DE GRUYTER DOI 10.1515/snde-2012-0018 — Stud Nonlinear Dyn E 2013; 17(5): 521–549

Elena Goldman*, Jouahn Nam, Hiroki Tsurumi and Jun Wang

Regimes and long memory in realized volatility

Abstract: In this paper we model regimes and long memory in the dynamics of realized volatilities of intraday ETF and stock returns. We estimate threshold fractionally integrated (TARFIMA) models using Bayesian Markov Chain Monte Carlo (MCMC) algorithms with efficient jump. We also introduce a test based on posterior distributions of the mean squared forecast errors for model selection. Our findings are that the TARFIMA model that accounts for a different degree of long memory, persistence and variance in two regimes outperforms ARFIMA and other models using 5 day forecasts.

Keywords: Bayesian model selection; forecasting; realized volatility; threshold regimes.

***Corresponding author: Elena Goldman,** Department of Finance and Economics, Lubin School of Business Pace University, One Pace Plaza, NY 10038, USA, Tel.: +212-618-6516, e-mail: egoldman@pace.edu
Jouahn Nam: Department of Finance and Economics, Lubin School of Business Pace University, One Pace Plaza, NY, USA
Hiroki Tsurumi: Department of Economics, Rutgers University, New Brunswick, NJ, USA
Jun Wang: Department of Economics and Finance, Baruch College, One Bernard Baruch Way, New York, NY, USA

1 Introduction

Volatility is an important measure in many financial applications such as option pricing, risk management and portfolio optimization. In addition volatility is a tradable asset where buyers and sellers gain or lose depending on volatility realized until maturity. Since the Chicago Board Options Exchange (CBOE) introduced option trading on the volatility of the S&P500 index (VIX), many volatility-based derivatives have been introduced in the option markets. Trading in volatility derivatives attracted many investors since it provided more tools for managing portfolio risk, option positions and exploiting the leverage effect of negative correlation between volatility movements and stock index returns.[1] As the volatility derivative markets grew, research on volatilities has expanded, and realized volatility based on high-frequency returns has become a popular measure of volatility.

Since Andersen and Bollerslev (1997) introduced the concept of realized volatility as a square root of the quadratic variation of intra-daily returns there has been considerable interest in its distribution, modeling, and forecasting. Fleming, Kirby, and Ostdiek (2003) show the economic value of using realized volatility for portfolio optimization. A risk-averse investor can improve the performance of her mean-variance efficient portfolio when she switches from daily to intra-daily measure of variances and covariances for volatility timing. Andersen et al. (2001a,b) provide distribution of logarithmic transform of realized volatility of foreign exchange returns and stock returns, respectively. Andersen et al. (2003) (ABDL 2003) used a fractional autoregressive model in order to account for the long memory properties of realized volatility and found that these forecasts are better than forecasts produced by many popular models of volatility, e.g., daily GARCH, RiskMetrics, and FIEGARCH. However spikes in realized volatility and nonlinearities in the dynamics are not incorporated in these models.

Barndorff-Nielsen and Shephard (2004) introduced a bipower variation based on the product of absolute values of consecutive intra-daily returns. Using this measure they separated jumps from the quadratic variation. Maheu and McCurdy (2002) emphasized nonlinear features of realized volatility for the DM/$ exchange rate and found that 1 day forecasts from a duration dependent Markov-switching ARMA model are better than from linear models. However their model is restrictive to two regimes with an unobservable state variable determining the regime and no analysis of long memory properties. Additional papers that used

1 See, e.g., Hafner and Wallmeier (2008) who analyzed variance swap positions for optimal portfolio investments.

Markov-switching models for international equity returns and volatilities include Ang and Bekaert (2002, 2004), Ramchand and Susmel (1998), Guidolin and Timmermann (2002), and Liu and Maheu (2008) among others.

As an approximation to a long memory ARFIMA process several papers used a heterogeneous autoregressive model (HAR) [see Andersen, Bollerslev, and Diebold (2007), Liu and Maheu (2008), McAleer and Medeiros (2008) and Corsi (2009) among others]. We do not consider this approximation for long memory, since we show an efficient way of joint estimation of the fractionally differenced parameters d and threshold parameters r in ARFIMA and TARFIMA models considered below. However, for the future research it would be interesting to compare HAR and ARFIMA models.

There are several papers that provide alternative measures of realized volatility adjusting for microstructure noise [see Bandi, Russell, and Yang (2008), Ait-Sahalia and Mancini (2008), etc]. In the presence of microstructure noise if sampled at high frequency intervals the realized volatility can become a biased estimate of daily volatility since noise can be time dependent and correlated with closing prices. However, as suggested in Andersen et al. (2001a,b) if the data is sampled at a lower frequency (they suggested 5 min) the realized volatility is a good estimate in practice. In this paper, we do not study microstructure noise and possible adjustments, rather, we proceed with 5-min intra-daily returns (excluding holidays and non-trading hours) obtained from the NYSE Trade and Quote (TAQ) database.

In this paper, we use a multiple regime threshold fractionally integrated autoregressive and moving average (TARFIMA) model to analyze the dynamics of the quadratic variation in intra-daily returns. The advantage of the TARFIMA model is that it is more flexible than a linear model since it is characterized by regimes with different persistence, long memory and variance, thus accounting for asymmetric dynamics in regimes. For the periods of extreme volatility, such as during a financial crisis, the TARFIMA model with regimes based on past volatility seems to be more appropriate than linear models as TARFIMA explicitly accounts for the dynamic properties of a high volatility regime. The use of the Markov-switching model with long memory seems to be not as appropriate since its regimes are randomly determined by an unobservable variable, while for the TARFIMA model regimes are determined by the previous day level of volatility or some other observable variables.[2] Threshold models allow the explicit estimation of threshold levels of volatility that change with regimes.

The TARFIMA model can capture the difference in dynamics of high and low volatility regimes. One of the prominent features of the data are frequent spikes in volatility. Although the measure based on the product of absolute consecutive returns removes some jumps there are still many spikes in the data that can be observed in a regime of high volatility. The TARFIMA model introduced here is quite general because the parameters of ARMA, long memory and variance change with regime and because it allows us to find multiple threshold regimes.[3] In particular, we investigate nonlinearities in the dynamics of quadratic variations of selected ETF and stock returns. We analyse four major ETFs (DIA, QQQQ, SPY and VXX) and three stocks from the Dow Jones Index (GE, MSFT and HPQ).

We find that the logarithm of volatility process can be characterized by two regimes of high and low volatility. The persistence, long memory and variance may change with each regime. This model helps us understand better the dynamics of intra-daily returns and improves forecasts compared to fractional ARMA (ARFIMA) models since forecasts based on estimated thresholds regimes of high or low volatility can be predicted separately. This information indicates the difference in investors behavior in high and low volatility regimes. We find that the high volatility regime is characterized by a higher variation in realized volatility which arises because of increased activity in the market. On the other hand, the lower volatility regime has generally higher long memory parameter. We find that the regimes of high volatility are short-lived and

2 One could use past level of the return rather than volatility as a threshold variable. This choice of the threshold variable may reflect the leverage effect, which is the asymmetric effect of negative and positive returns on volatility.

3 Many recent papers use threshold autoregressive (TAR) models in economic and financial time series to model the dynamics of short-term interest rates, real exchange rates, unemployment rate, stock prices, production, and inventories. See, e.g., Tsay (1989, 1998), Hansen (1997), Koop and Potter (1999), Phann, Schotman, and Tschering (1996), Forbes, Kalb, and Kofman (1999), Goldman and Agbeyegbe (2007), Dufrenot, Guegan, and Peguin-Feissolle (2008) among others.

mean-reverting. Also abrupt shifts in volatility may induce long memory characteristics, thus the results of ARFIMA and TARFIMA model may be different in terms of long memory parameters.

We also find that spikes above estimated threshold in realized volatility captured by upper regimes make about 15% of observations and are important in modeling and forecasting.[4]

The TARFIMA model is estimated using Bayesian Markov Chain Monte Carlo (MCMC) algorithms with efficient jump as an extension of the algorithm in Goldman and Agbeyegbe (2007). The algorithm is extended for the fractional integration parameter d. The Metropolis-Hastings algorithm with efficient jump allows us to estimate jointly parameters of all regimes and threshold parameters. For model selection and specification several in-sample techniques are used such as information criteria, significance of parameters and estimation with and without truncation restrictions imposed on thresholds.

The out-of-sample forecast performance is used in this paper for comparison of TARFIMA and ARFIMA models. We suggest comparison based on the distributions of mean squared errors (MSE). We illustrate a test of the dominance of the cumulative density for one model over other models. Using MCMC such a test is easy to implement and it incorporates the whole distribution of MSE rather than the point estimate frequently used in the literature for forecast evaluation.

Our paper contributes to the literature of modeling realized volatility as the first to apply a TARFIMA model with estimated thresholds, changing long memory parameters and variance. We also provide a test for MSE to select the best model. We show that TARFIMA model provides a better fit to the realized volatility processes than the benchmark linear fractionally integrated model for 5 day forecasts (as well as longer term forecasts) in most cases. So far in the realized volatility literature fractional autoregressive models are known to produce forecasts that are better than the forecasts of other models (ABDL 2003). Our results show that TARFIMA models tend to outperform ARFIMA models for realized volatility of individual equity returns in the DJ index. Thus TARFIMA models with several regimes may provide more benefit to investors in the areas of portfolio optimization and risk management.

The plan of the paper is as follows. In Section 2 we provide the TARFIMA and ARFIMA models, a Bayesian test of long memory and stationarity and model selection using posterior mean squared errors (MSE). We also show numerical examples illustrating biases in estimated misspecified nested TARFIMA models (TARFI, TARMA and ARFIMA). We show that both long memory and regime changes are important features of the TARFIMA model and ignoring one of these features produces biases in other parameters. In Section 3 we explain estimation results and stationarity tests for TARFIMA, TARFI, TARMA and ARFIMA models of logarithms of realized volatility. Each of the models considered is nested in a more general TARFIMA model and the results are sensitive to omitting some features of the data. In Section 4 we perform forecasts and model comparisons for TARFIMA and ARFIMA models. In Section 5 we give concluding remarks and plans for future research. Appendix with Bayesian posterior distributions and MCMC algorithms is in Section 6.

2 Model

The quadratic variation or realized variance RV was introduced by Andersen and Bollerslev (1997) as the sum of squared intra-day returns:

$$RV_t = \sum_{\tau} r^2_{t+\tau\Delta}$$

4 In additional results not reported in this paper we found that realized variance generally produces a better forecast than bipower variation for the stocks in the Dow Jones Index, since the former contains more information about the spikes incorporated in the dynamics of a high volatility regime.

where $r_{t+\tau\Delta}$ are returns for small intra-daily periods of length Δ within a day t. Five-minute returns are accepted by many studies as a norm.[5] They define the realized volatility as the square root of the quadratic variation $RV^{1/2}$.

Barndorff-Nielsen and Shephard (2004) suggested to use the normalized sum of products of absolute values of consecutive returns (which they called the bipower variation BV) in order to separate jumps from the total daily variation. In the literature that measures separately the continuous part of total realized variance measured by BV_t and jumps (RV_t–BV_t) the statistics for determining the significance of jumps is based on the size of the relative jump as defined by Huang and Tauchen (2005). This implies that on the days when volatility is high the jump may not be detected if the size of the relative jump is not large, while on low volatility days a smaller increase in volatility may qualify as a jump.[6]

In our study instead of separating BV and jumps we look at several regimes of the RV and compare the dynamics of high and low volatility regimes. Since we also know from previous studies (e.g., ABDL 2003) that realized volatility exhibits long memory we propose using a threshold fractionally integrated model with regimes that are characterized by different time series dynamics, variance and long memory parameters in each regime.

The regimes are determined by an observed threshold variable. In our case it is the past level of realized volatility. The lagged realized volatility above the threshold level will determine the high volatility regime.[7] In this model the highest volatility regime will be short-lived since overall variance turns our to be mean-reverting as discussed in Section 3. The current level of volatility determines the forecast of the next period regime and contains valuable information for portfolio diversification.[8]

In addition, the distribution of the logarithm of realized volatility may be better captured by a mixture of several normal distributions from several regimes rather than one normal distribution. The mixture of normal distributions can approximate well any distribution (Ferguson 1973). Thus, the threshold model with several regimes is expected to better fit the unknown distribution of the data.[9]

In the next two subsections we describe the ARFIMA and TARFIMA models, tests of stationarity, model choice and numerical examples.

2.1 Fractionally integrated ARMA (ARFIMA) model

Let $y_t = ln(RV_t^{1/2})$ or $ln(BV_t^{1/2})$. The fractionally integrated ARMA model (ARFIMA) is given by

$$\begin{gathered}(1-B)^d y_t = x_t\gamma + u_t \\ u_t = \frac{\Theta(B)}{\Phi(B)}\varepsilon_t \\ \Theta(B) = 1+\theta_1 B+\cdots+\theta_q B^q \\ \Phi(B) = 1-\phi_1 B-\cdots-\phi_p B^p \\ \varepsilon_t \sim \mathrm{N}(0,\sigma^2),\end{gathered} \tag{1}$$

where d is a long memory fractional integration parameter, γ is a vector of regression parameters for x explanatory variables[10], σ^2 is variance (showing variation in realized volatility), $\Phi(B)$ and $\Theta(B)$ are p-th order and q-th order polynomials in the backward shift (lag) operator B, respectively. For $d>-1$ the difference operator $(I-B)^d$ has the binomial expansion

5 Higher frequencies may result in microstructure problems (see Ait-Sahalia, Mykland, and Zhang 2005).
6 See Bollerslev, Law, and Tauchen (2008) for the data example illustrating this point.
7 Alternatively past level of returns or any other observable variable could be determining the regimes.
8 See, e.g., Ang and Bekaert (2002, 2004) showing the benefits of switching regime models for portfolio diversification.
9 Although the logarithm of realized variance looks normal, statistical tests (e.g., Jarque Bera test) often reject normality.
10 In the simplest case x is a constant term.

$$(I-B)^d=\sum_{j=0}^{\infty}\pi_j B^j, \tag{2}$$

where

$$\pi_j=\prod_{0<k\le j}\frac{k-1-d}{k}, \quad j=0,1,2,\ldots.$$

When $p=q=0$, Brockwell and Davis (1987) state that $\{y_t\}$ is a covariance stationary process if $-0.5<d<0.5$. If $0<d<1$ the ARFIMA model is called a long memory process.

2.2 Fractionally integrated threshold ARMA model (TARFIMA)

Let us consider the following dynamic threshold ARFIMA model (TARFIMA) with multiple regimes for logarithm of realized variance or bipower variation $y_t=ln(RV_t^{1/2})$ or $ln(BV_t^{1/2})$

$$(1-B)^{d^{(j)}}y_t=x_t\gamma^{(j)}+u_t, \tag{3}$$

$$u_t=\frac{\Theta^{(j)}(B)}{\Phi^{(j)}(B)}\varepsilon_t \qquad \varepsilon_t\sim \mathrm{N}(0,(\sigma^{(j)})^2), \tag{4}$$

where $j=1, \ldots, s+1$ with $s+1$ regimes defined by

$$\begin{cases} j=1 & y_{t-\delta}<r_1 \\ j=2 & r_1\le y_{t-\delta}<r_2 \\ \vdots & \vdots \\ j=s+1 & y_{t-\delta}\ge r_s \end{cases}. \tag{5}$$

We assume in the model above that there are s thresholds $r_1, r_2, \ldots, r_s$ separating $s+1$ regimes for y_t. The threshold variable y_{t-1} is an observable previous day volatility level with delay parameter, δ, set to one ($\delta=1$).[11] The error term follows ARMA process u_t~ARMA($p^{(j)}$, $q^{(j)}$), where $\Phi^{(j)}(B)$ and $\Theta^{(j)}(B)$ are $p^{(j)}$-th order and $q^{(j)}$-th order polynomials in the backward shift operator B, respectively.

We note that each parameter in this model $\{\gamma^{(j)}, \phi^{(j)}, \theta^{(j)}, \sigma^{2(j)}\}$ takes $s+1$ values depending on the regime j where $y_{t-\delta}$ belongs. This allows a change in dynamics, persistence, and variance of realized volatility depending on regime.

2.3 Testing nonstationarity and mean reversion

In the ARFIMA and TARFIMA models there are two sources of nonstationarity: fractional integration parameter d and the roots of $\Phi(B)$. Let ρ denote the maximum absolute value of the inverse roots of the AR parameters in the error term u_t. We will test for the mean reversion of realized volatility y_t looking at the long memory parameter $d^{(j)}$ and at the autoregressive root parameter $\rho^{(j)}$ in each regime. A simple Bayesian unit root test using the maximum absolute value of inverted roots of AR parameters ρ was introduced in Goldman et al. (2001) and we use the same method in this paper. It is important to test the null hypothesis of nonstationarity,

11 One can use any lag of volatility or a combination of several lags, but our analysis shows that the previous day provided the best fit for the model using MBIC criterion.

since nonstationarity implies high persistence and long memory of time series. The nonstationarity hypothesis in our setting is jointly given by:

$$H_0: \rho \geq 1 \quad \text{versus} \quad H_1: 0 \leq \rho < 1$$

and

$$H_0: |d| \geq 0.5 \quad \text{versus} \quad H_1: -0.5 < d < 0.5.$$

Although covariance stationarity requires $d<0.5$ there is still mean reversion if values of $0<d<1$ and the AR process is stationary. Diebold and Inoue (2001) define a long memory mean-reverting fractionally integrated process for $0<d<1$. Mean reversion for $0.5<d<1$ was illustrated for real exchange rates in Diebold, Husted, and Rudebusch (1991) and formally proved for ARFIMA(p, d, q) time series in Chung (2001). Both papers showed that impulse response functions converge to zero.

In the Bayesian approach it is common to use the highest posterior density intervals (HPDI) for hypothesis testing. One could construct a joint HPDI involving both parameters (e.g., for $\rho+d^2$), however, intuitively it is easier to interpret these two parameters separately. If the 95% HPDI includes 1 for ρ or includes 1 for d we would not reject a unit root. If the HPDI for ρ does not include 1 and HPDI for d is within [−0.5, 0.5] we would conclude that the process is stationary. If the HPDI for ρ does not include 1 and HPDI for d does not include 1 we would conclude that the process is mean-reverting even if $d>0.5$. Finally if the HPDI for d is positive and does not include 0 we conclude that the time series exhibits long memory.

2.4 Numerical examples

Tables 1 and 2 present numerical examples illustrating sensitivity of estimated parameters to different model specifications. We also examine the role of d and ρ in the stationarity tests.

In example 1 (Table 1) we generated a TARFIMA model with parameters: γ={0, 0.33}, ϕ={0.3, 0}, θ={−0.1, 0}, σ={0.05, 0.08}, d={0.3, 0.2}, r=0.4. In this example both ρ[12] and d have small values and are in the

Table 1 Numerical example 1.

Parameters					Mean (st. dev.)
		TARFIMA	TARMA	TARFI	ARFIMA
$\gamma^{(1)}$	0	−0.009 (0.010)	0.145 (0.094)	−0.011 (0.009)	0.003 (0.007)
$\gamma^{(2)}$	0.33	0.302 (0.073)	0.329 (0.105)	0.305 (0.072)	
$\phi^{(1)}$	0.3	0.199 (0.274)	0.926 (0.028)		−0.967 (0.026)
$\phi^{(2)}$	0	0.172 (0.367)	0.786 (0.116)		
$\theta^{(1)}$	−0.1	−0.094 (0.263)	−0.477 (0.049)		0.981 (0.024)
$\theta^{(2)}$	0	−0.139 (0.389)	−0.419 (0.090)		
$\sigma^{(1)}$	0.05	0.051 (0.003)	0.052 (0.003)	0.052 (0.003)	0.057 (0.003)
$\sigma^{(2)}$	0.08	0.076 (0.009)	0.079 (0.009)	0.076 (0.009)	
$d^{(1)}$	0.3	0.373 (0.043)		0.446 (0.038)	0.550 (0.025)
$d^{(2)}$	0.2	0.175 (0.065)		0.180 (0.071)	
r	0.4	0.399 (0.002)	0.398 (0.004)	0.399 (0.004)	
n2/n	0.193	0.174	0.174	0.174	

This table presents posterior means and standard deviations of estimated parameters in a simulated TARFIMA model. The true parameters are γ={0, 0.33}, ϕ={0.3, 0}, θ={−0.1, 0}, σ={0.05, 0.08}, d={0.3, 0.2}, r=0.4. The data were generated with n=1001 obs, n2/n=0.193 is % of observations in the upper regime. Forty lags were used for modeling long memory.

12 Note that in AR(1) process $\rho=|\phi|$.

Table 2 Numerical example 2.

Parameters		Mean (st. dev.)			
		TARFIMA	TARMA	TARFI	ARFIMA
$\gamma^{(1)}$	0	−0.078 (0.208)	−7.880 (0.416)	0.009 (0.013)	−0.338 (0.333)
$\gamma^{(2)}$	1	0.923 (0.221)	−7.386 (0.484)	−0.009 (0.026)	
$\phi^{(1)}$	0.95	0.972 (0.008)	0.999 (0.003)		0.924 (0.058)
$\phi^{(2)}$	0.9	0.901 (0.027)	0.997 (0.002)		
$\theta^{(1)}$	−0.1	−0.079 (0.041)	0.453 (0.0202)		−0.085 (0.124)
$\theta^{(2)}$	0.1	0.156 (0.050)	0.425 (0.029)		
$\sigma^{(1)}$	0.1	0.103 (0.004)	0.146 (0.006)	0.191 (0.008)	0.216 (0.010)
$\sigma^{(2)}$	0.4	0.402 (0.021)	0.449 (0.023)	0.560 (0.028)	
$d^{(1)}$	0.6	0.538 (0.0334)		0.9995 (0.0004)	0.602 (0.170)
$d^{(2)}$	0.5	0.395 (0.056)		0.9987 (0.0015)	
r	0	0.032 (0.030)	−0.737 (0.130)	−3.380 (0.190)	
n2/n	0.369	0.377	0.385	0.432	

This table presents posterior means and standard deviations of estimated parameters in a simulated TARFIMA model. The true parameters are γ={0, 1}, ϕ={0.95, 0.9}, θ={−0.1, 0.1}, σ={0.1, 0.4}, d={0.6, 0.5}, r=0. The data were generated with n=1001 obs, n2/n=0.193 is % of observations in the upper regime. Forty lags were used for modeling long memory.

stationary region. From the results we can see that the TARFIMA model estimates all the parameters well. For the generated TARFIMA model data we also estimated TARMA (no long memory: set d=0), TARFI (no ARMA: set $\phi=\theta=0$) and ARFIMA (no regimes). All the models pass the test of stationarity but show biases in estimated parameters due to omission of some important variables. For example, both TARMA and TARFI models identify the threshold r well but in case of TARMA the autoregressive parameters ϕ and the absolute inverse roots of the AR parameters ρ are overestimated while for TARFI the long memory parameters d are overestimated. A simpler ARFIMA model significantly overestimates ρ, θ and d. The results of ARFIMA model spuriously show d>0.5 and high persistence in AR due to simplified model neglecting change in regimes.

In example 2 (Table 2) the parameters of a generated TARFIMA model are γ={0, 1}, ϕ={0.95, 0.9}, θ={−0.1, 0.1}, σ={0.1, 0.04}, d={0.6, 0.5}, r=0. Here the parameters indicate high persistence with ρ<1 and d>0.5 in regime 1. This process is still mean reverting. In this example estimated TARMA and TARFI model again show overestimated parameters ρ (for TARMA) and d (for TARFI). Moreover, TARMA model indicates nonstationarity using 95% HPDI as it includes 1. The 95% HPDI for d in the TARFI model is very close to 1 as well.[13] As for ARFIMA model it estimates ρ and d well on average. However, as the standard deviations of these parameters are much larger for ARFIMA model the 95% confidence intervals are wider. In particular 95% HPDI for the d parameter includes 1. In this case we would mistakenly conclude that model is non-stationary and not mean reverting due to wrong model specification.

As we can see from both examples if the true generating process is TARFIMA then both ρ and d are important in the stationarity tests. If we estimate a model without d the parameter ρ is biased upward. Alternatively, if we estimate the model without ρ then d is biased upward. Finally a model ignoring regimes generally results in biased parameters as well.

On the other hand, when the true generating process is either ARFIMA or TARMA the TARFIMA model nesting both of these models works well.[14]

As for data applications for ETFs in Section 3 we generally find that both ρ and d are significant in the model and we focus on TARFIMA and ARFIMA models.

13 For example, the 95% HPDI for $d^{(1)}$ is (0.998577, 0.999990).

14 Results of these simulated examples are available upon request.

2.5 Model choice

After models are estimated using MCMC algorithms with efficient jump, explained in Appendix, we compare models based on the distribution of their out-of-sample mean squared errors (MSE). MSE is the most common overall measure of forecast accuracy measuring average squared deviations of forecasted values from observed values.[15] Using MCMC we can get the posterior distributions of the MSE for each model as explained below. The model with MSE distribution that dominates all other distributions using cumulative density function (CDF) is chosen over other models.[16]

The distributions of mean squared errors are obtained as follows. We estimate the model for the sample of data 1. . .T and make h-step ahead repeated forecasts based on the fixed sample estimates of parameters. At each i-th draw of all parameters $\theta^{(i)}$ we find the predicted values of

$$\begin{aligned}
&\hat{y}^{(i)}_{T+1} | (y_1,\dots,y_T,\theta^{(i)})\\
&\hat{y}^{(i)}_{T+2} | (y_1,\dots,y_T,\hat{y}^{(i)}_{T+1},\theta^{(i)})\\
&\vdots\\
&\hat{y}^{(i)}_{T+h} | (y_1,\dots,y_T,\hat{y}^{(i)}_{T+1},\dots,\hat{y}^{(i)}_{T+h-1},\theta^{(i)}).
\end{aligned}$$

Using MCMC algorithms we obtain the draws of MSE for the predicted values of $\hat{y}_{T+j}$, $j=1,\dots,h$:

$$MSE^{(i)}=\frac{1}{h}\sum_{j=1}^{h}(\hat{y}^{(i)}_{T+j}-y_{T+j})^2,$$

where $\hat{y}^{(i)}_{T+j}$ is the i-th MCMC draw of the predicted value at time $T+j$ and y_{T+j} is the realized value at $T+j$.

The mean forecast errors (ME) are average differences between predicted values from MCMC draws and actual values of y_{T+j}, $j=1,\dots,h$:

$$ME^{(i)}=\frac{1}{h}\sum_{j=1}^{h}(\hat{y}^{(i)}_{T+j}-y_{T+j}).$$

The mean error measures average bias (positive or negative) and is another measure of accuracy. Using MCMC we obtained the distribution of MSE and ME for each out-of-sample observation.

The CDF of the distribution of MSE closest to zero indicates which model has the best forecast accuracy. The model selection is based on *min* MSE for the posterior modes or medians of their distributions. Since the distribution of MSE taking only positive values is heavily skewed to the right it makes sense to use the mode or median rather than the mean of distribution.[17]

If we have m posterior predictive densities for repeated 1 day or 5 day forecasts we can find the average of the modes (or medians) of all distributions. For example:

$$\mathrm{MMSE}_i=\frac{1}{m}\sum_{j=1}^{m}\mathrm{mode}_j(\mathrm{MSE}|M_i),$$

where MMSE_i is the average of modes of forecasts for model M_i. We choose the minimum of MMSE:

$$\min_{i\in\{1,..,3\}} \mathrm{MMSE}_i$$

among three competing models.

15 The mean squared errors are commonly used in the classical statistics. The Bayesian analogue of this popular measure is used in this paper.

16 We note here the advantage of using a Bayesian approach where posterior distributions of any parameters of interest are easily obtained using MCMC draws as explained below. It is common in current forecasting literature to provide the distribution of forecasts rather that a point forecast.

17 The mean is highly affected by skewness compared to median or mode of asymmetric distributions.

3 Data analysis

The models are estimated for log realized volatilities of four ETFs: DIA, QQQQ, SPY and VXX. The data obtained from the TAQ database are 5-min returns from January 20, 1998 to October 29, 2010. The data for QQQQ starts on March 10, 1999 while the data for VXX starts on January 30, 2009. As a result the corresponding daily realized volatilities have 3217 observations of trading days for DIA and SPY, 2937 observations for QQQQ and 442 observations for VXX. For DIA, QQQQ and SPY the last 188 observations were left for out-of-sample forecast evaluation. For VXX due to smaller sample size the last 50 observations were left for forecasts. For comparison we also estimated models and provided forecasts for three selected stocks in the Dow Jones Index. Data for GE and MSFT are for the period January 4, 1993 to December 31, 2004 (3024 observations). Data for HPQ are for the period May 6, 2002 to December 31, 2004 (671 observations). We used the last 360 observations for forecast evaluation.[18]

Figure 1 shows the square roots of quadratic variations $RV^{1/2}$ for ETFs. We can observe similar dynamics and multiple spikes in realized volatilities of DIA, SPY and QQQQ. We observe the most extreme spikes in all indices for the financial crisis in 2008–2009, for the Asian crisis and the burst of the dot-com bubble between 1998–2002 and for the 1 day "flash crash" in May 2010. For the VXX a limited amount of data is available and we only see the May 6, 2010 "flash crash" spike. A long period of low volatility is observed between 2003 and 2007. Since most of our ETF data includes two crises, a one day largest crash on May 6, 2010 and a relatively long period of low volatility it is not surprising that we use a model with different regimes. The results from our paper could be applicable to other periods and assets where volatility may change significantly and abruptly.

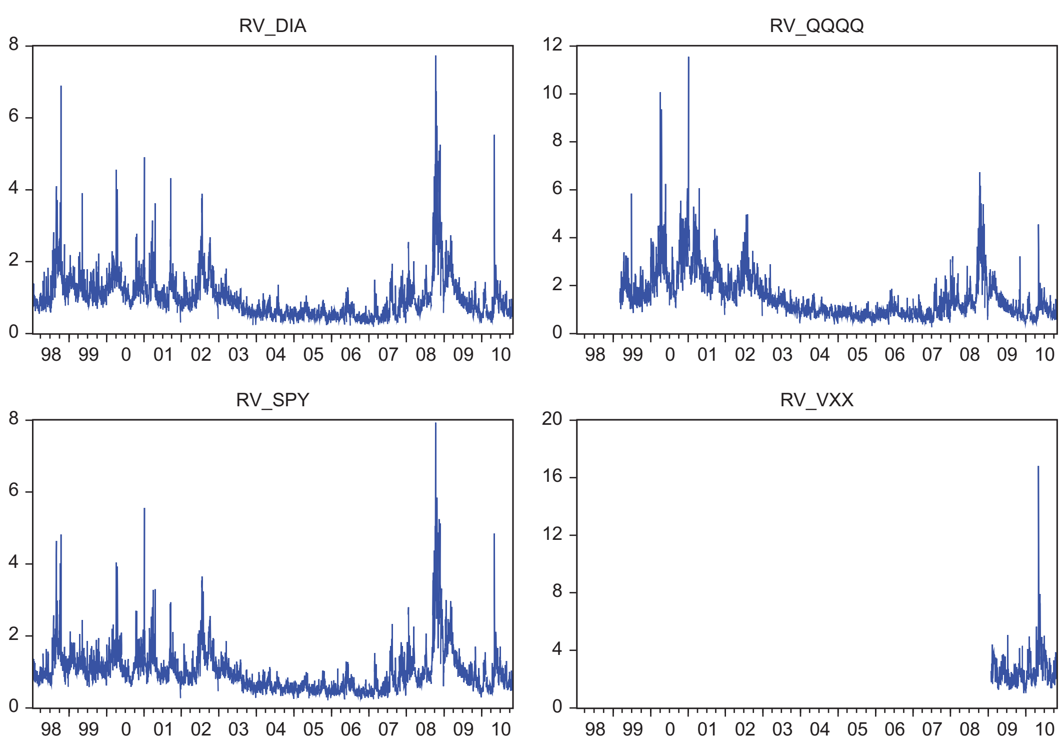

Figure 1 Realized volatility of DIA, QQQQ, SPY and VXX.

18 We present only the forecasting results for individual stocks in the paper since the estimated parameters and stationarity tests are similar to ETFs. The stocks intra-daily data were only available to authors till December 2004.

Figure 2 shows the logarithms of realized volatilities $ln(RV^{1/2})$. Although graphs are smoother one can still see the spikes and periods of high and low volatility. The advantage of using the TARFIMA model with two regimes is that it explicitly allows us to account for the dynamics in regimes of high and low volatility. The change in regimes is modeled once volatility is above a certain threshold level estimated in the model.

Table 3 gives summary statistics of the $ln(RV^{1/2})$ showing number of observations (number of trading days), mean, standard deviation, first order autocorrelation, skewness, kurtosis, and estimated fractional-differencing parameter d^e with asymptotic standard error using Geweke and Porter-Hudak (GPH) (1983) log – periodogram. Following ABDL 2003 using d^e we can examine long memory properties of realized volatility. We see that the GPH fractional parameter for ETFs ranges from 0.50 to 0.58 with asymptotic standard errors depending on number of observations and ranging from 0.025 to 0.056. The periodogram estimates of d^e all indicate long memory processes with statistically significant $d>0$ and close to $d=0.5$. However, we also notice that most series are highly autocorrelated with the first degree autocorrelation AR(1) around 0.7–0.9. The GPH estimator is simple to apply, but as pointed out in Baillie (1996) among others it is substantially biased in the presence of significant autocorrelation. Therefore, we prefer to use joint estimation of the fractional parameter with ARMA parameters using MCMC as described in the previous section. Below we compare estimated results and forecasting performance of ARFIMA and TARFIMA models.

The results of the estimated TARFIMA model (3)–(5) are given in Table 4. The second and third columns give thresholds for $ln(\sqrt{RV})$ with posterior standard deviations in brackets and corresponding annualized thresholds for realized volatility. Regime 2 is characterized by volatility levels above 23% for DIA and SPY, 44% for QQQQ and 55% for VXX. The fourth column gives the percentage of observations in the regime of low volatility below the threshold. The fifth and sixth columns give the estimated maximum absolute value of autoregressive roots in each regime, where regime 1 is below the threshold and regime 2 is above the threshold. The last two columns give estimates of fractional integration parameters d for each regime. We provide mean and standard deviation for each parameter in each regime.[19] Using MBIC information criterion we find the number of regimes and orders of ARMA model in each regime. For all the ETFs we found that models with

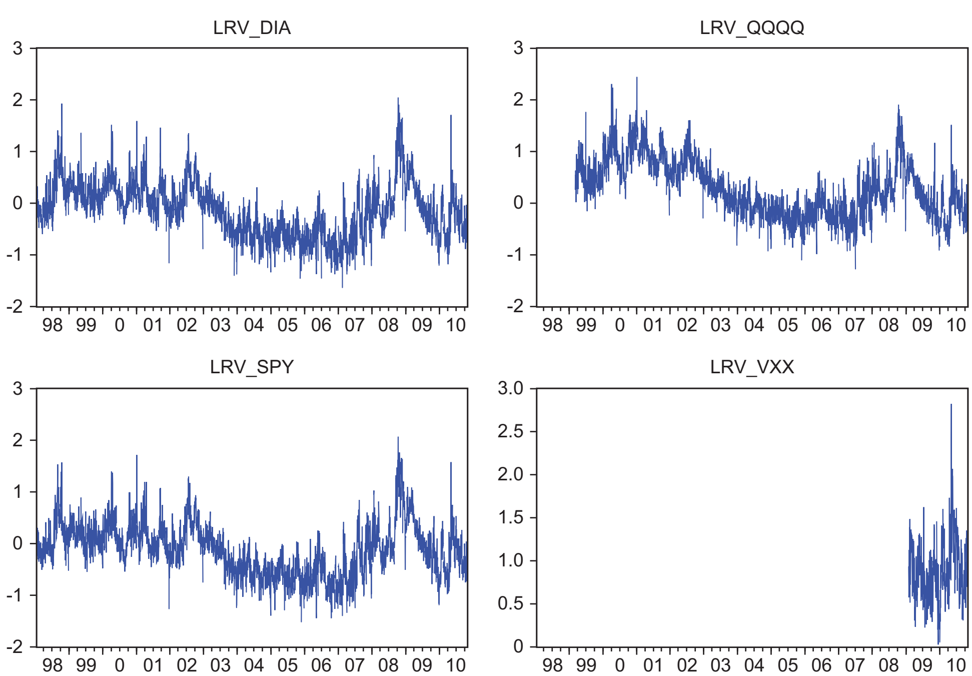

Figure 2 Log realized volatility of DIA, QQQQ, SPY and VXX.

19 The 95% HPDIs are omitted here to save space. They are available from the authors on request.

Table 3 Summary statistics of $log(\sqrt{RV})$ for ETFs.

ETF	N obs	Mean	St. dev.	AR(1)	Skewness	Kurtosis	d^e (as. std.)
DIA	3217	−0.138	0.520	0.862	0.284	3.145	0.547 (0.025)
QQQQ	2931	0.277	0.546	0.881	0.370	2.648	0.546 (0.026)
SPY	3217	−0.123	0.504	0.859	0.342	3.288	0.583 (0.025)
VXX	442	0.883	0.352	0.658	0.957	5.684	0.500 (0.056)

This table presents mean, standard deviation, first order autocorrelation, skewness, kurtosis and fractional integration parameter d^e for log realized volatilities. The data are for DIA (1/20/1998–10/29/2010) 3217 obs; QQQQ (3/10/1999–10/29/2010) 2931 obs; SPY (1/20/1998–10/29/2010) 3217 obs; VXX (1/30/2009–10/29/2010) 442 obs. GPH log-periodogram was used to estimate the fractional integration parameter d^e based on $m=N_{obs}^{4/5}$ lowest-frequency periodogram ordinates. The asymptotic standard error for d^e estimates is $\pi(24m)^{-1/2}$.

two regimes are preferred to three regimes and the orders of ARMA(1,1) were selected for the TARFIMA model in all regimes except for VXX where ARMA(1,1) coefficients were not significant and the model was simplified setting AR and MA parameters to zero.[20]

The long memory fractional integration parameters are positive and significant in both regimes. The 95% HPDI for parameter d is within the interval (0,1) for all cases.[21] The 95% HPDI includes 0.5 for all series. Therefore, the long memory parameter shows high persistence. As discussed in Section 2.3 volatility is still mean-reverting if $d<1$ and $\rho<1$ which is the case for our data.

The dynamics in the lower regime is generally more persistent than in the upper regime as can be seen from fractional parameters $d^{(1)}$ and $d^{(2)}$ for DIA, QQQQ and SPY given in Table 4. VXX has higher long memory parameter in the upper regime.[22] On the other hand, the AR parameters are similar around 0.3 for DIA, QQQQ and SPY and are zero for VXX. The implication of the TARFIMA model is that volatility is mean-reverting. Volatility reverts from the upper to the lower regime of normal volatility. By normal volatility we mean volatility below the estimated threshold values given in the second and third columns of Table 4. The percentage of observations in the lower volatility regime is between 85% and 90%. This means that shorter-lived spikes form a significant regime accounting for 10–15% of observations.

Table 4 Posterior means and standard deviations for TARFIMA model of $ln(\sqrt{RV})$ for ETFs.

	Threshold r	Vol	n1/n	$\rho^{(1)}$	$\rho^{(2)}$	$d^{(1)}$	$d^{(2)}$
DIA	0.395 (0.010)	23.56%	0.848	0.328 (0.087)	0.243 (0.199)	0.673 (0.040)	0.561 (0.076)
QQQQ	1.035 (0.006)	44.69%	0.899	0.328 (0.079)	0.236 (0.163)	0.658 (0.036)	0.420 (0.083)
SPY	0.383 (0.013)	23.28%	0.845	0.351 (0.096)	0.369 (0.270)	0.646 (0.039)	0.537 (0.090)
VXX	1.254 (0.036)	55.63%	0.864	–	–	0.445 (0.061)	0.720 (0.145)

This table presents posterior means and standard deviations (in brackets) for the parameters of two regime TARFIMA model of log realized volatilities. Vol is annualized volatility at the threshold level r, n1/n is % of observations in the lower regime. The data are for DIA (1/20/1998–10/29/2010) 3217 obs; QQQQ (3/10/1999–10/29/2010) 2931 obs; SPY (1/20/1998–10/29/2010) 3217 obs; VXX (1/30/2009–10/29/2010) 442 obs. For DIA, QQQQ and SPY the last 188 observations were left for out-of-sample forecast evaluation. For VXX the last 50 observations were left for forecasts.

20 The results of orders are not surprising since the ARFIMA(p,d,q) model is equivalent to the infinite lag ARMA model. Thus, the fractional integration parameter d takes care of higher order lags.

21 For VXX the 95% HPDI in regime 2 is [0.459, 0.994] with the upper bound close to 1. Since number of in-sample observations for VXX is relatively small and the estimation of TARFIMA model includes many parameters the results have higher standard deviations and wider confidence intervals.

22 Due to smaller number of observations for VXX the standard errors are larger.

We also found that σ^2, the variance of log volatility $ln(RV^{1/2})$ in equations (3)–(5), is significantly higher for the upper regime compared to the lower regime. Spikes in realized volatility are characterized by a higher variance of volatility, while a lower regime of volatility has more persistence[23] and smaller variance.

The results of estimated ARFIMA model (1) are given in Table 5. The second column gives estimated maximum absolute value of the inverse roots of the AR parameters (ρ) and their standard deviations. The third column gives estimates of fractional integration parameters d. For all ETFs the orders of ARMA(1,1) were selected using the significance of parameters and MBIC criterion.

The results for the TARFIMA model are different from a one regime ARFIMA model in terms of long memory parameter. Long memory parameter is higher for ARFIMA model compared to each regime of TARFIMA for 3 ETFs: DIA, QQQQ and SPY. VXX exhibits higher long memory in the upper regime of the TARFIMA model. We noted earlier that we have smaller number of observation for VXX (only about 50 observations in the upper regime and 300 in the lower regime). However, since VXX itself represents a volatility index it is different from other ETFs. We are looking at the dynamics of volatility of volatility and higher long memory in the upper regime may indicate the difference in the behavior of investors in VXX and equity ETFs. Since the VXX data are limited we would leave it for further research to investigate VXX compared to ETFs.

Table 6 compares the results for ARFIMA, TARMA, TARFI and TARFIMA models. TARMA model which accounts for different dynamics in ARMA parameters in two regimes shows different persistence in two regimes. We see that ρ (the maximum absolute value of the inverse roots of the AR parameters) is much higher in the lower regime showing higher persistence and showing mean reversion in the upper regime.[24] This model does not account for long memory. Since the 95% HPDI for the roots does not include one in any case we conclude that all series are stationary for the TARMA model.

TARFI model allows different parameters of long memory $d^{(1)}$ and $d^{(2)}$ in two regimes but does not account for ARMA in the error term. We see that for DIA, SPY and QQQQ there is some difference in the long memory parameters with parameters being slightly higher in the lower regime.[25] For all models the 95% HPDI for d does not include one and centers approximately at d=0.5. Thus the stability (mean-reversion) condition is satisfied.

Finally, as discussed above, for the most general TARFIMA model both conditions for stationarity in terms of ρ<1 and d<1 are satisfied.

Overall, we found that volatility is mean-reverting and there is less degree of long memory for the two regime TARFIMA model compared to ARFIMA model. This results indicate that changes in regimes and long memory are related and if regimes are not taken into account the persistence in d increases. Unlike long memory parameters the AR parameters are similar for ARFIMA and TARFIMA models. As observed in the literature abrupt shifts in volatility may induce long memory characteristics [see e.g., Andersen and Bollerslev (1998)].

Table 5 Posterior means and standard deviations for ARFIMA model of $ln(\sqrt{RV})$ for ETFs.

	Max abs inverse AR root ρ	Fractional parameter d
DIA	0.297 (0.055)	0.743 (0.049)
QQQQ	0.321 (0.048)	0.780 (0.053)
SPY	0.335 (0.060)	0.728 (0.049)
VXX	0.352 (0.213)	0.587 (0.153)

This table presents posterior means and standard deviations of the parameters for ARFIMA model of log realized volatilities. The data are for DIA (1/20/1998–10/29/2010) 3217 obs; QQQQ (3/10/1999–10/29/2010) 2931 obs; SPY (1/20/1998–10/29/2010) 3217 obs; VXX (1/30/2009–10/29/2010) 442 obs. For DIA, QQQQ and SPY the last 188 observations were left for out-of-sample forecast evaluation. For VXX the last 50 observations were left for forecasts.

23 Except for VXX.
24 For VXX there is no difference in dynamics.
25 However, considering the 95% HPDI the difference in parameters is not significant.

Table 6 Tests of stationarity for different specifications for ETFs.

	ARFIMA	TARMA	TARFI	TARFIMA
DIA				
$\rho^{(1)}$	0.309 (0.216, 0.396)	0.986 (0.977, 0.995)		0.282 (0.087, 0.492)
$\rho^{(2)}$		0.362 (0.221, 0.516)		0.220 (0.000, 0.626)
$d^{(1)}$	0.764 (0.676, 0.863)		0.520 (0.490, 0.551)	0.659 (0.576, 0.742)
$d^{(2)}$			0.481 (0.393, 0.574)	0.576 (0.425, 0.748)
Stationarity is not rejected in all specifications. Higher persistence in lower regime.				
QQQQ				
$\rho^{(1)}$	0.321 (0.225, 0.411)	0.994 (0.990, 0.998)		0.328 (0.164, 0.471)
$\rho^{(2)}$		0.008 (0.00002, 0.02770)		0.236 (0.000, 0.541)
$d^{(1)}$	0.780 (0.678, 0.885)		0.510 (0.482, 0.537)	0.658 (0.590, 0.730)
$d^{(2)}$			0.416 (0.297, 0.538)	0.420 (0.267, 0.594)
Stationarity is not rejected in all specifications. Higher persistence in lower regime.				
SPY				
$\rho^{(1)}$	0.335 (0.224, 0.448)	0.980 (0.966, 0.992)		0.351 (0.166, 0.543)
$\rho^{(2)}$		0.480 (0.329, 0.640)	(0.000, 0.851)	0.369
$d^{(1)}$	0.728 (0.635, 0.828)		0.518 (0.485, 0.550)	0.646 (0.572, 0.722)
$d^{(2)}$			0.492 (0.399, 0.585)	0.537 (0.375, 0.718)
Stationarity is not rejected in all specifications. Higher persistence in lower regime.				
VXX				
$\rho^{(1)}$	0.352 (0.000, 0.742)	0.629 (0.292, 0.952)	(0.002, 0.943)	0.474
$\rho^{(2)}$		0.666 (0.366, 0.937)	(0.000, 0.872)	0.399
$d^{(1)}$	0.587 (0.354, 0.928)		0.479 (0.268, 0.777)	0.535 (0.216, 0.903)
$d^{(2)}$			0.701 (0.483, 0.947)	0.509 (0.153, 0.910)
Stationarity is not rejected in all specifications. Higher persistence in upper regime for TARFI model.				

This table presents posterior means and 95% HPDIs (in brackets) for parameters of ARFIMA, TARMA, TARFI and TARFIMA models of log realized volatilities.

We also found different long memory and volatility of volatility for regimes of TARFIMA model. This information may also indicate the difference in investors behavior in high and low volatility regimes. The high volatility regime is characterized by higher variation in realized volatility which arises because of increased activity of the market. On the other hand, the regimes of high volatility are short-lived and mean-reverting.

In the following section we discuss the forecasting performance of each model.

4 Forecasting and model comparison

In this section, we compare the forecasting performance of the TARFIMA model with linear ARFIMA model. Estimates of each model for ETFs were first obtained for the in-sample period 1/20/1998 to 02/02/2010 and then used for repeated 1 and 5 day out-of-sample forecasts for the last 188 days (02/03/2010–10/29/2010). For VXX the in-sample period is 1/30/2009–08/19/2010 and the out-of sample 50 observations are from 08/20/2010 to 10/29/2010. For individual stocks GE and MSFT the in-sample period was 1/1/1994–7/29/2003 and out-of-sample period was 7/30/2003–12/31/2004. Data for HPQ are for the period 5/6/2002–7/29/2003 and out-of-sample period was 7/30/2003–12/31/2004. Finally daily VIX volatility index from CBOE website for the out-of sample period was used for the SPY 22 day forecast comparison.

ABDL 2003 used the fractionally integrated autoregressive model in order to account for long memory properties of realized volatility and found that their forecasts are better than forecasts produced by many popular models of volatility, e.g., daily GARCH, RiskMetrics, and FIEGARCH. Our TARFIMA model allows use of the information about the thresholds to predict regimes of high or low volatility and different dynamics in each regime.

In Tables 7–9 we present average posterior modes and medians of mean squared forecast errors (MSE) and mean forecast errors (ME) for 1 and 5 day forecasts for ETFs. Table 7 shows results for the TARFIMA model, Table 8 shows results for the ARFIMA model and Table 9 compares results for both models based on MSE.

From Tables 7 and 8 we see that both models show a small negative bias since ME is negative in most cases.[26] The bias may be explained since large spikes in realized volatility such as the "flash crash" are hard to predict and in such case volatility is underpredicted. Average ME for ARFIMA and TARFIMA models are similar.

Table 7 Posterior forecast MSE, ME and in sample MBIC for TARFIMA model for ETFs.

	1 day MSE		1 day ME		5 day MSE		5 day ME		MBIC
	Mode	Med	Mode	Med	Mode	Med	Mode	Med	Mean
DIA	0.142	0.167	−0.063	−0.026	0.205	0.220	−0.041	−0.014	−5597.19
QQQQ	0.111	0.146	−0.061	−0.019	0.191	0.219	−0.024	0.003	−5181.54
SPY	0.113	0.141	−0.069	−0.027	0.186	0.204	−0.051	−0.020	−5667.45
VXX	0.231	0.387	−0.147	−0.032	0.516	0.608	−0.088	0.001	−570.402

Notes: Averages of MSE modes and medians are shown for 1 and 5 day forecasts.
MBIC is evaluated at posterior mean.

Table 8 Posterior forecast MSE, ME and in sample MBIC for ARFIMA model for ETFs.

	1 day MSE		1 day ME		5 day MSE		5 day ME		MBIC
	Mode	Med	Mode	Med	Mode	Med	Mode	Med	Mean
DIA	0.144	0.167	−0.064	−0.027	0.201	0.221	−0.036	−0.010	−5587.65
QQQQ	0.111	0.146	−0.064	−0.021	0.189	0.221	−0.027	0.004	−5194.53
SPY	0.112	0.140	−0.067	−0.027	0.180	0.205	−0.043	−0.011	−5636.64
VXX	0.200	0.401	−0.150	−0.026	0.487	0.652	−0.085	0.017	−573.06

Notes: Averages of MSE modes and medians are shown for 1 and 5 day forecasts.
MBIC is evaluated at posterior mean.

Table 9 Comparison of MSE and MBIC for TARFIMA and ARFIMA models for ETFs.

	1 day MSE		5 day MSE		MBIC
	Mode	Med	Mode	Med	Mean
DIA	0.986	1.000	1.020	0.995	1.002
QQQQ	1.000	1.000	1.011	0.991	0.997
SPY	1.009	1.007	1.033	0.995	1.005
VXX	1.155	0.965	1.060	0.933	0.995

Notes: Averages of MSE modes and medians are shown for 1 and 5 day forecasts.
MBIC is evaluated at posterior mean.
MSE is for TARFIMA model relative to ARFIMA model for which MSE is set to 1.
MBIC is for TARFIMA model relative to ARFIMA model for which it is set to 1 (here we maximize MBIC).

26 However the standard deviations of the ME are large enough to include zero in the HPDIs.

In terms of the MSE and MBIC criterion, Table 9 summarizes results for both models for ETFs. Relative MSE is normalized relative to the ARFIMA model. Using the average modes and medians of MSE listed in Table 9 we can compare the forecasting performance of TARFIMA and ARFIMA at 1 and 5 day forecasting horizons based on minimum average mode (or median) of MSE as defined in Section 2.5. As we can see from Table 9 the results are mixed for 1 and 5 day forecasts. For example, for the 5 day forecasts using the modes of the distributions of MSE we would choose the ARFIMA model, while using the median (or mean) we would choose the TARFIMA model. Since the distribution of MSE is skewed to the right (MSE takes only positive values) it is not clear whether we should use the mode or the median. Therefore below we further investigate the performance of models based on the comparison of the whole distributions of MSE .

Better out-of-sample forecasts do not always imply better in-sample fit to which we turn next. The MBIC criterion results are mixed showing that ARFIMA is preferred in two cases and TARFIMA is preferred in the other two cases. However, the TARFIMA model has more that twice number of parameters compared to ARFIMA and in spite of higher penalty MBIC results are generally close for the two models.

Table 10 reports relative MSE and relative MBIC for three selected stocks: GE, MSFT and HPQ. Here the results of MSE for all horizons favor TARFIMA model for GE and MSFT and favor ARFIMA model for HPQ. The sample for these data ends in 2004 and there are no extreme spikes observed in the financial crisis of 2008–2009 and for the flash crash. However, the data for MSFT and GE contains more extreme spikes than HPQ due to contagion of Asian crisis in 1998.[27] This may explain the preference for TARFIMA model for stocks when there are extreme events such as Asian crisis. For the in-sample evaluation in spite of more than twice number of parameters for the TARFIMA model the in-sample MBIC criterion is maximized for TARFIMA model for all stocks.

Tables 7–10 are summary statistics of forecast performances. They do not give the whole pictures of the forecasts. Accordingly, we present the graphs of the forecasts as compared to actual volatilities.

Figure 3 shows out-of-sample posterior means of 1 and 5 day forecasts and actual data. The in-sample data was presented in Figure 1. We can notice that forecasts look very similar and close to actual data except for the "flash crash" on May 6, 2010.[28] Figure 4 shows a distributions of 1 and 5 day forecasts for QQQQ for a single day and the predictive distributions of two models look similar for the 1 day forecast with somewhat increased difference for a 5 day forecast.

Figure 5 examines annualized SPY realized volatility and its 22 day posterior mean forecasts compared to implied volatility measured by VIX.[29] As is commonly found VIX is generally above the historical realized volatility which could be justified since investors pay a risk premium when buying volatility. As a result there

Table 10 Comparison of MSE and MBIC for TARFIMA and ARFIMA models for GE, MSFT and HPQ.

Firm	N obs	1 day MSE		5 day MSE		MBIC
		Mode	Median	Mode	Median	Mean
GE	3024	0.857	0.964	0.928	0.935	1.014
MSFT	3024	0.975	0.994	0.959	0.967	1.004
HPQ	671	1.084	1.056	1.149	1.223	1.825

Notes: Data for GE and MSFT are for the period 1/4/1993–12/31/2004.
Data for HPQ are for the period 5/6/2002–12/31/2004.
The last 360 observations were left for out-of-sample evaluation.
Averages of MSE modes and medians are shown for 1 and 5 day forecasts.
MBIC is evaluated at posterior mean.
MSE is for TARFIMA model relative to ARFIMA model for which MSE is set to 1.
MBIC is for TARFIMA model relative to ARFIMA model for which it is set to 1 (here we maximize MBIC).

27 HPQ data starts in 2002.
28 The 95% HPDIs not presented here include actual data except for the flash crash.
29 The data for daily VIX is from CBOE website http://www.cboe.com.

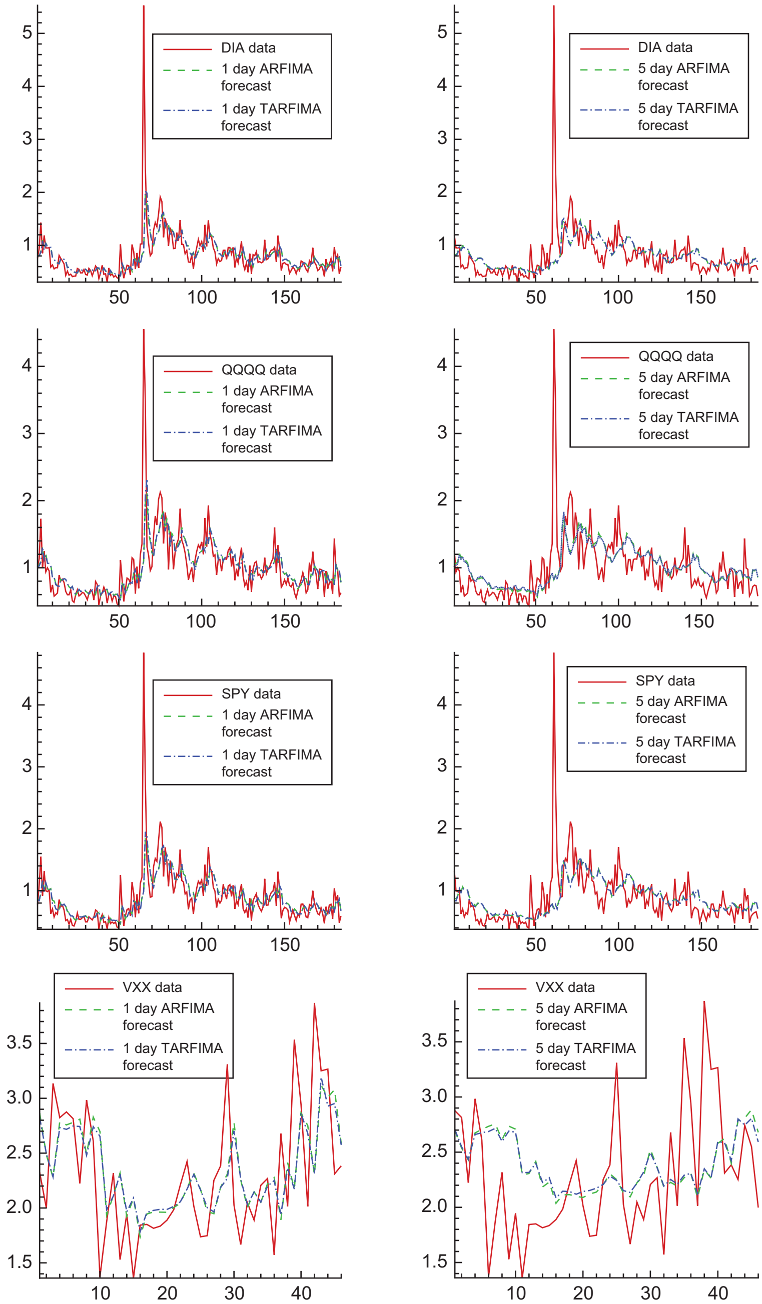

Figure 3 Posterior means of 1 and 5 day forecasts of realized volatility for ETFs.

is a positive spread between VIX and our 22 day forecasts. VIX and our forecasts seem to be co-moving with a spread between them.[30]

Figure 6 shows posterior probability densities (PDFs) for standardized logarithms of realized volatility forecasts for three stocks: GE, MSFT and HPQ. The distributions are close to normal with slight asymmetries. The forecast residuals can be approximated as normally distributed. Modeling the error term as normally

30 One of the potential applications of threshold type models is to estimate this model for the spread and see when the spread is above or below threshold levels making trading in options profitable. We leave this for future research.

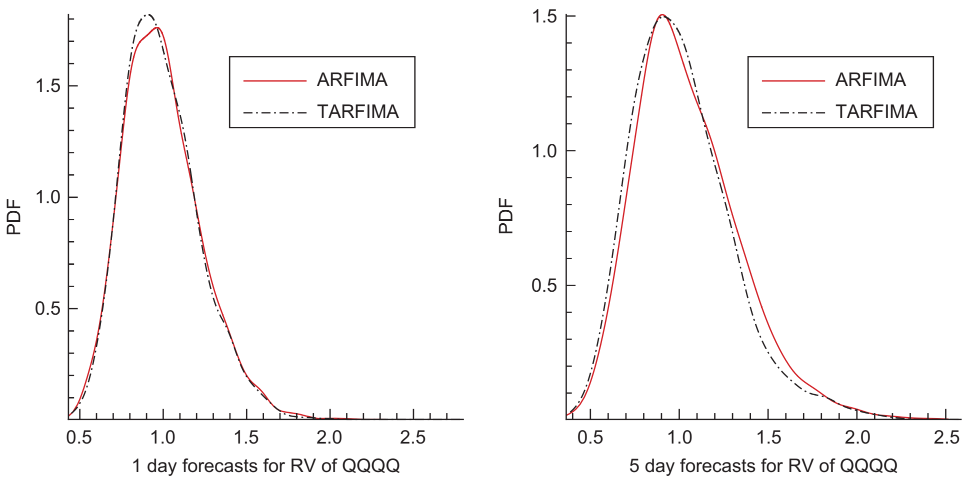

Figure 4 Posterior pdfs of log realized volatility forecasts for QQQQ.

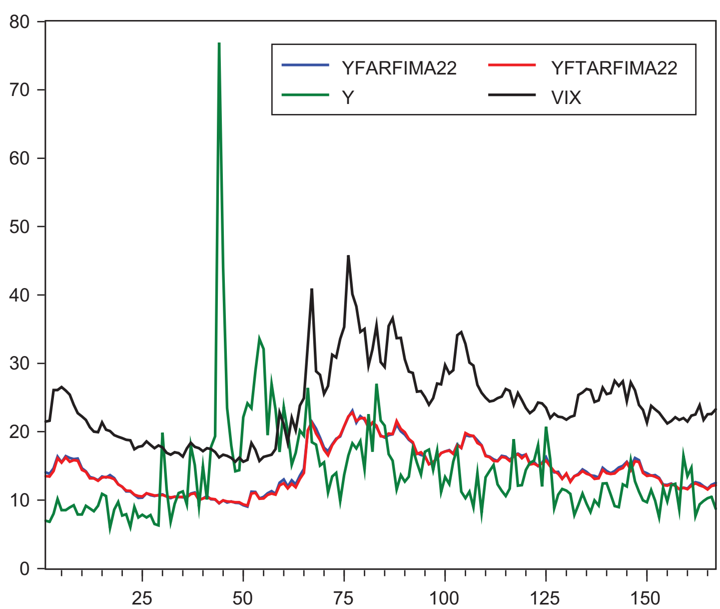

Figure 5 Posterior means of 22 day forecasts, VIX and actual annualized realized volatility data for SPY.

distributed is a fairly standard assumption for the $log(RV^{1/2})$ in the literature [see Andersen et al. (2001a,b, 2003) and Thomakos and Wang (2003) among others]. Although the Jarque Bera (JB) test rejects normality for the logarithm of realized volatility data it was shown in Thomakos and Wang (2003) that the JB test is oversized when the series exhibits long memory or autocorrelation and results in more bias than other tests of normality. Since we found long memory and autocorrelation of realized volatility for all series the JB test is not appropriate. Other tests, such as the Kolmogorov-Smirnov test, may or may not reject normality and could also be sensitive to autocorrelation. The residuals of TARFIMA and ARFIMA models (correcting for autocorrelation) are closer to the normal distribution than original data. We note the advantage of using the TARFIMA model which provides unconditional distribution of the error term as a mixture of normal distributions with two different variances corresponding to upper and lower regimes. Thus theoretically it should give a better approximation to the true distribution of the data.

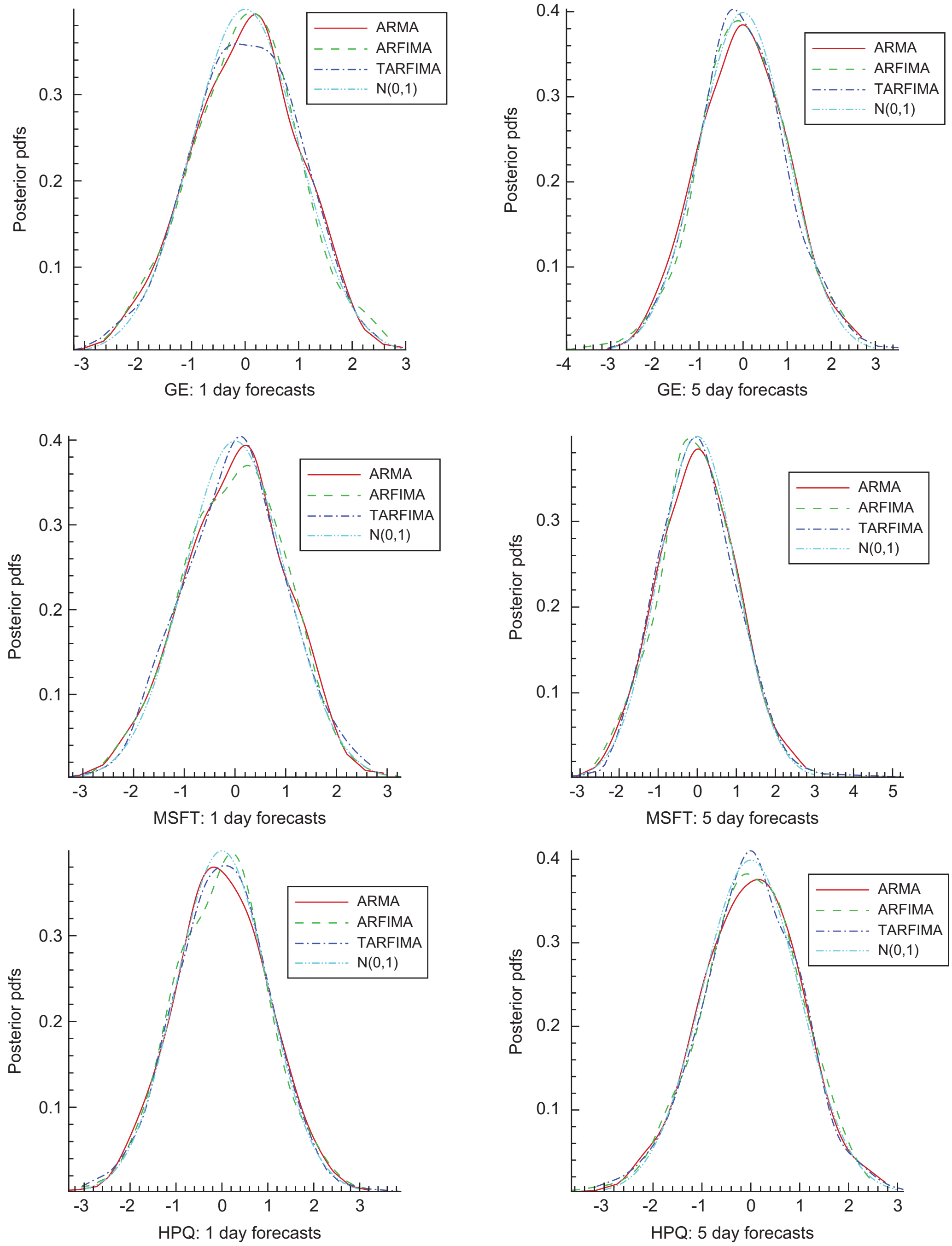

Figure 6 Posterior pdfs of normalized log realized volatility forecasts for GE, MSFT and HPQ.

In Figure 7 we show posterior cumulative densities (CDFs) of the medians of MSE for the ARFIMA and TARFIMA models.[31] The CDFs are calculated on the same set of horizontal points for MSE using the first degree Taylor approximation.[32] From the visual comparison of CDFs we see that the distribution of medians

31 Using modes of MSE instead of medians provides similar results.

32 We thank Evgeny Goldman for suggesting the following way to provide a common grid for two CDF graphs. Let a_i and b_i denote samples of xs for two CDF graphs. The grid of common x's is given by $x_i=\Delta x_i$, $i=1,\ldots,nt$, where number of points in the grid $nt=min((a_{max}-a_{min})/\Delta,\ (b_{max}-b_{min})/\Delta))$. The first degree Taylor approximation for two CDFs is given by $F_a(x_i)=F(a_i)+(x_i-a_i)\dfrac{F(a_i)-F(a_{i-1})}{a_i-a_{i-1}}$ and $F_b(x_i)=F(b_i)+(x_i-b_i)\dfrac{F(b_i)-F(b_{i-1})}{b_i-b_{i-1}}$ correspondingly.

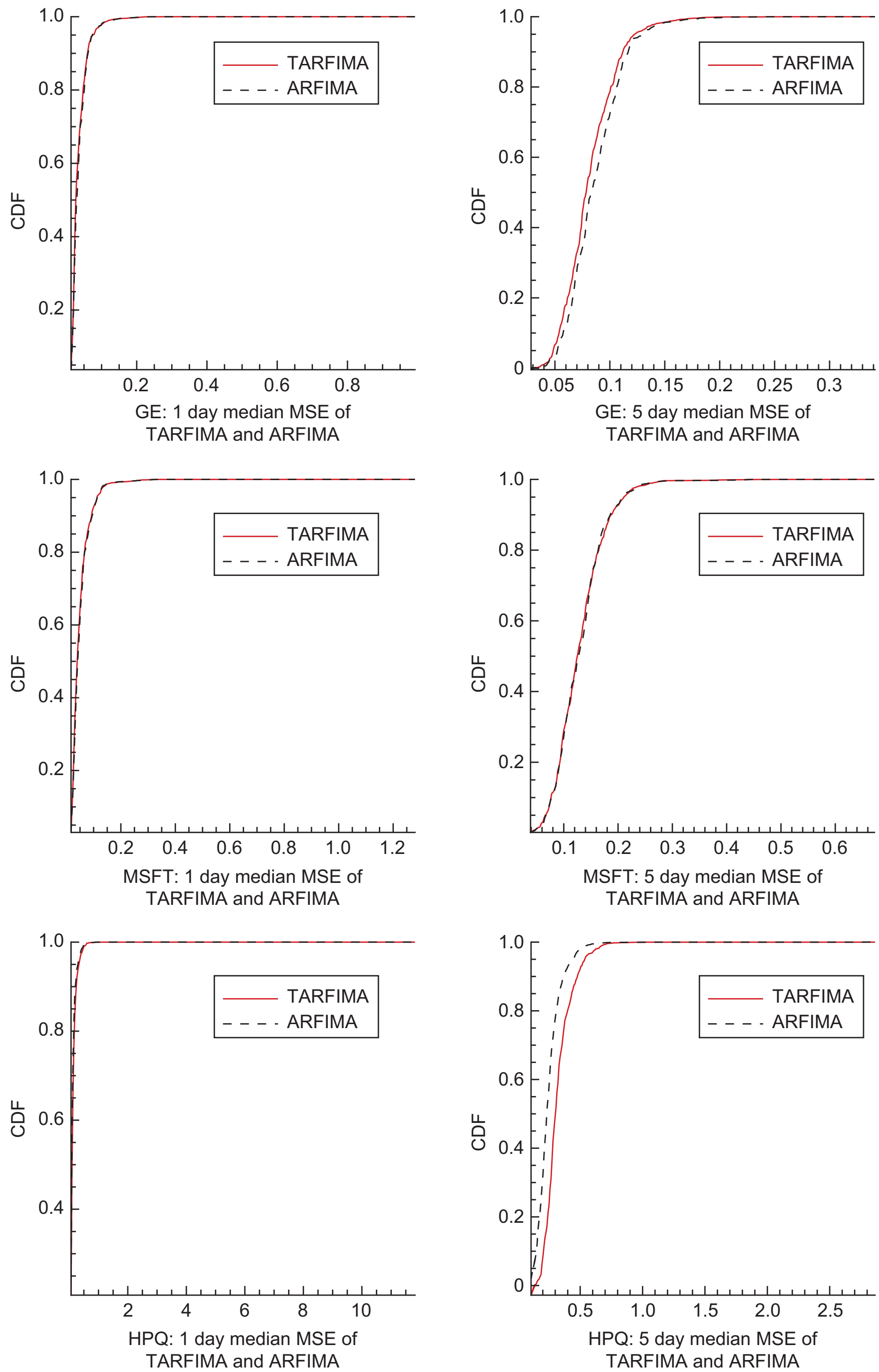

Figure 7 Posterior cdfs of the median MSE(ARFIMA) and median MSE(TARFIMA).

of MSE is closer to zero for the TARFIMA model for the 5 day forecasts of GE and for the ARFIMA model for the 5 day forecasts of HPQ. In other cases TARFIMA and ARFIMA CDFs look very close.

Figure 8 shows the PDFs of the differences between the CDFs in Figure 7 for ARFIMA and TARFIMA models. The pdfs of the difference provide a statistical test of the difference in the distributions in Figure 7. Positive values indicate that the ARFIMA model outperforms TARFIMA in MSE, while negative difference shows the superiority of the TARFIMA model. If x is the distance between CDF of ARFIMA and TARFIMA, then we can define the probability of the TARFIMA model being selected over ARFIMA as given by $P(x<0)$.

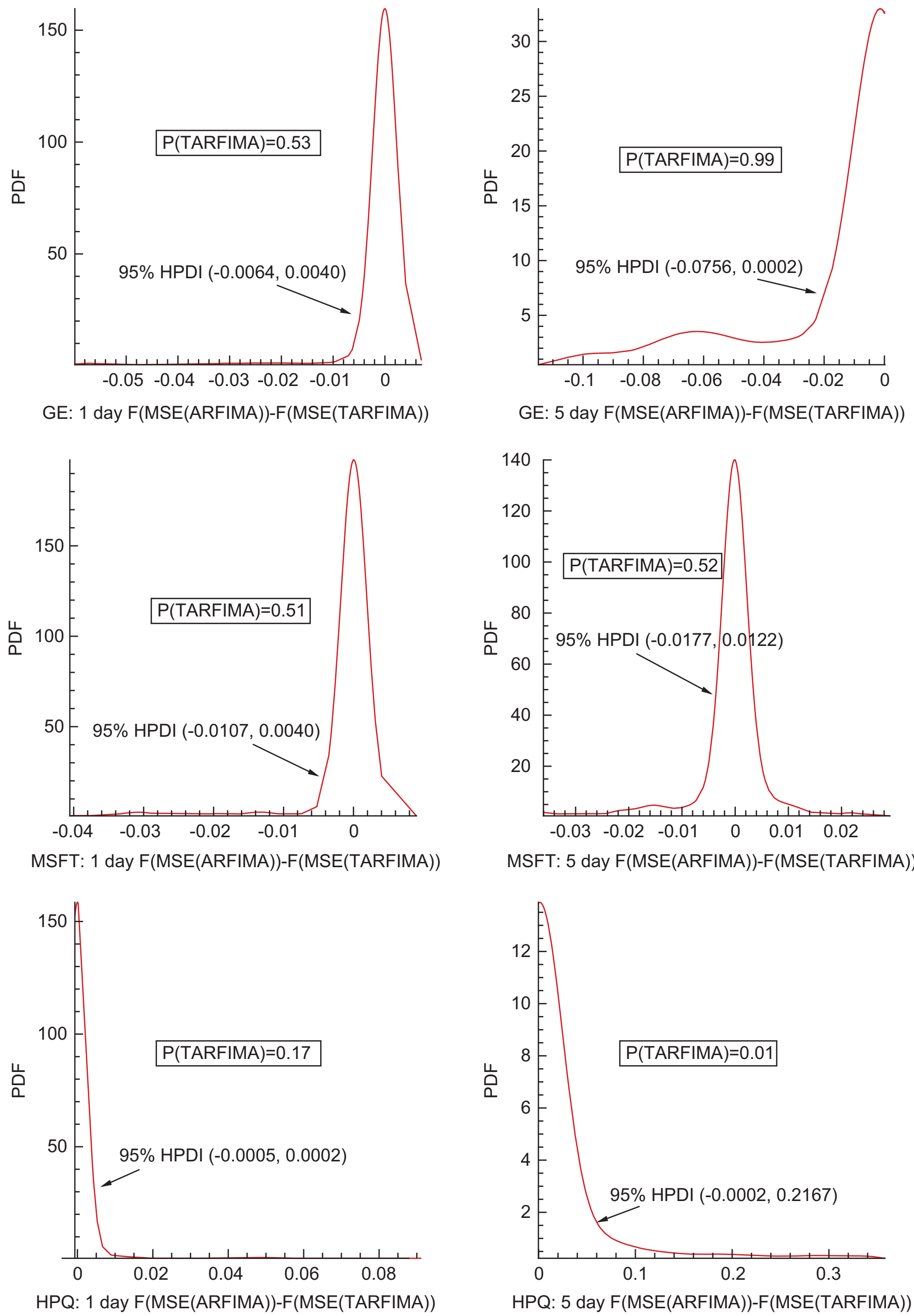

Figure 8 Posterior pdfs of the median F(MSE(ARFIMA))-F(MSE(TARFIMA)).

The results for three companies are as follows. The probability that the TARFIMA model is superior is 0.53 and 0.99 for GE for 1 and 5 day forecasts, respectively. Thus, TARFIMA is a preferred model for GE especially for a longer-term forecast where the distribution of MSE of TARFIMA dominates that of ARFIMA with 99% probability. For MSFT the probability of selecting TARFIMA is 0.51 and 0.52 for 1 and 5 day forecasts. We select TARFIMA models but probabilities of ARFIMA and TARFIMA models are close to 0.5 thus giving similar MSE distributions. For HPQ the probabilities of TARFIMA are 0.17 and 0.01, thus we select the ARFIMA model with high probability in each case. These results confirm the results in Table 10 where TARFIMA was selected for GE and MSFT and ARFIMA for HPQ based on average of MSE medians or modes.

Figure 9 and Table 11 present results for probabilities of ARFIMA and TARFIMA models for ETFs using the same method of distribution of CDF differences. The results show that TARFIMA model is selected for all four ETFs for 5 day forecasts. The probabilities of TARFIMA model for DIA, QQQQ, SPY and VXX are 0.598, 0.748, 0.520 and 0.989, correspondingly.[33] On the other hand, ARFIMA is selected for three ETFs for a 1 day forecast with exception of VXX for which TARFIMA is preferred to a 1 day forecast as well. Our overall conclusion is

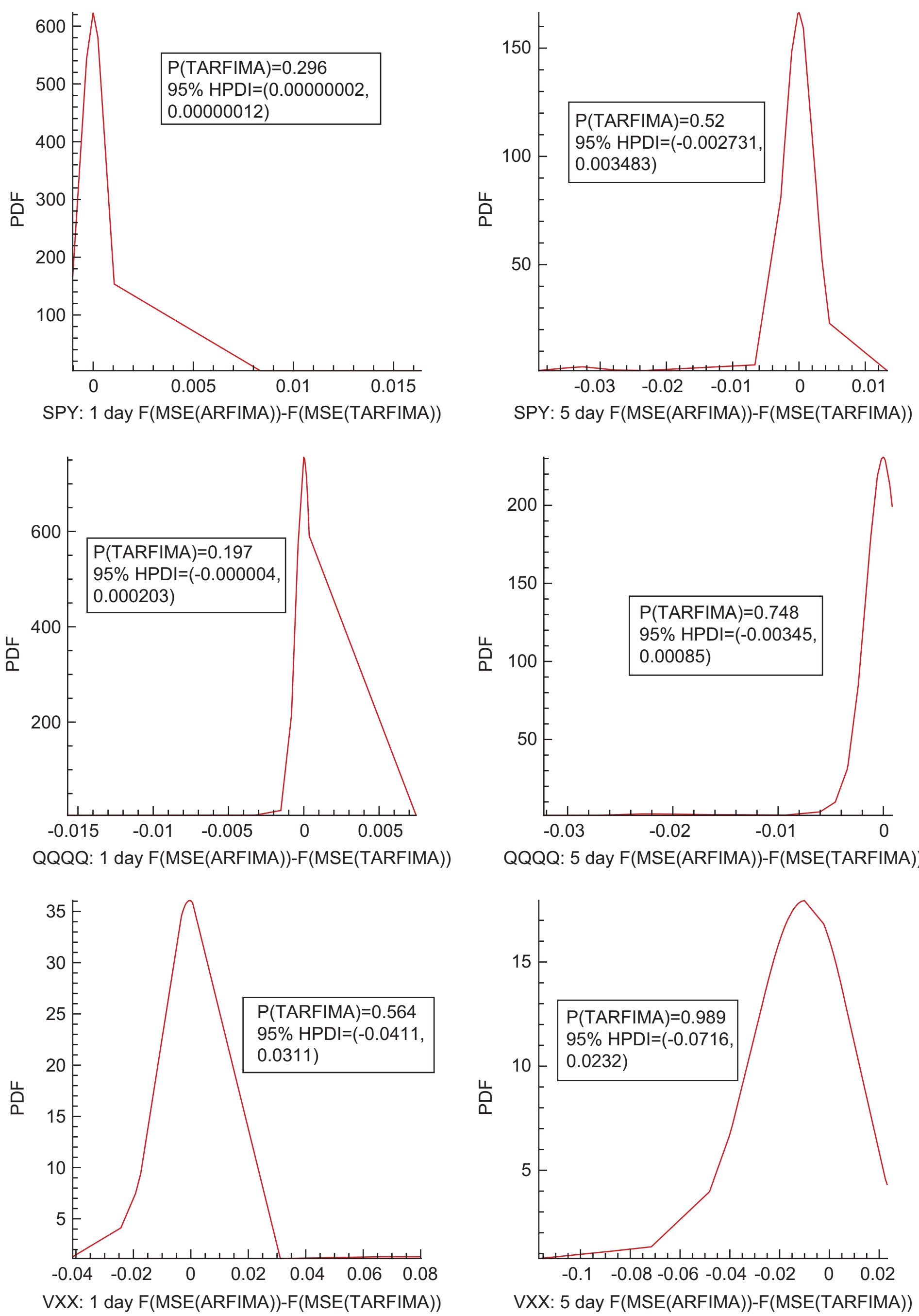

Figure 9 Posterior pdfs of the median F(MSE(ARFIMA))-F(MSE(TARFIMA)).

33 Generally these results hold for longer term forecasts such as 10 days as well.

Table 11 Model choice using the distribution of *F*(*MSE*(*ARFIMA*))–*F*(*MSE*(*TARFIMA*)) for ETFs.

ETF	1 day forecast		5 day forecast	
	95% HPDI	*P*(*ARFIMA*)	95% HPDI	*P*(ARFIMA)
DIA	(7.E-08, 9.E-08)	0.834	(9.22.E-04, 2.123.E-03)	0.402
QQQQ	(-4.E-06, 2.03.E-04)	0.803	(−3.45.E-03, 8.5.E-04)	0.252
SPY	(2.E-08, 1.2.E-07)	0.704	(−2.73.E-03, 3.48.E-03)	0.480
VXX	(−0.0519, 0.0102)	0.436	(−0.1080, 0.0016)	0.011

Notes: *P*(*ARFIMA*) is the probability of ARFIMA model computed as *P(F*(*MSE*(*ARFIMA*))–*F*(*MSE*(*TARFIMA*)>0).

that both models are useful is forecasting the realized volatilities of individual stocks and ETFs with general preference for TARFIMA model for longer term forecasting.

The TARFIMA model has more than twice the number of parameters than the ARFIMA model, thus it is better to use large samples for the TARFIMA model. On the other hand, the TARFIMA model captures the differences in persistence, long memory and volatility in two regimes, thus it can add information for better forecasts in large samples, especially for multi-day forecasts compared to 1 day forecasts if sample size is sufficiently large to allow better precision in estimation of parameters.

The TARFIMA model longer term forecasts are generally superior to linear models because of information about the threshold and regime conditioned on which the forecast is made. As discussed in Section 3 the difference in dynamics reflects the difference in investors' behavior in high and low volatility regimes.

The method of model selection using the posterior distribution of the differences between CDFs of MSE is new in this paper. It is better than methods based on point MSE estimate comparison since the latter does not account for uncertainty of MSE estimation. The posterior predictive densities of forecasts and posterior distributions of MSE provide information about uncertainty and tails of forecasts and provide straightforward tests of model selection based on the dominance of the distribution of MSE rather than point estimates of MSE. This is an advantage of the Bayesian MCMC approach used in this paper since posterior distributions of parameters and functions of interest are easily obtainable from MCMC chain.[34]

The model choice based on MSE distributions was used in this paper based on the popularity of MSE for forecasts comparisons. The CDF comparison test explained above can be applied in a straightforward manner to alternative measures of forecast accuracy, such as mean error or mean absolute error. Moreover we can compare more than two models based on their CDF distributions and rank them using probabilities of selecting each model as above for TARFIMA and ARFIMA.

In the classical approach it has become common to use the Diebold and Mariano (1995) test of superior predictive ability with many simulated forecasts for obtaining the distribution of MSE. In our future research we would like to compare the Bayesian approach of model selection with the Diebold and Mariano (1995) test.

5 Conclusion

In this paper, we used a threshold fractionally integrated autoregressive and moving average (TARFIMA) model to analyze the dynamics and forecasts of square root of quadratic variation in intra-daily returns also known as realized volatility. We find that the logarithmic transform of volatility processes of four ETFs and selected three stocks in the Dow Jones Index can be characterized by regimes of high and low volatility. The persistence, long memory and variance of volatilities may change with each regime. For the periods of extreme volatility, such as during a financial crisis, the TARFIMA model with regimes based on

34 The GAUSS codes are available from the authors.

past volatility seems to be more appropriate than linear models as TARFIMA explicitly accounts for the dynamic properties of a high volatility regime. The use of the Markov-switching model with long memory seems to be not as appropriate since its regimes are randomly determined by an unobservable variable, while for the TARFIMA model regimes are determined by the previous day level of volatility or some other observable variables.

In addition, we also compared the forecast effectiveness of TARFIMA and ARFIMA models. We find that the TARFIMA model that accounts for a different degree of long memory, persistence and variance in two regimes in longer term forecasts exceeding five days performs better than the ARFIMA model. A new test based on the posterior distributions of mean squared forecast errors (MSE) is used for model comparison.

Given that the ARFIMA model is known to produce forecasts that are better than the forecasts by the ARMA and other models (ABDL 2003), our results show that TARFIMA models may provide more benefit to investors in the areas of portfolio optimization and risk where longer term forecasts are important.

Future research may be directed to studying how TARFIMA models can better accommodate jumps, for example, using market news and individual company news announcements. We could apply Lee and Mykland (2008) nonparametric tests for jumps since they showed that their test outperforms tests based on bipower variation. TARFIMA models can be applied to other financial markets and used for identifying regimes of high volatility. Also we can can compare our Bayesian approach of model selection with Diebold and Mariano (1995) and other tests in the classical approach.

6 Appendix

6.1 Bayesian estimation of TARFIMA and ARFIMA models

Let the prior probability density function for the TARFIMA model be given by

$$\begin{aligned}\pi(\gamma,\phi,\theta,\sigma^2,r,d)\propto&\prod_{j=1}^{s+1}\mathrm{N}(\gamma_0^{(j)},\Sigma_\gamma^{(j)})\,\mathrm{N}(\phi_0^{(j)},\Sigma_\phi^{(j)})\,\mathrm{N}(\theta_0^{(j)},\Sigma_\theta^{(j)})\\&\times IG(\nu_0^{(j)},\delta_0^{(j)})\,I(r^{(j)}\in[r_{low}^{(j)},r_{up}^{(j)}])I(d^{(j)}\in[-0.5,\,1])\end{aligned} \tag{6}$$

where $\gamma, \phi, \theta, \sigma^2$ and d are

$$\begin{aligned}\gamma&=(\gamma^{(1)},\ldots,\gamma^{(s+1)})\\\phi&=(\phi_1^{(1)},\ldots,\phi_{p^{(1)}}^{(1)},\ldots,\phi_1^{(s+1)},\ldots,\phi_{p^{(s+1)}}^{(s+1)})\\\theta&=(\theta_1^{(1)},\ldots,\theta_{q^{(1)}}^{(1)},\ldots,\theta_1^{(s+1)},\ldots,\theta_{q^{(s+1)}}^{(s+1)})\\\sigma^2&=((\sigma^{(1)})^2),\ldots,(\sigma^{(s+1)})^2)\\d&=(d^{(1)},\ldots,d^{(s+1)}).\end{aligned}$$

We assume a uniform prior for threshold parameters r where each $r^{(j)}$ is constrained so that minimum m% of observations are within each regime. We use m=10% of total number of observations as a minimum sample size in each regime.[35] We also assume uniform prior distribution for d in the parameter space between [−0.5, 1]. Other parameters have proper normal-inverted gamma priors with large variances in normal distributions.[36]

The posterior pdf is

35 m depends on the number of observations, if the sample size is small higher m is recommended. For example, Koop and Potter (1999) used 15% as a minimum sample size in each regime.

36 This priors are standard in the literature for ARMA models [see Chib and Greenberg (1994) among others].

$$p(\gamma,\phi,\theta,\sigma^2,r,d|Y,X)\propto\pi(\gamma,\phi,\theta,\sigma^2,r,d)\times\prod_{j=1}^{s+1}\prod_{t\in T_j}\frac{1}{\sigma^{(j)}}\phi\left(\frac{(1-B)^{d^{(j)}}y_t-g(Z_t))}{\sigma^{(j)}}\right),\tag{7}$$

where for every

$$\begin{aligned}t\in T_j&=\{t:r_{j-1}\le z_{t-\delta}<r_j\}\\ u_t&=(1-B)^{d^{(j)}}y_t-x_t\gamma^{(j)}\\ \varepsilon_t&=(1-B)^{d^{(j)}}y_t-g(Z_t)\\ g(Z_t)&=x_t\gamma^{(j)}-\sum_{i=1}^{p^{(j)}}\phi_i^{(j)}u_{t-i}-\sum_{i=1}^{q^{(j)}}\theta_i^{(j)}\varepsilon_{t-i}.\end{aligned}$$

MCMC algorithms for ARMA model were developed by Chib and Greenberg (1994). The autoregressive AR and variance parameters are drawn using a Gibbs sampler and moving average MA parameters are drawn using a Metropolis-Hastings algorithm. Chib and Greenberg (1994) [as well as Nakatsuma (2000)] use the constrained nonlinear maximization algorithm in the MA block. Alternatively one can use a Metropolis-Hastings algorithm with a random walk Markov Chain as was done e.g., in Goldman and Tsurumi (2005). The random walk draws speed up the computational time of the MCMC algorithms without losing much of the acceptance rate of the Metropolis-Hastings algorithm. A simple intuitive explanation of the Metropolis-Hastings algorithm is given in Chib and Greenberg (1995).

In the literature for ARFIMA models there are several ways to handle the fractional difference $(1-B)^d y_t$. Sowell (1992) used the exact maximum likelihood. Hauser (1999) compares Sowell's procedure with others and shows Sowell's procedure tends to dominate the others. Bayesian estimation of ARFIMA models was done by Koop et al. (1997), Gil-Alana (2001), and Gouskova (2002), among others. Shimizu and Tsurumi (2009) compare several procedures from a Bayesian perspective. They show that when the sample size is large the posterior pdf of d is insensitive to the choice of n_{lags}, where n_{lags} is the truncation point of the infinite series (2).

In this paper, the ARFIMA and TARFIMA models are estimated using Metropolis-Hastings MCMC algorithms with efficient jump. In order to estimate threshold parameters $r=(r_1,\ldots,r_s)$ and fractional difference parameters $d=(d_1,\ldots,d_{s+1})$ we employ an algorithm with the efficient Metropolis jumping rule that is described in Gelman et al. (2004) and was used for a threshold ARMA model in Goldman and Agbeyegbe (2008). The algorithm described below allows efficient simultaneous estimation of multiple thresholds and fractional difference parameters.[37]

We briefly describe the algorithm for the case of two and three regimes. We estimate parameters in blocks: (i) regression parameters, γ; (ii) AR coefficients ϕ, (iii) MA coefficients θ; (iv) variance σ^2 parameters; (v) threshold parameters r; and (vi) fractional difference parameters d.

We describe the procedure below.

(i) Choose initial values for γ, ϕ, θ, σ^2, r, d. Start from crude estimates of mean or mode of the posterior distribution. Generally it is hard to find a good approximation for mean or mode of threshold distribution because of an unconventional shape of the likelihood function. For starting values we simply divide the sample into regimes with equal samples. In case we have two regimes we use the mean value of y_t as a starting point $r^{(0)}$, if there are three regimes we divide sample into three equal subsamples to find starting

37 Compared to existing methods of estimating multiple thresholds this method is more efficient, since we avoid estimation on a grid of points. Simultaneous estimation of multiple thresholds and difference parameters using many dimensions of grid search is virtually impossible. The Bayesian analogue of a grid search is a Griddy Gibbs sampler withing MCMC which was done for a single threshold parameter in Phann, Schotman, and Tschering (1996).

values $r_1^{(0)}$ and $r_2^{(0)}$. Using initial values for d in each regime the series y_t is fractionally differenced using the binomial expansion (2) with some limited number of lags n_{lags}.

Let the started points be denoted by $\gamma^{(0)}$, $\phi^{(0)}$, $\theta^{(0)}$, $\sigma^{2\,(0)}$, $r^{(0)}$, $d^{(0)}$. Given initial values $r_1^{(0)}$, $r_2^{(0)}$ the samples $\{y_t\}$ and $\{x_t\}$ are separated into regimes based on $y_{t-\delta}$:

$$\begin{cases} j=1 & y_{t-\delta}<r_1 \\ j=2 & r_1\leq y_{t-\delta}<r_2 \\ j=3 & y_{t-\delta}\geq r_2 \end{cases}$$

(ii) Once the data are separated into regimes given thresholds r the model is transformed into an arranged ARFIMA model.[38] After y_t is differenced using fractional parameters d in each regime and using equation (2) with fixed n_{lags} the model becomes an arranged ARMA model (with different parameters in each regime), the estimation of which using MCMC is standard [see Chib and Greenberg (1994) who used independence chains or Goldman and Tsurumi (2005) who used random walk algorithm]. We draw parameters of ARMA in each regime block by block using random walk Markov Chain. We draw parameters of variance σ^2 using the standard inverted gamma distribution. The acceptance rates of ARMA blocks are controlled by multiplying the variance of the proposal density with a scaling constant.

(iii) **Threshold parameters**

We use the following procedure with efficient jump.

Let the superscript, (i), denote the i-th draw. Each threshold parameter $r_j^{(i)}$ $(1\leq j\leq s)$ can be drawn either in a separate block given other threshold parameters, or all thresholds $\{r^{(i)}=(r_1^{(i)},\ldots,r_s^{(i)})\}$ could be drawn in one block, in the latter case the acceptance rate is lower.

We generate

$$r_j^{(i)}\sim N(r_j^{(i-1)},\ stdr_j^{(i-1)}),$$

where $stdr_j^{(i-1)}$ is initially selected as a constant C_0, such that the proposal normal distribution covers all threshold parameter space (e.g., C_0=quarter-distance between upper and lower bound for each regime). After sufficient number of draws mmm we set $stdr_j^{(i-1)}$ equal to the standard deviation of the sample of accepted draws $\{r_j^{(l)},\ l=1,\ldots,i-1\}$ multiplied by a scaling constant C. The variance of the proposal density is therefore proportional to the variance matrix estimated from the simulation.

$$stdr^{(i)}=C\star stdr(\{r^{(l)}\}),\quad l=n_0,\ldots,i-1.$$

Variance is adjusted using a scaling constant C, so that the acceptance rate is reasonable. Gelman et al. (2004) suggest optimal acceptance rate of 44% for 1 parameter and 23% for many parameters.

If $r_j^{(i)}$ does not satisfy the condition

$$r_j^{low}<r_j^{(i)}<r_j^{up},$$

where r_j^{low} and r_j^{up} are defined so that regimes below and above $r_j^{(i)}$ have at least m% of observations, then generate $r_j^{(i)}$ again until it falls within upper and lower bounds. We set m=5%.

We accept $r^{(i)}=(r_1^{(i)},\ldots,r_s^{(i)})$ with probability

$$a_r=\min\left\{\frac{p(\gamma^{(i)},\phi^{(i)},\theta^{(i)},r^{(i)}\,|\,\text{data})}{p(\gamma^{(i)},\phi^{(i)},\theta^{(i)},r^{(i-1)}\,|\,\text{data})},1\right\}.$$

Otherwise set $r^{(i)}=r^{(i-1)}$.

---

38 We construct an arranged ARFIMA model in a similar way as Tsay (1989) and others constructed arranged autoregressive model sorting data y_t by regimes.

Alternatively, one can construct a separate block and acceptance rate for each threshold.

(iv) **Fractional difference parameters**

We use the following procedure with efficient jump.

Each difference parameter $d_j^{(i)}$ $(1 \le j \le s+1)$ can be drawn either in a separate block given other difference parameters, or all parameters $\{d^{(i)}=(d_1^{(i)},\dots,d_{s+1}^{(i)})\}$ could be drawn in one block (in the latter case the acceptance rate is lower). We generate

$$d_j^{(i)} \sim N(d_j^{(i-1)}, stdd_j^{(i-1)}),$$

where $stdd_j^{(i-1)}$ is initially selected as a constant C_0, such that the proposal normal distribution covers all parameter space [e.g., C_0=quarter-distance between upper and lower bounds (–0.5, 1)]. After sufficient number of draws *mmm* we set $stdd_j^{(i-1)}$ equal to the standard deviation of the sample of accepted draws $\{d_j^{(l)},\ l=1,\dots,i-1\}$ multiplied by a scaling constant C. The variance of the proposal density is therefore proportional to the variance matrix estimated from the simulation.

$$stdd^{(i)}=C \star stdd(\{d^{(l)}\}), \quad l=n_0,\dots,i-1$$

Variance is adjusted using a scaling constant C, so that acceptance rate is reasonable.

If $d_j^{(i)}$ does not satisfy the condition

$$-0.5<d_j^{(i)}<1,$$

then generate $d_j^{(i)}$ again until it falls within upper and lower bounds.

We accept $d^{(i)}=(d_1^{(i)},\dots,d_{s+1}^{(i)})$ with probability

$$a_d=\min\left\{\frac{p(\gamma^{(i)}, \phi^{(i)}, \theta^{(i)}, r^{(i)}, d^{(i)}|\text{data})}{p(\gamma^{(i)}, \phi^{(i)}, \theta^{(i)}, r^{(i)}, d^{(i-1)}|\text{data})}, 1\right\}.$$

Otherwise set $d^{(i)}=d^{(i-1)}$.

Alternatively, one can construct a separate block and acceptance rate for each parameter d.

After d is drawn the series y_t is fractionally differenced using binomial expansion (2) with some limited number of lags n_{lags}. We change n_{lags} and make sure that the posterior pdf of d is insensitive to $n_{lags.}$[39]

After drawing threshold parameters r and filtering y_t with current draws of $d^{(j)}$ [where regime j is defined in (5)] the MCMC algorithm steps become identical to the steps in the ARMA model.

As with any MCMC procedure, we make N draws of the parameters in each of the blocks, and we burn the first m draws. Out of the remaining $N-m$ draws, we keep every h-th draw. We check convergence by testing that the draws attain mean and covariance stationarity.[40]

6.2 Choice of the number of regimes and orders

Estimation of a TARFIMA model involves the choice of (i) number of regimes and delay parameter and (ii) orders of ARMA(p,q) process in each regime. *ARFIMA*(p,d,q) model choice involves orders p and q.

For each model the orders p, q, the number of regimes $s+1$ and the delay parameter δ are chosen using the significance of coefficients of lags, marginal likelihood, and a modified Bayesian information criterion

39 We found that n_{lags}=40 which corresponds to 40 days was sufficient. Since in this paper we generally work with large samples sizes the problem of truncation does not arise.

40 For example, we use Kolmogorov Smirnov and the filtered fluctuation tests [these tests are studied and compared in Goldman, Valieva, and Tsurumi (2008)].

(MBIC) discussed in Goldman and Tsurumi (2005). This criterion is a Bayesian analogue of Akaike information criterion given by

$$AIC=-2\sum_{j=1}^{n_{regimes}} ln(L_j(p_j, q_j, s, \delta))+2(\nu+1),$$

where $ln(L_j())$ is a log-likelihood function for regime j and ν are degrees of freedom.

The modified Bayesian information criterion is given by:

$$MBIC=-2\ln m(x)+2(\nu+1),$$

where the marginal likelihood $m(x)$ is computed by the Laplace-Metropolis estimation and evaluated at either posterior mean or mode.[41] The MBIC criterion is an in-sample information criterion showing overall fit of the models penalizing for additional parameters.

For the choice of number of regimes we find the smallest MBIC. In addition we perform sensitivity analysis where estimation of thresholds r_j is done with and without restriction $r^{low(j)}<r_j<r^{up(j)}$, where upper and lower bounds are determined by using the minimum percentage of observation for each regime. We look at the sensitivity of posterior densities of r_j to imposing the minimum 10%, 5% of observations and no restriction. Typically we find two regimes with at least 10% of observations in each regime.[42]

To test whether the dynamics changes with regime one can simulate posterior distributions of differences in parameters of interest in upper and lower regimes. If there is considerable difference in some parameter's distributions it supports the hypothesis of non-linearity of series y_t.

Acknowledgments: The authors would like to thank the editor, anonymous referee, econometrics seminar participants at Rutgers University and the Seminar on Bayesian Inference in Econometrics and Statistics (SBIES) for their insightful comments. We thank Dr. Evgeny Goldman for his comments on the comparison of the mean squared errors distributions and Gayatri Jaganath for her excellent research assistance.

References

Ait-Sahalia, and Mancini. 2008. "Out of Sample Forecasts of Quadratic Variation." *Journal of Econometrics* 18: 17–33.
Ait-Sahalia, Y., P. Mykland, and L. Zhang. 2005. "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise." *Review of Financial Studies* 18: 351–416.
Andersen, T. G., and T. Bollerslev. 1997. "Intraday Periodicity and Volatility Persistence in Financial Markets." *Journal of Empirical Finance* 4: 115–158.
Andersen, T. G., and T. Bollerslev. 1998. "Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies." *The Journal of Finance* 53: 219–265.
Andersen, T. G., T. Bollerslev, and F. X. Diebold. 2007. "Roughing it Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility." *Review of Economics and Statistics* 89 (4): 701–720.
Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys. 2001a. "The Distribution of Exchange Rate Volatility." *Journal of the American Statistical Association* 96: 42–55.
Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens. 2001b. "The Distribution of Realized Stock Volatility." *Journal of Financial Economics* 61: 43–76.
Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys. 2003. "Modeling and Forecasting Realized Volatility." *Econometrica* 71: 579–625.
Ang, A., and G. Bekaert. 2002. "International Asset Allocation with Regime Shifts." *Review of Financial Studies* 15: 1137–1187.
Ang, A., and G. Bekaert. 2004. "How do Regimes Affect Asset Allocation?" *Financial Analysts Journal* 60 (2): 86–99.

41 Alternatively one can use Chib and Jeliazkov (2001) estimator of marginal likelihood.
42 We estimated models with two and three regimes, but using both in-sample and out-of-sample criteria all results were in favor of two regimes. Therefore, we present only results with two regimes in the paper.

Baillie, R. T. 1996. “Long Memory Processes and Fractional Integration in Econometrics.” *Journal of Econometrics* 73: 5–59.
Bandi, F. M., J. R. Russell, and C. Yang. 2008. “Realized Volatility Forecasting and Option Pricing.” *Journal of Econometrics* 147: 34–46.
Barndorff-Nielsen, O., and N. Shephard. 2004. “Power and Bi-Power Variation with Stochastic Volatility and Jumps.” *Journal of Financial Econometrics* 2 (2): 1–37.
Bollerslev, T., T. H. Law, and G. Tauchen. 2008. “Risk, Jumps, and Diversification.” *Journal of Econometrics* 144: 234–256.
Brockwell, P. J., and R. A. Davis. 1987. *Time Series: Theory and Methods,* 464–465. New York: Springer-Verlag.
Chib, S., and E. Greenberg. 1994. “Bayes Inference in Regression Models with ARMA(p,q) errors.” *Journal of Econometrics* 64: 183–206.
Chib, S., and E. Greenberg. 1995. “Understanding the Metropolis-Hastings Algorithm.” *The American Statistician* 49 (4): 327–335.
Chib, S., and I. Jeliazkov. 2001. “Marginal Likelihood from the Metropolis-Hastings Output.” *Journal of the American Statistical Association* 96: 270–281.
Chung, C. F. 2001. “Calculating and Analyzing Impulse Responses for the Vector ARFIMA Model.” *Economics Letters* 71: 17–25.
Corsi, F. 2009. “A Simple Approximate Long-memory Model of Realized Volatility.” *Journal of Financial Econometrics* 7: 174–196.
Diebold, F. X., and R. Mariano. 1995. “Comparing Predictive Accuracy.” *Journal of Business and Economic Statistics* 13: 253–263.
Diebold, F. X., and A. Inoue. 2001. “Long Memory and Regime Switching.” *Journal of Econometrics* 105: 131–159.
Diebold, F. X., S. Husted, and G. D. Rudebusch. 1991. “Real Exchange Rates under the Gold Standard.” *Journal of Political Economy* 99: 1252–1271.
Dufrenot, G., D. Guegan, and A. Peguin-Feissolle. 2008.“Changing-Regime Volatility: A Fractionally Integrated SETAR Model.” *Applied Financial Economics* 18: 519–526.
Fleming, J., C. Kirby, and B. Ostdiek. 2003. “The Economic Value of Volatility Timing using ‘Realized’ Volatility.” *Journal of Financial Economics* 67 (3): 473–509.
Ferguson, T. S. 1973. “A Bayesian Analysis of Some Nonparametric Problems.” *Annals of Statistics* 2: 615–629.
Forbes, C. S., G. Kalb, and P. Kofman. 1999. “Bayesian Arbitrage Threshold Analysis.” *Journal of Business and Economic Statistics, American Statistical Association* 17 (3): 364–372.
Geweke, J., and S. Porter-Hudak. 1983. “The Estimation and Application of Long Memory Time Series Models.” *Journal of Time Series Analysis* 4: 221–238.
Gil-Alana, L. A. 2001. “Estimation of Fractionally ARIMA Models for the UK Unemployment.” *Annales D’Economice et de Statistique* 62: 128–137.
Goldman, E., and H. Tsurumi. 2005. “Bayesian Analysis of a Doubly Truncated ARMA-GARCH Model.” *Studies in Nonlinear Dynamics and Econometrics* 9 (2): Article 5.
Goldman, E., and T. Agbeyegbe. 2007. “Estimation of Threshold Time Series Models using Efficient Jump MCMC.” In *Bayesian Statistics and its Applications*, edited by S. K. Upadhyay, U. Singh, and Dipak Dey, 241–253. New Delhi: Anamaya Publishers.
Goldman, E., S. Radchenko, T. Nakatsuma, and H. Tsurumi. 2001. “A Bayesian Test of Stationarity in a Regression Model with an ARMA Error Term.” *Proceedings of the Annual Meeting of the American Statistical Association*, August 5–9.
Gouskova, E. 2002. “ARFIMA Modeling of Financial Assets Volatility.” Ph.D. dissertation, Rutgers University.
Guidolin, M., and A. Timmermann. 2002. “Optimal Portfolio Choice under Regime Switching, Skew and Kurtosis Preferences.” Working Paper, UC San Diego.
Hafner, R., and M. Wallmeier. 2008. “Optimal Investments in Volatility.” *Financial Markets and Portfolio Management* 22: 147–167.
Hansen, B. E. 1997. “Inference in TAR Models.” *Studies in Nonlinear Dynamics and Econometrics* 2 (1): 1–14.
Hauser, M. A. 1999. “Maximum Likelihood Estimation for ARFIMA Models: A Monte Carlo Study.” *Journal of Statistical Planning and Inference* 80 (1): 229–255.
Huang, X., and G. Tauchen. 2005. “The Relative Contributions of Jumps to Total Variance.” *Journal of Financial Econometrics* 3: 456–499.
Koop, G., and S. M. Potter. 1999. “Dynamic Asymmetries in U.S. Unemployment.” *Journal of Business and Economic Statistics* 17: 298–312.
Koop, G., E. Ley, J. Osiewalski, and M. F. J. Steel. 1997. “Bayesian Analysis of Long Memory and Persistence using ARFIMA Models.” *Journal of Econometrics* 76: 149–169.
Lee, S. S., and P. A. Mykland. 2008. “Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics.” *The Review of Financial Studies* 21: 2535–2562.
Liu, C., and J. M. Maheu. 2008. “Are There Structural Breaks in Realized Volatility?” *Journal of Financial Econometrics* 6: 326–360.
Maheu, J. M., and T. H. McCurdy. 2002. “Non-Linear Features of Realized FX Volatility.” *Review of Economics and Statistics* 84: 668–681.
McAleer, M., and M. C. Medeiros. 2008. “A Multiple Regime Smooth Transition Heterogeneous Autoregressive Model for Long Memory and Asymmetries.” *Journal of Econometrics* 147: 104–119.
Phann, G. A., P. C. Schotman, and R. Tschering. 1996. “Non-Linear Interest Rate Dynamics and Implications for the Term Structure. *Journal of Econometrics* 74: 149–176.
Ramchand, L., and R. Susmel. 1998. “Volatility and Cross Correlations Across Major International Markets.” *Journal of Empirical Finance* 5: 397–416.

Shimizu and Tsurumi. 2009. “Comparison of Several Likelihood Functions used in Fractionally Integrated ARFIMA Models.” Working Paper, Rutgers University.

Sowell, F. 1992. “Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models.” *Journal of Econometrics* 53: 165–188.

Thomakos, D. D., and T. Wang. 2003. “Realized Volatility in the Futures Markets.” *Journal of Empirical Finance* 10: 321–353.

Tsay, R. S. 1989. “Testing and Modeling Threshold Autoregressive Processes.” *Journal of the American Statistical Association* 84: 231–240.

Tsay, R. S. 1998. “Testing and Modeling Multivariate Threshold Models.” *Journal of the American Statistical Association* 93: 1188–1202.