Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting

First version: November 26, 2014
This version: March 10, 2015

Tim Bollerslev[a], Andrew J. Patton[b], Rogier Quaedvlieg[c,]

[a]*Department of Economics, Duke University, NBER and CREATES*
[b]*Department of Economics, Duke University*
[c]*Department of Finance, Maastricht University*

Abstract

We propose a new family of easy-to-implement realized volatility based forecasting models. The models exploit the asymptotic theory for high-frequency realized volatility estimation to improve the accuracy of the forecasts. By allowing the parameters of the models to vary explicitly with the (estimated) degree of measurement error, the models exhibit stronger persistence, and in turn generate more responsive forecasts, when the measurement error is relatively low. Implementing the new class of models for the S&P500 equity index and the individual constituents of the Dow Jones Industrial Average, we document significant improvements in the accuracy of the resulting forecasts compared to the forecasts from some of the most popular existing models that implicitly ignore the temporal variation in the magnitude of the realized volatility measurement errors.

Keywords: Realized volatility; Forecasting; Measurement Errors; HAR; HARQ.
JEL: C22, C51, C53, C58

1. Introduction

Volatility, and volatility forecasting in particular, plays a crucial role in asset pricing and risk management. Access to accurate volatility forecasts is of the utmost importance for

☆Bollerslev gratefully acknowledges the support from an NSF grant to the NBER and CREATES funded by the Danish National Research Foundation (DNRF78). Quaedvlieg's research was financially supported by a grant of the Dutch Organization for Scientific Research (NWO). We would like to thank Nour Meddahi, as well as seminar participants at the Duke Financial Econometrics Lunch Group and Maastricht University for their helpful comments. We would also like to thank Bingzhi Zhao for providing us with the data.

many financial market practitioners and regulators. A long list of competing GARCH and stochastic volatility type formulations have been proposed in the literature for estimating and forecasting financial market volatility. The latent nature of volatility invariably complicates implementation of these models. The specific parametric models hitherto proposed in the literature generally also do not perform well when estimated directly with intraday data, which is now readily available for many financial assets. To help circumvent these complications and more effectively exploit the information inherent in high-frequency data, Andersen, Bollerslev, Diebold, and Labys (2003) suggested the use of reduced form time series forecasting models for the daily so-called realized volatilities constructed from the summation of the squared high-frequency intraday returns.[1]

Set against this background, we propose a new family of easy-to-implement volatility forecasting models. The models directly exploit the asymptotic theory for high-frequency realized volatility estimation by explicitly allowing the dynamic parameters of the models, and in turn the forecasts constructed from the models, to vary with the degree of estimation error in the realized volatility measures.

The realized volatility for most financial assets is a highly persistent process. Andersen, Bollerslev, Diebold, and Labys (2003) originally suggested the use of fractionally integrated ARFIMA models for characterizing this strong dependency. However, the simple and easy-to-estimate approximate long-memory HAR (Heterogeneous AR) model of Corsi (2009) has arguably emerged as the preferred specification for realized volatility based forecasting. Empirically, the volatility forecasts constructed from the HAR model, and other related reduced-form time series models that treat the realized volatility as directly observable, generally perform much better than the forecasts from traditional parametric GARCH and stochastic volatility models.[2]

Under certain conditions, realized volatility (RV) is consistent (as the sampling frequency goes to zero) for the true latent volatility, however in any given finite sample it is, of course, subject to measurement error. As such, RV will be equal to the sum of two components: the true latent Integrated Volatility (IV) and a measurement error. The dynamic modeling

[1]The use of realized volatility for accurately measuring the true latent integrated volatility was originally proposed by Andersen and Bollerslev (1998), and this approach has now become very popular for both measuring, modeling and forecasting volatility; see, e.g., the discussion and many references in the recent survey by Andersen, Bollerslev, Christoffersen, and Diebold (2013).

[2]Andersen, Bollerslev, and Meddahi (2004) and Sizova (2011) show how minor model misspecification can adversely affect the forecasts from tightly parameterized volatility models, thus providing a theoretical explanation for this superior reduced-form forecast performance.

of RV for the purposes of forecasting the true latent IV therefore suffers from a classical errors-in-variables problem. In most situations this leads to what is known as an attenuation bias, with the directly observable RV process being less persistent than the latent IV process. The degree to which this occurs obviously depends on the magnitude of the measurement errors; the greater the variance of the errors, the less persistent the observed process.[3]

Standard approaches for dealing with errors-in-variables problems treat the variance of the measurement error as constant through time.[4] In contrast, we explicitly take into account the temporal variation in the errors when modeling the realized volatility, building on the asymptotic distribution theory for the realized volatility measure developed by Barndorff-Nielsen and Shephard (2002). Intuitively, on days when the variance of the measurement error is small, the daily RV provides a stronger signal for next day's volatility than on days when the variance is large (with the opposite holding when the measurement error is large). Our new family of models exploits this heteroskedasticity in the error, by allowing for time-varying autoregressive parameters that are high when the variance of the realized volatility error is low, and adjusted downward on days when the variance is high and the signal is weak. Our adjustments are straightforward to implement and can easily be tailored to any autoregressive specification for RV. For concreteness, however, we focus our main discussion on the adaptation to the popular HAR model, which we dub the HARQ model. But, in our empirical investigation we also consider a number of other specifications and variations of the basic HARQ model.

Our empirical analysis relies on high-frequency data from 1997-2013 and corresponding realized volatility measures for the S&P 500 index and the individual constituents of Dow Jones Industrial Average. By explicitly incorporating the time-varying variance of the measurement errors into the parameterization of the model, the estimated HARQ models exhibit more persistence in “normal times” and quicker mean reversion in “erratic times” compared

[3]Alternative realized volatility estimators have been developed byBarndorff-Nielsen, Hansen, Lunde, and Shephard (2008); Zhang, Mykland, and Aït-Sahalia (2005); Jacod, Li, Mykland, Podolskij, and Vetter (2009) among others. Forecasting in the presence of microstructure “noise” has also been studied by Aït-Sahalia and Mancini (2008); Andersen, Bollerslev, and Meddahi (2011); Ghysels and Sinko (2011); Bandi, Russell, and Yang (2013). The analysis below effectively abstracts from these complications, by considering a coarse five-minute sampling frequency and using simple RV. We consider some of these alternative estimators in Section 4.1 below.

[4]General results for the estimation of autoregressive processes with measurement error are discussed in Staudenmayer and Buonaccorsi (2005). Hansen and Lunde (2014) have also recently advocated the use of standard instrumental variable techniques for estimating the persistence of the latent IV process, with the resulting estimates being significantly more persistent than the estimates for the directly observable RV process.

to the standard HAR model with constant autoregressive parameters.[5] Applying the HARQ model in an extensive out-of-sample forecast comparison, we document significant improvements in the accuracy of the forecasts compared to the forecasts from a challenging set of commonly used benchmark models. Interestingly, the forecasts from the HARQ models are not just improved in times when the right-hand side RVs are very noisy, and thus contain little relevant information, but also during tranquil times, when the forecasts benefit from the higher persistence afforded by the new models. Consistent with the basic intuition, the HARQ type models also offer the largest gains over the standard models for the assets for which the temporal variation in the magnitudes of the measurement errors are the highest.

The existing literature related to the dynamic modeling of RV and RV-based forecasting has largely ignored the issue of measurement errors, and when it has been considered, the errors have typically been treated as homoskedastic. Andersen, Bollerslev, and Meddahi (2011), for instance, advocate the use of ARMA models as a simple way to account for measurement errors, while Asai, McAleer, and Medeiros (2012) estimate a series of state-space models for the observable RV and the latent IV state variable with homoskedastic innovations. The approach for estimating stochastic volatility models based on realized volatility measures developed by Dobrev and Szerszen (2010) does incorporate the variance of the realized volatility error into the estimation of the models, but the parameters of the estimated models are assumed to be constant, and as such the dynamic dependencies and the forecasts from the models are not directly affected by the temporal variation in the size of the measurement errors. The motivation for the new family of HARQ models also bears some resemblance to the GMM estimation framework recently developed by Li and Xiu (2013). The idea of the paper is also related to the work of Bandi, Russell, and Yang (2013), who advocate the use of an "optimal," and possibly time-varying, sampling frequency when implementing RV measures, as a way to account for heteroskedasticity in the market microstructure "noise." In a similar vein, Shephard and Xiu (2014) interpret the magnitude of the parameter estimates associated with different RV measures in a GARCH-X model as indirect signals about the quality of the different measures: the lower the parameter estimate, the less smoothing, and the more accurate and informative the specific RV measure.

The rest of the paper is structured as follows. Section 2 provides the theoretical motivation

[5]The persistence of the estimated HARQ models at average values for the measurement errors is very similar to the unconditional estimates based on Hansen and Lunde (2014), and as such also much higher than the persistence of the standard HAR models. We discuss this further below.

for the new class of models, together with the results from a small scale simulation study designed to illustrate the workings of the models. Section 3 reports the results from an empirical application of the basic HARQ model for forecasting the volatility of the S&P 500 index and the individual constituents of the Dow Jones Industrial Average. Section 4 provides a series of robustness checks and extensions of the basic HARQ model. Section 5 concludes.

2. Realized Volatility-Based Forecasting and Measurement Errors

2.1. Realized Variance and High-Frequency Distribution Theory

To convey the main idea, consider a single asset for which the price process P_t is determined by the stochastic differential equation,

$$d\log(P_t) = \mu_t dt + \sigma_t dW_t, \tag{1}$$

where μ_t and σ_t denote the drift and the instantaneous volatility processes, respectively, and W_t is a standard Brownian motion assumed to be independent of σ_t. For simplicity and ease of notation, we do not include jumps in this discussion, but the main idea readily extends to discontinuous price processes, and we investigate this in Section 4 below. Following the vast realized volatility literature, our aim is to forecast the latent Integrated Variance (IV) over daily and longer horizons. Specifically, normalizing the unit time interval to a day, the one-day integrated variance is formally defined by,

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds. \tag{2}$$

The integrated variance is not directly observable. However, the Realized Variance (RV) defined by the summation of high-frequency returns,

$$RV_t \equiv \sum_{i=1}^{M} r_{t,i}^2, \tag{3}$$

where $M = 1/\Delta$, and the Δ-period intraday return is defined by $r_{t,i} \equiv log(P_{t-1+i\Delta}) - log(P_{t-1+(i-1)\Delta})$, provides a consistent estimator as the number of intraday observations increases, or equivalently $\Delta \to 0$ (see, e.g., Andersen and Bollerslev, 1998).

In practice, data limitations invariably put an upper bound on the value of M. The resulting estimation error in RV may be characterized by the asymptotic (for $\Delta \to 0$) distribution theory of Barndorff-Nielsen and Shephard (2002),

$$RV_t = IV_t + \eta_t, \qquad \eta_t \sim N(0, 2\Delta IQ_t), \tag{4}$$

where $IQ_t \equiv \int_{t-1}^{t} \sigma_s^4 ds$ denotes the Integrated Quarticity (IQ). In parallel, to the integrated variance, the integrated quarticity may be consistently estimated by the Realized Quarticity (RQ),

$$RQ_t \equiv \frac{M}{3} \sum_{i=1}^{M} r_{t,i}^4. \tag{5}$$

2.2. The ARQ model

The consistency of RV for IV, coupled with the fact that the measurement error is serially uncorrelated under general conditions, motivate the use of reduced form time series models for the observable realized volatility as a simple way to forecast the latent integrated volatility of interest.[6]

To illustrate, suppose that the dynamic dependencies in IV may be described by an AR(1) model,

$$IV_t = \phi_0 + \phi_1 IV_{t-1} + u_t. \tag{6}$$

If u_t and the measurement error η_t are both *i.i.d.*, with variances σ_u^2 and σ_η^2, then it follows by standard arguments that RV follows an ARMA(1,1) model, with AR-parameter equal to ϕ_1 and (invertible) MA-parameter equal to,

$$\theta_1 = \frac{-\sigma_u^2 - \left(1+\phi_1^2\right)\sigma_\eta^2 + \sqrt{\sigma_u^4 + 2\left(1+\phi_1^2\right)\sigma_u^2\sigma_\eta^2 + \left(1-\phi_1^2\right)^2\sigma_\eta^4}}{\phi_1 \sigma_\eta^2}, \text{ for } \sigma_\eta^2 > 0 \text{ and } \phi_1 \neq 0. \tag{7}$$

It is possible to show that θ is increasing (in absolute value) in the variance of the measurement error, σ_η^2, and that $\theta_1 \to 0$ as $\sigma_\eta^2 \to 0$ or $\phi_1 \to 0$.

Now suppose that instead of an ARMA(1,1) model, the researcher estimates an approximate and easy-to-implement AR(1) model for RV,

$$IV_t + \eta_t = \beta_0 + \beta_1(IV_{t-1} + \eta_{t-1}) + u_t. \tag{8}$$

[6] A formal theoretical justification for this approach is provided by Andersen, Bollerslev, Diebold, and Labys (2003). Further, as shown by Andersen, Bollerslev, and Meddahi (2004), for some of the most popular stochastic volatility models used in the literature, simple autoregressive models for RV provide close to efficient forecasts for IV.

The measurement error on the left hand side is unpredictable, and merely results in an increase in the standard errors of the parameter estimates. The measurement error on the right hand side, however, directly affects the parameter β_1, and in turn propagates into the forecasts from the model. Again, assuming for simplicity that u_t and η_t are both *i.i.d.*, so that $Cov(RV_t, RV_{t-1}) = \phi_1 Var(IV_t)$ and $Var(RV_t) = Var(IV_t) + \sigma^2_\eta$, the population value for β_1 may be expressed as,

$$\beta_1 = \phi_1 \left(1 + \frac{\sigma^2_\eta}{Var(IV_t)}\right)^{-1}. \tag{9}$$

The estimated autoregressive coefficient for RV will therefore be smaller than the ϕ_1 coefficient for IV.[7] This discrepancy between β_1 and ϕ_1 is directly attributable to the well-known attenuation bias arising from the presence of measurement errors. The degree to which β_1 is attenuated is a direct function of the measurement error variance: if $\sigma^2_\eta = 0$, then $\beta_1 = \phi_1$, but if σ^2_η is large, then β_1 goes to zero and RV is effectively unpredictable.[8]

The standard expression for β_1 in equation (9) is based on the assumption that the variance of the measurement error is constant. However, from equation (4) the variance pertaining to the estimation error in RV generally changes through time: there are days when IQ is low and RV provides a strong signal about the true IV, and days when IQ is high and the signal is relatively weak. The OLS-based estimate of β_1 will effectively be attenuated by the *average* of this measurement error variance. As such, the assumption of a constant AR parameter is suboptimal from a forecasting perspective. Instead, by explicitly allowing for a time-varying autoregressive parameter, say $\beta_{1,t}$, this parameter should be close to ϕ_1 on days when there is little measurement error, while on days where the measurement error variance is high, $\beta_{1,t}$ should be low and the model quickly mean reverting.[9]

The AR(1) representation for the latent integrated volatility in equation (6) that under-

[7]The R^2 from the regression based on RV will similarly be downward biased compared to the R^2 from the infeasible regression based on the latent IV. Andersen, Bollerslev, and Meddahi (2005) provide a simple adjustment for this unconditional bias in the R^2.

[8]As previously noted, Hansen and Lunde (2014) propose the use of an instrumental variable procedure for dealing with this attenuation bias and obtain a consistent estimator of the autoregressive parameters of the true latent IV process. For forecasting purposes, it is β_1 and not ϕ_1 that is the parameter of interest, as the explanatory variable is the noisy realized variance, not the true integrated variance.

[9]These same arguments carry over to the θ_1 parameter for the ARMA(1,1) model in equation (7) and the implied persistence as a function of the measurement error variance. Correspondingly, in the GARCH(1,1) model, in the usual notation of that model, the autoregressive parameter given by $\alpha + \beta$ should be constant, while the values of α and β should change over time so that β is larger (smaller) and α is smaller (larger) resulting in more (less) smoothing when the variance of the squared return is high (low).

lies these ideas merely serves as an illustration. Hence, rather than relying directly on the expression in equation (9) involving the inverse of the measurement error variance, in our practical implementation we use a more flexible and robust specification in which we allow the time-varying AR parameter to depend linearly on an estimate of $IQ^{1/2}$. We term this specification the ARQ model for short,[10]

$$RV_t = \beta_0 + \underbrace{(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2})}_{\beta_{1,t}} RV_{t-1} + u_t. \tag{10}$$

For ease of interpretation, we suggest demeaning $RQ^{1/2}$ so that the estimate of β_1 has the interpretation of the average autoregressive coefficient, directly comparable to β_1 in equation (8). The simple specification above has the advantage that it can easily be estimated by standard OLS, so both estimation and forecasting is straightforward and fast. Importantly, the value of the autoregressive $\beta_{1,t}$ parameter will vary with the estimated measurement error variance. (In our additional empirical investigations reported in Section 4.3 below, we also consider models that allow the intercept, or β_0, to vary with RQ.) In particular, assuming that $\beta_{1Q} < 0$ it follows that uninformative days with large measurement errors will have smaller impact on the forecasts than days where RV is estimated precisely and $\beta_{1,t}$ is larger. If RQ is constant over time, the ARQ model reduces to a standard AR(1) model. Thus, the greater the temporal variation in the measurement error variance, the greater the expected benefit of modeling and forecasting the volatility with the ARQ model, a prediction we confirm in our empirical analysis below.

2.3. The HARQ model

The AR(1) model in equation (8) is too simplistic to satisfactorily describe the long-run dependencies in most realized volatility series. Instead, the Heterogeneous Autoregression (HAR) model of Corsi (2009) has arguably emerged as the most popular model for daily realized volatility based forecasting,

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t, \tag{11}$$

[10] IQ is notoriously difficult to estimate in finite samples, and its inverse even more so. The use of the square-root as opposed to the inverse of RQ imbues the formulation with a degree of built-in robustness. However, we also consider a variety of other estimators and transformations of IQ in the robustness section below.

where $RV_{t-j|t-h} = \frac{1}{h}\sum_{i=j}^{h} RV_{t-i}$. The choice of a daily, weekly and monthly lag on the right-hand-side conveniently captures the approximate long-memory dynamic dependencies observed in most realized volatility series.

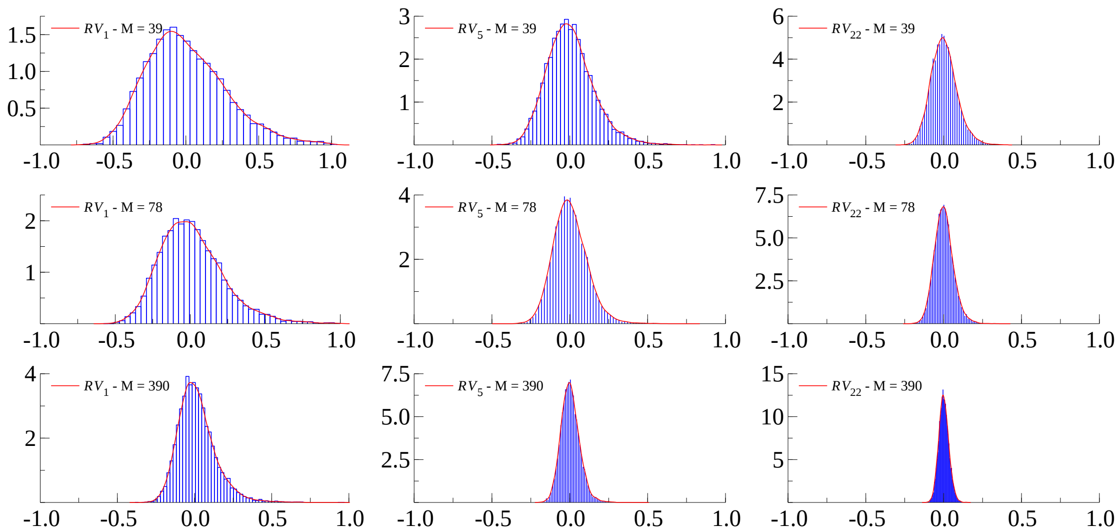
Of course, just like the simple AR(1) model discussed in the previous section, the beta coefficients in the HAR model are affected by measurement errors in the realized volatilities. In parallel to the ARQ model, this naturally suggests the following extension of the basic HAR model that directly adjust the coefficients in proportion to the magnitude of the corresponding measurement errors,

$$\begin{aligned} RV_t = \beta_0 + \underbrace{(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2})}_{\beta_{1,t}} RV_{t-1} + \underbrace{(\beta_2 + \beta_{2Q} RQ_{t-1|t-5}^{1/2})}_{\beta_{2,t}} RV_{t-1|t-5} \\ + \underbrace{(\beta_3 + \beta_{3Q} RQ_{t-1|t-22}^{1/2})}_{\beta_{3,t}} RV_{t-1|t-22} + u_t, \end{aligned} \tag{12}$$

where $RQ_{t-1|t-k} = \frac{1}{k}\sum_{j=1}^{k} RQ_{t-j}$. Of course, the magnitude of the (normalized) measurement errors in the realized volatilities will generally decrease with the horizon k as the errors are averaged out, indirectly suggesting that adjusting for the measurement errors in the daily lagged realized volatilities is likely to prove more important than the adjustments for the weekly and monthly coefficients. Intuitively, this also means that in the estimation of the standard HAR model some of the weight will be shifted away from the noisy daily lag to the "cleaner," though older, weekly and monthly lags that are less prone to measurement errors.

To directly illustrate how the measurement errors manifest over different sampling frequencies and horizons, Figure 1 plots the simulated RV measurement errors based on ten, five, and one-"minute" sampling ($M = 39, 78, 390$) and horizons ranging from "daily," to "weekly," to "monthly" ($k = 1, 5, 22$); the exact setup of the simulations are discussed in more detail in Section 2.4 and Appendix A. To facilitate comparison across the different values of M and k, we plot the distribution of $RV/IV - 1$, so that a value of 0.5 may be interpreted as an estimate that is 50% higher than the true IV. Even with an observation every minute ($M = 390$), the estimation error in the daily ($k = 1$) simulated RV can still be quite substantial. The measurement error variance for the weekly and monthly (normalized) RV are, as expected, much smaller and approximately 1/5 and 1/22 that of the daily RV. Thus, the attenuation bias in the standard HAR model will be much less severe for the weekly and monthly coefficients.

Figure 1: Estimation Error of RV



Note: The figure shows the simulated distribution of $RV/IV - 1$. The top, middle and bottom panels show the results for $M = 39, 78$, and 390, respectively, while the left, middle and right panels show the results for daily, weekly, and monthly forecast horizons, respectively.

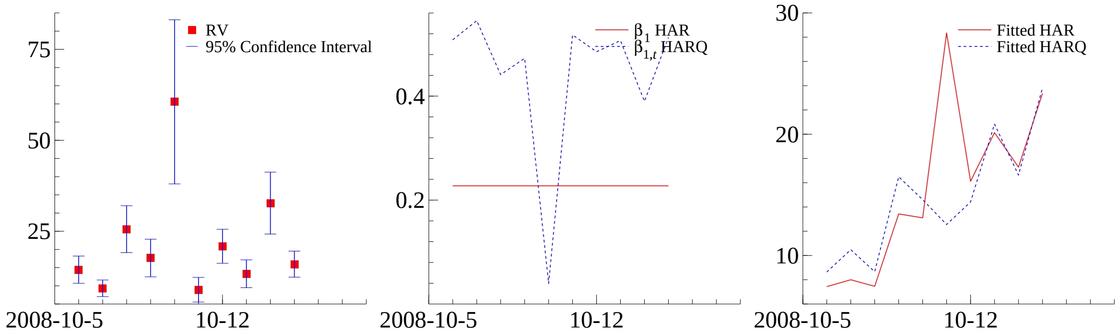
Motivated by these observations, coupled with the difficulties in precisely estimating the β_Q adjustment parameters, we will focus our main empirical investigations on the simplified version of the model in equation (12) that only allows the coefficient on the daily lagged RV to vary as a function of $RQ^{1/2}$,

$$RV_t = \beta_0 + \underbrace{(\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2})}_{\beta_{1,t}} RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t. \tag{13}$$

We will refer to this model as the HARQ model for short, and the model in equation (12) that allows all of the parameters to vary with an estimate of the measurement error variance as the "full HARQ" model, or HARQ-F.

To illustrate the intuition and inner workings of the HARQ model, Figure 2 plots the HAR and HARQ model estimates for the S&P 500 for ten consecutive trading days in October 2008; further details concerning the data are provided in the empirical section below. The left panel shows the estimated RV along with 95% confidence bands based in the asymptotic approximation in (4). One day in particular stands out: on Friday, October 10 the realized volatility was substantially higher than for any of the other ten days, and importantly, also

Figure 2: HAR vs. HARQ



Note: The figure illustrates the HARQ model for ten successive trading days. The left-panel shows the estimated RVs with 95% confidence bands based on the estimated RQ. The middle panel shows the $\beta_{1,t}$ estimates from the HARQ model, together with the estimate of β_1 from the standard HAR model. The right panel shows the resulting one-day-ahead RV forecasts from the HAR and HARQ models.

far less precisely estimated, as evidenced by the wider confidence bands.[11]. The middle panel shows the resulting β_1 and $\beta_{1,t}$ parameter estimates. The level of $\beta_{1,t}$ from the HARQ model is around 0.5 on "normal" days, more than double that of β_1 of just slightly above 0.2 from the standard HAR model. However, on the days when RV is estimated imprecisely, $\beta_{1,t}$ can be much lower, as illustrated by the precipitously drop to less than 0.1 on October 10, as well as the smaller drop on October 16. The rightmost panel shows the effect that this temporal variation in $\beta_{1,t}$ has on the one-day-ahead forecasts from the HARQ model. In contrast to the HAR model, where the high RV on October 10 leads to an increase in the fitted value for the next day, the HARQ model actually forecasts a *lower* value than the day before. Compared to the standard HAR model, the HARQ model allows for higher average persistence, together with forecasts closer to the unconditional volatility when the lagged RV is less informative.

2.4. Simulation Study

To further illustrate the workings of the HARQ model, this section presents the results from a small simulation study. We begin by demonstrating non-trivial improvements in the in-sample fits from the ARQ and HARQ models compared to the standard AR and HAR models. We then then show how these improved in-sample fits translates into superior out-of-

[11]October 10 was marked by a steep loss in the first few minutes of trading followed by a rise into positive territory and a subsequent decline, with all of the major indexes closing down just slightly for the day, including the S&P 500 which fell by 1.2%.

sample forecasts. Finally, we demonstrate how these improvements may be attributed to the increased average persistence of the estimated ARQ and HARQ models obtained by shifting the weights of the lags to more recent observations.

Our simulations are based on the two-factor stochastic volatility diffusion with noise previously analyzed by Huang and Tauchen (2005), Gonçalves and Meddahi (2009) and Patton (2011), among others. Details about the exact specification of the model and the parameter values used in the simulation are given in Appendix A. We report the results based on $M = 39, 78, 390$ "intraday" return observations, corresponding to ten, five, and one-"minute" sampling frequencies. We consider five different forecasting models: AR, HAR, ARQ, HARQ and HARQ-F. The AR and HAR models help gauge the magnitude of the improvements that may realistically be expected in practice. All of the models are estimated by OLS based on $T = 1,000$ simulated "daily" observations. Consistent with the OLS estimation of the models, we rely on a standard MSE measure to assess the in-sample fits,

$$MSE(RV_t, X_t) \equiv (RV_t - X_t)^2,$$

where X_t refers to the fit from any one of the different models. We also calculate one-day-ahead out-of-sample forecasts from all of the models. For the out-of-sample comparisons we consider both the $MSE(RV_t, F_t)$, and the QLIKE loss,

$$QLIKE(RV_t, F_t) \equiv \frac{RV_t}{F_t} - log\left(\frac{RV_t}{F_t}\right) - 1,$$

where F_t refers to the one-day-ahead forecasts from the different models.[12] To facilitate direct comparisons of the in- and out-of-sample results, we rely on a rolling window of 1,000 observations for the one-step-ahead forecasts and use these same 1,000 forecasted observations for the in-sample estimation. All of the reported simulation results are based on 1,000 replications.

Table 1 summarizes the key findings. To make the relative gains stand out more clearly, we standardize the relevant loss measures in each of the separate panels by the loss of the HAR model. As expected, the ARQ model systematically improves on the AR model, and the HARQ model similarly improves on the HAR model. This holds true both in- and out-

[12]Very similar out-of-sample results and rankings of the different models are obtained for the MSE and QLIKE defined relative to the true latent integrated volatility within the simulations; i.e., $MSE(IV_t, F_t)$ and $QLIKE(IV_t, F_t)$, respectively.

Table 1: Simulation Results

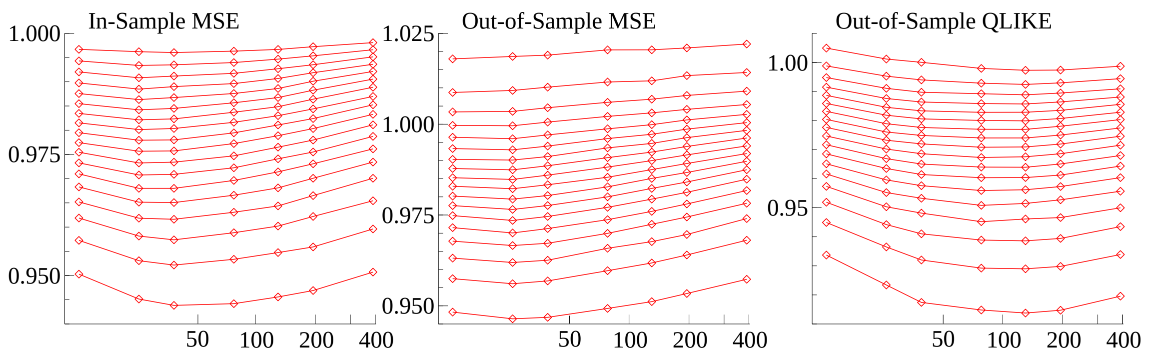
	AR	HAR	ARQ	HARQ	HARQ-F
M			In-Sample MSE		
39	1.0291	1.0000	0.9980	0.9773	0.9718
78	1.0285	1.0000	0.9996	0.9791	0.9735
390	1.0277	1.0000	1.0064	0.9851	0.9793
			Out-of-Sample MSE		
39	1.0438	1.0000	1.0166	0.9878	0.9900
78	1.0425	1.0000	1.0188	0.9901	0.9920
390	1.0413	1.0000	1.0268	0.9968	0.9985
			Out-of-Sample QLIKE		
39	1.0893	1.0000	1.0258	0.9680	0.9850
78	1.0881	1.0000	1.0186	0.9644	0.9821
390	1.0841	1.0000	1.0187	0.9678	0.9859
			Persistence		
39	0.4303	0.6593	0.6552	0.8132	0.9200
78	0.4568	0.6736	0.6876	0.8328	0.9449
390	0.4739	0.6803	0.6913	0.8297	0.9621
			Mean Lag		
39		5.6598		4.2410	4.6956
78		5.4963		4.1026	4.5968
390		5.3685		4.1196	4.6530

Note: The table reports the MSE and QLIKE losses for the different approximate models. The average losses are standardized by the loss of the HAR model. The bottom panel reports the average estimated persistence and mean lag across the different models. The two-factor stochastic volatility model and the exact design underlying the simulations are further described in the Appendix.

of-sample. The difficulties in accurately estimating the additional adjustment parameters for the weekly and monthly lags in the HARQ-F model manifest in this model generally not performing as well out-of-sample as the simpler HARQ model that only adjusts for the measurement error in the daily realized volatility. Also, the improvements afforded by the (H)ARQ models are decreasing in M, as more high-frequency observations help reduce the magnitude of the measurement errors, and thus reduce the gains from exploiting them.

Figure 3 further highlights this point. The figure plots the simulated quantiles of the ratio distribution of the HARQ to HAR models for different values of M. Each line represents one quantile, ranging from 5% to 95% in 5% increments. For all criteria, in- and out-of-sample MSE and out-of-sample QLIKE, the loss ratio shows a U-shaped pattern, with the gains of the HARQ model relative to the standard HAR model maximized somewhere between 2- and 10-"minute" sampling. When M is very large, the measurement error decreases and the gains

Figure 3: Distribution of HARQ/HAR ratio



Note: The figures depicts the quantiles ranging from 0.05 to 0.95 in increments of 0.05 for the simulated MSE and QLIKE loss ratios for the HARQ model relative to the standard HAR model. The horizontal axis shows the number of observations used to estimate RV, ranging from 13 to 390 per "day."

from using information on the magnitude of the error diminishes. When M is very small, estimating the measurement error variance by RQ becomes increasingly more difficult and the adjustments in turn less accurate. As such, the adjustments that motivate the HARQ model are likely to work best in environments where there is non-negligible measurement error in RV, and the estimation of this measurement error via RQ is at least somewhat reliable. Whether this holds in practice is an empirical question, and one that we study in great detail in the next section.

The second half of Table 1 reports the persistence for all of the models defined by the estimates of $\beta_1+\beta_2+\beta_3$, as well as the mean lags for the HAR, HARQ and HARQ-F models. In the HAR models, the weight on the first lag equals $b_1 = \beta_1 + \beta_2/5 + \beta_3/22$, on the second lag $b_2 = \beta_2/5 + \beta_3/22$, on the sixth lag $b_6 = \beta_3/22$, and so forth, so that these mean lags are easily computed as $22\sum_{i=1}^{22} ib_i/\sum_{i=1}^{22} b_i$. For the HARQ models, this corresponds to the mean lag at the *average* measurement error variance. The mean lag gives an indication of the location of the lag weights. The lower the mean lag, the greater the weight on more recent RVs.

The results confirm that at the mean measurement error variance, the HARQ model is far more persistent than the standard HAR model. As M increases, and the measurement error decreases, the gap between the models narrows. However, the persistence of the HARQ model is systematically higher, and importantly, much more stable across the different values of M. As M increases and the measurement error decreases, the loading on RQ diminishes, but this changes little in terms of the persistence of the underlying latent process that is being approximated by the HARQ model.[13] The result pertaining to the mean lags reported in the bottom panel further corroborates the idea that on average, the HARQ model assigns more weight to more recent RVs than the does the standard HAR model.

3. Modeling and Forecasting Equity Return Volatility

3.1. Data

We focus our empirical investigations on the S&P 500 aggregate market index. High-frequency futures prices for the index are obtained from Tick Data Inc. We complement our analysis of the aggregate market with additional results for the 27 Dow Jones Constituents as of September 20, 2013 that traded continuously from the start to the end of our sample. Data on these individual stocks comes from the TAQ database. Our sample starts on April 21, 1997, one thousand trading days (the length of our estimation window) before the final decimalization of NASDAQ on April 9, 2001. The sample for the S&P 500 ends on August 30, 2013, while the sample for the individual stocks ends on December 31, 2013, yielding a total of 3,096 observations for the S&P 500 and 3,202 observations for the DJIA constituents. The first 1,000 days are only used to estimate the models, so that the in-sample estimation results and the rolling out-of-sample forecasts are all based on the same samples.

Table 2 provides a standard set of summary statistics for the daily realized volatilities. Following common practice in the literature, all of the RVs are based on five-minute returns.[14] In addition to the usual summary measures, we also report the first order autocorrelation corresponding to β_1 in equation (8), the instrumental variable estimator of Hansen and Lunde (2014) denoted AR-HL, and the estimate of β_1 from the ARQ model in equation (10) corresponding to the autoregressive parameter at the average measurement error variance. The AR-HL estimates are all much larger than the standard AR estimates, directly highlighting the importance of measurement errors. By exploiting the heteroskedasticity in the measurement errors, the ARQ model allows for far greater persistence on average than the standard AR model, bridging most of the gap between the AR and AR-HL estimates.

[13] Interestingly, the HARQ-F model is even more persistent. This may be fully attributed to an increase in the monthly lag parameter, combined with a relatively high loading on the interaction of the monthly RV and RQ.

[14] Liu, Patton, and Sheppard (2015) provide a recent discussion and empirical justification for this common choice. In some of the additional results discussed below, we also consider other sampling frequencies and RV estimators. Our main empirical findings remain intact to these other choices.

Table 2: Summary Statistics

Company	Symbol	Min	Mean	Median	Max	AR	AR-HL	ARQ
S&P 500		0.043	1.175	0.629	60.563	0.651	0.953	0.983
Microsoft	MSFT	0.166	3.087	1.824	59.164	0.718	0.952	0.889
Coca-Cola	KO	0.049	2.011	1.154	54.883	0.618	0.949	0.834
DuPont	DD	0.093	3.327	2.165	81.721	0.707	0.950	0.956
ExxonMobil	XOM	0.114	2.348	1.476	130.667	0.668	0.947	0.997
General Electric	GE	0.131	3.440	1.794	173.223	0.681	0.915	0.987
IBM	IBM	0.115	2.464	1.340	72.789	0.657	0.959	0.890
Chevron	CVX	0.105	2.286	1.483	139.984	0.653	0.966	0.954
United Technologies	UTX	0.126	2.793	1.658	92.105	0.648	0.943	0.883
Procter & Gamble	PG	0.085	2.007	1.064	80.124	0.587	0.866	0.786
Caterpillar	CAT	0.207	3.810	2.401	127.119	0.727	0.954	0.896
Boeing	BA	0.167	3.371	2.147	79.760	0.630	0.936	0.822
Pfizer	PFE	0.176	2.822	1.809	60.302	0.570	0.933	0.837
Johnson & Johnson	JNJ	0.062	1.680	0.999	58.338	0.613	0.946	0.933
3M	MMM	0.140	2.278	1.358	123.197	0.495	0.952	0.748
Merck	MRK	0.127	2.758	1.718	223.723	0.372	0.913	0.708
Walt Disney	DIS	0.135	3.641	2.030	129.661	0.629	0.916	0.772
McDonald's	MCD	0.090	2.678	1.680	130.103	0.390	0.937	0.672
JPMorgan Chase	JPM	0.114	5.420	2.552	261.459	0.716	0.832	0.940
Wal-Mart	WMT	0.148	2.761	1.443	114.639	0.611	0.925	0.810
Nike	NKE	0.192	3.431	1.980	84.338	0.581	0.943	0.785
American Express	AXP	0.088	4.603	2.184	290.338	0.602	0.948	0.949
Intel	INTC	0.208	4.654	2.674	89.735	0.731	0.949	0.968
Travelers	TRV	0.098	3.579	1.637	273.579	0.646	0.933	0.915
Verizon	VZ	0.145	2.788	1.637	99.821	0.646	0.952	0.859
The Home Depot	HD	0.171	3.798	2.161	133.855	0.633	0.946	0.992
Cisco Systems	CSCO	0.234	5.120	2.742	96.212	0.715	0.939	0.942
UnitedHealth Group	UNH	0.222	4.145	2.304	169.815	0.616	0.920	0.846

Note: The table provides summary statistics for the daily RVs for each of the series. The column . labeled AR reports the standard first order autocorrelation coefficients, the column labeled AR-HL gives the instrumental variable estimator of Hansen and Lunde (2014), while β_1 refers to the corresponding estimates from the ARQ model in equation (10)

3.2. In-Sample Estimation Results

We begin by considering the full in-sample results. The top panel in Table 3 reports the parameter estimates for the S&P 500, with robust standard errors in parentheses, for the benchmark AR and HAR models, together with the ARQ, HARQ and HARQ-F models. For comparison purposes, we also include the AR-HL estimates, even though they were never intended to be used for forecasting purposes. The second and third panel report the R^2, MSE and QLIKE for the S&P500, and the average of those three statistics across the 27 DJIA individual stocks. Further details about the model parameter estimates for the individual stocks are available in Appendix A.

As expected, all of the β_{1Q} coefficients are negative and strongly statistically significant.

Table 3: In-Sample Estimation Results

	AR	HAR	AR-HL	ARQ	HARQ	HARQ-F
β_0	0.4109	0.1123		0.0892	-0.0098	-0.0187
	(0.1045)	(0.0615)		(0.0666)	(0.0617)	(0.0573)
β_1	0.6508	0.2273	0.9529	0.9830	0.6021	0.5725
	(0.1018)	(0.1104)	(0.0073)	(0.0782)	(0.0851)	(0.0775)
β_2		0.4903			0.3586	0.4368
		(0.1352)			(0.1284)	(0.1755)
β_3		0.1864			0.0976	0.0509
		(0.1100)			(0.1052)	(0.1447)
β_{1Q}				-0.5139	-0.3602	-0.3390
				(0.0708)	(0.0637)	(0.0730)
β_{2Q}						-0.1406
						(0.3301)
β_{3Q}						0.0856
						(0.3416)
R^2	0.4235	0.5224	0.3323	0.5263	0.5624	0.5628
MSE	3.1049	2.5722	3.5964	2.5512	2.3570	2.3546
QLIKE	0.2111	0.1438	0.1586	0.1530	0.1358	0.1380
$\overline{R}^2$ Stocks	0.3975	0.4852	0.2935	0.4676	0.5090	0.5139
$\overline{MSE}$ Stocks	17.4559	14.9845	20.0886	15.2782	14.1702	14.0154
$\overline{QLIKE}$ Stocks	0.2095	0.1496	0.1759	0.1804	0.1470	0.1547

Note: The table provides in-sample parameter estimates and measures of fit for the various benchmark and (H)ARQ models. The top two panels report the actual parameter estimates for the S&P500 with robust standard errors in parentheses, together with the R^2s, MSE and QLIKE losses from the regressions. The bottom panel summarizes the in-sample losses for the different models averaged across all of the individual stocks.

This is consistent with the simple intuition that as the measurement error and the current value of RQ increases, the informativeness of the current RV for future RVs decreases, and therefore the $\beta_{1,t}$ coefficient on the current RV decreases towards zero. Directly comparing the AR coefficient to the autoregressive parameter in the ARQ model also reveals a marked difference in the estimated persistence of the models. By failing to take into account the time-varying nature of the informativeness of the RV measures, the estimated AR coefficients are sharply attenuated.

The findings for the HARQ model are slightly more subtle. Comparing the HAR model with the HARQ model, the HAR places greater weight on the weekly and monthly lags, which are less prone to measurement errors than the daily lag, but also further in the past. These increased weights on the weekly and monthly lags hold true for the S&P500 index, and for every single individual stock in the sample. By taking into account the time-varying nature of the measurement error in the daily RV, the HARQ model assigns a greater average weight to the daily lag, while down-weighting the daily lag when the measurement error is large. The HARQ-F model parameters differ slightly from the HARQ model parameters, as the weekly and monthly lags are now also allowed to vary. However, the estimates for β_{2Q} and β_{3Q} are not statistically significant, and the improvement in the in-sample fit compared to the HARQ model is minimal.

To further corroborate the conjecture that the superior performance of the HARQ model is directly attributable to the measurement error adjustments, we also calculated the mean lags implied by the HAR and HARQ models estimated with less accurate realized volatilities based on coarser sampled 10- and 15-minute intraday returns. Consistent with the basic intuition of the measurement errors on average pushing the weights further in the past, the mean lags are systematically lower for the models that rely on the more finely sampled RVs. For instance, the average mean lag across all of the individual stocks for the HAR models equal 5.364, 5.262 and 5.003 for 15-, 10- and 5-minute RVs, respectively. As the measurement error decreases, the shorter lags become more accurate and informative for the predictions. By comparison, the average mean lag across all of the stocks for the HARQ models equal 4.063, 3.877 and 3.543 for 15-, 10- and 5-minute RVs, respectively. Thus, on average the HARQ models always assign more weight to the more recent RVs than the standard HAR models, and generally allow for a more rapid response, except, of course, when the signal is poor.

3.3. Out-of-Sample Forecast Results

Many other extensions of the standard HAR model have been proposed in the literature. To help assess the forecasting performance of the HARQ model more broadly, in addition to the basic AR and HAR models considered above, we therefore also consider the forecasts from three alternative popular HAR type formulations.

Specifically, following Andersen, Bollerslev, and Diebold (2007) we include both the the HAR-with-Jumps (HAR-J) and the Continuous-HAR (CHAR) models in our forecast comparisons. Both of these models rely on a decomposition of the total variation into a continuous and a discontinuous (jump) part. This decomposition is most commonly implemented using the Bi-Power Variation (BPV) measure of Barndorff-Nielsen and Shephard (2004b), which affords a consistent estimate of the continuous variation in the presence of jumps. The HAR-J model, in particular, includes a measure of the jump variation as an additional explanatory variable in the standard HAR model,

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \beta_J J_{t-1} + u_t, \tag{14}$$

where $J_t \equiv \max[RV_t - BPV_t, 0]$, and the BPV measure is defined as,

$$BPV_t \equiv \mu_1^{-2} \sum_{i=1}^{M-1} |r_{t,i}||r_{t+1,i}|, \tag{15}$$

with $\mu_1 = \sqrt{2/\pi} = \mathbb{E}(|Z|)$, and Z a standard normally distributed random variable. Empirically, the jump component has typically been found to be largely unpredictable. This motivates the alternative CHAR model, which only includes measures of the continuous variation on the right hand side,

$$RV_t = \beta_0 + \beta_1 BPV_{t-1} + \beta_2 BPV_{t-1|t-5} + \beta_3 BPV_{t-1|t-22} + u_t. \tag{16}$$

Several empirical studies have documented that the HAR-J and CHAR models often perform (slightly) better than the standard HAR model.

Meanwhile, Patton and Sheppard (2015) have recently argued that a Semivariance-HAR (SHAR) model sometimes performs even better than the HAR-J and CHAR models. Building on the semi-variation measures of Barndorff-Nielsen, Kinnebrock, and Shephard (2010), the SHAR model decomposes the total variation in the standard HAR model into the variation due to negative and positive intraday returns, respectively. In particular, let $RV_t^- \equiv \sum_{i=1}^{M} r_{t,i}^2 \mathbb{I}_{\{r_{t,i}<0\}}$ and $RV_t^+ \equiv \sum_{i=1}^{M} r_{t,i}^2 \mathbb{I}_{\{r_{t,i}>0\}}$, the SHAR model is then defined as:

$$RV_t = \beta_0 + \beta_1^+ RV_{t-1}^+ + \beta_1^- RV_{t-1}^- + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t. \tag{17}$$

Like the HARQ models, the HAR-J, CHAR and SHAR models are all easy to estimate and implement.

We focus our discussion on the one-day-ahead forecasts for the S&P500 index starting on April 9, 2001 through the end of the sample. However, we also present summary results for the 27 individual stocks, with additional details available in Appendix A. The forecast are based on re-estimating the parameters of the different models each day with a fixed length Rolling Window (RW) comprised of the previous 1,000 days, as well as an Increasing Window (IW) using all of the available observations. The sample sizes for the increasing window for the S&P500 thus range from 1,000 to 3,201 days.

The average MSE and QLIKE for the S&P500 index are reported in the top panel in

Table 4: Out-of-Sample Forecast Losses

			AR	HAR	HAR-J	CHAR	SHAR	ARQ	HARQ	HARQ-F
S&P 500										
MSE	RW		0.9166	1.0000	0.9176	0.9583	0.8375	**0.8115**	0.8266	0.9750
	IW		1.2315	1.0000	0.9676	0.9707	0.9012	0.9587	**0.8944**	0.9312
QLIKE	RW		1.4559	1.0000	1.0062	1.0124	**0.9375**	0.9570	0.9464	0.9934
	IW		1.7216	1.0000	0.9716	0.9829	0.8718	1.1845	0.8809	**0.8686**
Individual Stocks										
MSE	RW	Avg	1.1505	1.0000	1.0151	1.0080	1.0083	0.9659	**0.9349**	1.0149
		Med	1.1730	1.0000	1.0115	1.0158	1.0020	0.9864	**0.9418**	1.0263
	IW	Avg	1.2130	1.0000	1.0040	1.0013	0.9947	1.0371	**0.9525**	1.0071
		Med	1.2161	1.0000	1.0028	1.0010	0.9968	1.0396	**0.9525**	0.9660
QLIKE	RW	Avg	1.4204	1.0000	1.0018	0.9999	0.9902	1.1498	**0.9902**	1.1516
		Med	1.4044	1.0000	0.9976	1.0025	0.9941	1.1781	**0.9916**	1.1051
	IW	Avg	1.5803	1.0000	0.9930	1.0148	0.9829	1.2024	**0.9487**	0.9843
		Med	1.5565	1.0000	0.9959	1.0163	0.9887	1.1732	**0.9550**	0.9630

Note: The table reports the ratio of the losses for the different models relative to the losses of the HAR model. The top panel shows the results for the S&P500. The bottom panel reports the average and median loss ratios across all of the individual stocks. The lowest ratio in each row is highlighted in bold.

Table 4, with the results for the individual stocks summarized in the bottom panel.[15,16] The results for S&P500 index are somewhat mixed, with each of the three "Q" models performing the best for one of the loss functions/window lengths combinations, and the remaining case being won by the SHAR model. The lower panel pertaining to the individual stocks reveals a much cleaner picture: across both loss functions and both window lengths, the HARQ model systematically exhibits the lowest average and median loss. The HARQ-F model fails to improve on the HAR model, again reflecting the difficulties in accurately estimating the weekly and monthly adjustment parameters. Interestingly, and in contrast to the results for the S&P 500, the CHAR, HAR-J and SHAR models generally perform only around as well as the standard HAR model for the individual stocks.

In order to formally test whether the HARQ model significantly outperforms all of the other models, we use a modification of the Reality Check (RC) of White (2000). The standard RC test determines whether the loss from the best model from a set of competitor models is significantly lower than a given benchmark. Instead, we want to test whether the loss of

[15]Due to the estimation errors in RQ, the HARQ models may on are occasions produce implausibly large or small forecasts. Thus, to make our forecast analysis more realistic, we apply an "insanity filter" to the forecasts; see, e.g., Swanson and White (1997). If a forecast is outside the range of values of the target variable observed in the estimation period, the forecast is replaced by the unconditional mean over that period: "insanity" is replaced by "ignorance." This same filter is applied to all of the benchmark models. In practice this trims fewer than 0.1% of the forecasts for any of the series, and none for many.

[16]Surprisingly, the rolling window forecasts provided by the AR model have *lower* average MSE than the HAR model. However, in that same setting the ARQ model also beats the HARQ model.

a given model (HARQ) is lower than that from the best model among a set of benchmark models. As such, we adjust the hypotheses accordingly, testing

$$H_0 : \min_{k=1,...,K} \mathbb{E}[L^k(RV, X) - L^0(RV, X)] \leq 0,$$

versus

$$H_A : \min_{k=1,...,K} \mathbb{E}[L^k(RV, X) - L^0(RV, X)] > 0,$$

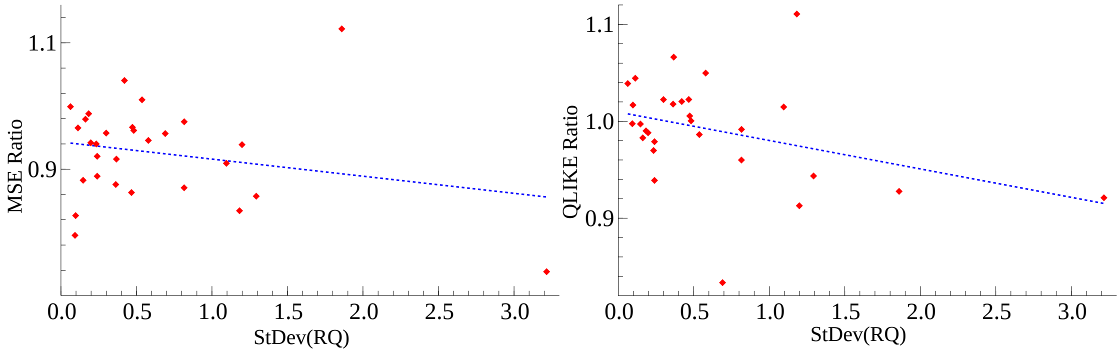
where L^0 denotes the loss of the HARQ model, and L^k, $k = 1, ..., K$ refers to the loss from all of the other benchmark models. A rejection of the null therefore implies that the loss of the HARQ model is significantly lower than all benchmark models. As suggested by White (2000), we implement the Reality Check using the stationary bootstrap of Politis and Romano (1994) with 999 re-samplings and an average block length of five. (The results are not sensitive to this choice of block-length.)

For the S&P500 index, the null hypothesis is rejected at the 10% level for the MSE loss with a p-value of 0.063, but not for QLIKE where the p-value equals 0.871. For the individual stocks, we reject the null in favor of the HARQ model under the MSE loss for 44% (63%) of stocks at the 5% (10%) significance level, respectively, and for 30% (37%) of the stocks under QLIKE loss. On the other hand, none of the benchmark models significantly outperforms the other models for more than one of the stocks. We thus conclude that for a large fraction of the stocks, the HARQ model significantly beats a challenging set of benchmark models commonly used in the literature.

3.4. Dissecting the Superior Performance of the HARQ Model

Our argument as to why the HARQ model improves on the familiar HAR model hinges on the model's ability to place a larger weight on the lagged daily RV on days when RV is measured relatively accurately (RQ is low), and to reduce the weight on days when RV is measured relatively poorly (RQ is high). At the same time, RV is generally harder to measure when it is high, making RV and RQ positively correlated. Moreover, days when RV is high often coincide with days that contain jumps. Thus, to help alleviate concerns that the improvements afforded by the HARQ model are primarily attributable to jumps, we next provide evidence that the model works differently from any of the previously considered models that explicitly allow for distinct dynamic dependencies in the jumps. Consistent with the basic intuition underlying the model, we demonstrate that the HARQ model achieves

Figure 4: Individual Stock Loss Ratios



Note: The graph plots the rolling window forecast MSE and QLIKE loss ratios of the HARQ model to the HAR model against the standard deviation of RQ ($StDev(RQ)$) for each of the individual stocks.

the greatest forecast improvements in environments where the measurement error is highly heteroskedastic. In particular, in an effort to dissect the forecasting results in Table 4, Table 5 further breaks down the results in the previous table into forecasts for days when the previous day's RQ was very high (Top 5% RQ) and the rest of the sample (Bottom 95% RQ). As this breakdown shows, the superior performance of the HARQ model isn't merely driven by adjusting the coefficients when RQ is high. On the contrary, most of the gains in the QLIKE loss for the individual stocks appear to come from "normal" days and the increased persistence afforded by the HARQ model on those days. The results for the HARQ-F model underscores the difficulties in accurately estimating all of the parameters for that model, with the poor performance mostly stemming from the high RQ forecast days. These results also demonstrate that the HARQ model is distinctly different from the benchmark models. The CHAR and HAR-J models primarily show improvements on "high" RQ days, whereas most of the SHAR model's improvements occur for the quieter "normal" days.

If the measurement error variance is constant over time, the HARQ model reduces to the standard HAR model. Correspondingly, it is natural to expect that the HARQ model offers the greatest improvements when the measurement error is highly heteroskedastic. The results in Figure 4 corroborates this idea. The figure plots the MSE and QLIKE loss ratio of the HARQ model relative to the HAR model for the rolling window (RW) forecasts against the standard deviation of RQ ($StDev(RQ)$) for each of the 27 individual stocks. Although $StDev(RQ)$ provides a very noisy proxy, there is obviously a negative relation between the improvements afforded by the HARQ model and the heteroskedasticity in the measurement

Table 5: Stratified Out-of-Sample Forecast Losses

			AR	HAR	HAR-J	CHAR	SHAR	ARQ	HARQ	HARQ-F
						Bottom 95% RQ_t				
						S&P 500				
MSE	RW		1.0745	1.0000	0.9874	0.9629	0.9038	0.9202	**0.8997**	0.9419
	IW		1.1566	1.0000	0.9777	0.9745	0.9139	0.9697	**0.9051**	0.9116
QLIKE	RW		1.5431	1.0000	1.0040	1.0252	**0.9298**	1.1146	1.0145	1.2032
	IW		1.7730	1.0000	0.9718	0.9849	0.8620	1.2025	0.8730	**0.8575**
						Individual Stocks				
MSE	RW	Avg	1.2081	1.0000	0.9939	1.0044	0.9869	1.0324	**0.9557**	0.9928
		Med	1.1987	1.0000	0.9957	1.0034	0.9915	1.0459	**0.9693**	0.9802
	IW	Avg	1.2823	1.0000	0.9944	1.0053	0.9865	1.0814	**0.9633**	0.9581
		Med	1.2114	1.0000	0.9971	1.0067	0.9937	1.0675	**0.9697**	0.9684
QLIKE	RW	Avg	1.4379	1.0000	0.9962	1.0067	0.9888	1.1309	**0.9787**	1.1212
		Med	1.4399	1.0000	0.9958	1.0127	0.9934	1.1180	**0.9820**	1.0673
	IW	Avg	1.6204	1.0000	0.9979	1.0257	0.9787	1.1920	**0.9353**	0.9639
		Med	1.5941	1.0000	0.9995	1.0258	0.9833	1.1616	**0.9395**	0.9532
						Top 5% RQ_t				
						S&P 500				
MSE	RW		0.8992	1.0000	0.9099	0.9578	0.8302	0.7995	0.8186	**0.7789**
	IW		1.2410	1.0000	0.9663	0.9702	0.8995	0.9573	**0.8930**	0.9337
QLIKE	RW		1.4311	1.0000	1.0615	**0.9869**	1.0049	1.2250	1.0310	1.2755
	IW		1.3650	1.0000	0.9703	0.9691	0.9397	1.0594	**0.9357**	0.9456
						Individual Stocks				
MSE	RW	Avg	1.1426	1.0000	1.0228	1.0112	1.0163	0.9461	**0.9268**	1.0222
		Med	1.1591	1.0000	1.0212	1.0175	1.0073	0.9614	**0.9217**	1.0383
	IW	Avg	1.1933	1.0000	1.0053	0.9981	0.9983	1.0243	**0.9476**	1.0124
		Med	1.1984	1.0000	1.0052	1.0007	0.9979	1.0417	**0.9479**	0.9677
QLIKE	RW	Avg	1.3380	1.0000	1.0308	**0.9633**	1.0056	1.3161	1.0916	1.3535
		Med	1.3408	1.0000	0.9998	**0.9377**	1.0097	1.3052	1.0846	1.3250
	IW	Avg	1.3112	1.0000	0.9564	**0.9350**	1.0130	1.2864	1.0464	1.1301
		Med	1.3049	1.0000	0.9636	**0.9371**	1.0111	1.2186	1.0180	1.0045

Note: The table reports the same loss ratios given in Table 4 stratified according to RQ. The bottom panel shows the ratios for days following a value of RQ in the top 5%. The top panel shows the results for the remaining 95% of the days. The lowest ratio in each rows is indicated in bold.

error variance. This is true for both of the loss ratios, but especially so for the QLIKE loss.[17]

3.5. Longer Forecast Horizons

Our analysis up until now has focused on forecasting daily volatility. In this section, we extend the analysis to longer weekly and monthly horizons. Our forecast will be based on

[17]These same negative relations between the average gains afforded by the HARQ model and the magnitude of the heteroskedasticity in the measurement error variance hold true for the increasing window (IW) based forecasts as well.

direct projection, in which we replace the daily RVs on the left-hand-side of the different models with the weekly and monthly RVs.

Our previous findings indicate that for the one-day forecasts, the daily lag is generally the most important and receives by far the largest weight in the estimated HARQ models. Anticipating our results, when forecasting the weekly RV the weekly lag increases in importance, and similarly when forecasting the monthly RV the monthly lag becomes relatively more important. As such, allowing only a time-varying parameter for the daily lag may be sub-optimal for the longer-run forecasts. Hence, in addition to the HARQ and HARQ-F models previously analyzed, we also consider a model in which we only adjust the lag corresponding to the specific forecast horizon. We term this model the HARQ-h model. Specifically, for the weekly and monthly forecasts analyzed here,

$$RV_{t+5|t} = \beta_0 + \beta_1 RV_{t-1} + \underbrace{(\beta_2 + \beta_{2Q} RQ_{t-1|t-5}^{1/2})}_{\beta_{2,t}} RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t \tag{18}$$

and

$$RV_{t+22|t} = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + u_t, + \underbrace{(\beta_3 + \beta_{3Q} RQ_{t-1|t-22}^{1/2})}_{\beta_{3,t}} RV_{t-1|t-22} + u_t, \tag{19}$$

respectively. Note that for the daily horizon, the HARQ and HARQ-h models coincide. Table 6 reports the in-sample parameter estimates for the S&P 500 index for each of the different specifications. The general pattern is very similar to that reported in Table 3. Compared to the standard HAR model, the HARQ model always shifts the weights to the shorter lags. Correspondingly, the HARQ-h model shifts most of the weight to the lag that is allowed to be time-varying. Meanwhile, the HARQ-F model increases the relative weight of the daily and weekly lags, while reducing the weight on the monthly lag, with the estimates for $\beta_{1,t}$ and $\beta_{2,t}$ both statistically significant and negative. The mean lags reported in the bottom panel also shows, that aside from the monthly HARQ-22 model, all of the HARQ specifications on average assign more weight to the more immediate and shorter lags than do the standard HAR models. The estimated HARQ-F models have the shortest mean lags among all of the models.

Turning to the out-of-sample forecast results for the weekly and monthly horizons reported in Tables 7 and 8, respectively, the CHAR and HAR-J benchmark models now both struggle

Table 6: In-Sample Weekly and Monthly Model Estimates

	$h=5$				$h=22$			
	HAR	HARQ	HARQ-F	HARQ-h	HAR	HARQ	HARQ-F	HARQ-h
β_0	0.1717	0.0977	0.0576	0.0170	0.3417	0.2914	0.2845	0.2930
	(0.0432)	(0.0429)	(0.0392)	(0.0466)	(0.0276)	(0.0297)	(0.0339)	(0.0346)
β_1	0.1864	0.4078	0.3408	0.1898	0.1049	0.2547	0.2124	0.1043
	(0.0597)	(0.0717)	(0.0699)	(0.0492)	(0.0502)	(0.0595)	(0.0617)	(0.0495)
β_2	0.3957	0.3159	0.5623	0.6825	0.3342	0.2802	0.4537	0.3364
	(0.0768)	(0.0762)	(0.1056)	(0.0980)	(0.0662)	(0.0658)	(0.0959)	(0.0678)
β_3	0.2709	0.2172	0.0862	0.1609	0.2695	0.2332	0.1122	0.3225
	(0.0655)	(0.0659)	(0.0852)	(0.0725)	(0.0540)	(0.0557)	(0.0658)	(0.0509)
β_{1Q}		-0.2182	-0.1488			-0.1476	-0.1032	
		(0.0420)	(0.0415)			(0.0278)	(0.0309)	
β_{2Q}			-0.4404	-0.5648			-0.3158	
			(0.1514)	(0.1246)			(0.1132)	
β_{3Q}			0.2173				0.2458	-0.1847
			(0.2508)				(0.1519)	(0.1353)
				Mean Lag				
S&P500	5.2626	4.0952	3.0520	3.9564	5.9369	4.9173	3.6799	6.3185
Stocks	6.2593	5.0054	4.6427	4.4950	7.1939	6.0099	5.7646	7.9598

Note: The top panel reports the in-sample parameter estimates for the S&P 500 for the standard HAR model and the various HARQ models for forecasting the weekly ($h = 5$) and monthly ($h = 22$) RVs. Newey and West (1987) robust standard errors allowing for serial correlation up to order 10 ($h = 5$), and 44 ($h = 22$), respectively, are reported in parentheses. The bottom panel reports the mean lag implied by the estimated S&P 500 models, as well as the mean lags averaged across the models estimates for each of the individual stocks.

to beat the HAR model. The SHAR model, on the other hand, offers small improvements over the HAR model in almost all cases. However, the simple version of the HARQ model substantially outperforms the standard HAR model in all of the different scenarios, except for the monthly rolling window MSE loss. The alternative HARQ-F and HARQ-h specifications sometimes perform even better, although there does not appear to be a single specification that systematically dominates all other.

The HARQ-F model, in particular, performs well for the increasing estimation window forecasts, but not so well for the rolling window forecasts. This again underscores the difficulties in accurately estimating the extra adjustment parameters in the HARQ-F model based on "only" 1,000 observations. The HARQ-h model that only adjusts the lag parameter corresponding to the forecast horizon often beats the HARQ model that only adjusts the daily lag parameter. At the weekly horizon, the HARQ and HARQ-h models also both perform better than the standard HAR model. For the monthly horizon, however, it appears more important to adjust the longer lags, and as a result the HARQ-F and HARQ-h models typically both do better than the HARQ model. Of course, as the forecast horizon increases,

Table 7: Weekly Out-of-Sample Forecast Losses

			AR	HAR	HAR-J	CHAR	SHAR	ARQ	HARQ	HARQ-F	HARQ-h
							S&P500				
MSE	RW		1.1450	1.0000	1.4030	0.9919	**0.9018**	1.0798	0.9475	1.2138	0.8884
	IW		1.3509	1.0000	1.1549	0.9673	**0.8365**	1.0861	0.9031	0.9171	0.9232
QLIKE	RW		1.5589	1.0000	1.3047	1.0417	0.9350	1.1892	**0.9159**	1.2529	0.9491
	IW		1.8801	1.0000	1.0898	0.9870	0.8735	1.3717	0.8537	**0.7540**	0.7996
							Individual Stocks				
MSE	RW	Avg	1.2902	1.0000	1.0580	0.9960	0.9864	1.0985	0.9838	1.0234	**0.9765**
		Med	1.2859	1.0000	1.0504	0.9948	0.9904	1.1109	0.9806	1.0051	**0.9517**
	IW	Avg	1.4259	1.0000	1.0500	1.0003	0.9955	1.2126	0.9627	0.9601	**0.9477**
		Med	1.4322	1.0000	1.0435	1.0005	0.9922	1.2110	0.9596	0.9378	**0.9311**
QLIKE	RW	Aveg	1.6564	1.0000	1.1034	1.0124	0.9820	1.2111	**0.9309**	1.0665	0.9873
		Med	1.6554	1.0000	1.0980	1.0156	0.9827	1.2010	**0.9422**	1.0673	0.9869
	IW	Avg	1.9062	1.0000	1.0894	1.0279	0.9770	1.4147	**0.9066**	0.8529	0.8530
		Med	1.8762	1.0000	1.0721	1.0265	0.9781	1.4044	**0.9186**	0.8420	0.8579

Note: The table reports the same loss ratios for the weekly forecasting models previously reported for the one-day-ahead forecasts in Table 4. The top panel shows the results for the S&P 500, while the bottom panel gives the average and median ratios across the individual stocks. The lowest ratio in each row is indicated in bold.

Table 8: Monthly Out-of-Sample Forecast Losses

			AR	HAR	HAR-J	CHAR	SHAR	ARQ	HARQ	HARQ-F	HARQ-h
							S&P500				
MSE	RW		1.1407	1.0000	0.9841	0.9642	**0.9558**	1.0964	1.0708	1.3485	1.2191
	IW		1.2411	1.0000	1.0312	1.0107	1.0119	1.1456	0.9667	**0.9339**	0.9832
QLIKE	RW		1.2455	1.0000	1.0552	0.9919	**0.9532**	1.0518	0.9808	1.1150	1.0450
	IW		1.4159	1.0000	1.0773	0.9937	0.9842	1.2144	0.9368	**0.8448**	0.8843
							Individual Stocks				
MSE	RW	Avg	1.2246	1.0000	1.0173	1.0159	0.9924	1.0969	0.9953	1.0198	**0.9756**
		Med	1.2613	1.0000	1.0118	1.0105	0.9949	1.1005	0.9965	0.9963	**0.9619**
	IW	Avg	1.4127	1.0000	1.0181	1.0123	0.9907	1.2366	0.9770	**0.9723**	0.9815
		Med	1.4052	1.0000	1.0172	1.0145	0.9927	1.2182	0.9692	**0.9480**	0.9705
QLIKE	RW	Avg	1.4125	1.0000	1.0385	1.0143	0.9909	1.1335	0.9485	0.9127	**0.8804**
		Med	1.4300	1.0000	1.0367	1.0125	0.9928	1.1228	0.9481	0.8778	**0.8635**
	IW	Avg	1.6612	1.0000	1.0360	1.0257	0.9885	1.3519	0.9371	**0.8185**	0.8278
		Med	1.6294	1.0000	1.0224	1.0296	0.9912	1.3619	0.9442	**0.8245**	0.8442

Note: The table reports the same loss ratios for the monthly forecasting models previously reported for the one-day-ahead forecasts in Table 4. The top panel shows the results for the S&P 500, while the bottom panel gives the average and median ratios across the individual stocks. The lowest ratio in each row is indicated in bold.

the forecasts become smoother and closer to the unconditional volatility, and as such the relative gains from adjusting the parameters are invariably reduced.

4. Robustness

4.1. Alternative Realized Variance Estimators

The most commonly used 5-minute RV estimator that underly all of our empirical results discussed above provides a simple way of mitigating the contaminating influences of market microstructure "noise" arising at higher intraday sampling frequencies.[18] However, a multitude of alternative RV estimators that allow for the use of higher intraday sampling frequencies have, of course, been proposed in the literature. In this section we consider some of the most commonly used of these robust estimators, namely: sub-sampled, two-scales, kernel, and pre-averaged RV, each described in more detail below. Our implementation of these alternative estimators will be based on 1-minute returns.[19] We begin by showing that the HARQ model based on the simple 5-minute RV outperforms the standard HAR models based on these alternative 1-minute robust RV estimators. We also show that despite the increased efficiency afforded by the use of a higher intraday sampling frequency, the HARQ models based on these alternative RV estimators still offer significant forecast improvements relative to the standard HAR models based on the same robust RV estimators. To allow for a direct comparison across the different estimators and models, we always take daily 5-minute RV as the forecast target. As such, the set-up mirrors that of the CHAR model in equation (16) with the different noise-robust RV estimators in place of the jump-robust BPV estimator.

The subsampled version of RV ($SS\text{-}RV$) was introduced by Zhang, Mykland, and Aït-Sahalia (2005). Subsampling provides a simple way to improve on the efficiency of the standard RV estimator, by averaging over multiple time grids. Specifically, by computing the 5-minute RV on time grids with 5-minute intervals starting at 9:30, 9:31, 9:32, 9:33 and 9:34, the $SS\text{-}RV$ estimator is obtained as the average of these five different RV estimators. The two-scale RV ($TS\text{-}RV$) of Zhang, Mykland, and Aït-Sahalia (2005), bias-corrects the $SS\text{-}RV$ estimator through a jackknife type adjustment and the use of RV at the highest possible frequency. It may be expressed as,

$$TS\text{-}RV = SS\text{-}RV - \frac{M}{M^{(all)}} RV^{(all)}, \tag{20}$$

[18] As previously noted, the comprehensive comparisons in Liu, Patton, and Sheppard (2015) also show that HAR type models based on the simple 5-minute RV generally perform quite well in out-of-sample forecasting.

[19] Since we only have access to 5-minute returns for the S&P 500 futures contract, our results in this section pertaining to the market index will be based on the SPY ETF contract. We purposely do not use the SPY in the rest of the paper, as it is less actively traded than the S&P 500 futures for the earlier part of our sample.

where $M^{(all)}$ denotes the number of observations at the highest frequency (here 1-minute), and $RV^{(all)}$ refers to the resulting standard RV estimator. The realized kernel (RK), developed by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008), takes the form,

$$RK = \sum_{h=-H}^{H} k\left(\frac{h}{H+1}\right)\gamma_h, \quad \gamma_h = \sum_{j=|h|+1}^{M} r_{t,i} r_{t,i-|h|}, \tag{21}$$

where $k(x)$ is a kernel weight function, and H is a bandwidth parameter. We follow Barndorff-Nielsen, Hansen, Lunde, and Shephard (2009) and use a Parzen kernel with their recommended choice of bandwidth. Finally, we also implement the pre-averaged RV ($PA\text{-}RV$) estimator of Jacod, Li, Mykland, Podolskij, and Vetter (2009), defined by,

$$PA\text{-}RV = \frac{1}{\sqrt{M}\theta\psi_2} \sum_{i=1}^{M-H+2} \bar{r}_{t,i}^2 - \frac{\psi_1}{2M\theta^2\psi_2} RV, \tag{22}$$

where $\bar{r}_{t,i} = \sum_{j=1}^{H-1} g(j/H) r_{t,j}$. For implementation we follow Jacod, Li, Mykland, Podolskij, and Vetter (2009) in choosing the weighting function $g(x) = min(x, 1-x)$, along with their recommendations for the data-driven bandwidth H, and tuning parameter θ. The ψ_i parameters are all functionals of the weighting function $g(x)$.

Table 9 compares the out-of-sample performance of the standard HAR model based on these more efficient noise-robust estimators with the forecasts from the HARQ model that uses 5-minute RV. (For comparison, the first column of this table shows the performance of the HAR model using 5-minute RV; the entries here are the inverses of those in the "HARQ" column in Table 4.) As the table shows, the HARQ model based on the 5-minute RV easily outperforms the forecasts from the standard HAR models based on the 1-minute robust RV estimators. In fact, in line with the results of Liu, Patton, and Sheppard (2015), for most of the series the HAR models based on the noise-robust estimators do not systematically improve on the standard HAR model based on the 5-minute RV. While still inferior, the HAR model based on $SS - RV$ gets closest in performance to the HARQ model, while the standard HAR model based on the other three estimators generally perform far worse. Of course, the $TS - RV$, RK, and $PA - RV$ estimators were all developed to allow for the consistent estimation of IV through the use of ever finely sampled returns. Thus, it is possible that even finer sampled RVs than the 1-minute frequency used here might outweigh the additional complexity of the estimators and result in better out-of-sample forecasts.

Table 9: HAR Models based on Noise-Robust RVs versus HARQ Model

			RV	SS-RV	TS-RV	RK	PA-RV
					S&P 500		
MSE	RW		1.2574	1.0801	1.3472	1.3443	1.3521
	IW		1.1290	1.1882	1.2468	1.1769	1.1604
QLIKE	RW		1.0025	1.0149	1.1476	1.0493	1.0331
	IW		1.1487	1.1424	1.2637	1.4182	1.3608
					Individual Stocks		
MSE	RW	Average	1.0714	1.0523	1.1641	1.0446	1.0476
		Median	1.0628	1.0430	1.1665	1.0604	1.0542
	IW	Average	1.0531	1.0502	1.1319	1.0603	1.0579
		Median	1.0499	1.0544	1.1206	1.0695	1.0479
QLIKE	RW	Average	1.0100	1.0438	1.0890	1.0615	1.1107
		Median	1.0064	1.0457	1.0923	1.0587	1.0988
	IW	Average	1.0552	1.0535	1.1297	1.1557	1.1425
		Median	1.0471	1.0491	1.1240	1.1520	1.1458

Note: The table reports the loss ratios of the HAR model using 1-minute noise-robust RV estimators and the 5-minute RV used previously, compared to the loss of the HARQ model using 5-minute RV. The S&P 500 results are based on returns for the SPY.

Further along these lines, all of the noise-robust 1-minute RV estimators are, of course, still subject to some measurement errors. To investigate whether adjusting for these errors remains useful, we also estimate HARQ models based on each of the alternative RV measures.[20] Table 10 shows the ratios of the resulting losses for the $HARQ$ to HAR models using a given realized measure. For ease of comparison, we also include the previous results for the simple 5-minute RV estimator. The use of higher 1-minute sampling and the more efficient RV estimators, should in theory result in smaller measurement errors. Consistent with this idea, the improvements afforded by the HARQ model for the 1-minute noise-robust estimators are typically smaller than for the 5-minute RV. However, the HARQ models for the 1-minute RVs still offer clear improvements over the standard HAR models, with most of the ratios below one.[21]

[20]Since the measurement error variance for all of the estimators are proportional to IQ (up to a small noise term, which is negligible at the 1-minute level), we continue to rely on the 5-minute RQ for estimating the measurement error variance in these HARQ models.

[21]Chaker and Meddahi (2013) have previously explored the use of $RV^{(all)}$ as an estimator (up to scale) for the variance of the market microstructure noise in the context of RV-based volatility forecasting. Motivated by this idea, we also experimented with the inclusion of an $RV^{(all)1/2} \cdot RV$ interaction term in the HAR and HARQ models as a simple way to adjust for the temporal variation in the magnitude of the market microstructure noise. The out-of-sample forecasts obtained from these alternative specifications were generally inferior to the forecasts from the basic HARQ model. Additional details of these results are available in Appendix D.

Table 10: HARQ versus HAR Models based on Noise-Robust RVs

			RV	SS-RV	TS-RV	RK	PA-RV
					S&P 500		
MSE	RW		0.7953	1.0059	0.8606	0.7873	0.7896
	IW		0.8857	0.8837	0.9749	0.8956	0.8952
QLIKE	RW		0.9975	1.0585	0.9776	0.9711	1.0504
	IW		0.8705	0.9195	0.9317	0.8971	0.9262
					Individual Stocks		
MSE	RW	Average	0.9333	0.9496	0.9547	1.0072	0.9698
		Median	0.9409	0.9593	0.9521	1.0012	0.9719
	IW	Average	0.9496	0.9582	0.9723	0.9755	0.9700
		Median	0.9525	0.9590	0.9730	0.9710	0.9771
QLIKE	RW	Average	0.9901	0.9462	0.9804	1.0874	0.9748
		Median	0.9936	0.9600	0.9831	0.9973	0.9793
	IW	Average	0.9477	0.9474	0.9665	0.9492	0.9443
		Median	0.9550	0.9447	0.9624	0.9431	0.9445

Note: The table reports the loss ratios of the HARQ models to the HAR models for the 1-minute noise-robust estimators. The S&P 500 results are based on returns for the SPY.

4.2. Alternative Quarticity Estimators

The integrated quarticity is notoriously difficult to estimate. To investigate the sensitivity of our results to this additional layer of estimation uncertainty, we consider a number of alternative estimators of IQ in place of the standard RQ estimator used in the HARQ models throughout the rest of the paper.

One such estimator is provided by the Tri-Power Quarticity of Barndorff-Nielsen and Shephard (2006),

$$TPQ_t \equiv M\mu_{4/3}^{-3} \sum_{i=1}^{M-2} |r_{t,i}|^{4/3} |r_{t+1,i}|^{4/3} |r_{t+2,i}|^{4/3}, \tag{23}$$

where $\mu_{4/3} \equiv 2^{2/3}\Gamma(7/6)/\Gamma(1/2) = \mathbb{E}(|Z|^{4/3})$. In contrast to the standard RQ estimator, TPQ remains consistent for IQ in the presence of jumps. Further along these lines, we also consider the jump-robust MedRQ estimator developed by Andersen, Dobrev, and Schaumburg (2012, 2014),

$$MedRQ \equiv \frac{3\pi}{9\pi + 72 - 52\sqrt{3}} \frac{M^2}{M-2} \sum_{i=1}^{M-2} \text{median}\left(|r_{t,i}|, |r_{t,i+1}|, |r_{t,i+2}|\right)^4, \tag{24}$$

as well as the Truncated RQ estimator based on the ideas of Mancini (2009), formally defined by,

$$TrRQ \quad \equiv M \sum_{i=1}^{M} |r_{t,i}^4| \mathbb{I}_{\{|r_{t,i}| \leq \alpha_i M^{\varpi}\}}, \tag{25}$$

Table 11: Alternative IQ estimators.

IQ-estimator			RQ	TPQ	MedRQ	TrRQ	RQ_{15min}	Bootstrap
					S&P500			
MSE	RW		1.0000	1.0497	1.0254	1.0208	1.0590	**0.9925**
	IW		1.0000	1.1635	1.0328	**0.9948**	0.9805	0.9981
QLIKE	RW		1.0000	1.0933	1.1227	0.9971	1.0231	**0.9933**
	IW		1.0000	**0.9814**	1.0548	1.2060	1.0414	0.9998
					Individual Stocks			
MSE	RW	Avg	1.0000	1.0403	1.0139	1.0531	1.0586	**0.9936**
		Med	**1.0000**	1.0497	1.0201	1.0378	1.0238	1.0003
	IW	Avg	1.0000	1.0211	1.0191	1.0491	1.0229	**0.9994**
		Med	1.0000	1.0220	1.0259	1.0558	1.0125	**0.9998**
QLIKE	RW	Avg	1.0000	1.0040	1.0089	1.0541	1.0254	**0.9995**
		Med	1.0000	0.9980	**0.9968**	1.0376	1.0015	0.9992
	IW	Avg	**1.0000**	1.0050	1.0047	1.0390	1.0014	1.0001
		Med	**1.0000**	1.0040	1.0065	1.0303	0.9989	1.0001

Note: The table reports the out-of-sample forecast losses from the HARQ model using different IQ estimators. All of the losses are reported relative to the losses from the $HARQ$ model based on the the standard RQ estimator used throughout the rest of the paper. The top panel shows the results for the S&P500, while the bottom panel reports the average and median ratios across each of the individual stocks. The lowest ratio in each row is indicated in bold.

with the tuning parameters α_i and ϖ implemented as in Bollerslev, Todorov, and Li (2013). It has also previously been suggested that the integrated quarticity may be more accurately estimated using a coarser time grid than that used for estimating integrated variance; see, e.g., Bandi and Russell (2008). To this end, we consider estimating RQ based on coarser 15-minute returns. The approximation in equation (4) that motivates the HARQ model is, of course, asymptotic (for $M \to \infty$). Finally, we also consider the wild bootstrap of Gonçalves and Meddahi (2009). Specifically, re-sampling the high-frequency returns 999 times for each of the days in the sample, we use the sample variance of the relevant RV across the bootstrap replications as the estimator of the measurement error variance.

For ease of comparison, Table 11 summarizes the out-of-sample forecast losses from the HARQ models based on each of these different IQ estimators relative to the losses for the HARQ model based on the standard 5-minute RQ estimator used in the rest of the paper. As the results show, the performance of the HARQ models based on the different IQ estimators are generally close. The only alternative estimator that possibly improves on the standard RQ estimator in a systematic fashion is the bootstrapped variance of RV. However, the improvements are at best minor.

Table 12: Alternative HARQ Specifications

			Alternative RQ Transformations					Adding $RQ^{1/2}$	
			RQ	$RQ^{1/2}$	$RQ^{-1/2}$	RQ^{-1}	$Log(RQ)$	HAR	HARQ
					S&P 500				
MSE	RW		1.0037	**1.0000**	1.2123	1.2334	1.3313	1.1552	1.0004
	IW		1.0344	**1.0000**	1.1166	1.1357	1.0736	1.1166	1.1402
QLIKE	RW		**0.9484**	1.0000	1.0952	1.0950	1.8104	0.9919	**0.9731**
	IW		1.0222	**1.0000**	1.1327	1.3217	2.0107	1.0452	1.0089
					Individual Stocks				
MSE	RW	Avg	1.0108	**1.0000**	1.0808	1.0931	1.0329	1.0339	1.0207
		Med	1.0112	**1.0000**	1.0577	1.0664	1.0336	1.0050	**0.9904**
	IW	Avg	1.0189	**1.0000**	1.0495	1.0644	1.0143	0.9979	**0.9895**
		Med	1.0198	**1.0000**	1.0403	1.0598	1.0082	0.9986	**0.9863**
QLIKE	RW	Avg	**0.9973**	1.0000	1.0678	1.0814	1.3723	1.0377	1.0639
		Med	**0.9847**	1.0000	1.0458	1.0579	1.3324	1.0219	1.0133
	IW	Avg	1.0263	**1.0000**	1.0961	1.1155	1.2903	1.0394	1.0081
		Med	1.0241	**1.0000**	1.0778	1.0886	1.2084	1.0279	**0.9937**

Note: The table reports the out-of-sample forecast losses for HARQ models based on different transformations of RQ. All of the losses are reported relative to those for the HARQ model the rely on $RQ^{1/2}$. The left panel reports the results based on alternative RQ interaction terms. The right panel reports the results from models that include $RQ^{1/2}$ as an explanatory variable. The top panel pertains to the S&P 500, while the bottom panel gives the average and median ratios across all of the individual stocks. The lowest ratio in each row is indicated in bold.

4.3. Alternative HARQ Specifications

The HARQ model is designed to allow for temporal variation in the degree of attenuation bias based on an estimate of the heteroskedastic measurement error variance. The exact specification of the model, and the interaction of RV with the square root of RQ, is, however, somewhat ad hoc and primarily motivated by concerns related to numerical stability. Equation (9), in particular, naturally suggests allowing the parameters of the model to vary with the inverse of RQ, as opposed to $RQ^{1/2}$. But, RQ may be close to zero, and as such the inverse of RQ is prone to amplify any estimation errors, resulting in inaccurate estimates of the inverse of IQ, and in turn unstable time-varying autoregressive parameter estimates. To further investigate this issue, we consider the out-of-sample forecasts from alternative HARQ specifications in which we substitute RQ, $RQ^{-1/2}$, RQ^{-1}, and $log(RQ)$ in place of $RQ^{1/2}$ in equation (2.3). All of the HARQ specifications that we have considered so far have also restricted the intercept in the models, or β_0, to be constant. We now consider two alternative specifications, where we add $RQ^{1/2}$ as an additional explanatory variable to the standard HAR and HARQ models, thereby allowing for a time-varying intercept in the HAR(Q) model. Table 12 reports the out-of-sample forecast results from each of these alternative HARQ specifications. For ease of comparison, we again normalize all of the losses relative to the losses of the HARQ model based on $RQ^{1/2}$ analyzed throughout. The first set of columns show the results for the alternative RQ transformations. The models based on $RQ^{1/2}$ and RQ perform roughly the same. Meanwhile, as expected, the modified HARQ models that rely on negative powers of RQ, or the log-transform of RQ, all result in numerically unstable parameter estimates and correspondingly inferior out-of-sample forecasts.[22]

The last two columns of the table give the results from directly including $RQ^{1/2}$ in the HAR and HARQ models. While the HAR model with $RQ^{1/2}$ tend to performs worse than the HARQ model, it does improve on the standard HAR model, as it is able to mimic the HARQ model and reduce the forecasts in situations with large measurement errors. However, the model does so less effectively than the HARQ model forecasts. The HARQ model that includes $RQ^{1/2}$ does improve on the standard HARQ model for some of the forecast scenarios, but performs worse in others, and in no case is the forecast improvement very large. Overall, we conclude that the simple HARQ model in equation (2.3) appears the more stable and generally superior model compared to any of these alternative specifications.

4.4. Alternative Q-Models

The motivation behind the HARQ model is equally applicable to other realized volatility based forecasting models, including the benchmark models analyzed in our forecast comparisons. In particular, the HAR-J model defined in equation (14) is readily modified in a manner completely analogous to the HARQ model, resulting in the HARQ-J model,

$$RV_t = \beta_0 + (\beta_1 + \beta_{1Q} RQ_{t-1}^{1/2}) RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + \beta_J J_{t-1} + u_t. \qquad (26)$$

The CHAR model in (16) relies on the jump-robust bi-power variation (BPV) measure in place of the realized volatility for predicting the future volatility. As shown by Barndorff-Nielsen and Shephard (2006), the asymptotic variance of BPV equals $2.61 \Delta IQ_t$. This asymptotic variance is naturally estimated by the Tri-Power Quarticity (TPQ) previously defined in equation (23), which remains consistent for IQ in the presence of jumps. Correspondingly, we define the CHARQ model as,

$$RV_t = \beta_0 + (\beta_1 + \beta_{1Q} TPQ_{t-1}^{1/2}) BPV_{t-1} + \beta_2 BPV_{t-1|t-5} + \beta_3 BPV_{t-1|t-22} + u_t. \qquad (27)$$

[22] We also experimented with "higher-order" models, allowing the autoregressive parameters to depend on multiple RQ terms; e.g., both $RQ^{1/2}$ and RQ. However, the difficulties in accurately estimating the parameters for these more elaborate models, again translate into inferior out-of-sample forecasts.

Table 13: Alternative Q-Model In-Sample Estimates

	HAR-J	HARQ-J	CHAR	CHARQ	SHAR	SHARQ
β_0	0.1208	0.0045	0.1361	-0.0064	0.0692	-0.0766
	(0.0606)	(0.0561)	(0.0595)	(0.0618)	(0.0667)	(0.0613)
β_1	0.3599	0.6035	0.2657	0.5834		
	(0.0891)	(0.0882)	(0.0958)	(0.0967)		
β_2	0.4341	0.3519	0.4980	0.4189	0.4176	0.3527
	(0.1300)	(0.1285)	(0.1489)	(0.1524)	(0.1223)	(0.1260)
β_3	0.1856	0.1057	0.1751	0.1131	0.1530	0.0822
	(0.1068)	(0.1034)	(0.1201)	(0.1138)	(0.1013)	(0.0997)
β_J	-1.0033	-0.3393				
	(0.3668)	(0.2857)				
β_1^+					-0.3734	-0.2027
					(0.1772)	(0.2054)
β_1^-					1.1282	1.5723
					(0.2773)	(0.2658)
β_{1Q}		-0.3266		-0.5410		
		(0.0617)		(0.1800)		
β_{1Q}^+						-1.3227
						(0.3632)
β_{1Q}^-						0.2485
						(0.1316)
R^2	0.5376	0.5638	0.5347	0.5526	0.5751	0.5972
MSE	2.4908	2.3495	2.5064	2.4097	2.2887	2.1693
QLIKE	0.1538	0.1336	0.1442	0.1377	0.3315	0.2154
$\overline{R}^2$ Stocks	0.4913	0.5110	0.4891	0.5106	0.4986	0.5239
$\overline{MSE}$ Stocks	14.8224	14.0916	14.9265	14.1891	14.5431	13.6915
$\overline{QLIKE}$ Stocks	0.1492	0.1470	0.1509	0.1449	0.1496	0.1534

Note: The table reports the S&P 500 in-sample parameter estimates and measures of fit for the different benchmark models and HARQ-adaptations discussed in the main text. The bottom panel shows the average R^2s, MSEs, and $QLIKE$s across the individual stocks.

The asymptotic distribution of the RV_t^+ and RV_t^- measures included in the SHAR model in equation (17) is unknown. However, a measure that is strongly correlated with their asymptotic variances should work well in terms of adjusting the parameters for measurement errors, as the estimated regression coefficients will automatically make up for any proportional differences. Hence, as a shortcut we simply rely on RQ to proxy the asymptotic variance of both RV^+ and RV^-, defining the SHARQ model,

$$RV_t = \beta_0 + (\beta_1^+ + \beta_{1Q}^+ RQ_{t-1}^{1/2}) RV_{t-1}^+ + (\beta_1^- + \beta_{1Q}^- RQ_{t-1}^{1/2}) RV_{t-1}^- + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t. \quad (28)$$

Table 13 reports the parameter estimates for each of these Q-models and their baseline counterparts. The general pattern directly mirrors that of the estimates for the HARQ and HAR models. All of the models shift the weights from the weekly and monthly lags to the

Table 14: Alternative Q-Model Out-of-Sample Forecast Losses

			HARQ	HARQ-J	CHARQ	SHARQ
			S&P500			
MSE	RW		0.8266	0.9243	0.8951	1.4412
	IW		0.8944	0.9335	1.0609	1.1027
QLIKE	RW		0.9464	0.9653	1.0235	1.4576
	IW		0.8809	0.9015	0.8825	1.2849
			Individual Stocks			
MSE	RW	Avg	0.9349	0.9397	0.9525	1.1308
		Med	0.9418	0.9513	0.9539	1.0840
	IW	Avg	0.9525	0.9666	0.9451	1.0870
		Med	0.9525	0.9662	0.9548	1.0554
QLIKE	RW	Avg	0.9902	0.9902	0.9879	1.1706
		Med	0.9916	0.9952	0.9900	1.1609
	IW	Avg	0.9487	0.9548	0.9306	1.1305
		Med	0.9550	0.9594	0.9277	1.1154

Note: The table reports the loss ratios for the alternative Q-model specifications discussed in the main text. All of the losses are reported relative the relevant baseline models without the Q-adjustment terms. The top panel shows the results for the S&P 500, while the bottom panel reports the average and median ratios across all of the individual stocks.

daily lag, with higher measurement error variances pulling the daily lag parameters closer to zero. Looking specifically at the HAR-J model, the parameter associated with the jump component is significantly negative: on days where part of the total RV is attributable to jumps, the next day's RV is reduced by -1.003 times the jump component. In the HARQ-J model, however, the measurement error subsumes a large portion of this jump variation: the β_J coefficient is reduced by two-thirds, and is no longer statistically significant. This same result holds true for the individual stocks, where the jump parameters are significant at the 5% level for 60% of the stocks for the HAR-J models, compared to only 10% of the stocks for the HARQ-J models. Also, comparing the SHAR and SHARQ models, the latter shifts the weight even further away from the positive part of RV to the negative part, so that only RV^- is significant in the SHARQ model.

The out-of-sample forecast results from each of these different Q-models are summarized in Table 14. For comparison purposes, we also include the results for the basic HARQ model. To facilitate the interpretation of the results, we report the loss ratios with respect to the relevant baseline models, that is the losses for the HARQ model is reported relative to the losses for the standard HAR model, the CHARQ model losses relative to the CHAR model losses, and so forth. The improvements obtained for the HARQ-J and CHARQ models are generally in line with those for the basic HARQ model. This is true both for the S&P 500 losses given in the top panel and the average losses across the individual stocks reported in the bottom panel. The SHARQ model, however, does not improve on the standard SHAR model. Of course, in contrast to the other Q-models, our specification of the SHARQ model relies on an imperfect proxy for the asymptotic variance of the measurement errors. This might help explain the relatively poor performance of that model, and also indirectly highlight the importance of using a proper approximation for the distribution of the measurement errors to guide the adjustments of the autoregressive parameters and the forecasts from the models.

5. Conclusion

We propose a simple-to-implement new class of realized volatility based forecasting models. The models improve on the forecasts from standard volatility forecasting models, by explicitly accounting for the temporal variation in the magnitude of the measurement errors in the realized volatilities and the way in which the errors attenuate the parameters of the models. A particularly simple member of this new class of models, which we term the HARQ model, outperforms the forecasts from several other commonly used models. This holds true both in simulations and in- and out-of-sample forecasts of the volatility of the S&P 500 aggregate market portfolio and a number of individual stocks.

The new models developed here may usefully be applied in many other situations. The volatility risk premium, for example, defined as the difference between the so-called risk neutral expectation of the future volatility of the aggregate market portfolio and the actual statistical expectation of the market volatility, has recently received a lot of attention in the literature. The risk neutral expectation of the volatility is readily inferred from options prices in an essentially model-free manner. The actual volatility forecasts, however, invariable depends on the model used for constructing the forecasts. Bekaert and Hoerova (2014) and Conrad and Loch (2015) have both recently demonstrated how the use of different realized volatility based forecasting models, including versions of the HAR, HAR-J and CHAR models analyzed here, can materially affect the estimates of the volatility risk premium and the interpretation thereof. The HARQ models, of course, hold the promise of even more accurate forecasts and better volatility risk premium estimates, and in turn new insights and a deeper understanding of the economic mechanisms behind the temporal variation in the premium.

All of the forecasting models developed here are univariate. However, most practical questions related to risk measurement and management are intrinsically multivariate in nature, requiring the forecasts of both asset return variances and covariances. Building on the distributional results of Barndorff-Nielsen and Shephard (2004a), the realized volatility based Vech HAR model of Chiriac and Voev (2010) may readily be extended to incorporate the effect of the measurement errors in the realized variances and covariance in a manner analogous to the one employed here for the univariate HARQ models. The multivariate HEAVY model of Noureldin, Shephard, and Sheppard (2012) and the Realized Beta GARCH model of Hansen, Lunde, and Voev (2014) may similarly be extended to allow the parameters of the models to vary with the degree of the measurement errors in the realized covariance matrix. We leave further work along these lines for future research.

6. References

Aït-Sahalia, Y., Mancini, L., 2008. Out of sample forecasts of quadratic variation. Journal of Econometrics 147 (1), 17–33.

Andersen, T. G., Bollerslev, T., 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review 39 (4), 885–905.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., Diebold, F. X., 2013. Financial risk measuremet for financila risk management. In: Handbook of the Economics of Finance (eds. G. Constantinides, M. Harris and R. Stulz). Elsevier Inc., pp. 1127–1220.

Andersen, T. G., Bollerslev, T., Diebold, F. X., 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., Bollerslev, T., Diebold, F. X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71 (2), 579–625.

Andersen, T. G., Bollerslev, T., Meddahi, N., 2004. Analytical evaluation of volatility forecasts. International Economic Review 45 (4), 1079–1110.

Andersen, T. G., Bollerslev, T., Meddahi, N., 2005. Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. Econometrica 73 (1), 279–296.

Andersen, T. G., Bollerslev, T., Meddahi, N., 2011. Realized volatility forecasting and market microstructure noise. Journal of Econometrics 160 (1), 220–234.

Andersen, T. G., Dobrev, D., Schaumburg, E., 2012. Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics 169 (1), 75–93.

Andersen, T. G., Dobrev, D., Schaumburg, E., 2014. A robust neighborhood truncation approach to estimation of integrated quarticity. Econometric Theory 30 (01), 3–59.

Asai, M., McAleer, M., Medeiros, M. C., 2012. Modelling and forecasting noisy realized volatility. Computational Statistics and Data Analysis 56 (1), 217–230.

Bandi, F. M., Russell, J. R., 2006. Separating microstructure noise from volatility. Journal of Financial Economics 79 (3), 655–692.

Bandi, F. M., Russell, J. R., 2008. Microstructure noise, realized variance, and optimal sampling. Review of Economic Studies 75 (2), 339–369.

Bandi, F. M., Russell, J. R., Yang, C., 2013. Realized volatility forecasting in the presence of time-varying noise. Journal of Business and Economic Statistics 31 (3), 331–345.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., Shephard, N., 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica 76 (6), 1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., Shephard, N., 2009. Realized kernels in practice: Trades and quotes. Econometrics Journal 12 (3), C1–C32.

Barndorff-Nielsen, O. E., Kinnebrock, S., Shephard, N., 2010. Measuring downside risk: realised semivariance. In: Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle (eds. T. Bollerslev, J. Russell and M. Watson). Oxford University Press, pp. 117–136.

Barndorff-Nielsen, O. E., Shephard, N., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society: Series B 64 (2), 253–280.

Barndorff-Nielsen, O. E., Shephard, N., 2004a. Econometric analysis of realized covariation: high frequency based covariance, regression and correlation in financial economics. Econometrica 72 (3), 885–925.

Barndorff-Nielsen, O. E., Shephard, N., 2004b. Power and bipower variation with stochastic volatility and jumps. Journal of Financial Econometrics 2 (1), 1–37.

Barndorff-Nielsen, O. E., Shephard, N., 2006. Econometrics of testing for jumps in financial economics using bipower variation. Journal of Financial Econometrics 4 (1), 1–30.

Bekaert, G., Hoerova, M., 2014. The vix, the variance premium and stock market volatility. Journal of Econometrics 183 (2), 181–192.

Bollerslev, T., Todorov, V., Li, S. Z., 2013. Jump tails, extreme dependencies, and the distribution of stock returns. Journal of Econometrics 172 (2), 307–324.

Chaker, S., Meddahi, N., 2013. Volatility forecasting when the noise variance is time-varying. Working Paper, Bank of Canada and University of Toulouse.

Chiriac, R., Voev, V., 2010. Modeling and forecasting multivaraite realized volatility. Journal of Applied Econometrics 26 (6), 922–947.

Conrad, C., Loch, K., 2015. The variance risk premium and fundamental uncertainty. Working Paper, Heidelberg University.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7 (2), 174–196.

Dobrev, D. P., Szerszen, P. J., 2010. The information content of high-frequency data for estimating equity return models and forecasting risk. Working Paper, Federal Reserve Board, Washington D.C.

Doornik, J., 2009. Object-oriented Matrix Programming Using Ox. Timberlake Consultants Press.

Ghysels, E., Sinko, A., 2011. Volatility forecasting and microstructure noise. Journal of Econometrics 160 (1), 257–271.

Gonçalves, S., Meddahi, N., 2009. Bootstrapping realized volatility. Econometrica 77 (1), 283–306.

Hansen, P. R., Lunde, A., 2006. Realized variance and market microstructure noise. Journal of Business and Economic Statistics 24 (2), 127–161.

Hansen, P. R., Lunde, A., 2014. Estimating the persistence and the autocorrelation function of a time series that is measured with error. Econometric Theory 30 (1), 60–93.

Hansen, P. R., Lunde, A., Voev, V., 2014. Realized beta garch: A multivariate garch model with realized measures of volatility. Journal of Applied Econometrics 29 (5), 774–799.

Huang, X., Tauchen, G., 2005. The relative contribution of jumps to total price variance. Journal of Financial Econometrics 3 (4), 456–499.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M., Vetter, M., 2009. Microstructure noise in the continuous case: the pre-averaging approach. Stochastic Processes and their Applications 119 (7), 2249–2276.

Li, J., Xiu, D., 2013. Spot variance regressions. Working Paper, Duke University and University of Chicago.

Liu, L., Patton, A. J., Sheppard, K., 2015. Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. Journal of Econometrics forthcoming.

Mancini, C., 2009. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. Scandinavian Journal of Statistics 36 (2), 270–296.

Newey, W. K., West, K. D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55 (3), 703–08.

Noureldin, D., Shephard, N., Sheppard, K., 2012. Multivariate high-frequency-based volatility (heavy) models. Journal of Applied Econometrics 27 (6), 908–933.

Patton, A. J., 2011. Data-based ranking of realised volatility estimators. Journal of Econometrics 161 (2), 284–303.

Patton, A. J., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics forthcoming.

Politis, D. N., Romano, J. P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89 (428), 1303–1313.

Shephard, N., Xiu, D., 2014. Econometric analysis of multivariate realised qml: efficient positive semi-definite estimators of the covariation of equity prices. Working Paper, Harvard University and University of Chicago.

Sizova, N., 2011. Integrated variance forecasting: Model-based versus reduced-form. Journal of Econometrics 162 (2), 294–311.

Staudenmayer, J., Buonaccorsi, J. P., 2005. Measurement error in linear autoregressive models. Journal of the American Statistical Association 100 (471), 841–852.

Swanson, N. R., White, H., 1997. Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. International journal of Forecasting 13 (4), 439–461.

White, H., 2000. A reality check for data snooping. Econometrica 68 (5), 1097–1126.

Zhang, L., Mykland, P. A., Aït-Sahalia, Y., 2005. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. Journal of the American Statistical Association 100 (472), 1394–1411.

Appendix A. Simulation Design

Our simulations are based on the two-factor stochastic volatility model previously analyzed by Huang and Tauchen (2005) among others,

$$
\begin{aligned}
d\log S_t &= \mu dt + \sigma_{ut}\nu_t \left(\rho_1 dW_{1t} + \rho_2 dW_{2t} + \sqrt{1-\rho_1^2-\rho_2^2}\, dW_{3t}\right) \\
\nu_t^2 &= \text{s-exp}\left\{\beta_0 + \beta_1 \nu_{1t}^2 + \beta_2 \nu_{2t}^2\right\} \\
d\nu_{1t}^2 &= \alpha_1 \nu_{1t}^2 dt + dW_{1t} \\
d\nu_{2t}^2 &= \alpha_2 \nu_{2t}^2 dt + (1+\phi\nu_{2t}^2) dW_{2t} \\
\sigma_{ut} &= C + Ae^{-at} + Be^{-b(1-t)},
\end{aligned}
$$

where s-exp denotes the exponential function with a polynomial splined at high values to avoid explosive behavior. We follow Huang and Tauchen (2005) in setting $\alpha = 0.03$, $\beta_0 = -1.2$, $\beta_1 = 0.04$, $\beta_2 = 1.5$, $\alpha_1 = -0.00137$, $\alpha_2 = -1.386$, $\phi_1 = 0.25$, and $\rho_1 = \rho_2 = -0.3$. We initialize the persistent factor ν_1 by drawing $\nu_{1,0} \sim N(0, \frac{-1}{2\alpha_1})$ from its unconditional distribution. The ν_2 factor is initialized at 0. The intraday volatility pattern is modeled by means of the diurnal U-shape σ_{ut} function. Following Andersen, Dobrev, and Schaumburg (2012), we set $A = 0.75$, $B = 0.25$, $C = 0.88929198$, and $a = b = 10$, respectively. The simulations are generated using an Euler scheme based on 23,400 intervals for each of the $T = 2{,}000$ "days" in the sample. We then aggregate these prices to sparsely sampled $M = 39, 78, 390$ return observations per day, corresponding to 10-, 5- and 1-"minute" returns.

To allow for empirically more realistic high-frequency prices, we further add "noise" to the simulated price process. In line with the empirical evidence in Bandi and Russell (2006) and Hansen and Lunde (2006), we allow the variance of the noise to increase with the volatility of the simulated efficient price. In particular, mirroring the design in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008), on each day we generate an *i.i.d.* noise term $u_{t,i} \sim N(0, \omega_t^2)$ with $\omega_t^2 = \xi^2 \int_{t-1}^{t} \sigma_{us}^2 \nu_s^2 ds$, so that the variance of the noise is constant throughout the day, but changes from day to day. This noise is then added to the $S_{t,i}$ price process to obtain the time series of actual high-frequency simulate prices $S_{t,i}^* = S_{t,i} + u_{t,i}$.