Forecasting S&P 500 Volatility: Long Memory, Level Shifts, Leverage Effects, Day-of-the-Week Seasonality, and Macroeconomic Announcements*

Martin Martens†
Finance Department
Erasmus University Rotterdam

Dick van Dijk‡
Econometric Institute
Erasmus University Rotterdam

Michiel de Pooter§
Federal Reserve Board

October 25, 2008

Abstract

We evaluate the forecasting performance of time series models for realized volatility, which accommodate long memory, level shifts, leverage effects, day-of-the-week and holiday effects, as well as macroeconomic news announcements. Applying the models to daily realized volatility for the S&P 500 futures index we find that explicitly accounting for these stylized facts of volatility improves out-of-sample forecast accuracy for horizons up to 20 days ahead. Capturing the long-memory feature of realized volatility by means of a flexible high-order AR-approximation instead of a parsimonious but stringent fractionally integrated specification also leads to improvements in forecast accuracy, especially for longer horizon forecasts.

Key words: Realized volatility, long memory, day-of-the-week effect, leverage effect, volatility forecasting, model confidence set, macroeconomic news announcements.

JEL Classification Code: C22, C53, G15

---

*We would like to thank seminar participants at the "New Frontiers in Financial Volatility Modelling" conference in Florence (May 2003) and the CIRNAO-CIREQ conference on "Realized Volatility" in Montreal (November 2003) for providing useful comments. We also want to thank Torben Andersen, Tim Bollerslev and an anonymous referee for detailed comments and suggestions. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other employee of the Federal Reserve System.

A web appendix to this paper which contains results for the two-time scales and bipower variation volatility estimators is available on `http://people.few.eur.nl/djvandijk/#research`

†Finance Department, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: `mmartens@few.eur.nl`

‡Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: `djvandijk@few.eur.nl` (corresponding author)

§Division of International Finance, Federal Reserve Board, 20th Street and Constitution Avenue NW, Washington, DC 20551, United States, e-mail: `michiel.d.depooter@frb.gov`

1 Introduction

Accurately measuring and forecasting financial volatility is of crucial importance for asset and derivative pricing, asset allocation and risk management. Merton (1980) already noted that the variance over a fixed period can be estimated arbitrarily accurately by the sum of squared intra-period returns, provided the data are available at a sufficiently high sampling frequency. With transaction prices becoming more widely available, Andersen and Bollerslev (1997, 1998a) kick-started a flurry of research on the use of high-frequency intra-day data for measuring and forecasting daily volatility. The so-called "realized" variance (RV), defined as the sum of squared intra-day returns, indeed has been found to provide a much more accurate measure of daily volatility than the popular daily squared return.

A large part of the existing literature focuses on using the available intra-day data in the best possible way for measuring daily volatility. Several alternatives for the standard RV estimator have been suggested to isolate the more predictable continuous volatility component from the jump component using bipower variation, see Barndorff-Nielsen and Shephard (2004), or to reduce the impact of market microstructure noise on the estimator, see Zhang, Mykland, and Aït-Sahalia (2005), Bandi and Russell (2006, 2008) and Hansen and Lunde (2006), among others.

In this paper we take a different perspective, in the sense that we take the daily volatility measures as given, and focus on modeling and forecasting realized volatility instead. We address two issues in particular. First, we compare different models that have been considered for modeling the apparent long-memory characteristic of (realized) volatility as suggested by the slow (hyperbolic) decay of its autocorrelation function, see Andersen, Bollerslev, Diebold, and Labys (2001b) and Andersen, Bollerslev, Diebold, and Ebens (2001a). Several studies employ Autoregressive Fractionally Integrated (ARFI) models for this purpose (Ebens, 1999, Andersen, Bollerslev, Diebold, and Labys, 2003a, Koopman, Jungbacker, and Hol, 2005, Martens and Zein, 2004, Pong, Shackleton, Taylor, and Xu, 2004, among others), while other studies consider the Heterogenous Autoregressive (HAR) model of Corsi (2004), see Andersen, Bollerslev, and Diebold (2007a), for example. We contrast these models with flexible, unrestricted high-order AR-approximations.

Second, we aim to examine the importance of explicitly incorporating several other stylized facts of (realized) volatility in these models for the purpose of forecasting. We consider four typical features in particular. First, we allow for the so-called leverage effect, which refers to the asymmetric relation between news (as measured by lagged unexpected returns) and volatility, as documented by Black (1976), Pagan and Schwert (1990), and Engle and Ng (1993). Using daily squared returns as a volatility proxy, these (and other) studies show that a negative return tends to increase subsequent volatility much more than a positive return of the same magnitude. Using realized volatility measures it is less clear whether a leverage effect will be present as a realized volatility measure will already contemporaneously capture negative return shocks. However, Bollerslev, Litvinova, and Tauchen (2006) and Bollerslev, Kretschmer, Pigorsch, and Tauchen (in press) provide evidence of a prolonged leverage effect for S&P500 futures returns. Second, we allow for the presence of irregular level shifts, which have been documented to occur in financial volatility by Lamoureux and Lastrapes (1990) and Andreou and Ghysels (2002), among others. Interestingly, neglecting such occasional structural breaks can spuriously suggest the presence of long memory, as shown by Diebold and Inoue (2001), among others. Andersen and Bollerslev (1997) on the other hand suggest that long-memory is more likely to be an inherent feature of the return-generating process than an artefact stemming from structural streaks. Third, we consider seasonality effects. Baillie and Bollerslev (1989) and Harvey and Huang (1991), among others, find that average volatility is not constant across the different days of the week but displays a rather pronounced U-shaped pattern with volatility being lowest on Wednesdays.

Furthermore, volatility tends to be different on and around holidays. Fourth and finally, we account for the effects of macroeconomic news announcements, following the empirical evidence that volatility is higher on days with such announcements, see Andersen, Bollerslev, Diebold, and Vega (2003b, 2007b) for recent accounts. In fact, Andersen and Bollerslev (1998b) explain the day-of-the-week effect in Deutsche Mark-Dollar volatility by the clustering of macroeconomic news announcements on specific weekdays (in particular on Fridays). Allowing for volatility to be different on days with news releases can disentangle calendar and announcement effects.

The main contribution of this paper is that we consider models for realized volatility that simultaneously capture long memory, level shifts, leverage effects, day-of-the-week effects and macroeconomic news announcements, where we focus in particular on the importance of accounting for these features for the purpose of forecasting. To the best of our knowledge, we are the first to consider such comprehensive nonlinear models. The small number of previous studies that have examined the presence of nonlinearities in realized volatility all are limited in one way or another. Ebens (1999), Oomen (2002) and Giot and Laurent (2004) incorporate leverage effects in a long memory model for various stock indexes. Only Giot and Laurent (2004) consider out-of-sample forecasting and only at the one-day horizon. Maheu and McCurdy (2002) use a regime-switching model for the DM/$ exchange rate, but do not consider any other nonlinearities or long memory and only forecast one-day-ahead.

We use the different models for producing volatility forecasts at various horizons for S&P 500 index-futures, which renders two main conclusions. First, accounting for the effects of leverage, seasonality, and macroeconomic announcements generally improves the forecast performance compared to models without these effects for forecast horizons of up to 20 days. The forecasts obtained from nonlinear models improve upon those from a linear ARFI model on all evaluation criteria considered. For example, for one-day ahead forecasts the R^2 from a regression of realized volatility on the volatility forecast increases from 60.0% to 62.6%. Comparing the different nonlinearities it follows that the largest gain in forecasting performance comes from incorporating the leverage effect. Interestingly, we find that explicitly accounting for level shifts does not improve forecast accuracy.

Second, it seems beneficial to model the long-memory feature of realized volatility by means of a flexible high-order AR-approximation instead of either a parsimonious but stringent fractionally integrated specification or a heterogenous autoregressive model, especially for longer horizon forecasts. For 10-day ahead forecasts, for example, the R^2 is 73.2% for an AR(22) model, compared to 72.2% for the nonlinear ARFI model and 70.6% for the HAR model. Our conclusions are robust to using different daily volatility measures, in particular to using the two-time scales and bipower variation estimators.

The remainder of this paper is structured as follows. First, we discuss the S&P 500 data in Section 2, highlighting the stylized facts which motivate our forecasting models. In Section 3 we describe the nonlinear long-memory models which incorporate possible leverage effects, level shifts, seasonal behaviour, and effects of macroeconomic news announcements. In Section 4, we discuss the out-of-sample forecasting results. Section 5 concludes.

2 Data

We construct our measure of daily realized volatility for the S&P 500 index using intraday transaction prices for futures contracts as traded on the Chicago Mercantile Exchange (CME) from 8:30AM to 3:15PM (Eastern Standard Time minus 1 hour, EST-1). The realized variance for day t, denoted s_t^2, is computed as the sum of M squared intraday returns plus the squared overnight return between the closing price on day $t-1$ and the opening price on day t. Given that the S&P 500 index futures contracts are highly liquid, we adopt the popular five-minute sampling frequency for the intraday returns ($J = 81$), using the last transaction price in each interval.[1] The sample period runs from January 3, 1994 until December 29, 2006. We exclude days on which the market was opened only during the morning, leaving a total of $T = 3265$ daily observations. The very large negative overnight return for the period September 11-17, 2001 is excluded as well.

Table 1 contains descriptive statistics of the (log) realized variance measure, as well as for daily returns r_t, for squared daily returns, and for daily returns standardized with the realized standard deviation; r_t/s_t. A number of interesting features emerge from this table, which closely correspond with the distributional characteristics for realized volatility documented in Andersen, Bollerslev, Diebold, and Ebens (2001a) and Andersen, Bollerslev, Diebold, and Labys (2001b), among others. First, comparing the daily squared returns with the realized variance shows that these have an identical mean of 1.181. We would expect this to be the case, as both are unbiased measures of the true volatility. However, the standard deviation of the realized variance is at 1.771 much smaller than the standard deviation of the squared returns, which equals 2.894. It is precisely this characteristic that shows that realized variance is a less noisy estimate of true volatility than the daily squared return. Second, the realized variance is heavily skewed and exhibits excess kurtosis. By contrast, the logarithm of realized volatility, $\log(s_t^2)$, is much more symmetrically distributed and has much lower kurtosis. For this reason we will consider time series models for the log realized volatility. Third, the daily S&P 500 returns are skewed and leptokurtic. Standardized returns r_t/s_t, however, exhibit much less skewness and excess kurtosis and are in fact fairly close to being normally distributed. Fourth, as documented in

- insert Table 1 and Figure 1 about here -

other studies (and therefore not explicitly shown here), the sample autocorrelation function of the realized volatility measure exhibits a slow hyperbolic decay, indicative for the presence of long memory. A further point is that the persistence in the autocorrelation functions is much stronger for the (log) realized variance than for the daily squared returns.

Fifth, from the time series plot of the daily log realized variance in Figure 1, it appears that volatility is higher on average during 1998-2003, approximately, than during the first and last few years of the sample period. This suggests the possibility that multiple structural breaks in the level of RV have occurred, as documented by Andreou and Ghysels (2002). It is difficult to pin down when exactly these level shifts occurred though, and they appear to be most adequately characterized as gradual changes.

Sixth, the scatter plot of $\log(s_t^2)$ against r_{t-1} in Figure 2 reveals a rather pronounced relationship between current volatility and lagged returns beyond that already captured in contemporaneous realized volatility. To examine the possible presence of a leverage effect, we estimate the "news impact curve" (Engle and Ng, 1993)

$$\log(s_t^2) = \beta_0 + \beta_1|r_{t-1}| + \beta_2\mathtt{I}[r_{t-1} < 0] + \beta_3|r_{t-1}|\mathtt{I}[r_{t-1} < 0], \tag{1}$$

where $\mathtt{I}[A]$ is an indicator function for the event A, being equal to 1 if A occurs, and 0 otherwise. The fit from this regression is included in Figure 2 as well, along with the fit from a symmetric version of this news impact curve, obtained by setting $\beta_2 = \beta_3 = 0$ in (1). It is clearly seen that the impact of negative lagged returns is larger than the effect of positive returns of equal magnitude. Also, the parametric form in (1) appears to be quite reasonable, as can be seen by comparing the fit from this regression with a nonparametric regression of log realized volatility on the lagged return, also shown in Figure 2.

- insert Figure 2 about here -

Seventh, Table 2 shows the overall mean for the return and volatility measures on different types of days by distinguishing between regular days, days following a holiday[2] and days with macroeconomic news announcements.[3] It is clear that returns and volatility are both higher after holidays and that volatility during the Christmas period is only roughly half of its level during regular days. Volatility is notably higher on announcement days, especially on days when employment figures or decisions regarding the federal funds rate are released when realized variance is on average 1.636 and 1.521, respectively, both substantially above its non-announcement day mean of 1.133.

- insert Table 2 about here -

Finally, Table 3 shows the overall mean and the mean on different days of the week for returns and variance measures. The top panel in the table confirms the common finding based on daily returns that Mondays and Fridays exhibit higher volatility than other days. Interestingly, this pattern is quite different for the realized variance. Thursdays and Fridays exhibit the highest volatility but Mondays no longer have an above average volatility. In fact, for realized variance the mean is lowest on Mondays. These observed patterns can to a large extent be explained by making the distinction between days with and without macro releases, similar as in Andersen and Bollerslev (1998b). The middle and bottom panels in Table 3 allow for a direct comparison. Squared returns and realized variance are both clearly higher on announcement days than non-announcement days, especially on Thursdays and Fridays. For the sample of announcement days, squared returns and realized variance increase substantially over the days of the week (leaving Mondays, with 2 announcements only, out of consideration). This largely explains the higher average values for realized variance on these days in the top panel: on non-announcement days (Panel B), day-of-the-week effects in RV seem to be absent altogether. It is interesting to note that squared returns are still higher at the beginning of the week with averages of 1.245 and 1.331 for Mondays and Tuesdays, respectively.

- insert Table 3 about here -

3 Nonlinear Long Memory Models

We consider a range of forecasting models for the logarithmic realized volatility $y_t = \log(s_t^2)$, which can all be put in the general form

$$\phi(L)(y_t - \mu_t) = \varepsilon_t, \tag{2}$$

where $\phi(L)$ is a polynomial in the lag operator L defined as $L^k y_t \equiv y_{t-k}$ for any integer k, μ_t is the possibly time-varying mean, and ε_t is a white noise process. The models are distinguished by the specifications of $\phi(L)$ and μ_t. We consider three possibilities for the autoregressive dynamics, where the most crucial feature to capture is the long memory of the log realized variance, that is, the slow decay of its autocorrelation function. First, we employ Autoregressive Fractionally Integrated (ARFI) models, where

$$\phi(L) = \phi^*(L)(1-L)^d,$$

where the order of integration d is allowed to take non-integer values, and $\phi^*(L) = 1 - \phi_1^* L - \cdots - \phi_p^* L^p$ is a p-th order lag polynomial assumed to have all roots outside the unit circle. For values of $0 < d < 1$, the series y_t exhibits long memory and for that reason ARFI models have been popular in the context of modeling and forecasting realized volatility. Second, we consider AR(p) models, where $\phi(L)$ in (2) is simply an unrestricted $p-$th order lag polynomial. Using longer lag lengths allows the AR(p) model to mimic long memory-type behaviour more closely. Here we consider lag lengths $p = 1$, 5, and 22, in order to examine the relevance of this issue from a forecasting perspective. Third, we use the Heterogeneous Autoregressive (HAR) model proposed by Corsi (2004), given by

$$y_t = \mu_t + \phi_d(y_{t-1} - \mu_{t-1}) + \phi_w(y_{t-1}^{(5)} - \mu_{t-1}^{(5)}) + \phi_m(y_{t-1}^{(22)} - \mu_{t-1}^{(22)}) + \varepsilon_t, \tag{3}$$

where $y_{t-1}^{(j)}$ and $\mu_{t-1}^{(j)}$ are the average realized volatilities and (unknown) means, respectively, over the previous j days. Note that the HAR model can be obtained from an AR(22) model by imposing restrictions on the autoregressive parameters in the lag polynomial $\phi(L)$. Given its (implicit) long lag length, the HAR model is capable of capturing long memory features quite accurately. At the same time, it is much more parsimonious than an unrestricted AR(22) model. The HAR model has been successfully applied to forecast realized volatility by Andersen, Bollerslev, and Diebold (2007a) among others.

It is common practice to set the mean μ_t equal to a constant, that is, $\mu_t = c$. However, to capture the salient features of S&P realized volatility, discussed in the previous section, we extend the model to allow for gradual level shifts, day-of-the-week and holiday effects, announcement effects and nonlinear effects of lagged returns. In the most general specification we set

$$\begin{aligned}\mu_t = c(t) &+ \beta_1|r_{t-1}| + \beta_2\mathtt{I}[r_{t-1}<0] + \beta_3|r_{t-1}|\mathtt{I}[r_{t-1}<0] \\ &+ \delta_1 D^*_{1,t} + \delta_2 D^*_{2,t} + \delta_4 D^*_{4,t} + \delta_5 D^*_{5,t} + \delta_6 D_{\mathrm{HOL},t} + \delta_7 D_{\mathrm{XMS},t} \\ &+ \delta_8 D_{\mathrm{EMP},t} + \delta_9 D_{\mathrm{PPI},t} + \delta_{10} D_{\mathrm{CPI},t} + \delta_{11} D_{\mathrm{FF},t} \\ &+ \delta_{12} D_{\mathrm{EMP},t+1} + \delta_{13} D_{\mathrm{PPI},t+1} + \delta_{14} D_{\mathrm{CPI},t+1} + \delta_{15} D_{\mathrm{FF},t+1}, \quad (4)\end{aligned}$$

where $D^*_{s,t} \equiv D_{s,t} - D_{3,t}$, $s = 1, 2, 4, 5$ are "centered" daily dummy variables, with $D_{s,t} = 1$ when calendar day t corresponds with weekday s (1=Monday, 2=Tuesday, etc.) and $D_{s,t} = 0$ otherwise. The dummies $D_{\mathrm{HOL},t}$ and $D_{\mathrm{XMS},t}$ are 1 when day t follows a holiday or is part of the Christmas period, respectively. In our choice of macroeconomic announcements we follow Ederington and Lee (1993), among others, by including dummies for the release of the Employment report ($D_{\mathrm{EMP},t}$), Consumer Price Index ($D_{\mathrm{CPI},t}$) and Producer Price Index ($D_{\mathrm{PPI},t}$).[4] In addition, we include a dummy for meeting days of the Federal Open Market Committee (FOMC) on which decisions regarding the Federal Funds target rate are announced ($D_{\mathrm{FF},t}$).[5] Similar to Bomfim (2003) we also include "calm-before-the-storm" dummies ($D_{\mathrm{EMP},t+1}$, $D_{\mathrm{CPI},t+1}$, $D_{\mathrm{PPI},t+1}$, $D_{\mathrm{FF},t+1}$) to allow for volatility to be depressed on pre-announcement days.[6] Finally, we capture gradual level shifts in volatility by specifying the parameter $c(t)$ as an arbitrary deterministic trend function which is estimated semi-parametrically using kernel regression, see Beran and Ocker (2001) for further details. The models with μ_t specified as in (4), that is, with Semi-Parametric mean (SP), holiday and day-of-the-week Dummies (D), macro news (pre-)Announcement dummies (A), lagged Return (R), and Leverage effects (L) will be denoted SP[Z]-DARL, where Z is either ARFI, AR(p) or HAR.

We also estimate an alternative model, where (cf. Ebens, 1999) terms involving the lagged returns are not included in the conditional mean μ_t, but as "exogenous regressors" (X), leading to the model (denoted SP[Z]-DAXRL)

$$\phi(L)(y_t - \mu_t) = \beta_1 |r_{t-1}| + \beta_2 \mathtt{I}[r_{t-1} < 0] + \beta_3 |r_{t-1}| \mathtt{I}[r_{t-1} < 0] + \varepsilon_t, \tag{5}$$

where μ_t is now given by

$$\begin{aligned}\mu_t = c(t) &+ \delta_1 D^*_{1,t} + \delta_2 D^*_{2,t} + \delta_4 D^*_{4,t} + \delta_5 D^*_{5,t} + \delta_6 D_{\mathrm{HOL},t} + \delta_7 D_{\mathrm{XMS},t} \\ &+ \delta_8 D_{\mathrm{EMP},t} + \delta_9 D_{\mathrm{PPI},t} + \delta_{10} D_{\mathrm{CPI},t} + \delta_{11} D_{\mathrm{FF},t} \\ &+ \delta_{12} D_{\mathrm{EMP},t+1} + \delta_{13} D_{\mathrm{PPI},t+1} + \delta_{14} D_{\mathrm{CPI},t+1} + \delta_{15} D_{\mathrm{FF},t+1}.\end{aligned} \tag{6}$$

To gauge the relative importance of the different nonlinear features of realized volatility, we also estimate several restricted versions of the full models. In particular, we consider (i) a model without the leverage effect but including the lagged absolute return ($\beta_2 = \beta_3 = 0$ in (4) and (5); SP[Z]-DA(X)R), (ii) a model without any effect of lagged returns at all ($\beta_1 = \beta_2 = \beta_3 = 0$ in (4) and (5); SP[Z]-DA), (iii) a model without any effect of lagged returns and without (pre-)announcement effects (imposing in addition that $\delta_i = 0$, for $i = 8, \ldots, 15$ in (4) and (6); SP[Z]-D) and (iv) a model without any effect of lagged returns, (pre-)announcement effects and seasonal effects (thereby also imposing $\delta_1 = \cdots = \delta_7 = 0$ in (4) and (6); SP[Z]). Finally, we also estimate all models without allowing for structural changes (by imposing that $c(t)$ is equal to a constant, $c(t) = c$). Those models are indicated by omitting the prefix SP. In total, 70 different models are considered.

4 Forecasting Volatility

The period from January 2, 1998 through December 29, 2006 ($P = 2232$ observations) is used to judge the relevance of modeling the nonlinearities in S&P 500 realized volatility for out-of-sample forecasting purposes. All models are estimated recursively, using an expanding window of data.[7,8] Volatility forecasts for 1- to 20-days ahead are constructed by means of Monte Carlo simulation.[9] In addition to forecasts for logarithmic realized volatility, which is the dependent variable in the various models, we also construct forecasts for the realized variance and realized standard deviation, where we ensure that these forecasts are unbiased.[10] In the forecast evaluation below, we concentrate on forecasts of the standard deviation. Results for forecasts of the (logarithmic) variance are qualitatively similar and are available upon request. Furthermore, h-days ahead forecasts refer to realized standard deviations over the next h days, that is, $\widehat{s}_{t+h|t} = \sqrt{\sum_{j=1}^{h} \widehat{s}^2_{t+j|t}}$ (instead of daily realized standard deviation h-days ahead).

4.1 Alternative forecasting models

For comparison purposes, we also produce forecast for two popular models that only use daily returns and treat volatility as a latent variable. First, Riskmetrics uses historical volatility with exponentially declining weights,

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2, \tag{7}$$

with $\lambda = 0.94$.

Second, Glosten, Jagannathan, and Runkle (1993) (GJR) incorporate leverage effects into the popular Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model. Also including day-of-week dummies, calm-before-the-storm dummies and announcement dummies, the GJR-GARCH(1,1)-DA model is specified as

$$\begin{aligned}
r_t &= c + u_t && (8)\\
u_t|\Psi_{t-1} &\sim N(0,\sigma_t^2) && (9)\\
\sigma_t^2 &= \omega + \delta_1 D_{1,t}^* + \delta_2 D_{2,t}^* + \delta_4 D_{4,t}^* + \delta_5 D_{5,t}^* + \delta_6 D_{\mathrm{HOL},t} + \delta_7 D_{\mathrm{XMS},t} \\
&\quad + \delta_8 D_{\mathrm{EMP},t} + \delta_9 D_{\mathrm{PPI},t} + \delta_{10} D_{\mathrm{CPI},t} + \delta_{11} D_{\mathrm{FF},t} \\
&\quad + \delta_{12} D_{\mathrm{EMP},t+1} + \delta_{13} D_{\mathrm{PPI},t+1} + \delta_{14} D_{\mathrm{CPI},t+1} + \delta_{15} D_{\mathrm{FF},t+1} \\
&\quad + \alpha u_{t-1}^2 + \gamma u_{t-1}^2 \mathtt{I}[u_{t-1} < 0] + \beta\sigma_{t-1}^2, && (10)
\end{aligned}$$

where Ψ_{t-1} is the information set containing all daily information up to day t, and the error terms u_t are assumed to follow a conditional normal distribution with mean zero and variance σ_t^2.

4.2 Forecast evaluation

Evaluation of (realized) volatility forecasts has received considerable attention recently, see Andersen, Bollerslev, and Meddahi (2004, 2005), Patton (2006) and Patton and Sheppard (2007) for recent contributions. We use several out-of-sample performance measures to evaluate and compare the various forecasting models. First, the quality of individual forecasts is assessed by the Mincer-Zarnowitz type regression of the observed h-day realized standard deviation on the corresponding forecast, that is,

$$s_{t+h|t+1} = \sqrt{\sum_{j=1}^{h} s_{t+j}^2} = b_0 + b_1 \widehat{s}_{t+h|t} + \nu_t, \tag{11}$$

where b_0 and b_1 should be equal to 0 and 1, respectively, for the forecast to be considered optimal. Following Patton and Sheppard (2007), we estimate the Mincer-Zarnowitz regression using Generalized Least Squares (GLS), effectively using the form $\frac{s_{t+h|t+1}}{\widehat{s}_{t+h|t}} = b_0 \frac{1}{\widehat{s}_{t+h|t}} + b_1 + \nu_t^*$. In addition, a number of popular error metrics are computed, namely the Mean Squared Prediction Error (MSPE; $MSPE = \frac{1}{P}\sum_{t=R}^{R+P-1}(s_{t+h|t+1} - \widehat{s}_{t+h|t})^2$ where R denotes the first forecast origin and P denotes the number of forecasts), the Mean Absolute Error (MAE; $MAE = \frac{1}{P}\sum_{t=R}^{R+P-1}|s_{t+h|t+1} - \widehat{s}_{t+h|t}|$), and the Heteroskedasticity-adjusted MSPE (HMSPE; $HMSPE = \frac{1}{P}\sum_{t=R}^{R+P-1}(1 - \frac{s_{t+h|t+1}}{\widehat{s}_{t+h|t}})^2$). In most cases we will focus the discussion on the MSPE results, with Patton (2006) showing this is the most reliable metric given that the forecast target is a proxy for volatility. Finally, we also report the Mean Error (ME; $ME = \frac{1}{P}\sum_{t=R}^{R+P-1}(s_{t+h|t+1} - \widehat{s}_{t+h|t})$) to assess the bias in the volatility forecasts.

4.3 Forecast comparison

To test for significant differences in forecast accuracy between competing models we apply the Model Confidence Set (MCS) approach developed by Hansen *et al.* (2003, 2005). Given a set of forecasting models, $\mathcal{M}_0$, the goal of the MCS procedure is to identify the MCS $\widehat{\mathcal{M}}_{1-\alpha}^* \subset \mathcal{M}_0$, which is the set of models that contains the "best" forecasting model given a level of confidence α.

Starting with the full set of models, $\mathcal{M} = \mathcal{M}_0$, the MCS procedure repeatedly tests the null hypothesis of equal forecasting accuracy,

$$H_{0,\mathcal{M}}: \ \mathsf{E}[d_{ij,t}] = 0, \ \text{for all } i, j \in \mathcal{M},$$

where $d_{ij,t} = L_{i,t} - L_{j,t}$ is the loss differential between models i and j in the set, with L an appropriate loss function. The MCS procedure sequentially eliminates the worst performing model from $\mathcal{M}$ as long as the null is rejected. This trimming of models is repeated until the null is not rejected any longer and the surviving set of models then form the Model Confidence Set, $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Hansen *et al.* (2005) discuss three different test statistics to test the null hypothesis $H_{0,\mathcal{M}}$. We follow Hansen *et al.* (2003) by applying the MCS procedure with two of these; the range statistic, T_R, and the semi-quadratic statistic, T_{SQ}. Both test statistics are based on the following t-statistics (using $\bar{d}_{ij} = \frac{1}{P}\sum_{t=R}^{R+P-1} d_{ij,t}$),

$$t_{ij} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\text{var}}(d_{ij,t})/P}} \quad \text{for } i, j \in \mathcal{M}, \tag{12}$$

and are given by

$$T_R \equiv \max_{i,j\in\mathcal{M}} |t_{ij}| \quad \text{and} \quad T_{SQ} \equiv \sum_{i,j\in\mathcal{M}} t_{ij}^2. \tag{13}$$

Following Hansen *et al.* (2005), we implement the MCS procedure using the stationary block bootstrap of Politis and Romano (1994) with an average block length of 20 days and we choose the squared forecast error for the loss function L.

The MCS procedure assigns p-values to each model in the initial set. For a given model $i \in \mathcal{M}_0$, the MCS p-value, $\widehat{p}_i$, is the threshold confidence level that determines whether the model belongs to the MCS or not. It holds that $i \in \widehat{\mathcal{M}}^*_{1-\alpha}$ if and only if $\widehat{p}_i \geq \alpha$. We report results for confidence levels of 10% and 25%.

In addition to applying the MCS approach to identify the best performing models from a set of models, we also apply it pairwise to test for significant differences between two competing models, like the Diebold and Mariano (1995) test of equal forecast accuracy. In this case the reported p-values are based on the T_{SQ} test.

Note that a number of issues need to be addressed when testing for significant differences in forecast error metrics between models. These include comparing forecast which are based on estimated parameters, the choice of estimation window and comparing forecasts from nested models, see West (2006) for a discussion. As noted by Hansen *et al.* (2005), the MCS procedure is also potentially affected when it is used for comparing forecasts which are based on estimated parameters, but it nevertheless seems informative to apply the MCS procedure here.

4.4 Forecast results

Results for the one-day ahead forecasts are presented in Table 4. To save space we only report results for a selection of the models considered, which are chosen to reflect our two main research objectives.[11] In order to provide insight in the added value of incorporating the different nonlinear features into the forecasting model, we provide results for all ARFI specifications that do not allow for structural change. For comparing the different autoregressive specifications, we examine the AR(p)- and HAR-DAXRL specifications, which include all nonlinear features except level shifts. The latter is considered via the SPARFI-DAXRL model.

- insert Table 4 about here -

Several interesting conclusions emerge from Table 4. First, we confirm the findings of previous studies (e.g. Koopman *et al.*, 2005), in that the models for realized volatility produce volatility forecasts that are superior to popular models based on daily data. For example, the R^2 of the Mincer-Zarnowitz (MZ) regression (11) for the "basic" ARFI model is 60.0% compared to 51.9% for Riskmetrics. Including the leverage effect, seasonal dummies and (pre-)announcement dummies in the GJR-GARCH(1,1) model increases its regression R^2 to 57.5%, but this is still short of the R^2 of even the simplest ARFI model.

Second, explicitly accounting for the various stylized facts of volatility in the ARFI model enhances forecast accuracy.[12] The R^2-s of the MZ regression show an increasing pattern when adding non-linearities. In general, allowing for a leverage effect contributes most of all features to the improvement in forecast performance over the linear ARFI model, where including this effect exogenously (ARFI-DAXRL) seems better than including it in μ_t (ARFI-DARL). The MSPE, MAE and HMSPE criteria confirm this conclusion. Incorporating (pre-)announcement effects increases the R^2 by 0.3% to 60.6%, despite there being only 387 announcement days out of 2232 days in total. Unreported results show that on these days MSPE, MAE and HMSPE are about twice as high as on non-announcement days. Interestingly, allowing for (pre-)announcement effects not only improves forecast accuracy on announcement days, but also (and sometimes even more) on non-announcement days.

According to all three criteria the most accurate forecasts are those obtained from the ARFI-DAXRL specification. It is also interesting to note that the estimate of the forecast coefficient b_1 in the MZ regression declines from 1.024 for the basic ARFI specification to 0.979 for the most elaborate ARFI-DAXRL specification. This is due to the fact that introducing nonlinear components in the ARFI model renders more volatile forecasts, as demonstrated by an (unreported) increase in the forecast variance from 0.162 for the basic ARFI model to 0.216 for the ARFI-DAXRL specification. In addition, the first-order autocorrelation of the forecasts declines from 0.968 to 0.790 for these models. Hence, we find that models for RV that produce smoother forecasts are less accurate, which is opposite from the conclusion reached by Koopman *et al.* (2005).

Third, incorporating structural changes adds little or nothing in terms of forecast accuracy. The R^2 of the MZ-regression are identical for the ARFI-DAXRL and SPARFI-DAXRL models, while according to the MSPE and MAE criteria including level shifts in fact leads to slightly less accurate forecasts.

Fourth, at least for forecasting purposes it seems better not to model the long-memory feature of volatility with a parsimonious but strict ARFI-specification but rather in a more flexible manner by means of a long AR-approximation. This conclusion follows from observing in Table 4 that the AR(22)-DAXRL model gives slightly better results with a higher R^2 of 62.9% and slightly lower values for MSPE. The AR(22)-specification also outperforms the HAR model in these respects. The MCS p-values reported in the table allow for a comparison between the forecasts of the AR(22)-DAXRL model with those of all other models (on a one-to-one basis). We observe that the differences in MSPE with most ARFI specifications are fairly small and give rise to moderately large p-values, such that the null hypothesis of equal forecast accuracy cannot be rejected at conventional significance levels (except for the simplest ARFI specifications). By contrast, p-values tend to be smaller when forecasts are compared by means of MAE or HMSPE, such that at conventional significance levels we reject the null hypothesis in favor of the AR(22)-DAXRL specification producing significantly more accurate forecasts.

The results for 5-, 10- and 20-days ahead forecasts of realized standard deviation are summarized in Table 5. The benefits of modelling the nonlinearities remain present for longer horizons, although not surprisingly their importance diminishes somewhat. Comparing the ARFI and ARFI-DAXRL models, the MZ regression R^2 increases from 71.6% to 74.0% for the 5-day horizon, from 70.3% to 72.2% for the 10-day horizon, and from 64.8% to 65.9% for the 20-day horizon. The AR(22) model that includes all non-linearities remains the best forecasting model at all horizons. Note that in particular for the 5-day horizon the AR(22) model provides significantly more accurate forecasts than the HAR and ARFI specifications (as judged by the p-values below the MSPE values in Table 5), again showing the advantage of employing a flexible high-order AR model for accommodating the long memory behavior of RV.

- insert Table 5 about here -

To compare forecast accuracy among multiple models we apply the MCS procedure to the models listed in Table 4 with Table 6 showing the results based on the squared error loss function. As expected from the relatively small differences in MSPE discussed above, for all horizons except 5-days ahead the model confidence sets consists of several models in addition to the top-performing AR(22)-DAXRL specification.[13] We do observe though that generally the confidence set does not include the models based on daily returns and, for horizons beyond one-day ahead, the SPARFI-DAXRL model that allows for level shifts.

Table 7 compares different models, all including the full array of nonlinearities, as well as the two models that are based on daily returns. The confidence sets are now much smaller with again the models that allow for level shifts never being included for longer forecast horizons.

- insert Tables 6 and 7 about here -

4.5 Value-at-Risk forecasts

As an alternative method for evaluating the forecasts of the different models from a more economic point of view, we consider Value-at-Risk (VaR) estimates which are constructed using the volatility forecasts of the different models, see also Giot and Laurent (2004). Under the 1998 Market Risk Amendment (MRA) to the Basle Capital Accord, commercial banks are required to reserve capital to cover their market risk exposure. The required market risk capital for time $t+1$ (MRC_{t+1}) is determined by a bank's 99% VaR estimate calculated for a 10-day holding period (VaR_t^{10}) and is defined as the larger of the previous day's VaR estimate or the average of the estimates over the last sixty business days times a multiplication term[14]

$$MRC_{t+1} = \max(VaR_t^{10}, S_t \times \frac{1}{60} \sum_{i=0}^{59} VaR_{t-i}^{10}), \tag{14}$$

using the one-step ahead volatility forecasts from the different models.

Under the MRA, the accuracy of a bank's VaR estimates is judged by the accuracy of the 1-day ahead 99% VaR estimates.[15] We evaluate the accuracy of these estimates using the interval forecast evaluation techniques proposed by Christoffersen (1998) and the method set forth by Lopez (1999) using regulatory loss functions. The Christoffersen method is a test for correct conditional coverage, jointly testing the null hypotheses that the percentage of times that the actual loss on a portfolio exceeds the VaR estimate (denoted by a VaR 'violation') equals one minus the significance level used in the VaR calculation (unconditional coverage) together with serial independence for these VaR violations (independence).

Defining the indicator variable I_{t+1} as

$$I_{t+1} = \begin{cases} 1 & \text{if } r_{t+1} < VaR_t^1 \\ 0 & \text{if } r_{t+1} \geq VaR_t^1, \end{cases} \tag{15}$$

where r_{t+1} is the return over day $t+1$ and VaR_t^1 is the 100(1-α)% VaR estimate for day $t+1$ made on day t. A VaR violation corresponds with $I_{t+1} = 1$. The likelihood of observing x violations in a series of length R under the null hypothesis of accurate unconditional coverage is given by $L_0 = \alpha^x(1-\alpha)^{R-x}$. The corresponding likelihood under the alternative is $L_1 = \hat{\alpha}^x(1-\hat{\alpha})^{R-x}$, where $\hat{\alpha}$ is the "observed" probability of a violation, $\hat{\alpha} = x/R$. The null hypothesis of correct unconditional coverage can then be tested by means of the standard likelihood ratio statistic

$$LR_{uc} = 2[\log(L_1) - \log(L_0)], \tag{16}$$

which has an asymptotic $\chi^2(1)$ distribution.

The test statistic for the test of independence is also $\chi^2(1)$ distributed and is given by

$$LR_{ind} = 2[\log(L_2) - \log(L_1)], \tag{17}$$

where L_1 is the likelihood function under independence as given above, and L_2 is the likelihood function under the alternative of first-order Markov dependence. The latter is given by $L_2 = (1-\pi_{01})^{R_{00}}\pi_{01}^{R_{01}}(1-\pi_{11})^{R_{10}}\pi_{11}^{R_{11}}$ where R_{ij} is the number of observations with value i followed by value j and where π_{01} and π_{11} are the transition probabilities which are estimated as $\hat{\pi}_{01} = \frac{R_{01}}{R_{00}+R_{01}}$ and $\hat{\pi}_{11} = \frac{R_{11}}{R_{10}+R_{11}}$ respectively.

Testing for correct conditional coverage boils down to testing for correct unconditional coverage and independence simultaneously. The likelihood functions under the null and alternative are given by L_0 and L_2, respectively and, hence, the likelihood ratio statistic for correct conditional coverage is simply the sum of the two previous statistics,

$$LR_{cc} = LR_{uc} + LR_{ind}, \tag{18}$$

which is asymptotically $\chi^2(2)$ distributed.

An alternative method proposed by Lopez (1999) for evaluating VaR forecasts is based on the use of loss functions that are more closely related to the regulatory VaR requirements. By choosing a specific loss function, one can assign a numerical score to each individual VaR estimate that reflects any specific concerns that one may have. For example, it seems natural to not only consider the number of VaR violations but also the magnitude of violations since the latter can be quite substantial. Therefore, we consider the loss function[16]

$$C_{t+1} = \begin{cases} 1 + (r_{t+1} - VaR_t^1)^2 & \text{if } r_{t+1} < VaR_t^1 \\ 0 & \text{if } r_{t+1} \geq VaR_t^1, \end{cases} \tag{19}$$

which includes the squared magnitude of the VaR violation, $(r_{t+1} - VaR_t^1)^2$. Given a sample of P VaR estimates the total loss for the sample is given by $C = \sum_{i=1}^{P} C_{t+1}$.

To assess the relative performance of each volatility forecasting model, we compute the average and standard deviation of the capital requirement MRC_{t+1} for each model over the entire forecast evaluation period. Furthermore, we determine the unconditional coverage, $\hat{\alpha}$, together with the interval evaluation test statistics.

Finally, we compute the total loss, C, for the sample of P one-day VaR estimates as well as the average score (defined as the total score divided by the number of violations), and the maximum daily score. The results are presented in Table 8.

- insert Table 8 about here -

We first of all observe that the average capital requirement is comparable across the different models. However, the long memory models typically have considerably less fluctuation in the required level of capital than the popular Riskmetrics method, as judged from the second column in Table 8. This is despite the fact that the actual VaR estimates of the Riskmetrics model are much less responsive to the variation in returns, as is evident from Figure 3, which shows timeseries of the VaR estimates of the Riskmetrics, GJR-GARCH(1,1)-DA and AR(22)-DAXRL models.

- insert Figure 3 about here -

All models have a higher unconditional coverage than expected, leading to strong rejections of the null of correct unconditional coverage in all cases. By contrast, the null of independence is not rejected for any of the models under consideration. Based on the quadratic magnitude loss function, the nonlinear realized volatility models again perform well when compared to the GARCH type models. Figure 4 shows the 1-day VaR estimates from the latter two models together with the AR(22)-DAXRL model.

- insert Figure 4 about here -

What is most obvious from these graphs is that the GARCH-type VaR estimates track the general pattern, or low-frequency movements in volatility quite well. The main advantage of the long memory model is that, in addition, it captures a much larger part of the high-frequency movements in volatility.

5 Concluding Remarks

In this paper we evaluated the forecasting performance of several time series models for realized volatility, by examining the importance of accounting for well-known stylized facts from the volatility literature, in particular long memory, level shifts, leverage effects, day-of-the-week and holiday effects, and macroeconomic announcement effects.

Our analysis rendered two main conclusions. First, the out-of-sample results demonstrate that for forecast horizons of up to 20 days, accounting for the effects of leverage, seasonality, and macroeconomic announcements generally improves the forecast performance compared to models without these effects. Including level shifts does not lead to more accurate forecasts. Second, it seems beneficial to model the long-memory feature of realized volatility by means of a flexible high-order AR-approximation instead of a parsimonious but stringent fractionally integrated specification or a heterogenous autoregressive model, especially for longer horizon forecasts.

References

Andersen, T. G. and T. Bollerslev (1997), Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns, *Journal of Finance*, **52**, 975–1005.

Andersen, T. G. and T. Bollerslev (1998a), Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts, *International Economic Review*, **39**, 885–905.

Andersen, T. G. and T. Bollerslev (1998b), Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies, *Journal of Finance*, **53**, 219–265.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2007a), Roughing It Up: Including Jump Components in the Measurement, Modeling and Forecasting of Return Volatility, *Review of Economics and Statistics*, **89**, 701–720.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001a), The Distribution of Realized Stock Return Volatility, *Journal of Financial Economics*, **61**, 43–76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001b), The Distribution of Realized Exchange Rate Volatility, *Journal of the American Statistical Association*, **96**, 42–55.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003a), Modeling and Forecasting Realized Volatility, *Econometrica*, **71**, 579–625.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and C. Vega (2003b), Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange, *American Economic Review*, **93**, 38–62.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and C. Vega (2007b), Real-Time Price Discovery in Global Stock, Bond and Foreign Exchange Markets, *Journal of International Economics*, **73**, 251–277.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2004), Analytic Evaluation of Volatility Forecasts, *International Economic Review*, **45**, 1079–1110.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2005), Correcting the Errors: Volatility Forecast Evaluation Using on High-Frequency Data and Realized Volatilities, *Econometrica*, **73**, 279–296.

Andreou, E. and E. Ghysels (2002), Detecting Multiple Breaks in Financial Market Volatility Dynamics, *Journal of Applied Econometrics*, **17**, 579–600.

Baillie, R. and T. Bollerslev (1989), The Message in Daily Exchange Rates: A Conditional Variance Tale, *Journal of Business and Economic Statistics*, **7**, 279–305.

Bandi, F. M. and J. R. Russell (2006), Separating Microstructure Noise from Volatility, *Journal of Financial Economics*, **79**, 655–692.

Bandi, F. M. and J. R. Russell (2008), Microstructure Noise, Realized Variance, and Optimal Sampling, *Review of Economic Studies*, **75**, 339–369.

Barndorff-Nielsen, O. E. and N. Shephard (2004), Power and Bipower Variation with Stochastic Volatility and Jumps, *Journal of Financial Econometrics*, **2**, 1–48.

Beran, J. (1995), Maximum Likelihood Estimation of the Differencing Parameter for Invertible Short and Long Memory Autoregressive Integrated Moving Average Models, *Journal of the Royal Statistical Society*, **B 57**, 659–672.

Beran, J. and D. Ocker (2001), Volatility of Stock Market Indices - An Analysis Based on SEMIFAR Models, *Journal of Business & Economic Statistics*, **19**, 103 – 116.

Black, F. (1976), Studies of Stock Market Volatility Changes, *Proceedings of the American Statistical Association, Buinsess and Economic Statistics Section*, 177–181.

Bollerslev, T., U. Kretschmer, C. Pigorsch, and G. Tauchen (in press), A Discrete-Time Model for Daily S&P500 Returns and Realized Variations: Jumps and Leverage Effects, *Journal of Econometrics*.

Bollerslev, T., J. Litvinova, and G. Tauchen (2006), Leverage and Volatility Feedback Effects in High-Frequency Data, *Journal of Financial Econometrics*, **4**, 353–384.

Bomfim, A. N. (2003), Pre-Announcement Effects, News Effects, and Volatility; Monetary Policy and the Stock Market, *Journal of Banking & Finance*, **27**, 133–151.

Christoffersen, P. F. (1998), Evaluating Interval Forecasts, *International Economic Review*, **39**, 841–862.

Corsi, F. (2004), A Simple Long Memory Model of Realized Volatility, *Working paper, University of Southern Switserland*.

Diebold, F. X. and A. Inoue (2001), Long Memory and Structural Change, *Journal of Econometrics*, **105**, 131–159.

Diebold, F. X. and R. S. Mariano (1995), Comparing Predictive Accuracy, *Journal of Business & Economic Statistics*, **13**, 253–263.

Ebens, H. (1999), Realized Stock Volatility, *unpublished manuscript, Johns Hopkins University*.

Ederington, L. H. and J. H. Lee (1993), How Markets Process Information: News Releases and Volatility, *Journal of Finance*, **48**, 1161–1191.

Engle, R. F. and V. K. Ng (1993), Measuring and Testing the Impact of News on Volatility, *Journal of Finance*, **48**, 1749–1778.

Giot, P. and S. Laurent (2004), Modelling Daily Value-at-Risk Using Realized Volatility and ARCH Type Models, *Journal of Empirical Finance*, **11**, 379–398.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993), On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, *Journal of Finance*, **48**, 1779–1801.

Hansen, P. R. and A. Lunde (2006), Realized Variance and Market Microstructure Noise, *Journal of Business and Economic Statistics*, **24**, 127–218.

Hansen, P. R., A. Lunde, and J. M. Nason (2003), Choosing the Best Volatility Model: The Model Confidence Set Approach, *Oxford Bulletin of Economics and Statistics*, **65**, 839–861.

Hansen, P. R., A. Lunde, and J. M. Nason (2005), Model Confidence Sets for Forecasting Models, *Federal Reserve Bank of Atlanta, Working paper 2005-7*.

Harvey, C. R. and R. D. Huang (1991), Volatility in the Foreign Currency Futures Market, *Review of Financial Studies*, **4**, 543–569.

Koopman, S. J., B. Jungbacker, and E. Hol (2005), Forecasting Daily Variability of the S&P 100 Stock Index Using Historical Realised and Implied Volatility Measurements, *Journal of Empirical Finance*, **12**, 445–475.

Lamoureux, C. G. and W. D. Lastrapes (1990), Persistence in Variance, Structural Change, and the GARCH Model, *Journal of Business and Economic Statistics*, **8**, 225–234.

Lopez, J. A. (1999), Methods for Evaluating Value-at-Risk Estimates, *Federal Reserve Bank of San Francisco*, **2**, 3–17.

Maheu, J. M. and T. H. McCurdy (2002), Nonlinear Features of Realized FX Volatility, *Review of Economics and Statistics*, **84**, 668–681.

Martens, M. and J. Zein (2004), Predicting Financial Volatility: High-Frequency Time-Series Forecasts Vis-à-Vis Implied Volatility, *Journal of Futures Markets*, **24**, 1005–1028.

Merton, R. C. (1980), On Estimating the Expected Return on the Market; An Exploratory Investigation, *Journal of Financial Economics*, **8**, 323–361.

Oomen, R. C. A. (2002), Modelling Realized Variance when Returns are Serially Correlated, *Working paper, University of Warwick*.

Pagan, A. R. and G. W. Schwert (1990), Alternative Models for Conditional Stock Volatility, *Journal of Econometrics*, **45**, 267–290.

Patton, A. J. (2006), Volatility Forecast Comparison Using Imperfect Volatility Proxies, *Working paper, London School of Economics*.

Patton, A. J. and K. Sheppard (2007), Evaluating Volatility and Correlation Forecasts, in T. Andersen, R. Davis, J.-P. Kreiss, and T. Mikosch (eds.), *Handbook of Financial Time Series*, Springer-Verlag, Berlin, *forthcoming*.

Politis, D. N. and J. P. Romano (1994), The Stationary Bootstrap, *Journal of the American Stastical Association*, **89**, 1303–1313.

Pong, S., M. B. Shackleton, S. J. Taylor, and X. Xu (2004), Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models, *Journal of Banking and Finance*, **28**, 2541–2563.

Basle Committee on Banking Supervision (1996), Supervisory Framework for the Use of Backtesting in Conjunction with the Internal Models Approach to Market Risk Capital Requirements, *Bank for International Settlements Publication No. 22*.

West, K. D. (2006), Forecast Evaluation, in G. Elliot, C. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, North-Holland, pp. 99–134.

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

Notes

[1]We repeat the entire empirical analysis using the two-time scales estimator of Zhang *et al.* (2005) and the bipower variation estimator of Barndorff-Nielsen and Shephard (2004). The results for the two-time scales estimator, which are available in the web appendix (Appendix B), are very similar to the ones reported here for the standard realized variance measure, suggesting that microstructure noise effects do not affect our main conclusions. The same holds for the bipower variation estimator, see Appendix C in the web appendix, so that our results also appear to be robust to the treatment of jumps.

[2]Holidays include New Year's Day, Martin Luther King Day, President's Day, Good Friday, Memorial Day, Independence Day, Labor Day, Thanksgiving and Christmas Day. We also consider the week between Christmas and New Year separately.

[3]As discussed below, we consider announcements of employment, producer and consumer prices, and Federal funds target rate decisions.

[4]As pointed out by the referee, one would expect CPI and PPI inflation measures to be highly correlated, such that the first announced figure captures most if not all relevant information and, possibly, affects stock market volatility. This is not what we find, however. In our sample of 156 months, CPI is announced before PPI only 12 times, suggesting that the CPI announcement dummy may be redundant in our models. However, for most models and estimation windows, we find that the CPI dummy coefficient is larger than the PPI coefficient (0.10 compared to 0.05, approximately). In addition, the CPI coefficient estimate generally is significant with t-statistics ranging between 2 and 3, while the PPI coefficient often is not significantly different from zero.

[5]As of 1994 the Federal Open Market Committee started to announce the decision about changing the target rate following FOMC meetings. There are eight regularly scheduled FOMC meetings each year. During our sample period there have been five unscheduled meetings: April 18, 1994; October 15, 1998; January 3, 2001; April 18, 2001 and September 17, 2001. For these days we also set $D_{\mathrm{FF},t} = 1$, but only once these dates are included in the in-sample estimation period. Furthermore, for two-day FOMC meetings we set $D_{\mathrm{FF},t} = 1$ only for the second day.

[6]Bomfim (2003) examines the news effect of monetary policy decisions on S&P 500 index returns and therefore only includes a calm-before-the-storm effect surrounding FOMC meetings. Here we also allow volatility to be different on pre-announcement days for employment, PPI, and CPI.

[7]To estimate the parameters in the ARFI models we use the Beran (1995) approximate maximum likelihood (AML) estimator. We employ the Akaike Information Criterion (AIC) in the most general ARFI specification to select the appropriate autoregressive order p. The selected lag order $p = 2$ is subsequently imposed in the nested models.

[8]The estimation results largely confirm the stylized facts of realized volatility discussed in Section 2. Based on the full sample period we find significant seasonality effects due to the day-of-the-week (even after accounting for macroeconomic announcements), holidays and the Christmas period. The effects of macroeconomic announcements are even larger, with the employment announcement being the most important, both in terms of magnitude and significance. The dummy coefficient estimate ($\hat{\delta}_8$) generally varies around 0.36 (with t-statistics in the range of 5-6), compared to 0.22 (3-4) for the Federal funds target rate coefficient, 0.11 (2-3) for the CPI coefficient and 0.05-0.08 (1-2) for the PPI coefficient. The relative importance of the employment report also holds from a forecasting perspective, in the sense that forecast accuracy deteriorates most when the employment announcement is omitted from the volatility models. These results are not surprising, given that the employment report is usually one of the first macroeconomic releases that becomes available during a calendar month. We also find significant 'pre-announcement' effects for the employment report and in particular for Federal funds target rate decisions, with the estimate of δ_{15} generally being around -0.15. The leverage effect is strong, with negative lagged returns leading to a substantial increase in volatility while positive returns generally do not have significant effects. The estimate of the fractional differencing parameter d in the ARFI models ranges between 0.51 and 0.55 but with a significantly negative first-order autoregressive parameter coefficient ranging between -0.28 and -0.20, jointly reflecting the long memory of realized volatility. Similarly, we find small but slowly decaying coefficients in the AR(22) specifications. Detailed estimation results for the various ARFI

specifications are reported in Table A.1 in the web appendix.

[9]The general AR representation in (2) can be rewritten as $y_t = \hat{\mu}_t + \sum_{j=1}^{t-1} \phi_j (y_{t-j} - \hat{\mu}_{t-j}) + \varepsilon_t$, where the values of ϕ_j vary across the ARFI, AR(p) and HAR models. The 1-step ahead forecast, $y_{t+1|t}$, is obtained by sampling B independent random draws $z_{t+1}^{(i)}$, $i = 1, \ldots, B$, from a standard normal distribution, which are multiplied by the residual standard deviation $\hat{\sigma}_e$. The resulting shocks $e_{t+1}^{(i)} = z_{t+1}^{(i)} \hat{\sigma}_e$ are fed into the model to obtain a realization $y_{t+1|t}^{(i)} = \hat{\mu}_{t+1} + \sum_{j=1}^{t-1} \pi_j (y_{t+1-j} - \hat{\mu}_{t+1-j}) + e_{t+1}^{(i)}$. Finally, the 1-step ahead forecast $y_{t+1|t}$ is the mean across all B realizations, $y_{t+1|t} = \frac{1}{B} \sum_{i=1}^{B} y_{t+1|t}^{(i)}$. For multiple step ahead forecasts from models which include lagged returns in μ_t, these returns are simulated as well, by multiplying the standard deviation in the i-th path by another draw v_t from a standard normal distribution, e.g. $r_{t+1}^{(i)} = \sqrt{\exp(y_{t+1|t}^{(i)})} v_t$. Note that v_t represents the standardized returns, for which the assumption of normality does not seem unreasonable as shown in Section 2.

[10]This is achieved by applying the appropriate transformation to all simulated paths of log realized volatility individually, and then averaging. For example, the 1-step ahead forecast of the realized standard deviation is computed as $s_{t+1|t} = \frac{1}{B} \sum_{i=1}^{B} \sqrt{\exp(y_{t+1|t}^{(i)})}$.

[11]Detailed results for the complete suite of models are available upon request.

[12]The same conclusion holds for the AR(p) and HAR models, see the results in Table A.2 in the web appendix.

[13]Model confidence sets based on the alternative MAE and HMSPE loss functions are generally much smaller, often including only the AR(22)-DAXRL specification, together with the HAR-DAXRL or ARFI-DAXRL model, which suggests that the MSPE measure is affected by outliers. We also considered confidence sets starting from only the ARFI specifications, but based on MSPE and MAE is it difficult to distinguish between the different nonlinear specifications. Interestingly though, the confidence sets tend to be much smaller when these are based on the outlier and heteroscedasticity-robust HMSPE loss function. In particular, the ARFI-DAXRL model is the only model that is always included and it is often the sole model in the confidence set. Tables A.4-A.6 in the web appendix give detailed results.

[14]The value of the multiplication factor S_t is determined by the number of times that the actual losses on a portfolio exceed the 1-day 99% VaR estimates (so called VaR 'violations'). Three zones for an increasing number of violations are distinguished and the value of S_t increases from a minimum value of 3 to a possible maximum value of 4 across the different zones, see the Basle Committee on Banking Supervision (1996) for more details. In the evaluation of the VaR estimates, however, we fix S_t to the value 3. Note further that under the conditions of the MRA the VaR estimates are evaluated in dollar terms whereas we will consider the VaR in percentage terms.

[15]It is suggested by the Basle committee to use an evaluation period of at least 250 business days. Here we use the entire forecast evaluation sample.

[16]Lopez (1999) discusses several possible loss functions, one of which is the binomial loss function given in (15). The loss function given in the main text is in line with the guidelines by the Basle Committee which state that both the number as well as the magnitude of violations are a matter of concern to regulators.

Table 1: Descriptive statistics for daily S&P 500 return and realized volatility

	Mean	Med	Min	Max	Std.dev.	Skew	Kurt
Returns	0.025	0.054	−7.706	5.814	1.086	−0.134	7.022
Standardized returns	0.084	0.074	−2.831	3.009	0.998	0.016	2.641
Squared returns	1.181	0.317	0.000	59.379	2.894	8.683	127.379
Realized variance	1.181	0.672	0.031	33.430	1.771	6.882	83.671
Log realized variance	−0.340	−0.398	−3.485	3.509	0.952	0.370	3.033

Notes: The table contains summary statistics for daily S&P 500 returns and realized volatility measures (based on 5-minute intraday returns). The sample period covers January 3, 1994 until December 29, 2006 (3265 observations). Standardized returns are obtained by dividing the raw returns by the realized standard deviation.

Table 2: Descriptive statistics for daily S&P 500 return and realized volatility

	ALL	NONE	HOL	XMS	ANN	EMP	PPI	CPI	FF
Returns	0.025	0.003	0.046	0.007	0.131	0.170	0.116	0.048	0.244
Standardized returns	0.084	0.064	0.121	0.029	0.182	0.230	0.101	0.193	0.258
Squared returns	1.181	1.111	1.694	0.666	1.454	1.660	1.429	1.375	1.312
Realized variance	1.181	1.133	1.168	0.510	1.472	1.636	1.439	1.270	1.521
Log realized variance	−0.340	−0.369	−0.347	−0.992	−0.144	0.005	−0.238	−0.226	−0.135
Number of obs.	3265	2577	114	52	558	154	152	156	107

Notes: The table contains sample averages of daily S&P 500 returns and realized volatility measures (based on 5-minute intraday returns). The sample period covers January 3, 1994 until December 29, 2006 (3265 observations). ALL indicates all days in the sample period; NONE indicates days without announcements, not following a holiday, and not in the Christmas period. HOL indicates days following a holiday. XMS denotes days during the Christmas period; ANN indicates all days with one or more macroeconomic announcements; EMP, PPI, CPI and FF indicate days with an announcement of employment, PPI, CPI, and the Federal Funds target rate, respectively. Standardized returns are obtained by dividing the raw returns by the realized standard deviation.

Table 3: Day-of-the-week effects in S&P 500 return and realized volatility

	Overall	MON	TUE	WED	THU	FRI
All days						
Returns	0.025	0.046	0.042	0.012	0.000	0.028
Standardized returns	0.084	0.126	0.077	0.072	0.060	0.087
Squared returns	1.181	1.254	1.245	1.062	1.142	1.207
Realized variance	1.181	1.060	1.110	1.190	1.234	1.307
Log realized variance	−0.340	−0.461	−0.374	−0.310	−0.307	−0.253
Non-announcement days						
Returns	0.004	0.051	0.006	−0.005	−0.025	−0.018
Standardized returns	0.063	0.130	0.043	0.056	0.040	0.034
Squared returns	1.124	1.245	1.331	1.026	1.075	0.871
Realized variance	1.121	1.047	1.130	1.143	1.182	1.104
Log realized variance	−0.380	−0.467	−0.378	−0.327	−0.320	−0.413
Announcement days						
Returns	0.131	-1.309	0.187	0.117	0.199	0.099
Standardized returns	0.182	-0.871	0.215	0.169	0.217	0.169
Squared returns	1.454	3.853	0.895	1.279	1.669	1.728
Realized variance	1.472	5.210	1.029	1.472	1.648	1.621
Log realized variance	−0.144	1.476	−0.357	−0.212	−0.200	−0.005
Number of obs.	558	2	132	95	74	255

Notes: The table contains daily means for S&P 500 returns and (log) realized variance (based on 5-minute intraday returns). The sample period covers January 3, 1994 until December 29, 2006 (3265 observations). Standardized returns are obtained by dividing the raw returns by the realized standard deviation. The three panels distinguish between statistics computed using all days (top panel), days without any macro news announcements (middle panel) and days on which at least one macro figure is released (bottom panel). The final row in the table shows the total number of announcement days and how these are dispersed across the days of the week.

Table 4: Out-of-sample forecast evaluation, January 1998-December 2006 - standard deviation, one-day ahead

	ME	MSPE	MAE	HMSPE	b_0	b_1	R^2
Riskmetrics	−0.043	0.152	0.259	0.096	0.088	0.879	0.519
	(0.008)	[0.000]	[0.000]	[0.000]	(0.015)	(0.017)	
GJR-G-DA(1,1)	−0.047	0.133	0.243	0.094	0.070	0.889	0.575
	(0.008)	[0.002]	[0.000]	[0.001]	(0.016)	(0.018)	
ARFI	0.002	0.122	0.222	0.080	−0.024	1.024	0.600
	(0.007)	[0.089]	[0.009]	[0.002]	(0.015)	(0.017)	
ARFI-D	0.001	0.121	0.220	0.078	−0.015	1.014	0.603
	(0.007)	[0.112]	[0.036]	[0.003]	(0.014)	(0.017)	
ARFI-DA	0.001	0.120	0.220	0.075	−0.010	1.009	0.606
	(0.007)	[0.114]	[0.021]	[0.034]	(0.014)	(0.016)	
ARFI-DAR	0.000	0.120	0.219	0.075	−0.009	1.008	0.606
	(0.007)	[0.137]	[0.034]	[0.012]	(0.014)	(0.016)	
ARFI-DARL	−0.001	0.117	0.217	0.074	−0.009	1.008	0.617
	(0.007)	[0.356]	[0.065]	[0.004]	(0.014)	(0.016)	
ARFI-DAXRL	−0.018	0.115	0.216	0.069	0.005	0.979	0.626
	(0.007)	[0.224]	[0.009]	[0.951]	(0.013)	(0.015)	
SPARFI-DAXRL	−0.026	0.116	0.220	0.069	0.022	0.956	0.626
	(0.007)	[0.169]	[0.001]	[0.990]	(0.013)	(0.015)	
HAR-DAXRL	0.010	0.116	0.213	0.075	−0.015	1.028	0.622
	(0.007)	[0.316]	[0.541]	[0.000]	(0.014)	(0.017)	
AR(1)-DAXRL	0.031	0.142	0.241	0.107	−0.057	1.090	0.536
	(0.008)	[0.000]	[0.000]	[0.000]	(0.018)	(0.022)	
AR(5)-DAXRL	0.004	0.115	0.213	0.074	−0.015	1.022	0.625
	(0.007)	[0.488]	[0.813]	[0.005]	(0.014)	(0.016)	
AR(22)-DAXRL	−0.005	0.114	0.214	0.071	−0.002	1.000	0.629
	(0.007)				(0.013)	(0.016)	

Notes: The table presents estimates of regressions of the realized standard deviation for the S&P 500 on a constant and one-day ahead forecasts from different models. The regression is $s_{t+1} = b_0 + b_1 \widehat{s}_{t+1|t} + u_t$, where $\widehat{s}_{t+1|t}$ is the one-day ahead forecast of the realized standard deviation. The forecast evaluation period covers January 2, 1998 – December 29, 2006 ($P = 2232$). SPARFI-DAXRL refers to the full ARFI model including level shifts, seasonal dummies, (pre-)announcement dummies and leverage effects as part of μ_t, ARFI-DA(X)RL refers to the ARFI model without level shifts but including seasonal dummies, (pre-)announcement dummies and leverage effects as part of μ_t (as exogenous variables), ARFI-DAR refers to the model with seasonal dummies, (pre-)announcement dummies and only symmetric effects of the lagged absolute return as part of μ_t, ARFI-DA to the model with seasonal dummies, (pre-)announcement dummies but without the lagged absolute return, ARFI-D to the model with only seasonal dummies and ARFI to the model without any dummies or lagged absolute returns. HAR-DAXRL is the full HAR model and AR(p)-DAXRL is the full AR(p) model, where $p = 1$, 5 or 22. Figures in brackets below b_j, $j = 0, 1$, and ME are heteroskedasticity and autocorrelation-consistent standard errors. Figures in straight brackets below MSPE, MAE and HMSPE are MCS p-values of testing equal forecast accuracy using the T_{SQ} test, comparing the relevant model with the AR(22)-DAXRL model, where values below a certain confidence level (e.g. 5%) indicate that the AR(22)-DAXRL model is significantly more accurate.

Table 5: Out-of-sample forecast evaluation, January 1998-December 2006 - standard deviation, five, ten and twenty-days ahead

	5-days ahead				10-days ahead				20-days ahead			
	MSPE	b_0	b_1	R^2	MSPE	b_0	b_1	R^2	MSPE	b_0	b_1	R^2
Riskmetrics	0.097	0.104	0.890	0.628	0.092	0.112	0.881	0.631	0.096	0.147	0.866	0.602
	[0.000]	(0.012)	(0.014)		[0.000]	(0.012)	(0.014)		[0.000]	(0.012)	(0.014)	
GJR-G-DA(1,1)	0.084	0.021	0.966	0.667	0.080	−0.005	1.005	0.659	0.083	−0.055	1.063	0.620
	[0.001]	(0.014)	(0.016)		[0.056]	(0.015)	(0.017)		[0.243]	(0.017)	(0.018)	
ARFI	0.071	−0.046	1.030	0.716	0.070	−0.060	1.043	0.703	0.077	−0.078	1.067	0.648
	[0.016]	(0.012)	(0.013)		[0.134]	(0.012)	(0.014)		[0.588]	(0.014)	(0.015)	
ARFI-D	0.070	−0.043	1.027	0.721	0.069	−0.057	1.040	0.706	0.077	−0.075	1.065	0.650
	[0.034]	(0.012)	(0.013)		[0.194]	(0.012)	(0.013)		[0.664]	(0.014)	(0.015)	
ARFI-DA	0.070	−0.039	1.022	0.723	0.068	−0.055	1.037	0.710	0.076	−0.074	1.062	0.652
	[0.042]	(0.012)	(0.013)		[0.248]	(0.012)	(0.013)		[0.698]	(0.014)	(0.015)	
ARFI-DAR	0.070	−0.040	1.024	0.722	0.069	−0.056	1.038	0.709	0.077	−0.076	1.065	0.650
	[0.044]	(0.012)	(0.013)		[0.241]	(0.012)	(0.013)		[0.633]	(0.014)	(0.015)	
ARFI-DARL	0.070	−0.025	1.005	0.719	0.070	−0.035	1.013	0.702	0.078	−0.051	1.036	0.643
	[0.022]	(0.011)	(0.013)		[0.081]	(0.012)	(0.017)		[0.378]	(0.014)	(0.015)	
ARFI-DAXRL	0.067	0.004	0.952	0.740	0.069	0.009	0.937	0.722	0.081	0.015	0.928	0.659
	[0.013]	(0.011)	(0.012)		[0.042]	(0.011)	(0.012)		[0.079]	(0.012)	(0.013)	
SPARFI-DAXRL	0.072	0.038	0.911	0.731	0.079	0.059	0.880	0.703	0.100	0.087	0.849	0.625
	[0.003]	(0.010)	(0.011)		[0.007]	(0.011)	(0.012)		[0.007]	(0.012)	(0.012)	
HAR-DAXRL	0.068	−0.037	1.028	0.726	0.069	−0.049	1.039	0.706	0.078	−0.075	1.072	0.643
	[0.036]	(0.012)	(0.013)		[0.063]	(0.012)	(0.014)		[0.091]	(0.015)	(0.016)	
AR(1)-DAXRL	0.133	−0.601	1.599	0.536	0.162	−0.891	1.857	0.402	0.191	−0.276	1.279	0.151
	[0.001]	(0.033)	(0.034)		[0.064]	(0.058)	(0.056)		[0.000]	(0.090)	(0.086)	
AR(5)-DAXRL	0.067	−0.046	1.032	0.730	0.068	−0.104	1.083	0.710	0.080	−0.244	1.218	0.640
	[0.147]	(0.012)	(0.013)		[0.346]	(0.014)	(0.015)		[0.325]	(0.019)	(0.019)	
AR(22)-DAXRL	0.064	−0.013	0.988	0.748	0.064	−0.021	0.991	0.732	0.074	−0.041	1.013	0.669
		(0.011)	(0.012)			(0.012)	(0.013)			(0.014)	(0.015)	

Notes: The table presents estimates of regressions of realized standard deviation for the S&P 500 on a constant and either five, ten or twenty-days ahead forecasts from different models. The regression is $s_{t+h|t+1} = b_0 + b_1 \widehat{s}_{t+h|t} + u_t$, where $\widehat{s}_{t+h|t}$ is the forecast of h-days realized standard deviation with $h \in \{5, 10, 20\}$. The forecast evaluation period covers January 2, 1998 – December 29, 2006 ($P = 2232$). See Table 4 for further details.

Table 6: Model Confidence Sets, January 1998-December 2006 - standard deviation, one, five, ten and twenty-days ahead

	1-day ahead			5-days ahead			10-days ahead			20-days ahead		
	MSPE	T_R	T_{SQ}	MSPE	T_R	T_{SQ}	MSPE	T_R	T_{SQ}	MSPE	T_R	T_{SQ}
Riskmetrics	0.152	0.000	0.000	0.097	0.003	0.002	0.092	0.004	0.018	0.096	0.015	0.063
GJR-G-DA(1,1)	0.133	0.062	0.055	0.084	0.029	0.040	0.080	0.254**	0.249**	0.083	0.456**	0.490**
ARFI	0.122	0.600**	0.214*	0.071	0.037	0.087	0.070	0.254**	0.329**	0.077	0.456**	0.490**
ARFI-D	0.121	0.655**	0.268**	0.070	0.099	0.162*	0.069	0.324**	0.414**	0.077	0.456**	0.490**
ARFI-DA	0.120	0.655**	0.424**	0.070	0.099	0.162*	0.068	0.484**	0.496**	0.076	0.685**	0.681**
ARFI-DAR	0.120	0.655**	0.333**	0.070	0.099	0.162*	0.069	0.324**	0.428**	0.077	0.456**	0.490**
ARFI-DARL	0.117	0.692**	0.695**	0.070	0.099	0.162*	0.070	0.254**	0.329**	0.078	0.456**	0.490**
ARFI-DAXRL	0.115	0.692**	0.695**	0.067	0.099	0.162*	0.069	0.324**	0.428**	0.081	0.456**	0.490**
SPARFI-DAXRL	0.116	0.655**	0.665**	0.072	0.037	0.102*	0.079	0.053	0.122*	0.100	0.025	0.143*
HAR-DAXRL	0.116	0.692**	0.695**	0.068	0.099	0.162*	0.069	0.324**	0.428**	0.078	0.456**	0.490**
AR(1)-DAXRL	0.142	0.002	0.004	0.133	0.003	0.002	0.162	0.001	0.001	0.191	0.000	0.006
AR(5)-DAXRL	0.115	0.692**	0.695	0.067	0.144*	0.169*	0.068	0.484**	0.496**	0.080	0.456**	0.490**
AR(22)-DAXRL	0.114	1.000**	1.000**	0.064	1.000**	1.000**	0.064	1.000**	1.000**	0.074	1.000**	1.000**

Notes: The table presents model confidence sets based on MSPE for the selection of models reported in Table 4 for one, five, ten or twenty-days ahead forecasts. For each forecast horizon the MSPE is reported as well MCS p-values according to the test statistics T_R and T_{SQ}. Two stars indicate that a model belongs to the model set $\widehat{\mathcal{M}}^*_{0.25}$ whereas models with one star belong to $\widehat{\mathcal{M}}^*_{0.10}$. The forecast evaluation period covers January 2, 1998-December 29, 2006 ($P = 2232$).

Table 7: Model Confidence Sets, January 1998-December 2006 - standard deviation, one, five, ten and twenty-days ahead

	1-day ahead			5-days ahead			10-days ahead			20-days ahead		
	MSPE	T_R	T_{SQ}	MSPE	T_R	T_{SQ}	MSPE	T_R	T_{SQ}	MSPE	T_R	T_{SQ}
Riskmetrics	0.152	0.000	0.000	0.097	0.002	0.001	0.092	0.003	0.006	0.096	0.011	0.064
GJR-G-DA	0.133	0.055	0.032	0.084	0.013	0.002	0.080	0.236*	0.103*	0.083	0.339**	0.417**
ARFI-DAXRL	0.115	0.727**	0.816**	0.067	0.033	0.089	0.069	0.236*	0.221*	0.081	0.339**	0.417**
SPARFI-DAXRL	0.116	0.665**	0.664**	0.072	0.015	0.030	0.079	0.030	0.028	0.100	0.011	0.025
HAR-DAXRL	0.116	0.727**	0.816**	0.068	0.033	0.089	0.069	0.236*	0.221*	0.078	0.339**	0.417**
SPHAR-DAXRL	0.118	0.139*	0.262**	0.076	0.003	0.001	0.083	0.004	0.008	0.101	0.011	0.014
AR(1)-DAXRL	0.142	0.000	0.000	0.133	0.002	0.001	0.162	0.001	0.001	0.191	0.000	0.001
SPAR(1)-DAXRL	0.138	0.000	0.000	0.125	0.002	0.001	0.135	0.000	0.000	0.139	0.000	0.001
AR(5)-DAXRL	0.115	0.727**	0.816**	0.067	0.144*	0.169*	0.068	0.351**	0.342**	0.080	0.339**	0.417**
SPAR(5)-DAXRL	0.116	0.665**	0.628**	0.073	0.013	0.009	0.082	0.012	0.013	0.103	0.011	0.007
AR(22)-DAXRL	0.114	1.000**	1.000**	0.064	1.000**	1.000**	0.064	1.000**	1.000**	0.074	1.000**	1.000**
SPAR(22)-DAXRL	0.115	0.727**	0.816**	0.070	0.025	0.058	0.076	0.044	0.049	0.095	0.011	0.033

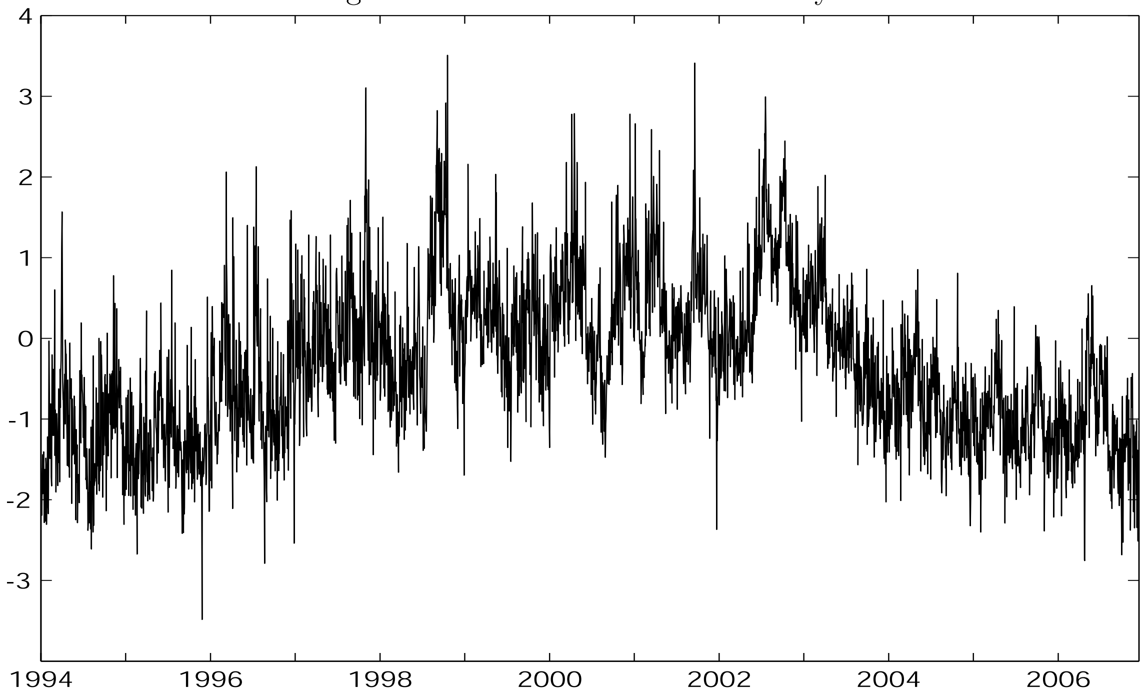
Notes: The table presents model confidence sets based on MSPE for different models, each with a '-DAXRL' specification and the Riskmetrics and GJR-GARCH(1,1)-DA model for one, five, ten or twenty-days ahead forecasts. For each forecast horizon the MSPE is reported as well MCS p-values according to the test statistics T_R and T_{SQ}. Two stars indicate that a model belongs to the model set $\widehat{\mathcal{M}}^*_{0.25}$ whereas models with one star belong to $\widehat{\mathcal{M}}^*_{0.10}$. The forecast evaluation period covers January 2, 1998-December 29, 2006 ($P = 2232$).

Table 8: Value-at-Risk evaluation for the out-of-sample period January 1998-December 2003

	MRC-μ	MRC-σ	$\hat{\alpha}$	LR_{uc}	LR_{ind}	LR_{cc}	C	$\bar{C}$	$\max(C_t)$
Riskmetrics	23.049	9.055	0.019	13.530 (0.000)	1.390 (0.238)	14.920 (0.001)	82.023	1.953	16.231
GJR-G-DA(1,1)	22.888	7.219	0.015	5.092 (0.024)	1.042 (0.307)	6.134 (0.047)	65.729	1.933	8.165
ARFI	23.049	7.471	0.017	8.882 (0.003)	1.935 (0.164)	10.817 (0.004)	68.001	1.790	12.140
ARFI-D	23.051	7.503	0.018	12.292 (0.000)	0.081 (0.776)	12.373 (0.002)	73.265	1.787	14.597
ARFI-DA	23.084	7.511	0.017	9.967 (0.002)	0.142 (0.707)	10.108 (0.006)	71.268	1.827	13.545
ARFI-DAR	23.072	7.494	0.018	12.292 (0.000)	0.081 (0.776)	12.373 (0.002)	77.103	1.881	16.725
ARFI-DARL	23.241	7.736	0.019	14.816 (0.000)	0.038 (0.845)	14.854 (0.001)	78.350	1.822	18.860
ARFI-DAXRL	24.168	8.483	0.018	11.104 (0.001)	0.109 (0.741)	11.213 (0.004)	69.758	1.744	15.636
SPARFI-DAXRL	24.553	9.184	0.018	11.104 (0.001)	0.109 (0.741)	11.213 (0.004)	68.614	1.715	15.291
HAR-DAXRL	23.029	7.832	0.022	23.495 (0.000)	0.004 (0.948)	23.499 (0.000)	82.812	1.690	16.314
ARX(1)-DAXL	22.368	2.788	0.022	25.092 (0.000)	0.012 (0.914)	25.104 (0.000)	91.300	1.826	18.365
ARX(5)-DAXL	23.151	7.027	0.020	17.530 (0.000)	0.011 (0.915)	17.542 (0.000)	78.866	1.753	15.827
ARX(22)-DAXL	23.559	8.221	0.021	20.426 (0.000)	0.000 (0.984)	20.426 (0.000)	77.410	1.647	13.781

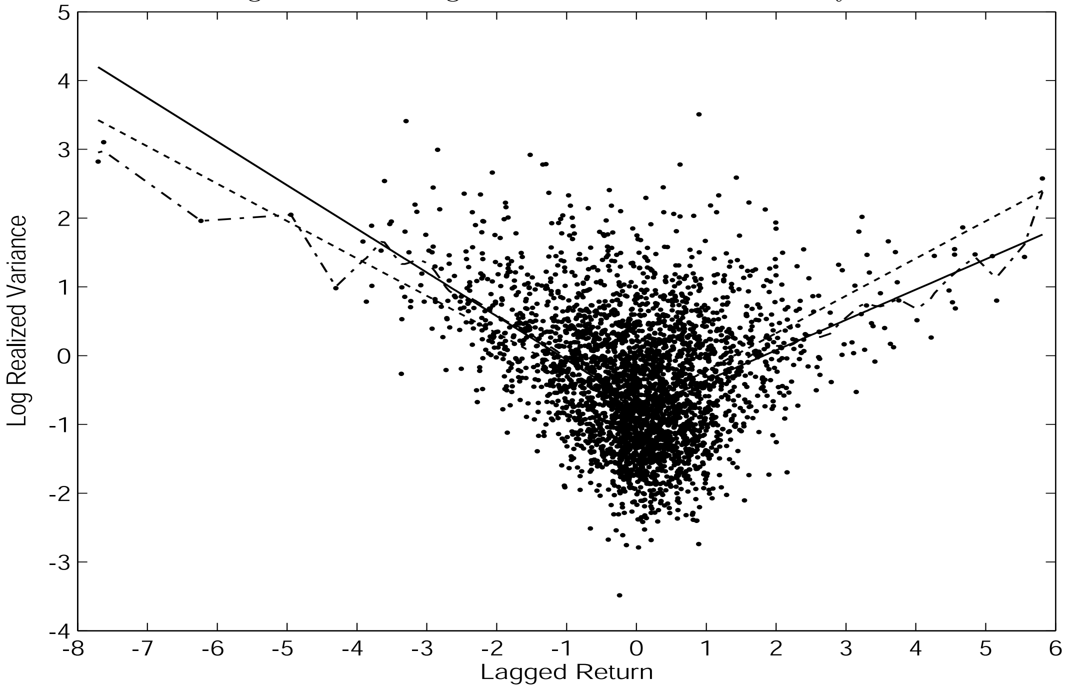
Notes: The table presents results for VaR estimates generated under the conditions of the Basle Committee MRA for the forecast evaluation period January 2, 1998-December 29, 2006 ($P = 2232$). The first two columns show the average and standard deviation of the required capital to cover market risk exposure (in percentage terms). Column 3 shows the average percentage number of exceptions defined as $\hat{\alpha} = x/P$ where x is the number of exceptions. Columns 4-6 show the interval forecast evaluation test statistics of correct unconditional coverage (uc), independence (ind) and correct conditional coverage (cc) (p-values are between brackets). Columns 7-9 give the total score C based on (19), the average score, $\bar{C} = C/x$ and the maximum individual score (all in percentage terms).

Figure 1: Realized S&P 500 volatility



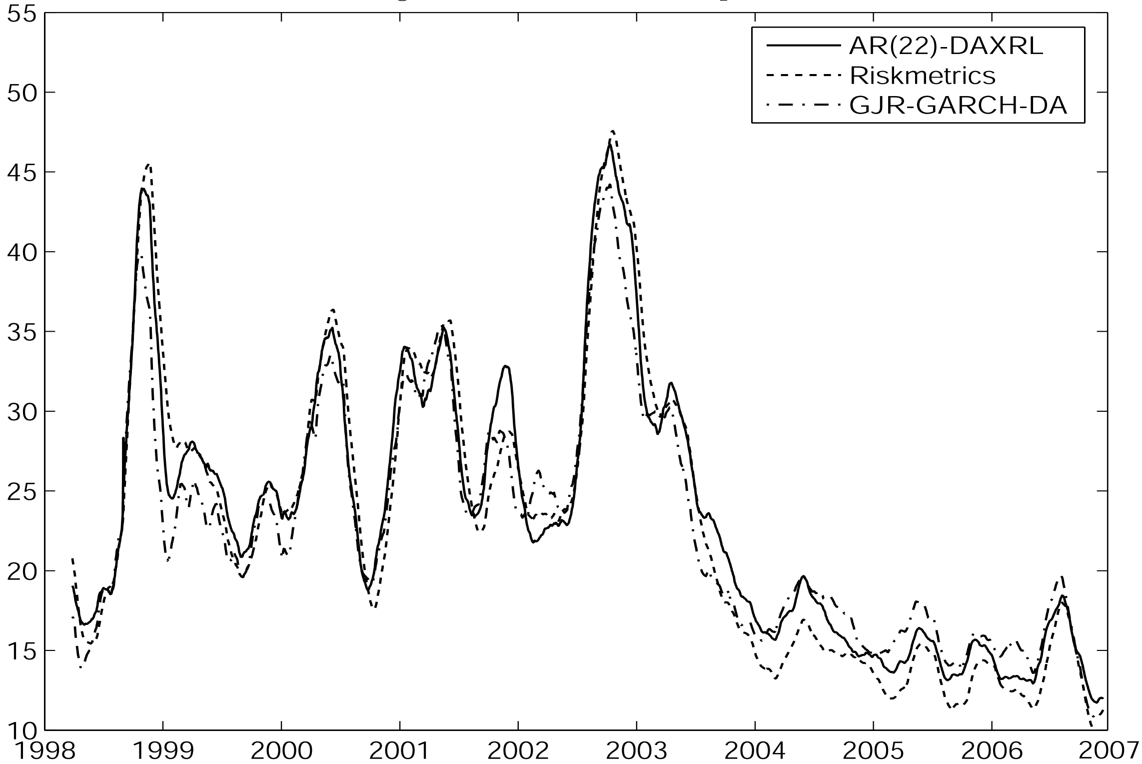
Notes: The graph shows the daily log realized variance for S&P 500 returns based on 5-minute returns for the period from January 3, 1994 until December 29, 2006 (3265 observations).

Figure 2: Leverage effects in realized volatility



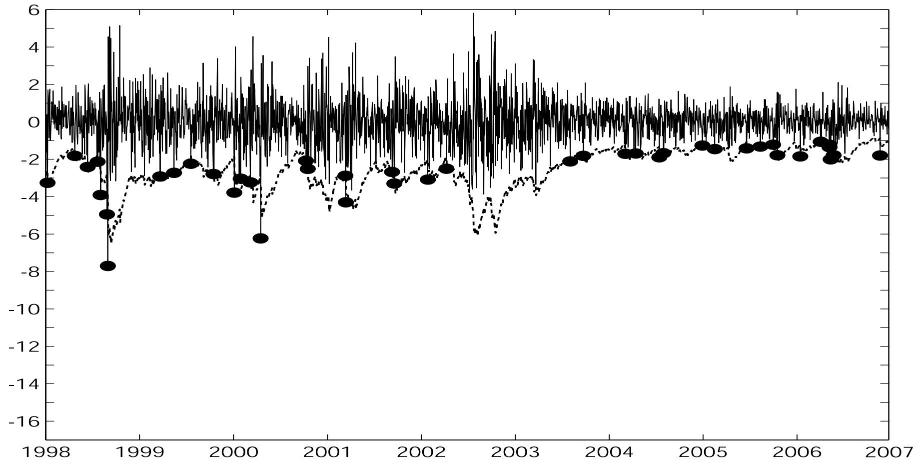
Notes: Scatter plot of daily log realized variance and lagged returns, based on observations for the period from January 3, 1994, until December 29, 2006 (3265 observations). The solid line is the fit of the news impact curve (1), where log realized volatility is regressed on a constant, the lagged absolute return, a dummy for negative returns and an interaction term of this dummy with the lagged absolute return. The dashed line is the fit of a symmetric news impact curve, i.e. (1) with $\beta_2 = \beta_3 = 0$. The dot-dashed line is the fit from a nonparametric regression of log realized volatility on the lagged return.

Figure 3: Market Risk Capital

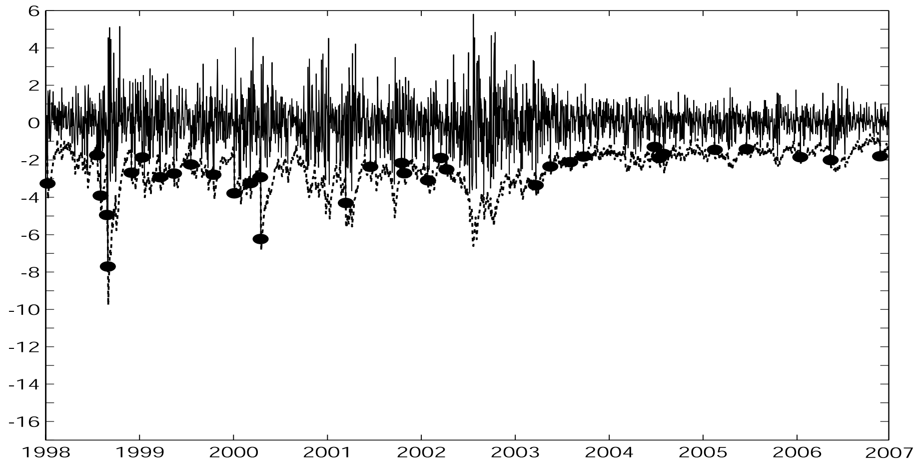


Notes: The graph shows the required capital (in percentage terms) to cover market risk exposure which is calculated as $MRC_{t+1} = \max(VaR_t^{10}, S_t \times \frac{1}{60}\sum_{i=0}^{59} VaR_{t-i}^{10})$ based on volatility forecasts from the Riskmetrics, GJR-GARCH(1,1)-DA and AR(22)-DAXRL models from January 2, 1998, until December 29, 2006 ($P = 2232$). VaR_t^{10} is the 99% VaR estimate for a 10-day holding period. The first sixty 1-day VaR estimates were used to construct the initial history needed to calculate MRC_{t+1}.

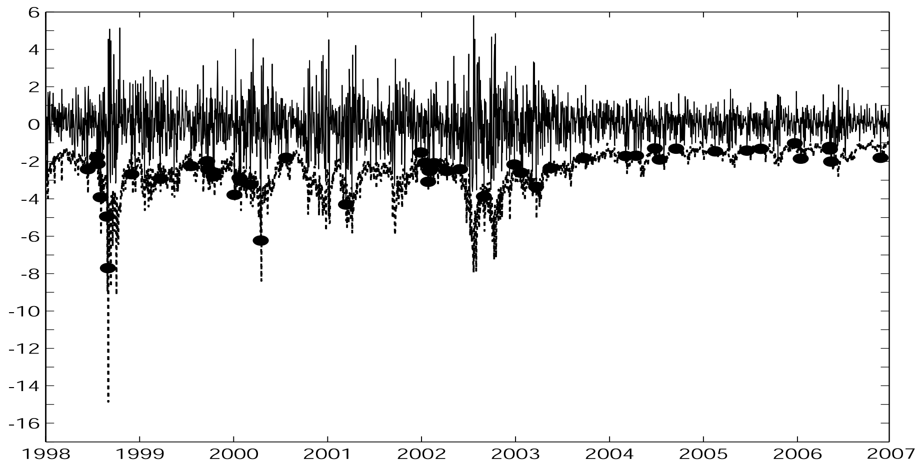
Figure 4: Value-at-Risk estimates



(a) Riskmetrics



(b) GJR-GARCH(1,1)-DA



(c) AR(22)-DAXRL

Notes: Realized returns (solid line) and 1-day 99% Value-at-Risk estimates based on volatility forecasts for the Riskmetrics, GJR-GARCH(1,1)-DA and AR(22)-DAXRL models (dotted lines) for the period from January 2, 1998, until December 29, 2006 ($P = 2232$). Black dots indicate VAR exceptions.