Bespoke Realized Volatility: Tailored Measures of Risk for Volatility Prediction

This version: April 22, 2024

Andrew J. Patton[a,*], Haozhe Zhang[b]

[a] *Department of Economics, Duke University*
[b] *Two Sigma Investments, LP*

Abstract

Standard realized volatility (RV) measures estimate the latent volatility of an asset price using high frequency data with no reference to how or where the estimate will subsequently be used. This paper presents methods for "tailoring" the estimate of volatility to the application in which it will be used. For example, if the volatility measure will be used in a specific parametric forecasting model, it may be possible to exploit that information and construct a better measure of volatility. We use methods from machine learning to estimate optimal "bespoke" RVs for heterogeneous autoregressive (HAR) and GARCH-X forecasting applications. We apply the methods to 886 U.S. stock returns and find that bespoke RVs significantly improve out-of-sample forecast performance. We find that, across a variety of volatility models, the bespoke RV places more weight on data from the end of the trade day, and that the optimal bespoke weights can be well-approximated by a simple parametric function.

Keywords: Volatility forecasting, Machine learning, High frequency data
JEL: C22, C51, C53, C58

[⋆] For helpful comments we thank Tim Bollerslev, Anna Bykhovskaya, Adam Clements, Peter Hansen, Nikolaus Hautsch, Ole Maneesoonthorn, Jia Li, Michael Smith, Jun Yu, and seminar participants at Chicago Booth, Duke, Macquarie University, Singapore Management University, University of Melbourne, Midwest Econometrics Group, University of Sydney Time Series and Forecasting Symposium, Virtual Time Series Seminar, and Advances in Financial Econometrics: A Conference in Honor of Torben Andersen at Copenhagen Business School. The first author also thanks UNSW Sydney, where part of this work was completed, for their hospitality. The views expressed herein are solely the views of the authors and are not necessarily the views of Two Sigma Investments, LP or any of its affiliates. They are not intended to provide, and should not be relied upon for, investment advice.

[*] Corresponding author: Department of Economics, Duke University, 213 Social Sciences Building, Durham NC 27708-0097, USA. andrew.patton@duke.edu

1. Introduction

Discussing the empirical success of models based on high-frequency measures of volatility, Andersen, Bollerslev, Diebold and Labys (2003) make the point that *"[t]he essence of forecasting is quantification of the mapping from the past and present into the future. Hence, quite generally, superior estimates of present conditions translate into superior forecasts of the future."* In the subsequent two decades a large literature on high frequency financial econometrics has emerged, containing many studies confirming that using more accurate measures of volatility in a forecasting model indeed leads to better predictions.

Advances in the financial econometrics literature in the last two decades have produced measures of volatility that are more efficient (e.g. Ait-Sahalia, Mykland and Zhang, 2005; Bandi and Russell, 2008), robust to micro-structure noise (e.g. Zhang, Mykland and Aït-Sahalia, 2005; Barndorff-Nielsen, Hansen, Lunde and Shephard, 2008; Jacod, Li, Mykland, Podolskij and Vetter, 2009), and robust to jumps in the price process (e.g. Barndorff-Nielsen and Shephard, 2004; Mancini, 2009; Andersen, Dobrev and Schaumburg, 2012). These advances share the feature that they seek an improved measure of volatility for later use in a variety of unknown applications. We consider the construction of a volatility measure from a different perspective: Can a volatility measure be improved by tailoring it for its eventual use in a volatility forecasting model?

The distinction between an "all purpose" estimator of volatility and one that is tailored to a specific application mimics the distinction between supervised and unsupervised machine learning algorithms. In unsupervised learning, data are analyzed without the use of an outcome measure. This is analogous to the above-mentioned volatility measures, where we seek a good (somehow defined) measure of volatility that works in a variety of unknown future applications. In supervised learning the algorithm is tailored to the problem at hand. This paper focuses on the case that we know that the resulting measure will be used in a specific forecasting model, to predict a specific asset's volatility, with a specific forecast horizon. We exploit that information to obtain a "bespoke" measure of volatility for that application. In so doing, we may obtain a worse general-purpose estimator of volatility, but it is hoped that it is a better measure of volatility for the specific purpose of volatility prediction.

To make this idea concrete, consider the widely-used heterogeneous autoregressive (HAR) model of Corsi (2009):

$$\begin{aligned} RV_t &= \beta_0 + \beta_d RV_{t-1} + \beta_w \frac{1}{4}\sum_{j=2}^{5} RV_{t-j} + \beta_m \frac{1}{16}\sum_{j=6}^{21} RV_{t-j} + e_t & (1) \\ \text{where} \quad RV_{t-j} &\equiv \sum_{i=1}^{M} r_{i,t-j}^2 & (2) \end{aligned}$$

and $r_{i,t-j}$ is the i^{th} high frequency return on day $t-j$. In standard applications, RV_t is constructed as the sum of squared five-minute returns over day t, as in equation (2), which for stocks on the New York Stock Exchange means the sum of $M = 78$ such returns. Realized volatility can be shown to be consistent as the sampling interval shrinks to zero (Jacod, 2018), it is fully efficient under some regularity conditions (Jacod and Protter, 1998; Jacod, 2008), and works well in a variety of empirical applications (Liu, Patton and Sheppard, 2015). We study whether we can obtain better forecasts by altering the construction of realized volatility from that in equation (2) to exploit the knowledge that it will subsequently be used in the model in equation (1).[1]

We exploit recent advances in the estimation of deep neural networks (DNNs) to flexibly construct "bespoke" measures of volatility for use in a forecasting model. We find that being completely flexible in the construction leads to poor out-of-sample forecast performance, even when the tuning parameters of the estimation algorithm are carefully selected. However after imposing some economically motivated structure on the tailoring we obtain forecasts that significantly outperform benchmark forecasts.

Our empirical analysis uses high frequency data on all stocks that were ever a constituent of the S&P 500 index over the period January 1995 to December 2019, a total of 886 securities. In our main analysis we take the HAR model of Corsi (2009) as the predictive model, and in Section 4 we further consider the GARCH-X model (Engle,

[1] Importantly, we only consider tailoring the terms on the right-hand-side of equations like (1); we leave the target variable as standard RV, with the motivation that it is a good measure of the unknown true volatility.

2002) and refinements of the HAR model, such as the "continuous HAR" of Andersen, Bollerslev and Diebold (2007) and the "semi HAR" of Patton and Sheppard (2015). In all cases we find significant improvements in out-of-sample forecast performance when using RVs tailored to the application.

We investigate the sources of the predictive gains from using bespoke RVs and find two primary channels. First, we find that models using bespoke RVs place greater weight on more recent lags than those using standard RVs. As more recent data tends to be more useful for prediction, this makes models using bespoke RVs more responsive to news. Second, we find that the weights attached to intra-daily returns in the bespoke RV are different from flat (as they are for standard RV) and also different from a "time-of-day" pattern motivated by measurement error considerations. Instead, the weights are low at the start of the trade day, consistent with measurement error considerations, increase slowly until the middle of the day, and then sharply increase over the last two hours of the trade day. This pattern is consistent with an information channel: returns from the latter part of the day is closest to the returns that forms part of the target variable, and thus are particularly valuable for forecasting. We find evidence in support of this explanation via multi-step-ahead forecasts.

We consider an extension and a simplification of our main "bespoke RV" approach. Firstly, we extend the model to allow the bespoke weights to depend on the sign and size of the high frequency return, as well as the time of day. For this purpose we consider the model proposed in Chen and Ghysels (2011), as well as a more direct extension of our "one-dimensional" bespoke RV. We find that the benchmark HAR model of Corsi (2009) is significantly beaten by the model of Chen and Ghysels (2011), but both are significantly beaten by our bespoke RV HAR model. The "two-dimensional bespoke RV" extension yields the best performance of all models considered, but it does not significantly outperform the one-dimensional bespoke RV, revealing that the optimal function of high frequency returns is not different from the simple quadratic, as used in our main analysis.

Finally, we consider a simplification of our proposed approach, where we impose the quadratic function for high frequency returns, and approximate the optimal bespoke weights using a simple parametric function. The resulting model has only seven parameters, is very easy to estimate, and has forecast performance comparable to the best, computationally demanding, model.

Our work is related to the enormous literature on using high frequency data for volatility forecasting, as reviewed in Bollerslev, Engle and Nelson (1994), Poon and Granger (2003) and Andersen, Bollerslev, Christoffersen and Diebold (2006). It is also linked to the growing literature in applying machine learning methods in econometrics (e.g. Chernozhukov, Hansen and Spindler, 2015; Mullainathan and Spiess, 2017; Athey and Imbens, 2019) and finance (Gu, Kelly and Xiu, 2020; Freyberger, Neuhierl and Weber, 2020; Bianchi, Büchner and Tamoni, 2021; Patton and Weller, 2022). Within this growing machine learning in economics literature, our work is particularly related to applications of these methods for volatility forecasting (e.g. Bucci, 2020; Filipović and Khalilzadeh, 2021; Li and Tang, 2021; Christensen, Siggaard and Veliyev, 2022; Patton and Zhang, 2022; Reisenhofer, Bayer and Hautsch, 2022). Our study of high-frequency returns for predicting lower-frequency volatility also links our analysis to the "mixed data sampling" (MIDAS) models introduced by Ghysels, Santa-Clara and Valkanov (2004), and used for volatility forecasting by Ghysels, Santa-Clara and Valkanov (2006).

The rest of the paper is structured as follows: Section 2 introduces the bespoke RVs models in detail and draws connections to standard RV estimation and the standard HAR model. Section 3 presents the results on the out-of-sample forecasting performance of the competing models when applied to 886 U.S. equities. Section 4 presents results using alternative forecasting models, demonstrating the generalizability of bespoke RVs. Section 5 concludes. A supplemental appendix contains some additional details and results.

2. Constructing bespoke realized volatilities

The standard realized variance estimator, given in equation (2), can be shown (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2002; Andersen, Bollerslev, Diebold and Labys, 2001) to be consistent for the true latent quadratic variation of an asset price process.[2] This measure has also been found to be useful for forecasting future volatility, as it is a more accurate measure of current volatility than squared daily returns, as used in ARCH/GARCH models (Engle, 1982; Bollerslev, 1986).

A key feature of the usual RV is that it is an *equal-weighted* sum of the high frequency squared returns. In constructing our "bespoke RVs" we will relax this assumption and estimate the optimal weight attached to each high frequency return. When using five-minute returns on stocks traded on the New York Stock Exchange, we have 78 such returns to consider.

2.1. Bespoke RVs for HAR models

The HAR model can be interpreted as a autoregression of order 21 with parameter equality constraints to reduce the number of free parameters from 21 to three (plus the intercept). In the most flexible bespoke RV forecast, we relax both the equal weights in the construction of RV and the HAR parameter constraints on the AR(21) process to obtain:

$$\begin{aligned} RV_t &= \beta_0 + \sum_{j=1}^{21} \widetilde{RV}_{t-j}(\boldsymbol{\gamma}_j) + e_t && (3) \\ \text{where} \quad \widetilde{RV}_{t-j}(\boldsymbol{\gamma}_j) &\equiv \gamma_{i,j} r_{i,t-j}^2 && (4) \end{aligned}$$

This simple-looking model is very flexible, with a total of $1 + 21 \times 78 = 1,639$ free parameters. Being linear, it is possible to estimate this model via standard OLS, however with the sample sizes available in practice OLS unsurprisingly performs very poorly. Instead, we treat this model as a single-layer neural network model and estimate it using

[2] Extensions of standard RV that are robust to market microstructure effects and/or jumps are discussed in the introduction.

methods from machine learning. We describe the estimation method in detail in the Section 2.3.

We next consider a hybrid between the fully flexible model in equation (3) and the restrictive standard HAR: we impose the constraint that the "daily," "weekly," and "monthly" lags in the model satisfy the parameter equality constraints in the HAR structure, but we allow each of these terms to be flexible functions of the underlying high-frequency returns:

$$RV_t \;=\; \beta_0 + \widetilde{RV}_{t-1}(\boldsymbol{\gamma}_d) + \frac{1}{4}\sum_{j=2}^{5}\widetilde{RV}_{t-j}(\boldsymbol{\gamma}_w) + \frac{1}{16}\sum_{j=6}^{21}\widetilde{RV}_{t-j}(\boldsymbol{\gamma}_m) + e_t \qquad (5)$$

This model has "only" $1 + 3 \times 78 = 235$ parameters and is thus much more parsimonious than the fully flexible specification. Nevertheless, this remains a large number of parameters, and we consider a variety of methods for regularizing them in estimation.

The final bespoke RV we consider is one that imposes some smoothness on the weights attached to the high frequency returns. Research on the empirical characteristics of intra-daily asset returns (see, e.g., Wood, McInish and Ord, 1985; Harris, 1986; Andersen and Bollerslev, 1998) shows that market conditions vary over the trade day, but generally smoothly. We impose this smoothness by using a cubic spline (see Judd, 1998, for further details on this interpolation method) for the intra-day weights:

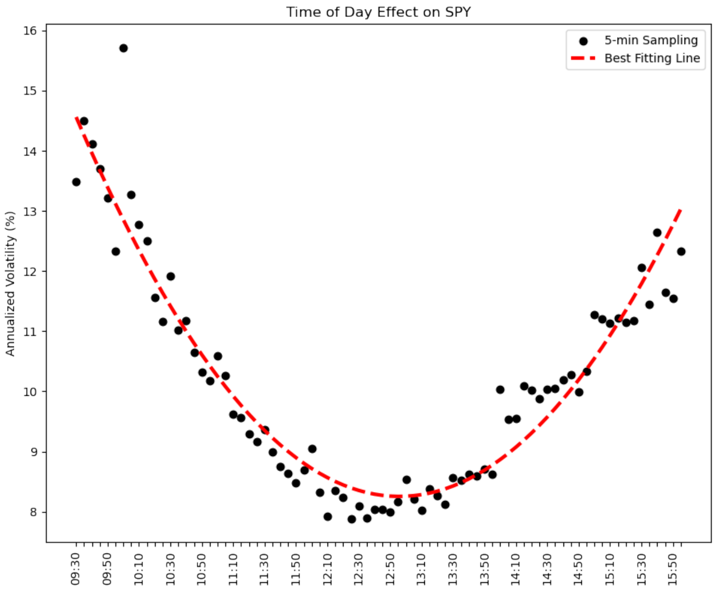
$$RV_t \;=\; \beta_0 + \widetilde{RV}_{t-1}(g(\mathbf{c}_d,\tau)) + \frac{1}{4}\sum_{j=2}^{5}\widetilde{RV}_{t-j}(g(\mathbf{c}_w,\tau)) + \frac{1}{16}\sum_{j=6}^{21}\widetilde{RV}_{t-j}(g(\mathbf{c}_m,\tau)) + e_t \quad (6)$$

where $g(\mathbf{c},\tau)$ returns a 78×1 vector of weights based on a cubic spline with knots given by τ. With knots every hour or half-hour, the parameter vector $\mathbf{c}$ is of length 7 or 13, leading to a total of either 22 or 40 free parameters.

2.2. A time-of-day adjusted RV

We consider one additional estimator, lying in between standard RV and bespoke RV, motivated by two well-known features of volatility. Firstly, volatility has a prounounced diurnal pattern, with volatility being highest at the open and close of the trade day, and

Figure 1: Time-of-day effect on SPY



Note: This figure illustrates the time-of-day pattern for the SPY exchange traded fund from 1995 to 2019. The black dots are the average 5-min sample intraday volatility (annualized) and the dashed red line is the best fitting line.

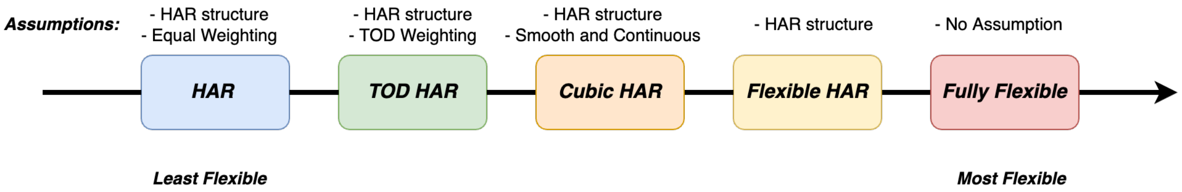
lowest around the middle, see Andersen and Bollerslev (1997, 1998) for example. Figure 1 confirms this pattern for the SPY in our sample period. Secondly, the estimation error in sample variance is increasing with the level of the variance. For example, in a simple i.i.d. setting, the asymptotic variance of the sample variance is proportional to the true variance squared:

$$\sqrt{T}(\hat{\sigma}_T^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$$

Combining the diurnal pattern in volatility with the fact that the accuracy of volatility changes with its level suggests a simple alternative to standard RV: a “time-of-day” adjusted RV, where we use the inverse of the average level of volatility as a weight function:

$$\begin{aligned} RV_t^{TOD} &= \sum_{i=1}^{M} \omega_i r_{i,t}^2 && (7) \\ \text{where} \quad \omega_i^{-1} &= \frac{1}{T}\sum_{t=1}^{T} r_{i,t}^2 && (8) \end{aligned}$$

Figure 2: Forecasting models, from simplest to most complex



Note: This figure illustrates the relationship among the five models considered in the out-of-sample analysis.

To avoid contaminating our out-of-sample comparisons, we estimate the RV_t^{TOD} weights, ω_i, using only the training sample.

Using RV_t^{TOD} in the familiar HAR specification yields:

$$RV_t = \beta_0 + \beta_d RV_{t-1}^{TOD} + \beta_w \frac{1}{4}\sum_{j=2}^{5} RV_{t-j}^{TOD} + \beta_m \frac{1}{16}\sum_{j=6}^{21} RV_{t-j}^{TOD} + e_t \tag{9}$$

Note that RV_t^{TOD} is *not* a bespoke RV; it is not customized for a specific forecasting model, rather it is a simple alternative to equal-weighting intra-daily returns, motivated by measurement error considerations. We consider this volatility measure as another benchmark that our proposed bespoke RVs must beat in order to be considered a success.

Figure 2 summarizes the models that will be considered in our out-of-sample analysis in the next section. These range from the benchmark HAR model using standard RV as a predictor variable as the simplest specification, to the "fully flexible" model that imposes the fewest constraints.

2.3. Estimation of bespoke RVs

The specifications in equations (3) and (5) are both linear and could easily be estimated via OLS, but this is unlikely to result in good out of sample performance (and we demonstrate this empirically below) as the model is severely over-parameterized. The specification in equation (6) is nonlinear, and requires a numerical optimization method. We estimate all of these models using the gradient descent approach, described below, designed for training deep neural networks. Such models are very flexible and users of these models are aware of the need for methods to tame the over-fitting problem. As is common in the machine learning literature, we split our sample period into "training," "validation," and "testing" samples to estimate the models, select the hyperparameters, and compute out-of-sample forecasts.

2.3.1. Stochastic gradient descent methods

We adopt the mini-Batch stochastic gradient descent algorithm (Robbins and Monro, 1951; Bilmes, Asanovic, Chin and Demmel, 1997) with a gradually decreasing of learning rate. See Goodfellow, Bengio and Courville (2016) for an accessible discussion of this optimization method and related methods. We estimate the parameters using the following algorithm:

1. Initialize the model with HAR model coefficients estimated on the training sample.[3]
2. For each combination of hyperparameters, described below, estimate the model in the training sample.
3. Evaluate predictive performance on validation sample, and select the hyperparameters that lead to the best forecasting performance.
4. Fix the hyperparameters at their optimal values, and estimate the model parameters on the entire in-sample data (training and validation samples combined).
5. Evaluate the predictions on the testing sample (the out-of-sample period).

There are a total of five hyperparameters in the optimization algorithm for the "fully flexible" and "flexible HAR" methods, and six for the "cubic HAR" method:

- $LearningRate \in \{0.1, 0.01\}$. The learning rate controls the step size of each mini-Batch gradient update in the optimization algorithm.

- $BatchSize \in \{512, 2048\}$. The batch size controls how many observations we use to compute the gradient direction for each update in the optimization algorithm.

- $NumberofEpochs \in \{50, 100, 250\}$. This controls the number of times that the algorithm works through the training data set to compute the gradients for the update.

[3] We also considered using many random starting and then ensembling the resulting estimates, as is common in the machine learning literature, and the results are qualitatively the same.

- $StepSize \in \{0.1, 0.5\}$. The step size controls how frequently we reduce the learning rate (if at all). The numbers are as a fraction of the number of epochs. For example, if the step size is 0.1 and the the number of epochs is 100, then we reduce the learning rate by gamma (described below) after every $0.1 \times 100 = 10$ epochs.

- $Gamma \in \{0.1, 0.25, 1\}$. Gamma controls how much to reduce the learning rate. For example, if the learning rate is 0.1 and gamma is 0.25, then when it is time to reduce learning rate, the updated learning rate will be $0.1 \times 0.25 = 0.025$.

- $NumNodes \in \{7, 13\}$. This parameter controls how flexible the spline function is.

We tune the hyperparameters separately for each stock in our analysis. Section S.9 in the supplemental appendix reports summary statistics on how the optimal hyperparameters vary across stocks.

In estimating the cubic spline model, we need to implement a cubic spline layer, similar to the linear and convolution layers in PyTorch, the Python package we use for estimation, where the layer can initialize values for the base points (K) and generate cubic spline interpolations for the final desired number of points (M). The parameters for the cubic spline layer are simply the K initialized base points. We then use the standard back-propagation and mini-Batch stochastic gradient descent framework, where we gradually find the optimal parameter values by iterating through all the batches and epochs.

In Figure S.2 we present the model architecture for the cubic spline model. (The architecture for the other two models is simliar, but without the second and third layers.) We initialize cubic spline interpolating nodes for daily, weekly and monthly lags separately, and then use the cubic spline layer to generate the initial bespoke weights. Then we combine the initial bespoke weights with the lagged high frequency returns squares and use a linear layer with a constant to construct the forecast. After this, we optimize the interpolating nodes and the constant term through the miniBatch gradient descent and reach the optimal bespoke weights for the cubic HAR model.

2.3.2. Regularized regression methods

In addition to the stochastic gradient descent methods described above, we also consider more familiar regularization methods for the “flexible HAR” model in equation (5), which has 235 free parameters. Denoting the usual, non-penalized, objective function as $L_T(\boldsymbol{\beta})$, we consider the penalized loss:

$$\bar{L}_T(\boldsymbol{\beta};\alpha,\lambda) = L_T(\boldsymbol{\beta}) + \alpha(\lambda\|\boldsymbol{\beta}\|_1 + (1-\lambda)\|\boldsymbol{\beta}\|_2)$$

This formulation allows us to nest four standard shrinkage methods: No penalty ($\alpha = 0$), ridge regression ($\lambda = 0,\ \alpha \geq 0$), LASSO ($\lambda = 1,\ \alpha \geq 0$), and elastic net ($\lambda \in [0,1],\ \alpha \geq 0$). We consider the following values for these hyperparameters:

$$\begin{aligned} \alpha &\in 0, 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5, 10, 50, 100, 1000 \\ \lambda &\in 0, 0.25, 0.5, 0.75, 1 \end{aligned}$$

3. Out-of-sample forecast performance of bespoke realized volatilities

3.1. Data description

Our empirical analysis is based on high frequency stock prices from the Trades and Quotes (TAQ) database, spanning the period from January 1995 to December 2019, a total of 6,293 days. We include every stock that was ever a constituent of the S&P 500 index during this period, and we follow Barndorff-Nielsen, Hansen, Lunde and Shephard (2009) for the data cleaning process, retaining only stocks with at least 2,000 observations in the sample period. This yields a total of 886 different stocks for our analysis. Following Liu, Patton and Sheppard (2015), we use five-minute sampling for our high frequency returns throughout.

We follow common practice in the machine learning literature (see, e.g., Christensen, Siggaard and Veliyev, 2022) and split the available sample period for a given stock into a training sample (first 60% of data), a validation sample for choosing hyperparameters (next 20%), and a test sample for out-of-sample comparisons (final 20%). The first 80%

of the sample is the full in-sample period. For models that do not involve any hyperparameters search we simply estimate the models on the in-sample period and evaluate on the test sample. For models that involve hyperparameters, we estimate the models on the training sample, and select the hyperparameters based on the validation sample performance. We then we re-estimate the models using the optimal hyperparameters on the full in-sample period, and evaluate on the out-of-sample period.

3.2. Optimal weights for bespoke RVs

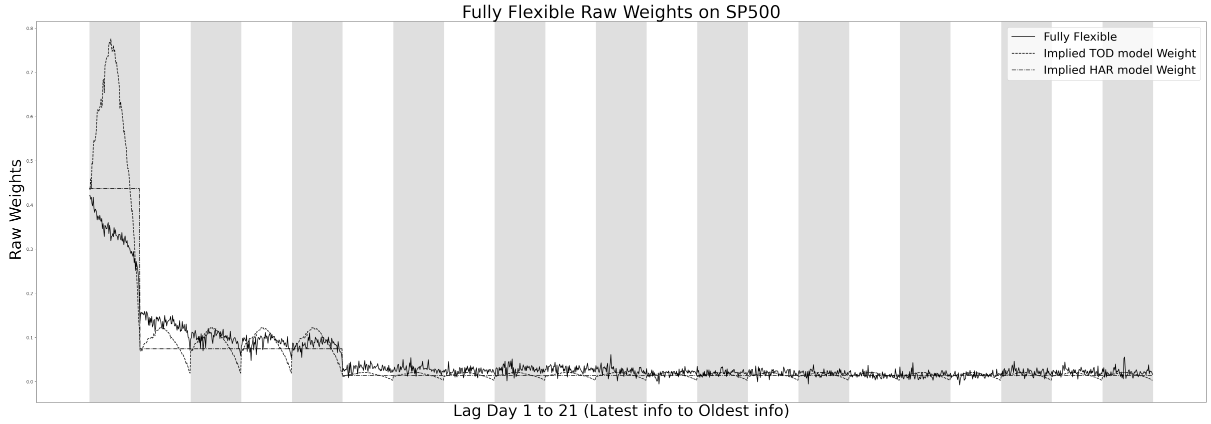
This section presents the optimal bespoke RV weights across the three degrees of tailoring that we consider (fully flexible, flexible HAR, and cubic HAR). In Figure 3 we present the bespoke weights implied by the "fully flexible" model, averaged across all 886 stocks in our sample, and for comparison we also present the weights implied by the standard HAR (which appear as a step function) and the TOD-HAR (which appear as an inverse-U layered on a step function).

The weights for the fully-flexible bespoke RV are similar to the weights for standard RV for daily lags 7 to 21, while they are lower for the first daily lag, and slightly above for lags 2 through 6. For all lags, the estimated weights appear to have non-negligible estimation error, which is unsurprising given that each of these 1,638 weights are freely estimated. Foreshadowing our forecast comparison results, the estimation error in these weights make the fully flexible model perform significantly worse than the benchmark HAR forecasts out of sample.

Figure 4 presents the optimal bespoke weights for the "flexible HAR" model, which imposes that RVs lagged 2 through 5 periods share the same "weekly" weight function, and the RVs lagged 6 through 21 share the same "monthly" weight function, while the first lag gets its own "daily" weight function. We find that the daily lag weights, despite still being somewhat noisy, clearly display an up-weighting the end of day information. For the weekly and monthly lags weights, we find them again being somewhat noisy, but roughly showing a flat shape with slight down-weighting at the beginning of the trade day.

Finally, we present the bespoke RV weights when we impose smoothness across the

Figure 3: Cross-sectional average optimal fully flexible weights



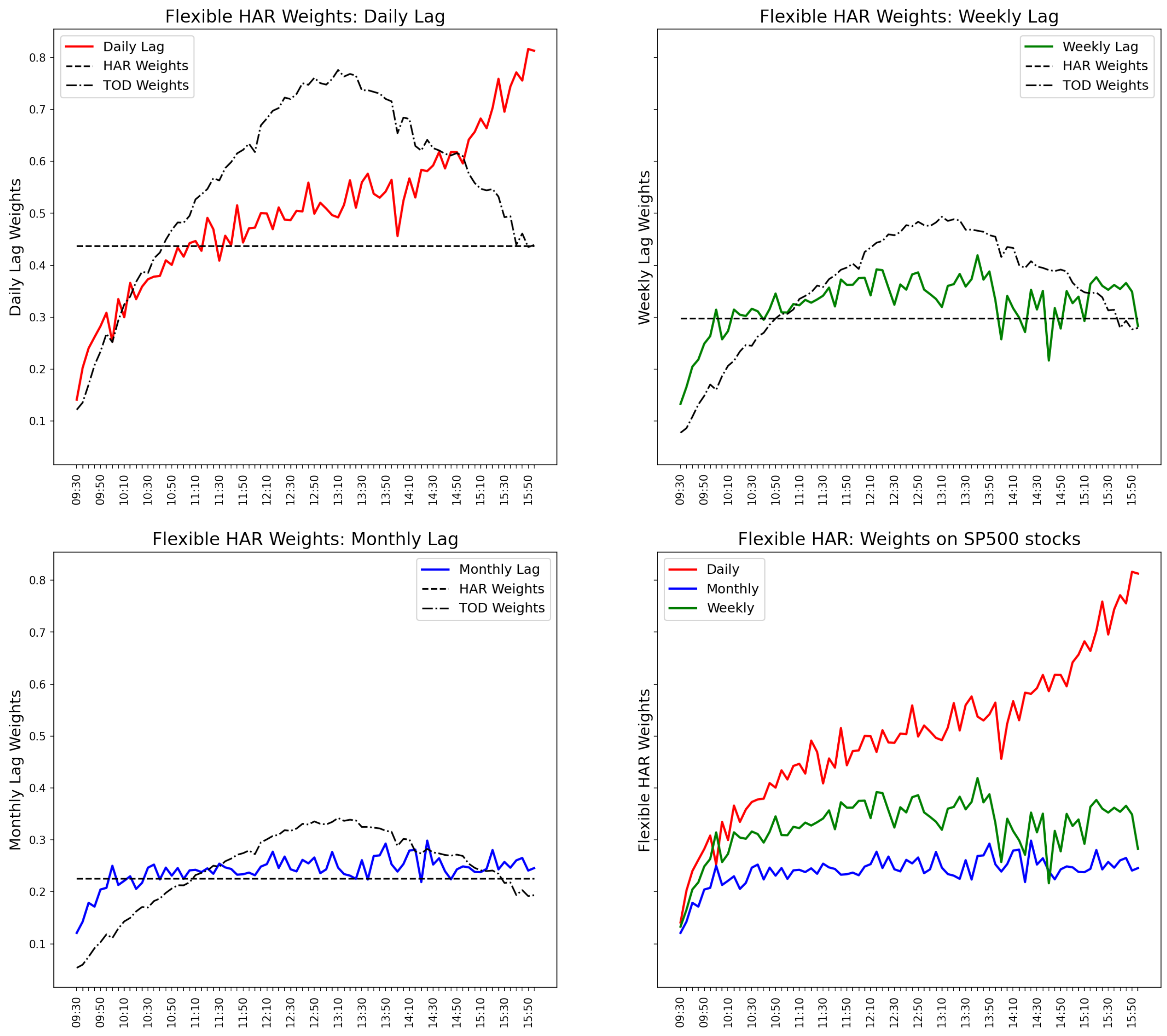
Note: This figure depicts the cross-sectional average optimal weights implied by the "fully flexible" model, along with the equal-weighting scheme and the time-of-day weighting scheme. The x-axis runs from the most-recent high-frequency return to least recent, and each of the 21 days is marked by a gray or white region.

trade day using a cubic spline. Figure 5 presents the optimal bespoke weights for the daily, weekly and monthly weights in the cross section of S&P500. The most prominent feature of the optimal weights is the strong increase in weights in the daily lag towards the end of the trade day. We investigate the source of this feature in Section 3.4 below. It is also noteworthy that the cubic HAR daily lag weights roughly track the TOD weights until lunch time, but are much higher at the end of day.The weekly and monthly weights are broadly similar to each other, and hard to distinguish from either flat or TOD weights. Rather than estimate confidence intervals for these estimated weights, which is made more difficult due to our use of regularized estimation, we instead use out-of-sample forecast performance to determine whether these weights are indeed different from either of these benchmarks. This approach is consistent with the paper's focus on forecast performance of competing models and estimation methods.

3.3. Comparing out-of-sample forecast performance

We now present results comparing the forecast performance of the models using bespoke RVs with the benchmark HAR model using standard RV, as well as the HAR model using the time-of-day weighted RV. For our main analysis we measure forecast accuracy

Figure 4: Cross-sectional average optimal flexible HAR weights



Note: This figure depicts the cross sectional average optimal weights implied by Flexible HAR model, along with its comparisons with the equal-weighting scheme and the time-of-day weighting scheme in the S&P500 cross section. The upper left corner depicts the daily lag weights, upper right is the weekly lag weights, the lower left is the monthly lag weights, and the lower right presents all three for ease of comparison.

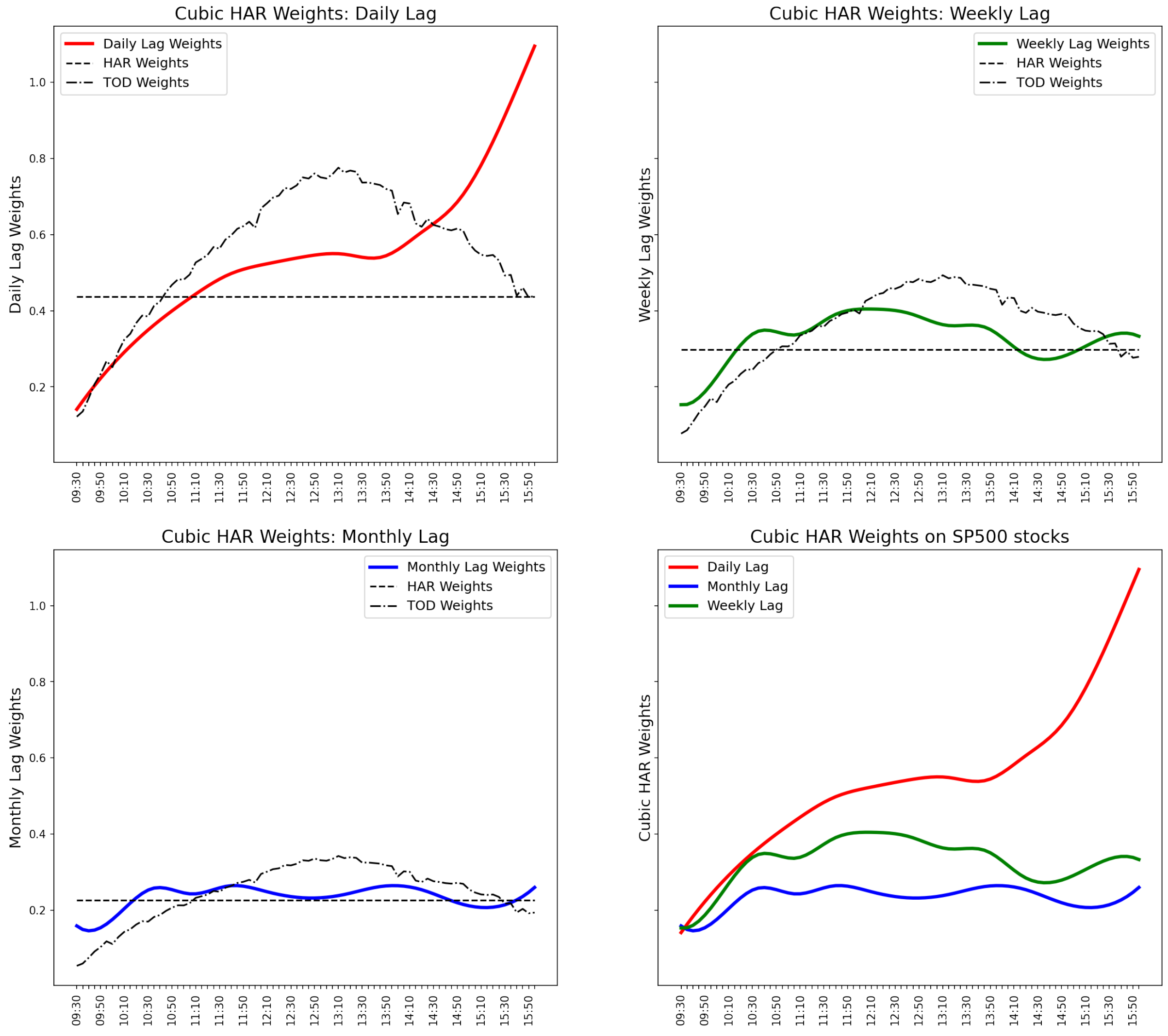
using the QLIKE loss function:

$$L_{QLIKE}(RV,\widehat{RV}) = RV/\widehat{RV} - \log RV/\widehat{RV} - 1 \tag{10}$$

We report corresponding results using the quadratic loss function, $L_{MSE}(RV,\widehat{RV}) = (RV - \widehat{RV})^2$, in the supplemental appendix.[4]

[4] Consistent with the power analyses in Patton and Sheppard (2009), the rankings using quadratic loss are similar to those using QLIKE but the significance of the loss differences is generally weaker.

Figure 5: Cross-sectional average optimal cubic HAR weights



Note: This figure depicts the cross sectional average optimal weights implied by the Cubic HAR regressions, along with its comparisons with the equal-weighting scheme and the time-of-day weighting scheme in the S&P500 cross section. The upper left corner depicts the daily lag weights, upper right is the weekly lag weights, the lower left is the monthly lag weights, and the lower right presents all three for ease of comparison.

We compare the predictive accuracy of competing models using Giacomini-White (2006) tests for each individual asset in our sample, and a panel GW test for all stocks jointly.[5] GW tests are ideal for comparing forecasts from models estimated with regularization methods, such as stochastic gradient descent, LASSO, and ridge regression, as the estimation method is explicitly considered part of the forecasting method, along with the

[5] For the individual GW tests we use Newey and West (1987) standard errors allowing for autocorrelation up to 10 lags. For the panel GW tests we additionally cluster by stock.

forecasting model and the sample period used for estimation. The same model estimated using a different method (e.g., with different regularization techniques) will produce different forecasts, and the performance of the resulting forecasts can be formally compared via the GW approach.

Table 1 presents forecast comparison results comparing the baseline HAR model with various competing methods. The first four rows present comparisons with the methods presented in Figure 2, namely the TOD-HAR and three bespoke HAR models, while the bottom four rows present comparisons with different, more familiar, methods for estimating the "flexible HAR" model.

The first row of Table 1 compares the baseline HAR model with a fully flexible bespoke HAR models. We see that the baseline model outperforms the fully flexible model in 714 out of 886 individual comparisons, losing in only 172, and of those 714 "wins" 321 are statistically significant at the 5% level. Pooling the individual stocks and conducting the comparison jointly, the last column reports a panel Giacomini-White (2006) t-statistic of -9.3, which is strong evidence that the baseline HAR model outperforms the fully flexible model overall. These comparisons lead to the conclusion that fully flexible model is worse than the simple, familiar, and parsimonious HAR model.

The second row of Table 1 compares the baseline HAR with the "flexible HAR," which imposes the HAR structure on the lag parameters, but allows bespoke weights on each of the daily, weekly, and monthly lags. We see that this model performs much better than the fully flexible model: it out-performs the baseline HAR for 677 out of 886 stocks, and has a panel GW statistic of 4.5, indicating strongly significant out-performance. The third row of Table 1 imposes more structure on the bespoke RV, using a cubic spline to ensure that the bespoke weights are smooth through the trade day. We see that this improves forecast performance even further, with a panel GW statistic of 21.7. Thus imposing some economically-motivated structure on the bespoke RV leads us to a model that out-performs the baseline model by an even greater margin than that by which the fully flexible model under-performs.[6]

[6] The form of the QLIKE loss function mitigates the right-skewness of RV, which can cause problems

Table 1: Forecast performance of HAR vs other models for 886 S&P 500 stocks

	GW Losses		**GW Wins**		**GW t-stat**
HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	172	53	714	321	-9.3
Flexible HAR	677	430	209	38	4.5
Cubic HAR	731	470	155	41	21.7
TOD HAR	680	458	206	63	18.6
Ridge	445	185	441	175	-2.3
LASSO	222	42	664	299	-6.1
Elastic net	454	182	432	174	-7.2
OLS	245	46	641	272	-5.7

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the baseline HAR model against competing models across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model out-performs the HAR model, while a negative t-statstic indicates the opposite.

Finally, the fourth row of Table 1 considers the non-bespoke, but computationally simple, TOD-HAR, where the realized variances are computing using time-of-day weights. We see that this model also significantly out-performs the baseline HAR, with a panel GW statistic of 18.6, and almost the same number of "wins" as the flexible (bespoke) HAR. These results suggest that this simple "off-the-rack" alternative RV is a significant improvement over standard equal-weighted RV for volatility forecasting.

The bottom four rows of Table 1 present alternative methods for regularizing the flexible HAR model parameters. Recall that the flexible HAR model has a total of 235 parameters, and all of them appear linearly, meaning that familiar methods like OLS and ridge regression can be employed in place of our preferred stochastic gradient descent (SGD) optimization method. We see, however, that all of these methods lead to significantly worse performance than the baseline HAR: the panel GW t-statistics are

when using OLS and MSE for evaluation. Some researchers choose to use OLS for the *logarithm* of RV, which is more symmetrically distributed, though this transformation must be undone if the target is truly the level of volatility. We estimated the HAR and cubic HAR models on logRV and find that the panel GW t-statistic is -7.2 in favor of cubic HAR. In individual tests, "log cubic HAR" beats (is beaten by) "log HAR" in 738 (148) out of the 886 stocks, with 563 (59) of these significant at the 5% level. Thus the outperformance we observe using QLIKE on levels of RV in Table 1 is very similar to what is obtained using MSE on log RV.

Table 2: Forecast performance of Cubic HAR vs other models for 886 S&P 500 stocks

	GW Losses		**GW Wins**		**GW t-stat**
Cubic HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	106	39	780	571	-10.8
Flexible HAR	277	72	609	204	-12.8
TOD HAR	358	95	528	205	-2.9
HAR	155	41	731	470	-21.7
Ridge	114	32	772	434	-3.3
LASSO	48	15	838	595	-6.9
Elastic net	118	26	768	433	-10.2
OLS	54	17	832	564	-6.4

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the cubic HAR model against competing models across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite.

all less than -2, while the flexible HAR model estimated using SGD significantly *out*-performs the baseline model.[7] This is evidence that the choice of regularization method can have a large impact on the results: in our application, SGD is significantly better than ridge, lasso, and elastic net. This is related to recent work by Shen and Xiu (2024) on regularization methods for machine learning.

From the comparisons with the baseline model in Table 1, it appears that the cubic HAR is the best-performing model, but strictly those comparisons do not guarantee this interpretation is correct. Table 2 changes the reference model to the cubic HAR and compares all of the other methods to it. We see that the panel GW t-statistic is uniformly below -2, and indeed in several comparisons it is much smaller than that, confirming that the cubic HAR model is indeed the best-performing model on average.

We next consider the gains from bespoke RV for multi-step ahead volatility forecasting. Table 3 presents results for forecast horizons ranging from one day to 60 days. In

[7] In Section S.3 of the supplemental appendix we consider shrinking the estimated OLS parameters towards the benchmark HAR parameters rather than towards zero, as is done here. We find that the optimal degree of shrinkage for this design is large, and the estimated parameters are shrunk almost all the way towards the original HAR parameters. The cubic HAR model is shown to significantly out-perform these alternative shrinkage estimators as well.

Panel A we see that using bespoke RV significantly improves forecast accuracy for all horizons, with the panel GW t-statistic less than -5 in all cases. The improvement is generally declining with the horizon, with the t-statistics decreasing in magnitude, and the proportion of "wins" in individual comparisons decreasing as well. Nevertheless, these results represent strong evidence that "bespoke" RVs are preferred to standard RVs, when employed in the HAR model, even for relatively long forecast horizons. In the next section we examine the optimal weights across various horizons to understand the source of this outperformance.

In Panel B of Table 3 we compare the bespoke RV to the simple TOD-RV. Table 1 revealed that TOD-RV is significantly better, on average, than standard RV and so this is a much tougher competitor. We find statistically significant gains from using bespoke RV in around half of the horizons considered, and in only one horizon is the ranking reversed, though in that case the difference in forecast performance is not significant. This panel thus also confirms that bespoke RVs provide important improvements in forecast performance across a range of forecast horizons.

3.4. Understanding the optimal bespoke weights

In this section we seek to understand the sources of the out-performance of models using bespoke RV relative to the benchmark methods. We first investigate whether by tailoring the measure of risk to the forecasting problem at hand the model can place greater weight on the risk measure and thus react to news more quickly. To measure this, we compute the average effective regression coefficients for the daily, weekly, and monthly lagged RVs.[8] Table 4 presents the results, and shows that the bespoke RVs are more responsive than the standard RVs. The standard HAR has an average coefficient on daily lagged RV of 0.438, while the bespoke cubic HAR has an average coefficient of 0.466. In the other direction, we see that the poor-performing fully flexible HAR has an average daily coefficient of only 0.312. When the sum of the coefficients is less than one, we can interpret that sum as the weight on lagged information, and the difference

[8] We adjust the regression weights to ensure that the bespoke RVs and the standard RVs have the same unconditional mean during the in-sample period. This makes comparing coefficient magnitudes meaningful.

Table 3: Multi-day ahead volatility forecasting

Horizon (days)	GW Losses Total	GW Losses Signif	GW Wins Total	GW Wins Signif	GW t-stat Panel
Panel A: Cubic vs HAR					
1	155	41	731	470	-21.7
2	194	31	692	366	-11.6
3	222	38	664	325	-10.5
4	247	41	639	285	-8.8
5	250	36	636	265	-18.6
20	348	81	538	248	-8.6
60	413	154	473	223	-5.4
Panel B: Cubic vs TOD HAR					
1	358	95	528	205	-2.9
2	449	120	437	146	-1.2
3	487	120	399	136	-2.6
4	496	129	390	133	1.0
5	500	131	386	124	-4.6
20	484	160	402	143	-0.3
60	458	175	428	189	-4.9

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the cubic HAR model against the HAR (Panel A) and TOD-HAR (Panel B) models for 886 S&P 500 stocks, for various forecast horizons. A positive panel GW t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite.

from one as the effective weight on the unconditional average (the intercept). This too lines up with the relative forecast performances documented in the previous subsection, and is consistent with bespoke RVs (appropriately disciplined) providing more responsive forecasts.

We next seek to understand the reason cubic HAR weights take the shape that they do. These weights were presented above in Figure 5 and present two key questions. Firstly, why do the weights on the daily lag rise in the afternoon? Secondly, are the weights on the weekly and monthly lags significantly different from either the flat weights from the standard RV or the time-of-day (TOD) weights?

We conjecture that the rising weights in the daily lagged RV come from the proximity of the afternoon to the target day. That is, returns realized in the afternoon are the

Table 4: Average regression coefficients for different models

	Daily	**Weekly**	**Monthly**	**Sum**
HAR	0.438	0.298	0.225	0.960
TOD HAR	0.465	0.295	0.204	0.964
Cubic HAR	0.466	0.294	0.217	0.978
Flexible HAR	0.442	0.298	0.224	0.963
Fully Flexible	0.312	0.384	0.346	1.041

Note: This table presents the average, across 886 S&P 500 stocks, coefficients on daily, weekly and monthly lagged RVs.

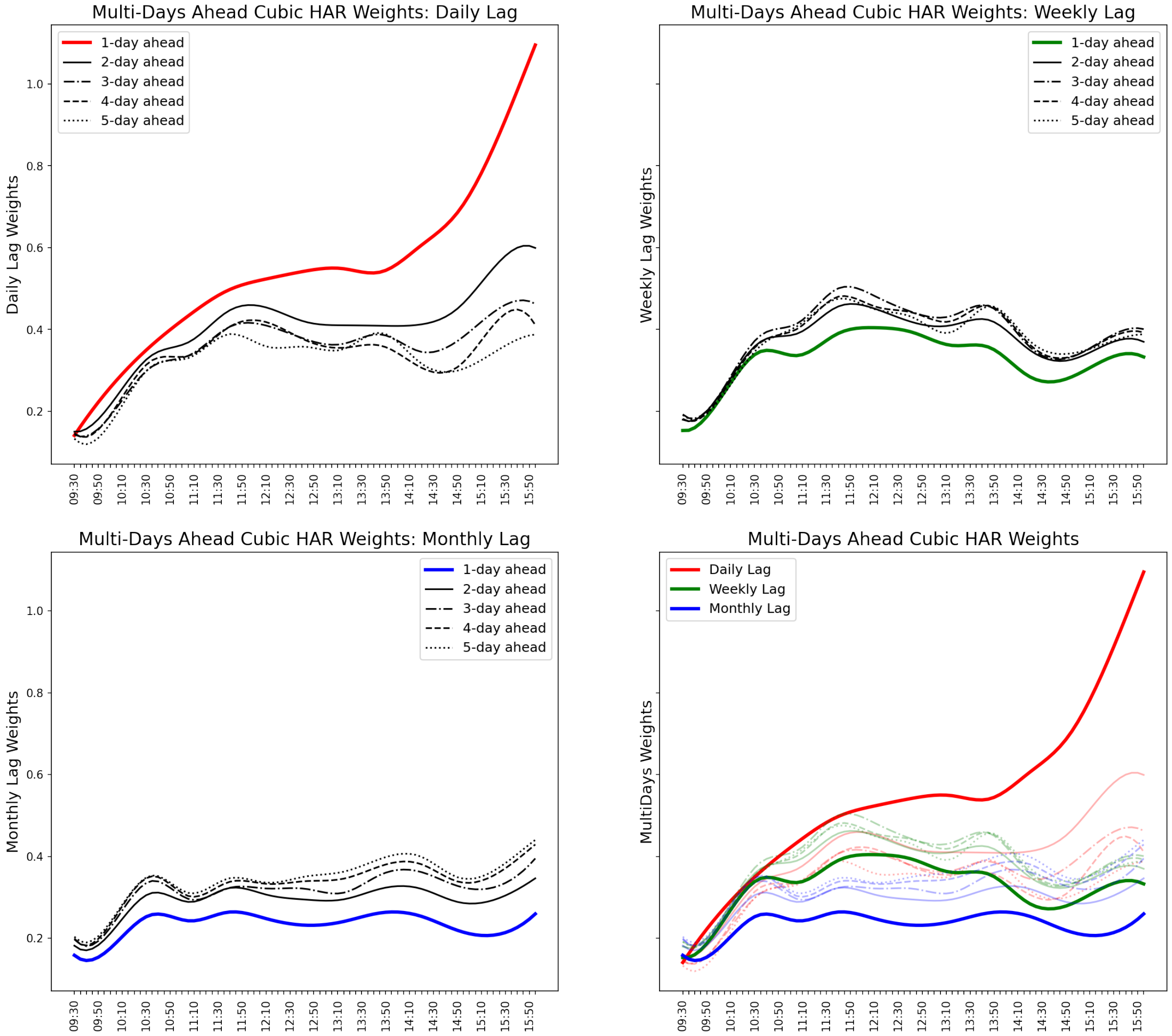
closest to, and presumably the most informative for, the returns that arise during the following day. We test this conjecture by examining the weights on the daily lag returns for longer forecast horizons. If the conjecture is correct, then the weights on the daily lagged RV should rise by less in the afternoon for longer forecast horizons. Figure 6 presents the estimated weights for forecast horizons ranging from one to five days, and we see that the weights on afternoon returns are monotonically declining as the horizon increases, consistent with this being an information effect. The weights on weekly and monthly lagged RVs increase slightly with the forecast horizon, offsetting the declining weight given to the daily lagged RVs.

Next we seek to determine whether the weights on the weekly and monthly lags significantly different from either the flat weights from the standard RV or the time-of-day (TOD) weights. We do this via an out-of-sample forecast comparison of the cubic HAR with a hybrid model that estimates the weights on the lagged daily data but imposes either equal weights or TOD weights on the lagged weekly and monthly data. We estimate these hybrid "cubic-EW" and "cubic-TOD" models using the same SGD algorithm as the original cubic HAR model.

Table 5 shows that the hybrid cubic-TOD model significantly out-performs all competing models, according to the panel GW t-statistics, aside from the cubic-EW model. The t-statistic for that comparison is 1.1, indicating no significant difference in performance on average.[9] While this analysis does not allow us to determine whether the

[9] We informally break this statistical tie by considering the number of significant differences for indi-

Figure 6: Multi-day ahead cubic HAR weights



Multi-Days Ahead Cubic HAR Weights: Daily Lag

Multi-Days Ahead Cubic HAR Weights: Weekly Lag

Multi-Days Ahead Cubic HAR Weights: Monthly Lag

Multi-Days Ahead Cubic HAR Weights

Note: This figure depicts the cross sectional average optimal weights implied by the cubic HAR regressions for multi-days ahead forecasts in the S&P 500 cross section. Note that the upper left is the daily lag, upper right is the weekly lag, and the lower left is the monthly lag. Also the colored lines are representing weights for one day ahead forecasting.

Table 5: Forecast performance of Cubic-TOD vs other models on 886 S&P 500 stocks

	GW Losses		**GW Wins**		**GW t-stat**
Cubic-TOD vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	100	35	786	567	-10.9
Flexible HAR	281	69	605	228	-14.2
Cubic	402	130	484	150	-3.7
Cubic-EW	417	123	469	136	1.1
TOD HAR	318	74	568	219	-8.1
HAR	159	37	727	471	-21.0

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the cubic-TOD model against competing models across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model out-performs the cubic-TOD model, while a negative t-statstic indicates the opposite.

optimal weights for weekly and monthly lagged information are equal or TOD shaped, Table 5 does show that it is preferable to impose either of those shapes than to try to estimate them from data.

4. Additional analyses

In this section we consider extensions of the methods presented above. Firstly, we consider bespoke RV for models other volatility forecasting models. We consider the "continuous" HAR model of Andersen, Bollerslev and Diebold (2007), the "semi" HAR model of Patton and Sheppard (2015), and the GARCH-X model of Engle (2002). The former two models are refinements of the original HAR model, while the latter lies outside the HAR class of models.

We next consider two methods for obtaining a more flexible bespoke RV, where the weights are a function of both the time of day and also the return sign and size. The first of these is the model of Chen and Ghysels (2011), and the second is a more direct extension of our "one-dimensional" bespoke RV for the HAR model.

Finally, we consider a simplification of the our machine-learning based approach,

vidual stocks: we observe that cubic-TOD significantly out-performs cubic-EW for 136 stocks, compared with 123 for cubic-EW.

exploiting the fact that the estimated optimal weights appear to be simple functions of the time of day, and thus potentially amenable to a parametric approximation. We show that this simplified method performs comparably to the more complicated method introduced above, and requires only a fraction of the computational effort.

4.1. Bespoke RV for alternative forecasting models

The analysis in the previous section focused on tailoring realized variance (RV) measures for application in the heterogeneous autoregressive (HAR) model of Corsi (2009). This section shows that out-of-sample forecast performance can similarly be improved when tailoring RV measures for use in refinements of the original HAR model: the "continuous" HAR (CHAR) model of Andersen, Bollerslev and Diebold (2007), which decomposes volatility into continuous and jump components, and the "semi" HAR (SHAR) model of Patton and Sheppard (2015), which decomposes volatility into upside and downside movements.

The CHAR model of Andersen, Bollerslev and Diebold (2007) forecasts future realized volatility using only the continuous component of volatility, discarding the component coming from jumps, as that component is found to be nearly unpredictable. This model uses bi-power variation (BPV) (Barndorff-Nielsen and Shephard, 2004) to estimate the continuous component of volatility:

$$\begin{aligned} RV_t &= \beta_0 + \beta_d BPV_{t-1} + \beta_w \frac{1}{4}\sum_{j=2}^{5} BPV_{t-j} + \beta_m \frac{1}{16}\sum_{j=6}^{21} BPV_{t-j} + e_t \quad (11) \\ \text{where } BPV_t &= \mu_1^{-2}\sum_{i=1}^{M-1} |r_{t,i}||r_{t,i+1}|, \quad \text{and} \quad \mu_1 \equiv \sqrt{2/\pi} \end{aligned}$$

We tailor the measure of continuous volatility to the CHAR model by flexibly estimating the weights attached to each product $|r_{t,i}||r_{t,i+1}|$. Motivated by the results in Tables 2 and 5, we impose that the daily weights are smooth by using a cubic spline, and that the weekly and monthly weights use time-of-day (TOD) weights:

$$RV_t = \beta_0 + \widetilde{BPV}_{t-1}(\gamma) + \beta_w \frac{1}{4}\sum_{j=2}^{5} BPV_{t-j}^{TOD} + \beta_m \frac{1}{16}\sum_{j=6}^{21} BPV_{t-j}^{TOD} + e_t \qquad (12)$$

Table 6: Bespoke RV for CHAR and SHAR models

Model	GW Losses Total	 Signif	GW Wins Total	 Signif	GW t-stat Panel
CHAR: Basic vs Bespoke	120	38	766	606	-19.2
SHAR: Basic vs Bespoke	164	40	722	492	-11.1

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the CHAR and SHAR models with their bespoke counterparts, across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the original model out-performs the bespoke version, while a negative t-statstic indicates the opposite.

where $\widetilde{BPV}_t(\boldsymbol{\gamma}) = \sum_{i=1}^{M-1} \gamma_i |r_{t,i}||r_{t,i+1}|$, and γ_i comes from a cubic spline with hourly or half-hourly nodes. BPV_t^{TOD} is defined analogously to RV_t^{TOD} in equation (7).

Next we turn to the SHAR of Patton and Sheppard (2015), which decomposes realized variance into positive and negative realized semivariances (Barndorff-Nielsen, Kinnebrock and Shephard, 2010):

$$RV_t = \beta_0 + \beta_{d,p} RV_{t-1}^{+} + \beta_{d,n} RV_{t-1}^{-} + \beta_w \frac{1}{4}\sum_{j=2}^{5} RV_{t-j} + \beta_m \frac{1}{16}\sum_{j=6}^{21} RV_{t-j} + e_t \tag{13}$$

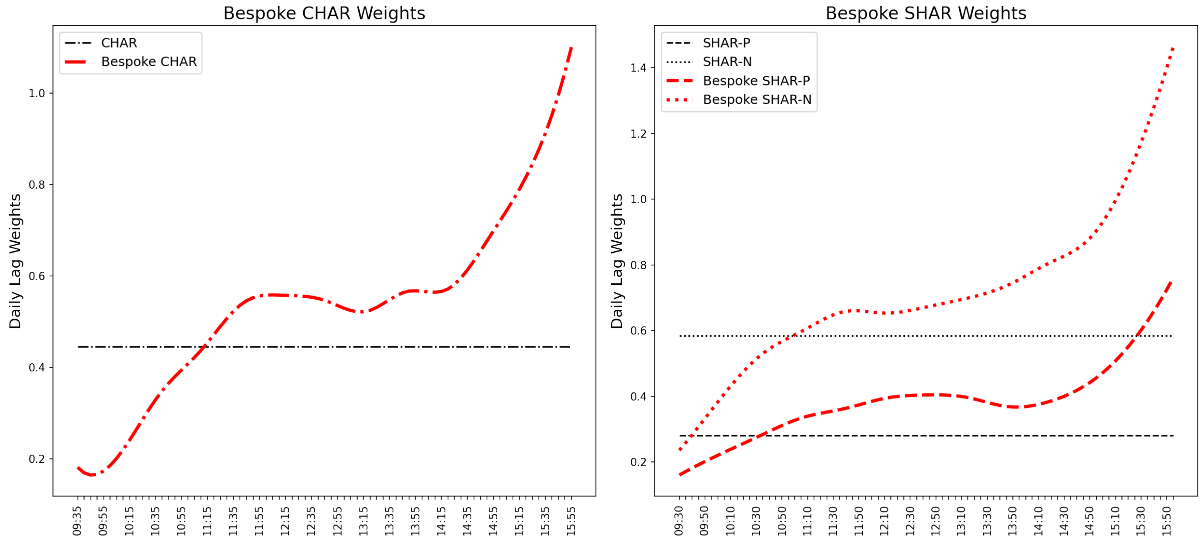
where $RV_t^{+} = \sum_{i=1}^{M} \max(0, r_{i,t})^2$ and $RV_t^{-} = \sum_{i=1}^{M} \min(0, r_{i,t})^2$ are the positive and negative realized semivariances. We tailor the measures of semivariance by using a cubic spline to aggregate the high-frequency positive and negative returns. As for the bespoke CHAR model, we again only estimate the weights for the daily lag, and impose TOD weights for the weekly and monthly lags:

$$RV_t = \beta_0 + \widetilde{RV}_{t-1}^{+}(\boldsymbol{\gamma}_p) + \widetilde{RV}_{t-1}^{-}(\boldsymbol{\gamma}_n) + \beta_m \frac{1}{4}\sum_{j=2}^{5} RV_{t-j}^{TOD} + \beta_w \frac{1}{16}\sum_{j=6}^{21} RV_{t-j}^{TOD} + e_t \tag{14}$$

where $\widetilde{RV}_t^{+}(\boldsymbol{\gamma}_p) = \sum_{i=1}^{M} \gamma_{i,p} \max(r_{t,i}, 0)^2$, $\widetilde{RV}_t^{-}(\boldsymbol{\gamma}_n) = \sum_{i=1}^{M} \gamma_{i,n} \min(r_{t,i}, 0)^2$, and $(\gamma_{i,p}, \gamma_{i,n})$ come from cubic splines with hourly or half-hourly nodes. RV_t^{TOD} is from equation (7).

We estimate the bespoke CHAR and SHAR models using the same methods as the models in Section 3. Table 6 presents out-of-sample forecast comparisons of the original models with their bespoke counterparts. We see that forecasts based on bespoke

Figure 7: Bespoke CHAR and SHAR weights



Note: This figure depicts the cross sectional average optimal weights implied by the CHAR and SHAR models along with their bespoke versions.

volatility measures significantly out-perform their benchmark alternatives. The panel GW t-statistics are less than -10 in both comparisons, and the bespoke models out-perform for over 700 of the 866 individual comparisons. These results are qualitatively as strong as the out-performance of cubic HAR over standard HAR reported in Table 1, and provide evidence that the idea of tailoring the risk measure to the forecasting model is not specific to the HAR model of Corsi (2009).

After seeing this strong forecasting performance, we are again interested in the optimal weights that lead to the improved forecasting performance. To this end, we visualize the average optimal weights from bespoke CHAR and SHAR models, and compare them with the weights from original SHAR and CHAR models, which are flat. Figure 7 shows that the optimal weights for these models both display the rise in weights in the afternoon that was observed for bespoke RV for the HAR model in Figure 5. Also, consistent with Patton and Sheppard (2015), we find that the weights on the negative semivariances are greater than those on the positive semivariances, both for the benchmark SHAR model and the bespoke SHAR model, indicating that negative high frequency returns are more important for forecasting one-day-ahead volatility than positive high frequency returns.

4.2. Bespoke GARCH-X

Next we explore bespoke measures of volatility for a model outside of the HAR family of models. We consider the GARCH-X model of Engle (2002), which replaces the lagged squared return in the GARCH model (Bollerslev, 1986) with the lagged realized variance.[10] Assuming a zero mean, the GARCH-X model is

$$r_t = \sqrt{h_t} z_t \tag{15}$$
$$h_t = \omega + \beta h_{t-1} + \alpha RV_{t-1}$$

where $z_t \sim iid\ (0, 1)$.

We construct a bespoke RV for the GARCH-X model by flexibly estimating the weights attached to the high frequency squared returns in RV_t, again using a cubic spline to impose smoothness as a function of the time of day.

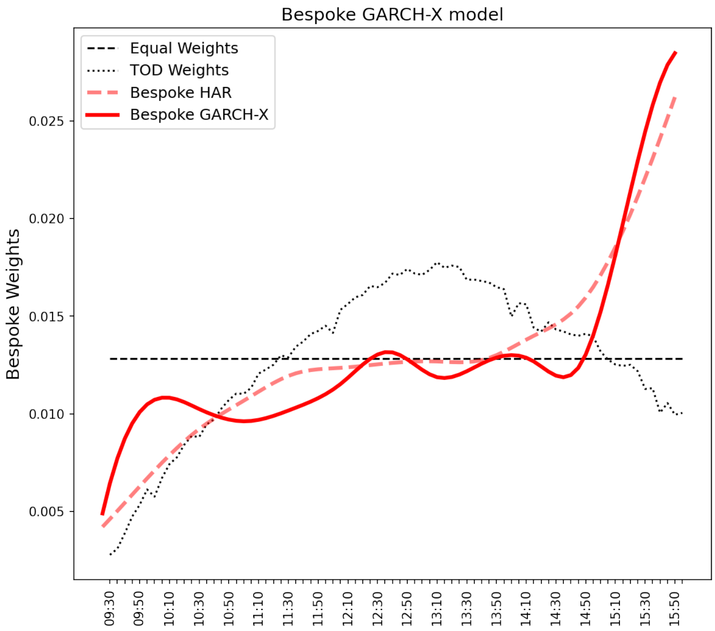
$$h_t = \omega + \beta h_{t-1} + \alpha \widetilde{RV}_t(\boldsymbol{\gamma}) \tag{16}$$

where $\widetilde{RV}_t(\boldsymbol{\gamma})$ is constructed as in the cubic HAR model in equation (6). Given the impressive performance of the simple time-of-day adjusted RV measure in the HAR analysis, we also consider a GARCH-X model with RV^{TOD} on the right-hand side in place of standard RV.

Figure 8 presents the bespoke GARCH-X model weights, as well as the bespoke HAR weights and the TOD weights for comparison. We observe that the bespoke GARCH-X weights are markedly different from both the equal weights of standard RV and the TOD weights, and are quite similar to the optimal bespoke HAR weights, starting the day low and close to the TOD weight, then rising to around the same level as the equal weighted case, and then rising strongly in the afternoon. As discussed in Section 3.4, this increase is likely driven by the additional information contained in that period for returns in the

[10] When focusing on one-day-ahead volatility forecasting, the HEAVY (Shephard and Sheppard, 2010) and realized GARCH (Hansen, Huang and Shek, 2012) models both reduce to a GARCH-X type model. For multi-step forecasting one of these models, or some other extension of the GARCH-X model, is required.

Figure 8: Bespoke GARCH-X weights



Note: This figure depicts the cross sectional average optimal weights implied by the Bespoke GARCH-X model weights.

subsequent day.

Next, we turn to forecast comparisons for this application. We evaluate the forecast performance using the QLIKE loss function and using r_t^2 as a proxy for true volatility, following the spirit of the GARCH-type models.[11] Table 7 reports the forecast comparison results, and confirms that bespoke RVs outperform both the standard RV and RV-TOD when used in GARCH-X models. The panel GW t-statistic comparing basic and bespoke GARCH-X forecasts is -9.1, indicating very strong statistical significance of the forecast improvement. Bespoke RV also significantly outperforms TOD-RV, with a GW t-statistic of -4.3. These results confirm that tailoring the risk measure to the predictive model in which it will be used, whatever form that model takes, leads to improved out-of-sample volatility forecasts.

[11] In Section S.8 of the supplemental appendix we redo the analysis using RV as the volatility proxy, and find slightly more rejections of the null hypothesis of equal forecast accuracy, consistent with this more accurate proxy yielding more powerful tests.

Table 7: Bespoke RV for GARCH-X models

	GW Losses		GW Wins		GW t-stat
Model	**Total**	**Signif**	**Total**	**Signif**	**Panel**
GARCH-X: Basic vs Bespoke	336	68	550	186	-9.1
GARCH-X: TOD vs Bespoke	431	87	455	126	-4.3

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the GARCH-X model with a bespoke counterpart, and with a model using RV^{TOD}, across 886 S&P 500 stocks. A negative panel GW t-statistic indicates that the bespoke model out-performs the competitor, while a positive t-statstic indicates the opposite.

4.3. Flexible models of the news impact curve

The "news impact curve" was introduced by Engle and Ng (1993) as a way to measure the reaction of future volatility to past returns. In the ARCH/GARCH models of Engle (1982) and Bollerslev (1986), the news impact curve is a parabola centered on zero, imposing that positive and negative shocks have an equal impact on future volatility. Engle and Ng (1993), Glosten, Jagannathan and Runkle (1993), and others have found that negative returns lead to greater future volatility than equally-sized positive returns. Chen and Ghysels (2011) propose a flexible model to estimate the news impact curve, as well as allowing for flexible weights on lagged returns. Their most general formulation, adapted to the notation of this paper, is:

$$RV_t = \beta_0 + \sum_{j=1}^{21} \sum_{i=1}^{78} \psi_{ij}(\theta) \text{NIC}(r_{i,t-j}; \theta) + e_t \tag{17}$$

where ψ_{ij} is the weight attached to the i^{th} return on day $t - j$, NIC is a function that determines how past returns affect future volatility, and both functions depend on a parameter vector θ. This framework nests the "fully flexible" model considered above, in that it not only flexibly weights all 21 × 78 past five-minute returns, it estimates what function of those returns best fits the data, whereas the "fully flexible" model imposes that the news impact curve is a parabola. In this section we implement the preferred specifications of Chen and Ghysels (2011) and compare them with the bespoke HAR models discussed above.

Chen and Ghysels (2011) propose modeling the weight function, ψ, as the product of two beta kernels:

$$\begin{aligned}
\psi_{ij}(\theta) &= \text{Beta}(j,21;\theta_1,\theta_2)\times\overline{\text{Beta}}(i,78;\theta_3,\theta_4) && (18)\\
\text{where}\quad \text{Beta}(k,K;\alpha,\beta) &= \left(\frac{k}{K+1}\right)^{\alpha-1}\left(1-\frac{k}{K+1}\right)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (19)\\
\overline{\text{Beta}}(k,K;\theta_3,\theta_4) &\equiv \frac{\text{Beta}(k,K;\alpha,\beta)}{\sum_{i=1}^{K}\text{Beta}(i,K;\alpha,\beta)}\\
\text{for}\quad (\alpha,\beta) &> 0 \quad\text{and}\quad k=1,2,...,K
\end{aligned}$$

The first beta function determines how observations from different days are weighted, and the second beta function (which is normalized to sum to one, for identification) determines how observations from within the trade day are weighted. Our preferred specification, the "cubic HAR" model, uses a step function to weight observations across days (in line with the original HAR specification of Corsi, 2009) and a cubic spline to weight observations within the trade day. Chen and Ghysels (2011) find that their best specification for the NIC function is one they call "ASYMGJR":

$$\text{NIC}(r) = b\cdot r^2 + c\cdot r^2\mathbf{1}\{r\leq 0\} \tag{20}$$

which allows positive and negative returns to have differing impacts on future volatility. The symmetric version of this model imposes $c = 0$.

We estimate the symmetric and asymmetric versions of the Chen-Ghysels model on our sample of stocks, and report the results of tests of out-of-sample forecast performance in Table 8. In the first row we consider the symmetric NIC function, and so differences in forecast performance are purely attributable to differences in how lagged high-frequency returns are weighted. We observe that the cubic HAR model significantly outperforms the symmetric CG model, with a panel GW t-statistic of less than -25. Thus this sample of stocks is better fit using the HAR-like step function to weight different days, and the cubic spline within each day. In the second row we consider the asymmetric CG model, and we see that it has better performance than the symmetric version, but is still

Table 8: Comparison with Chen and Ghysels (2011)

	GW Losses		**GW Wins**		**GW t-stat**
Cubic HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
CG symmetric	195	35	676	354	-26.5
CG asymmetric	226	45	646	298	-17.0

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the cubic HAR model against the symmetric and asymmtric models of Chen and Ghysels (2011), see equation (20), across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite.

soundly beaten by the cubic HAR model.[12,13] We infer that the flexibility of allowing positive and negative returns to affect future volatility is not enough to overcome the loss in performance from using the beta weight functions rather than the weights adopted in the Cubic HAR model. In the next section, we propose a refinement of that model to see whether sign information can indeed be useful in constructing a bespoke RV.

4.4. Two-dimensional bespoke RV

Here we consider an extension of our baseline model and allow the bespoke RV weights to vary across lags and also across the value of the high frequency return, in the spirit of Chen and Ghysels (2011) discussed in the previous section, though with a different functional form and different estimation method. We continue to embed the bespoke

[12] Comparing the models of Chen and Ghysels (2011) directly with the benchmark HAR model, we find that the symmetric model is beaten by the HAR model, with a GW t-statistic of -4.4, while the asymmetric Chen-Ghysels model significantly beats the benchmark HAR model, with a panel GW t-statistic of 6.4.

[13] It is worth noting that Chen and Ghysels (2011) apply their model to stock indices (S&P 500 and the Dow Jones Industrial Average) rather than individual stocks as we have done here.

RVs in a HAR model, but now with a two-dimensional weighting function:

$$\begin{aligned} RV_t &= \beta_0 + \beta_d \sum_{i=1}^{78} g(i;\gamma_d,\tau) g\left(\frac{r_{i,t-1}}{\sigma_{i,t-1}};\alpha_d,\mathbf{s}\right)\sigma^2_{i,t-1} \\ &\quad +\beta_w \frac{1}{4}\sum_{j=2}^{5}\sum_{i=1}^{78} g(i;\gamma_w,\tau) g\left(\frac{r_{i,t-j}}{\sigma_{i,t-j}};\alpha_w,\mathbf{s}\right)\sigma^2_{i,t-1} \\ &\quad +\beta_m \frac{1}{16}\sum_{j=6}^{21}\sum_{i=1}^{78} g(i;\gamma_m,\tau) g\left(\frac{r_{i,t-j}}{\sigma_{i,t-j}};\alpha_m,\mathbf{s}\right)\sigma^2_{i,t-1} \end{aligned} \tag{21}$$

where $g(\cdot;\gamma,\tau)$ is a cubic spline with knots given by τ and estimated parameter γ. As in the cubic HAR model, we use hourly or half-hourly knots for the time-of-day weights. For the second dimension, we use a cubic spline for the standardized return, $r_{i,t}/\sigma_{i,t}$. We set the knots to be $0, \pm 2, \pm 5$, distinguishing between large and moderate returns, and allowing the impact to differ depending on whether the return is positive or negative. As return volatility is known to vary predictably as a function of the time of day, recall Figure 1, as well as being strongly persistent across time, we estimate $\sigma_{i,t}$ in two steps. We first estimate a simple HAR model to get the one-day-ahead predicted volatility, $\widehat{RV}_t$, and then we compute standardized five-minute returns as

$$\tilde{r}_{i,t} = \frac{r_{i,t}}{\sqrt{\widehat{RV}_t/78}} \tag{22}$$

We estimate the time-of-day component, $\tilde{\omega}_i$, as in equation (8), but using the standardized five-minute returns, $\tilde{r}_{i,t}$ rather than the raw returns. We combine these components to get $\sigma_{i,t}$:

$$\sigma^2_{i,t} = \tilde{\omega}^2_i \times \widehat{RV}_t \tag{23}$$

Note that if the spline for the standardized returns simplifies to be a parabola, then the $1/\sigma_{i,t}$ term inside the spline and the $\sigma^2_{i,t}$ outside the spline cancel out, and the model simplifies to the "cubic HAR" model described in Section 2.1.

We estimate this "two-dimensional bespoke HAR" model using the the mini-Batch stochastic gradient descent algorithm described in Section 2.3.1. The results of forecast comparison tests are reported in Table 9.

Table 9: Two-dimensional bespoke RV

	GW Losses		**GW Wins**		**GW t-stat**
2D bespoke HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
HAR	182	45	704	432	-13.1
TOD HAR	313	82	573	179	-2.0
Cubic HAR	376	99	510	155	-1.1
Bespoke SHAR	415	126	466	156	-3.0
CG symmetric	224	46	647	313	-13.9
CG asymmetric	248	61	624	268	-9.8

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the two-dimensional bespoke HAR model, see equation (22), against various competing models across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model outperforms the two-dimensional bespoke HAR model, while a negative t-statstic indicates the opposite.

Table 9 reveals that the 2D bespoke HAR is, on average, the best-performing model considered, with all panel GW t-statistics being negative, indicating it has lower average loss than the competing models. The 2D bespoke HAR significantly beats the baseline HAR model, with a panel GW t statistic of -13, and it beats both of the Chen and Ghysels (2011) specifications almost as strongly. We compare the 2D bespoke HAR model with the bespoke semi HAR model introduced in Section 4.1 as the bespoke semi HAR represents a different way to incorporate sign information into the predictive model. This model inflicts the most losses on the 2D bespoke HAR (415 out of 886) but it still significantly outperformed across the panel, with a t-statistic of -3.

The only model that 2D bespoke HAR fails to significantly outperform is the Cubic HAR model, for which the panel GW t-statistic is -1.1. Since the 2D bespoke HAR reduces to the Cubic HAR when the spline for standardized returns is a parabola, this statistical tie in forecast performance reveals that the optimal transformation of high-freqency returns is not different from the simple squared return. We attribute this finding to our focus on prediction rather than estimation: in out-of-sample forecasting an effect has to be informative enough to overcome the increased estimation error that capturing it incurs. In our sample it turns out that a simple quadratic function of returns, or something close to it, produces the best forecasts.

4.5. Parametric bespoke RV

Finally, we exploit the fact that the optimal bespoke RV weights for the baseline HAR model reveal a relatively simple pattern, see Figure 5, one that might be well approximated using a simple polynomial function of the time of day. We consider this simplification of our original approach here. We propose a "parametric bespoke RV" that imposes the use of squared returns, motivated by the results in the previous section, and uses a simple cubic function of the time of day:

$$\begin{aligned} \widetilde{RV}_t(\gamma) &= \sum_{i=1}^{78} \left(\gamma_0 + \gamma_1 i + \gamma_2 i^2 + \gamma_3 i^3\right) r_{i,t}^2 \\ &\equiv \gamma_0 RV_t + \gamma_1 RV_t^{Lin} + \gamma_2 RV_t^{Quad} + \gamma_3 RV_t^{Cub} \end{aligned} \tag{24}$$

Note that the new realized variances, RV_t^{Lin}, RV_t^{Quad}, and RV_t^{Cub}, are simple functions of squared returns, and do not require optimization.[14] If we use this approximation in place of the more flexible approach, and exploit the knowledge gained from Section 3.4 that the optimal weights for the weekly and monthly lags are the simple time-of-day weights, we obtain a HAR model with three additional regressors:

$$\begin{aligned} RV_t = \beta_0 &+ \beta_d RV_{t-1} + \beta_d^L RV_{t-1}^{Lin} + \beta_d^Q RV_{t-1}^{Quad} + \beta_d^C RV_{t-1}^{Cub} \\ &+ \beta_w \frac{1}{4} \sum_{j=2}^{5} RV_{t-j}^{TOD} + \beta_m \frac{1}{16} \sum_{j=6}^{21} RV_{t-j}^{TOD} + e_t \end{aligned} \tag{25}$$

The model has only seven parameters and can be estimated with standard methods: OLS if the quadratic loss function is adopted, or simple numerical optimization if QLIKE is adopted. Unlike the larger models considered in Section 2.1, whose parameters are estimated using regularized methods, the model in equation (26) can be studied using standard inference methods for time series data, see Hamilton (1994) and White (2001) for example.

[14] In unreported results, we also experimented with a step function approximation, where the weights are flat as a function of the time of day aside for a jump for the last hour of trade. This corresponds to augmenting the standard HAR with an additional regressor, the realized volatility computed over the last hour of the trade day. Despite the intuitive appeal of this model, the forecasts it produced were significantly worse than those of the flexible approach and the parametric polynomial approach in equation (25), and so we did not proceed in that direction.

Table 10: Forecast performance of parametric bespoke HAR vs other models on 886 S&P 500 stocks

	GW Losses		**GW Wins**		**GW t-stat**
Parametric Bespoke vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
HAR	149	33	737	521	-25.4
TOD HAR	339	61	547	168	-9.2
Cubic HAR	479	159	407	110	-3.4

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the "parametric bespoke HAR" model against competing models across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model out-performs the parametric bespoke HAR model, while a negative t-statstic indicates the opposite.

In Table 10 we show that the parametric bespoke RV loses the majority of individual GW tests (479 losses to 407 wins; 159 significant losses to 110 significant wins) but it wins in the panel GW test, with a t-statistic of -3.4. Overall, a reasonable conclusion is that its performance is on par with the more flexible approach, and it certainly wins computationally: the model has just 7 free parameters, and can be estimated quickly and easily.

Of course, some "data snooping" is happening here: Having seen the results of the flexible approach on our sample of data, e.g. in Figure 5, it is not surprising that a more restrictive method can replicate the good out-of-sample forecast performance so long as it can capture the main features detected by the more flexible approach. For future researchers, the benefits of this section are that they can obtain optimal or near-optimal weights with a simple parametric method, thereby avoiding a difficult computational problem, and can be assured that such an approach yields forecasts that are optimal in a much broader class of models.

5. Conclusion

This paper proposes to tailor the measure of risk, such as realized variance (RV), to the specific forecasting model and the specific asset of interest in order to improve the model's forecasting performance. The resulting "bespoke RV" will not necessarily be a good "all purpose" measure of risk, but it will optimally draw on the available high frequency information to improve, if possible, the forecasting performance of the model.

We use data on all 886 stocks that were ever a constituent of the S&P 500 index over the period 1995 to 2019, and we exploit recent advances in the estimation of deep neural networks (DNNs) to flexibly construct "bespoke" measures of volatility. We find that being completely flexible in the construction of the bespoke measure leads to poor out-of-sample forecast performance, however after imposing some economically motivated structure we obtain forecasts that significantly outperform the benchmark forecasts. Our main analyses focus on bespoke RVs for the heterogeneous autoregressive (HAR) model of Corsi (2009), and as extensions we consider the GARCH-X model (Engle, 2002), the "continuous HAR" of Andersen et al. (2007), and the "semi HAR" of Patton and Sheppard (2015). In all four cases we find significant improvements in out-of-sample forecast performance when using RVs tailored to the application.

We consider an extension of the model to allow the bespoke weights to depend on the sign and size of the high-frequency return, as well as the time of day, and find that the optimal transformation of high-frequency returns, from a forecasting perspective, is not different from the simple quadratic function. We also consider a simplification of the model, exploiting the fact that optimal bespoke weights appear easy to approximate parametrically, and find that a "parametric bespoke RV" achieves comparable forecast performance at significantly lower computational cost.

Opening the black box to understand the sources of forecast improvements from using bespoke RVs, we find two main channels. Firstly, we find that using a bespoke RV in place of a standard RV leads the model to put more weight on the risk measure, increasing the responsiveness of the forecast. Secondly, we find that the optimal bespoke RVs, across all four models, place higher weight on returns from the afternoon, which is in contrast with both the standard equal-weighted RV and an RV based on time-of-day information. We find evidence that this increased afternoon weight comes from an information channel: afternoon returns are closest to the target date, and thus more informative about the future volatility.

This paper leaves open several avenues for future work. We focused exclusively on univariate volatility models, and the important extension to multivariate models opens up questions about the optimal degree of customization potentially differing between variance and correlations, as well as the usual empirical challenges of moving to high dimension models. We focused on linear bespoke RVs, and the extension to nonlinear versions could yield further improvements, although our results suggest that imposing some structure on the bespoke RV is important for achieving forecast gains. Finally, our focus is on volatility models, but any forecasting model that uses a variable constructed from other variables or data sources, e.g. macroeconomic forecasting models using pre-constructed indices of prices or economic activity, may benefit from tailoring that input. We look forward to pursuing, or reading about, these ideas in the future.

References

Ait-Sahalia, Y., Mykland, P.A., Zhang, L., 2005. How often to sample a continuous-time process in the presence of market microstructure noise. Review of Financial Studies 18, 351–416.

Andersen, T.G., Bollerslev, T., 1997. Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance 4, 115–158.

Andersen, T.G., Bollerslev, T., 1998. Deutsche mark–dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies. Journal of Finance 53, 219–265.

Andersen, T.G., Bollerslev, T., Christoffersen, P.F., Diebold, F.X., 2006. Volatility and correlation forecasting. Handbook of Economic Forecasting 1, 777–878.

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, 701–720.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001. The distribution of realized exchange rate volatility. Journal of the American Statistical Association 96, 42–55.

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

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

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annual Review of Economics 11, 685–725.

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

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, 1481–1536.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2009. Realized kernels in practice: Trades and quotes. Econometrics Journal 12, 1–32.

Barndorff-Nielsen, O.E., Kinnebrock, S., Shephard, N., 2010. Measuring downside risk: Realised semivariance, in: Bollerslev, T., Russell, J.R., Watson, M.W. (Eds.), Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle. Oxford University Press, Oxford, 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, 253–280.

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

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. Review of Financial Studies 34, 1046–1089.

Bilmes, J., Asanovic, K., Chin, C.W., Demmel, J., 1997. Using PHiPAC to speed error back-propagation learning, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4153–4156.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. Handbook of Econometrics 4, 2959–3038.

Bucci, A., 2020. Realized volatility forecasting with neural networks. Journal of Financial Econometrics 18, 502–531.

Chen, X., Ghysels, E., 2011. News – good or bad – and its impact on volatility predictions over multiple horizons. Review of Financial Studies 24, 46–81.

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Valid post-selection and post-regularization inference: An elementary, general approach. Annual Reviews in Economics 7, 649–688.

Christensen, K., Siggaard, M., Veliyev, B., 2022. A machine learning approach to volatility forecasting. Journal of Financial Econometrics forthcoming.

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

Engle, R., 2002. New frontiers for ARCH models. Journal of Applied Econometrics 17, 425–446.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica , 987–1007.

Engle, R.F., Ng, V.K., 1993. Measuring and testing the impact of news on volatility. Journal of Finance 48, 1749–1778.

Filipović, D., Khalilzadeh, A., 2021. Machine learning for predicting stock return volatility. Swiss Finance Institute Research Paper .

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326–2377.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS touch: Mixed data sampling regression models. Working Paper .

Ghysels, E., Santa-Clara, P., Valkanov, R., 2006. Predicting volatility: How to get most out of returns data sampled at different frequencies. Journal of Econometrics 131, 59–95.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48, 1779–1801.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press.

Hansen, P.R., Huang, Z., Shek, H.H., 2012. Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics 27, 877–906.

Harris, L., 1986. A transaction data study of weekly and intradaily patterns in stock returns. Journal of Financial Economics 16, 99–117.

Jacod, J., 2008. Asymptotic properties of realized power variations and related functionals of semimartingales. Stochastic Processes and their Applications 118, 517–559.

Jacod, J., 2018. Limit of random measures associated with the increments of a Brownian semimartingale. Journal of Financial Econometrics 16.

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, 2249–2276.

Jacod, J., Protter, P., 1998. Asymptotic error distributions for the euler method for stochastic differential equations. Annals of Probability , 267–307.

Judd, K.L., 1998. Numerical Methods in Economics. MIT Press.

Li, S.Z., Tang, Y., 2021. Forecasting realized volatility: An automatic system using many features and many machine learning algorithms. Available at SSRN 3776915 .

Liu, L.Y., Patton, A.J., Sheppard, K., 2015. Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. Journal of Econometrics 187, 293–311.

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

Mullainathan, S., Spiess, J., 2017. Machine learning: An applied econometric approach. Journal of Economic Perspectives 31, 87–106.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Patton, A., Zhang, H., 2022. Re-imgaing volatility: Computer vision approach for realized volatility forecasting. Working Paper .

Patton, A.J., Sheppard, K., 2009. Evaluating volatility and correlation forecasts, in: Handbook of Financial Time Series. Springer, pp. 801–838.

Patton, A.J., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics 97, 683–697.

Patton, A.J., Weller, B.M., 2022. Risk price variation: The missing half of empirical asset pricing. Review of Financial Studies 35, 5127–5184.

Poon, S.H., Granger, C.W.J., 2003. Forecasting volatility in financial markets: A review. Journal of Economic Literature 41, 478–539.

Reisenhofer, R., Bayer, X., Hautsch, N., 2022. Harnet: A convolutional neural network for realized volatility forecasting. arXiv preprint arXiv:2205.07719 .

Robbins, H., Monro, S., 1951. A stochastic approximation method. Annals of Mathematical Statistics , 400–407.

Shen, Z., Xiu, D., 2024. Can machines learn weak signals? University of Chicago, Becker Friedman Institute for Economics Working Paper .

Shephard, N., Sheppard, K., 2010. Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. Journal of Applied Econometrics 25, 197–231.

White, H., 2001. Asymptotic theory for econometricians (2nd ed.). Academic Press.

Wood, R.A., McInish, T.H., Ord, J.K., 1985. An investigation of transactions data for nyse stocks. Journal of Finance 40, 723–739.

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, 1394–1411.

Supplemental Appendix for

Bespoke Realized Volatility: Tailored Measures of Risk for Volatility Prediction

by Andrew J. Patton and Haozhe Zhang

This version: April 22, 2024

This appendix contains nine sections, presenting additional details and analyses relevant to the main paper.

S.1. Data cleaning details

We use trade data from the NYSE Trade and Quote (TAQ) database. We follow Barndorff-Nielsen et al. (2009) to clean this data, using the following rules:

1. Keep only entries with time step between 9:30 am to 4:00 pm (when the exchange is open).
2. Delete entries with zero transaction prices.
3. Retain entries originating from NYSE and NASDAQ only.
4. Delete entries with corrected trades (CORR $\neq 0$).
5. Delete entries with unusual sale condition (COND with letter code except for letters E and F).
6. We use the median price if multiple entries have the same time stamp.

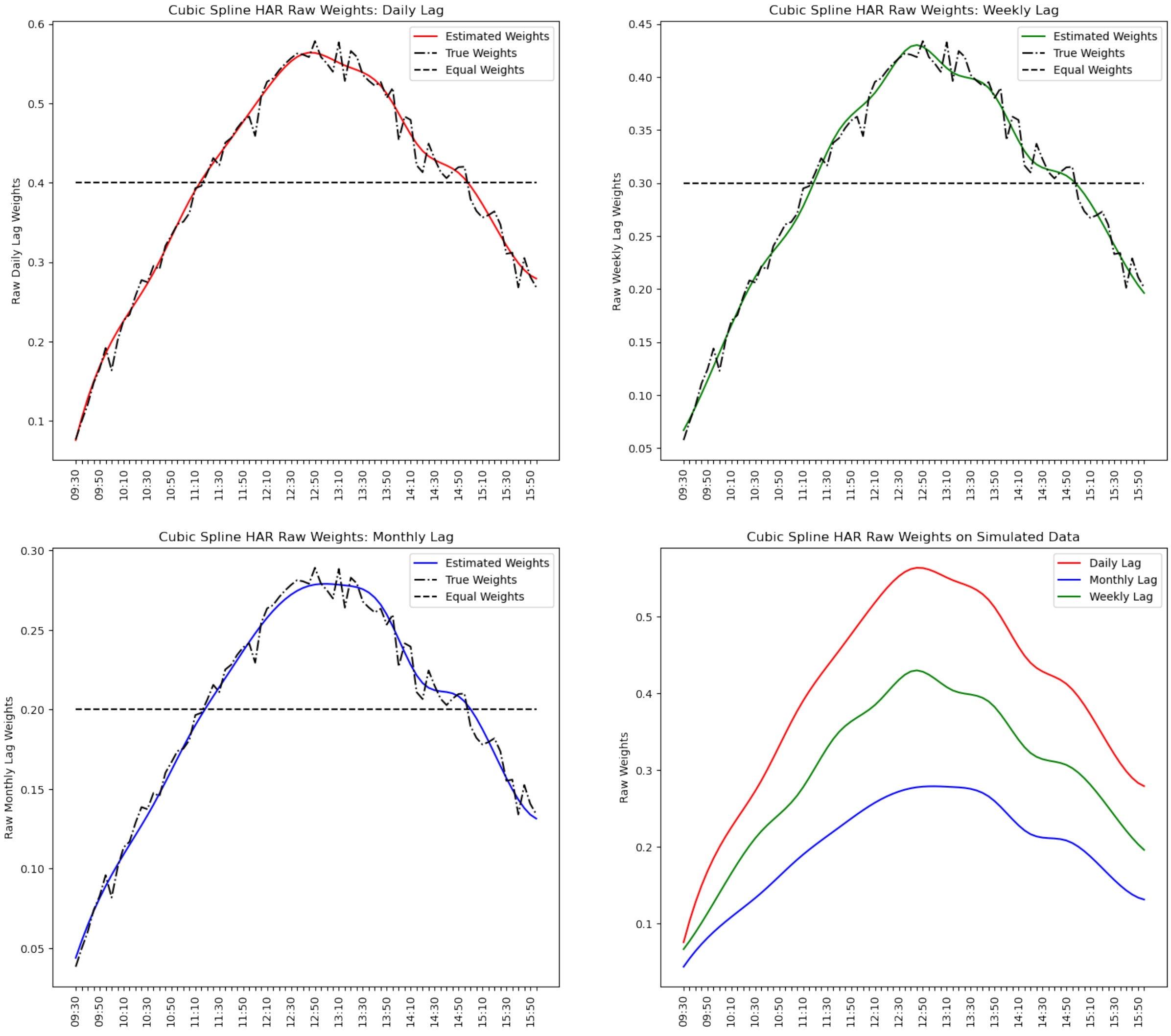
S.2. Simulation study

In this section we simulate data to illustrate the ability of the proposed Cubic HAR model to recover the true coefficients. We assume that high-frequency returns, $r_{i,t}$, are i.i.d. N(0,1/M), where M is the number of high-frequency returns each trade day. We combine these high frequency returns to obtain realized variances using time-of-day (TOD) weights, ω_i, as in equation 7 of the main paper, and we use a HAR model to simulate the time series dynamics of RV.

$$RV_t = \beta_0 + \beta_d \sum_{i=1}^{M} \omega_i r_{i,t-1}^2 + \beta_w \frac{1}{4} \sum_{j=2}^{5} \sum_{i=1}^{M} \omega_i r_{i,t-j}^2 + \beta_m \frac{1}{16} \sum_{j=6}^{21} \sum_{i=1}^{M} \omega_i r_{i,t-j}^2 + \epsilon_t \tag{S.1}$$

where $\epsilon_t \sim N(0,1)$. We set $\beta_0 = 0$, $\beta_d = 0.4$, $\beta_w = 0.3$, $\beta_m = 0.3$, and $M = 78$. We impose that ω_i follows the empirical TOD weights presented in Figure 4. Then we use the cubic HAR model with the exact same set-up (hyperparameter grid and estimation algorithm) to estimate the coefficients on the simulated data. Figure S.1 shows that when the true data generating process indeed is a linear combination with the TOD weights, our cubic HAR model will recover the true weights almost perfectly. (As the true weights in this simulation are not a smooth function of time, the cubic spline model clearly cannot obtain a perfect fit.) This simulation results give us confidence that the proposed Cubic HAR has enough estimation power to recover the optimal weighting scheme for realized variance forecasting.

Figure S.1: Cubic HAR vs True Weights: Simulated Data



Note: This figure compares the estimated optimal weights (solid line) based on simulated data based with the true weights (dash-dotted line).

S.3. Alternative target for shrinkage

In the analysis in Section 3.2 of the main paper, we consider shrinkage methods (Ridge, LASSO and elastic net) that shrink the estimated parameters towards zero. This is a standard shrinkage target in high-dimensional estimation, but in our application an interesting alternative target is to shrink the parameters towards the benchmark HAR parameters. To do this, we re-write the regularized regression models as:

$$RV_t = \beta_0 + \widetilde{RV}_{t-1}(\beta^d_{HAR} + \boldsymbol{\gamma}^d) + \frac{1}{4}\sum_{j=2}^{5}\widetilde{RV}_{t-j}(\beta^w_{HAR} + \boldsymbol{\gamma}^w) + \frac{1}{16}\sum_{j=6}^{21}\widetilde{RV}_{t-j}(\beta^m_{HAR} + \boldsymbol{\gamma}^m) + e_t \tag{S.1}$$

That is, we split the coefficients on high frequency returns into the HAR coefficients and a perturbation term to be estimated. Then, instead of regularizing the entire coefficient, we only penalize the perturbation terms, and shrink these towards zero. That is, the penalty term becomes:

$$\alpha(\lambda\|\boldsymbol{\gamma}\|_1 + (1-\lambda)\|\boldsymbol{\gamma}\|_2) \tag{S.2}$$

where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}^d, \boldsymbol{\gamma}^w, \boldsymbol{\gamma}^m]$. The estimation procedure and hyperparameters grid are identical to those in the original regularized regression analysis. We compare the cubic HAR model with these alternative shrinkage estimators in Table S.1. (The OLS and HAR results are identical to those in Table 2 of the main paper, and are included here for ease of comparison.) With this alternative target the optimal degree of shrinkage is found to be large, and the Ridge, LASSO and elastic net models are all shrunk almost all the way to the benchmark HAR parameters. Given that, it is unsurprising that the cubic HAR continues to significantly out-perform these alternative regularized models, as shown in Table S.1, which have performance comparable to the benchmark HAR model.

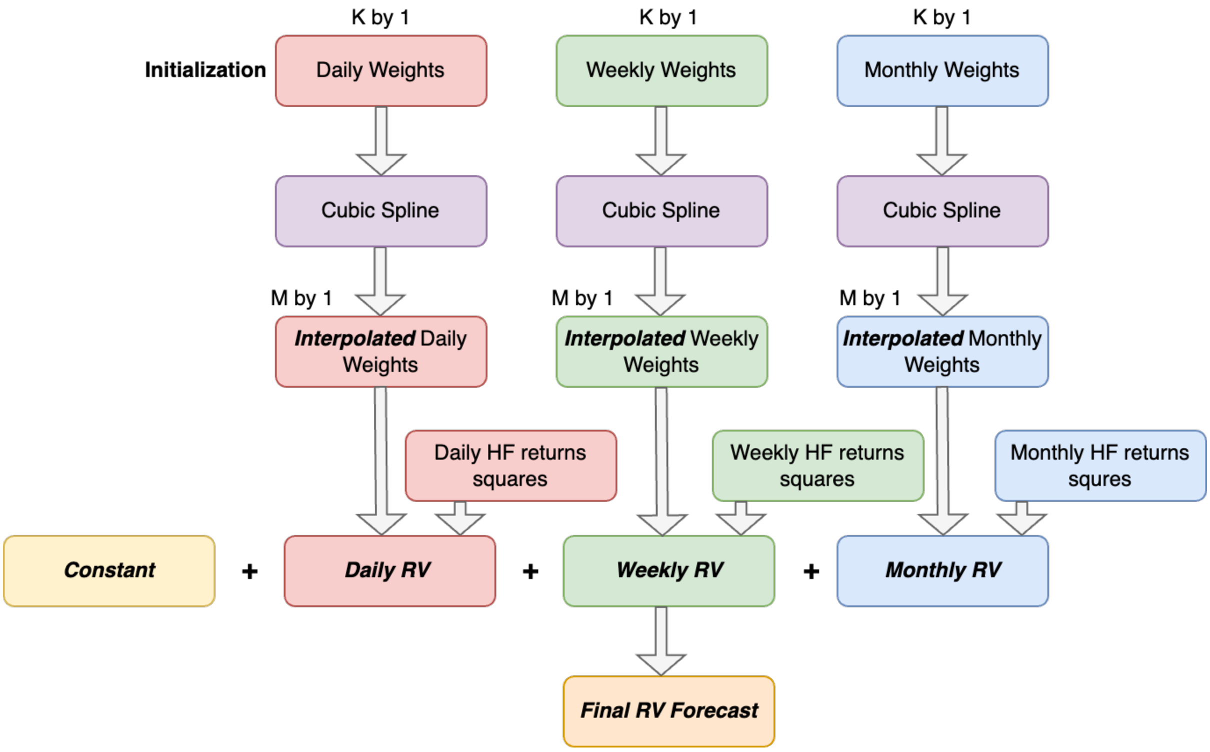
Table S.1: Alternative regularized regression models: Shrink towards HAR

	GW Losses		**GW Wins**		**GW t-stat**
CubicHAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Ridge-Alt	166	44	720	423	-17.3
LASSO-Alt	165	46	721	437	-4.2
Elastic net-Alt	170	46	716	412	-16.4
OLS	54	17	832	564	-6.4
HAR	155	41	731	470	-21.7

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the cubic HAR model against competing models across 886 S&P 500 stocks. A positive panel GW t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite. The models labeled "-Alt" shrink the estimated coefficients towards the HAR model coefficients, rather than towards zero as in Table 2 of the main paper.

S.4. Cubic HAR Model Architecture

Figure S.2: Cubic HAR Model Architecture

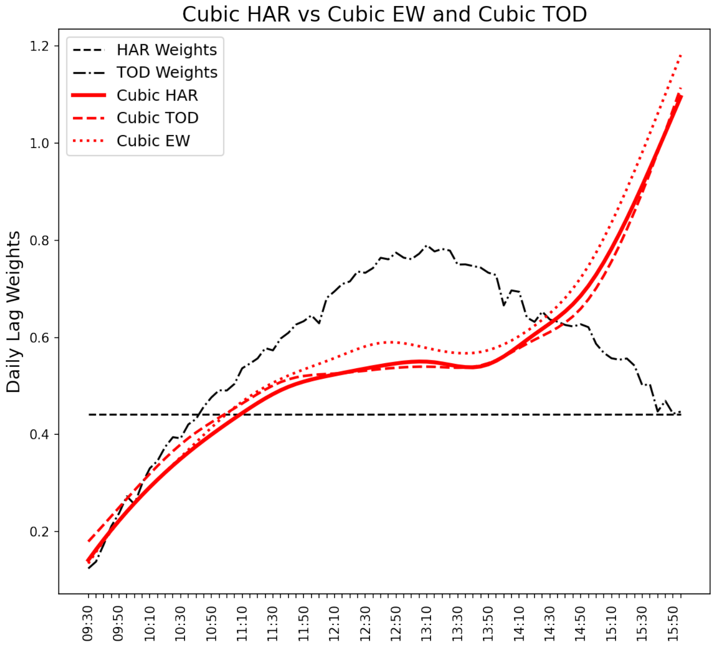


Note: This figure illustrates the cubic HAR model architecture. In particular, it shows how we use miniBatch gradient descent to iteratively solve the optimal parameters $(3 \times K + 1)$ and then use them to get the cubic spline interpolated weights for constructing final RV forecast. Here K represents the number of basis points for cubic spline interpolation or the true number of parameters, and M represents the desired number of points as cubic spline interpolation output.

S.5. Visualizing the optimal Cubic TOD and Cubic EW weights

This section presents the optimal daily weights for the cubic HAR, cubic TOD and cubic EW models, which differ in what shapes are estimated or imposed on the weekly and monthly weights. We observe the cubic TOD and cubic EW daily weights are almost identical to the cubic HAR weights for daily lags, which shows that restricting the shapes of weekly and monthly weights has almost zero affect on the estimated daily weights.

Figure S.3: Cubic TOD and Cubic EW weights



Note: This figure compares estimated cubic TOD and cubic EW weights against the cubic HAR model weights, averaged across the 886 S&P 500 stocks in our sample.

S.6. Forecast comparisons using mean-squared error loss

Table S.2: HAR vs other models performance on S&P500. MSE loss.

	GW Losses		GW Wins		GW t-stat
HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	421	169	465	106	0.8
Flexible HAR	726	408	160	11	3.3
Cubic HAR	714	406	172	7	1.9
TOD HAR	653	348	233	62	2.2
Ridge	462	186	424	120	-0.1
LASSO	203	46	683	338	-4.1
Elastic net	434	166	452	136	-2.0
OLS	64	7	822	507	-3.3

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the baseline HAR model against competing models across 886 S&P 500 stocks, using MSE loss. A positive panel GW t-statistic indicates that the competing model out-performs the HAR model, while a negative t-statstic indicates the opposite. This table is related to Table 1 of the main paper, which uses QLIKE loss.

Table S.3: Cubic HAR vs other models performance on S&P 500. MSE loss.

	GW Losses		GW Wins		GW t-stat
Cubic HAR vs:	**Total**	**Signif**	**Total**	**Signif**	**Panel**
Fully Flexible	151	16	735	383	-0.7
Flexible HAR	411	55	475	90	0.0
TOD HAR	405	69	481	196	-1.0
HAR	172	7	714	406	-1.9
Ridge	171	12	715	349	-1.7
LASSO	66	2	820	542	-5.2
Elastic net	159	8	727	372	-3.4
OLS	23	3	863	664	-3.3

Note: This table reports individual and panel Giacomini-White (2006) tests comparing the cubic HAR model against competing models across 886 S&P 500 stocks, using MSE loss. A positive panel GW t-statistic indicates that the competing model out-performs the cubic HAR model, while a negative t-statstic indicates the opposite. This table is related to Table 2 of the main paper, which uses QLIKE loss.

S.7. Multi-days ahead forecasting comparisons: TOD HAR vs HAR

Table S.4: Multidays Ahead Volatility Forecasting: TOD HAR vs HAR

TOD vs HAR	**GW Losses**		**GW Wins**		**GW t-stat**
Horizon	**Total**	**Signif**	**Total**	**Signif**	**Panel**
1-Day	206	63	680	458	-18.6
2-Day	253	72	633	349	-12.5
3-Day	257	74	629	314	-10.1
4-Day	264	74	622	293	-13.6
5-Day	270	71	616	298	-11.5
20-Day	253	68	633	345	-6.8
60-Day	345	126	541	255	-0.2

Note: In this table, we report the individual and panel Diebold Mariano tests results of TOD HAR model against the HAR model in the S&P500 cross section for longer horizon forecasting. Note that negative test statistics favors the TOD HAR model.

S.8. GARCH-X with RV as the target variable

The GARCH-X model in Section 4.2 of the main paper uses, effectively, the daily squared return as the volatility proxy when evaluating forecast accuracy. Realized variance is known to be a more accurate volatility proxy (see, e.g., Andersen and Bollerslev (1998) and Andersen et al. (2003)), and more accurate proxies lead to more powerful forecast comparisons, (see, e.g., Patton (2011)). Here we consider the estimating and evaluating the GARCH-X model replacing the squared daily return with 5-minute realized variance. Table S.5 presents results corresponding to Table 8 in the main paper. We see that GARCH-X based on bespoke RV continues to significantly outperform both the standard GARCH-X model, and the model use time-of-day (TOD) RV, with panel GW t-statistics less than -4 for both comparisons. The main difference between Table S.5 and Table 8 is that the number of individual GW tests that reject the null (listed in the "Signif" columns of the table) is greater, ranging from 68 to 186 in Table 8 and from 75 to 293 here. This increase in significance is consistent with 5-minute RV being a more accurate volatility proxy.

Table S.5: Bespoke RV for GARCH-X models, with RV as the target variable

	GW Losses		**GW Wins**		**GW t-stat**
Model	**Total**	**Signif**	**Total**	**Signif**	**Panel**
GARCH-X: Basic vs Bespoke	281	75	605	293	-2.6
GARCH-X: TOD vs Bespoke	528	151	358	84	-7.4

In Figure S.4 we plot the cross sectional average optimal weights for the bespoke GARCH-X model, estimated with 5-minute RV as the volatility proxy. This figure is similar to Figure 8 in the main paper.

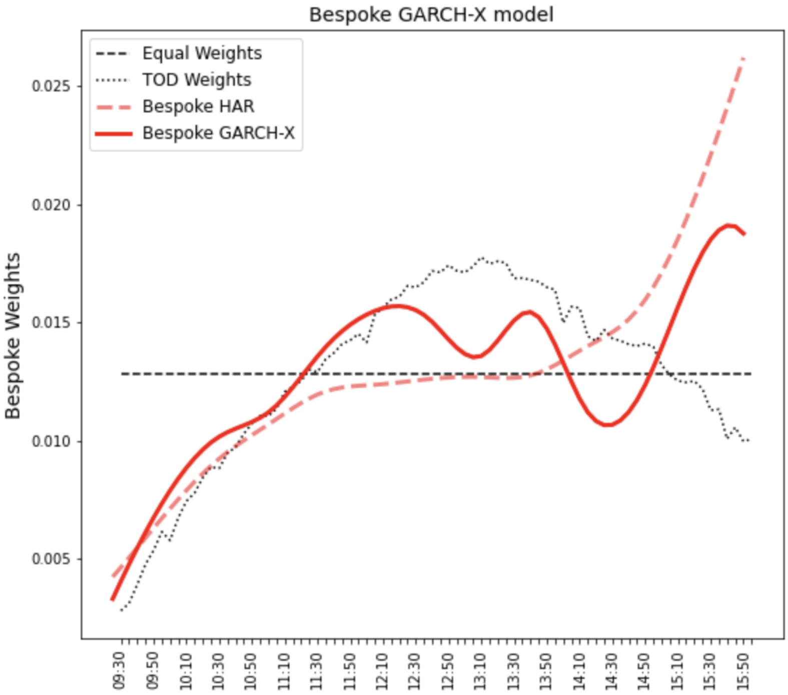
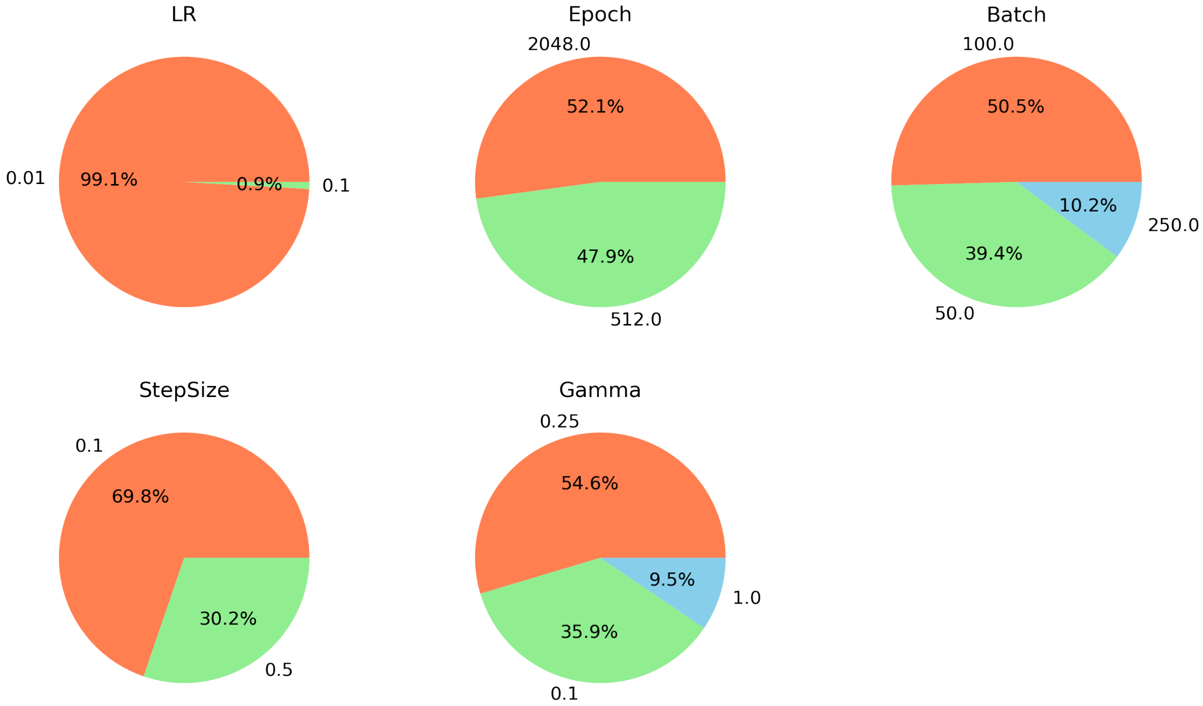


Figure S.4: Bespoke GARCH-X with RV target variable

S.9. Visualization of the hyper-parameters for deep learning based models

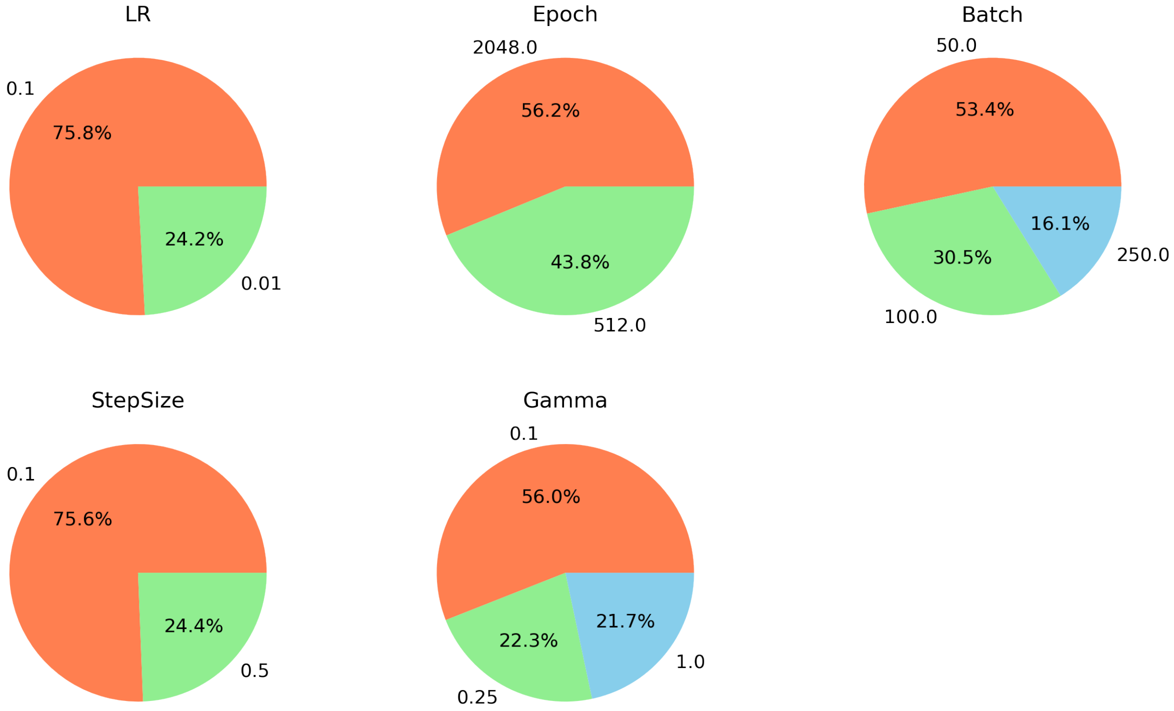
In this section, we also report visualization of the cross sectional variation in the deep learning based models.

Figure S.5: Fully flexible model hyperparameters



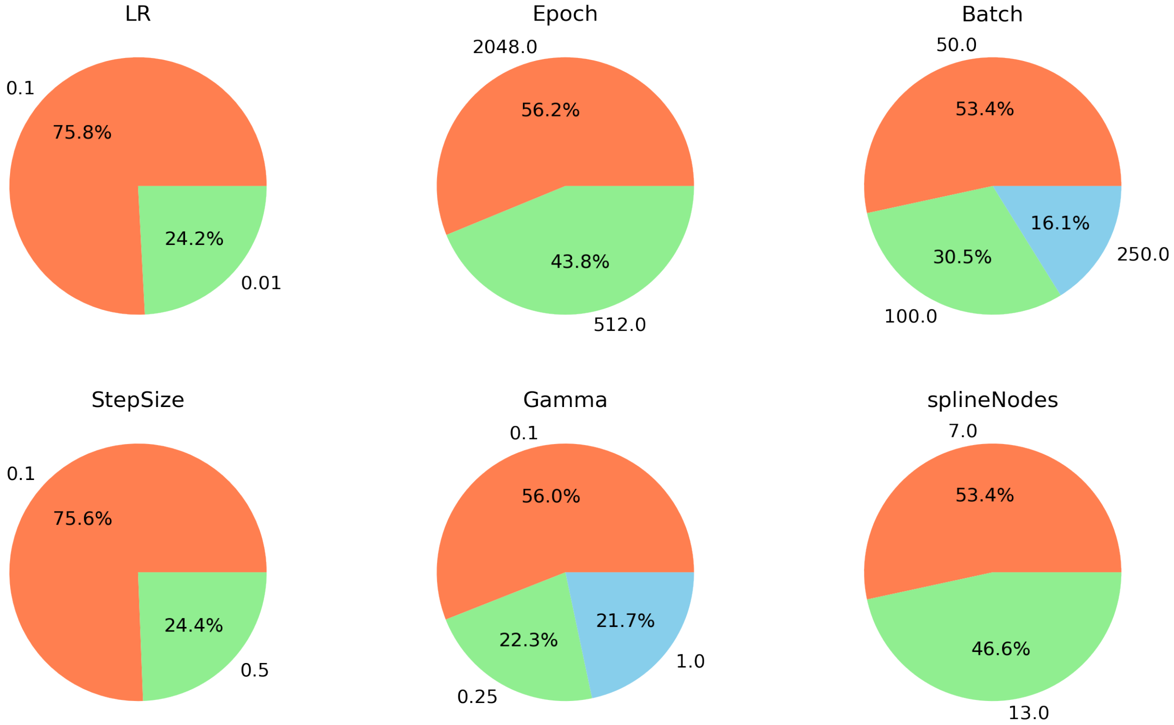
Note: This figure displays the cross-sectional variation of the optimal hyperparameters for the fully flexible model.

Figure S.6: Flexible HAR model hyperparameters



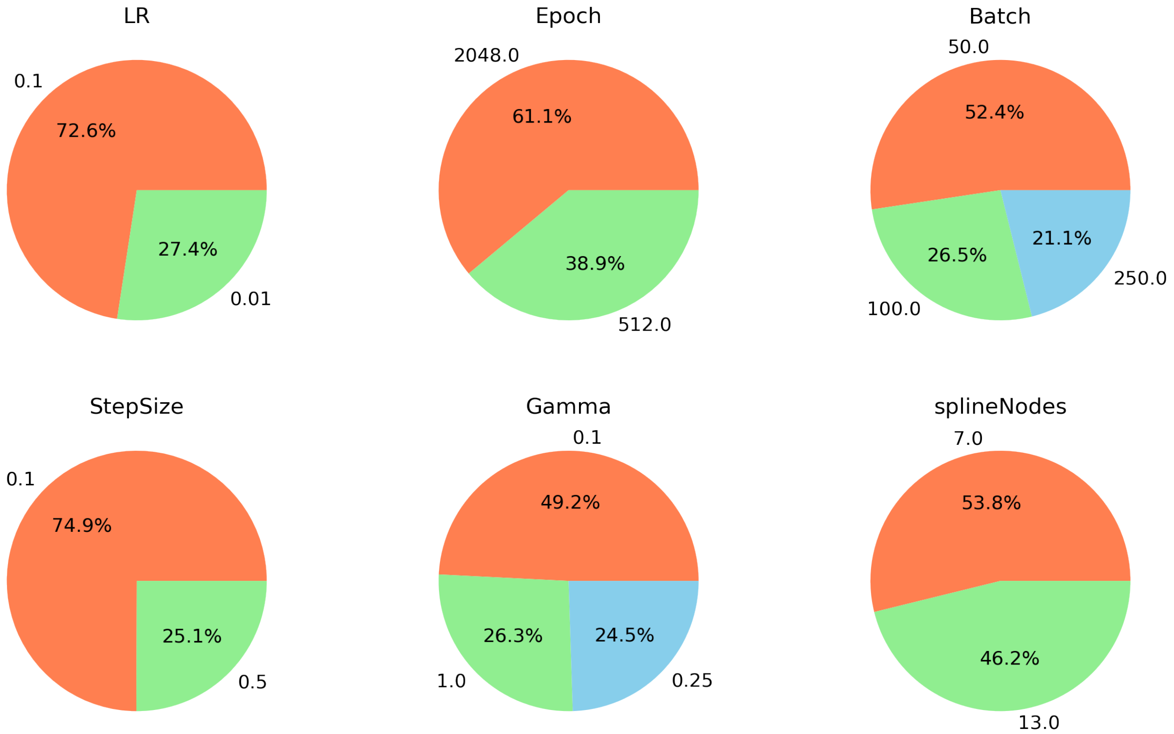
Note: This figure displays the cross-sectional variation of the optimal hyperparameters for the flexible HAR model.

Figure S.7: Cubic HAR model hyperparameters



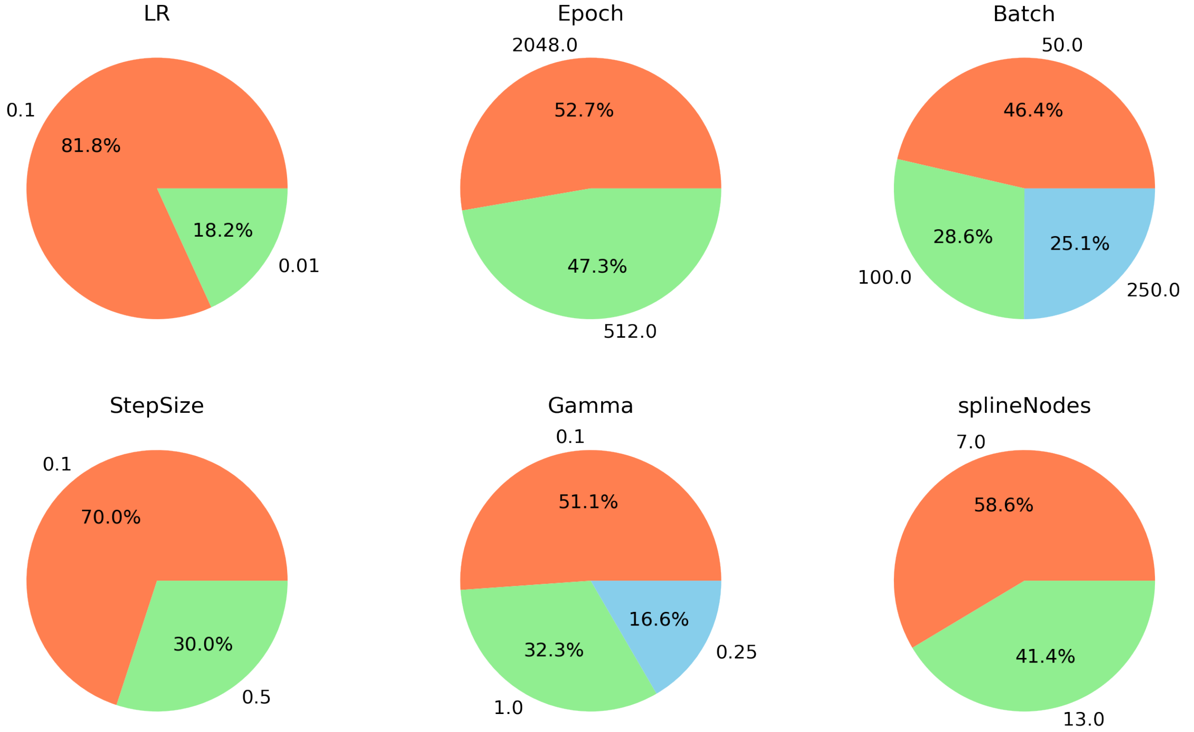
Note: This figure displays the cross-sectional variation of the optimal hyperparameters for the cubic HAR model.

Figure S.8: Cubic-TOD model hyperparameters



Note: This figure displays the cross-sectional variation of the optimal hyperparameters for the cubic-TOD model.

Figure S.9: Cubic-EW model hyperparameters



Note: This figure displays the cross-sectional variation of the optimal hyperparameters for the cubic-EW model.

Additional References

[SuppApp1] Andersen, T.G., Bollerslev, T., 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International economic review , 885–905.

[SuppApp7] Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71, 579–625.

[SuppApp12] Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2009. Realized kernels in practice: Trades and quotes. Econometrics Journal 12, 1–32.

[SuppApp4] Patton, A.J., 2011. Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics 160, 246–256.