Evaluating the Predictive Accuracy of Volatility Models

Jose A. Lopez

Economic Research Department
Federal Reserve Bank of San Francisco
101 Market Street
San Francisco, CA 94705-1579
(415) 977-3894
jose.a.lopez@sf.frb.org

Draft Date: June 30, 1999

ABSTRACT:

Statistical loss functions that generally lack economic content are commonly used for evaluating financial volatility forecasts. In this paper, an evaluation framework based on loss functions tailored to a user's economic interests is proposed. According to these interests, the user specifies the economic events to be forecast, the criterion with which to evaluate these forecasts, and the subsets of the forecasts of particular interest. The volatility forecasts from a model are then transformed into probability forecasts of the relevant events and evaluated using the specified criteria (i.e., a probability scoring rule and calibration tests). An empirical example using exchange rate data illustrates the framework and confirms that the choice of loss function directly affects the forecast evaluation results.

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

Acknowledgments: The views expressed here are those of the author and do not necessarily reflect those of the Federal Reserve Bank of San Francisco or the Federal Reserve System. I thank Frank Diebold, Tim Bollerslev, Peter Christoffersen, Roberto Mariano, Lee Ohanian, Esther Ruiz and Ken West for their comments as well as 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 and the Federal Reserve Bank of New York.

Author biography:

Jose A. Lopez is currently an economist in the Research Department at the Federal Reserve Bank of San Francisco. He holds a Ph.D. in economics from the University of Pennsylvania. His current research focuses on financial volatility forecasting and its applications in the risk measurement and management systems of financial institutions.

Evaluating the Predictive Accuracy of Volatility Models

ABSTRACT:

Statistical loss functions that generally lack economic content are commonly used for evaluating financial volatility forecasts. In this paper, an evaluation framework based on loss functions tailored to a user's economic interests is proposed. According to these interests, the user specifies the economic events to be forecast, the criterion with which to evaluate these forecasts, and the subsets of the forecasts of particular interest. The volatility forecasts from a model are then transformed into probability forecasts of the relevant events and evaluated using the specified criteria (i.e., a probability scoring rule and calibration tests). An empirical example using exchange rate data illustrates the framework and confirms that the choice of loss function directly affects the forecast evaluation results.

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

"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 twenty years. GARCH models, pioneered by Engle (1982) and Bollerslev (1986), are the volatility models most commonly used, although numerous alternatives exist. Specifically, stochastic volatility models, which arose from the theoretical finance literature, are increasingly employed in empirical research. For recent surveys of volatility models, see Bollerslev, Engle and Nelson (1994) as well as Diebold and Lopez (1995).

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. 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. Several such loss functions based on specific economic questions have been proposed in the literature. For example, West, Edison and Cho (1993) propose a loss function based on the expected utility from an investment decision requiring volatility forecasts. Engle *et al.* (1993) propose an economic loss function based on trading profits earned in a simulated options market; see also Engle, Kane and Noh (1996) and Noh, Engle and Kane (1994).

However, since specific economic loss functions are frequently unavailable, volatility forecast evaluation is typically conducted by minimizing a statistical loss function, such as mean squared error (MSE). Numerous studies have used MSE to evaluate the performance of volatility models; for example, see Taylor (1987); Akgiray (1989); Dimson and Marsh (1990); Pagan and Schwert (1990); Lee (1991); West and Cho (1994); Kroner, Kneafsey and Claessens (1995); Bollerslev and Ghysels (1996); and Brailsford and Faff (1996). Such forecast evaluations are problematic since they almost exclusively use squared asset returns as a proxy for the latent volatility process. As discussed by Andersen and Bollerslev (1997), squared returns are noisy estimators of the actual variance dynamics and will thus limit the inference available regarding the forecast accuracy.

In this paper, a forecast evaluation framework based on an alternative set of statistical loss functions is proposed. Rather than selecting a simple statistical criterion as the appropriate evaluation measure, this framework permits the forecast user to tailor their choice of criterion according to their economic interests. Specifically, the framework is based on evaluating probability forecasts generated from an underlying volatility model. Probability forecasts are the object of interest for a large, general class of loss functions based on decision-theoretic problems, such as deciding whether to take a specific action; see Granger and Pesaran (1996) as well as Zellner and Hong (1991) for further discussion.

The user tailors the forecast evaluation by specifying three key elements of this framework. The first element is the economic events of interest, such as a central bank’s interest in whether an exchange rate will leave a specified target zone. Volatility forecasts are readily transformed into such probability forecasts using the underlying model's assumed (or estimated) distribution for the innovation term. Once generated, the forecasts are evaluated using scoring rules, which measure the accuracy of probability forecasts with respect to whether the forecasted events occur. Forecast users can select the scoring rule best suited to their particular interests. The most common scoring rule (and the one used in this paper) is the quadratic probability score, which is the analog of MSE for probability forecasts and an explicit measure of forecast accuracy. Furthermore, forecast users can employ calibration tests to evaluate subsets of the probability forecasts that are of particular interest, such as forecasts that lie above a certain decision-theoretic trigger point (say, 75 percent). 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 subsets of the unit interval. The specification of these three elements forms an evaluation framework that, although based on statistical loss functions, can be tailored to provide economically meaningful inference for a wide range of decision-theoretic problems.

The paper is structured as follows. In Section II, the loss functions 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 -- the events to be forecast, the scoring rule used to evaluate the probability forecasts, and the subsets of the forecasts of particular interest. Section IV presents an empirical example using daily exchange rates to illustrate the evaluation procedure. Four different sets of probability forecasts are used to evaluate the forecasts from nine volatility models. The example provides three insights into volatility forecast evaluation. First, the evaluations are sensitive to the loss function chosen, and thus forecast users should consider their choice carefully. Second, the disparity between in-sample and out-of-sample evaluation results clearly indicate that in-sample diagnostic tests must be supplemented by out-of-sample analysis. Third, although not capable of providing definitive answers on which forecasts minimize the relevant loss functions, the proposed forecast evaluation framework provides useful and economically meaningful information for reducing the choice set. Section V concludes and suggests directions for further research.

II. Previous Evaluations of Volatility Forecasts

Volatility models are generally expressed as

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t = \sqrt{h_t}\, z_t, \qquad z_t \sim \text{ i.i.d. } D(0, 1),$$

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; thus, $\varepsilon_t \mid \Omega_{t-1} \sim D\left(0, h_t\right)$. Variance dynamics are introduced into the model via the specification of h_t; for example, the GARCH(1,1) model sets $h_t = \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}$. Generally, the parameters of the 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*. (Note that, in a convenient abuse of terminology, the in-sample, fitted conditional variances are frequently referred to as in-sample forecasts).

Several in-sample diagnostics, such as analysis of standardized residuals, conditional moment tests as per Wooldridge (1990) or the news impact curve proposed by Engle and Ng (1993), are available for comparing volatility model specifications. However, out-of-sample forecast accuracy provides an alternative and potentially more useful comparison; i.e., a model that generates accurate forecasts must be providing a reasonable approximation of the underlying data generating process. In the literature, volatility forecasts have been evaluated using economic and statistical loss functions. Economic loss functions, such as the utility-based loss function proposed by West *et al.* (1993) or the profit-based loss function proposed by Engle *et al.* (1993), provide the most meaningful forecast evaluations since they directly incorporate the user's decision structure into the evaluation.

However, such specific economic loss functions are rarely available, and purely statistical loss functions, such as the mean-squared error (MSE) criterion, are more commonly used to evaluate both in-sample and out-of-sample volatility forecasts. Focusing on out-of-sample forecasts, MSE is the average squared difference between the actual conditional variance h_{T+t} and the corresponding volatility forecast $\hat{h}_{T+t}$. However, since h_{T+t} is not directly observable, the squared innovation ε^2_{T+t} is used as a proxy. Thus, $\text{MSE} = \frac{1}{T^*}\sum_{t=1}^{T^*}\left(\varepsilon^2_{T+t} - \hat{h}_{T+t}\right)^2$, 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. (Note that this analysis can be extended to k-step-ahead forecasts.)

Although MSE could in some instances be the appropriate loss function for evaluating conditional variance forecasts, the use of the ε^2_{T+t} proxy is problematic. Specifically, even though ε^2_{T+t} is an unbiased estimator of h_{T+t}, it is generally an imprecise or "noisy" estimator due to its asymmetric distribution. For example, if $z_{T+t} \sim N(0, 1)$, then $E\left[\varepsilon^2_{T+t} \mid \Omega_{T+t-1}\right] = h_{T+t}E\left[z^2_{T+t} \mid \Omega_{T+t-1}\right] = h_{T+t}$ since $z^2_{T+t} \sim \chi^2_{(1)}$. 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,

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

that is, even if one is willing to accept a proxy that is up to 50% different from h_{T+t}, ε^2_{T+t} would fulfill this condition only 25% of the time. (See Andersen and Bollerslev (1997) for a discussion of alternative proxies based on high-frequency data.) A further difficulty with MSE evaluation in a heteroskedastic environment, as noted by Bollerslev *et al.* (1994), is that the symmetric nature of this loss function does not sufficiently penalize nonpositive variance forecasts.

Several alternative statistical loss functions are proposed in the literature, although the same criticisms cited above would apply to them; for example, mean absolute error (MAE), $\text{MAE} = \frac{1}{T^*}\sum_{t=1}^{T^*}\left|\varepsilon^2_{T+t} - \hat{h}_{T+t}\right|$. Two loss functions that penalize volatility forecasts asymmetrically are the logarithmic loss (LL) function employed by Pagan and Schwert (1990),

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

which penalizes inaccurate variance forecasts more heavily when ε^2_{T+t} is low, and the heteroskedasticity-adjusted MSE (HMSE) of Bollerslev and Ghysels (1996),

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

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

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

Although these functions may coincide with a specific forecast user's economic loss function, they are generally devoid of economic content. Hence, economic inference based on such forecast evaluations is limited. Below, a forecast evaluation framework based on probability forecasts is proposed. This framework is also based on statistical loss functions; however, analyzing probability forecasts generated from volatility forecasts permits the evaluation to be tailored more closely to the user's economic interests.

III. Generating and Evaluating Probability Forecasts

The proposed framework permits the user to tailor the forecast evaluation to their particular interests in two stages. The first stage consists of transforming volatility forecasts into probability forecasts of events of interest. (Note that this type of evaluation defines a volatility model as a combination of both the variance dynamics specified by h_t and the distribution of the standardized residuals specified by D.) This focus on probability forecasts can be linked to a large, general class of economic loss functions commonly used in decision-theoretic problems. As discussed by Granger and Pesaran (1996), probability forecasts are the key input to loss functions that gauge the economic value of taking an action, such as whether or not to purchase an asset. In the second stage, the probability forecasts are evaluated using the statistical criteria that most closely correspond to the user's interests. The first part of this section describes how probability forecasts are generated from volatility forecasts, and the remainder describes the evaluation criteria employed.

3a. Transforming Volatility Forecasts into Probability Forecasts

Given that $\varepsilon_t \mid \Omega_{t-1} \sim D\left(0, h_t\right)$, volatility forecasts can be easily transformed into probability forecasts. The distribution D can be assumed or estimated as per Engle and Gonzalez-Rivera (1991), and one-step-ahead probability forecasts can be easily generated by integrating over this distribution. The appropriate range of integration will depend on the economic event of interest; i.e., the user can tailor the forecast evaluation to the region of $D\left(0, h_t\right)$ of particular interest. This transformation of the volatility forecasts permits forecast users with different (but unspecified) economic loss functions to introduce their particular interests into the forecast evaluation, even though they will all use the same statistical tools.

Multi-step-ahead volatility forecasts are not as easily transformed into probability forecasts. As shown by Baillie and Bollerslev (1992), variance dynamics cause the distribution of ε_{T+k} (k>1) to be a nontrivial function of Ω_T. However, a number of methods are available to address this difficulty; for example, Baillie and Bollerslev (1992) use the Cornish-Fisher approximation method to construct k-step-ahead confidence intervals for a GARCH(1,1) process. Alternatively, simulation methods, as used by Geweke (1989) as well as Bollerslev and Mikkelsen (1996), can be used to generate the desired k-step-ahead probability forecasts. Since the forecast evaluation tools described below are appropriate for transformed volatility forecasts of all horizons, the following discussion will focus on one-step-ahead forecasts to simplify both notation and exposition.

The probability forecast notation is as follows. Volatility models are fit to the in-sample observations $\left\{y_t\right\}_{t=1}^{T}$, and P_t is the in-sample probability forecast for time t based on the parameters estimated over the entire sample and the necessary elements of Ω_T. Out-of-sample, P_{T+t} for t = 1,...,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 equaling one if the event occurs and zero otherwise, are scalar variables.

In the following discussion and empirical exercise, two general types of binomial events, innovation and level events, will be examined. 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 hedging strategies with respect to ε_{T+t}. For such users, the economic event of interest is $\varepsilon_{T+t} \in \left[L_{\varepsilon,T+t}, \ U_{\varepsilon,T+t}\right]$, and the associated probability forecast given the one-step-ahead forecast $\hat{h}_{T+t}$ is

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

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 tailor the volatility forecast evaluation by specifying the appropriate $\left[L_{\varepsilon,T+t}, \ U_{\varepsilon,T+t}\right]$ interval.

The second type of economic events of interest are based on the behavior of the level term y_{T+t}. Such events would be of interest, for example, to a central bank forecasting whether an exchange rate will remain within a target zone. In such cases, the event of interest is $y_{T+t} \in \left[L_{y,T+t}, \ U_{y,T+t}\right]$, and the probability forecast is

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

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

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. For the central bank target zone example, $\left[L_{y,T+t}, \ U_{y,T+t}\right]$ would equal the constant interval [L,U] for all t. Granger, White and Kamstra (1989) examine interval forecasts based on y_{T+t}, in which case the endpoints $\left\{\left[L_{y,\,T+t},\ U_{y,\,T+t}\right]\right\}_{t=1}^{T^*}$ are set to generate a constant α% confidence interval around the corresponding conditional mean forecasts; see Christoffersen (1996) for further discussion on evaluating interval forecasts.

Once the desired probability forecasts are generated, forecast users can proceed to evaluate them using probability scoring rules and calibration tests. These two forecast evaluation criteria, as well as the Diebold-Mariano (1995) tests of comparative predictive accuracy, are described in the following subsections; see Diebold and Lopez (1996) 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). Scoring rules measure the statistical accuracy of the forecasted probabilities. (Note that a scoring rule should be "proper"; i.e., it should not cause the forecaster to modify their actual forecasts in order to improve their expected score.) Just as with the statistical loss functions described in Section II, a variety of proper scoring rules are available; see Murphy and Daan (1985) as well as Winkler (1993) for further discussion. However, a key difference between these two classes of statistical loss functions with respect to volatility forecasts is that scoring rules are based on comparing a model's forecasts to observable events, not a proxy for the true, unobservable variance.

In practice, the forecast user should select the proper scoring rule that addresses their particular interests most directly. In the following discussion and empirical exercise, the quadratic probability score (QPS), the most common scoring rule, is used. The QPS, originally developed by Brier (1950), is the analog of MSE for probability forecasts and thus implies a quadratic loss function. 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). This type of statistical loss function would be most relevant to economic agents facing decision-theoretic problems that explicitly value the accuracy of the probability forecast inputs.

3c. Comparing the Predictive Accuracy of Probability Forecasts

Scoring rules measure the accuracy of probability forecasts; hence, if the QPS for $\left\{P_{A,\, T+t}\right\}_{t=1}^{T^*}$ is closer to zero than the QPS for $\left\{P_{B,\, T+t}\right\}_{t=1}^{T^*}$, then the forecasts from model A are more accurate than those from model B. However, whether this outcome is statistically significant or an artifact of the sample at hand is unclear. Diebold and Mariano (1995) propose several tests for determining whether the expected losses (under a general loss function) induced by two sets of *point* forecasts are statistically different. These tests are readily generalized to probability forecasts.

The null hypothesis under the 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, $E\left[d_{T+t}\right] = E\left[g\left(P_{A,\,T+t},\ R_{T+t}\right) - g\left(P_{B,\,T+t},\ R_{T+t}\right)\right] = 0$. 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. (Note that to test this hypothesis, $\left\{d_{T+t}\right\}_{t=1}^{T^*}$ must be covariance stationary, a condition determined empirically.) Several statistics, both asymptotic and finite-sample, are proposed for testing this null hypothesis. In this paper, the asymptotic mean statistic is used and defined as

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

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.

3d. Calibration Tests of Probability Forecasts

The calibration of probability forecasts is defined as the degree of equivalence between an event's observed and predicted frequencies of occurrence within subsets of the unit interval. For example, a forecaster is providing perfectly calibrated rain forecasts if it rains on 10% of the days for which a 10% chance of rain is forecast. A simple measure of calibration is the calibration plot, which graphs Pr(event occurs | π_j) against π_j, where π_j is the midpoint of one of the J mutually exclusive and collectively exhaustive subsets of the unit interval specified by the user. The degree to which this graph lies on the 45° line is a visual measure of calibration. Seillier-Moiseiwitsch and Dawid (1993) present test statistics that formalize this analysis.

The null hypothesis of these tests is that the predicted frequency of occurrence for a specified binomial event equals the observed frequency of occurrence and hence the forecasts are well calibrated. The relevant test statistics are constructed by dividing the out-of-sample probability forecasts into the J subsets of the unit interval relevant to the forecast user; for example, if the deciles {[0, 0.1); [0.1, 0.2); ...; [0.9, 1]} are of interest as in Diebold and

Rudebusch (1989), then J = 10. Let π_j denote the midpoint of subset j, T_j^* the number of forecasts in subset j, and R_j the number of observed events paired with the forecasts in subset j. The test statistics are the subset j calibration statistics,

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

and the overall calibration statistic,

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

where $R_0 = \sum_{j=1}^{J} R_j$, $e_0 = \sum_{j=1}^{J} e_j$ and $w_0 = \sum_{j=1}^{J} w_j$. 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. 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 tests and this flexible specification of the J subsets of interest provide forecast users with an additional statistical tool with which to tailor their evaluations of the transformed volatility forecasts.

In summary, a statistical framework for volatility forecast evaluation that permits the user to tailor the evaluation to their economic interests is proposed. As discussed by Murphy (1973), purely statistical and purely economic evaluations of forecasts represent the end points of a continuum. By permitting the user to specify the event to be forecast, a probability scoring rule and a partitioning of the unit interval for calibration tests, this framework permits evaluations that incorporate elements of both types. In the following section, an empirical application illustrating this framework for one-step-ahead volatility forecasts is presented. Evaluations of k-step-ahead forecasts using this framework are reserved for future research.

IV. Evaluating Exchange Rate Volatility Forecasts

To illustrate the proposed evaluation framework, one-step-ahead volatility forecasts of several foreign exchange rates are evaluated with respect to four different economic events of interest. Within this empirical exercise, no one set of forecasts is found to minimize all of the probability-based loss functions. In fact, not only does the minimizing set of forecasts vary across exchange rates and events, 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 function employed and the usefulness of out-of-sample forecasting results for model comparison and specification.

4a. Exchange Rate Data

The four exchange rates examined are the logged daily Canadian dollar (CD), Deutschemark (DM) and Japanese yen (YN) 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 1995. The in-sample period used for parameter estimation is 1980-1993 (3516 observations), and the out-of-sample period is 1994-1995 (499 observations).

Given established unit root results for exchange rates (as per Baillie and Bollerslev (1989) and others) and after standard unit root testing, the log exchange rates are modeled as I(1) processes; that is,

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

where $y_t \equiv 100 \log(s_t)$ and s_t is one of the spot exchange rates. The moving-average term is included due to the presence of a small, but significant, first-order serial correlation in Δy_t. Table 1 presents the in-sample, least squares estimates of these conditional mean parameters, higher-order moments of the in-sample and out-of-sample ε_t series and the portmanteau statistics for up to twentieth-order serial correlation in the ε_t and ε_t^2 series. The kurtosis estimates and the $Q^2(20)$ statistics indicate the presence of conditional heteroskedasticity in all four series.

4b. Volatility Models and Forecasts

To address the presence of variance dynamics in these series, three categories of volatility models are used: simple models, GARCH models and stochastic volatility models. These models are estimated over the in-sample period, and the parameter estimates are used to generate volatility forecasts over the out-of-sample period. (The parameter estimates and relevant standard errors are reported in the Appendix.)

Three simple volatility models are examined. The first model is the Gaussian homoskedastic model (homo.), which ignores variance dynamics and assumes $h_t = \sigma^2$. Although, as shown in Table 2, the squared standardized residuals from this model still display significant conditional heteroskedasticity, it is included in the evaluation exercise as a benchmark. As in West and Cho (1994), two autoregressive models of ε_t are estimated using ordinary least squares and an assumed conditionally normal distribution. The autoregressive ε_t^2 model assuming a lag-order of 10 specifies $h_t = \omega + \sum_{i=1}^{10} \alpha_i \varepsilon_{t-i}^2$ and is denoted as AR10sq.

(Note that 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 (AR10ab), as developed by Davidian and Carroll (1987) and Schwert (1989), is based on $\left|\varepsilon_t\right|$ and is specified as

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

From the second category, five models are specified and estimated via maximum likelihood methods. The first three models share the same GARCH(1,1) specification (i.e., $h_t = \omega + \alpha\, \varepsilon_t^2 + \beta\, h_{t-1}$), but use different distributional assumptions. Specifically, the three assumed, symmetric forms of D are the normal distribution, the t-distribution and the generalized exponential distribution (GED); all of which are special cases of the generalized t-distribution discussed in McDonald and Newey (1988). Both the t- and GED-distributions permit fatter tails than the normal distribution, an empirical regularity often found in financial time series; see Figure 1 for a comparison of these distributions based on the estimated shape parameters for the BP series. These models are denoted as GARCH, GARCH-t and GARCH-g, respectively.

The fourth and fifth models are derived from the calibrated, exponential-smoothing model currently used in the volatility forecasting methodology proposed in the J.P. Morgan's Riskmetrics^{TM} system; see J.P. Morgan (1996). The assumed variance dynamics are

$$h_t = (1-\lambda)\sum_{i=1}^{\infty} \lambda^i \varepsilon_{t-i}^2 = (1-\lambda)\,\varepsilon_{t-1}^2 + \lambda h_{t-1};$$

note that this calibrated model is equivalent to the IGARCH model of Engle and Bollerslev (1986) with the ω parameter set to zero. The λ parameter is calibrated according to the degree of persistence desired. For this exercise, the commonly-used value of λ=0.94 is specified, and the conditional mean parameters are those estimated using the homoskedastic model. The two exponential-smoothing models used in this exercise differ only in their assumed distribution D; the e.smooth-N model uses the standard normal, and the e.smooth-t model uses the t-distribution with the same shape parameter as estimated for the GARCH-t model.

The final model is the AR(1) stochastic volatility model (s.v.) used by Harvey, Ruiz and Shephard (1994); that is, after removing the conditional mean based on least squares parameter estimates,

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

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

where $\nu_t \perp \eta_t$. Thus, the innovation term is subject to two independent shocks. Estimation of the model is conducted using the Kalman filter on the measurement equation $\log \varepsilon_t^2 = \alpha_t + \log \nu_t^2$ and the transition equation $\alpha_t = \varphi \alpha_{t-1} + \eta_t$. To use quasi-maximum likelihood methods, the measurement equation is rewritten as $\log \varepsilon_t^2 = \omega + \alpha_t + \xi_t$, where $\omega = E\left[\log \nu_t^2\right]$ and ξ_t is assumed to be distributed $N\left(0, \pi^2/2\right)$. (Note that alternative estimation techniques have been proposed; see Jacquier, Polson and Rossi, 1994). The log-likelihood function to be maximized is

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

where a_t and G_t are the Kalman filter estimates of α_t and its conditional variance, respectively, and $\theta = \left[\varphi, \sigma_\eta^2\right]$.

One-step-ahead volatility forecasts are easily obtained from the simple and GARCH(1,1) models, and the corresponding probability forecasts are derived via integration. One-step-ahead volatility forecasts from the s.v. model are obtained using the Kalman filter, but converting them into probability forecasts is not as straightforward. The standardized innovation term $z_t = e^{\eta_t/2}\nu_t$ is the product of 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 is both difficult and computationally intensive. To alleviate this problem, this distribution is empirically approximated by simulating 10,000 draws from it and creating 0.5% quantiles. The probability that $z_t \in \left(l_t, u_t\right)$ is approximated as $P_t = \hat{F}\left(u_t\right) - \hat{F}\left(l_t\right)$, where $\hat{F}$ is the empirical cumulative distribution function of z_t. (The magnitude of this approximation error is small and can be made arbitrarily small by increasing the number of simulated draws.)

Standard diagnostics of the in-sample and out-of-sample standardized residuals are presented in Table 2. In-sample, almost all of the volatility models, except for the homo. and s.v. models, consistently remove the serial correlation in the squared standardized residuals in all four series. Out-of-sample, only the homo. model fails to consistently do so. These results indicate that all of the volatility models, with the possible exception of the s.v. model, may be appropriate specifications of the variance dynamics. In fact, as shown in Table 3 for the BP series, the in-sample and out-of-sample volatility forecasts are highly correlated, especially within the GARCH class of models. These results are similar for the other three series.

Before evaluating the generated probability forecasts, Tables 4 and 5 report the in-sample and out-of-sample forecast evaluation results using the statistical loss functions from Section II. The in-sample results clearly indicate that the loss function chosen directly impacts the forecast evaluation results. MSE in minimized for all four series by the AR10sq model, as expected. However, as indicated by the S test results, it does not do so at the 5% level of statistical significance. The MAE and LL loss functions are minimized by the s.v. model significantly; i.e., most of the alternative models' forecasts fail the one-on-one S test at the 5% level. The results for HMSE indicate no clear pattern, and the GMLE loss function is minimized, but not significantly, in all cases by the GARCH model.

The out-of-sample forecast evaluation results vary considerably in comparison to the in-sample results. No one set of forecasts minimize a specific loss function across all four series, and few of the forecasts that minimize a loss function do so significantly. (West and Cho (1995) find qualitatively similar results for MSE using different statistical tests and various forecast horizons.) Of the twenty forecast evaluations reported in Table 5, in only eight cases is the minimizing set of forecasts the same across the in-sample and out-of-sample periods, and six of these cases are the MAE and LL with respect to the s.v. model. These results indicate that forecast evaluation based on such loss functions varies considerably; hence, the forecast user should do so only after careful consideration of how these loss functions reflect their particular economic interests.

4c. Specifying Economic Events

In the next subsection, the four economic events used for evaluating the volatility forecasts from the nine models are described. In order to simplify this sample evaluation, only the event to be forecast is varied. The QPS is the probability scoring rule used throughout, and the subsets of interest are the deciles [0, 0.1), [0.1, 0.2),...,[0.9, 1].

The first two events are based on the behavior of Δy_t, the one-day return. The first event is $\Delta y_t \in \left[-0.5 , 1 \right]$, which is a relatively high probability event, as shown in Table 6. (Note that, after adjusting for the specified conditional mean, this event is $\varepsilon_t \in \left[-0.5 - \mu - \theta\varepsilon_{t-1}, 1 - \mu - \theta\varepsilon_{t-1}\right]$.) Such events are of interest, for example, to spot traders deciding how to structure their market positions or to options traders holding out-of-the-money positions. The second event is related to the value-at-risk estimates now commonly used in managing the market risk of trading portfolios; see Lopez (1997) for further discussion. In such cases, traders generally wish to know a specified lower-tail quantile of their portfolio returns distribution in order to manage their market risk exposure. For this exercise, the lower five percent quantile of the empirical, in-sample distribution of Δy_t is specified as the threshold loss (TL) of interest to the forecast users, and the event to be forecast is $\Delta y_t < \mathrm{TL}$. As shown in Table 6, the observed in-sample frequency of occurrence of this event is constructed to be five percent, and the out-of-sample frequency is relatively close to five percent for all four currencies.

The last two events focus on the behavior of the level term y_t. The third event is a $\pm\gamma\%$ move in y_t; that is, $y_t \in \left[(1-\gamma/100)\,y_{t-1},\ (1+\gamma/100)\,y_{t-1}\right]$. Such an event would be of interest, for example, to a portfolio manager deciding when to transact in a foreign-currency denominated asset. The parameter γ is set at 2, except for the case of YN in which it is set to 0.2 (due to the high level of y_t). The last event is based on the idea of a central bank forecasting whether an exchange rate will remain within a specified target zone; i.e., the event of interest is $y_t \in [\,L,\ U\,]$. In this evaluation exercise, L and U are set as ±5% of y_T, the last in-sample observation. The observed frequencies of occurrence for event 3 are relatively high, while event 4 occurs only about 15 percent of the time, except for the YN series which is much higher. For the purpose of illustration, Figure 2 plots the out-of-sample probability forecasts from the GARCH-g model for the BP series for all four events of interest.

4d. Forecast Evaluation Results

Table 7 presents the in-sample QPS results for the four exchange rates and the four specified events. The minimum QPS value for each column in each panel is underlined. An interesting result is that the GARCH-g model minimizes the QPS value in half of the 16 cases, although the significance of this result is low given the poor S test results. That is, many of models generate QPS values that are greater than that of the GARCH-g model, but not significantly so. Overall, these results suggest that the GARCH-g model seems to characterize well the variance dynamics and the distribution of z_t for these series. The fitted GED distributions fit these series well both in the center of the distribution, as per events 1 and 3, and in the tails, as per event 2.

Turning to the out-of-sample QPS results presented in Table 8, the results are less conclusive. Overall, only three matches with the in-sample results occur. The GARCH-g model generates the minimum QPS values in only four cases, and this result (as well as all other minimum QPS values) is not generally significant at the 5% level. (Although not shown, sub-sample results with the out-of-sample period divided in half provide similar results.) Thus, the out-of-sample results indicate, much like the results in Table 5, that the chosen volatility models provide forecasts that are practically indistinguishable under these event-based, loss functions.

However, by focusing on probability forecasts, calibration tests can be used as an additional tool for evaluating the competing sets of out-of-sample volatility forecasts. The calibration test results are presented in the companion columns of Table 8. Probability forecasts that are well calibrated at the five percent significance level are denoted with '+' symbol. Clearly, in many cases, the calibration test reduces the number of models under consideration by rejecting the probability forecasts that are not calibrated over the user-specified subsets of the unit interval. (Once again, subsample results are consistent with the full-sample results.) For example, for BP with respect to event 1, forecasts from the GARCH-g model generate the lowest QPS value; the S test eliminates three competing models since they are significantly greater than this value; and the calibration test eliminates three additional model. Hence, in this case, although the forecasts from the GARCH-g, e.smooth-N and s.v. models are indistinguishable using this statistical evaluation framework, the number of models to consider has been reduced from nine to three. Although, as shown in Table 8, this evaluation procedure does not distinguish this well between competing volatility forecasts in all cases, it can be used to assist the forecast user in limiting the choice of volatility forecasts to consider.

In conclusion, this sample forecast evaluation provides three clear results. First, the different model rankings presented in Tables 4, 5, 7 and 8 clearly indicate that forecast evaluations are directly affected by the loss function selected. In the absence of economic loss functions, forecast users should carefully select the statistical loss functions that best match their interests. Second, the disparity between the in-sample and out-of-sample results indicate that tests of model fit and model specification should not be considered sufficient analysis. Out-of-sample forecast evaluation provides additional, useful information for volatility model specification and model selection. Third, the proposed framework for evaluating generated probability forecasts can generally reduce the number of competing volatility forecasts under consideration, a result that should be of particular interest to forecast users who must select among volatility models.

V. Conclusions

The forecast evaluations most meaningful to agents facing the issue of volatility model selection are those conducted under economically meaningful loss functions. In the general absence of such loss functions, researchers rely on statistical loss functions, such as MSE. Although reasonable for point forecasts, MSE and related statistical loss functions are problematic for volatility forecasts. In addition, since model rankings are highly dependent on the loss function chosen and these loss functions are generally devoid of economic content, the results from such analysis is of questionable worth.

In this paper, a framework for volatility forecast evaluation based on statistical tools that can be tailored to the economic interests of the forecast user is proposed. Specifically, the forecasts from a volatility model are transformed (via integration or simulation) into probability forecasts of events of interest. These probability forecasts are then evaluated using a proper scoring rule selected by the user and calibration tests over the subsets of the unit interval that are of interest to the user.

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 standard loss function for volatility forecast evaluation must be replaced with a more thoughtful selection. Yet, even under the appropriate loss function, forecast evaluations may provide comparative results that are not significantly different, as shown by the S test results. (In fact, such results for one-step-ahead volatility forecasts may not be surprising in light of Nelson (1992) as well as Nelson and Foster (1994), who find that short-term volatility forecasting is robust to a variety of misspecifications.) An advantage of the proposed framework is that calibration tests provide further analysis of the transformed volatility forecasts using criteria tailored to the user's interests.

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 probability-based loss functions are available. Secondly, the properties of the QPS and other scoring rules must be examined in light of West (1996); that is, uncertainty due to parameter estimation could be incorporated into this analysis. Other avenues for research are evaluating volatility forecasts from other models (i.e., different specifications for h_t and D) and for other financial time series under different loss functions. A systematic exploration of the theoretical and empirical aspects of the proposed evaluation framework should permit the formulation of more useful volatility model specifications and volatility forecasts.

References

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

Andersen, T.G. and Bollerslev, T., 1997. "Answering the Critics: Yes, ARCH Models Do Provide Good Volatility Forecasts," Working Paper #227, Department of Finance, Kellogg Graduate School of Management, Northwestern University.

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., Engle, R.F. and Nelson, D.B., 1994. "ARCH Models," in Engle, R.F. and McFadden, D., eds. *The Handbook of Econometrics, Volume 4*, 2959-3038. Amsterdam: North-Holland.

Bollerslev, T. and Ghysels, E., 1996. "Periodic Autoregressive Conditional Heteroskedasticity," *Journal of Business and Economic Statistics*, 14, 139-157.

Bollerslev, T. and Mikkelsen, H.O., 1996. "Modeling and Pricing Long Memory in Stock Market Volatility," *Journal of Econometrics*, 73, 151-184.

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.

Brailsford, T.J. and Faff, R.W., 1996. "An Evaluation of Volatility Forecasting Techniques," *Journal of Banking and Finance*, 20, 419-438.

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

Christoffersen, P.F., 1998. "Evaluating Interval Forecasts," *International Economic Review*, 39, 841-862.

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

Diebold, F.X. and Lopez, J.A., 1995. "Modeling Volatility Dynamics," in Hoover, K., ed., *Macroeconometrics: Developments, Tensions and Prospects*, 427-466. Boston: Kluwer Academic Press.

Diebold, F.X. and Lopez, J.A., 1996. "Forecast Evaluation and Combination," in Maddala, G.S. and Rao, C.R., eds., *Handbook of Statistics, Volume 14: Statistical Methods in Finance*, 241-268. Amsterdam: North-Holland.

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.

Dimson, E. and Marsh, P., 1990. "Volatility Forecasting without Data-Snooping," *Journal of Banking and Finance*, 14, 399-421.

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 Bollerslev, T., 1986. "Modeling the Persistence of Conditional Variances," *Econometric Reviews*, 5, 1-50.

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*, 393-415. Greenwich, CT: JIA Press.

Engle, R.F., Kane, A. and Noh, J., 1996. "Index-Option Pricing with Stochastic Volatility and the Value of Accurate Variance Forecasts," *Review of Derivatives Research*, 1, 139-158.

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

Geweke, J., 1989. "Exact Predictive Densities in Linear Models with ARCH Disturbances," *Journal of Econometrics*, 44, 307-325.

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

Granger, C.W.J. and Pesaran, M.H., 1996. "A Decision Theoretic Approach to Forecast Evaluation," Working Paper #96-23, University of California, San Diego.

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.

J.P. Morgan, 1996. *J.P. Morgan/Reuters Riskmetrics^{TM} - Technical Document.* New York: J.P. Morgan. (http://www.jpmorgan.com/RiskManagement/RiskMetrics/pubs.html).

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.

Lopez, J.A., 1997. "Regulatory Evaluation of Value-at-Risk Models," Federal Reserve Bank of New York Staff Report #33.

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

McDonald, J.B. and Newey, W.K., 1988. "Partially Adaptive Estimation of Regression Models via the Generalized T-Distribution," *Econometric Theory*, 4, 428-457.

Murphy, A.H., 1973. "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12, 595-600.

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*, 379-437. Boulder, Colorado: Westview Press.

Nelson, D.B., 1992. “Filtering and Forecasting with Misspecified ARCH Models: I,” *Journal of Econometrics*, 52, 61-90.

Nelson, D.B., and Foster, D.P., 1994. “Asymptotic Filtering Theory for Univariate ARCH Models,” *Econometrica*, 62, 1-41.

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.

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

West, K.D., 1996. “Asymptotic Inference about Predictive Ability,” *Econometrica*, 64, 1067-1084.

West, K.D. and Cho, D., 1994. “The Predictive Accuracy of Several Models of Exchange Rate Volatility,” *Journal of Econometrics*, 69, 367-391.

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.

Wooldridge, J.M., 1990. “A Unified Approach to Robust Regression-Based Specification Tests,” *Econometric Theory*, 6, 17-43.

Zellner, A. and Hong, C., 1991. "Bayesian Methods for Forecasting Turning Points in Economic Time Series: Sensitivity of Forecasts to Asymmetry of Loss Structures," in Lahiri, H. and Moore, G., eds., *Leading Economic Indicators: New Approaches and Forecasting Records*, 129-140. Cambridge, U.K.: Cambridge University Press.

Table 1. Conditional Mean Estimates and Residual Analysis for the First-Differenced Exchange Rate Series

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

$$\varepsilon_t | \Omega_{t-1} \sim N(0, \omega)$$

	BP	CD	DM	YN
Conditional Mean Parameters				
μ	0.0115 (0.0131)	0.0004 (0.0047)	0.0002 (0.0126)	-0.0218 (0.0114)
θ	0.0814* (0.0168)	0.0496* (0.0166)	0.0365* (0.0169)	0.0331* (0.0169)
In-sample Moments of ε_t				
variance	0.5170	0.0701	0.5203	0.4256
kurtosis	5.3181* (0.0826)	6.3905* (0.0826)	4.7579* (0.0826)	5.4200* (0.0826)
Q(20)	26.550	14.158	27.925	28.220
$Q^2(20)$	577.26*	522.10*	391.10*	225.45*
Out-of-sample Moments of ε_t				
variance	0.2377	0.0967	0.4493	0.5390
kurtosis	5.0975* (0.2193)	5.9554* (0.2193)	4.8715* (0.2193)	6.3198* (0.2193)
Q(20)	61.776*	25.474	22.573	23.958
$Q^2(20)$	63.408*	37.965*	106.534*	45.200*

Note: The four exchange rate series are the logged daily spot rates from 1980 to 1995. The in-sample period is 1980-1993 (3516 observations), and the out-of-sample period is 1994-1995 (499 observations). The conditional mean parameters of the in-sample, first-differenced series are estimated using least squares. Asymptotic standard errors are listed in parentheses. The test for excess kurtosis is relative to the standard normal distribution; this statistic is asymptotically distributed N(0,24/T). The portmanteau statistics for up to 20th-order serial correlation, Q(20) and $Q^2(20)$, are for the ε_t series and the ε_t^2 series, respectively. * indicates significance at the 5% level.

Table 2. Standard Diagnostic Tests of the In-Sample and Out-of-Sample Standardized Residuals

	In-Sample			Out-of-Sample		
	Kurtosis	Q(20)	$Q^2(20)$	Kurtosis	Q(20)	$Q^2(20)$
Panel A. BP						
homo.	5.32	26.62	581.68*	5.13	62.34*	63.42*
AR10sq	4.58	22.70	30.42	4.75	38.36*	20.77
AR10ab	4.51	21.14	21.78	5.13	31.51*	15.66
GARCH	4.43	22.94	17.75	5.00	30.50	20.40
GARCH-t	4.45	27.03	16.87	5.05	26.03	20.30
GARCH-g	4.45	28.98	16.87	5.05	25.21	20.45
e.smooth	4.50	23.91	16.38	5.63	21.57	15.62
s.v.	4.47	22.47	35.69*	5.58	30.19	28.49
Panel B. CD						
homo.	6.39	14.29	524.34*	5.96	25.47	37.96*
AR10sq	5.58	14.81	36.20*	5.52	22.95	20.82
AR10ab	5.83	14.52	21.85	6.00	23.89	20.27
GARCH	6.11	16.07	10.79	5.93	23.48	18.39
GARCH-t	6.30	17.25	13.58	5.89	23.73	17.69
GARCH-g	6.20	17.09	12.20	5.91	23.80	18.00
e.smooth	6.77	18.12	44.59*	5.79	22.12	16.38
s.v.	6.77	16.28	176.98*	6.16	21.85	21.53
Panel C. DM						
homo.	4.76	28.08	393.50*	4.86	22.83	106.54*
AR10sq	4.21	30.79	34.85*	4.51	18.56	37.97*
AR10ab	4.21	31.08	24.57	4.59	17.83	33.22*
GARCH	4.12	35.65*	14.85	4.38	16.21	28.74
GARCH-t	4.13	36.23*	13.88	4.40	16.20	28.38
GARCH-g	4.12	37.08*	14.17	4.41	16.41	28.74
e.smooth	4.34	42.19*	14.41	4.55	15.90	24.63
s.v.	4.30	35.24	52.01*	4.43	16.38	35.64*
Panel D. YN						
homo.	5.42	28.09	226.12*	6.32	23.93	45.19*
AR10sq	5.27	28.64	35.09*	6.35	22.67	21.83
AR10ab	5.93	29.48	39.37*	6.09	24.41	21.86
GARCH	5.51	29.84	33.00*	6.21	23.35	13.45
GARCH-t	5.87	33.34*	31.30	6.40	25.12	12.90
GARCH-g	5.75	35.97*	31.18	6.35	25.21	13.30
e.smooth	7.26	30.44	27.84	6.81	25.40	12.54
s.v.	6.84	34.75*	51.38*	5.99	29.15	29.49

Note: The first column in each panel presents the kurtosis of the in-sample and out-of-sample standardized residuals. The columns labeled Q(20) and $Q^2(20)$ are the portmanteau statistics for these standardized residuals and their squares. * denotes significance at the five percent level for a $\chi^2(20)$ distribution.

Table 3. Correlations between the Volatility Forecasts for BP

In-sample

	AR10sq	AR10ab	GARCH	GARCH-t	GARCH-g	e.smooth	s.v.
AR10sq	1.00						
AR10ab	0.94	1.00					
GARCH	0.86	0.84	1.00				
GARCH-t	0.87	0.85	0.99	1.00			
GARCH-g	0.87	0.85	0.99	0.99	1.00		
e.smooth	0.84	0.83	0.99	0.99	0.99	1.00	
s.v.	0.69	0.78	0.84	0.83	0.83	0.85	1.00

Out-of-sample

	AR10sq	AR10ab	GARCH	GARCH-t	GARCH-g	e.smooth	s.v.
AR10sq	1.00						
AR10ab	0.96	1.00					
GARCH	0.84	0.84	1.00				
GARCH-t	0.86	0.85	0.99	1.00			
GARCH-g	0.85	0.85	0.99	1.00	1.00		
e.smooth	0.82	0.82	0.99	0.99	0.99	1.00	
s.v.	0.62	0.71	0.82	0.81	0.81	0.84	1.00

Table 4. In-Sample Forecast Evaluation Results for Statistical Loss Functions

	MSE	MAE	LL	HMSE	GMLE
Panel A. BP					
homo.	1.1539	0.5665*	8.6182*	4.3204	0.3396*
AR10sq	1.0935	0.5473*	8.2178*	3.4809	0.2542
AR10ab	1.0996	0.5249*	7.8512*	4.3371*	0.2571
GARCH	1.0944	0.5464*	7.9304	3.4159*	0.2416
GARCH-t	1.1106	0.5604*	7.8821	3.1699	0.2437
GARCH-g	1.1127	0.5495*	7.7035	3.4653*	0.2426
e.smooth	1.0973	0.5458*	7.9064*	4.1656*	0.2572*
s.v.	1.1188	0.5047	7.4904	5.9245*	0.3004*
Panel B. CD					
homo.	0.0264	0.0794*	8.5695*	5.3887	-1.6586
AR10sq	0.0248	0.0749*	7.9613*	4.2311	-1.7953
AR10ab	0.0249	0.0713*	7.3964*	5.9917	-1.8100
GARCH	0.0255	0.0770*	7.4381	5.1159	-1.8246
GARCH-t	0.0261	0.0790*	7.4158	5.1547	-1.8184
GARCH-g	0.0256	0.0769*	7.6094	5.3606	-1.8236
e.smooth	0.0254	0.0747	7.4049	7.3534	-1.7894
s.v.	0.0256	0.0696	7.1285	10.0064	-1.7339
Panel C. DM					
homo.	1.0168	0.5706	7.4755*	3.7563	0.3464*
AR10sq	0.9750	0.5536*	7.1299*	3.0720	0.2683
AR10ab	0.9761	0.5330	6.7509	3.8699*	0.2644
GARCH	0.9772	0.5570*	7.0919*	3.0957*	0.2582
GARCH-t	0.9824	0.5694*	7.3830*	2.8496	0.2604
GARCH-g	0.9813	0.5574*	7.3371	3.1214*	0.2585
e.smooth	0.9839	0.5552*	6.8452	3.9607*	0.2807
s.v.	0.9861	0.5212	6.5459	4.7767*	0.2857*
Panel D. YN					
homo.	0.8004	0.4852*	8.8365*	4.4207	0.1452*
AR10sq	0.7746	0.4756*	8.5198*	4.2530	0.0885
AR10ab	0.7823	0.4451*	7.9554*	7.1143*	0.1140
GARCH	0.7841	0.4763*	8.3220*	4.5366	0.0861
GARCH-t	0.7994	0.4741*	8.3996	5.4410	0.0923
GARCH-g	0.8020	0.4801*	8.4988	4.9428	0.0896
e.smooth	0.7892	0.4771*	8.2154*	8.0179	0.1279*
s.v.	0.8045	0.4228	7.4526	13.5137*	0.2282*

Note: The columns represent the values of the statistical loss functions from Section II for the models' in-sample, fitted conditional variances. The minimum value within each column of each panel is underlined. * indicates that this particular set of in-sample forecasts rejects the null hypothesis of equality with the minimum value using the S test at the five percent significance level.

Table 5. Out-of-Sample Forecast Evaluation Results for Statistical Loss Functions

	MSE	MAE	LL	HMSE	GMLE
Panel A. BP					
homo.	0.3089*	0.4553*	11.8837*	1.1567	-0.1994*
AR10sq	0.2423	0.3466*	10.1319*	1.5552	-0.3888
AR10ab	0.2318	0.3034*	9.1532*	2.5910	-0.4523
GARCH	0.2275	0.2999*	8.9282	2.4515	-0.4611
GARCH-t	0.2290	0.3067*	8.8628	2.3033	-0.4582*
GARCH-g	0.2263	0.2981*	8.4016	2.4919	-0.4701
e.smooth	0.2257	0.2650	8.0144	5.8356	-0.4424
s.v.	0.2299	0.2672	8.2337	5.8601	-0.4099
Panel B. CD					
homo.	0.0470	0.0953	7.7033	9.5930	-1.2779
AR10sq	0.0465	0.1010	7.8258	7.3429	-1.3103
AR10ab	0.0469	0.0979	7.5533	11.2557	-1.2133
GARCH	0.0478	0.1123	8.1170	6.9315	-1.2686
GARCH-t	0.0482	0.1146	8.9795	6.1234	-1.2806
GARCH-g	0.0480	0.1125	8.2514	6.7614	-1.2715
e.smooth	0.0467	0.1085	8.3924	6.0445	-1.3083
s.v.	0.0478	0.0943	7.3589	15.9774	-1.0905
Panel C. DM					
homo.	0.7853	0.5343	8.0663	2.9008	0.2130
AR10sq	0.7430	0.4950*	7.4763*	2.8007	0.1266
AR10ab	0.7470	0.4707*	7.0276	3.9370	0.1347
GARCH	0.7402	0.4862*	7.2766	3.0493	0.1151
GARCH-t	0.7451	0.4981*	7.1669	2.8330	0.1198
GARCH-g	0.7452	0.4867*	7.0650	3.0989	0.1174
e.smooth	0.7453	0.4778	6.9420	4.3004	0.1378
s.v.	0.7527	0.4516	6.8011	4.9428	0.1581
Panel D. YN					
homo.	1.5456	0.5595	8.2659	8.5368	0.4073
AR10sq	1.5322	0.5701*	8.2178*	7.5746	0.3488
AR10ab	1.5265	0.5373	7.7350*	9.6895	0.3726
GARCH	1.4995	0.5750*	7.8797	6.3216	0.2922
GARCH-t	1.5127	0.5806*	7.7722	6.8564	0.2938
GARCH-g	1.5192	0.5847*	7.8844	6.4474	0.2921
e.smooth	1.5034	0.6009	8.2284	7.7505	0.3138
s.v.	1.5571	0.5175	7.3176	14.9116	0.5133

Note: The columns represent the values of the statistical loss functions from Section II for the models' out-of-sample volatility forecasts. The minimum value within each column of each panel is underlined. * indicates that this particular set of in-sample forecasts rejects the null hypothesis of equality with the minimum value using the S test at the five percent significance level.

Table 6. In-Sample and Out-of-Sample Observed Event Frequencies

Number of in-sample observations:	3515
Number of out-of-sample observations:	499

	Event 1	Event 2	Event 3	Event 4
Panel A. BP				
In-sample	72.4%	5.0%	83.1%	10.6%
Out-of-sample	84.4%	1.6%	92.8%	19.0%
Panel B. CD				
In-sample	96.8%	5.0%	89.0%	12.5%
Out-of-sample	94.4%	6.6%	94.8%	12.4%
Panel C. DM				
In-sample	72.7%	5.0%	91.8%	13.2%
Out-of-sample	72.9%	4.2%	79.4%	12.6%
Panel D. YN				
In-sample	77.4%	5.0%	89.4% [a]	36.5%
Out-of-sample	72.1%	4.8%	84.6% [a]	83.8%

[a] The parameter γ is set to 0.2, instead of 2 as for the series, due to the high level of the Japanese yen.

Note: Event 1 refers to the innovation event $\Delta y_t \in [-0.5, 1]$ or, equivalently, $\varepsilon_t \in [-0.5 - \mu - \theta\varepsilon_{t-1}, 1 - \mu - \theta\varepsilon_{t-1}]$. Event 2 refers to the innovation event $\Delta y_t < \mathrm{TL}$ or, equivalently, $\varepsilon_t < \mathrm{TL} - \mu - \theta\varepsilon_{t-1}$. TL is the lower five percent quantile of the in-sample, empirical distribution of Δy_t. Event 3 refers to the level event $y_t \in [(1-\gamma/100)y_{t-1}, (1+\gamma/100)y_{t-1}]$, where γ is set at 2 for BP, CD and DM and at 0.2 for YN. Event 4 refers to the level event $y_t \in [L, U]$, where L and U are ±5% of the last in-sample observation y_T.

Table 7. In-sample QPS Results

	Event 1	Event 2	Event 3	Event 4
Panel A. BP				
homo.	0.2024*	0.0474	0.1183	0.0178
AR10sq	0.1945	0.0465	0.1150	0.0177
AR10ab	0.1934	0.0465	0.1146	0.0177
GARCH	0.1946	0.0463	0.1146	0.0177
GARCH-t	0.1998*	0.0477*	0.1194*	0.0180
GARCH-g	0.1934	0.0463	0.1140	0.0177
e.smooth-N	0.1953	0.0464	0.1154	0.0178
e.smooth-t	0.1978*	0.0473	0.1184*	0.0179
s.v.	0.1961	0.0464	0.1150	0.0177
Panel B. CD				
homo.	0.0314	0.0476*	0.0982	0.0089
AR10sq	0.0302	0.0458	0.0910	0.0090
AR10ab	0.0303	0.0455	0.0902	0.0091
GARCH	0.0305	0.0460	0.0909	0.0090
GARCH-t	0.0324*	0.0478*	0.0958*	0.0090
GARCH-g	0.0304	0.0457	0.0899	0.0090
e.smooth-N	0.0305	0.0460	0.0917	0.0090
e.smooth-t	0.0315*	0.0471*	0.0944*	0.0089
s.v.	0.0307	0.0457	0.0915	0.0090
Panel C. DM				
homo.	0.2014*	0.0473	0.0701	0.0212
AR10sq	0.1938	0.0465	0.0682	0.0210
AR10ab	0.1921	0.0467	0.0683	0.0209
GARCH	0.1926	0.0466	0.0686	0.0209
GARCH-t	0.1982*	0.0473*	0.0719*	0.0212
GARCH-g	0.1912	0.0465	0.0683	0.0208
e.smooth-N	0.1929	0.0468*	0.0692	0.0209
e.smooth-t	0.1962*	0.0473*	0.0715*	0.0211
s.v.	0.1927	0.0469	0.0689	0.0209
Panel D. YN				
homo.	0.1788*	0.0473	0.0949*	0.0055
AR10sq	0.1739	0.0466	0.0932	0.0056
AR10ab	0.1711	0.0470	0.0933	0.0056
GARCH	0.1724	0.0468	0.0931	0.0056
GARCH-t	0.1762*	0.0469	0.0961*	0.0057
GARCH-g	0.1697	0.0467	0.0926	0.0055
e.smooth-N	0.1724	0.0470	0.0942	0.0057
e.smooth-t	0.1772*	0.0473	0.0971*	0.0058
s.v.	0.1720*	0.0475	0.0947	0.0055

Note: The columns contain the QPS values for the models' in-sample probability forecasts for the four specified events. The minimum QPS value within each column of each panel is underlined. * indicates that this particular set of in-sample forecasts rejects the null hypothesis of equality with the minimum QPS value using the S test at the five percent significance level; i.e., the models thus marked have QPS values significantly greater than the minimum value.

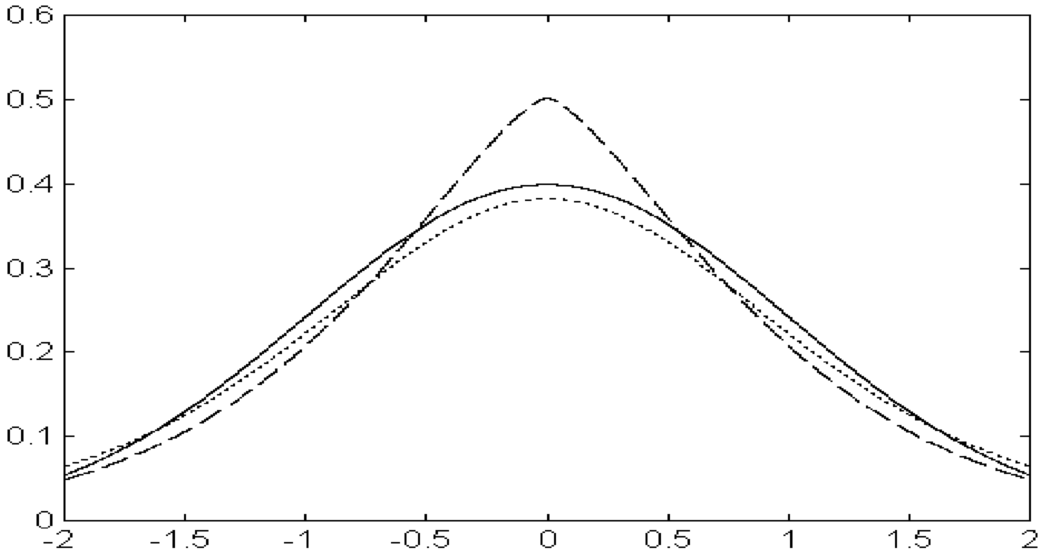
Table 8. Out-of-sample QPS Results

	Event 1		Event 2		Event 3		Event 4	
		Calib.		Calib.		Calib.		Calib.
Panel A. BP								
homo.	0.1600*		0.0178		0.0919		0.0135	
AR10sq	0.1392*		0.0163	0.0734	0.0121			
AR10ab	0.1338	0.0161	0.0696	0.0115				
GARCH	0.1330	0.0160	0.0682	0.0114				
GARCH-t	0.1411*		0.0171	0.0757*		0.0119		
GARCH-g	0.1312	+	0.0160	0.0677	0.0112			
e.smooth-N	0.1318	+	0.0161	0.0672	+	0.0109		
e.smooth-t	0.1338	0.0165	0.0672	+	0.0111			
s.v.	0.1332	+	0.0159	0.0683	+	0.0113	+	
Panel B. CD								
homo.	0.0538	+	0.0615	+	0.0504	+	0.0290	
AR10sq	0.0542	+	0.0625	+	0.0503	+	0.0278	
AR10ab	0.0543	+	0.0624	+	0.0505	+	0.0283	
GARCH	0.0551	+	0.0641	0.0518	+	0.0279		
GARCH-t	0.0571	0.0666	0.0551*		0.0279			
GARCH-g	0.0545	+	0.0632	+	0.0511	+	0.0279	
e.smooth-N	0.0544	+	0.0634	0.0510	+	0.0280		
e.smooth-t	0.0557	0.0652	0.0508	+	0.0280			
s.v.	0.0540	+	0.0620	+	0.0508	+	0.0296	+
Panel C. DM								
homo.	0.2005	0.0402	+	0.1509	0.0104			
AR10sq	0.1939	+	0.0400	+	0.1429	+	0.0104	
AR10ab	0.1928	+	0.0406	+	0.1428	+	0.0104	
GARCH	0.1923	+	0.0402	+	0.1416	+	0.0105	
GARCH-t	0.1961	0.0410	0.1462*		0.0105			
GARCH-g	0.1918	+	0.0401	+	0.1417	+	0.0105	
e.smooth-N	0.1929	+	0.0407	+	0.1424	+	0.0104	
e.smooth-t	0.1942	+	0.0412	0.1424	+	0.0104		
s.v.	0.1929	+	0.0408	+	0.1442*		0.0103	+
Panel D. YN								
homo.	0.2008	+	0.0457	+	0.1295	+	0.0091	
AR10sq	0.1993	+	0.0454	+	0.1272	+	0.0091	
AR10ab	0.1977	+	0.0453	+	0.1274	+	0.0092	
GARCH	0.1976	+	0.0451	+	0.1256	+	0.0092	
GARCH-t	0.2007	0.0461	0.1304	0.0097				
GARCH-g	0.1976	+	0.0451	+	0.1248	+	0.0090	
e.smooth-N	0.2016	+	0.0459	0.1288	+	0.0093		
e.smooth-t	0.2042	0.0470	0.1289	+	0.0094			
s.v.	0.2012	0.0456	+	0.1327	0.0093			

Note: The event columns contain the QPS values for the models' out-of-sample probability forecasts for the four specified events. The minimum QPS value within each column of each panel is underlined. * indicates that this particular set of out-of-sample forecasts rejects the null hypothesis of equality with the minimum QPS value using the S test at the five percent significance level; i.e., the models thus marked have QPS values significantly greater than the minimum value. The companion columns labeled 'calib.' present the results from the Z_0 calibration test using the deciles of the unit interval (i.e., [0,0.1); [0.1,0.2); ...; [0.9,1]) as the subsets of interest. The '+' symbol indicates that the forecasts do *not* reject the null hypothesis and are thus considered to be well calibrated over the selected subsets.

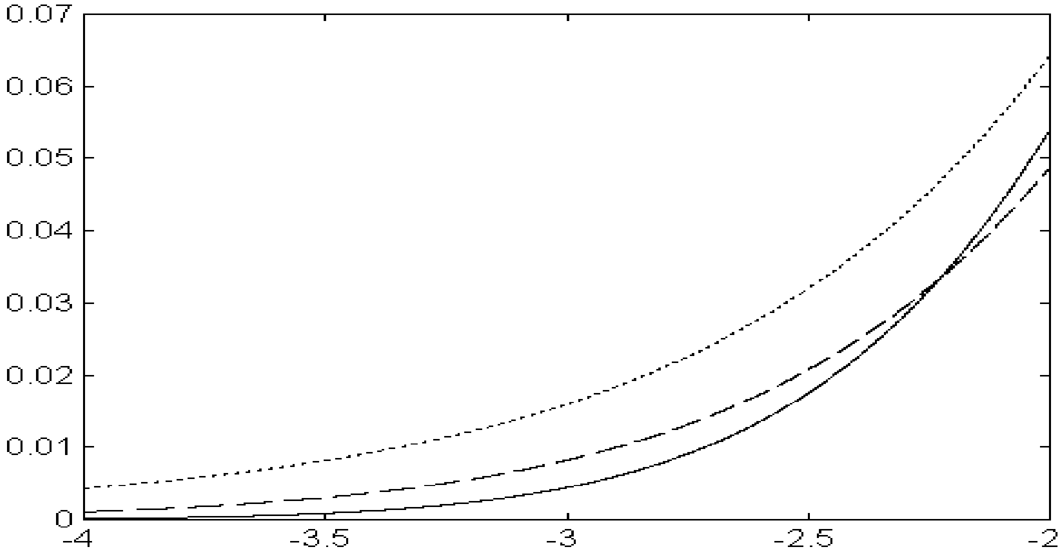
Figure 1. Alternative Distributions of the BP Standardized Residuals

Figure 1a.



...

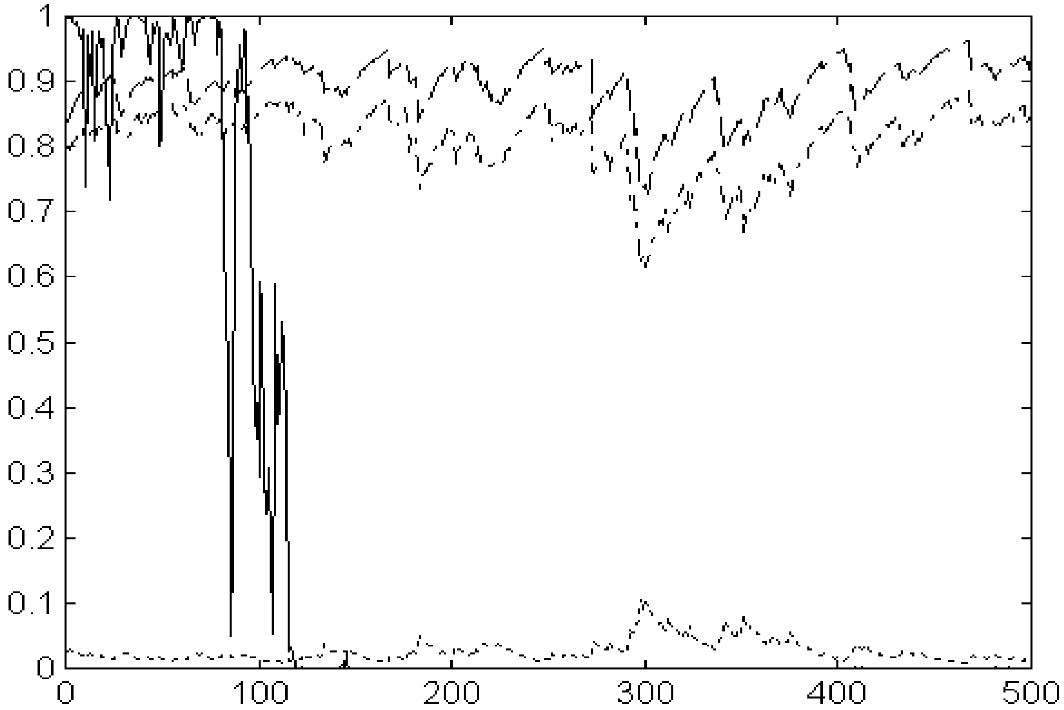
Figure 1b.



...

Note: The standard normal distribution is plotted with a solid line. The estimated t- and GED-distributions are plotted with dotted and dashed lines, respectively. The estimated shape parameter (i.e., degrees of freedom) for the t-distribution is 5.7, and the estimated shape parameter for the GED-distribution is 1.4.

Figure 2. Out-of-Sample Probability Forecasts from the GARCH-g Model for BP



...

Note: The out-of-sample probability forecasts from the GARCH-g model for events 1 through 4 are plotted with a dot-dash line, a dotted line, a dashed line and a solid line, respectively.

APPENDIX: Volatility Model Parameter Estimates

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

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

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

$$h_t = \sigma^2$$

Parameter	BP	CD	DM	YN
μ	0.0118 (0.0236)	0.0035 (0.0201)	0.0005 (0.0228)	-0.0216 (0.0224)
θ	0.0815* (0.0257)	0.0499* (0.0206)	0.0362* (0.0143)	0.0334* (0.0153)
σ^2	0.5168* (0.0244)	0.0700* (0.0028)	0.5203* (0.0237)	0.4255* (0.0191)

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 five percent level.

Table A.2. Volatility Model Parameters for the AR10sq Model

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

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

$$h_t = \omega + \sum_{i=1}^{10} \alpha_i \varepsilon_{t-i}^2 .$$

Parameter	BP	CD	DM	YN
μ	0.0118	0.0035	0.0005	-0.0216
	(0.0236)	(0.0201)	(0.0228)	(0.0224)
θ	0.0815*	0.0499*	0.0362*	0.0334*
	(0.0257)	(0.0206)	(0.0143)	(0.0153)
ω	0.2600*	0.0355*	0.2781*	0.2578*
	(0.0169)	(0.0037)	(0.0263)	(0.0230)
α_1	0.0726*	0.0244	0.0515*	0.0414*
	(0.0169)	(0.0169)	(0.0169)	(0.0168)
α_2	0.0183	0.0363*	0.0160	0.0232
	(0.0169)	(0.0170)	(0.0169)	(0.0169)
α_3	0.0180	0.0406*	0.0325	0.0363*
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_4	0.0379*	0.0093	0.0353*	-0.0163
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_5	0.0437*	0.0224	0.0596*	0.0282
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_6	0.0924*	0.0560*	0.0709*	0.0869
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_7	0.0773*	0.0247	0.0406*	0.0152
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_8	0.0492*	0.0908*	0.0606*	0.0779
	(0.0169)	(0.0170)	(0.0169)	(0.0169)
α_9	0.0494*	0.0593*	0.0462*	0.0541
	(0.0169)	(0.0170)	(0.0169)	(0.0169)
α_{10}	0.0391*	0.1314*	0.0539*	0.0443
	(0.0168)	(0.0169)	(0.0169)	(0.0169)

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

Table A.3. Volatility Model Parameters for the AR10ab Model

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

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

Parameter	BP	CD	DM	YN
μ	0.0118	0.0035	0.0005	-0.0216
	(0.0236)	(0.0201)	(0.0228)	(0.0224)
θ	0.0815*	0.0499*	0.0362*	0.0334*
	(0.0257)	(0.0206)	(0.0143)	(0.0153)
ω	0.2447*	0.0794*	0.2487*	0.2611*
	(0.0168)	(0.0069)	(0.0169)	(0.0201)
α_1	0.0835*	0.0291	0.0652*	0.0450*
	(0.0169)	(0.0169)	(0.0169)	(0.0168)
α_2	0.0462*	0.0027	0.0312	0.0399*
	(0.0169)	(0.0170)	(0.0169)	(0.0168)
α_3	0.0270	0.0498*	0.0499*	0.0335*
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_4	0.0356*	0.0167	0.0498*	0.0055
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_5	0.0390*	0.0605*	0.0457*	0.0358*
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_6	0.1044*	0.0895*	0.0877*	0.0727*
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_7	0.0453*	0.0491*	0.0457*	0.0289
	(0.0169)	(0.0171)	(0.0169)	(0.0169)
α_8	0.0548*	0.0963*	0.0731*	0.0727*
	(0.0169)	(0.0170)	(0.0169)	(0.0169)
α_9	0.0476*	0.0744*	0.0442*	0.0583*
	(0.0168)	(0.0170)	(0.0169)	(0.0169)
α_{10}	0.0575*	0.1198*	0.0453*	0.0536*
	(0.0168)	(0.0169)	(0.0169)	(0.0169)

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

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

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

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

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

Parameter	BP	CD	DM	YN
μ	0.0067	0.0010	0.0060	-0.0135
	(0.0111)	(0.0037)	(0.0111)	(0.0104)
θ	0.0704*	0.0425*	0.0349	0.0231
	(0.0178)	(0.0196)	(0.0181)	(0.0185)
ω	0.0083*	0.0018*	0.0153*	0.0180*
	(0.0033)	(0.0002)	(0.0029)	(0.0048)
α	0.0510*	0.1432*	0.0810*	0.0587*
	(0.0203)	(0.0320)	(0.0259)	(0.0242)
β	0.9332*	0.8430*	0.8919*	0.8993*
	(0.0810)	(0.0588)	(0.0733)	(0.1225)

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 five percent level.

Table A.5. Volatility Model Parameters for the GARCH(1,1)-t Model

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

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

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

Parameter	BP	CD	DM	YN
μ	-0.0002	-0.0033	0.0127	0.0054
	(0.0116)	(0.0035)	(0.0101)	(0.0095)
θ	0.0457*	0.0320	0.0319	0.0290
	(0.0205)	(0.0187)	(0.0167)	(0.0168)
ω	0.0095*	0.0011*	0.0134*	0.0083*
	(0.0033)	(0.0005)	(0.0037)	(0.0030)
α	0.0588*	0.1290*	0.0792*	0.0546*
	(0.0105)	(0.0182)	(0.0114)	(0.0101)
β	0.9266*	0.8691*	0.9018*	0.9247*
	(0.0137)	(0.0190)	(0.0138)	(0.0149)
v	5.7176*	5.7167*	5.7162*	5.6939*
	(0.0549)	(0.0518)	(0.0425)	(0.0567)

Note: These parameters values are maximum likelihood estimates. Asymptotic standard errors, presented in parentheses, are derived from the value of the numerical Hessian at the estimated parameter values. * represents significance at the five percent level.

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

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

$$\varepsilon_t \mid \Omega_{t-1} \sim \text{GED}\left(0, h_t, \eta\right)$$

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

Parameter	BP	CD	DM	YN
μ	0.0009	-0.0021	0.0145	0.0073
	(0.0100)	(0.0033)	(0.0096)	(0.0088)
θ	0.0389*	0.0327	0.0265	-0.0068
	(0.0173)	(0.0167)	(0.0168)	(0.0146)
ω	0.0089*	0.0014*	0.0138*	0.0115*
	(0.0030)	(0.0004)	(0.0035)	(0.0040)
α	0.0548*	0.1333*	0.0789*	0.0581*
	(0.0095)	(0.0166)	(0.0106)	(0.0113)
β	0.9284*	0.8572*	0.8966*	0.9156*
	(0.0127)	(0.0170)	(0.0131)	(0.0186)
η	1.3838*	1.3663*	1.4071*	1.1665*
	(0.0467)	(0.0417)	(0.0463)	(0.0363)

Note: These parameters values are maximum likelihood estimates. Asymptotic standard errors, presented in parentheses, 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

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

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

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

$$\nu_t \perp \eta_t$$

Measurement equation: $\log \varepsilon_t^2 = -1.27 + \alpha_t + \xi_t$

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

Parameter	BP	CD	DM	YN
μ	0.0118	0.0035	0.0005	-0.0216
	(0.0236)	(0.0201)	(0.0228)	(0.0224)
θ	0.0815*	0.0499*	0.0362*	0.0334*
	(0.0257)	(0.0206)	(0.0143)	(0.0153)
ϕ	0.9789*	0.9648*	0.9778*	0.9378*
	(0.0069)	(0.0100)	(0.0068)	(0.0303)
σ_η^2	0.0173*	0.0484*	0.0186*	0.0609*
	(0.0071)	(0.0107)	(0.0066)	(0.0298)

Note: The conditional mean parameters are estimated using ordinary least squares, and the standard errors are presented in parentheses. * indicates significance at the 5% 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 the method proposed by Dunsmuir (1979).