ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Does the Option Market Produce Superior Forecasts of Noise-Corrected Volatility Measures?

Gael M. Martin, Andrew Reidy and Jill Wright

June 2007

Working Paper 5/07

Does the option market produce superior forecasts of noise-corrected volatility measures?

Gael M. Martin, Andrew Reidy and Jill Wright*

June 7, 2007

Abstract

This paper presents a comprehensive empirical evaluation of option-implied and returns-based forecasts of volatility, in which recent developments related to the impact on measured volatility of market microstructure noise are taken into account. The paper also assesses the robustness of the performance of the option-implied forecasts to the way in which those forecasts are extracted from the option market. Using a test for superior predictive ability, model-free implied volatility, which aggregates information across the volatility 'smile', and at-the-money implied volatility, which ignores such information, are both tested as benchmark forecasts. The forecasting assessment is conducted using intraday data for three Dow Jones Industrial Average (DJIA) stocks and the S&P500 index over the 1996-2006 period, with future volatility proxied by a range of alternative noise-corrected realized measures. The results provide compelling evidence against the model-free forecast, with its poor performance linked to both the bias and excess variability that it exhibits as a forecast of actual volatility. The positive bias, in particular, is consistent with the option market factoring in a substantial premium for volatility risk. In contrast, implied volatility constructed from liquid at-the-money options is given strong support as a forecast of volatility, at least for the DJIA stocks. Neither benchmark is supported for the S&P500 index. Importantly, the qualitative results are robust to the measure used to proxy future volatility, although there is some evidence to suggest that any option-implied forecast may perform less well in forecasting the measure that excludes jump information, namely bi-power variation.

Keywords: Volatility Forecasts; Quadratic Variation; Intraday Volatility Measures; Model-free Implied Volatility; Superior Predictive Ability Test; Volatility Risk Premium.

JEL Classifications: C10, C53, G12.

*Department of Econometrics and Business Statistics, Monash University. Corresponding author: Gael Martin, Department of Econometrics and Business Statistics, P.O. Box, 11E, Monash University, Victoria, 3800, Australia. *(Email: Gael.Martin@Buseco.monash.edu.au).* This research has been supported by Australian Research Council Discovery Grant No. DP0664121. The authors would like to thank a co-editor and three referees for very extensive and constructive comments on an earlier draft of the paper.

1 Introduction

In recent years, many studies have investigated the relative performance of option-implied and returns-based forecasts of the future volatility of an asset. Since the advent of the realized volatility literature (e.g. Barndorff-Nielsen and Shephard, 2002, Andersen *et al.,* 2003), the measurable proxy used for the unobserved asset volatility has almost exclusively been constructed from high-frequency intraday returns. The most common such measure has been based on the sum of squared returns over small, regular intervals, such as 5 or 30 minutes (e.g. Poteshman, 2000, Blair, Poon and Taylor, 2001, Neely, 2003, Martens and Zein, 2004, Pong, Shackleton and Taylor, 2004, Jiang and Tian, 2005, and Koopman, Jungbacker and Hol, 2005), with such time intervals deemed to be sufficiently small to provide an accurate estimate of volatility over the time period of interest (a day, say), whilst, at the same time, avoiding much of the bias induced by the microstructure noise present in transactions data.[1] Studies that have adopted the realized volatility proxy have produced more definitive results, overall, than earlier work which used squared (or absolute) daily returns as the volatility measure (e.g. Day and Lewis, 1995). Nevertheless, conclusions have still been mixed, with the information content of option prices sometimes deemed to be superior to (or to subsume) that of historical returns (e.g. Blair *et al.,* 2001, Jiang and Tian, 2005) and sometimes not (e.g. Neely, 2003, and Martens and Zein, 2004).

The primary aim of this paper is to reassess the relative importance of option and spot prices in the prediction of future volatility by exploiting very recent developments related to the measurement of volatility in the presence of the empirical regularity of microstructure noise. The forecasting assessments are performed using a range of measures of future volatility that are alternatives to the conventional estimator based on squared returns sampled at an arbitrarily chosen regular interval. The first three such measures are designed to cater explicitly for microstructure noise, namely: the two scales realized volatility estimator of Zhang, Mykland and Ait-Sahalia (2005) and Ait-Sahalia, Mykland and Zhang (2005); the realized kernel estimator of Barndorff-Nielsen *et al.* (2005, 2006a, 2007); and the optimal sampling frequency estimator of Bandi and Russell (2006). As a fourth alternative, and in the spirit of the analysis conducted in Busch, Christensen and Nielsen (2006) and Anderson and Vahid (2007), only the continuous path component of future volatility is measured, via the bi-power variation estimator of Barndorff-Nielsen and Shephard (2004). The bi-power calculations are corrected for microstructure noise using the approach proposed in Andersen, Bollerslev and Diebold (2005). Finally, we pursue

[1] Jiang and Tian (2005) make some adjustment to the conventional realized variance measure to accommodate autocorrelation in intraday returns; see also Andersen *et al.* (2003).

the method of Large (2007), whereby a consistent estimator of quadratic variation, in the presence of microstructure noise, is constructed from a scaled function of the number of discrete price movements from transaction to transaction.

A secondary aim of our paper is to assess the robustness of the performance of option-implied forecasts to the way in which they are extracted from the option market. In particular, we compare the predictive performance of the 'model free' (MF) implied volatility of Britten-Jones and Neuberger (2000) and Jiang and Tian (2005), with implied volatility forecasts extracted from at-the-money market option prices.[2] The comparison is conducted both for individual DJIA stocks, on which American-style options are written, and the S&P500 index, for which the options are European.

In the case of the index, at-the-money (ATM) volatility forecasts are produced via the Black-Scholes (Black and Scholes (BS), 1973) option pricing model. A priori, and as was first argued by Jiang and Tian, one might expect the MF implied volatility to be a more accurate forecast of true volatility than the volatility implied by the empirically misspecified BS model.[3] Moreover, as the MF volatility is an estimate of quadratic variation in both the continuous and jump component of returns, it may be expected to produce a better prediction than BS for that reason alone, as long as the realized measure of future volatility itself incorporated jump information.

On the other hand, the fact that the MF quantity is a *risk-adjusted* expectation of actual volatility means that MF implied volatility incorporates any non-zero premium for volatility risk (or jump risk) that is factored into market option prices. As demonstrated by Bollerslev and Zhou (2006), under the assumption of a particular stochastic volatility specification, a non-zero volatility risk premium unambiguously leads to MF values that are biased forecasts of true volatility. To the extent that the BS volatility, for which no risk premium is *formally* incorporated, is less affected by this bias, it may actually out-perform the more flexibly specified MF alternative. Further, with the MF volatility being based on the full spectrum of option strike prices, i.e., using information from the volatility 'smile', it is necessarily more influenced than BS by the more extreme and noisy away-from-the-money option prices that prevail in high volatility periods in particular. The influence of these values may serve to further disconnect the MF volatility from the true underlying volatility process and thereby offset any accuracy gains associated with

[2] An option contract is said to be in-the-money if its immediate exercise would lead to a positive cash flow, that is, if the current value of the spot price exceeds the value of the strike price. Similarly, the option is out-of-the-money if the spot price is less than the strike price and at-the-money if the two prices are equal.

[3] The BS model assumes that returns on the underlying asset are normal with constant variance; assumptions that conflict with virtually all empirical evidence on financial returns.

the use of more options-based information.

In the case of the American options written on the DJIA stocks, neither the BS nor the MF formula is strictly appropriate. Rather than approximating the American price with the BS formula, as has often been done in past work (e.g. Christensen and Prabhala, 1998), we extract an ATM forecast using published option-market volatilities, calculated using a binomial tree method that caters for early exercise. For the MF calculation however, we *do* invoke an approximation by applying the European formula, with this approximation necessarily introducing some measurement error into the MF calculations. The spirit of the comparison in the stock option case, however, remains the same as for the index options: which form of option-implied volatility is given more support as a forecast of the volatility of the underlying, one that exploits the distributional information in the volatility smile, or one that does not?

To assess the relative performance of returns- and options-based forecasts of volatility, we take a different approach from previous analyses by using the test for superior predictive ability (SPA) of Hansen (2005) and Hansen and Lunde (2005a). That is, we address the question of whether any forecast method out-performs a particular options-based forecast while taking appropriate account of the fact that *multiple* forecast models are legitimate competitors. We use, in turn, the MF and ATM (or BS) implied volatility as the benchmark forecast, and document the robustness of the test results to the way in which microstructure noise, and random jumps, are handled in the measurement of future volatility. We also use different versions of the MF measure, ranging from a measure based on the full (empirically available) moneyness spectrum, to a measure based on a very truncated representation of that spectrum. Returns-based forecasts are produced both *directly*, via time series models for the volatility proxy itself, and *indirectly,* via generalized autoregressive conditional heteroscedastic (GARCH)-type models for daily returns. In the spirit of much of the recent literature, and as tallies with the features of our empirical data, we include long memory autoregressive fractionally integrated moving average (ARFIMA) models for the volatility proxy, in addition to short memory ARMA specifications. We also consider both short memory and long memory fractionally integrated GARCH (FIGARCH) models for daily returns, as well as certain asymmetric specifications.[4]

The forecasting assessment is conducted using a comprehensive set of intraday spot and option price data for three DJIA stocks - International Business Machines (IBM),

[4]The empirical work is conducted using Time Series Modelling 4.17 (www.timeseriesmodelling.com.), Ox (www.nuff.ox.ac.uk/Users/Doornik) and the SPA module for OX made publicly available by P. Hansen (http://www.stanford.edu/~prhansen/).

Microsoft (MSFT) and General Electric (GE) - and the S&P500 index, over the 1996 to 2006 period. Given that the noise adjustments to be discussed have their main motivation in the context of traded assets, we produce a more limited set of results for the index, with the primary focus for this particular data set being on the relative performance of the alternative option-implied forecasts. Analysis of the index data also enables MF-related results to be checked against results that use the VIX implied volatility as benchmark, where the latter is constructed by the Chicago Board Option Exchange (CBOE) using the MF methodology.

In quantifying the impact on the ranking of volatility models of different proxies of the true unobservable volatility, we expand upon the theme in Hansen and Lunde (2006a). In the latter work, the conventional realized volatility estimator, as proxy, is compared with squared daily returns, with the more accurate former measure found to produce a more reliable ranking of models in simulation experiments; see also Blair *et al.* (2001) and Hansen and Lunde (2005a). A further link with Hansen and Lunde is the way in which we conduct the SPA test for a criterion identified as 'robust' by these authors, namely mean squared forecast error (MSFE) constructed for variance quantities. Our work is also related to that of Andersen, Bollerslev, and Meddahi (2005), in which the R^2 of regression-based evaluations of alternative forecasting models are adjusted (upwards) to cater for the error-in-variables problem associated with proxying the unobserved forecast variable with a realized volatility measure that is biased in the presence of microstructure noise.

Other related work that assesses the relative forecasting performance of various noise-corrected realized volatility measures includes Anderson, Bollerslev and Meddahi (2006) and Ghysels and Sinko (2006). Neither of these analyses, however, includes options-based forecasts or assesses forecasting performance using the SPA approach. Bandi, Russell and Yang (2006) consider a range of noise-corrected measures, but evaluate those measures according to the profits/losses that option dealers would incur from pricing options on the basis of the alternative volatility forecasts. Bandi, Russell and Zhu (2006) and De Pooter, Martens and van Dijk (2006) also use an economic (rather than statistical) criterion function, gauging the impact of alternative volatility measurement on portfolio allocation decisions.

An outline of the remainder of the paper is as follows. In Section 2, we present the continuous time jump diffusion model for asset prices that underlies our analysis, and discuss the measurement of volatility within that context. The issues associated with forecasting (measured) volatility and evaluating alternative forecasts are addressed in Section 3. In Section 4, all aspects of the empirical investigation are outlined, including the

details of the construction of realized and option-implied volatility measures. The results represent strong evidence against the superiority of the MF implied volatility forecast. In contrast, the ATM implied volatility is given support as the benchmark forecast, at least for the three individual equity series investigated. *Both* options-based forecasts are rejected as superior benchmarks in the case of the S&P500 index. The qualitative results are robust to the measure used to proxy future volatility, apart from some results that suggest that option-implied forecasts may perform less well when the volatility measure excludes jump information. Section 5 concludes.

2 Measurement of Volatility

Denoting by $p(t)$ the logarithm of the asset price $P(t)$ at time t, we assume a continuous time jump diffusion process,

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + \kappa(t)dq(t),\ t \geq 0, \tag{1}$$

where $\mu(t)$ is a continuous (locally bounded) function, $\sigma(t)$ is a strictly positive volatility process, $W(t)$ is standard Brownian motion, and $\kappa(t)dq(t)$ is a random jump process that allows for occasional jumps in $p(t)$ of size $\kappa(t)$. The *quadratic variation* (QV) for the return over one day (say),

$$r_t = p(t) - p(t-1) \tag{2}$$

is then given by

$$QV_t = \int_{t-1}^{t} \sigma^2(s)ds + \sum_{t-1<s\leq t} \kappa^2(s). \tag{3}$$

That is, QV_t is equal to the sum of the *integrated volatility* of the continuous sample path component,

$$IV_t = \int_{t-1}^{t} \sigma^2(s)ds \tag{4}$$

and the sum of the $q(t)$ squared jumps that occur over day t. Denoting by p_{t_i} the ith logarithmic price that is observed on day t, and $r_{t_i} = p_{t_i} - p_{t_{i-1}}$ as the ith transaction return, it is now well known (see, in particular, Barndorff-Nielsen and Shephard, 2002, and Andersen *et al.*, 2003) that

$$RV_t = \sum_{t_{i-1},t_i\in[t-1,t]} r_{t_i}^2 \xrightarrow{p} QV_t \tag{5}$$

where RV_t is referred to as *realized volatility*[5].

[5] As is quite common in the literature, we use the term 'volatility' to refer to either a variance or a standard deviation quantity. Exactly which type of quantity is being referenced in any particular instance will be made clear by both the context and the notation.

Three comments can be made about the consistency result in (5). Firstly, the result in (5) is contingent upon *observed* price data adhering to the model in (1). In practice, observed prices should be viewed as reflecting both the process in (1) and a process that results from market microstructure noise. Secondly, the sample quantity RV_t will reflect both the continuous and jump components of the asset price process. In particular, only in the absence of jumps ($\kappa(t) = 0$) will realized volatility estimate integrated volatility alone. Thirdly, in practice, prices are not continuous random variables, but move in discrete numbers of ticks. This discreteness can be viewed as one component of the microstructure noise referred to in the first point. We take up these points in Sections 2.1, 2.2 and 2.3 respectively .

2.1 Realized Volatility Calculation in the Presence of Microstructure Noise

As highlighted in Barndorff-Nielsen *et al.* (2005, 2006a, 2007), Zhang *et al.* (2005), Ait-Sahalia *et al.* (2005) and Bandi and Russell (2006), amongst others, observed transactions data do not adhere to (1), due to a range of factors collectively referred to as *market microstructure*. That is, the true price is distorted by effects that include price discreteness, separate trading prices for buyers and sellers (the bid-ask spread) and the information asymmetry of market participants. Due to the presence of such factors, the 'true' latent logarithmic price process, $p^*(t)$, may be assumed to follow (1), but is observed with error. Hence, a suitable model for the observed ith logarithmic price on day t, is

$$p_{t_i} = p^*_{t_i} + \varepsilon_{t_i}, \tag{6}$$

where ε_{t_i} is assumed (at least initially) to be an $i.i.d.$ white noise component. The ith *observed* transaction return, r_{t_i}, is thus given by the sum of the latent return, $r^*_{t_i} = p^*_{t_i} - p^*_{t_{i-1}}$, and a first order moving average (MA) process, $\eta_{t_i} = \varepsilon_{t_i} - \varepsilon_{t_{i-1}}$. It is straightforward to show (see Zhang *et al.* 2005) that

$$E\left(RV_t | \, p^*(t_i)\right) = \sum_{t_{i-1}, t_i \in [t-1,t]} r^{*2}_{t_i} + 2n\sigma^2_\varepsilon, \tag{7}$$

where n denotes the number of transaction returns observed on day t. Hence, realized volatility constructed from the *observed* returns is a biased representation of $\sum_{t_{i-1}, t_i \in [t-1,t]} r^{*2}_{t_i}$ and, hence, a biased estimator of quadratic variation. Moreover, the bias is $O(n)$, meaning that bias is proportional to the number of returns used to construct the realized volatility measure. Defining

$$\widehat{\sigma^2_\varepsilon} = \frac{1}{2n} RV_t, \tag{8}$$

Zhang *et al.* (2005) also demonstrate that as $n \to \infty$, $n^{1/2}(\widehat{\sigma_\varepsilon^2} - \sigma_\varepsilon^2) \to N(0, E(\varepsilon^4))$. That is, (scaled) realized volatility constructed from observed transactions data is a consistent estimator, not of quadratic variation, but of the variance of the microstructure noise, σ_ε^2; see also Bandi and Russell (2005).

Given the clear deficiency of the realized volatility estimator based on all observed data, alternative estimators that adjust for the impact of noise have been suggested. We include three such estimators in our empirical analysis, referring readers to the relevant papers for more details about the construction of these specific estimators and discussion of related variants.

2.1.1 The Two-Scale Realized Volatility (TSRV) Estimator

The TSRV estimator of Zhang *et al.* (2005) and Ait-Sahalia *et al.* (2005) is based on a weighted difference between two estimators: 1) an average of realized volatilities calculated essentially as per (5), but over moving windows of subgrids defined on a 'slow' time scale (only observations several transactions apart are used); and 2) realized volatility calculated on a 'fast' time scale, as per (5) with all transactions used. More specifically, the full grid of observational points on day t, $G = \{t_0, t_2, \ldots, t_i, t_{i+1}, \ldots, t_n\}$, is partitioned into K nonoverlapping subgrids $G^{(k)}$, $k = 1, 2, 3, \ldots, K$, where

$$G^{(k)} = \{t_{k-1}, t_{k-1+K}, t_{k-1+2K}, \ldots, t_{k-1+n_K K}\}, \tag{9}$$

with $n_K = [\frac{n-K-1}{K}]$. Realized volatility is then constructed from returns over successive time points in $G^{(k)}$, denoted by $t_{i,-}$ and t_i respectively,

$$RV_t^{(k)} = \sum_{t_{i,-}, t_i \in G^{(k)}} r_{t_i}^2, \tag{10}$$

and the TSRV estimator then defined as

$$TSRV_t = \left(\frac{n}{(K-1)n_K}\right)\left(\overline{RV}_t^{(K)} - \frac{n_K}{n} RV_t\right), \tag{11}$$

where $\overline{RV}_t^{(K)} = \frac{1}{K}\sum_{k=1}^{K} RV_t^{(k)}$, RV_t is as defined in (5) and the scale factor $\left(\frac{n}{(K-1)n_K}\right)$ is used to improve the performance of the estimator when K is large.

The TSRV measure is shown to be a consistent estimator of quadratic variation, in the presence of microstructure noise. In the spirit of recent work (e.g. Hansen and Lunde, 2006b) in which the increased prevalence of time dependent noise has been documented, we accommodate dependent noise via the modification to (11) suggested by Ait-Sahalia *et al.* (2005),

$$TSRV2_t = \left(\frac{n}{(K-J)n_K}\right)\left(\overline{RV}_t^{(K)} - \frac{n_K}{n_J}\overline{RV}_t^{(J)}\right). \tag{12}$$

The elements in the average defining $\overline{RV}_t^{(J)} = \frac{1}{J}\sum_{j=1}^{J} RV_t^{(j)}$ are defined analogously to $RV_t^{(k)}$ in (10), but with $1 < J < K$), and $n_J = [\frac{n-J-1}{J}]$.[6]

2.1.2 The Realized Kernel (RKERN) Estimator

Barndorff-Nielsen *et al.* (2005, 2006a, 2007) develop kernel estimators of the quadratic variation, with the weights used in constructing the kernel chosen to ensure that the resultant estimator is consistent in the presence of microstructure noise, and the autocorrelation in transaction returns that it induces[7]. Consistent with the definition of $RV_t^{(k)}$ above, we define

$$RCV_t^{(k)}(h) = \sum_{t_{i,-},t_i,t_{i+h,-},t_{i+h}\in G^{(k)}} r_{t_i} r_{t_{i+h}}, \quad h = -H,\ldots,-1,0,1,2,\ldots H,$$

as the realized autocovariance function constructed from returns observed over pairs of successive time points in $G^{(k)}$ in (9), $k = 1,2,3,\ldots,K$, with the returns being $|h|$ time points apart.[8] When $h = 0$, we regain the variance quantity, $RV_t^{(k)}$. The averaged (or 'subsampled') version of $RCV_t^{(k)}(h)$ is then given by $\overline{RCV}_t^{(K)}(h) = \frac{1}{K}\sum_{k=1}^{K} RCV_t^{(k)}(h)$, analogously with the averaged version of $RV_t^{(k)}$ above. A symmetric version of the realized kernel (RKERN) estimator is given by

$$\begin{aligned} RKERN_t &= \sum_{h=-H}^{H} w(\frac{h-1}{H})\overline{RCV}_t^{(K)}(h) = w_0\overline{RV}_t^{(K)} \\ &\quad + \sum_{h=1}^{H} w(\frac{h-1}{H})\left\{\overline{RCV}_t^{(K)}(h) + \overline{RCV}_t^{(K)}(-h)\right\}, \end{aligned} \tag{13}$$

with the particular form chosen for the weights, w_h, $h = 2,3,\ldots,H$, determining the precise version of the estimator. In the empirical work we report results based on the

[6] Following Zhang *et al.* (2005) we use $K = cn^{2/3}$, where $c = \left(16\sigma_\varepsilon^4/TE\left(\eta^2\right)\right)^{1/3}$ and $\eta^2 = \frac{4}{3}\int_{t-1}^{t}\sigma^4(s)ds$. The term σ_ε^4 is square of the variance of the noise, while $\int_{t-1}^{t}\sigma^4(s)ds$ is the integrated quarticity. σ_ε^2 is estimated as in (8), but using transactions that are approximately one-minute apart. This modified estimate of the noise variance is an attempt to reduce the impact of dependent noise; see Barndorff-Neilsen *et al.* (2006a). The term in the denominator, $E(\eta^2)$, is estimated as $\widehat{E(\eta^2)} = \frac{4}{3}\left[RV_t(\Delta)\right]^2$ using $\Delta \approx 30$ minutes, and we use $J = \max(1, \frac{K}{4})$. See Barndorff-Neilsen *et al.* (2006b) for further discussion of some of these computational issues.

[7] Although the kernel estimator is introduced within the context of general semimartingales, the properties of the estimator are demonstrated under the assumption of a model without random jumps (i.e. with $\kappa(t) = 0$ in (6)). In Barndorff-Nielsen *et al.* (2006a and 2007) the properties of kernel estimators under a non-$i.i.d$ assumption for the noise process are investigated.

[8] The notation $r_{t_{i+h}}$ denotes the return over successive time-points in the sub-grid $G^{(k)}$, where that return is $|h|$ time points distant from r_{t_i} according to the sub-grid $G^{(k)}$.

cubic kernel estimator, in which $w(0) = w(1) = 1$; $w(\frac{h-1}{H}) = 1 - 3(\frac{h-1}{H})^2 + 2(\frac{h-1}{H})^3$, $h = 2, 3, \ldots H$. [9]

2.1.3 The Optimally Sampled Realized Volatility (OSRV) Estimator

Bandi and Russell (2006) propose an estimator that optimally balances the noise-induced bias associated with an increase in the number of transactions used in the construction of realized volatility, with the increased efficiency produced by higher sampling frequency. Specifically, they define the optimally sampled realized volatility (OSRV) estimator,

$$OSRV_t = \sum_{j=0}^{M_t^*} r_{t+j\delta_t,\delta_t}^2, \tag{14}$$

based on M_t^* discretely sampled δ_t-period returns, $r_{t,\delta_t} = p(t) - p(t - \delta_t)$, where the sampling frequency, $\delta_t = 1/M_t^*$, is chosen to minimize the mean squared error (MSE) of $OSRV_t$ as an estimator of quadratic variation. Under certain conditions[10], the MSE is shown to be a function of M_t^*, the second and fourth moments of the noise process, the integrated variance, $\int_{t-1}^{t} \sigma^2(s)ds$, and the integrated quarticity, $\int_{t-1}^{t} \sigma^4(s)ds$. Given sample estimates of all population moments, M_t^* is chosen so as to minimize MSE where, as indicated by the notation, M_t^* (and, hence, δ_t) varies with t.[11]

2.2 Realized Bi-Power Variation

With regard to the role of the continuous and jump components of the asset price process in the calculation of realized measures, Barndorff-Nielsen and Shephard (2004) focus on the separate identification and estimation of integrated volatility, exclusive of jumps.

[9] The weights $w(0) = w(1) = 1$ ensure that the kernel is asymptotically unbiased, with inclusion of the additional terms in the kernel ($h = 2, 3, \ldots, H$) serving to reduce the variance. The value of $H = c_K \sqrt{\frac{\widehat{\sigma_\varepsilon^2}}{\widehat{\int_{t-1}^{t} \sigma^4(s)ds}}} n$ is chosen to (approximately) minimize the asymptotic variance of the estimator, where c_K is specified exactly as in Barndorff-Nielsen *et al.* (2007) for the cubic kernel case, with K determined as per Footnote 6. Note that we adopt a subsampled version of the kernel estimator despite the results in Barndorff-Nielsen *et al.*, which indicate that the subsampling can increase the asymptotic variance of the estimator. The estimates of the noise variance (σ_ε^2) and integrated volatility used in the construction of H are the same as those used in the construction of K, as detailed in Footnote 6. See Barndorff-Nielsen *et al.* (2006a) for discussion of the connection between the kernel estimator and the two-scale estimator of Zhang et al. (2005).

[10] In particular, with reference to (1), it is assumed that $\mu(t) = \kappa(t) = 0$.

[11] Following Bandi and Russell (2006), we approximate the optimal value of M_t^* as $M_t^* \sim \left(\frac{\widehat{\int_{t-1}^{t} \sigma^4(s)ds}}{\widehat{\sigma_\varepsilon^4}} \right)^{1/3}$. The numerator and denominator are both calculated as explained in Footnote 6. Returns on day t are thus sampled less frequently (M_t^* is smaller), the larger is the squared variance of the noise in the data relative to the quarticity of the underlying efficient price process.

Defining *realized bi-power variation* as

$$BPV_t = \frac{\pi}{2} \sum_{t_{i-1}, t_i \in [t-1,t]} |r_{t_i}| \left| r_{t_{i-1}} \right|, \tag{15}$$

they show that as $n \to \infty$, $BPV_t \xrightarrow{p} IV_t = \int_{t-1}^{t} \sigma(s) ds$, i.e. that realized bi-power variation consistently estimates the integrated variance of the continuous sample path component of the price process in (1). Analogous to the realized volatility estimator in (5), for very large n the statistic in (15) is adversely affected by the presence of microstructure noise. To at least partially offset this bias, Andersen, Bollerslev and Diebold (2005), and Huang and Tauchen (2005) propose a modification of (15), whereby the sum of absolute adjacent returns is replaced with the sum of the corresponding one-period staggered returns. In the empirical section we implement an averaged version of this modified estimator,

$$\overline{BV}_t^{(K)} = \frac{1}{K} \sum_{k=1}^{K} BV_t^{(k)}, \tag{16}$$

where $BV_t^{(k)} = \frac{\pi n}{2n-4k} \sum_{t_{i,-}, t_i, t_{i+2,-}, t_{i+2} \in G^{(k)}} |r_{t_i}| \left| r_{t_{i+2}} \right|$, and k and K are defined with respect to the transaction grid in (9).[12]

A-priori one would anticipate that option-implied forecasts, to the extent that such forecasts incorporate jump information, may be less accurate in forecasting (16) than in forecasting other realized volatility measures. This issue is investigated in Section 4.

2.3 Realized Volatility for Discrete Prices

To address the fact that prices move in discrete numbers of ticks, Large (2007) proposes an estimator of quadratic variation that focusses on the number and direction of price changes during the day, rather than the magnitude of such changes, as measured by intraday returns. The estimator, which we refer to as the 'alternation' estimator, is given by

$$ALT_t = n^{(ch)} tick^2 \frac{C}{A}, \tag{17}$$

where $n^{(ch)} \in \mathbb{N}$ is the number of price changes in a day and $tick$ is the price tick (i.e. the minimum amount by which the price can change on the relevant exchange). Defining an alternation as a price change that occurs in the opposite direction to the previous price

[12] As pointed out by a referee, the subsampling process may affect the robustness of the bi-power measure to jumps.

change, and a continuation as a price change in the same direction, A then denotes the number of alternations and C the number of continuations, with $A + C = n^{(ch)}$.[13]

Without the presence of microstructure noise, the estimator $n^{(ch)}tick^2$ is a consistent estimator of quadratic variation, whilst in the presence of noise the value of $n^{(ch)}tick^2$ is asymptotically biased. Given that the presence of noise implies an excess of alternations, multiplication by the fraction C/A produces a consistent estimator in the presence of noise. The modified version of the alternation estimator that we apply in the empirical investigation (see also Barndorff-Nielsen and Shephard, 2005), and which we denote by the acronym ALTM, is given by

$$ALTM_t = \overline{RV}_t^{(K)} \frac{C}{A}, \tag{18}$$

which is simply the (average of the) realized volatility measure in (10) multiplied by C/A in order to correct for the upward bias induced by the noise.[14]

3 Forecasting Volatility

3.1 Overview

Since the advent of the realized volatility literature, not only has focus shifted from daily returns to the use of a measurable proxy for volatility based on intraday day returns, but emphasis is also now given to production of *direct* forecasts produced from standard time series models; see Andersen, Bollerslev and Meddahi (2004) for relevant discussion. In particular, the stylized empirical properties of the (logarithmic) realized volatility measures are such that long-memory Gaussian ARFIMA models for this (transformation of) realized volatility have become the mainstay of empirical work. As such, the interest is now in the merit of these *direct* forecasts of some proxy of future volatility, compared with *indirect* forecasts based on low-frequency (usually daily) returns, in particular returns produced via the ubiquitous GARCH-type specifications. Such returns-based specifications are then compared with forecasts from the options market, with the relative predictive performance of the latter thereby assessed.

In this paper *eight* volatility measures are used in the comparative analysis, including the six volatility measures outlined in Section 2 namely TSRV and TSRV2 in (11) and (12) respectively, RKERN in (13), OSRV in (14), BV in (16) and ALTM in (18). A measure based on fixed 5 minute sampling, denoted by RV(5), is also included as being representative of the type of measure used in literature prior to the development of the

[13] The first price of the day is defined as an alternation.

[14] See Oomen (2006) for a related measure based on a discrete jump process.

more formal noise-(and/or jump-)adjusted measures. As an intermediate type of measure we also include a subsampled (or averaged) version of RV(5), denoted by RVAV(5).[15] All measures are used both as proxies for the latent volatility and as the basis for forecasting future volatility. Following Hansen and Lunde (2005b) we extend all eight within-day volatility measures to 24-hour measures by taking a weighted average of the within-day measure and the squared overnight (close-to-open) return, where the weights are determined empirically using a mean squared error (MSE) criterion.[16]

Details of the models used to produce the direct and indirect returns-based forecasts follow, plus details of the production of alternative options-based forecasts.

3.2 Forecast Model Set

3.2.1 Indirect (Daily) Returns-Based Forecasts

In order to cater for the standard empirical features exhibited by daily returns on all three individual stocks and the S&P500 index, namely varying degrees of time-varying volatility, excess kurtosis, skewness, plus long memory in the squared returns, the forecast set includes forecasts produced from a range of GARCH-type specifications with a Student t conditional distribution.[17] Given $r_t = \mu + \varepsilon_t = \mu + \sigma_t e_t$, where r_t denotes the tth daily return in (2), μ the mean daily return, σ_t^2 the variance for day t and $e_t \sim Student\ t(0, 1, \nu)$, the following GARCH, threshold GARCH (TGARCH), power ARCH (PARCH) and fractionally integrated GARCH (FIGARCH) models are included in the initial forecast set:

$$
\begin{aligned}
GARCH(p,q) \quad &: \quad \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + ... + \alpha\varepsilon_{t-q}^2 + \beta\sigma_{t-p}^2 \\
TGARCH(p,q) \quad &: \quad \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \alpha\gamma s_t\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + ... + \alpha\varepsilon_{t-q}^2 + \alpha\gamma s_t\varepsilon_{t-q}^2 + \beta\sigma_{t-p}^2 \\
PARCH(p,q) \quad &: \quad \sigma_t^\delta = \omega + \alpha\left|\varepsilon_{t-1}\right|^\delta + \beta\sigma_{t-1}^\delta + ... + \alpha\left|\varepsilon_{t-q}\right|^\delta + \beta\sigma_{t-p}^\delta \\
FIGARCH(p,d,q) \quad &: \quad \beta(L)(\sigma_t^2 - \omega) = \left[\beta(L) - (1-L)^d\alpha(L)\right]\varepsilon_t^2.
\end{aligned}
$$

The notation L is used to denote the lag operator, with $\alpha(L)$ and $\beta(L)$ being polynomials of order q and p in L, $d > -0.5$ is the fractional parameter, $(1-L)^d = \sum_{j=0}^{\infty} b_j L^j$, with $b_0 = 1$ and $b_j = \frac{-d\Gamma(j-d)}{\Gamma(1-d)\Gamma(j+1)}$, and the remaining parameters satisfy the usual restrictions.

[15] The measure RV(5) is based on artificial returns five minutes apart. We experimented with both the previous tick and interpolation methods to construct these returns. The results were so similar (for the particular purpose at hand) that we report only the results using the interpolation method. The measure RVAV(5) is the averaged version of RV(5) based on successive subgrids of (artificial) prices spaced five minutes apart.

[16] See Hansen and Lunde (2005b) for precise details, including of the rule adopted for discarding outliers when calculating the weights.

[17] Details of all preliminary data analysis are available from the authors on request.

In the asymmetric TGARCH model, $s_{t+1} = 1$ if $\varepsilon_t < 0$ and 0 otherwise. The PARCH model nests the GARCH model when and $\delta = 2$. Maximum lag lengths of $p = q = 2$ are entertained for each model type.

3.2.2 Direct (Intraday) Returns-Based Forecasts

To cater for the long memory properties exhibited by all of the realized volatility measures, for each of the four time series under investigation, we produce direct forecasts using the following ARFIMA(p,d,q) model with Student t innovations (where the generic notation y_t refers to any of the volatility measures described in Section 2, and α its mean):

$$\phi(L)(1-L)^d\,(\ln y_t - \alpha) = \theta(L)u_t\ ; \qquad u_t \sim \textit{Student } t\ (0, \sigma^2\frac{\nu}{\nu-2}, \nu).$$

The autoregressive and moving average polynomials $\phi(L)$ and $\theta(L)$ are of lag length p and q respectively and $(1-L)^d$ is as defined earlier. For completeness we also produce forecasts via short memory ARMA (p,q) models. As with the GARCH models, the ARFIMA and ARMA models are estimated for lag lengths up to and including $p = q = 2$. In the model set we include both own-forecasts (i.e. a forecast for a particular measure based on a model estimated for that same measure) and cross-forecasts (i.e. forecasts based on other measures).

3.2.3 Option-Implied Forecasts

The BS option price model assumes that the asset price, $P(t)$, follows a geometric Brownian motion process with constant diffusion parameter σ. Under this distributional assumption, the BS price of a European call option with strike price X and maturity T is

$$BS(\sigma) = P_t^{(D)}\Phi(d_1) - X^{-i_t\tau}\Phi(d_2), \tag{19}$$

where $d_1 = \left(\ln(P_t^{(D)}/X) + (i_t + 0.5\sigma^2)\,\tau\right)/\sigma\sqrt{\tau}$, $d_2 = d_1 - \sigma\sqrt{\tau}$, $P_t^{(D)}$ = the (dividend-discounted) spot price at time t, i_t = the (annualized) risk free rate of return at time t, $\tau = T - t$ = the time to maturity (expressed as a proportion of a year) and $\Phi(.)$ = the cumulative normal distribution. An observed market option price at time t for a call option with maturity T and strike X, $C_t(T, X)$, can be used to produce an estimate of σ *implied* by $C_t(T, X)$, by equating $C_t(T, X)$ to the right-hand-side of (19) and solving for σ.

If the BS model were correct, the estimate of σ implied by $C_t(T, X)$ would be invariant to both X and τ. As is now standard knowledge however, implied volatilities across strike prices (or across 'moneyness', X/P_t, with P_t the current spot price) exhibit stylized

'smile' patterns, with these patterns varying, in turn, with the time to expiry, τ. Such patterns have been shown to be a manifestation of the misspecification of the BS model (e.g. Bakshi, Cao and Chen, 1997, Corrado and Su, 1997, Bates, 2000, Lim, Martin and Martin, 2005), with the downward skew shape for equities, in particular, being evidence that market option prices have factored in the negative skewness that characterizes equity returns.

It is with this misspecification issue in mind that Britten-Jones and Neuberger (2000) and Jiang and Tian (2005) motivate the MF implied volatility. As demonstrated by these authors, under the assumption of a diffusion process for the spot price a forecast of integrated variance for the period t to T can be determined from observed European call option prices with maturity T as follows

$$E_t^* \left[\int_t^T \sigma^2(s) ds \right] = 2 \int_0^\infty \frac{C_t(T,X) e^{i_t \tau} - \max\left[0, P_t^{(D)} e^{i_t \tau} - X\right]}{X^2} dX, \tag{20}$$

where E_t^* denotes the time t expectation with respect to the *risk-neutral* distribution of the asset price. Jiang and Tian point out that the result in (20) can be extended to jump-diffusion processes, in which case the method produces a forecast of quadratic variation. That is, in the case where the true latent price follows the model in (1), the implied variance is an estimate of (3), rather than an estimate of the integrated volatility in (4). Crucially, the calculation in (20) avoids the BS misspecification of the spot price process as geometric Brownian motion with a constant diffusion parameter. Instead, the right hand side of (20) harnesses the distributional information about $P(t)$ incorporated in the variation of the $C_t(T,X)$ across X. Details of how (20) is estimated using a finite number of strike prices are given in Section 4.1.

In the case of the individual DJIA stocks analysed in Section 4, on which American options are written, we continue to use the MF formula in (20) as a method for extracting information from the full spectrum of observed option prices. As noted earlier, the error associated with this approximation can be viewed as contributing to measurement error in the forecast. The ATM forecasts for the individual stocks are extracted from published option-market data in a manner described in Section 4.1, rather than via the inappropriate BS formula in (19).

3.3 Evaluation of Volatility Forecasts: Superior Predictive Ability (SPA) Testing

The forecast evaluation involves the assessment of multiple GARCH-type specifications for daily returns, ARFIMA (and ARMA) specifications for the realized measures based

on the intraday returns, and option-implied volatility forecasts. The assessment is to be performed for each of the eight volatility proxies, as measures of the latent, or actual, volatility quantity of interest, denoted by V_t^2. When the true latent price follows the model in (1), $V_t^2 = QV_t$. Only one proxy, BV, is consistent for IV_t when the true process contains random jumps.

For each proxy, alternative forecasts are compared with an option-implied benchmark using the SPA test of Hansen (2005) and Hansen and Lunde (2005a). Denoting by $\widehat{V}_t^2$ the realized proxy for the latent volatility at time t, and $f_{j,t}$ as the forecast of V_t^2 produced by the jth model (or forecast method), $j = 0, 1, 2, \ldots, m$, the SPA test is conducted via the following steps:

1. Based on rolling samples of fixed length R, $m + 1$ forecasts are produced for an evaluation period, $t = 1, 2, \ldots, N$.

2. Associated with each forecast method is a sequence of losses, $L_{j,t} = L(\widehat{V}_t^2, f_{j,t})$, $t = 1, 2, \ldots, N$. With $j = 0$ denoting the benchmark forecast, all m alternative forecasts are compared with the benchmark via the time series of loss differentials, $D_{j,t} = L_{0,t} - L_{j,t}$, $j = 1, 2, \ldots, m$, $t = 1, 2, \ldots, N$.

3. A test of whether or not the benchmark model is outperformed by any other model is conducted by testing $H_0 : E(D_{j,t}) \leq 0$ for all $j = 1, 2, \ldots, m$ against $H_A : E(D_{j,t}) > 0$ for at least one $j = 1, 2, \ldots, m$, using the test statistic $SPA = \max_{j=1,2,\ldots,m} \frac{\sqrt{N}\overline{D}_j}{\widehat{\omega}_{jj}}$, where $\overline{D}_j = \frac{1}{N}\sum_{t=1}^{N} D_{j,t}$ and $\widehat{\omega}_{jj}$ is a consistent estimator of $\omega_{jj} = \lim_{n\to\infty} var(\sqrt{N}\overline{D}_j)$, $j = 1, 2, \ldots, m$.

In short, a large value for the SPA test statistic represents evidence against the null hypothesis and indicates that at least one model in the model set significantly outperforms the benchmark model. As detailed clearly in Hansen (2005) and Hansen and Lunde (2005a), the null distribution of the test statistic needs to be approximated numerically, the bootstrap method used to this end taking into account the time series dependence in the loss differentials. The p-value associated with the observed test statistic is calculated as the proportion of times the bootstrap draws produce a statistic that exceeds the observed value. Given the need to recentre the bootstrap draws around the true (but unobserved) value of $E(D_{j,t})$, alternative p-values are produced corresponding to alternative estimates of $E(D_{j,t})$. In the empirical section we report results based the estimated p-value that is consistent for the true p-value.

Crucially, this test procedure caters explicitly for the *multiple* models included in the comparison. Hence, the results are not subject to the criticism of data-mining, whereby a

sequence of pair-wise comparisons between a benchmark model and any set of comparators has a high probability of leading to incorrect rejection of a true null due to an implicit inflation of the size associated with the overall procedure.[18]

4 Empirical Analysis Using U.S. Stock Market Data

4.1 Computational Details

The numerical analysis is performed using equity and option data for IBM, GE and MSFT over the ten year period from 30 June, 1996 to 30 June, 2006. Results are also produced for the S&P500 index using data over the same period, but with only the RV(5) and BV measures used as forecast variables of interest. All equity data has been supplied by the Securities Industries Research Centre of Asia Pacific (SIRCA) on behalf of Reuters, with the raw data then cleaned using the methods of Brownless and Gallo (2005). The VIX data is extracted from the CBOE website (www.cboe.com). All ATM, BS and MF calculations are based on the implied volatility surface data provided by IVOLATILTY (www.ivolatility.com). The surface data consists of implied volatilities for options with values of moneyness (X/P_t) ranging from 0.5 to 1.5 in steps of 0.1, and with varying times to maturity. The raw option data from which the surface is constructed is end-of-day out-of-the-money (OTM) put and call quote data.[19] For the individual stocks, we take as our estimate of ATM volatility (denoted by $\widehat{ATM}$), the value on the surface associated $X/P_t = 1$ and one month (22 trading days) to maturity. For the S&P500 index, on which European options are written, the corresponding value on the surface is taken as an estimate of BS volatility (denoted by $\widehat{BS}$).

Given maximum and minimum strike values $X_{\max}$ and $X_{\min}$ respectively, the estimate of MF implied volatility in (20) is given by

$$\begin{aligned}\widehat{MF} &= E_t^* \left[\int_t^T \sigma^2(s) ds \right] \approx 2 \int_{X_{\min}}^{X_{\max}} \frac{C_t(T,X) e^{i_t(T-t)} - \max\left[0, P_t^{(D)} e^{i_t(T-t)} - X\right]}{X^2} dX \\ &\approx \sum_{j=1}^{M} \left[g(T,X_j) + g(T,X_{j-1}) \right] \Delta X, \end{aligned} \tag{21}$$

[18] See Hsu (1996), White (2000), Sullivan, Timmermann and White (2003) and Romano and Wolf (2005) for other size-controlled multiple comparison tests. Other approaches to forecast evaluation include Granger and Pesaran (2000), Giacomini and White (2006), Hansen, Lund and Nason (2003), Giacomini and Komunjer (2005) and Corradi and Swanson (2006).

[19] For American options a binomial tree method is used, while the Black-Scholes model is used to produce the implied volatilities for European options. For more details on the construction of the surface, see
http://www.ivolatility.com/doc/IVolatility_Data_detailed.pdf.

where $\Delta X = (X_{\max} - X_{\min})/M$, $X_j = X_{\min} + j\Delta X$ for $0 \leq j \leq M$ and $g(T, X_j) = (C_t(T, X_j)e^{i_t(T-t)} - \max\left[0, P_t^{(D)}e^{i_t(T-t)} - X_j\right])/X_j^2$. Given the finite number of points on the moneyness spectrum of the IVOLATILITY surface, a procedure similar to that used by Jiang and Tian (2005) is adopted, with steps as follows: 1) Extract the IVOLATILTY one-month implied volatilities for the available range of moneyness values: $0.5 < X/P_t < 1.5$ in steps of 0.1[20]; 2) Use linear interpolation between these values to produce a smooth function of implied volatilities and use this function to extract implied volatilities at the M grid points X_j; 3) Use the BS model in (19) to translate the X_j into 'observed' prices $C_t(T, X_j)$; 4) Use the full set of M X_j and $C_t(T, X_j)$ values to estimate MF integrated volatility as in (21).[21] The forecasts $\widehat{ATM}$, $\widehat{BS}$ and $\widehat{MF}$ all represent forecasts of volatility over the next 22 trading days (by construction), and thereby avoid the so-called 'telescoping' problem highlighted by Christensen *et al.,* 2001, amongst others.

Rolling one day ahead forecasts are produced for the period 29 August, 2001 to 30 May, 2006. Forecasts for 22 days ahead (one-month) are produced from the same starting point, but with the final date extended accordingly. The 22-day-ahead forecast is the average of the one-day-ahead, two-day-ahead, up to 22-day-ahead forecasts, with the average then expressed as an annualized figure. Correspondingly, the variance measure being forecast corresponds to the (annualized) average of the daily variance values over the forecast period. Each returns-based forecast is produced using both daily and intraday observations from $R = 1000$ days. The first year of observations (30 June, 1996 to 30 June, 1997) is used to set pre-sample values in the estimation of all long-memory models. All models are estimated using conditional maximum likelihood, with the infinite lag structure in the long memory models truncated at the lag determined by the number of sample observations plus the number of pre-sample observations.[22] Each option-implied

[20] Note that this curve itself has been produced via an initial interpolation procedure given the quoted option prices for particular strikes.

[21] As pointed out by Jiang and Tian (2005), the BS model is simply being used as a mechanism to produce (artificially) a larger range of option prices than is available in practice, with the curve fitting procedure not requiring the BS model to be the 'true' model underlying the observed prices. That said, there is a slight inconsistency in the case of the American options, in that the artificial option prices are created using a formula (BS) that does not match that used to produce the initial implied volatility surface. Given that the IVOLATILITY surface is constructed from OTM put and call options only, one would not expect that a substantial premium for early exercise has been factored into the options. Hence, the mismatch between the inital prices used to construct the smile and the artificial, interpolated prices produced for use in (21) may not be too large. This also means that the prices used in (21) may not be too different from prices based on a European formula, and the approximation error in MF reduced accordingly.

[22] In the production of some of the rolling forecasts convergence problems occur, in particular for certain of the more highly parameterized GARCH-type models. When this occurs the models are re-estimated up to six times with different starting values each time. If the model still fails to converge then the forecasts for this date and model are marked as non-convergent. If a model produces only a few non-convergent

forecast is based on option prices observed on the day immediately prior to the forecast day (or period).

4.2 Empirical Results

4.2.1 SPA Tests of Option-Implied Forecasts for Individual Stocks

In this section we present the SPA test results for all three individual stocks, IBM, MSFT and GE, with both $\widehat{MF}$ and $\widehat{ATM}$ used as respective benchmarks. Comparative results for the S&P500 index are reported in Section 4.2.5. In the spirit of Hansen and Lunde (2006a) and Patton (2006) we use a 'robust' criterion, to measure the accuracy of forecast j, namely MSFE for variance quantities, with $L_{j,t} = \left[\widehat{V}_t^2 - f_{j,t}\right]^2$. We provide results for one and 22 days ahead in Table 1 and 2 respectively, with the maturity of the options used to construct $\widehat{MF}$ and $\widehat{ATM}$ matching the forecast horizon in the second case only. To aid in the interpretation of the large number of numerical results, in each table we group the eight measures, and associated results, according to the way in which the different volatility measures accommodate noise and/or jumps. Specifically we define: I. Measures that do not *formally* adjust for noise or jumps (No ADJ): RV(5) and RVAV(5); II. Measures that adjust for noise only (NOISE_ADJ): TSRV, TSRV2, RKERN, OSRV and ALTM; and III. The measure that adjusts for both noise and jumps (NOISE and JUMPS_ADJ): BV. We annotate the results in the following way: i) if a benchmark is not rejected at the 5% level, the SPA p-value appears in bold; ii) if a benchmark is not rejected *and* its MSFE loss is the smallest of that of all $m+1$ models in the choice set, the bolded p-value is allocated a # superscript; iii) in the case where either the $\widehat{MF}$ or $\widehat{ATM}$ benchmark is rejected, the 'most significant' forecast model according to the pair-wise 't statistics' is indicated by a superscript.[23]

The results in Table 1 provide little evidence that the MF implied volatility is an accurate forecast of actual volatility one day ahead. For IBM the SPA test rejects at the 5% level for all *eight* measures of volatility. In all cases, $\widehat{ATM}$ is the most 'significant' alternative, as based on the individual pair-wise 't statistics'. For MSFT and GE there is support for $\widehat{MF}$ using the ALTM measure, and a small amount of support in the case of GE using the RKERN measure also; however, in all other cases the $\widehat{MF}$ benchmark is rejected, with $\widehat{ATM}$ again the most 'significant' alternative in many instances. Both

forecasts then we simply remove these days from the out-of-sample dataset; however if a large number of days for a particular model are non-convergent then we remove that particular model from the model set used in the SPA test.

[23] It is important to remember that the 'most significant' forecast model is not necessarily the model with the smallest MSFE loss. Also, most importantly, the 'most signficant' alternative according to the pair-wise comparisons may itself be rejected as a benchmark model using the SPA test.

Table 1:

SPA p-values: forecasts based on a one-day-ahead forecast horizon. An option-implied volatility forecast is used as benchmark: $\widehat{MF}$ (model free) and $\widehat{ATM}$ (at-the-money). The SPA test is based on a mean squared forecast error (MSFE) loss criterion, for variance quantities. For each data set the number of models against which the benchmark model is compared (m), plus the number of observations in the forecast evaluation period from which the p- values and sample loss are calculated (N) are as follows: IBM: $m = 67$; $N = 1149$; MSFT: $m = 63$; $N = 1154$; GE: $m = 66$; $N = 1147$.

	IBM		MSFT		GE	
Benchmark:	$\widehat{MF}$	$\widehat{ATM}$	$\widehat{MF}$	$\widehat{ATM}$	$\widehat{MF}$	$\widehat{ATM}$
I. No ADJ						
RV(5)	$0.000^{(ATM)}$	**0.301**	$0.014^{(ATM)}$	**0.472**	$0.003^{(ATM)}$	**0.934**$^{\#}$
RVAV(5)	$0.000^{(ATM)}$	**0.207**	$0.012^{(ATM)}$	**0.474**	$0.000^{(ATM)}$	**0.941**$^{\#}$
II. NOISE_ADJ						
OSRV	$0.000^{(ATM)}$	**0.233**	$0.006^{(ATM)}$	**0.485**	$0.000^{(ATM)}$	**0.935**$^{\#}$
RKERN	$0.000^{(ATM)}$	**0.346**	$0.014^{(LMown)}$	**0.471**	**0.075**	**0.740**
TSRV1	$0.000^{(ATM)}$	**0.238**	$0.001^{(SMcross)}$	**0.306**	$0.041^{(LMcross)}$	**0.560**
TSRV2	$0.000^{(ATM)}$	**0.217**	$0.001^{(LMcross)}$	**0.337**	$0.015^{(ATM)}$	**0.626**
ALTM	$0.011^{(ATM)}$	**0.746**$^{\#}$	**0.364**	**0.382**	**0.121**	**0.404**
III. NOISE and JUMPS_ADJ						
BV	$0.000^{(ATM)}$	$0.002^{(SMcross)}$	$0.025^{(ATM)}$	**0.502**	$0.000^{(ATM)}$	**0.804**

(ATM): In this case, when the $\widehat{MF}$ benchmark is rejected, the $\widehat{ATM}$ forecast is the 'most significant' according to the pair-wise 't statistics'.

(LMown (cross)). In this case, when the $\widehat{MF}$ benchmark is rejected, a long-memory own (cross) forecast is the 'most significant' according to the pair-wise 't statistics'.

(SMcross). In this case, when the $\widehat{MF}$ or $\widehat{ATM}$ benchmark is rejected, a short-memory cross forecast is the 'most significant' according to the pair-wise 't statistics'.

$\#$ indicates that $\widehat{ATM}$ has the smallest MSFE loss of all $m + 1$ models in the choice set.

Table 2:

SPA p-values: forecasts based on a 22-day-ahead forecast horizon. An option-implied volatility forecast is used as benchmark: $\widehat{MF}$ (model free) and $\widehat{ATM}$ (at-the-money). The SPA test is based on a mean squared forecast error (MSFE) loss criterion, for variance quantities. For each data set the number of models against which the benchmark model is compared (m), plus the number of observations in the forecast evaluation period from which the $p-$ values and sample loss are calculated (N) are as follows: IBM: $m = 67$; $N = 1149$; MSFT: $m = 63$; $N = 1154$; GE: $m = 66$; $N = 1147$.

	IBM		MSFT		GE	
Benchmark:	$\widehat{MF}$	$\widehat{ATM}$	$\widehat{MF}$	$\widehat{ATM}$	$\widehat{MF}$	$\widehat{ATM}$
I. No ADJ						
RV(5)	$0.000^{(ATM)}$	**0.117**	$0.000^{(ATM)}$	**0.571**$^{\#}$	$0.000^{(ATM)}$	**0.964**$^{\#}$
RVAV(5)	$0.000^{(ATM)}$	$0.045^{(LMcross)}$	$0.000^{(ATM)}$	**0.961**$^{\#}$	$0.000^{(ATM)}$	**0.960**$^{\#}$
II. NOISE_ADJ						
OSRV	$0.000^{(ATM)}$	$0.043^{(LMcross)}$	$0.000^{(ATM)}$	**0.934**$^{\#}$	$0.000^{(ATM)}$	**0.946**$^{\#}$
RKERN	$0.000^{(ATM)}$	**0.208**	$0.000^{(ATM)}$	**0.585**$^{\#}$	$0.004^{(ATM)}$	**0.984**$^{\#}$
TSRV1	$0.000^{(ATM)}$	**0.060**	$0.000^{(ATM)}$	**0.870**$^{\#}$	$0.001^{(ATM)}$	**0.956**$^{\#}$
TSRV2	$0.000^{(ATM)}$	**0.052**	$0.000^{(ATM)}$	**0.940**$^{\#}$	$0.000^{(ATM)}$	**0.927**$^{\#}$
ALTM	$0.000^{(ATM)}$	**0.844**$^{\#}$	$0.038^{(ATM)}$	**0.579**$^{\#}$	**0.057**	**0.993**$^{\#}$
III. NOISE and JUMPS_ADJ						
BV	$0.000^{(ATM)}$	$0.000^{(LMcross)}$	$0.000^{(ATM)}$	**0.592**$^{\#}$	$0.000^{(ATM)}$	**0.816**$^{\#}$

(ATM): In this case, when the $\widehat{MF}$ benchmark is rejected, the $\widehat{ATM}$ forecast is the 'most significant' according to the pair-wise 't statistics'.

(LMcross). In this case, when the $\widehat{ATM}$ benchmark is rejected, a long-memory own cross forecast is the 'most significant' according to the pair-wise 't statistics'.

$\#$ indicates that $\widehat{ATM}$ has the smallest MSFE loss of all $m + 1$ models in the choice set.

long-memory and short-memory direct forecasts also feature in some cases.

Whilst the lack of support for the $\widehat{MF}$ benchmark may, superficially, be unsurprising, given the mismatch between option maturity (22 trading days) and forecast horizon (1 day), the results for the $\widehat{ATM}$ benchmark provide a startling refutation of the maturity explanation. In all but one case (the BV measure for IBM) $\widehat{ATM}$ is accepted as a superior forecast, with the p-values all exceeding 0.2, usually well and truly so. In four cases the $\widehat{ATM}$ is not only not rejected as benchmark, but also has the smallest MSFE loss of all models considered.

Most importantly, given one of the main focusses of this paper, apart from the odd exception and some variation in the magnitudes of the p-values, these qualitative results - strong support for $\widehat{ATM}$ and lack of support for $\widehat{MF}$ - are almost completely invariant to the measure used to proxy future volatility. This result is consistent with the robustness results reported by Ghysels and Sinko (2006), in the context of a more limited forecasting analysis of direct intraday returns-based forecasts. The only result that really stands out here is the inability of $\widehat{ATM}$ to forecast the "jump-free" BV measure for IBM, a result that contrasts with all other results in the table related to this benchmark. It is also worth noting that none of the results that indicate some support for $\widehat{MF}$ are associated with the BV measure.

Given the particular maturity associated with the option-implied forecasts - 22 trading days - one would anticipate an improved performance when the forecast horizon matches that maturity. As indicated by the results reported in Table 2, for the $\widehat{ATM}$ forecast of MSFT and GE volatility this is indeed the case, with the p-values for the $\widehat{ATM}$ benchmark uniformly higher for the 22 day forecast horizon than the corresponding p-values for the one day horizon, and close to one in many cases. Moreover, and as is not surprising given the strength of the test results for $\widehat{ATM}$, the latter forecast has the lowest MSFE for *all eight* forecast variables, for both series. The results for IBM are less clear-cut, although there is still support for the benchmark $\widehat{ATM}$ for the majority of forecast variables. In contrast, the results for the $\widehat{MF}$ benchmark are even weaker at the longer horizon, with only a single failure to reject $\widehat{MF}$ as the superior forecast, across all series and all measures, and that support for $\widehat{MF}$ being only marginal (p-value = 0.057). Once again, both option-implied volatilities fail to successfully predict the BV measure for IBM. Moreover, the p-value for the BV measure for MSFT, although supportive of the $\widehat{ATM}$ benchmark, is smaller than the majority of p-values for the other measures. In the case of GE the BV p-value is smaller than the p-values for *all* other measures.

As with the one-day-ahead predictions, there is some support for direct forecasts, in that for the three instances in which $\widehat{ATM}$ is rejected as the benchmark model, a

long memory direct forecast is the 'most significant' according to the pair-wise test. For the longer time horizon, short-memory direct forecasts do not feature at all. For neither forecast horizon is any support given to the GARCH-type forecasts based on daily returns. Indeed, although these figures are not reported here, this category of model is consistently ranked the lowest in terms of MSFE, for all series and measures, and for both forecast horizons.

In summary, the results of this section highlight a distinct contrast between the performance of the two alternative option-implied forecasts, $\widehat{MF}$ and $\widehat{ATM}$. They also give some support to the idea that both option-implied forecasts factor in jump information and thus do less well at forecasting the BV measure, in which such information has, in principle, been eliminated. The results do not support the proposition that $\widehat{MF}$, as an forecast of quadratic variation, forecasts those measures that include jump variation better than does $\widehat{ATM}$. For *no* measure, and for no series, is the support for $\widehat{MF}$ as benchmark stronger than the corresponding support for $\widehat{ATM}$.

In the following section we attempt to shed some light on the contrasting performances of $\widehat{MF}$ and $\widehat{ATM}$ via an examination of the option market information from which the options-based forecasts have been extracted. In Section 4.2.3 we shed further light on the issue via reference to the analysis in Bollerslev and Zhou (2006) and Bollerslev, Gibson and Zhou (2006) of the volatility risk premium.

4.2.2 Implied Volatility Curves

In Figure 1, Panels (a), (c) and (e) we plot one particular volatility measure, OSRV, for each series, against $\widehat{MF}$.[24] In the right-hand panels, (b), (d) and (f) respectively, we plot $\widehat{MF}$ against $\widehat{ATM}$ for each series. The intraday measure reported is for the 22-day-ahead forecast horizon and all volatility measures (both realized and option-implied) are graphed as annualized standard deviation figures.[25] Four features in Figure 1, common to all three series, are immediately apparent: 1) There are two distinct sub-periods: a high-volatility period from 28 August, 2001 to (approximately) 30 July, 2004, and a lower volatility period from 2 August, 2004 to 30 May, 2006;[26] 2) The $\widehat{MF}$ forecast tends to exceed realized volatility (overall), and by a greater amount in the high- than in the low-volatility period; 3) The $\widehat{MF}$ forecast tends to exceed the $\widehat{ATM}$ forecast, again by a greater amount in the high volatility period; 4) The $\widehat{MF}$ forecast is excessively noisy,

[24] Qualitatively similar results are produced for the other measures of volatility.

[25] All graphs in the paper present annualized standard deviation quantities in order to enable easy visual comparisons with the types of volatility graphs that usually appear in this literature.

[26] For the purposes of this illustration we omit the last 44 observations from the second MSFT subsample so that this second sub-period is accurately described as 'low-volatility'.

Table 3:

Summary statistics for the two option-implied forecasts, over the full sample and the high- and low-volatility sub-periods; realized volatility measured by OSRV.

	IBM		MSFT		GE	
Forecast:	$\widehat{MF}$	$\widehat{ATM}$	$\widehat{MF}$	$\widehat{ATM}$	$\widehat{MF}$	$\widehat{ATM}$
	Full sample period (28 August, 2001 to 30 May, 2006)					
$\widehat{E}(\widehat{V}_t^2 - f_t)$	-0.0611	-0.0389	-0.0457	-0.0227	-0.0418	-0.0178
$\widehat{var}(f_t)$	0.0061	0.0038	0.0107	0.0060	0.0079	0.0046
	High-volatility sample period (28 August, 2001 to 30 July, 2004)					
$\widehat{E}(\widehat{V}_t^2 - f_t)$	-0.0824	-0.0507	-0.0711	-0.0337	-0.0589	-0.0202
$\widehat{var}(f_t)$	0.0068	0.0044	0.0108	0.0060	0.0080	0.0049
	Low-volatility sample period (2 August, 2004 to 30 May, 2006)					
$\widehat{E}(\widehat{V}_t^2 - f_t)$	-0.0282	-0.0204	-0.0198	-0.0188	-0.0153	-0.0142
$\widehat{var}(f_t)$	1.12e-004	8.81e-005	7.58e-005	7.12e-005	3.14e-005	2.89e-005

relative to realized volatility, and more so than is the $\widehat{ATM}$ forecast, again in the high-volatility period in particular.

The empirical features of OSRV, $\widehat{MF}$ and $\widehat{ATM}$, for all three series, and for the full sample period and both sub-periods identified here, are summarized in Table 3. Using $\widehat{V}_t^2$ to represent OSRV and setting $f_t = \widehat{MF}$, $\widehat{ATM}$ (as variance quantities), we report sample estimates of the forecasting bias, $E(\widehat{V}_t^2 - f_t)$ and the variance of the forecast, $var(f_t)$. The numerical results clearly support the informal graphical evidence: $\widehat{MF}$ are both a more biased forecast and a noisier one than $\widehat{ATM}$, in particular over the high-volatility period. Specifically, both the variance and the (magnitude of the) bias of $\widehat{MF}$ is approximately twice as large as the corresponding statistics for $\widehat{ATM}$ in the high volatility period. In the low volatility period, however, the corresponding bias and variance figures for both forecasts are much more similar, for MSFT and GE in particular. Both options-based forecasts *overestimate* actual volatility.

From the high- and low-volatility sub-periods we reproduce, in turn, a representative sequence of implied volatility curves from which both $\widehat{MF}$ and $\widehat{ATM}$ have been con-

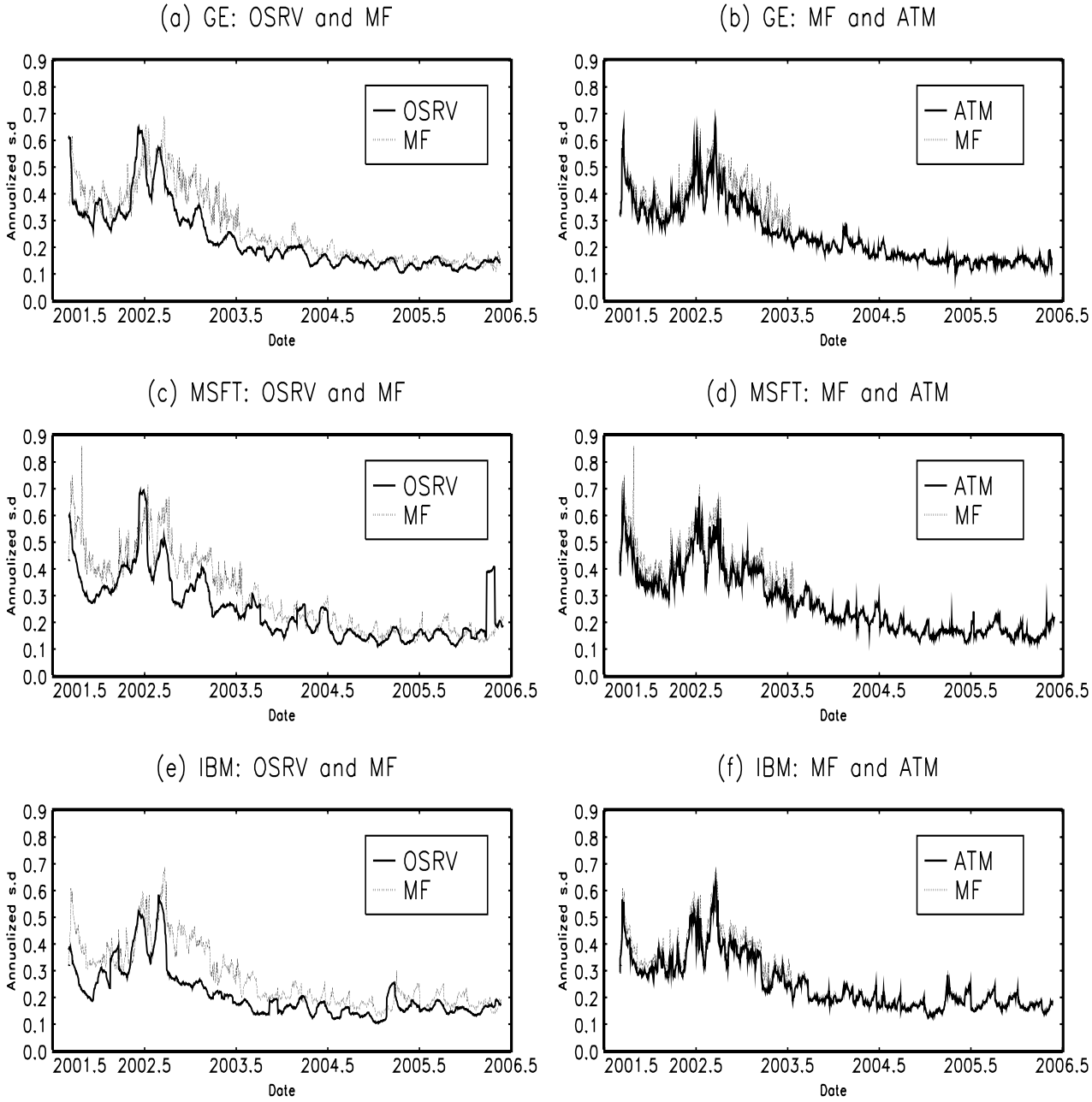

Figure 1: GE, MSFT and IBM Volatility (Annualized standard deviation): 29 August, 2001 to 30 May, 2006.

structed, as per the explanation in Section 4.1. In Figure 2, all three curves, on each of four representative days from the high volatility period, give higher implied volatility figures for each moneyness ratio, when compared with the comparable curves for the low volatility period in Figure 3. Moreover, the former also exhibit a much more pronounced curvature than the latter, with the volatilities associated with very low values for X/P_t (and, in some instances, those associated with very high values for X/P_t) exceeding the near-the-money volatilities ($X/P_t \approx 1$) by a large amount. This pattern reflects, in turn, both the *existence* of quotes for OTM put options (X/P_t low) and OTM calls (X/P_t high), plus the assignment of high values to some of those options. In a high volatility state the market thus places high value on options that pay off only if the asset price either rises or falls by a large amount, i.e. only if the present high volatility state persists. A positive liquidity premium, associated with the relative lack of liquidity in far-from-the-money options, may also contribute to some of the high volatilities observed at the extreme ends of the moneyness spectrum. Only on one of the chosen days (17 May, 2002) do *all three* implied volatility curves display the downward sloping skew pattern that is often a feature of equity option data.

Given that $\widehat{ATM}$ is equated to the ordinate of the volatility curve at $X/P_t = 1$, and $\widehat{MF}$ constructed from a formula that uses *all* ordinates, the reason why $\widehat{MF}$ tends to exceed $\widehat{ATM}$ by a large amount in the high-volatility period is clear. In addition, an examination of the sequence of implied volatility curves over the entire high-volatility period, of which the graphs in Figure 2 provide a snapshot, highlights a large degree of variation in the away-from-the-money volatilities in particular, a feature that contributes to the large variation in $\widehat{MF}$ reported in Table 3. Again, this noise is likely to be exacerbated by the lack of liquidity in the away-from-the-money options.

In contrast to the rather distinct smile shape that characterizes some of the curves in Figure 2, during the low volatility period higlighted in Figure 3, skewed curves, mostly with the typical negative slope, are more in evidence, with much less variation exhibited across the moneyness spectrum. The flat curves beyond certain narrow ranges around $X/P_t = 1$ indicate that no quotes on away-from-the-money options are made at the end of the relevant day, with the implied volatilities at these boundary points simply being extrapolated to the outer boundaries of 0.5 and 1.5; see Jiang and Tian (2005). In the low volatility state, options that have positive pay-offs only if P_t varies substantially from its current value, i.e. if volatility is high over the maturity of the option, are not traded. In this case, there is much less difference between the $\widehat{MF}$ and $\widehat{ATM}$ values, plus much less variation in the $\widehat{MF}$ values, than during the high volatility state.

In summary, close examination of the volatility smile information from which $\widehat{MF}$

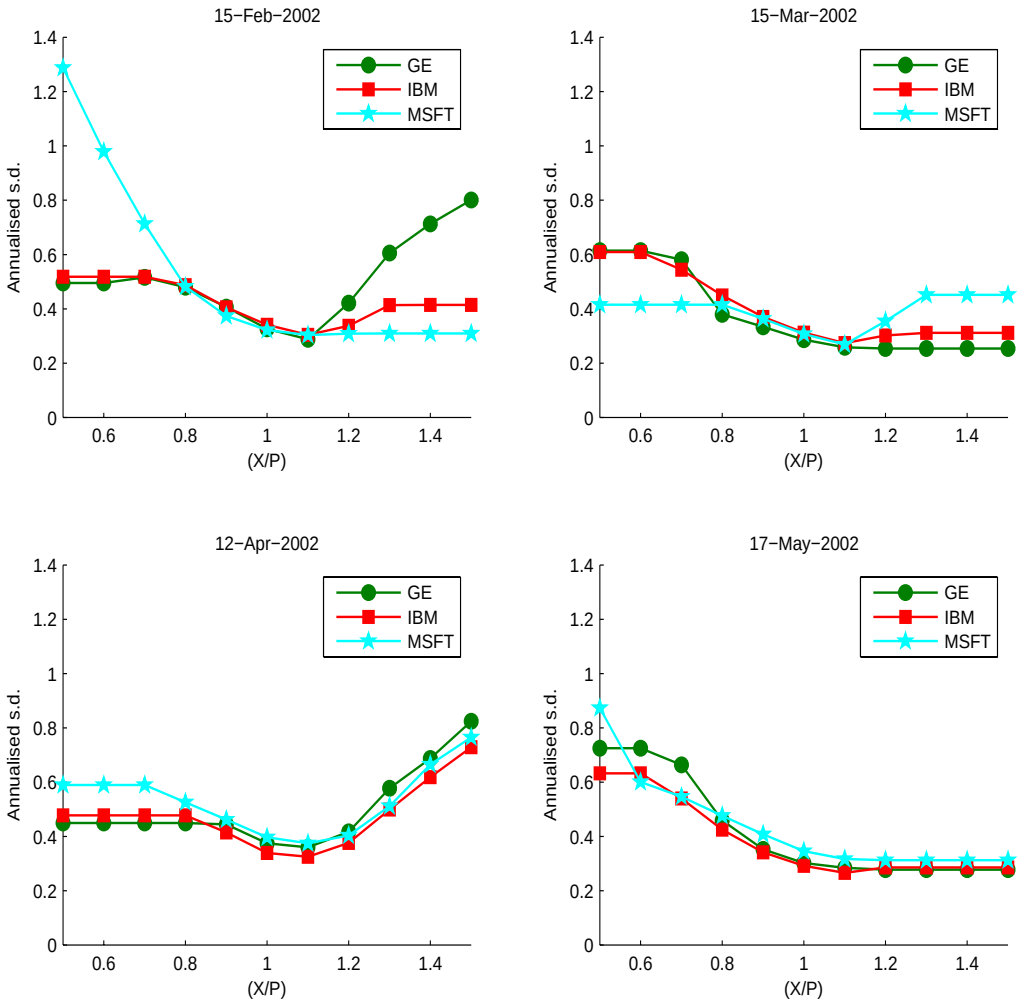

Figure 2: Implied volatility curves for representative days on four sequential months during the high-volatility period. Volatility is represented as an annualized standard deviation figure.

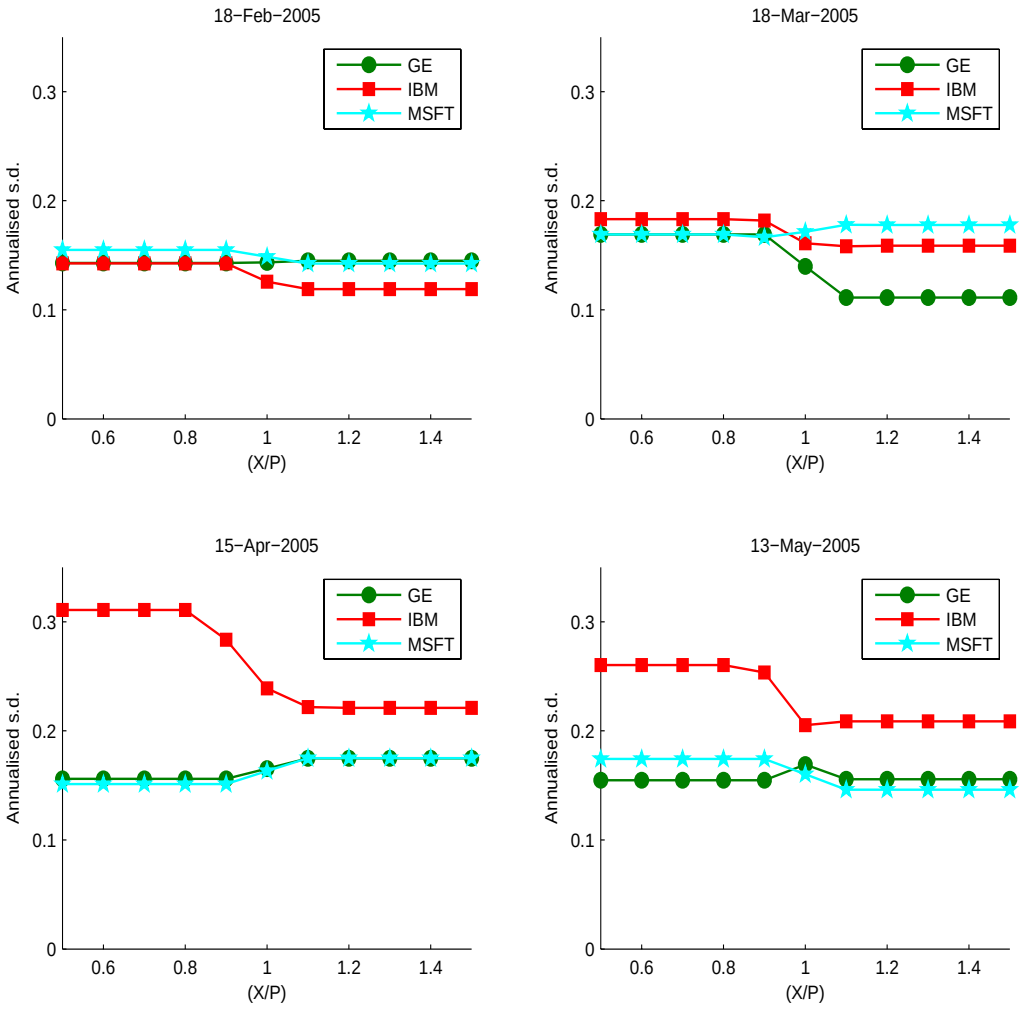

Figure 3: Implied volatility curves for representative days on four sequential months during the low-volatility period. Volatility is represented as an annualized standard deviation figure.

and $\widehat{ATM}$ are extracted provides some explanation for both the discrepancy between the two measures and for the added variability in the $\widehat{MF}$ measure, in particular in times of high volatility. The excessive variability in $\widehat{MF}$ would only be exacerbated by any measurement error associated with the application of the MF formula to the American-style option data. In the following section we draw upon the insights of Bollerslev and Zhou (2006) and Bollerslev *et al.* (2006) in order to provide an explanation for the *positive* bias in both measures and for the fact that the magnitude of that bias is larger in the high volatility period.

4.2.3 Forecasting Bias: Implied Volatility Risk Premium

Bollerslev and Zhou (2006) demonstrate that under the assumption of the square root stochastic volatility model of Heston (1993), the coefficients in the regression,

$$\int_t^T \sigma^2(s)ds = \phi_0 + \phi_1 E_t^* \left[\int_t^T \sigma^2(s)ds \right] + e_{t,T} \tag{22}$$

are functions of the parameters of the risk-neutralized version of the distribution with respect to which $E_t^*(.)$ in (22) is defined. We refer readers to Bollerslev and Zhou for details of the objective and risk-neutral distributions in question and the links between them. It is sufficient to note here that for standard values of the objective parameters, the *negative* market price of volatility risk that is observed empirically (e.g. Guo 1998, Eraker, 2004, Forbes, Martin and Wright, 2007) leads unambiguously to $\phi_1 < 1$. Translated into the option context, the negative price means that the risk-neutralized distribution for volatility reverts more slowly to a higher long-run mean, in comparison with the objective distribution. That is, option prices have a *positive* premium factored in, as a consequence of stochastic volatility. It is this positive premium that leads to the implied volatility measure exceeding, on average, the objective measure of volatility, with the bias in the forecasting regression in (22) being a manifestation of the deviation between the two forms of volatility. As Bollerslev and Zhou demonstrate via simulation experiments, this qualitative result is unaffected by the estimation of $\int_t^T \sigma^2(s)ds$ using observed intraday returns. The empirical results reported in the previous section, in which both option-implied forecasts have positive bias with respect to one particular estimate of $\int_t^T \sigma^2(s)ds$, namely OSRV, support this finding.[27]

The assumption of an underlying stochastic volatility process for returns is completely consistent with the implied volatility patterns observed in practice, including for the data analysed here. That is, implied volatility smiles/skews can be linked to the fat tails

[27] Again, the same qualitative results, although not reported, were obtained for the other realized measures.

(and/or skewness) that characterize empirical returns, characteristics that, in turn, can be associated with a stochastic volatility process (see, for e.g. Heston, 1993, Bakshi, Cao and Chen, 1997, and Bates, 2000). The particular shape of the implied volatility curve can be linked to features of the underlying stochastic volatility process, most notably the degree of volatility of volatility and the magnitude (and sign) of the instantaneous correlation between volatility and returns. The varying shapes observed over the sample period considered are suggestive of an underlying stochastic volatility model with time-varying parameters, although we attempt no formal investigation here of that observation. Certainly, the varying degree of bias, in particular between the high and low volatility periods, is indicative of a time-varying risk premium that is a positive function of the level of actual volatility. This empirical feature is consistent with the analysis in Bollerslev *et al.* (2006), in which the volatility risk premium is found to be a function of several macro-finance state variables, including the observed level of volatility itself.

It is the $\widehat{MF}$ measure which is formally consistent with an underlying stochastic volatility models for returns and, hence, legitimately affected by any volatility risk premium via its method of calculation, whereby all available smile information is used. The $\widehat{ATM}$ forecast, on the other hand, approximated by an implied volatility at a single point in the moneyness spectrum, does not *formally* factor in a risk premium and, as a consequence, exhibits less bias as a forecast of actual volatility, as attested to by the results in Table 3.[28]

In summary then, any potential additional forecast accuracy associated with the added flexibility of the assumptions underlying the $\widehat{MF}$ forecast appears to be offset by the bias and noise which beset its calculation in practice. As such, it is of interest to ascertain whether or not a truncated version of $\widehat{MF}$, which retains some of the smile information, but not all, manages to outperform $\widehat{ATM}$. We investigate this in the following section by reporting SPA test results for three modified versions of $\widehat{MF}$.

4.2.4 SPA Tests of Truncated MF Forecasts

In Table 4 we present the SPA p-values associated with the 22-day-ahead forecasts using 5 benchmarks: $\widehat{MF}$ and $\widehat{ATM}$, plus three truncated versions of $\widehat{MF}$, denoted by: $\widehat{MF}(1.5)$, $\widehat{MF}(2.0)$ and $\widehat{MF}(2.5)$. The benchmark $\widehat{MF}(1.5)$, for example, is the estimate of MF produced from implied volatilities within the moneyness range: $1 + 1.5 \times \widehat{ATM}/\sqrt{12}$. The benchmarks $\widehat{MF}(2.0)$ and $\widehat{MF}(2.5)$ are defined correspondingly.[29] We produce the

[28] See also Bates (1996) for early discussion of the robustness of option-implied volatility based on at-the-money options.

[29] Jiang and Tian (2005) and Bollerslev, Gibson and Zhou (2006) also use truncation in calculating MF implied volatilities.

test results for the full sample period, as well as results for the low-volatility period identified in Section 4.2.2, the idea here being that the reduced bias and variation in all MF estimates in this latter period may lead to these benchmarks being given more support by the SPA test. The results for benchmarks $\widehat{MF}$ and $\widehat{ATM}$ are re-produced under the expanded model set in which $\widehat{MF}(1.5)$, $\widehat{MF}(2.0)$ and $\widehat{MF}(2.5)$ are included as alternatives. Hence, the results in the columns headed $\widehat{MF}$ and $\widehat{ATM}$ in Table 4 differ slightly from the corresponding results reported in Table 2. In order to reduce the number of results reported, we focus on only three measures for each series: ALTM, RKERN and BV.

For the full sample period, the truncation of the smile used to estimate the MF implied volatility does nothing to improve its forecast performance in the case of IBM. The $\widehat{MF}(1.5)$ benchmark is given limited support for GE and MSFT (for the ALTM volatility measure in particular). However, overall, the $\widehat{ATM}$ forecast remains dominant, even when the model set is expanded to include the added variants of $\widehat{MF}$.[30] For the low-volatility period, as would be anticipated from the results recorded in Table 3, the performance of both forms of option-implied forecasts ($\widehat{ATM}$, plus all variants of $\widehat{MF}$) is more similar, overall, than is their performance for the full period. However, rather than the performance of the MF forecasts improving when assessed over the low volatility period, *both* the $\widehat{ATM}$ and $\widehat{MF}-$ type forecasts are now rejected as benchmarks in virtually all cases! Only for a single measure (ALTM for the IBM and MSFT series), is there any support for an option-implied forecast. Once again, it is the BV measure which has the smallest p-values overall, with the majority being zero to three decimal places. As was the case for the earlier results, there is some support indicated for long-memory direct forecasts; however this observation would need to be formally verified by conducting SPA tests of long memory benchmarks.

4.2.5 SPA Tests for the S&P500 Index

The small amount of work that has assessed the forecasting performance of the MF implied volatility has done so without formal account being taken of any alternative forecasting models; see, for example, Jiang and Tian (2005) and Bollerslev and Zhou (2006). The analysis has also focussed on the volatility of the S&P500 Index, with the MF implied volatility being proxied by the VIX in the case of Bollerslev and Zhou. The results reported in Jiang and Tian, in which the MF method is explicitly compared with the

[30] Note, that the fact an $\widehat{MF}$ variant is often the most 'significant' alternative according to a pair-wise 't test' is not inconsistent with the fact that this same variant may be rejected as a benchmark by the SPA test. This result simply highlights one of the dangers of conducting pair-wise comparisons.

Table 4:

SPA p-values: forecasts based on a 22-day-ahead forecast horizon. Alternative option-implied volatility forecasts are used as benchmark: $\widehat{ATM}$, $\widehat{MF}(1.5)$, $\widehat{MF}(2.0)$ and $\widehat{MF}(2.5)$ and $\widehat{MF}$. The SPA test is based on a mean squared forecast error (MSFE) loss criterion, for variance quantities, with three alternative measures used the actual volatility: ALTM, RKERN and BV. The measure on which the SPA test is based is denoted in parentheses for each series listed in the first column of the table. Results are produced for the full sample and low-volatility periods.

Benchmark:	$\widehat{ATM}$	$\widehat{MF}(1.5)$	$\widehat{MF}(2.0)$	$\widehat{MF}(2.5)$	$\widehat{MF}$
	Full Sample Period (28 August, 2001 to 30 May, 2006)				
IBM (RKERN)	**0.208**	$0.000^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(MF1.5)}$	$0.000^{(MF2)}$
IBM (ALTM)	**0.844**$^{\#}$	$0.000^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(MF2)}$	$0.000^{(MF2)}$
IBM (BV)	$0.000^{(LMcross)}$	$0.000^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(MF1.5)}$	$0.000^{(MF1.5)}$
MSFT (RKERN)	**0.649**$^{\#}$	$0.001^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(ATM)}$	$0.000^{(ATM)}$
MSFT (ALTM)	**0.753**	**0.316**	$0.003^{(MF1.5)}$	$0.017^{(MF1.5)}$	$0.017^{(MF1.5)}$
MSFT (BV)	**0.634**$^{\#}$	$0.001^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(ATM)}$	$0.000^{(ATM)}$
GE (RKERN)	**0.945**$^{\#}$	**0.595**	$0.004^{(MF1.5)}$	$0.001^{(MF2)}$	$0.000^{(MF2.5)}$
GE (ALTM)	**0.216**	**1.000**	**0.088**	$0.004^{(MF2)}$	$0.000^{(MF2.5)}$
GE (BV)	**0.855**$^{\#}$	$0.023^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(MF2)}$	$0.000^{(MF1.5)}$
	Low Volatility Period (2 August, 2004 to 30 May, 2006)				
IBM (RKERN)	$0.028^{(LMcross)}$	$0.000^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(ATM)}$	$0.000^{(ATM)}$
IBM (ALTM)	**0.287**	$0.000^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(ATM)}$	$0.000^{(ATM)}$
IBM (BV)	$0.000^{(LMcross)}$	$0.000^{(ATM)}$	$0.000^{(MF1.5)}$	$0.000^{(ATM)}$	$0.000^{(ATM)}$
MSFT (RKERN)	$0.007^{(LMcross)}$	$0.005^{(ATM)}$	$0.001^{(MF1.5)}$	$0.001^{(MF1.5)}$	$0.001^{(MF1.5)}$
MSFT (ALTM)	**0.140**	**0.163**	**0.075**	**0.153**	**0.147**
MSFT (BV)	$0.000^{(LMcross)}$	$0.001^{(LMcross)}$	$0.001^{(LMcross)}$	$0.001^{(LMcross)}$	$0.001^{(LMcross)}$
GE (RKERN)	$0.009^{(LMcross)}$	$0.008^{(LMcross)}$	$0.009^{(MF1.5)}$	$0.013^{(MF1.5)}$	$0.009^{(MF1.5)}$
GE (ALTM)	$0.002^{(LMcross)}$	$0.003^{(LMcross)}$	$0.002^{(LMcross)}$	$0.003^{(LMcross)}$	$0.002^{(LMcross)}$
GE (BV)	$0.000^{(SMcross)}$	$0.000^{(SMcross)}$	$0.000^{(SMcross)}$	$0.000^{(SMcross)}$	$0.000^{(SMcross)}$

(ATM): In this case, when a benchmark is rejected, the $\widehat{ATM}$ forecast is the 'most significant' according to the pair-wise 't statistics'.

(MF*): In this case, when a benchmark is rejected, the $\widehat{MF}$*forecast is the 'most significant' according to the pair-wise 't statistics'.

(LM(SM)cross): In this case, when a benchmark is rejected, a long-(short-)memory cross forecast is the 'most significant' according to the pair-wise 't statistics'.

# indicates that $\widehat{ATM}$ has the smallest MSFE loss of all models in the choice set.

BS method, give some support to MF. This result is thus in conflict with our SPA test results, which cast doubt on the usefulness of the MF method in forecasting the volatility of individual stocks. It is of interest, therefore, to assess the robustness of our SPA-based conclusions to the shift from individual equities to the index, in particular given that the MF formula is designed for the European-style option data associated with the index. Given that the different forms of noise adjustments that have been used in this paper have their prime motivation in the case of data on traded assets, rather than observations on a constructed index, we conduct SPA tests of the S&P500 implied volatility measures for the case where actual volatility is measured by RV(5) and BV only.[31]

In Figure 4, Panels (a), (b), (c) and (d), we plot, respectively, RV(5) and $\widehat{MF}$, RV(5) and $\widehat{BS}$, $\widehat{MF}$ and VIX, and $\widehat{MF}(2.5)$ and VIX, for the 22-day-ahead forecast horizon. As is evident from Panels (a) and (b), both implied volatility forecasts are very biased, even more so than was the case with the individual stocks. This is consistent with a substantial risk premium being factored into the index options. Panel (c) demonstrates the accuracy with which the VIX reproduces the MF method, with the truncated $\widehat{MF}(2.5)$ being virtually indistinguishable from the CBOE measure in Panel (d). SPA-based tests of all five benchmarks used in the previous section were conducted, in addition to the test for the VIX benchmark. The tests were conducted over the full and low volatility periods. The results (not reported here) provide a resounding rejection of *all* implied volatility benchmarks, with *all* $p-$ values (to three decimal places) being equal to zero.

5 Summary and Conclusions

This paper presents the first empirical evaluation of option-implied versus returns-based volatility forecasts that takes into account all of the important recent developments regarding market microstructure noise. The options-based component of the analysis also accommodates the concept of model-free implied volatility, in an attempt to separate the forecasting performance of the option market from the issue of misspecification of the option pricing model. The testing framework properly caters for the existence of multiple alternative forecasts, as well as the sampling variability in estimated forecast loss, via use of the superior predictive accuracy test.

The model-free implied volatility performs poorly as a forecast of future volatility, with this conclusion applying to both individual equities and the S&P500 stock index. In contrast, volatility extracted from at-the-money options is given strong support as

[31] In order to retain comparability with the earlier results, we continue to construct the BV measure with the noise adjustment as per (16).

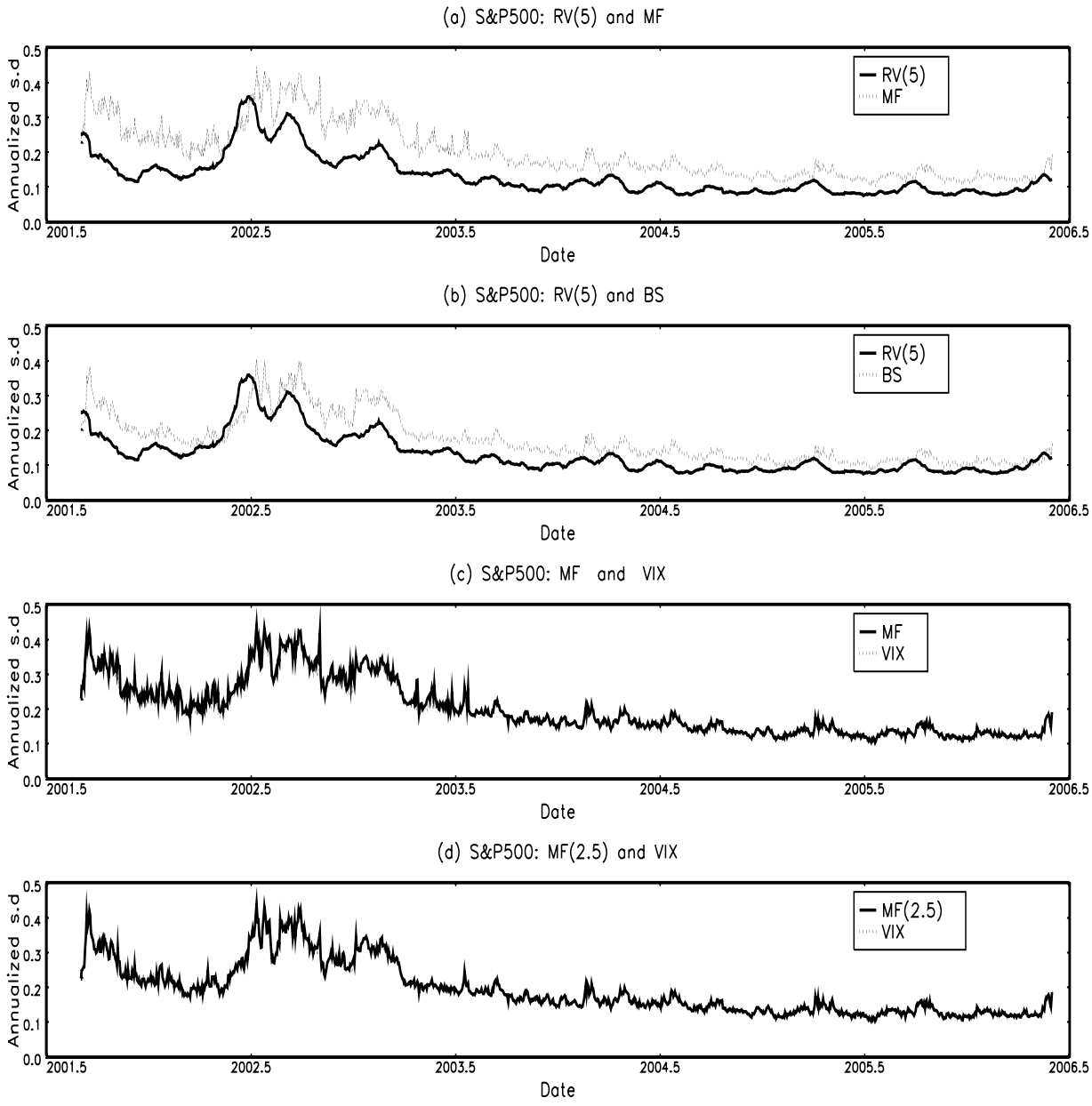

Figure 4: S&P500 Volatility (Annualized standard deviation): 29 August, 2001 to 30 May, 2006.

a superior forecast of individual stock volatility, in particular over a time horizon that matches the maturity of the options from which the implied volatility has been extracted. Like the model-free forecast, the at-the-money (Black-Scholes) forecast is rejected as a benchmark forecast in the case of the index. The qualitative results are, in the main, robust to the measure used to proxy future volatility. However, there is limited support for the idea that option prices do factor in jump information, given the slight tendency for both types of option-implied forecasts to do less well as a forecast of (noise-adjusted) bi-power variation. This observation requires more detailed investigation, however, before any definitive conclusions along these lines can be drawn.

The poor relative performance of the model-free implied volatility can be linked to both the bias and excess variability that it exhibits as a forecast of actual volatility, with the positive bias, in particular, being consistent with the option market factoring in a negative price for volatility risk. The at-the-money forecast, on the other hand, takes no account of the distributional information in the implied volatility patterns that characterize the option market. In so doing it can be viewed as missing vital information about the underlying asset price and its future volatility. It would appear, however, that this deficiency is more than offset by the reduction in forecast bias and variability that its more restrictive use of option market information entails.

Finally, some limited evidence has been produced that suggests that direct forecasts of realized volatility measures, based on long memory models, may also serve as useful forecasts of future volatility. In particular, it may be the case that certain measures used as the basis for producing forecasts may perform better than others, no matter what the variable (or measure) being forecast; i.e. that cross-forecasts may out-perform own-forecasts. An alternative exercise, in which the full range of alternative forecast models are ranked, rather than a particular benchmark model being assessed, could be implemented using the model confidence set methodology of Hansen and Lunde (2003). In so doing we could attempt to answer a different question from that addressed here: *which* form of model, or category of model, whether returns- or options-based, provides the best forecast of volatility?

References

[1] Ait-Sahalia, Y, Mykland, P.A. and Zhang, L. 2005. Ultra High Frequency Volatility Estimation with Dependent Noise, *National Bureau of Economic Research Working Paper 11380.*

[2] Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2003. Modelling and Forecasting Realized Volatility, *Econometrica,* 71: 579-625.

[3] Andersen, T.G., Bollerslev, T. and Meddahi, N. 2004. Analytical Evaluation of Volatility Forecasts, *International Economic Review,* 45: 1079-1110.

[4] Andersen, T.G., Bollerslev, T., Diebold, F.X. 2005. Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility, *National Bureau of Economic Research Working Paper 11775.*

[5] Andersen, T.G., Bollerslev, T. and Meddahi, N. 2005. Correcting the Errors: Volatility Forecast Evaluation using High Frequency Data and Realized Volatilities, *Econometrica,* 73: 279-296.

[6] Andersen, T.G., Bollerslev, T. and Meddahi, N. 2006. Realized Volatility Forecasting and Market Microstructure Noise, *Draft Paper.*

[7] Anderson, H.M. and Vahid, F. 2007. Forecasting the Volatility of Australian Stock Returns: do Common Factors Help?, *Journal of Business and Economic Statistics*, 25: 76 - 90.

[8] Bakshi, G., Cao, C. and Chen, Z. 1997. Empirical Performance of Alternative Option Pricing Models, *Journal of Finance,* 52: 2003-2049.

[9] Bandi, F.M. and Russell, J.R. 2005. Microstructure Noise, Realized Variance, and Optimal Sampling, *Draft Paper.*

[10] Bandi, F.M. and Russell, J.R. 2006. Separating Microstructure Noise from Volatility, *Journal of Financial Economics,* 79: 655-692.

[11] Bandi, F.M., Russell, J.R. and Yang, C. 2006. Realized Volatility Forecasting and Option Pricing. *Draft Paper.*

[12] Bandi, F.M., Russell, J.R. and Zhu, Y. 2006. Using High-frequency Data in Dynamic Portfolio Choice, Forthcoming in *Econometric Reviews.*

[13] Barndorff-Nielsen, O.E. and Shephard, N. 2002. Econometric Analysis of Realized Volatility and its Use in Estimating Stochastic Volatility Models, *Journal of the Royal Statistical Society B,* 64: 253-280.

[14] Barndorff-Nielsen, O.E. and Shephard, N. 2004. Power and Bipower Variation with Stochastic Volatility and Jumps, *Journal of Financial Econometrics,* 2: 1-37.

[15] Barndorff-Nielsen, O.E. and Shephard, N. 2005. Variations, Jumps, Market Frictions and High Frequency Data in Financial Econometrics, Forthcoming in *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress,* (eds. R. Blundell, P. Torsten and W. Newey), Econometric Society Monographs, Cambridge University Press.

[16] Barndorff-Nielsen, O.E., Lunde, A., Hansen, P.R. and Shephard, N. 2005. Realized Kernels can Consistently Estimate Integrated Variance: Correcting Realized Variance for the Effect of Market Frictions, *Draft Paper.*

[17] Barndorff-Nielsen, O.E., Lunde, A., Hansen, P.R. and Shephard, N. 2006a. Designing Realized Kernels to Measure the Ex-post Variation of Equity Prices in the Presence of Noise, *Draft Paper.*

[18] Barndorff-Nielsen, O.E., Lunde, A., Hansen, P.R. and Shephard, N. 2006b. Subsampling Realized Kernels, Empirical Appendix, *Draft Paper.*

[19] Barndorff-Nielsen, O.E., Lunde, A., Hansen, P.R. and Shephard, N. 2007. Subsampling Realized Kernels, *Draft Paper.*

[20] Bates, D. S. 1996. Testing Option Pricing Models, In Maddala, G.S. and Rao, C. R. (eds.), *Statistical Methods in Finance (Handbook of Statistics, v. 14).* Amsterdam: Elsevier Publishing.

[21] Bates, D.S. 2000. Post-87 Crash Fears in the S&P 500 Futures Option Market, *Journal of Econometrics,* 94: 181-238.

[22] Black, F. and Scholes, M., 1973 The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 81: 637-659.

[23] Blair, B.J, Poon, S-H. and Taylor, S.J. 2001. Forecasting S&P100 Volatility: the Incremental Information Content of Implied Volatilities and High Frequency Index Returns, *Journal of Econometrics*, 105, 5-26.

[24] Bollerslev, T. and Zhou, H. 2006. Volatility Puzzles: a Simple Framework for Gauging Return-Volatility Regressions, *Journal of Econometrics,* 131, 123-150.

[25] Bollerslev, T., Gibson M. and Zhou, H. 2006. Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities, *Working Paper, Division of Research and Statistics, Federal Reserve Board.*

[26] Britten-Jones, M. and Neuberger, A. 2000. Option Prices, Implied Price Processes and Stochastic Volatility. *The Journal of Finance,* LV: 839-866.

[27] Brownlees, C.T. and Gallo, G.M. 2005. Ultra High-Frequency Data Management, in *Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione* (ed. C. Provasi), Atti del Convegno S. Co., CLUEP, Padova.

[28] Busch, T., Christensen, B.J. and Nielsen, M.O. 2006. The Role of Implied Volatility in Forecasting Future Realized Volatility and Jumps in Foreign Exchange, Stock, and Bond Markets, *Working Paper, Centre for Analytical Finance, University of Aarhus.*

[29] Christensen, B.J., Hansen, C.S. and Prabhala, N.R. 2001. The Telescoping Overlap problem in Options Data, *Draft Paper.*

[30] Christensen B.J. and Prabhala, N.R. 1998. The Relation Between Implied and Realized Volatility, *Journal of Financial Economics,* 50: 125-150.

[31] Corradi, V. and Swanson, N.R. 2006. Predictive Density and Conditional Confidence Interval Accuracy Tests, *Journal of Econometrics,* 135: 187–228.

[32] Corrado, C.J. and Su, T. 1997. Implied Volatility Skews and Stock Index Skewness and Kurtosis implied by S&P 500 index Option Prices, *Journal of Derivatives*, Summer: 8-19.

[33] Day, T.E. and Lewis, C.M. 1995. Stock Market Volatility and the Information Content of Stock Index Options, in *ARCH, Selected Readings* (ed. R. Engle), Oxford University Press, Oxford.

[34] De Pooter, M., Martens, M. and van Dijk, D. 2006. Predicting the Daily Covariance Matrix for S&P100 Stocks using Intraday Data: But Which Frequency to Use? Forthcoming in *Econometric Reviews.*

[35] Eraker, B. 2004. Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices." *The Journal of Finance*, LIX: 1367-1403.

[36] Forbes C.S., Martin, G.M. and Wright J. 2007. Inference for a Class of Stochastic Volatility Models Using Option and Spot Prices: Application of a Bivariate Kalman Filter, *Econometric Reviews, Special Issue on Bayesian Dynamic Econometrics,* 26: 387-418.

[37] Ghysels, E. and Sinko, A. 2006. Comment on 'Realized Variance and Market Microstructure Noise', *Journal of Business and Economic Statistics*, 24: 192-194.

[38] Giacomini, R. and Komunjer, I. 2005. Evaluation and Combination of Conditional Quantile Forecasts, *Journal of Business and Economic Statistics*, 23: 416-431.

[39] Giacomini, R. and White, H. 2006, Tests of Conditional Predictive Ability, *Econometrica,* 74: 1545-1578.

[40] Granger, C.W.J. and Pesaran, M.H. 2000. A Decision-Theoretic Approach to Forecast Evaluation". In Chan, W.S., Li, W.K. and Tong, H. (Eds.), *Statistics and Finance: An Interface.* Imperial College Press, London.

[41] Guo, D. 1998. The Risk Premium of Volatility Implicit in Currency Options, *Journal of Business and Economic Statistics*, 16: 498-507.

[42] Hansen, P.R. 2005. A Test for Superior Predictive Ability, *Journal of Business and Economic Statistics*, 23: 365-380.

[43] Hansen, P.R. and Lunde, A. 2005a. A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?, *Journal of Applied Econometrics,* 20: 873-889.

[44] Hansen, P.R. and Lunde, A. 2005b. A Realized Variance for the Whole day Based on Intermittent High-Frequency Data, *Journal of Financial Econometrics,* 3: 525–554.

[45] Hansen, P.R and Lunde, A. 2006a. Consistent Ranking of Volatility Models, *Journal of Econometrics,* 131: 97-121.

[46] Hansen, P.R and Lunde, A. 2006b. Realized Variance and Market Microstructure Noise, *Journal of Business and Economic Statistics*, 24: 127-218.

[47] Hansen, P.R and Lunde, A. and Nason, J.M. 2003. Choosing the Best Volatility Models: The Model Confidence Set Approach, *Oxford Bulletin of Economics and Statistics,* 65: 839-861.

[48] Heston, S.L. 1993. A Closed-form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options, *The Review of Financial Studies*, 6: 327-343.

[49] Hsu, J.C. (1996), *Multiple Comparisons; Theory and Methods,* Chapman & Hall/CRC.

[50] Huang, X. and Tauchen, G. 2005. The Relative Contribution of Jumps to Total Price Variation, *Journal of Financial Econometrics*, 3: 456-499.

[51] Hull, J.C. (2000). *Options, Futures, and Other Derivative Securities*, 4rd ed., Prentice Hall, New Jersey.

[52] Koopman, S. J., Jungbacker, B. and Hol, E. 2005. Forecasting Daily Variability of the S&P100 Stock Index using Historical, Realized and Implied Volatility Measurements, *Journal of Empirical Finance,* 12: 445-475.

[53] Jiang, G.J. and Tian, Y.S. 2005. The Model-Free Implied Volatility and its Information Content, *The Review of Financial Studies*, 18: 1305-1342.

[54] Large, J. 2007. Estimating Quadratic Variation when Quoted Prices Jump by a Constant Increment, *University of Oxford, Draft Paper.*

[55] Lim, G.C., Martin, G.M. and Martin, V.L. 2005. Parametric Pricing of Higher Order Moments in S&P500 Options, *Journal of Applied Econometrics,* 20: 377-404.

[56] Martens, M. and Zein, J. 2004. Predicting Financial Volatility: High-Frequency Time Series Forecasts Vis-a-Vis Implied Volatility, *Journal of Futures Markets,* 24: 1005-1028.

[57] Oomen, R.C.A. 2006. Properties of Realized Variance under Alternative Sampling Schemes, *Journal of Business and Economic Statistics*, 24: 219-237.

[58] Neely, C.J. 2003. Forecasting Exchange Volatility: is Implied Volatility the Best we Can Do? *Working Paper, Federal Reserve Bank of St. Louis.*

[59] Patton, A. 2006. Volatility Forecast Comparison using Imperfect Volatility Proxies, *Research Paper 175, Quantitative Finance Research Centre, University of Technology, Sydney.*

[60] Pong, S., Shackleton, M.B. and Taylor, S.J. 2004. Forecasting Currency Volatility: a Comparison of Implied Volatilities and AR(FI)MA Models, *Journal of Banking and Finance,* 28: 2541-2563.

[61] Poteshman, A.M. 2000. Forecasting Future Volatility from Option Prices, *Department of Finance, University of Illinois Working Paper.*

[62] Romano, J.P. and Wolf, M. 2005. Stepwise Multiple Testing as Formalized Data Snooping, *Econometrica,* 73: 1237-1282.

[63] Sullivan, R., Timmermann, A. and White, H. 2003. Forecast Evaluation with Shared Data Sets, *International Journal of Forecasting,* 19: 217-227.

[64] White, H. 2000. A Reality Check for Data Snooping, *Econometrica* 68: 1097-1126.

[65] Zhang, L., Mykland, P.A. and Ait-Sahalia, Y. 2005. A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data, *Journal of the American Statistical Association,* 100: 1394-1411.