumed distribution of ε_{T-t},

$$P_{T-t} = \Pr\left(L_{\varepsilon,T-t} \le \varepsilon_{T-t} \le U_{\varepsilon,T-t}\right) = \Pr\left(\frac{L_{\varepsilon,T-t}}{\sqrt{\hat{h}_{T-t}}} \le z_{T-t} \le \frac{U_{\varepsilon,T-t}}{\sqrt{\hat{h}_{T-t}}}\right) = \int_{l_{\varepsilon,T-t}}^{u_{\varepsilon,T-t}} f\left(z_{T-t}\right) dz_{T-t},$$

where z_{T+t} is the standardized innovation, $f\left(z_{T-t}\right)$ is the functional form of $D(0, 1)$ and $\left[l_{\varepsilon,T+t},\ u_{\varepsilon,T-t}\right]$ is the standardized range of integration. Forecast users thus incorporate their loss functions directly into the probability forecasts by specifying the appropriate $\left[L_{\varepsilon,T+t},\ U_{\varepsilon,T-t}\right]$ interval.

The second category of economic loss functions is characterized by economic events based on the behavior of the level term y_{T+t}. Such loss functions would be relevant to a central bank forecasting whether an exchange rate will remain within a target zone or a portfolio manager comparing cross-country asset returns. In such cases, the event of interest is $y_{T-t} \in \left[L_{y,T-t},\ U_{y,T-t}\right]$, the probability forecast is $\Pr\left(L_{y,T-t} \le y_{T-t} \le U_{y,T-t}\right)$, and

$$P_{T-t} = \Pr\left(L_{y,T-t} \le y_{T-t} \le U_{y,T-t}\right) = \Pr\left(L_{y,T-t} - \hat{\mu}_{T-t} \le \varepsilon_{T-t} \le U_{y,T-t} - \hat{\mu}_{T-t}\right)$$

$$= \Pr\left(\frac{L_{y,T-t} - \hat{\mu}_{T-t}}{\sqrt{\hat{h}_{T-t}}} \le z_{T-t} \le \frac{U_{y,T-t} - \hat{\mu}_{T-t}}{\sqrt{\hat{h}_{T-t}}}\right) = \int_{l_{y,T-t}}^{u_{y,T-t}} f\left(z_{T-t}\right) dz_{T-t},$$

where $\hat{\mu}_{T-t}$ is the forecasted conditional mean and $\left[l_{y,T-t},\ u_{y,T-t}\right]$ is the standardized range of integration.[9] For example, a central bank's target zone would be $\left[L_{y,T-t},\ U_{y,T-t}\right] = [L, U]$.

Once the desired probability forecasts are generated, forecast users can evaluate them using probability scoring rules and calibration tests, statistical criteria that accommodate

---

[9] A special case of probability forecasts based on y_{T+t} is examined in Granger *et al.* (1989). Interval forecasts $\left\{\left(L_{y,T-t},\ U_{y,T-t}\right)\right\}_{t=1}^{T^*}$ are set to generate a constant α% confidence interval around the corresponding conditional mean forecasts. The proposed LIH statistic, which is based on the χ^2 goodness-of-fit test, is used to evaluate whether the forecasts adequately represent the observed outcomes. Dawid (1986) alludes to, but does not propose, such a statistic. Further tests for evaluating interval forecasts are proposed in Christoffersen (1995).

elements of users' loss functions. These two forecast evaluation criteria, as well as the Diebold-Mariano (1995) tests of comparative predictive accuracy, are briefly described in the following subsections; see Diebold and Lopez (1995b) for further discussion.

3b. Probability Scoring Rules

Probability scoring rules are primarily employed in the economics literature to evaluate business-cycle turning-point probabilities as in Diebold and Rudebusch (1989), Ghysels (1993), and Lahiri and Wang (1994).[10] Scoring rules measure the "goodness" of the forecasted probabilities, as defined by the forecast user. Thus, a user's economic loss function is used to select the probability scoring rule (i.e., statistical loss function) with which to evaluate the transformed volatility forecasts.

The quadratic probability score (QPS), developed by Brier (1950) for evaluating weather forecasts, is the most common probability scoring rule. The QPS is the analog of MSE for probability forecasts and thus implies a quadratic loss function.[11] The QPS over a forecast sample of size T^* is

$$QPS = \frac{1}{T^*} \sum_{t=1}^{T^*} 2\left(P_{T-t} - R_{T-t}\right)^2,$$

which implies that QPS $\in [0,2]$ and has a negative orientation (i.e., smaller values indicate more accurate forecasts). The QPS is a proper scoring rule, which means that forecasters should report their actual forecasts to minimize their expected QPS score.

In addition to being intuitively simple, QPS is a useful scoring rule because it highlights various attributes of probability forecasts. The three main attributes of probability forecasts are accuracy, calibration and resolution. Accuracy refers to the closeness, on average, of the predicted probabilities to the observed realizations and is directly measured by QPS. Calibration refers to the degree of equivalence between the forecasted and observed

---

[10] Probability scoring rules are also used in Fair (1993) to evaluate probability forecasts generated by stochastic simulation.

[11] Other scoring rules, such as the logarithmic score, with different implied loss functions are available; see Murphy and Daan (1985) for further discussion.

frequencies of occurrence. An overall measure of calibration is global squared bias (GSB),

$$GSB = 2\left(\bar{P} - \bar{R}\right)^2,$$

where $\bar{P} = \frac{1}{T^*}\sum_{t=1}^{T^*} P_{T+t}$, $\bar{R} = \frac{1}{T^*}\sum_{t=1}^{T^*} R_{T+t}$ and GSB $\in [0,2]$ with a negative orientation.

Calibration can also be examined in subsets of probability forecasts using the local squared bias (LSB) measure; that is, calibration can be examined within J mutually exclusive and collectively exhaustive subsets created by the forecast user. The LSB is

$$LSB = \frac{1}{T^*}\sum_{j=1}^{J} 2T_j^*\left(\bar{P}_j - \bar{R}_j\right)^2,$$

where $\bar{P}_j$ and $\bar{R}_j$ are the forecasted and observed frequencies of occurrence in subset j, T_j^* is the number of forecasts in subset j, and LSB $\in [0,2]$ with a negative orientation.[12]

The resolution (RES) of probability forecasts is the correspondence between $\bar{R}_j$ and $\bar{R}$; that is,

$$RES = \frac{1}{T^*}\sum_{j=1}^{J} 2T_j^*\left(\bar{R}_j - \bar{R}\right)^2.$$

Note that RES ≥ 0 and has a positive orientation; as the resolution increases, probability forecasts more accurately to describe the underlying process. Since RES depends only on the defined J subsets, it measures how well a series of events can be forecast with respect to the subsets relevant to the forecast user.

The QPS can be decomposed as $QPS = QPS_{\bar{R}} + LSB - RES$, where $QPS_{\bar{R}}$ is QPS evaluated with $P_{T+t} = \bar{R}\ \forall t$. Thus, QPS is a particularly useful scoring rule because it can be used to examine several attributes of probability forecasts. As suggested in Diebold and Rudebusch (1989), this decomposition alludes to a more general decomposition, $F\left[g\left(\bar{R}\right), LSB, RES\right]$, where the functions F and g are determined by the relevant economic

[12] Note that LSB = 0 implies GSB = 0, but not conversely.

loss function.[13] This decomposition underscores the main advantage of using probability-based criteria for evaluating volatility forecasts: a forecast user's economic loss function can be introduced into the evaluation process in several ways.

3c. Comparing the Predictive Accuracy of Probability Forecasts

Scoring rules measure the accuracy of probability forecasts; for example, if the QPS for $\{P_{A,T-t}\}_{t=1}^{T^*}$ is closer to zero than the QPS for $\{P_{B,T-t}\}_{t=1}^{T^*}$, then the forecasts from model A are more accurate than those from model B. However, the statistical significance of the difference between the two QPS values is unclear. Diebold and Mariano (1995) propose several tests for determining whether the expected losses induced by two sets of *point* forecasts under a general loss function are statistically different. These tests are readily generalized to probability forecasts.

The null hypothesis under a loss function g is $E\left[g\left(P_{A,T-t}, R_{T-t}\right)\right] = E\left[g\left(P_{B,T-t}, R_{T-t}\right)\right]$, or, equivalently, $E\left[d_{T-t}\right] = E\left[g\left(P_{A,T-t}, R_{T-t}\right) - g\left(P_{B,T-t}, R_{T-t}\right)\right] = 0$.[14] For QPS,

$$d_{T-t} = 2\left(P_{A,T-t} - R_{T-t}\right)^2 - 2\left(P_{B,T-t} - R_{T-t}\right)^2,$$

where $P_{A,T+t}$ and $P_{B,T+t}$ are the probability forecasts from two volatility models. Several statistics are proposed for testing this null hypothesis. The first is the asymptotic mean statistic defined as

$$S_1 = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T^*}}} \overset{a}{\sim} N(0,1),$$

where $\bar{d}$ is the sample mean of d_{T+t} and $\hat{f}_d(0)$ is the spectral density function at frequency zero estimated using a rectangular lag window.

---

[13] Winkler (1993) presents a family of asymmetric scoring rules that address this generalization indirectly by permitting users' loss functions to shape the appropriate scoring rule.

[14] To employ these tests, $\left\{d_{T-t}\right\}_{t=1}^{T^*}$ must be covariance stationary, a condition determined empirically.

The second and third statistics require d_{T+t} to be serially uncorrelated, an issue that must be empirically determined.[15] These statistics, which test the hypothesis that the median of d_{T+t} is zero, are based on the sign test and the Wilcoxon signed-rank test. The sign statistic is $S_2 = \sum_{t=1}^{T^*} I_{T+t}$, where I_{T+t} equals one if $d_{T+t} > 0$ and zero otherwise, and the Wilcoxon signed-rank statistic is $S_3 = \sum_{t=1}^{T^*} I_{T+t} \text{Rank}(|d_{T+t}|)$, where $\text{Rank}(|d_{T+t}|)$ is the rank of the absolute value of d_{T+t} in descending order. Note that S_3 further requires d_{T+t} to be symmetrically distributed about its median. Finite sample critical values are available for these two non-parametric tests, and the standardized forms of these statistics are asymptotically normally distributed.

3d. Calibration Tests of Probability Forecasts

As previously discussed, calibration is the degree of equivalence between an event's observed and predicted frequencies of occurrence within subsets of probability forecasts. For example, a forecaster is providing perfectly calibrated rain forecasts if it rained on 10% of the days for which a 10% chance of rain was forecast. A simple measure of calibration is the calibration plot, which graphs Pr(event occurs | P_j) against P_j for the forecast subsets j = 1,...,J created by the user. The degree to which the graph lies on the 45° line is a visual measure of calibration. Seillier-Moiseiwitsch and Dawid (1993) present test statistics that formalize this analysis.[16]

The null hypothesis is that the predicted frequency of occurrence for a specified binomial event equals the observed frequency of occurrence. The statistics are constructed by dividing the out-of-sample probability forecasts into J mutually exclusive and collectively exhaustive subsets according to the user's interests. Let π_j denote the midpoint of subset j,

[15] If serial correlation does exist, an adjustment can be made as described in Diebold and Mariano (1995).

[16] Although several components of the proposed forecast evaluation framework can be generalized to multinomial events, the calibration tests are designed exclusively for binomial events.

T_j^* the number of forecasts in subset j, and R_j the number of observed events paired with forecasts in subset j. The test statistics are the subset j calibration statistics,

$$Z_j = \frac{\left(R_j - T_j^* \pi_j\right)}{\left[T_j^* \pi_j (1 - \pi_j)\right]^{1/2}} = \frac{(R_j - e_j)}{w_j^{1/2}}, \quad j = 1,\ldots,J,$$

and the overall calibration statistic,

$$Z_0 = \frac{(R_0 - e_0)}{w_0^{1/2}},$$

where $R_0 = \sum_{j=1}^{J} R_j$, $e_0 = \sum_{j=1}^{J} e_j$ and $w_0 = \sum_{j=1}^{J} w_j$.[17] Each of these statistics is the square root of a χ^2 goodness-of-fit statistic for a binomial event with R_j observed outcomes and e_j expected outcomes.[18]

Under the null hypothesis and weak conditions on the distribution of the probability forecasts, these statistics are asymptotically normally distributed. If the null hypothesis for the forecasts in subset j is rejected, then these forecasts do not adequately represent the observed frequency of occurrence in that subset. The calibration test results complement the other sets of results by evaluating the transformed volatility forecasts using another element of the user's economic loss function.

In summary, a framework for volatility forecast evaluation under a general class of economic loss functions is proposed; such loss functions are used to specify events to be forecast, a probability scoring rule and subsets of probability forecasts. These three elements are used to generate probability forecasts, both in-sample and out-of-sample, and tailor statistical criteria for evaluating the forecasts. In the following section, an empirical

---

[17] Under weak conditions on the distribution of the probability forecasts, the Z_j statistics are asymptotically independent, which allows Z_0 to be constructed as the sum of the Z_j statistics.

[18] The similarity between LSB and the calibration tests is also noteworthy. Both statistics examine the differences between the observed and expected event frequencies within subsets of probability forecasts. However, the calibration statistics are directly linked to the χ^2 goodness-of-fit test and permit hypothesis testing.

application illustrating the evaluation procedure under a variety of loss functions is presented.

IV. Evaluating Exchange Rate Volatility Forecasts

To illustrate the proposed evaluation framework, one-step-ahead volatility forecasts of several foreign exchange rates are evaluated under economic loss functions implicitly based on volatility forecasts; that is, five implicit economic loss functions are used to specify the three elements needed for the probability-based forecast evaluations. As might be expected, no one set of forecasts minimizes all of the probability-based loss functions. In fact, not only does the minimizing set of forecasts vary across loss functions and exchange rates, it also varies across the in-sample and out-of-sample periods. These results highlight two key features of volatility forecast evaluation (and forecast evaluation, in general): the need for careful selection of the loss functions employed, and the usefulness of out-of-sample results for model specification.

4a. Exchange Rate Data

The three exchange rates examined are the logged daily Deutchsemark (DM) and Canadian dollar (CD) spot exchange rates with respect to the U.S. dollar and the U.S. dollar exchange rate with respect to the British pound (BP), as recorded by the Federal Reserve Bank of New York, from 1980 to 1991. The in-sample period used for model estimation is 1980-1989 (2508 observations), and the out-of-sample period is 1990-1991 (502 observations). The data series and their first differences are plotted in Figure 1.

Given established unit root results, the log exchange rates are modelled as I(1) processes;[19] that is,

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_t = y_{t-1} + v_t, \qquad \varepsilon_t \mid \Omega_{t-1} \sim D(0, h_t)$$

where $y_t \equiv 100 \log(s_t)$ and s_t is any of the three spot rates. Thus,

$$\Delta y_t = v_t + \varepsilon_t.$$

---

[19] Kim and Schmidt (1993) show that the Dickey-Fuller unit root tests are not seriously affected by the presence of conditional heteroskedasticity in finite samples, although the tests do overreject in certain cases. The results of the unit root tests for this dataset are available upon request.

For the DM and CD series, $v_t = \mu$, but since the BP series exhibits weak first-order serial correlation, $v_t = \mu + \theta \varepsilon_{t-1}$. Table 1 presents the in-sample least squares estimates of these conditional mean parameters, the first three moments of the in-sample and out-of-sample ε_t series and the portmanteau statistics for up to 15th order serial correlation in the ε_t and ε_t^2 series. The kurtosis estimates and the $Q^2(15)$ statistics indicate the presence of conditional heteroskedasticity in all three series.

4b. Volatility Models and Forecasts

Seven volatility models are estimated over the in-sample period, and the parameter estimates are used to generate volatility forecasts over the out-of-sample period.[20] The seven sets of in-sample and out-of-sample volatility forecasts for each series are graphed in Figure 2. These superimposed graphs highlight the relatively small degree of variation across the sets of forecasts. This graphical result suggests that several sets of forecasts may be indistinguishable when evaluated under the relevant loss function.

The first volatility model is the simple Gaussian homoskedastic model (homo.), which assumes $h_t = \sigma^2$. The next three models are drawn from the GARCH family of models: the GARCH(1,1) model with $h_t = \omega + \alpha \varepsilon_t^2 + \beta h_{t-1}$ and the IGARCH(1,1) model with $h_t = \omega + \alpha \varepsilon_t^2 + (1 - \alpha) h_{t-1}$, both assuming conditional normality, and the GARCH(1,1) model assuming the conditional t-distribution.[21] The models are denoted as GARCH, IGARCH and GARCH-t, and all four of these models are estimated using maximum likelihood.

As in West and Cho (1994), two autoregressive models of ε_t are estimated using

[20] The parameter estimates and relevant standard errors are reported in the Appendix.

[21] The estimated degrees of freedom for the DM, CD and BP series are 5.7, 6.4 and 6.6, respectively. These estimates are similar to those presented in the literature, such as Baillie and Bollerslev (1989).

ordinary least squares and an assumed conditionally normal distribution. The autoregressive ε_t^2 model assuming a lag-order of 12 (AR12sq) is $h_t = \omega + \sum_{i=1}^{12} \alpha_i \varepsilon_{t-i}^2$. Volatility forecasts from this model can be negative, and in such cases, the forecasts are set to a small positive number. The second autoregressive model (AR12ab), as developed by Davidian and Carroll (1987) and Schwert (1989), is based on $|\varepsilon_t|$ and is specified as

$$h_t = \frac{\pi}{2}\left(\omega + \sum_{i=1}^{12} \alpha_i |\varepsilon_{t-i}| \right).$$

The seventh model is the stochastic volatility model (s.v.) used by Harvey *et al.* (1994); that is, after removing $\hat{v}_t$ based on least squares parameter estimates,

$$\varepsilon_t = \exp(\alpha_t / 2)\, v_t, \quad v_t \sim N(0, 1)$$

$$\alpha_t = \phi\, \alpha_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2),$$

where $v_t \perp \eta_t$. Thus, the innovation term is subject to two independent shocks. Estimation of the model is conducted using the Kalman filter on the measurement equation $\log \varepsilon_t^2 = \alpha_t + \log v_t^2$ and the transition equation $\alpha_t = \phi\, \alpha_{t-1} + \eta_t$. To use quasi-maximum likelihood methods, the measurement equation is rewritten as $\log \varepsilon_t^2 = \omega + \alpha_t + \xi_t$, where $\omega = E\left[\log v_t^2\right]$ and ξ_t is assumed to be distributed $N(0, \pi^2/2)$.[22] The log-likelihood function to be maximized is

$$\ln L(\theta ; \varepsilon_1, ..., \varepsilon_T) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\sum_{t=1}^{T} \log(G_t) + \sum_{t=1}^{T} \frac{(\log \varepsilon_t^2 - \omega - a_t)^2}{G_t},$$

where a_t and G_t are the Kalman filter estimates of α_t and its conditional variance, respectively,

---

[22] Jacquier *et al.* (1994) and Kim and Shepard (1994) present alternative estimation techniques that employ more appropriate distributional assumptions. See Geweke (1994) for a Bayesian comparison of stochastic volatility models and GARCH models.

and $\theta = \left[\phi, \omega, \sigma_\eta^2\right]$

One-step-ahead volatility forecasts from the s.v. model are easily obtained using the Kalman filter, but converting them into probability forecasts is not as straightforward. The standardized innovation term $z_t = e^{\eta_t/2} v_t$ is the product a standard normal and a lognormal distribution with mean $e^{\sigma_\eta^2/8}$ and variance $e^{\sigma_\eta^2/4}\left(e^{\sigma_\eta^2/4} - 1\right)$. Integration over the distribution of this random variable, which is required for generating probability forecasts, is both difficult and computationally intensive. To avoid this problem, its distribution is empirically approximated by simulating 100,000 draws from it and creating 0.5% quantiles. The probability that $z_t \in (l_t, u_t)$ is approximated as $P_t = \hat{F}(u_t) - \hat{F}(l_t)$, where $\hat{F}$ is the empirical cumulative distribution function of z_t.[23]

Before evaluating the probability forecasts, Tables 2 and 3 report the forecast evaluation results using the statistical loss functions in Section II. In the top portion of each panel, the columns of the panels represent the five loss functions, the rows represent the seven sets of volatility forecasts, and the lowest loss function value in each column is underlined. The most obvious result is that the forecasts minimizing the loss functions vary considerably within the in-sample and out-of-sample periods. This variation holds across the two sample periods and the three exchange rates. Of the 15 possible in-sample vs. out-of-sample comparisons, the set of minimizing forecasts match only 8 times. Across the three series, one set of forecasts minimizes the same loss function in only three cases -- in-sample MSE (as expected), in-sample HMSE and out-of-sample LL. In short, no clear patterns emerge from these results.

The S tests are conducted to determine whether the differences between the smallest

[23] Although this estimated c.d.f. procedure introduces another source of error into the probability forecasts generated from the s.v. model, this error is irrelevant to the purpose of the exercise, which is to evaluate competing volatility forecasts, regardless of their source of origin.

loss function values and the other six values are statistically significant.[24] The summary results are reported in the bottom portion of each panel in Tables 2 and 3; for each loss function, the models whose forecasts *reject* the null hypothesis at the 10% significance level for more than one of the S tests are listed.[25] Once again, the results for a given loss function vary considerable; in some cases, the null hypothesis is never rejected while for others, the loss function values of several forecasts are clearly different from the minimum value. The overall inability to reject the null hypothesis (as well as the graphs in Figure 3) suggests that no one set of volatility forecasts minimizes all of these statistical loss functions.

4c. Specifying Economic Loss Functions

In this subsection, the implicit economic loss functions used to evaluate the volatility forecasts are described. The proposed evaluation framework uses these loss functions to specify the events to be forecast, the probability scoring rule used to evaluate the forecasts, and the subsets of these forecasts of particular interest. To simplify the evaluation exercise, only the events to be forecast will be varied. The QPS is the probability scoring rule used throughout, and the subsets are the quartiles {[0,0.25); [0.25,0.5); [0.5;0.75); [0.75,1]}.[26] In total, five loss functions based on probability forecasts are specified.

The first two events are based on the innovation term ε_t: (1). $\varepsilon_t \in [\gamma_1, \gamma_2]$ where $\gamma_1 < 0 < \gamma_2$, and (2). $\varepsilon_t \in [\gamma_2, \gamma_3]$. The first event examines an interval with a relatively high probability of occurrence, and in contrast. the second event examines the upper tail of the innovation's distribution. Such events would be of interest to spot traders deciding how to structure their market positions and to options traders holding out-of-the-money positions,

---

[24] West and Cho (1994) propose an alternative procedure for testing the null hypothesis that the root-MSE of different sets of volatility forecasts are equal. West (1994) notes that the S tests ignore the uncertainty caused by parameter estimation.

[25] The S test results as well as the diagnostics of the loss differential series are available upon request.

[26] Although selected to simplify the exercise, these subsets of the probability forecasts are reasonable. For example, an options trader or portfolio manager may take different actions based upon which quartile a probability forecast falls into.

respectively. For DM and BP, $(\gamma_1, \gamma_2, \gamma_3) = (-0.5, 1, 2.5)$. In order to define similar events for all three series, $(\gamma_1, \gamma_2, \gamma_3) = (-0.2, 0.4, 1)$ for CD because of its lower unconditional variance.

The last three events are based on the level term y_t. The third event is a $\pm 2\%$ move in y_t; that is, $y_t \in [0.98y_{t-1}, 1.02y_{t-1}]$. This event would be of interest to a portfolio manager deciding if and when to transact in a foreign-currency denominated asset. For comparison purposes, the fourth event is specified as just a $+2\%$ move in y_t. The last event is based on a central bank forecasting whether an exchange rate will remain within a specified target zone. The event of interest is $y_t \in [L, U]$, where the bounds are set by the central bank. In this evaluation exercise, L and U are set as $\pm\gamma_4\%$ of y_T, the last in-sample observation. For DM and BP, $\gamma_4 = 5$, and for CD, $\gamma_4 = 2.5$.

Table 4 presents the in-sample and out-of-sample occurrence rates (or $\bar{R}$'s) of the five events for the three exchange rates. As expected, the occurrence rates of the second event, a "tail" event for ε_t, are much lower than those of the first event, and the $\bar{R}$'s for the fourth event are roughly half of those for the third event since the empirical distributions of ε_t are symmetric. The last event exhibits the greatest variation in occurrence rates since it is based on a nonstationary series crossing one of two fixed points.

The GSB compares the predicted and actual frequencies of an event's occurrence; recall $GSB = 2(\bar{P} - \bar{R})^2$. The in-sample and out-of-sample GSB values for the five events are presented in the panels of Table 5. Most of the in-sample GSB values are close to zero, implying that the probability forecasts are well calibrated on average and that the models provide a reasonable in-sample fit. The out-of-sample GSB values are generally higher than in-sample, particularly for the last event, but they are still close to zero. Since they are far from the upper GSB bound of two, the out-of-sample forecasts are also well calibrated on average.

4d. Forecast Evaluation Results

The panels of Table 6 contain the in-sample QPS values for the three currencies. As before, the columns of the panels represent the five events, the rows represent the seven sets of generated probability forecasts, and the lowest value in each column is underlined. The panels also list the sets of forecasts for which the null hypotheses of the S tests are *rejected* at the 10% level for more than one of the tests.

For the two innovation events, the AR12ab forecasts minimize the in-sample QPS in 5 of 6 possible cases. However, the results for the three level events are not as clear within or across series. As with the statistical loss functions based on ε_{τ}^{2}, the forecast evaluation results are sensitive to the chosen loss function. The S test results also indicate that for a given probability-based loss function, the differences between the lowest QPS value and the other six QPS values are usually not statistically significant. Thus, no one set of forecasts clearly minimizes the specified loss functions, although several sets of forecasts can be rejected relative to the minimizing set. A forecast user may choose the minimizing set of forecasts, but should be aware that the choice is made with generally little statistical significance.

The out-of-sample evaluation results are presented in Table 7. Once again, considerable variation in the minimizing set of forecasts is seen within and across the series. The only probability-based loss function minimized across the series by one set of forecasts is event 5 minimized by the s.v. forecasts, but the reason for this result is not immediately clear. The S test results suggest that, although certain forecasts do reject the null hypotheses relative to the QPS-minimizing forecasts, no one set significantly minimizes any of the loss functions. As with the statistical loss functions, comparison of the in-sample and out-of-sample results indicates considerable variation. Of the 15 possible comparisons, only 7 matches occur; further highlighting the need for out-of-sample forecast evaluation for improving model specification.

Calibration tests are used to further evaluate the out-of-sample probability forecasts under the specified loss functions and thus complement the QPS results. Recall that for this exercise, the same forecast subsets are specified for the five economic loss functions. The

summary results for the overall calibration tests are presented in the bottom portion of the panels in Table 7; the lists contain the models whose Z_0 statistics *do not reject* the null hypothesis that the predicted and observed frequencies of occurrence are equal (i.e., the forecasts are well calibrated) at the 10% significance level. In 8 of the 15 cases, none of the forecasts are well calibrated.[27] However, for the other 7 cases, the set of forecasts minimizing the QPS is well calibrated. Thus, the calibration results certainly support choosing the QPS-minimizing forecasts as the most appropriate in these cases.

In summary, this empirical exercise presents two clear results. First, forecast evaluations are directly affected by the selected economic loss function. Forecast users must select the loss functions that most closely represent their economic interests, and the proposed evaluation framework is capable of incorporating a large class of economic loss functions. Second, the differences in evaluation results across the in-sample and out-of-sample periods strongly suggest that reliance on measures of in-sample fit are insufficient. Out-of-sample forecast evaluation under the relevant loss function is necessary for improving model specification.

V. Conclusions

Given the wide variety of volatility forecast users, it is unreasonable to evaluate such forecasts with a single, statistical loss function. Since forecast evaluations based on economic loss functions are undoubtedly more useful in differentiating among volatility models, a forecast evaluation framework that directly incorporates a general class of economic loss functions is proposed. These economic loss functions, implicitly or explicitly based on volatility forecasts, are translated into statistical loss functions and hypotheses tests based on probability forecasts using three key elements: the events to be forecast, a probability scoring rule with which to evaluate the forecasts, and the subsets of these forecasts of interest to the

---

[27] If different subsets are specified, different probability-based hypothesis tests are created, and the forecast evaluation results may differ. For example, for event 2, where the forecasts are generally below 0.5, a redefinition of the subsets as {[0,.125); [.125,.25); [.25,.75); [.75,1]} changes the results and certain forecasts no longer reject the null hypothesis. Complete results of the calibration tests are available upon request.

user. Using the event of interest and the distribution underlying a model's innovation term, volatility forecast are transformed into probability forecasts, and these forecasts are evaluated using statistical criteria tailored to the user's interests.

The empirical exercise in Section IV clearly indicates that the loss function directly influences the forecast evaluation results. Thus, the use of MSE as the loss function of choice for forecast evaluation and model selection must be replaced with a more thoughtful selection. Yet, even under the appropriate loss function, forecast evaluations may provide comparative results that are not statistically significant, as shown by the S test results. An advantage of the proposed framework is that calibration tests provide further analysis of volatility forecasts under the relevant loss function. The exercise also clearly shows that out-of-sample forecast evaluation is required to achieve reasonable model selection results.

This evaluation framework introduces a number of questions that require further analysis. Most immediately, the properties of the transformed volatility forecasts must be more clearly delineated. For example, further research is required to determine whether optimal forecasts under the implicit economic loss functions are available. Secondly, the properties of the QPS and other scoring rules must be examined in light of West (1994); that is, uncertainty due to parameter estimation should be incorporated into this analysis. Other avenues for research are evaluating volatility forecasts from other models (including those without assumed distributions for ε_t) and for other financial time series under different loss functions. A systematic exploration of the theoretical and empirical aspects of the proposed evaluation framework will undoubtedly provide more useful volatility forecasts and volatility model specifications.

References

Akgiray, V., 1989. "Conditional Heteroskedasticity in Time Series of Stock Returns: Evidence and Forecasts," *Journal of Business* 62: 55-80.

Andersen, T.G., 1994. "Volatility*," *Mathematical Finance*, 1994, 4, 75-102.

Baillie, R.T. and Bollerslev, T., 1989. "The Message in Daily Exchange Rates: A Conditional Variance Tale," *Journal of Business and Economic Statistics* 7:297-305.

Baillie, R.T, and Bollerslev, T., 1992. "Prediction in Dynamic Models with Time-Dependent Conditional Variances," *Journal of Econometrics* 52:91-113.

Bollerslev, T., 1986. "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics* 31:307-327.

Bollerslev, T. and Ghysels, E., 1994. "Periodic Autoregressive Conditional Heteroskedasticity," Working Paper #178, Department of Finance, J.L. Kellogg Graduate School of Management, Northwestern University.

Bollerslev, T. and Wooldridge, J.M. (1992), "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances," *Econometric Reviews* 11:143-179.

Bollerslev, T., Engle, R.F. and Nelson, D.B., 1994. "ARCH Models," in Engle, R.F. and McFadden, D., eds. *The Handbook of Econometrics, Volume 4*, Amsterdam: North-Holland, forthcoming.

Brier, G.W., 1950. "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 75:1-3.

Christoffersen, P.F., 1995. "Predicting Uncertainty in the Foreign Exchange Markets," Manuscript, Department of Economics, University of Pennsylvania.

Davidian, M. and Carroll, R.J., 1989. "Variance Function Estimation," *Journal of the American Statistical Association* 82:1079-1091.

Dawid, A.P., 1986. "Probability Forecasting," in Kotz, S. and Johnson, N.L., eds. *The Encyclopedia of Statistical Science, Volume 7*, New York: John Wiley & Sons, Inc.

Diebold, F.X. and Lopez, J.A., 1995a. "Modelling Volatility Dynamics," in Hoover, K., ed., *Macroeconometrics: Developments, Tensions and Prospects*. Amsterdam: Kluwer Academic Press, forthcoming.

Diebold, F.X. and Lopez, J.A., 1995b. "Forecast Evaluation and Combination," in Maddala, G.S. and Rao, C.R., eds., *Handbook of Statistics, Volume 14: Statistical Methods in Finance*. Amsterdam: North-Holland, forthcoming.

Diebold, F.X. and Mariano, R., 1995. "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-264.

Diebold, F.X. and Rudebusch, G.D., 1989. "Scoring the Leading Indicators," *Journal of Business* 62:369-391.

Donaldson, R.G., Kamstra, M., and Kim, H.Y., 1993. "Evaluating Alternative Models for the Conditional Volatility of Stock Returns: Evidence from International Data," Manuscript, Department of Economics, University of British Columbia.

Dunsmuir, W., 1979. "A Central Limit Theorem for Parameter Estimation in Stationary Vector Time Series and its Application to Models for a Signal Observed with Noise," *The Annals of Statistics* 7:490-506.

Engle, R.F., 1982. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica* 50:987-1008.

Engle, R.F., 1993. "Statistical Models for Financial Volatility," *Financial Analysts Journal* 49:72-78.

Engle, R.F. and Gonzalez-Rivera, G., 1991. "Semiparametric ARCH Models," *Journal of Business and Economic Statistics* 9:345-359.

Engle, R.F., Hong, C.-H., Kane, A. and Noh, J., 1993. "Arbitrage Valuation of Variance Forecasts with Simulated Options," in Chance, D.M. and Tripp, R.R., eds., *Advances in Futures and Options Research*. Greenwich, CT: JIA Press.

Engle, R.F., Kane, A. and Noh, J., 1993. "Index-Option Pricing with Stochastic Volatility and the Value of Accurate Variance Forecasts," Working Paper #4519, National Bureau of Economic Research.

Engle, R.F. and Ng, V., 1993. "Measuring and Testing the Impact of News on Volatility," *Journal of Finance* 48:1749-1778.

Fair, R.C., 1993. "Estimating Event Probabilities from Macroeconometric Models Using Stochastic Simulation," in Stock, J.H. and Watson, M.W., eds. *Business Cycles, Indicators and Forecasting*. Chicago: The University of Chicago Press.

Friedman, B.M. and Kuttner, K.N., 1992. "Time-Varying Risk Perceptions and the Pricing

of Risky Assets," *Oxford Economic Papers* 44: 566-598.

Geweke, J.F., 1994. "Bayesian Comparison of Econometric Models," Working Paper #532, Research Department, Federal Reserve Bank of Minneapolis.

Ghysels, E., 1993. "On Scoring Asymmetric Periodic Probability Models of Turning-Point Forecasts," *Journal of Forecasting* 12:227-238.

Granger, C.W.J., White, H. and Kamstra, M., 1989. "Interval Forecasting: An Analysis Based Upon ARCH-Quantile Estimators," *Journal of Econometrics* 40:87-96.

Harvey, A.C., Ruiz, E. and Shepard, N., 1994. "Multivariate Stochastic Variance Models," *Review of Economic Studies* 61:247-264.

Jacquier, E., Polson, N.G. and Rossi, P.E., 1994. "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics* 12:371-389.

Kim, K. and Schmidt, P., 1993. "Unit Root Tests with Conditional Heteroskedasticity," *Journal of Econometrics* 59:287-300.

Kim, S. and Shepard, N., 1994. "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," Manuscript, Nuffield College, Oxford University.

Kroner, K.F., Kneafsey, K.P. and Claessens, S., 1995. "Forecasting Volatility in Commodity Markets," *Journal of Forecasting*, 14, 77-96.

Lahiri, K. and Wang, J.G., 1994. "Predicting Cyclical Turning Points with Leading Index in a Markov Switching Model," *Journal of Forecasting* 13:245-263.

Lee, K.Y., 1991. "Are the GARCH Models Best in Out-of-Sample Performance?," *Economic Letters* 37:305-308.

Lee, K.Y.,1995. "The Time Aggregation Effect on the Predictability of Exchange Rate Volatility," Manuscript, First Economic Research Institute, Seoul, Korea.

Mandelbrot, B., 1963. "The Variation of Certain Speculative Prices," *Journal of Business* 36:394-419.

Murphy, A.H. and Daan, H., 1985. "Forecast Evaluation" in Murphy, A.H. and Katz, R.W., eds., *Probability, Statistics and Decision Making in the Atmospheric Sciences*. Boulder, Colorado: Westview Press.

Noh, J., Engle, R.F., and Kane, A., 1994. "Forecasting Volatility and Option Prices of the

S&P 500 Index," *Journal of Derivatives*, 2*, 17-30.

Pagan, A.R. and Schwert, G.W., 1990. "Alternative Models for Conditional Stock Volatility," *Journal of Econometrics* 45:267-290.

Schwert, G.W., 1989. "Why Does Stock Market Volatility Change over Time?", *Journal of Finance* 44:1115-1154.

Seillier-Moiseiwitsch, F., and Dawid, A.P., 1993. "On Testing the Validity of Sequential Probability Forecasts," *Journal of the American Statistical Association* 88:355-359.

Stockman, A.C., 1987. "Economic Theory and Exchange Rate Forecasts," *International Journal of Forecasting* 3:3-15.

Taylor, S.J., 1986. *Modelling Financial Time Series*. New York: John Wiley & Sons, Ltd.

Taylor, S.J., 1987. "Forecasting the Volatility of Currency Exchange Rates," *International Journal of Forecasting* 3:159-170.

West, K.D. and Cho, D., 1994. "The Predictive Accuracy of Several Models of Exchange Rate Volatility," Technical Working Paper # 152, National Bureau of Economic Research.

West, K.D., 1994. "Asymptotic Inference about Predictive Ability," Manuscript, Department of Economics, University of Wisconsin.

West, K.D., Edison, H.J. and Cho, D., 1993. "A Utility-Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics* 35:23-45.

Winkler, R.L., 1993. "Evaluating Probabilities: Asymmetric Scoring Rules," *Management Science*, 4, 1395-1405.

Table 1. Descriptive Statistics of the First-Differenced Exchange Rate Series

	DM	CD	BP
Conditional Mean Parameters			
μ	-0.0006 (0.0141)	-0.0003 (0.0052)	-0.0131 (0.0150)
θ	---	---	0.065 (0.020)
In-sample Moments of ε_t			
variance	0.500	0.067	0.499
skewness	-0.252	0.117	0.099
excess kurtosis	2.013* (0.098)	4.473* (0.098)	2.747* (0.098)
Q(15)	17.886	20.602	22.262
$Q^2(15)$	206.75*	352.13*	346.17*
Out-of-sample Moments of ε_t			
variance	0.545	0.063	0.471
skewness	-0.032	0.787	0.009
excess kurtosis	1.914* (0.219)	2.293* (0.219)	1.643* (0.219)
Q(15)	17.886	9.042	17.693
$Q^2(15)$	40.01*	139.67*	43.25*

Notes: The three exchange rate series are logged daily spot rates from 1980 to 1991. The in-sample period is 1980-1989 (2508 observations), and the out-of-sample period is 1990-1991 (502 observations). The conditional mean parameters of the first-differenced series are estimated using least squares. Standard errors are listed in parentheses. Excess kurtosis is expressed relative to the standard normal distribution; this statistic is asymptotically distributed $N(0,24/T)$. The portmanteau statistics for up to 15th-order serial correlation, Q(15) and $Q^2(15)$, are for the ε_t series and the ε_t^2 series, respectively. * indicates significance at the 5% level.

Table 2. In-Sample Forecast Evaluation Results for Statistical Loss Functions

Panel A. DM

	MSE	MAE	LL	HMSE	GMLE
homo.	1.0028	0.5551	29.5289	4.0159	0.3063
AR12sq	0.9614	0.5397	29.2085	3.0623	0.2202
AR12ab	0.9647	0.5160	28.6636	4.0287	0.2157
GARCH	0.9707	0.5431	29.1041	3.2275	0.2145
IGARCH	0.9915	0.5731	29.3290	2.9475	0.2213
GARCH-t	0.9725	0.5509	29.2003	3.0671	0.2152
s.v.	0.9736	0.5031	6.9911	5.0330	0.2435
S results:	---	AR12sq GARCH IGARCH GARCH-t	homo. AR12sq	s.v.	homo.

Panel B. CD

	MSE	MAE	LL	HMSE	GMLE
homo.	0.0290	0.0767	74.9056	6.4754	-1.7034
AR12sq	0.0271	0.0722	73.9293	6.8975	-1.9585
AR12ab	0.0274	0.0685	72.8760	6.5748	-1.8737
GARCH	0..0282	0.0753	73.4072	5.4510	-1.8850
IGARCH	0.0284	0.0763	73.4428	5.3429	-1.8848
GARCH-t	0.0279	0.0745	73.3611	5.7192	-1.8839
s.v.	0.0280	0.0747	9.1930	6.3231	-1.8132
S results:	GARCH-t	AR12sq GARCH IGARCH	GARCH IGARCH	---	---

Panel C. BP

	MSE	MAE	LL	HMSE	GMLE
homo.	1.1792	0.5511	8.3560	4.7456	0.3034
AR12sq	1.0987	0.5306	8.1105	3.7935	0.2194
AR12ab	1.1082	0.5074	7.7393	4.5016	0.2161
GARCH	2.2072	1.2415	12.645	0.8810	0.7071
IGARCH	2.4915	1.3748	13.041	0.8229	0.7762
GARCH-t	1.1259	0.5322	7.6856	3.5284	0.1989
s.v.	1.1294	0.4907	7.7143	6.0733	0.2536
S results:	GARCH IGARCH	homo. AR12sq AR12ab GARCH IGARCH GARCH-t	GARCH IGARCH	AR12sq AR12ab GARCH-t s.v.	---

Note: In the top portion of each panel, the columns represent the loss functions, the rows represent the forecasts, and the lowest value in each column is underlined. The bottom portion of each panel lists the model forecasts that *reject* the null hypotheses of more than one of the S tests at the 10% significance level.

Table 3. Out-of-Sample Forecast Evaluation Results for Statistical Loss Functions

Panel A DM

	MSE	MAE	LL	HMSE	GMLE
homo.	1.1634	0.5633	30.586	4.6592	0.3950
AR12sq	1.1585	0.5707	30.410	4.1546	0.3560
AR12ab	1.1516	0.5506	29.921	5.1821	0.3709
GARCH	1.1346	0.5742	30.353	3.4078	0.3252
IGARCH	1.1595	0.6093	30.609	2.8652	0.3297
GARCH-t	1.1373	0.5828	30.444	3.1849	0.3242
s.v.	1.1335	0.5383	6.8523	4.5950	0.3422
S results:	---	---	---	---	---

Panel B CD

	MSE	MAE	LL	HMSE	GMLE
homo.	0.0169	0.0753	75.289	3.7732	-1.7670
AR12sq	0.0155	0.0671	73.817	8.0314	-1.8625
AR12ab	0.0153	0.0625	72.633	7.3283	-1.9391
GARCH	0.0156	0.0668	72.917	6.0474	-1.9634
IGARCH	0.0157	0.0676	72.933	6.0903	-1.9612
GARCH-t	0.0156	0.0665	72.889	6.1965	-1.9576
s.v.	0.0151	0.0678	9.2695	4.1699	-1.9290
S results:	---	s.v.	AR12ab	---	---

Panel C BP

	MSE	MAE	LL	HMSE	GMLE
homo.	0.8239	0.5254	8.9497	3.3158	0.2551
AR12sq	0.8029	0.5086	8.5012	3.2571	0.1852
AR12ab	0.7954	0.4816	8.0976	4.1200	0.1892
GARCH	1.7014	1.1785	14.405	0.7652	0.6740
IGARCH	2.1063	1.3378	15.125	0.7451	0.7636
GARCH-t	0.8793	0.5480	9.2914	5.1983	0.3357
s.v.	0.8118	0.4645	7.6870	5.8899	0.2641
S results:	GARCH IGARCH	AR12sq GARCH IGARCH	GARCH IGARCH	AR12ab s.v.	GARCH IGARCH

Note: In the top portion of each panel, the columns represent the loss functions, the rows represent the forecasts, and the lowest value in each column is underlined. The bottom portion of each panel list the models whose forecasts *reject* the null hypotheses of more than one of the S tests at the 10% significance level.

Table 4. In-Sample and Out-of-Sample Observed Event Frequencies

Number of in-sample observations: 2508
Number of out-of-sample observations: 502

Panel A. DM

	Event 1	Event 2	Event 3	Event 4	Event 5
In-sample	73.8%	6.5%	95.7%	49.4%	11.5%
Out-of-sample	70.7%	6.8%	84.5%	41.2%	15.7%

Panel B. CD

	Event 1	Event 2	Event 3	Event 4	Event 5
In-sample	77.6%	5.0%	91.1%	43.9%	15.1%
Out-of-sample	75.5%	6.8%	81.9%	34.9%	48.0%

Panel C. BP

	Event 1	Event 2	Event 3	Event 4	Event 5
In-sample	72.9%	6.2%	82.3%	40.3%	21.4%
Out-of-sample	74.3%	6.2%	91.2%	48.2%	33.7%

Table 5. In-Sample and Out-of-Sample GSB Results

Panel A. DM

In-sample	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.0032	0.0002	0.0000	0.0003	0.0000
AR12sq	0.0021	0.0001	0.0000	0.0003	0.0000
AR12ab	0.0004	0.0000	0.0001	0.0002	0.0000
GARCH	0.0014	0.0001	0.0000	0.0001	0.0000
IGARCH	0.0026	0.0003	0.0001	0.0002	0.0000
GARCH-t	0.0068	0.0011	0.0019	0.0008	0.0000
s.v.	0.0000	0.0000	0.0002	0.0001	0.0129

Out-of-sample	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.0007	0.0001	0.0002	0.0000	0.1841
AR12sq	0.0005	0.0001	0.0002	0.0000	0.1817
AR12ab	0.0000	0.0000	0.0000	0.0002	0.1816
GARCH	0.0004	0.0002	0.0003	0.0000	0.1814
IGARCH	0.0013	0.0005	0.0012	0.0000	0.1808
GARCH-t	0.0041	0.0012	0.0051	0.0004	0.1805
s.v.	0.0001	0.0000	0.0002	0.0003	0.0202

Table 5. In-Sample and Out-of-Sample GSB Results (continued)

Panel B. CD

In-sample	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.0032	0.0001	0.0002	0.0001	0.0000
AR12sq	0.0013	0.0000	0.0001	0.0002	0.0000
AR12ab	0.0001	0.0000	0.0000	0.0004	0.0000
GARCH	0.0008	0.0001	0.0002	0.0002	0.0000
IGARCH	0.0010	0.0001	0.0003	0.0001	0.0000
GARCH-t	0.0042	0.0005	0.0023	0.0001	0.0000
s.v.	0.0027	0.0007	0.0016	0.0000	0.0219

Out-of-sample	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.0013	0.0000	0.0069	0.0004	0.1788
AR12sq	0.0001	0.0002	0.0024	0.0013	0.1787
AR12ab	0.0005	0.0005	0.0001	0.0030	0.1795
GARCH	0.0000	0.0002	0.0010	0.0022	0.1787
IGARCH	0.0000	0.0001	0.0012	0.0022	0.1786
GARCH-t	0.0009	0.0000	0.0050	0.0006	0.1775
s.v.	0.0005	0.0000	0.0134	0.0020	0.2103

Panel C. BP

In-sample	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.0026	0.0002	0.0009	0.0001	0.0000
AR12sq	0.0010	0.0001	0.0004	0.0000	0.0000
AR12ab	0.0001	0.0000	0.0000	0.0001	0.0000
GARCH	0.0754	0.0138	0.0289	0.0140	0.0000
IGARCH	0.0884	0.0138	0.0746	0.0177	0.0000
GARCH-t	0.0045	0.0006	0.0032	0.0006	0.0000
s.v.	0.0001	0.0000	0.0002	0.0001	0.0000

Out-of-sample	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.0041	0.0002	0.0006	0.0018	0.1024
AR12sq	0.0268	0.0001	0.0003	0.0011	0.1016
AR12ab	0.0003	0.0000	0.0000	0.0006	0.1024
GARCH	0.0812	0.0139	0.0655	0.0243	0.1023
IGARCH	0.0969	0.0142	0.0819	0.0305	0.1042
GARCH-t	0.0057	0.0006	0.0041	0.0040	0.1038
s.v.	0.0000	0.0000	0.0002	0.0005	0.0002

Note: The three panels correspond to the three exchange rate series. Within the panels, the columns represent the specified economic events, and the rows represent the models used to generate the in-sample and out-of-sample probability forecasts. Recall that $GSB \in [0, 2]$.

Table 6. In-sample QPS Results

Panel A. DM

	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.1964	0.0609	0.0416	0.2504	0.0082
AR12sq	0.1894	0.0608	0.0410	0.2497	0.0079
AR12ab	0.1870	0.0608	0.0411	0.2495	0.0077
GARCH	0.1878	0.0612	0.0414	0.2488	0.0077
IGARCH	0.1896	0.0619	0.0423	0.2489	0.0077
GARCH-t	0.1929	0.0618	0.0439	0.2495	0.0079
s.v.	0.1879	0.0608	0.0412	0.2498	0.1120
S results:	---	GARCH-t	GARCH-t	---	GARCH-t

Panel B. CD

	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.1773	0.0475	0.0802	0.2459	0.0129
AR12sq	0.1680	0.0468	0.0826	0.2442	0.0130
AR12ab	0.1652	0.0466	0.0752	0.2444	0.0130
GARCH	0.1672	0.0470	0.0766	0.2443	0.0131
IGARCH	0.1676	0.0470	0.0769	0.2442	0.0131
GARCH-t	0.1699	0.0471	0.0787	0.2440	0.0132
s.v.	0.1710	0.0474	0.0790	0.2443	0.1456
S results:	s.v.	GARCH-t	GARCH-t s.v.	---	GARCH-t

Panel C. BP

	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.1997	0.0585	0.1158	0.2248	0.0087
AR12sq	0.1940	0.0573	0.1101	0.2209	0.0082
AR12ab	0.1921	0.0572	0.1075	0.2196	0.0078
GARCH	0.2699	0.0715	0.1473	0.2440	0.0141
IGARCH	0.2821	0.0715	0.1908	0.2479	0.0150
GARCH-t	0.1964	0.0575	0.1094	0.2261	0.0086
s.v.	0.1925	0.0573	0.1057	0.2184	0.0073
S results:	GARCH IGARCH	GARCH IGARCH	homo. GARCH IGARCH	GARCH IGARCH GARCH-t	---

Note: In the top portion of each panel, the columns represent the specified economic events, the rows represent the forecasts, and the lowest QPS value in each column is underlined. The bottom portion of each panel lists the models whose forecasts *reject* the null hypotheses of more than one of the S tests at the 10% significance level.

Table 7. Out-of-sample QPS Results

Panel A. DM

	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.2077	0.0633	0.1302	0.2430	0.4156
AR12sq	0.2027	0.0628	0.1298	0.2413	0.4096
AR12ab	0.2030	0.0627	0.1308	0.2425	0.4106
GARCH	0.2012	0.0630	0.1289	0.2413	0.4082
IGARCH	0.2032	0.0636	0.1314	0.2418	0.4056
GARCH-t	0.2050	0.0638	0.1336	0.2419	0.4018
s.v.	0.2031	0.0624	0.1292	0.2440	0.1477
S results:	---	---	---	---	---
Z_0 results:	AR12ab GARCH IGARCH s.v.	---	homo. AR12sq AR12ab GARCH s.v.	homo. AR12sq AR12ab GARCH IGARCH GARCH-t s.v.	---

Panel B. CD

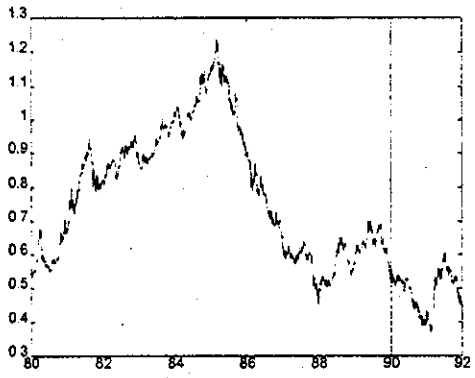
	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.1863	0.0632	0.1616	0.2297	0.5183
AR12sq	0.1715	0.0603	0.1390	0.2261	0.5190
AR12ab	0.1682	0.0603	0.1337	0.2275	0.5202
GARCH	0.1674	0.0605	0.1338	0.2276	0.5203
IGARCH	0.1674	0.0605	0.1342	0.2277	0.5202
GARCH-t	0.1685	0.0607	0.1385	0.2258	0.5181
s.v.	0.1711	0.0606	0.1402	0.2267	0.4605
S results:	---	GARCH-t	---	---	---
Z_0 results:	AR12sq AR12ab GARCH IGARCH GARCH-t s.v.	---	AR12sq AR12ab	homo. AR12sq AR12ab GARCH IGARCH GARCH-t s.v.	---

Panel C. BP

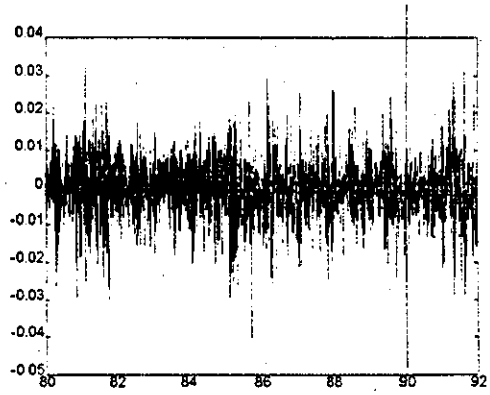
	Event 1	Event 2	Event 3	Event 4	Event 5
homo.	0.1944	0.0580	0.0798	0.2515	0.2849
AR12sq	0.2325	0.0581	0.0794	0.2508	0.2843
AR12ab	0.1843	0.0583	0.0782	0.2497	0.2870
GARCH	0.2664	0.0712	0.1416	0.2738	0.2675
IGARCH	0.2828	0.0715	0.1583	0.2802	0.2679
GARCH-t	0.1987	0.0597	0.0868	0.2500	0.2816
s.v.	0.1854	0.0582	0.0796	0.2482	0.0325
S results:	AR12sq GARCH IGARCH	GARCH IGARCH	GARCH IGARCH	---	---
Z_0 results:	AR12ab s.v.	---	---	---	---

Note: In the top portion of each panel, the columns represent the specified economic events, the rows represent the forecasts, and the lowest QPS value in each column is underlined. The middle position of each panel lists the models whose forecasts *reject* the null hypotheses of more than one of the S tests at the 10% significance level. The bottom portion of each panel lists the models whose forecasts are well calibrated with respect to the forecast subsets {[0,.25); [.25,.50); [.50,.75); [.75,1.0]}; i.e. these forecasts *do not reject* the null of the overall calibration test with respect to the specified subsets.

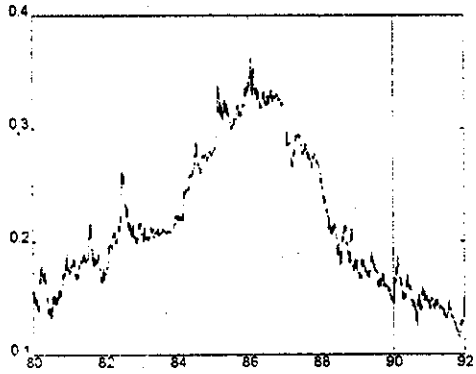
Figure 1. DM/$, 1980-1992



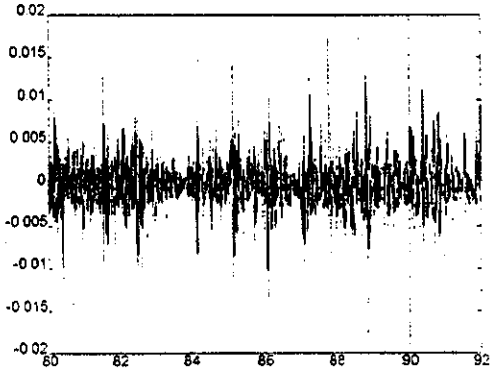
First difference



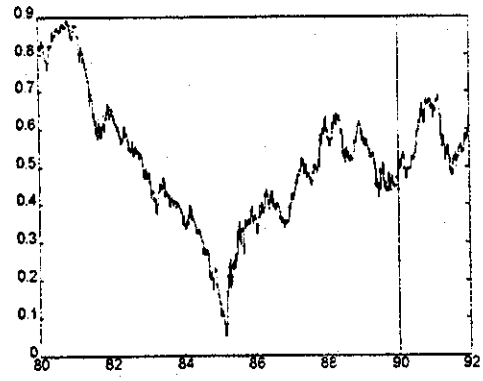
CD/$, 1980-1992



First difference



$/BP, 1980-1992



First difference

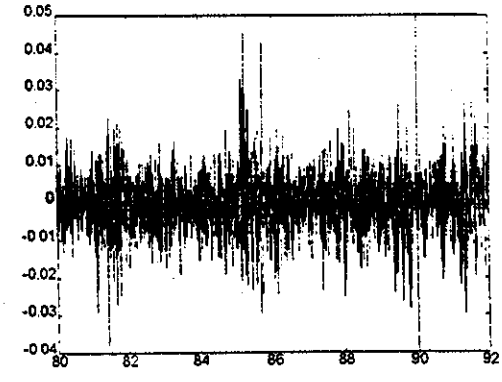
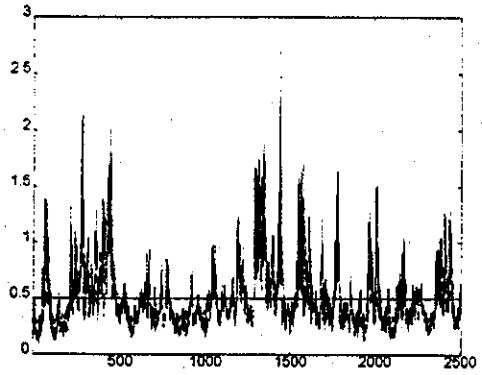
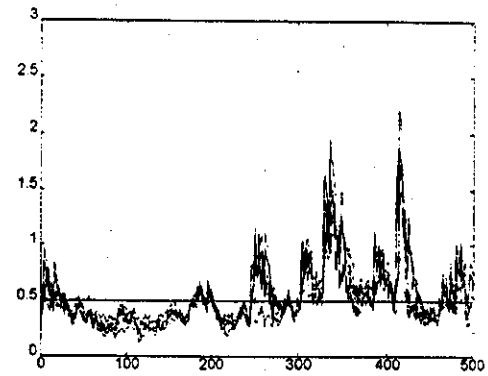


Figure 2.

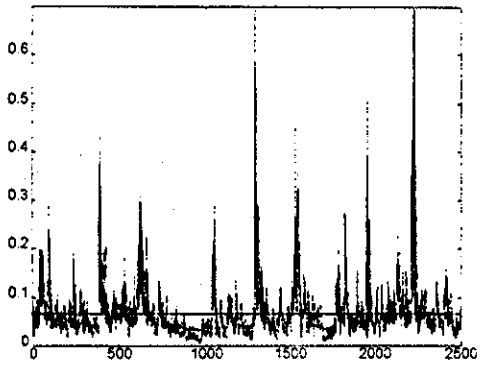
DM: In-sample Volatility Forecasts



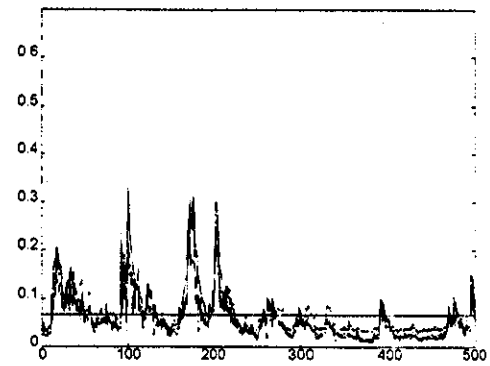
Out-of-Sample Volatility Forecasts



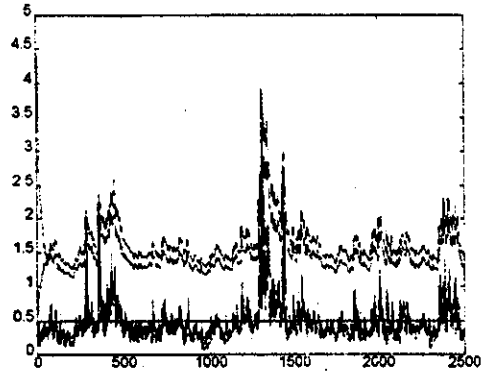
CD: In-sample Volatility Forecasts



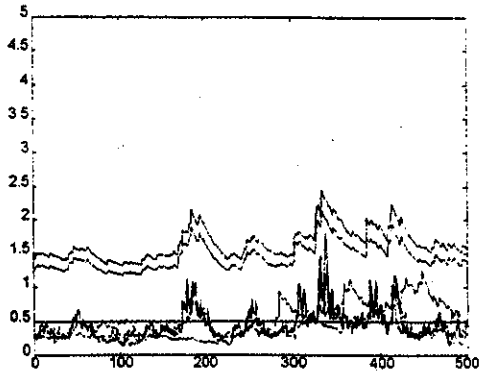
Out-of-Sample Volatility Forecasts



BP: In-sample Volatility Forecasts



Out-of-Sample Volatility Forecasts



APPENDIX: Volatility Model Parameter Estimates

Table A.1 Volatility Model Parameters for the Gaussian Homoskedastic Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\varepsilon_t \mid \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \sigma^2$$

Parameter	DM	CD	BP
μ	-0.0006	-0.0003	-0.0131
	(0.0141)	(0.0052)	(0.0141)
θ	---	---	0.0649*
	---	---	(0.0232)
σ^2	0.4997*	0.0670*	0.4985*
	(0.0223)	(0.0037)	(0.0217)

Note: These parameters values are maximum likelihood estimates. The standard errors in parentheses are the robust standard errors of Bollerslev and Wooldridge (1992). * indicates significance at the 5% level.

Table A.2. Volatility Model Parameters for the AR12sq Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\varepsilon_t \mid \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \sum_{i=1}^{12} \alpha_i \varepsilon_{t-i}^2 .$$

Parameter	DM	CD	BP
μ	-0.0006	-0.0003	0.0123
	(0.0141)	(0.0052)	(0.0150)
θ	---	---	0.0652*
	---	---	(0.0202)
ω	0.2596*	0.0324*	0.2244*
	(0.0318)	(0.0045)	(0.0316)
α_1	0.0272	0.1532*	0.0207
	(0.0200)	(0.0200)	(0.0200)
α_2	0.0289	0.0431*	0.0320
	(0.0200)	(0.0202)	(0.0200)
α_3	0.0510*	0.0946*	0.0357
	(0.0200)	(0.0202)	(0.0199)
α_4	0.0473*	0.0226	0.1013*
	(0.0200)	(0.0203)	(0.0199)
α_5	0.0675*	0.0407*	0.0957*
	(0.0200)	(0.0203)	(0.0200)
α_6	0.0867*	0.0293	0.0600*
	(0.0201)	(0.0203)	(0.0201)
α_7	0.0222	-0.0253	0.0313
	(0.0201)	(0.0203)	(0.0201)
α_8	0.0311	0.0242	0.0219
	(0.0200)	(0.0203)	(0.0200)
α_9	0.0159	0.0434*	-0.0020
	(0.0200)	(0.0203)	(0.0199)
α_{10}	0.0512*	0.0066	0.0683*
	(0.0200)	(0.0202)	(0.0199)
α_{11}	0.0494*	0.0587*	0.0654*
	(0.0200)	(0.0202)	(0.0200)
α_{12}	0.0047	0.0268	0.0208
	(0.0200)	(0.0200)	(0.0200)

Note: The conditional mean parameters are estimated using least squares. The conditional variance parameters are estimated using least squares on the squared residuals. Standard errors are presented in parentheses. * indicates significance at the 10% level.

Table A.3. Volatility Model Parameters for the AR12ab Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t \qquad \varepsilon_t \mid \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \frac{\pi}{2}\left(\omega + \sum_{i=1}^{12} \alpha_i \left| \varepsilon_{t-i} \right| \right)$$

Parameter	DM	CD	BP
μ	-0.0006	-0.0003	0.0123
	(0.0141)	(0.0052)	(0.0150)
θ	---	---	0.0652*
	---	---	(0.0202)
ω	0.2347*	0.0719*	0.2228*
	(0.0258)	(0.0083)	(0.0254)
α_1	0.0150	0.1310*	0.0399
	(0.0200)	(0.0200)	(0.0200)
α_2	0.0450*	0.0647*	0.0417*
	(0.0200)	(0.0201)	(0.0200)
α_3	0.0708*	0.0962*	0.0564*
	(0.0200)	(0.0202)	(0.0200)
α_4	0.0546*	0.0441*	0.0427*
	(0.0200)	(0.0203)	(0.0200)
α_5	0.0840*	0.0799*	0.1075*
	(0.0201)	(0.0203)	(0.0200)
α_6	0.0619*	0.0655*	0.0495*
	(0.0201)	(0.0203)	(0.0201)
α_7	0.0690*	-0.0201	0.0399
	(0.0201)	(0.0203)	(0.0201)
α_8	0.0404*	0.0408*	0.0295
	(0.0201)	(0.0203)	(0.0200)
α_9	0.0225	0.0030	0.0425*
	(0.0201)	(0.0203)	(0.0200)
α_{10}	0.0585*	0.0188	0.0880*
	(0.0200)	(0.0202)	(0.0200)
α_{11}	0.0221	0.0533*	0.0175
	(0.0200)	(0.0201)	(0.0200)
α_{12}	0.0083	0.0367	0.0177
	(0.0200)	(0.0200)	(0.0200)

Note: The conditional mean parameters are estimated using least squares. The conditional variance parameters are estimated using least squares on the absolute value of the residuals. Standard errors are presented in parentheses. * indicates significance at the 10% level.

Table A.4. Volatility Model Parameters for the Gaussian GARCH(1,1) Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\varepsilon_t \mid \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}$$

Parameter	DM	CD	BP
μ	0.0127	0.0022	-0.0189
	(0.0128)	(0.0042)	(0.0117)
θ	---	---	0.0618*
	---	---	(0.0193)
ω	0.0153*	0.0015*	0.0064*
	(0.0031)	(0.0002)	(0.0027)
α	0.0859*	0.1503*	0.0444*
	(0.0300)	(0.0395)	(0.0220)
β	0.8863*	0.8425*	0.9429*
	(0.0873)	(0.0729)	(0.0895)

Note: These parameters values are maximum likelihood estimates. The standard errors in parentheses are the robust standard errors of Bollerslev and Wooldridge (1992). * indicates significance at the 5% level.

Table A.5. Volatility Model Parameters for the IGARCH(1,1) Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\varepsilon_t \mid \Omega_{t-1} \sim N(0, h_t)$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + (1 - \alpha) h_{t-1}$$

Parameter	DM	CD	BP
μ	0.0145	0.0023	-0.0208
	(0.0129)	(0.0042)	(0.0120)
θ	---	---	0.0605*
	---	---	(0.0187)
ω	0.0072	0.0013	0.0026
	(0.0043)	(0.0014)	(0.0035)
α	0.1030*	0.1557*	0.0515*
	(0.0282)	(0.0396)	(0.0210)

Note: These parameters values are maximum likelihood estimates. The standard errors in parentheses are the robust standard errors of Bollerslev and Wooldridge (1992). * represents significance at the 5% level.

Table A.6 Volatility Model Parameters for the GARCH(1,1)-t Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\varepsilon_t \mid \Omega_{t-1} \sim t(0, h_t, v)$$

$$h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}$$

Parameter	DM	CD	BP
μ	0.0196	-0.0013	-0.0131
	(0.0121)	(0.0039)	(0.0121)
θ	---	---	0.0379*
	---	---	(0.0058)
ω	0.0129	0.0011	0.0069
	(0.0325)	(0.0025)	(0.0055)
α	0.0837*	0.1240*	0.0473*
	(0.0309)	(0.0277)	(0.0039)
β	0.8966*	0.8683*	0.9396*
	(0.0862)	(0.0670)	(0.0138)
v	5.7117	6.4091	6.6198
	(0.6948)	(0.7459)	(0.8670)

Note: These parameters values are maximum likelihood estimates. Except for v, the standard errors in parentheses are the robust standard errors of Bollerslev and Wooldridge (1992). The standard errors for v are derived from the value of the numerical Hessian at the estimated parameter values. * represents significance at the 5% level.

Table A.7. Volatility Model Parameters for the Stochastic Volatility Model

$$\Delta y_t = \mu + \theta \varepsilon_{t-1} + \varepsilon_t$$

$$\varepsilon_t = \exp(\alpha_t / 2) v_t, \quad v_t \sim N(0, 1)$$

$$\alpha_t = \phi \alpha_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2)$$

$$v_t \perp \eta_t$$

Measurement equation: $\log \varepsilon_t^2 = \omega + \alpha_t + \xi_t$

Transition equation: $\alpha_t = \phi \alpha_{t-1} + \eta_t$

Parameter	DM	CD	BP
μ	-0.0006	-0.0003	0.0123
	(0.0141)	(0.0052)	(0.0150)
θ	---	---	0.0652*
	---	---	(0.0202)
ϕ	0.9728*	0.9158*	0.9726*
	(0.0098)	(0.058)	(0.0113)
ω	-0.0275	-0.2663	-0.0293
σ_η^2	0.0234*	0.1648*	0.0258
	(0.0092)	(0.0618)	(0.0119)

Note: The conditional mean parameters are estimated using least squares, and the standard errors are presented in parentheses. * indicates significance at the 10% level. The conditional variance parameters are estimated using the Kalman filter and quasi-maximum likelihood methods as proposed in Harvey *et al.* (1994). These standard errors are estimated using results from Dunsmuir (1979).