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

Gael M. Martin, Andrew Reidy and Jill Wright*

September 29, 2007

Abstract

This paper assesses the robustness of the relative performance of spot- and option-based volatility forecasts to the treatment of microstructure noise. Robustness of the results to the method of constructing option-implied forecasts is also investigated. Using a test for superior predictive ability, model-free implied volatility, which exploits information in the volatility ‘smile’, and at-the-money implied volatility, which does not, are both tested as benchmark forecasts of a range of alternative volatility proxies. The results provide compelling evidence against the model-free forecast for three Dow Jones Industrial Average stocks, over a 2001 to 2006 evaluation period. In contrast, the at-the-money implied volatility forecast is given strong support for the three equities over this period. Neither benchmark is supported for the S&P500 index. Importantly, the main qualitative results are invariant to the method of noise correction used in measuring future volatility.

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

Over the past decade or so, 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 volatility has almost exclusively been constructed from high-frequency intraday data, most commonly as the sum of squared returns over small, regular intervals. (See, for e.g., Poteshman, 2000, Blair, Poon and Taylor, 2001, Neely, 2003, Martens and Zein, 2004, Pong, Shackleton, Taylor and Xu, 2004, Jiang and Tian, 2005, and Koopman, Jungbacker and Hol, 2005). Studies that have adopted the realized volatility proxy have produced more definitive results, overall, than earlier work that used squared (or absolute) daily returns as the volatility measure (e.g. Day and Lewis, 1995). Nevertheless, conclusions have still been mixed, with the option-based forecasts sometimes deemed to be superior to those based on historical returns (e.g. Blair et al., 2001, Jiang and Tian, 2005) and sometimes not (e.g. Neely, 2003, and Martens and Zein, 2004). Most notably, studies that use intraday data in measuring future volatility have not formally adjusted for the empirical regularity of microstructure noise, nor documented the influence of this phenomenon on the conclusions drawn.1

The primary aim of this paper is to reassess the predictive content of option and spot prices by taking into account recent developments related to the measurement of volatility in the presence of microstructure noise. Specifically, the forecasting assessment is performed using a range of future volatility proxies, including several that formally adjust for noise. The relative accuracy of the option-implied and returns-based forecasts is tested using all proxies, with a view to gauging the robustness of the results to the treatment of noise.2

A secondary aim is to assess the robustness of the conclusions to the way in which the option-implied forecasts are extracted from option market data. 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 that of implied volatility forecasts extracted from at-the-money (ATM) market option prices.3 A link is drawn between the

---

1Jiang and Tian (2005) do make some adjustment to the conventional realized variance measure to accommodate the autocorrelation in intraday returns that arises due to noise in the price process.

2In 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); see also Blair et al. (2001) and Hansen and Lunde (2005a). Other related work, in which noise-corrected proxies are assessed in a range of different contexts, includes Andersen, Bollerslev and Meddahi (2006), Ghysels and Sinko (2006), Bandi, Russell and Yang (2006), Bandi, Russell and Zhu (2006) and De Pooter, Martens and van Dijk (2006).

3An 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.
performance of the MF quantity and both its interpretation as a risk-adjusted expectation of actual volatility (see also Bollerslev and Zhou, 2006) and its use of volatility ‘smile’ information across the spectrum of option strike prices.

To assess predictive performance, we use the test for superior predictive ability (SPA) of Hansen (2005) and Hansen and Lunde (2005a), with an option-implied volatility designated as the benchmark forecast. That is, we address the specific question of whether any forecast method out-performs an options-based forecast while taking appropriate account of the fact that multiple forecast models are legitimate competitors. Returns-based forecasts are produced both directly, via time series models for the volatility proxies themselves, and indirectly, via generalized autoregressive conditional heteroscedastic (GARCH)-type models for daily returns.

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 using the SPA test are addressed in Section 3. In Section 4, all aspects of the empirical investigation are outlined, including the details of the construction of the option-implied volatility measures. The forecasting assessment is conducted using intraday spot and option price data for three Dow Jones Industrial Average (DJIA) stocks - International Business Machines (IBM), Microsoft (MSFT) and General Electric (GE) - and the S&P500 index, over the 1996 to 2006 period. In the case of the index, on which European-style options are written, ATM volatility forecasts are produced via the Black-Scholes (Black and Scholes, 1973) option pricing model. The empirical results for the DJIA stocks provide strong evidence against the superiority of the MF benchmark, with this forecast’s performance adversely affected by both large positive bias and high forecast error variance. In contrast, over the full forecast evaluation period of August, 2001 to May, 2006, the ATM implied volatility is given a great deal of support as a benchmark forecast for all three DJIA series, with the ATM volatility incurring the smallest loss of all forecasts considered, in many cases. Only when attention is directed to a more recent low-volatility sub-period is the ATM benchmark often out-performed, with some form of long memory direct forecast being the ‘most significant’ alternative in this case. Both options-based forecasts are rejected as superior benchmarks in the case of the S&P500

---

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/).

5 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.
index, over both the full and low-volatility periods. Crucially, the main qualitative results are robust to the measure used to proxy future volatility, apart from some limited evidence suggesting 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,

\[
dp(t) = \mu(t)dt + \sigma(t)dW(t) + \kappa(t)dq(t), \quad t \geq 0,
\]

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), \( r_t = p(t) - p(t - 1) \), is then given by

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

That is, \( QV_t \) is equal to the sum of the integrated volatility of the continuous sample path component, \( IV_t = \int_{t-1}^{t} \sigma^2(s)ds \), and the sum of the \( g(t) - q(t - 1) \) squared jumps that occur over day \( t \). Denoting by \( p_i \) the \( i \)th logarithmic price that is observed on day \( t \), and \( r_{ti} = p_t - p_{t-1} \) as the \( i \)th transaction return, it is now well known (see, in particular, Barndorff-Nielsen and Shephard, 2002, and Andersen et al., 2003) that

\[
RV_t = \sum_{t_{i-1},t_i \in [t-1,t]} r_{ti}^2 \overset{P}{\rightarrow} QV_t
\]

where \( RV_t \) is referred to as realized volatility\(^6\).

Three comments can be made about the consistency result in (3). Firstly, the result in (3) 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.

\(^6\)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.
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 logarithmic price on day $t$, is $p_t = p_t^* + \varepsilon_t$, where $\varepsilon_t$ is assumed (at least initially) to be an i.i.d. white noise component. The $i$th observed transaction return, $r_t^i$, is thus given by the sum of the latent return, $r_t^* = p_t^* - p_{t-1}^*$, and a first order moving average (MA) process, $\eta_t = \varepsilon_t - \varepsilon_{t-1}$. It is straightforward to show (see Zhang et al. 2005) that $E(RV_t | p_t^*(t_i)) = \sum_{t_{i-1}, t_i \in [t-1,t]} r_t^2 + 2n\sigma^2_\varepsilon$, 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 $RV_t$ 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 $\hat{\sigma}^2_{\varepsilon} = \frac{1}{2n}RV_t$, Zhang et al. (2005) also demonstrate that as $n \to \infty$, $n^{1/2}(\hat{\sigma}^2_{\varepsilon} - \sigma^2_\varepsilon) \to N(0, E(\varepsilon^4))$. 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, $\sigma^2_\varepsilon$; 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 (3), 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 (3) 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

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

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

$$RV^{(k)}_t = \sum_{t_{i,-}, t_i \in G^{(k)}} r_{t_i}^2,$$

and the TSRV estimator then defined as

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

where $\overline{RV}^{(K)}_t = \frac{1}{K} \sum_{k=1}^{K} RV^{(k)}_t$, $RV_t$ is as defined in (3) 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 (6) suggested by Ait-Sahalia et al. (2005),

$$TSRV_2t = \left(\frac{n}{(K-J)n_K}\right) \left(K^{(K)}_t - \frac{n_K}{n_J} \overline{RV}^{(J)}_t\right).$$

The terms $\overline{RV}^{(J)}_t$ and $n_J$ are defined analogously to $\overline{RV}^{(K)}_t$ and $n_K$ respectively, but with $1 < J < K$ replacing $K$ in both expressions.\footnote{Following Zhang et al. (2005) we use $K = cn^{2/3}$, where $c = (16\sigma_t^4/\text{TE} (\eta^2))^{1/3}$ and $\eta^2 = \frac{4}{3}\int_{t_{i-1}}^{t_i} \sigma^4(s)ds$. The term $\sigma_t^4$ is square of the variance of the noise, while $\int_{t_{i-1}}^{t_i} \sigma^4(s)ds$ is the integrated quarticity. $\sigma^2_t$ is estimated as $\hat{\sigma}^2_t = \frac{1}{3T} RV_t(1)_i$, where $RV_t(1)$ denotes realized volatility based on fixed 1 minute sampling. This modified estimate of the noise variance is an attempt to reduce the impact of dependent noise; see Barndorff-Nielsen et al. (2006a). The term in the denominator, $E(\eta^2^2)$, is estimated as $\hat{E}(\eta^2^2) = \frac{4}{3} [RV_t(30)]^2$, where $RV_t(30)$ denotes realized volatility based on fixed 30 minute sampling, and we use $J = \max(1, \frac{K}{4})$. See Barndorff-Nielsen et al. (2006b) for further discussion of some of these computational issues.}

### 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\(^8\). Consistent with the definition of \(RV_t^{(k)}\) in (5), we define \(RCV_t^{(k)}(h) = \sum_{t_i, \ldots, t_i + h, \ldots, t_i + K} r_t, r_{t + h}, \ h = -H, \ldots, -1, 0, 1, 2, \ldots H, \) as the realized autocovariance function constructed from returns observed over pairs of successive time points in \(G^{(k)}\) in (4), \(k = 1, 2, 3, \ldots, K, \) with the returns being \(|h|\) time points apart.\(^9\) 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

\[
RKERN_t = \overline{RCV}_t^{(K)} + \sum_{h=1}^{H} w \left( \frac{h - 1}{H} \right) \left\{ \overline{RCV}_t^{(K)}(h) + \overline{RCV}_t^{(K)}(-h) \right\},
\]

with the particular form chosen for the weights, \(w \left( \frac{h - 1}{H} \right), h = 1, 2, 3, \ldots, H, \) determining the precise version of the estimator. In the empirical work we report results based on the cubic kernel estimator, in which \(w \left( \frac{h - 1}{H} \right) = 1 - 3 \left( \frac{h - 1}{H} \right)^2 + 2 \left( \frac{h - 1}{H} \right)^3, h = 1, 2, 3, \ldots H.\(^{10}\)

**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,

\[
OSRV_t = \sum_{j=0}^{M^*_t} r_{t+j, \delta_t}^2,
\]

based on \(M^*_t\) discretely sampled \(\delta_t\)-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 \(OSRV_t\)

---

\(^8\)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 (1)). 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.

\(^9\)The notation \(r_{t+h}\) denotes the return over successive time-points in the sub-grid \(G^{(k)}\), where that return is \(|h|\) time points distant from \(r_t\), according to the sub-grid \(G^{(k)}\).

\(^{10}\)The value of \(H = c_K \sqrt{\int_{t-1}^{t} \sigma^2(s) ds}n \) 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 7. 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 (\(\sigma^2_t\)) and integrated volatility used in the construction of \(H\) are the same as those used in the construction of \(K\), as detailed in Footnote 7. 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).
as an estimator of quadratic variation. Under certain conditions\textsuperscript{11}, the MSE is shown to be a function of $M^*_t\delta_t$, 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, $M^*_t\delta_t$ is chosen so as to minimize MSE where, as indicated by the notation, $M^*_t$ (and, hence, $\delta_t$) varies with $t$\textsuperscript{12}.

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. Defining \textit{realized bi-power variation} as

$$BPV_t = \frac{\pi}{2} \sum_{t_{i-1}, i \in [t-1, t]} |r_{t_i}| |r_{t_{i-1}}|, \quad (10)$$

Barndorff-Nielsen and Shephard 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 (3), for very large $n$ the statistic in (10) 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 (10), 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,

$$\overline{BV}^{(K)}_t = \frac{1}{K} \sum_{k=1}^{K} BV^{(k)}_t, \quad (11)$$

where $BV^{(k)}_t = \frac{x}{2 \sqrt{\pi}} \sum_{t_{i-1}, t_{i+1}, t_{i+2} \in G^{(k)}} |r_{t_i}| |r_{t_{i+2}}|$; and $k$ and $K$ are defined with respect to the transaction grid in (4).\textsuperscript{13}

\textsuperscript{11}In particular, with reference to (1), it is assumed that $\mu(t) = \nu(t) = 0$.

\textsuperscript{12}Following Bandi and Russell (2006), we approximate the optimal value of $M^*_t\delta_t$ as $M^*_t \sim \left( \int_{t-1}^{t} \sigma^4(s)ds \right)^{1/3} / \sigma_t^2$. The numerator and denominator are both calculated as explained in Footnote 7. Returns on day $t$ are thus sampled less frequently ($M^*_t\delta_t$ is smaller), the larger is the squared variance of the noise in the data relative to the quarticity of the underlying efficient price process.

\textsuperscript{13}As pointed out by a referee, the influence of subsampling on the robustness of the bi-power measure to jumps has not yet been investigated in the literature.
A-priori one would anticipate that option-implied forecasts, to the extent that such forecasts incorporate jump information, may be less accurate in forecasting (11) 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 is given by \( n^{(ch)} \times \text{tick}^2 \), where \( n^{(ch)} \in \mathbb{N} \) is the number of price changes in a day and \( \text{tick} \) 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 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 \( A + C = n^{(ch)}.15 \)

Without the presence of microstructure noise, the estimator \( n^{(ch)} \times \text{tick}^2 \) is a consistent estimator of quadratic variation, whilst in the presence of noise the value of \( n^{(ch)} \times \text{tick}^2 \) 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 (to represent a ‘modified alternation’ estimator), is given by

\[
ALTM_t = \frac{1}{K} \sum_{k=1}^{K} \frac{RV_t^{(k)} C^{(k)}}{A^{(k)}},
\]

with \( RV_t^{(k)} \) as defined in (5) and \( C^{(k)} \) \( A^{(k)} \) denoting the number of continuations (alternations) on grid \( G^{(k)} \) in (4).16

3 Forecasting Volatility

3.1 Overview

Since the advent of the realized volatility literature, not only has focus shifted from daily returns to intraday day returns in the construction of volatility proxies, but emphasis is also

---

14 See also Busch, Christensen and Nielsen (2006) and Anderson and Vahid (2007).
15 The first price of the day is defined as an alternation.
16 See Oomen (2006) for a related measure based on a discrete jump process.
now given to production of direct forecasts via standard time series models for these proxies; 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 can 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 (6) and (7) respectively, RKERN in (8), OSRV in (9), BV in (11) and ALTM in (12). 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 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 RVA(5). All measures are used both as proxies for the latent volatility and as the basis for producing direct forecasts of 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.

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

17 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 RVA(5) is the averaged version of RV(5) based on successive subgrids of (artificial) prices spaced five minutes apart.

18 See Hansen and Lunde (2005b) for precise details, including of the rule adopted for discarding outliers when calculating the weights.
forecasts produced from a range of GARCH-type specifications with a Student $t$ conditional distribution.\(^{19}\) Given $r_t = \mu + \varepsilon_t = \mu + \sigma_t e_t$, where $r_t$ denotes the $t$th daily return, $\mu$ the mean daily return, $\sigma_t^2$ the variance for day $t$ and $e_t$ a (standardized) Student $t$ variable, $t(\nu)/\sqrt{\nu-2}$, the following GARCH, threshold GARCH (TGARCH), power ARCH (PARCH) and fractionally integrated GARCH (FIGARCH) models are included in the initial forecast set:

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

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 $b_j = \frac{-dT(j-d)}{T(1-d)T(j+1)}$, and the remaining parameters satisfy the usual restrictions. In the asymmetric TGARCH model, $s_t = 1$ if $\varepsilon_t < 0$ and 0 otherwise. The PARCH model nests the GARCH model when $\delta = 2$. 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 an ARFIMA(p,d,q) model, $\phi(L)(1 - L)^d (\ln y_t - \alpha) = \theta(L)u_t$, where $y_t$ refers to any of the volatility measures described in Section 2, $\alpha$ is its mean, and $u_t$ is a Student $t$ variable scaled by $\psi > 0$. 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 $\sigma$. Under this distributional assumption,
the BS price of a European call option with strike price $X$ and maturity $T$ is

$$BS(\sigma) = P_t^{(D)}(d_1) - X e^{-i_t \tau} \Phi(d_2), \quad (13)$$

where $d_1 = \left( \ln\left( \frac{P_t^{(D)}}{X} \right) + (i_t + 0.5\sigma^2) \tau \right) / \sigma \sqrt{\tau}$, $d_2 = d_1 - \sigma \sqrt{\tau}$, $P_t^{(D)}$ is the (dividend-discounted) spot price at time $t$, $i_t$ is the (annualized) risk free rate of return at time $t$, $	au = T - t$ is the time to maturity (expressed as a proportion of a year) and $\Phi(.)$ is 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 $\sigma$ implied by $C_t(T,X)$ by equating $C_t(T,X)$ to the right-hand-side of (13) and solving for $\sigma$.

If the BS model were correct, the estimate of $\sigma$ implied by $C_t(T,X)$ would be invariant to both $X$ and $\tau$. 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, $\tau$. 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

$$E_t^* \left[ \int_t^T \sigma^2(s) ds \right] = 2 \int_0^\infty C_t(T,X) e^{i_t \tau} \max \left[ 0, P_t^{(D)} e^{i_t \tau} - X \right] dX, \quad (14)$$

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 (14) 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 (2), rather than an estimate of the integrated volatility, $IV_t = \int_t^{T_t} \sigma^2(s) ds$. Crucially, the calculation in (14) avoids the BS misspecification of the spot price process as geometric Brownian motion with a constant diffusion parameter. Instead, the right hand side of (14) harnesses the distributional information about $P(t)$ incorporated in the variation of the $C_t(T,X)$ across $X$. Details of how (14) is estimated using a finite number of strike prices are given in Section 4.1.

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 in (14), 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?

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 $V_i^2$. When the true latent price follows the model in (1), $V_i^2 = QV_t$. 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 $\hat{V}_t^2$ the realized proxy for the latent volatility at time $t$, and $f_{jt}$ as the forecast of $V_t^2$ produced by the $j$th 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, $L_{jt} = L(\hat{V}_t^2, f_{jt})$, $t = 1, 2, \ldots, N$. With $j = 0$ denoting the benchmark forecast, all $m$ alternative forecasts are compared with the benchmark via the time series of loss differentials, $D_{jt} = L_{0,t} - L_{jt}$, $j = 1, 2, \ldots, m$, $t = 1, 2, \ldots, N$.

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

13
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, with the bootstrap method used to this end catering for 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 unknown) 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 pairwise 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.\footnote{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), Hansen, Lund and Nason (2003), Giacomini and Komunjer (2005), Corradi and Swanson (2006) and Giacomini and White (2006).}

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. 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 Brownlees 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.\footnote{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.} For the individual stocks, we take as our estimate of ATM volatility (denoted by 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 \( \text{BS} \)).

Given maximum and minimum strike values \( X_{\text{max}} \) and \( X_{\text{min}} \) respectively, the estimate of MF implied volatility in (14) (denoted by \( \text{MF} \)) is given by

\[
MF = 2 \int_{X_{\text{min}}}^{X_{\text{max}}} \frac{C_t(T, X)e^{i\tau(t-T)}}{X^2} dX - \max \left[ 0, P_t(D)e^{i\tau(t-T)} - X \right]
\]

\[
\approx \sum_{j=1}^{M} [g(T, X_j) + g(T, X_{j-1})] \Delta X,
\]

where \( \Delta X = (X_{\text{max}} - X_{\text{min}})/M \), \( X_j = X_{\text{min}} + j\Delta X \) for \( 0 \leq j \leq M \) and \( g(T, X_j) = (C_t(T, X_j)e^{i\tau(t-T)} - \max \left[ 0, P_t(D)e^{i\tau(t-T)} - X_j \right]) / X_j^2 \). 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 IVOLATILITY one-month implied volatilities for the available range of moneyness values: \( 0.5 < X/P_t < 1.5 \) in steps of 0.1 \(^{22}\); 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 (13) 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 (15).\(^{23}\)

The forecasts \( \text{ATM} \), \( \text{BS} \) and \( \text{MF} \) 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. That is, the first 22-day-ahead forecast is the average of the forecasts for one day ahead (29 August, 2001), two days ahead, up to 22 days ahead, with the average then expressed as an annualized figure. The final 22-day-ahead forecast is then the (annualized) average of the forecasts for 30 May, 2006, and up to 21

\(^{22}\)Note that this curve itself has been produced via an initial interpolation procedure given the quoted option prices for particular strikes.

\(^{23}\)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 initial prices used to construct the smile and the artificial, interpolated prices produced for use in (15) may not be too large. This also means that the prices used in (15) may not be too different from prices based on a European formula, and the approximation error in MF reduced accordingly.
trading days after that. The variance measure being forecast corresponds to the (annualized) average of the daily variance values over the relevant forecast horizon. Each returns-based forecast is produced using both daily and intraday observations from \( R = 1000 \) days.\(^{24}\) 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.\(^{25}\) Each option-implied forecast is based on option prices observed on the day immediately prior to the forecast day (or the 22-day-ahead forecast 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 \( MF \) and \( ATM \) 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 \( L_{j,t} = \left[ \hat{V}_t^2 - f_{j,t} \right]^2 \). We provide results for one and 22 days ahead in Table 1 and 2 respectively, with the maturity of the options used to construct \( MF \) and \( ATM \) matching the forecast horizon in the second case only. In both tables, the results for IBM, MSFT and GE are given respectively in Panels A, B and C. Across the columns of each table we order the eight measures \( (\hat{V}_t^2) \), and associated results, according to the extent to which each measure accommodates noise and/or jumps. Specifically we report results for measures that do not formally adjust for noise or jumps: \( RV(5) \) and \( RVA(5) \); measures that adjust for noise only: \( TSRV \), \( TSRV2 \), \( RKERN \), \( OSRV \) and \( ALT \); and the measure that adjusts for both noise and jumps: \( BV \). We annotate the results in the following way: i) if a benchmark is rejected at the 5% level, the SPA \( p \)-value appears in bold; ii) in

\(^{24}\)To minimize the impact on the empirical results of variation in the 24-hour scaling weights (from sample to sample), we estimate the weights for each volatility measure once only, using data over the forecast evaluation period. This single pair of weights is then used to re-scale the relevant measure over all rolling samples used to produce the one-step-ahead direct forecasts. Although this amounts to some ‘out-of-sample information’ being used in the production of the direct forecasts, the weights do not constitute specific observational information that could be viewed as unduly benefiting those forecasts relative to the competing classes of forecasts.

\(^{25}\)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 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.
the case where a benchmark is rejected, the ‘most significant’ forecast model according to
the pairwise ‘t statistics’ is indicated in parentheses in the line below; iii) 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.27

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, \( ATM \) is the most ‘significant’ alternative,
as based on the individual pairwise ‘t statistics’. For MSFT and GE there is support for
\( MF \) 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 \( MF \) benchmark is rejected, with
\( ATM \) again the most ‘significant’ alternative in many instances. Both long-memory and
short-memory direct forecasts also feature as the most significant alternatives in some cases.

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

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 \( ATM \) and lack of support for \( MF \) - 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 \( ATM \) to forecast the “jump-free” BV measure for IBM, a result that
contrasts with all other results in the table related to this benchmark.

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 \( ATM \) forecast of
MSFT and GE volatility this is indeed the case, with the \( p \)-values for the \( ATM \) benchmark
uniformly higher for the 22-day forecast horizon than the corresponding \( p \)-values for the

\( \footnote{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 significant’ alternative according to the pairwise
comparisons may itself be rejected as a benchmark model using the SPA test.}

\( \footnote{Note that Hansen and Lunde (2005) suggest the convention of assigning a value of one to the \( p \)-value in
this case.} \)
Table 1:

SPA p-values: forecasts based on a one-day-ahead forecast horizon. An option-implied volatility forecast is used as benchmark: \( MF \) (model free) and \( ATM \) (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: \( m = 66; N = 1147 \). The model set always includes the option-implied forecast that is the alternative to the one being tested as the benchmark. p-values that are associated with rejection of the benchmark forecast at the 5% level are highlighted in bold font. In the case of rejection, the ‘most significant’ alternative forecast, according to the pairwise ‘t statistics’, is reported in parentheses below the p-value. The acronym \( LM_{own(cross)} \) denotes a long-memory ARFIMA own (cross) forecast, while the acronym \( SM_{own(cross)} \) denotes a short-memory ARMA own (cross) forecast. In the case where a benchmark is not rejected, the superscript \( # \) indicates that the forecast also has the smallest MSFE loss of all \( m + 1 \) forecasts in the choice set.

Panel A: IBM data

<table>
<thead>
<tr>
<th>Measure to be forecast( ^{(a)} )</th>
<th>RV(5)</th>
<th>RVA(5)</th>
<th>OSRV</th>
<th>RKERN</th>
<th>TSRV1</th>
<th>TSRV2</th>
<th>ALTM</th>
<th>BV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>( MF ) (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>( ATM ) (most sig.)</td>
<td>0.301</td>
<td>0.207</td>
<td>0.233</td>
<td>0.346</td>
<td>0.238</td>
<td>0.217</td>
<td>0.764#</td>
</tr>
</tbody>
</table>

Panel B: MSFT data

<table>
<thead>
<tr>
<th>Measure to be forecast( ^{(a)} )</th>
<th>RV(5)</th>
<th>RVA(5)</th>
<th>OSRV</th>
<th>RKERN</th>
<th>TSRV1</th>
<th>TSRV2</th>
<th>ALTM</th>
<th>BV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>( MF ) (most sig.)</td>
<td>0.014</td>
<td>0.012</td>
<td>0.006</td>
<td>0.014</td>
<td>0.001</td>
<td>0.001</td>
<td>0.364</td>
</tr>
<tr>
<td></td>
<td>( ATM ) (most sig.)</td>
<td>0.472</td>
<td>0.474</td>
<td>0.485</td>
<td>0.471</td>
<td>0.306</td>
<td>0.337</td>
<td>0.382</td>
</tr>
</tbody>
</table>

Panel C: GE data

<table>
<thead>
<tr>
<th>Measure to be forecast( ^{(a)} )</th>
<th>RV(5)</th>
<th>RVA(5)</th>
<th>OSRV</th>
<th>RKERN</th>
<th>TSRV1</th>
<th>TSRV2</th>
<th>ALTM</th>
<th>BV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>( MF ) (most sig.)</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>0.075</td>
<td>0.041</td>
<td>0.015</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td>( ATM ) (most sig.)</td>
<td>0.934#</td>
<td>0.941#</td>
<td>0.935#</td>
<td>0.740</td>
<td>0.560</td>
<td>0.626</td>
<td>0.404</td>
</tr>
</tbody>
</table>

\( ^{(a)} \) All acronyms for the eight measures being forecast are defined in Section 4.2.1.
Table 2:
SPAp values: forecasts based on a 22-day-ahead forecast horizon. An option-implied volatility forecast is used as benchmark: MF (model free) and 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. The model set always includes the option-implied forecast that is the alternative to the one being tested as the benchmark. p-values that are associated with rejection of the benchmark forecast at the 5% level are highlighted in bold font. In the case of rejection, the ‘most significant’ alternative forecast, according to the pairwise ‘t statistics’, is reported in parentheses below the p-value. The acronym LMcross denotes a long-memory ARFIMA cross forecast In the case where a benchmark is not rejected, the superscript # indicates that the forecast also has the smallest MSFE loss of all m + 1 forecasts in the choice set.

Panel A: IBM data

<table>
<thead>
<tr>
<th>Measure to be forecast (a)</th>
<th>RV(5)</th>
<th>RVA(5)</th>
<th>OSRV</th>
<th>RKERN</th>
<th>TSRV1</th>
<th>TSRV2</th>
<th>ALTM</th>
<th>BV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>MF</td>
<td>ATM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>ATM</td>
<td>LMcross</td>
<td>0.117</td>
<td>0.043</td>
<td>0.208</td>
<td>0.060</td>
<td>0.052</td>
<td>0.844#</td>
</tr>
</tbody>
</table>

Panel B: MSFT data

<table>
<thead>
<tr>
<th>Measure to be forecast</th>
<th>RV(5)</th>
<th>RVA(5)</th>
<th>OSRV</th>
<th>RKERN</th>
<th>TSRV1</th>
<th>TSRV2</th>
<th>ALTM</th>
<th>BV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>MF</td>
<td>ATM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.038</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>ATM</td>
<td>LMcross</td>
<td>0.571#</td>
<td>0.961#</td>
<td>0.934#</td>
<td>0.585#</td>
<td>0.870#</td>
<td>0.940#</td>
</tr>
</tbody>
</table>

Panel C: GE data

<table>
<thead>
<tr>
<th>Measure to be forecast</th>
<th>RV(5)</th>
<th>RVA(5)</th>
<th>OSRV</th>
<th>RKERN</th>
<th>TSRV1</th>
<th>TSRV2</th>
<th>ALTM</th>
<th>BV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>MF</td>
<td>ATM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
<td>0.001</td>
<td>0.000</td>
<td>0.057</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>ATM</td>
<td>LMcross</td>
<td>0.964#</td>
<td>0.960#</td>
<td>0.946#</td>
<td>0.984#</td>
<td>0.956#</td>
<td>0.927#</td>
</tr>
</tbody>
</table>

(a) All acronyms for the eight measures being forecast are defined in Section 4.2.1.
one-day horizon, and close to one in many cases. Moreover, the ATM forecast has the lowest MSFE in the forecast set (again, as indicated by the \# superscript) for all eight forecast variables, for both the MSFT and GE series. The results for IBM are less clear-cut, although there is still support for the benchmark ATM for the majority of forecast measures. In contrast, the results for the MF benchmark are even weaker at the longer horizon, with only a single failure to reject MF as the superior forecast, across all series and all measures, and that support for MF being only marginal (p-value = 0.057). Once again, both option-implied volatilities fail to successfully predict the BV measure for IBM. The p-value for ATM forecast of the BV measure, in the case of GE, although very supportive of the ATM benchmark, is the smallest across the alternative measures. The corresponding p-value is amongst the smallest in the case of MSFT.

As with the one-day-ahead predictions, there is some support for direct forecasts, in that for the three instances in which ATM is rejected as the benchmark model, a long memory direct forecast is the ‘most significant’ according to the pairwise 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 amongst the lowest in terms of MSFE, for all series and measures, and for both forecast horizons.

In the following section we attempt to shed some light on the contrast between the support for the ATM benchmark and the (overall) lack of support for the MF benchmark, by examining the option market information from which the 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 MF.\(^{28}\) In the right-hand panels, (b), (d) and (f) respectively, we plot MF against 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.\(^{29}\) 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

---

\(^{28}\)Qualitatively similar results are produced for the other measures of volatility.

\(^{29}\)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.
from 2 August, 2004 to 30 May, 2006; 2) The $MF$ forecast tends to exceed realized volatility (overall), and by a greater amount in the high- than in the low-volatility period; 3) The $MF$ forecast tends to exceed the $ATM$ forecast, again by a greater amount in the high-volatility period; 4) The $MF$ forecast is excessively noisy, relative to realized volatility, and more so than is the $ATM$ forecast, again in the high-volatility period in particular.

The empirical features of OSRV, $MF$ and $ATM$, for all three series, and for the full sample period and both sub-periods identified here, are summarized in Table 3. Using $\hat{\nu}_t^2$ to represent OSRV, setting $f_t = MF, ATM$ (as variance quantities), and using the decomposition of the MSFE as $E\left[\hat{\nu}_t^2 - f_t\right]^2 = \text{var}(\hat{\nu}_t^2 - f_t) + \text{var}(\hat{\nu}_t^2 - f_t)$, we report sample estimates of the forecast bias and forecast error variance, $E(\hat{\nu}_t^2 - f_t)$ and $\text{var}(\hat{\nu}_t^2 - f_t)$ respectively, as well as the sample variance of the forecast itself, $\text{var}(f_t)$. The numerical results clearly support the informal graphical evidence: $MF$ is both a more biased forecast and has a larger forecast error variance than $ATM$, in particular over the high-volatility period. Most notably, the (magnitude of the) bias of $MF$ is approximately twice as large as that for $ATM$ in the high-volatility period, in the case of IBM and MSFT, and more than five times larger for the GE data set. 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 in both the high- and low-volatility sample periods.

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

---

30 For the purposes of this illustration we omit the last 44 observations from the second MSFT sub-sample so that this second sub-period is accurately described as ‘low-volatility’.
Figure 1: GE, MSFT and IBM Volatility (Annualized standard deviation): 29 August, 2001 to 30 May, 2006.
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.

<table>
<thead>
<tr>
<th></th>
<th>IBM</th>
<th>MSFT</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full sample period</strong> (28 August, 2001 to 30 May, 2006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E(\hat{V}_t^2 - f_t)$</td>
<td>-0.0343</td>
<td>-0.0190</td>
<td>-0.0313</td>
</tr>
<tr>
<td>$\text{var}(\hat{V}_t^2 - f_t)$</td>
<td>0.0021</td>
<td>0.0012</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\text{var}(f_t)$</td>
<td>0.0061</td>
<td>0.0038</td>
<td>0.0107</td>
</tr>
<tr>
<td><strong>High-volatility sample period</strong> (28 August, 2001 to 30 July, 2004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E(\hat{V}_t^2 - f_t)$</td>
<td>-0.0503</td>
<td>-0.0271</td>
<td>-0.0518</td>
</tr>
<tr>
<td>$\text{var}(\hat{V}_t^2 - f_t)$</td>
<td>0.0026</td>
<td>0.0018</td>
<td>0.0049</td>
</tr>
<tr>
<td>$\text{var}(f_t)$</td>
<td>0.0068</td>
<td>0.0044</td>
<td>0.0108</td>
</tr>
<tr>
<td><strong>Low-volatility sample period</strong> (2 August, 2004 to 30 May, 2006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E(\hat{V}_t^2 - f_t)$</td>
<td>-0.0095</td>
<td>-0.0066</td>
<td>-0.0065</td>
</tr>
<tr>
<td>$\text{var}(\hat{V}_t^2 - f_t)$</td>
<td>1.52e-004</td>
<td>1.32e-004</td>
<td>9.27e-005</td>
</tr>
<tr>
<td>$\text{var}(f_t)$</td>
<td>1.12e-004</td>
<td>8.81e-005</td>
<td>7.58e-005</td>
</tr>
</tbody>
</table>
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 \( ATM \) is equated to the ordinate of the volatility curve at \( X/P_t = 1 \), and \( MF \) constructed from a formula that uses all ordinates, the reason why \( MF \) tends to exceed \( ATM \) 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 \( MF \) reported in Table 3, which contributes, in turn, to the large forecast error variance. Again, this noise in the away-from-the-money volatilities is likely to be exacerbated by the lack of liquidity in options far from the money.

In contrast to the rather distinct smile shape that characterizes some of the curves in Figure 2, during the low-volatility period highlighted in Figure 3 the curves tend to be skewed, the majority having the negative slope that typifies equity option graphs, and all exhibiting much less variation across the moneyness spectrum than the curves in Figure 2. 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 \( MF \) and \( ATM \) values, plus much less variation in the \( MF \) values, than during the high volatility state.

In summary, close examination of the volatility smile information from which \( MF \) and \( ATM \) are extracted provides some explanation for both the discrepancy between the two measures and for the added variability in the \( MF \) measure, in particular in times of high volatility.\(^{31}\) 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.

---

\(^{31}\)The excessive variability in \( MF \) would only be exacerbated by any measurement error associated with the application of the MF formula to the American-style option data.
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.
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,

$$\int_t^T \sigma^2(s) ds = \phi_0 + \phi_1 \mathbb{E}^*_t \left[ \int_t^T \sigma^2(s) ds \right] + e_{t,T} \quad (16)$$

are functions of the parameters of the risk-neutralized version of the distribution with respect to which $E^*_t(.)$ in (16) 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 (16) 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.\(^3\)

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 (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

\(^{32}\text{Again, the same qualitative results, although not reported, were obtained for the other realized measures.}\)
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 \( MF \) measure that 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 \( ATM \) forecast, on the other hand, approximated by an implied volatility at a single point in the moneyness spectrum, does not \textit{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.\textsuperscript{33}

In summary then, any potential additional forecast accuracy associated with the added flexibility of the assumptions underlying the \( MF \) 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 \( MF \), which retains some of the smile information, but not all, manages to outperform \( ATM \). We investigate this in the following section by reporting SPA test results for three modified versions of \( MF \).

### 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: \( MF \) and \( ATM \), plus three truncated versions of \( MF \), denoted by: \( MF_{1.5} \), \( MF_{2.0} \) and \( MF_{2.5} \). The benchmark \( MF_{1.5} \), for example, is the estimate of \( MF \) produced from implied volatilities within the moneyness range: \( 1 \pm 1.5 \times ATM/\sqrt{12} \). The benchmarks \( MF_{2.0} \) and \( MF_{2.5} \) are defined correspondingly.\textsuperscript{34} We produce the test results for the full sample period (Panel A), as well as results for the low-volatility period identified in Section 4.2.2 (Panel B), 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 \( MF \) and \( ATM \) are re-produced under the expanded model set in which \( MF_{1.5} \), \( MF_{2.0} \) and \( MF_{2.5} \) are included as alternatives. Hence, the results in the rows headed \( MF \) and \( ATM \) in Table 4 differ in some cases 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: RKERN, ALTM 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 \( MF_{1.5} \) benchmark is given limited support for GE and MSFT (for the ALTM volatility measure in

\textsuperscript{33}See also Bates (1996) for early discussion of the robustness of option-implied volatility based on at-the-money options.

\textsuperscript{34}Jiang and Tian (2005) and Bollerslev, Gibson and Zhou (2006) also use truncation in calculating MF implied volatilities.
Table 4:

SPA p-values: forecasts based on a 22-day-ahead forecast horizon. Alternative option-implied volatility forecasts are used as benchmark: $MF$, $MF_{2.5}$, $MF_{2.0}$, $MF_{1.5}$ and $ATM$. The SPA test is based on a mean squared forecast error (MSFE) loss criterion, for variance quantities, with three alternative measures used for the actual volatility: RKERN, ALTM and BV. Results are produced for the full sample and low-volatility periods. $p$-values that are associated with rejection of the benchmark forecast at the 5% level are highlighted in bold font. In the case of rejection, the ‘most significant’ alternative forecast, according to the pairwise ‘t statistics’, is reported in parentheses below the $p$-value. The model set always includes all of the option-implied forecasts that are alternatives to the one being tested as the benchmark. Hence, the model set for each series underlying the results in Tables 1 and 2 is augmented by three here, to cater for the three additional versions of $MF$. The acronym $LM_{cross}$ denotes a long-memory ARFIMA cross forecast. In the case where a benchmark is not rejected, the superscript # indicates that the forecast also has the smallest MSFE loss of all $m + 1$ forecasts in the choice set.

### Panel A: Full Sample Period (28 August, 2001 to 30 May, 2006)

<table>
<thead>
<tr>
<th>Measure to be forecast(^{(a)})</th>
<th>IBM</th>
<th>MSFT</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bench.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$MF$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($MF_{2.0}$)</td>
<td>($MF_{2.0}$)</td>
<td>($MF_{2.5}$)</td>
<td></td>
</tr>
<tr>
<td>$MF_{2.5}$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($MF_{2.0}$)</td>
<td>($MF_{2.5}$)</td>
<td>($MF_{2.5}$)</td>
<td></td>
</tr>
<tr>
<td>$MF_{2.0}$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($MF_{2.5}$)</td>
<td>($MF_{2.0}$)</td>
<td>($MF_{2.0}$)</td>
<td></td>
</tr>
<tr>
<td>$MF_{1.5}$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td></td>
</tr>
<tr>
<td>$ATM$ (most sig.)</td>
<td>0.208</td>
<td>0.844#</td>
<td>0.649*</td>
</tr>
<tr>
<td>($LM_{cross}$)</td>
<td>($LM_{cross}$)</td>
<td>($LM_{cross}$)</td>
<td></td>
</tr>
</tbody>
</table>

### Panel B: Low Volatility Period (2 August, 2004 to 30 May, 2006)

<table>
<thead>
<tr>
<th>Measure to be forecast</th>
<th>IBM</th>
<th>MSFT</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bench.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$MF$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td></td>
</tr>
<tr>
<td>$MF_{2.5}$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td></td>
</tr>
<tr>
<td>$MF_{2.0}$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($MF_{1.5}$)</td>
<td>($MF_{1.5}$)</td>
<td>($MF_{1.5}$)</td>
<td></td>
</tr>
<tr>
<td>$MF_{1.5}$ (most sig.)</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td>($ATM$)</td>
<td></td>
</tr>
<tr>
<td>$ATM$ (most sig.)</td>
<td>0.028</td>
<td>0.287</td>
<td>0.007</td>
</tr>
<tr>
<td>($LM_{cross}$)</td>
<td>($LM_{cross}$)</td>
<td>($LM_{cross}$)</td>
<td></td>
</tr>
</tbody>
</table>

\(^{(a)}\) All acronyms for the measures being forecast are defined in Section 4.2.1.
particular). However, overall, the ATM forecast remains dominant, even when the model set is expanded to include the added variants of MF.\(^{35}\) 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 (ATM, plus all variants of 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 ATM and 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, a long-memory direct forecast sometimes features as the most significant alternative according to a pairwise test. This is most notable in the low volatility period. However the superiority of any particular long-memory forecast, taking into account the multiple alternative forecasts, 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 multiple alternative forecasts; see 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 compared with the 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.\(^{36}\)

\(^{35}\)Note, that the fact an MF variant is often the most ‘significant’ alternative according to a pairwise ‘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 pairwise comparisons.

\(^{36}\)In order to retain comparability with the earlier results, we continue to construct the BV measure with the noise adjustment as per (11). As pointed out by a co-editor, S&P500 index futures, being a traded asset, would be subject to similar microstructure influences to the traded stocks.
In Figure 4, Panels (a), (b), (c) and (d), we plot, respectively, RV(5) and $MF$, RV(5) and $BS$, $MF$ and $VIX$, and $MF_{2.5}$ 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 $MF_{2.5}$ 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 several 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 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, and over an evaluation period that includes a high-volatility state. 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. That said, there is limited support for the idea that option prices do factor in jump information, given the overall tendency for both types of option-implied forecasts to do less well as a forecast of (noise-adjusted) bi-power variation. This observation would need to be pursued more formally, however, before any definitive conclusions along these lines could 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
Figure 4: S&P500 Volatility (Annualized standard deviation): 29 August, 2001 to 30 May, 2006.
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, we reiterate that the focus of this paper has been on the assessment of the forecasting performance of particular option-implied benchmarks. Some limited evidence has been produced that suggests that direct forecasts of realized volatility measures may also serve as useful forecasts of future volatility. Most notably, the results suggest that 22-day-ahead forecasts based on long memory models may perform well during a low-volatility period. However, it has not been within the purview of this paper to investigate this matter formally via SPA tests of direct benchmarks. Indeed, given the large number of different possible direct forecasts, including both own-forecasts (constructed from the proxy being forecast) and cross-forecasts (constructed from alternative proxies), it may be that the SPA approach is not suitable for assessing the superiority (or otherwise) of such forecasts. Rather, 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 quite different question from that addressed here: which form of model, or category of model, whether returns- or options-based, direct or indirect, long memory or short memory, provides the best forecast of volatility?

References


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


