Does Adding Social Media Sentiment Upstage Admitting Ignorance when Forecasting Volatility*

Steven Lehrer†, Tian Xie‡, and Xinyu Zhang§

†Queen's University, NYU Shanghai, and NBER, lehrers@queensu.ca
‡Xiamen University, WISE and SOE, xietian@xmu.edu.cn
§Academy of Mathematics and Systems Science, Chinese Academy of Sciences, xinyu@amss.ac.cn

March 15, 2019

Abstract

One of the challenges in forecasting volatility is that the specification of the lag index to capture the persistence of volatility is not well motivated by theory and as such is plagued by model uncertainty. This study explores whether there are benefits for forecasting volatility from (i) acknowledging this specification uncertainty of the lag index in the heterogeneous autoregression model, and (ii) incorporating social media data. We first develop a model averaging heterogeneous autoregression estimator and prove that it is asymptotically optimal. Second, we use a deep learning algorithm on a 10% random sample of Twitter messages to construct a measure of sentiment at the one-minute level. Our empirical results contrasting alternative estimators finds that jointly incorporating model averaging techniques and sentiment measures from social media can significantly improve the forecasting accuracy of financial volatility.

JEL classification: C52, C53, G12, G17
Keywords: HAR, Model Averaging, Volatility Forecasting, Social Media Big Data

*We wish to thank Bruno Solnik, Jun Yu, Lu Zhang, seminar participants at the Young Econometricians around Pacific (YEAP) 2017 annual conference in Shanghai, Financial Econometrics and New Finance (FEFN) 2018 conference in Hangzhou, EcoSta 2018 annual conference in Hong Kong, ZEW Conference on Information and Communication Technologies, Chinese Academy of Sciences, Renmin University, Zhejiang University, and Xiamen University for helpful comments and suggestions. Lehrer wishes to thank SSHRC for research support. Zhang's research is supported by the Natural Science Foundation of China (71522004, 11471324 and 71631008). Xie's research is supported by the Natural Science Foundation of China (71701175), the Chinese Ministry of Education Project of Humanities and Social Sciences (17YJC790174), the Natural Science Foundation of Fujian Province of China (2018J01116), the Fundamental Research Funds for the Central Universities in China (20720171002, 20720171076, and 20720181050), and Educational and Scientific Research Program for Young and Middle-aged Instructor of Fujian Province (JAS170018). Xie is grateful to Jiarong Chen for research support. The usual caveat applies.

1 Introduction

Volatility is the most important variable in the pricing of derivative securities. Further, regulators require the financial industry to incorporate time-varying volatility in their risk management models. Given the interest in having accurate volatility forecasts, a burgeoning literature has developed estimators that often include lagged values of the dependent variable to forecast financial market volatility. The inclusion of lagged dependent variables is motivated by the well established property that volatility possesses a slow decaying autocorrelation function, usually termed long-memory.[1] To contrast the performance of different estimators developed for volatility forecasting, assessments are commonly structured to hold the test asset and estimation strategy fixed, but vary the model specification of the lag index.

In this paper, we first propose a model averaging estimator as a means to solve any potential uncertainty in the specification of the lag index of the heterogeneous autoregressive realized volatility model (HAR) of Corsi (2009). Since the HAR is easy to implement it has become one of the most popular estimators to forecast indices such as the Chicago Board Options Exchange Volatility Index (VIX). This estimator can also approximate the long memory and multiscaling properties of realized volatility.[2] We extend the prediction model averaging (PMA) estimator of Xie (2015) to the dynamic HAR framework.[3] A

[1]This phenomenon has also been uncovered by Dacorogna, Műller, Nagler, Olsen, and Pictet (1993), Andersen, Bollerslev, Diebold, and Labys (2001b) for the foreign exchange market and Andersen, Bollerslev, Diebold, and Ebens (2001a) for stock market returns.

[2]Fernandes, Medeiros, and Scharth (2014) perform a thorough statistical examination of the time-series properties of the VIX index and propose using the HAR-type models for modeling and forecasting purposes. See Vortelinos (2017) for further discussion.

[3]Corsi, Audrino, and Renò (2012) point out that most HAR-type models fixate on using the lag index $[1, 5, 22]$ to reflect the daily, weekly, and monthly contributions to the volatility process. Our approach differs from both Wang, Ma, Wei, and Wu (2016) and Craioveanu and Hillebrand (2012) who respectively considered a Bayesian model averaging approach to deal with specification uncertainty on the control vari-

model averaging estimator generates a weighted average model using all the approximation models and can be viewed as general case of forecast combination.[4] We prove that the model averaging HAR (henceforth MAHAR) estimator is asymptotically optimal in the sense of achieving the lowest possible mean squared error.

A secondary aim of this study is to examine whether broad measures of social media sentiment calculated by a deep learning algorithm improve volatility forecasts. The potential for machine learning and social media data to influence stock markets is tantalizing to many, and in response numerous companies have already rapidly developed products to measure the mood or sentiment on social media platforms for individual investors and hedge funds.[5] Beyond understanding if social media sentiment holds significant explanatory power, our application illustrates how one needs to properly aggregate social media data that is measured at a higher frequency than many measures of volatility.[6] This analysis informs volatility modelling a subject of interest to practionners in risk management and asset allocation.

To examine the empirical performance of the proposed MAHAR estimator, we first conduct a Monte Carlo study and second undertake an assessment of forecasting the VIX using data collected between 2013-2017. We compare MAHAR to the original HAR

ables in the various HAR-type models and a parallel computing method to investigate the lag structure of HAR-like models.

[4]Bates and Granger (1969) introduces forecast combination and Timmermann (2006) provides a recent survey of developments in forecast combination a popular method since numerous studies have found it yields better forecasts on average than methods based on the ex ante best individual forecasting model.

[5]The idea that sentiment and emotions of market participants can influence the performance of the stock market is often discussed in both the media and research studies in the field of behavioral finance. As such, firms are developing products that trade on this information posted on social media. Additional motivation for our study comes from Ted Merz, the content business manager at Bloomberg who recently stated, "Trump tweets cause volatility is unquestionable."

[6]This aggregation is consistent with Nofer and Hinz (2015), who find that one must properly account for spread of mood states on Twitter. It is worth noting that prior work (see e.g. Zhang, Fuehres, and Gloor, 2011 and Bollen, Mao, and Zheng, 2011) documenting associations between social media sentiment and financial market performance generally uses a less sophisticated sentiment analysis.

model of Corsi (2009) and several popular extensions to the multivariate model. These extensions include Fernandes et al. (2014) who allows for asymmetric effects on volatility dynamics and the estimator of Audrino and Knaus (2016), which uses the Lasso to determine the components of the lag index.[7] Put differently, the potential success of model averaging mirrors forecast combinations and is related to the evidence that a decision maker in almost all cases cannot identify ex ante the exact true data generating process, but different models play a complementary role in approximating it.[8]

Our empirical results find that the MAHAR estimator has significantly greater forecast accuracy than other HAR-type estimators. Further, we present evidence that suggests incorporating Twitter sentiment measures can greatly improve forecasts of the VIX index. Specifically, we find that including sentiment measures from Twitter significantly improves forecast accuracy in the short term and that their forecasting power declines as we increase the forecasting horizon.

The remainder of this paper is organized as follows. In the next section, we discuss how the VIX index is measured and provide a more detailed overview of existing HAR strategies used to forecast the VIX index. In Section 3 we propose the MAHAR strategy, prove that it is asymptotically optimal and compare the finite sample performance of MAHAR relative to other HAR estimators using a Monte Carlo study. Our empirical application and data that we utilize including how Twitter sentiment measures are calculated and aggregated to be measured at the same daily frequency as the VIX index is described in Section 4. We also discuss the results of the out of sample forecasting exer-

[7]In the Section 2, we review extensions to the HAR model that accommodate other features including jumps, leverage effects, and other nonlinear behaviors.

[8]Since the VIX is measured on a daily basis, we do not consider volatility prediction models based on high-frequency realized volatility measures and stochastic volatility models.

cise that provide strong evidence of the benefits of both the MAHAR estimator relative to other HAR estimators and incorporating social media data when forecasting volatility. Our conclusions are presented in Section 5.

2 Strategies to Forecast the VIX Index

The Chicago Board Options Exchange Volatility Index (VIX) is a popular measure of the stock market's expectation of volatility implied by S&P 500 index options. It is colloquially referred to as the "fear index" or the "fear gauge". The VIX index measures the market expectations of the near-term volatility implied by stock index option prices. The VIX index is calculated as a weighted blend of prices for a range of options on the S&P 500 index,[9] using the following formula

$$\mathrm{VIX} = 100 \cdot \sqrt{\frac{2}{T} \sum_i \frac{\Delta K_i}{K_i^2} e^{rT} Q(K_i) - \frac{1}{T} \left(\frac{F}{K_0} - 1 \right)^2}, \tag{1}$$

where T is time to expiration, F is the forward index level derived from the index options prices, K_i is the strike price of the ith out-of-the-money option, $\Delta K_i = (K_{i+1} - K_{i-1})/2$, K_0 is the first strike below the forward index level, r is the risk-free interest rate to expiration, and $Q(K_i)$ is the mid-quote for the option with strike of K_i.

Following Fernandes et al. (2014), we use the HAR model of Corsi (2009) to approximate the VIX index. This model postulates that h-step-ahead daily volatility y_{t+h} can be

[9]The VIX is quoted in percentage points and represents the expected range of movement in the S&P 500 index over the next year, at a 68% confidence level (i.e. one standard deviation from the mean under the normal probability density curve).

modeled as

$$y_{t+h} = \beta_0 + \beta_d \bar{y}_t^{(1)} + \beta_w \bar{y}_t^{(5)} + \beta_m \bar{y}_t^{(22)} + \epsilon_{t+h}, \tag{2}$$

where we define

$$\bar{y}_t^{(l)} \equiv l^{-1} \sum_{s=1}^{l} y_{t-s} \tag{3}$$

as the averages of the previous l periods of y from period t and $\{\epsilon_t\}_t$ is a zero mean innovation process. Practitioners often select a *lag index vector* $\boldsymbol{l} = [1, 5, 22]$, to mirror the daily, weekly, and monthly components of the volatility process being forecasted.

The HAR model in Equation (2) can be estimated using OLS and Fernandes et al. (2014) allow for a more persistent volatility process, $\boldsymbol{l} = [1, 5, 10, 22, 66]$. It is simple to incorporate a K-dimensional set of exogenous regressors $\boldsymbol{z}_t = [z_{1t}, ..., z_{Kt}]$ into Equation (2), generating

$$y_{t+h} = \beta_0 + \beta_d \bar{y}_t^{(1)} + \beta_w \bar{y}_t^{(5)} + \beta_m \bar{y}_t^{(22)} + \boldsymbol{z}_t \boldsymbol{\beta}_z + \epsilon_{t+h}, \tag{4}$$

where $\boldsymbol{\beta}_z$ represent the associated coefficients for $\boldsymbol{z}_t$. Note that all the exogenous variables $\boldsymbol{z}_t$ are measured h periods before the dependent variable y_{t+h}. We refer to the above HAR model with exogenous variables as HARX, and through the remainder of the text, X refers to specifications that additionally control for exogenous covariates.[10]

A popular variant of the HARX specification allows the exogenous variables $\boldsymbol{z}_t$ to

[10]The univariate HAR was first extended to allow for jumps, leverage effects, and other nonlinear behaviors to the multivariate context. Corsi, Pirino, and Renò (2010) use the C-Tz test for jumps detection and threshold bipower variation to estimate relevant parameters. Bollerslev, Litvinova, and Tauchen (2006) stress the importance of allowing for the asymmetric leverage effects. Patton and Sheppard (2015) extend the HAR model by incorporating the realized semivariances variables first proposed in Barndorff-Neilsen, Kinnebrouk, and Shephard (2010). More complicated nonlinear effects, including structural breaks and regime-switches, are modeled by McAleer and Medeiros (2008) and Scharth and Medeiros (2009). Corsi et al. (2012) provide a comprehensive survey on the development of HAR-type models.

have asymmetric effects on volatility dynamics and is known as the asymmetric HARX (henceforth AHARX) model. Following Fernandes et al. (2014), AHARX is expressed as

$$y_{t+h} = \beta_0 + \beta_d \bar{y}_t^{(1)} + \beta_w \bar{y}_t^{(5)} + \beta_m \bar{y}_t^{(22)} + \boldsymbol{z}_t^- \boldsymbol{\beta}_z^- + \boldsymbol{z}_t^+ \boldsymbol{\beta}_z^+ + \epsilon_{t+h}, \tag{5}$$

where $\boldsymbol{z}_t^- = [z_{1t}^-, ..., z_{Kt}^-]$ and $\boldsymbol{z}_t^+ = [z_{1t}^+, ..., z_{Kt}^+]$ for $k = 1, ..., K$, with

$$z_{kt}^- = \begin{cases} z_{kt}\mathbb{I}(z_{kt} < 0) & \text{if a return} \\ z_{kt}\mathbb{I}(\Delta z_{kt} < 0) & \text{if in levels} \end{cases} \qquad z_{kt}^+ = \begin{cases} z_{kt}\mathbb{I}(z_{kt} > 0) & \text{if a return} \\ z_{kt}\mathbb{I}(\Delta z_{kt} > 0) & \text{if in levels} \end{cases}$$

and $\mathbb{I}(\cdot)$ is an indicator function that equals one if the argument of the function is satisfied. Since the VIX index is calculated daily based on implied volatility, we focus on the *original* HAR models and do not consider the *innovation-type* HAR models.

Each HAR model defined in equations (2), (4), and (5) can be reexpressed in compact matrix form as

$$y_{t+h} = \boldsymbol{x}_t(\boldsymbol{l})\boldsymbol{\beta} + \epsilon_{t+h}, \tag{6}$$

where $\boldsymbol{\beta}$ is the $L \times 1$ coefficient vector and $\boldsymbol{x}_t(\boldsymbol{l})$ is the independent variable for some $L \times 1$ vector of lag indices $\boldsymbol{l} = (l_1, ..., l_p) \in \mathbb{Z}_+^p$ with p being the maximum positive order of lags, such that

$$\boldsymbol{x}_t(\boldsymbol{l}) \equiv \begin{cases} [1, \bar{y}_t^{(l_1)}, ..., \bar{y}_t^{(l_p)}] & \text{for model HAR in (2),} \\ [1, \bar{y}_t^{(l_1)}, ..., \bar{y}_t^{(l_p)}, \boldsymbol{z}_t] & \text{for model HARX in (4),} \\ [1, \bar{y}_t^{(l_1)}, ..., \bar{y}_t^{(l_p)}, \boldsymbol{z}_t^-, \boldsymbol{z}_t^+] & \text{for model AHARX in (5).} \end{cases} \tag{7}$$

Since Model (6) contains an infeasible y_{t+h}, in practice the series $\{y_t, 1 \leq t \leq T\}$ is presumably approximated by a HAR-type model with h-period-ahead forecast horizon if it satisfies

$$y_t = \boldsymbol{x}_{t-h}(\boldsymbol{l})\boldsymbol{\beta} + \epsilon_t \quad \text{for any } h \in \mathbb{Z}_+, \tag{8}$$

where ϵ_t is generic (weak) white noise. Currently, the choice of index vector $\boldsymbol{l}$ in the literature is generally made as either $\boldsymbol{l}' \equiv [1,5,22]$ or $\boldsymbol{l}'' \equiv [1,5,10,22,66]$. These choices for $\boldsymbol{l}$ appear to be made to mirror intuitive concepts like daily, weekly, and monthly components of the volatility process, rather than have a formal justification. Further, Corsi (2009) demonstrate that the out-of-sample performance of the standard HAR model with a fixed $\boldsymbol{l}'$ is often unstable and can depend on the level of noise contained in the underlying pricing series of the asset considered.

Since the arrival frequency of new information changes an asset's volatility, it will be important for forecast accuracy. However, the arrival frequency is unknown to the forecaster. Audrino and Knaus (2016) propose using the least absolute shrinkage and selection operator (Lasso) to determine the components of the lag index vector for the HAR model. For $\boldsymbol{l} = [1,2,...,p]$, the Lasso estimator considers the constrained least-squares estimates for Model (8) subject to the constraint $\sum_{j=1}^{L} |\beta_j| < c$ for some constant c. Audrino and Knaus (2016) show that the Lasso recovers the lag index structure of the HAR model asymptotically. This estimator crucially relies on the assumption that the HAR model is the true model that generates the data. We denote this estimator as LHAR and Audrino and Knaus (2016) find that LHAR does not exhibit gains in performance relative to the conventional HAR model. This result is not surprising since Belloni and Chernozhukov (2013) show that the Lasso can induce bias and asymptotic risk.[11] In the next section,

[11] In general, the benefits from applying the Lasso method instead of OLS exist in some cases where either the number of regressors exceeds the number of observations since it involves shrinkage, or in other cases

we propose an alternative estimator that does not make this strong assumption and can allow for more general specifications of the contents of the index vector than either the HAR, HARX or AHARX models.

3 A New Strategy to Forecast Volatility

To circumvent the issue of lag index uncertainty, we suggest applying a model averaging estimator to the HAR framework. We name this method as model averaging HAR (MAHAR) and the framework is flexible to incorporate control variables.[12] The starting point is to assume there is a set of M candidate models that approximate the unknown DGP. Within any candidate model, r denotes the maximal lag order of y_t available for $t = -r + 1, ..., T$, so that the vector of observations of the dependent variable $[y_1, ..., y_T]^\top$ in every candidate model is $[y_1, ..., y_T]^\top$. Thus,

$$\bar{\boldsymbol{y}}^{(l)} = \left[\bar{y}_1^{(l)}, ..., \bar{y}_T^{(l)}\right]^\top \quad and \quad \boldsymbol{X} = \left[\boldsymbol{1}, \bar{\boldsymbol{y}}^{(1)}, ..., \bar{\boldsymbol{y}}^{(r)}, \boldsymbol{Z}\right],$$

where $\bar{y}_t^{(l)}$ is defined in Equation (3) and $\boldsymbol{Z} = [Z_1, \ldots, Z_T]^\top$ is a $T \times q$ matrix containing explanatory regressors.[13] We assume q is fixed and $\boldsymbol{X}$ is assumed to have full column rank in the unrestricted model.[14] The unrestricted model can be expressed as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_T]^\top$.

The regressor matrix $\boldsymbol{X}^m$ of the m^{th} candidate model is formed by combining columns

where the number of parameters is not small relative to the sample size and some form of regularization is necessary. However, these cases do not apply to our VIX modeling exercises.

[12]Note, the technique derived in this paper can be easily applied to other forecasting models.

[13]These regressors may demonstrate significant forecasting power on y_t in a finite sample.

[14]Note this matrix $\boldsymbol{X}$ differs from the $\boldsymbol{Y}_L$ term in Zhang, Wan, and Zou (2013), since the former includes $\bar{\boldsymbol{y}}^{(l)}$ terms, whereas the latter consists of lagged $\boldsymbol{y}$ terms.

in $\left[\bar{\boldsymbol{y}}^{(1)}, ..., \bar{\boldsymbol{y}}^{(r)}\right]$, $\mathbf{1}$, and $\boldsymbol{Z}$. If we collect the lag index of the component in $\boldsymbol{X}$ into the vector $\boldsymbol{l}_{\text{full}} \equiv [1, 2, ..., r]$, the lag index vector of $\boldsymbol{X}^m$, denoted as $\boldsymbol{l}^m$, is simply a subset of $\boldsymbol{l}_{\text{full}}$. The m^{th} candidate model can be written as

$$\boldsymbol{y} = \boldsymbol{X}^m \boldsymbol{\beta}^m + \boldsymbol{\epsilon}^m,$$

where $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_T]^\top$. Note that all candidate models and even the unrestricted model can be misspecified. The least squares estimate of $\boldsymbol{\beta}^m$ is $\hat{\boldsymbol{\beta}}^m = \left(\boldsymbol{X}^{m\top} \boldsymbol{X}^m\right)^{-1} \boldsymbol{X}^{m\top} \boldsymbol{y}$. Following Hansen (2008), the optimal mean-square forecast is the conditional mean μ_{T+1}. Therefore, the least-squares forecast of y_{T+1} from the m^{th} approximation model is then $\hat{y}^m_{T+1} = \hat{\mu}^m_{T+1} = \boldsymbol{x}^{m\ \top}_{T+1} \hat{\boldsymbol{\beta}}^m$. Thus, $\hat{\boldsymbol{\mu}}^m$ is defined as the vector of the estimated conditional means from model m.

We obtain forecasts of y_{T+1} from all approximation models and define the vector

$$\hat{\boldsymbol{y}}_{T+1} \equiv \left[\hat{y}^1_{T+1}, \hat{y}^2_{T+1}, \cdots, \hat{y}^M_{T+1}\right]^\top. \tag{9}$$

The model averaging forecast is simply the weighted average of $\hat{\boldsymbol{y}}_{T+1}$, such that

$$\hat{y}_{T+1}(\boldsymbol{w}) \equiv \boldsymbol{w}^\top \hat{\boldsymbol{y}}_{T+1} = \sum_{m=1}^{M} w^m \hat{y}^m_{T+1},$$

where $\boldsymbol{w} = \left[w^1, ..., w^M\right]^\top$ is a weight vector in the unit simplex in $\mathbb{R}^M$

$$\mathcal{H} \equiv \left\{ \boldsymbol{w} \in [0, 1]^M : \sum_{m=1}^{M} w^m = 1 \right\}.$$

We require the weights to be non-negative and sum to one, since Hansen (2008) and Xie

(2015) suggest that relaxing the non-negativity restriction in least squares model averaging estimation can result in poor empirical performance.

The performance of model averaging forecast crucially depends on the criterion used to estimate the weight vector $\boldsymbol{w}$. The MAHAR criterion aims to balance the fit and the complexity of a model:

$$\text{MAHAR}(\boldsymbol{w}) = \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\right)^\top \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\right) \left(\frac{T + k(\boldsymbol{w})}{T - k(\boldsymbol{w})}\right), \tag{10}$$

where $k(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^m k^m$ is the effective number of parameters and the model averaging estimator of the conditional mean is

$$\hat{\boldsymbol{\mu}}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^m \hat{\boldsymbol{\mu}}^m. \tag{11}$$

The empirical weight vector $\hat{\boldsymbol{w}}$, is obtained from the MAHAR criterion

$$\hat{\boldsymbol{w}} = \underset{w \in \mathcal{H}}{\arg\min}\, \text{MAHAR}(\boldsymbol{w}).$$

With the MAHAR empirical weight vector, we can calculate the model averaging forecast $\hat{y}_{T+1}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{y}}_{T+1}$. The MAHAR estimator is an extension of the PMA estimator of Xie (2015) in the HAR framework.

3.1 Asymptotic Optimality of MAHAR

Proving the asymptotic optimality of the MAHAR estimator relies on applying results in mathematical statistics developed by Ing and Wei (2003), in which a crucial condition requires y_t to follow the stationary AR(∞) process

$$y_t = \sum_{l=1}^{\infty} a_l y_{t-l} + e_t, \quad i = ..., -1, 0, 1, ... \tag{12}$$

where e_t follows the i.i.d. process with mean 0 and variance σ^2, $\sum_{l=1}^{\infty} |a_l| < \infty$, and $1 + \sum_{l=1}^{\infty} a_l v^l$ is bounded away from zero for $|v| \leq 1$. For the results of Ing and Wei (2003) to be applicable to our setting, we first demonstrate that the HAR model can be rewritten as a restricted AR process. We now define the projection matrices $\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}$ and $\boldsymbol{M} \equiv \boldsymbol{I}_T - \boldsymbol{P}$.

Following Corsi (2009), the HAR model with lag order vector $\boldsymbol{l} = [l_1, l_2, ...l_p] \in \mathbb{Z}_+^p$ ($l_1 < l_2 < ... < l_p$) is closely related to the conventional AR(l_p) model. For simplicity, if we ignore any control variable, we can rewrite the HAR model

$$y_t = \beta_0 + \beta_1 \bar{y}_t^{(l_1)} + \beta_2 \bar{y}_t^{(l_2)} + ... + \beta_p \bar{y}_t^{(l_p)} + \epsilon_i,$$

as a restricted AR(l_p) model, in which

$$y_t = \beta_0 + \sum_{s=1}^{l_p} \phi_s^{\text{HAR}} y_{t-s} + \epsilon_i \tag{13}$$

with restrictions

$$\phi_s^{\text{HAR}} = \begin{cases} \frac{1}{l_1}\beta_1 + \frac{1}{l_2}\beta_2 + ... + \frac{1}{l_p}\beta_p & \text{for } s = 1 \text{ to } l_1 \\ \frac{1}{l_2}\beta_2 + ... + \frac{1}{l_p}\beta_p & \text{for } s = (l_1 + 1) \text{ to } (l_2 - 1) \\ & \vdots \\ \frac{1}{l_p}\beta_p & \text{for } s = (l_{p-1} + 1) \text{ to } l_p. \end{cases} \tag{14}$$

That is, the restrictions require the equality between certain specific coefficients among the set of l_p lag coefficients of AR(l_p).

Remark In the literature, it is conventional to set $\boldsymbol{l}' = [1,5,22]$ or $\boldsymbol{l}'' = [1,5,10,22,66]$. The former is proposed by Corsi (2009) to mirror the daily (d), weekly (w), and monthly (m) components of the volatility process, whereas the latter proposed in in Fernandes et al. (2014), mimics the daily, bi-weekly, monthly, and quarterly components of the time series process. Consider $\boldsymbol{l}'$, the HAR model

$$y_t = \beta_0 + \beta_{(d)}\bar{y}_t^{(1)} + \beta_{(w)}\bar{y}_t^{(5)} + \beta_{(m)}\bar{y}_t^{(22)} + \epsilon_t,$$

is simply a constrained AR(22) model

$$y_t = c + \sum_{s=1}^{22} \phi_s^{\text{HAR}} y_{t-s} + \epsilon_i,$$

where the restrictions imposed require

$$\phi_s^{\text{HAR}} = \begin{cases} \beta_d + \frac{1}{5}\beta_w + \frac{1}{22}\beta_m & \text{for } s = 1 \\ \frac{1}{5}\beta_w + \frac{1}{22}\beta_m & \text{for } s = 2, ..., 5 \\ \frac{1}{22}\beta_m & \text{for } s = 6, ..., 22. \end{cases}$$

Although a direct specification test of the validity of these restrictions may appear sensible, Audrino and Knaus (2016) points out this test usually leads to the rejection of the Null due to an excessive number of restrictions being imposed. Further, since the true DGP in (12) can never be achieved, every model used in practice is an approximation with specification error.

Let $\mu_t = E(y_t|y_{t-l}, y_{t-2}, \ldots)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T)^\top$. We consider the following squared error loss:

$$L_T(\boldsymbol{w}) = \|\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu}\|^2,$$

where $\hat{\boldsymbol{\mu}}(\boldsymbol{w})$ is the model averaging estimator of the conditional mean defined in Equation (11). Let $\boldsymbol{A}(\boldsymbol{w}) = \boldsymbol{I}_T - \boldsymbol{P}(\boldsymbol{w})$, $R_T(\boldsymbol{w}) = \|\boldsymbol{A}(\boldsymbol{w})\boldsymbol{\mu}\|^2 + \sigma^2 \text{tr}\{\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\}$, $\xi_T = \inf_{w \in \mathcal{H}_T} R_T(\boldsymbol{w})$, and $\boldsymbol{e} = (e_1, \ldots, e_T)^\top$. We demonstrate the asymptotic optimality of the MAHAR estimator in the following theorem.

Theorem 1 *The MAHAR estimator is asymptotically optimal such that*

$$\frac{L_T(\hat{\boldsymbol{w}})}{\inf_{w \in \mathcal{H}_T} L_T(\boldsymbol{w})} \xrightarrow{p} 1,$$

if the following conditions hold:

(C.1) $r^4 \xi_T^{-1} = o_p(1)$, $T^{-1}\boldsymbol{\mu}^\top \boldsymbol{\mu} = O_p(1)$, *and* $r^4 \boldsymbol{\mu}^\top \boldsymbol{P} \boldsymbol{\mu} \xi_T^{-2} = o_p(1)$ *;*

(C.2) $T^{-1/2}\boldsymbol{Z}^\top \boldsymbol{e} = O_p(1)$, $\mathcal{S}(T^{-1}\boldsymbol{Z}^\top \boldsymbol{Z}) = O_p(1)$, *and* $\mathcal{S}\left((T^{-1}\boldsymbol{Z}^\top \boldsymbol{M}\boldsymbol{Z})^{-1}\right) = O_p(1)$*;*

(C.3) there exist some positive constants α_1, α_2 *and* α_3 *such that* $|F_i(d_1) - F_i(d_2)| \leq \alpha_3 |d_1 - d_2|^{\alpha_1}$ *for all* i *when* $|d_1 - d_2| \leq \alpha_2$*; and*

(C.4) either $r^{6+\alpha_4} = O(T)$ *for some* $\alpha_4 > 0$ *and* $\sup_{-\infty < i < \infty} \mathbb{E} e_i^4 < \infty$*, or* $r^{2+\alpha_4} = O(T)$ *for some* $\alpha_4 > 0$ *and* $\sup_{-\infty < i < \infty} \mathbb{E} e_i^s < \infty$ *for all* s*,*

where $F_i(\cdot)$ *is the distribution function of* e_t*, and* $\mathcal{S}(\boldsymbol{B})$ *denotes the largest singular value of a matrix* $\boldsymbol{B}$*.*

See Appendix A for a detailed proof. Briefly, the first and third parts of Condition (C.1) place restrictions on the orders of r and ξ_T. The second part of Condition (C.1) concerns the average of μ_i^2's and a similar condition is used in Shao (1997). Condition (C.2) holds when $T^{-1}\boldsymbol{Z}^\top\boldsymbol{Z}$ and $T^{-1}\boldsymbol{Z}^\top\boldsymbol{M}\boldsymbol{Z}$ converge to positive definite matrices in probability and $\{Z_t e_t\}$ is a stationary and ergodic martingale difference sequence with finite fourth moments. Conditions (C.3) and (C.4) focus on properties of the least squares predictor with the observations generated by an AR(∞) process and were initially drawn from Ing and Wei (2003) and discussed in further detail in Zhang et al. (2013). Note that Condition (C.3) is mild and can be easily fulfilled by any distribution with a bounded density. Condition (C.4) requires a tradeoff between the existence of higher moments of e_i and the order r.

3.2 Monte Carlo Study

To examine the finite sample performance of the proposed MAHAR method, we conduct the following Monte Carlo study. We assume that the true series for y_t is generated by a long memory ARFIMA(p, d, q) process with a white noise term drawn from $N(0, \sigma^2)$. In our simulation design, we set $p = 0.8$, $q = 0.1$, $d = 0.3$, and $\sigma = 1$. We vary both the sample size ($T = 100, 200, 300,$ and 400) and forecast horizons ($h = 1, 2, 4$ and 8) across our study.

Since the true model form is unknown to researchers, they often construct forecasts of y_{t+h} using the following HAR model

$$y_{t+h} = \beta_0 + \beta_1 \bar{y}_t^{(1)} + ... + \beta_l \bar{y}_t^{(l)} + \epsilon_t,$$

for $l = 1, 2, ..., 10$.

Given $\max(\boldsymbol{l}) = 10$, we have in total $2^{10} = 1024$ approximation models; assuming each model includes a constant term. Xie (2015) points out that model screening should be considered to limit the number of approximation models whenever $\max(\boldsymbol{l})$ is a relatively large number, for both practical and theoretical considerations.[15] Thus, we compare the forecast accuracy of the MAHAR method with and without model screening, where we use the heteroscedasticity-robust model screening (HRMS) method of Xie (2017). In addition, we examine how the performance of the MAHAR method compares to both the conventional HAR model of Corsi (2009) and the Lasso-HAR method proposed in Audrino and Knaus (2016).

We compare the forecasting methods based on the following out-of-sample mean-square forecast error (MSFE),

$$\text{MSFE} = \frac{T}{\sigma^2}\left(\mathbb{E}\left(y_{t+h} - \hat{y}_{t+h}\right)^2 - \sigma^2\right).$$

In the above equation, we subtract the error variance σ^2 since it is also included in the leading term in the MSFE and is common across all forecast methods. The term (T/σ^2) is used to render the results scale-free. We compute the MSFE by computing averages across 10,000 simulation draws.

The results of the Monte Carlo experiments are presented in Table 1 where we use MAHAR_1 and MAHAR_2 to respectively denote with and without model screening. As the forecast horizon increases, all methods tend to yield higher MSFEs. Since we are pri-

[15]That said, in practice there remains the possibility that model screening accidentally removes useful candidate models, which would jeopardize forecasting accuracy.

marily interested in the relative performance of the four methods, we normalize all MSFEs by the MSFE of $MAHAR_2$ for all cases. Entries larger than one indicating performance of the respective estimator relative to MAHAR without model screening.

Table 1: Simulation Results

Horizon	HAR	LHAR	$MAHAR_1$	$MAHAR_2$
		$T = 100$		
1	1.1953	1.1326	1.0023	**1.0000**
2	1.1452	1.1163	1.0441	**1.0000**
4	1.1550	1.1451	1.0099	**1.0000**
8	1.1920	1.1803	**0.9979**	1.0000
		$T = 200$		
1	1.0667	1.0544	**0.9720**	1.0000
2	1.0528	1.0466	1.0244	**1.0000**
4	1.0316	1.0287	1.0062	**1.0000**
8	1.0326	1.0300	**0.9865**	1.0000
		$T = 300$		
1	1.0238	1.0145	**0.9839**	1.0000
2	1.0337	1.0305	1.0055	**1.0000**
4	1.0389	1.0380	1.0044	**1.0000**
8	1.0179	1.0170	**0.9997**	1.0000
		$T = 400$		
1	1.0167	1.0120	**0.9720**	1.0000
2	1.0284	1.0267	**0.9977**	1.0000
4	1.0134	1.0129	1.0018	**1.0000**
8	1.0158	1.0151	**0.9984**	1.0000

Note: Bold numbers with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the $MAHAR_2$ method presented in the last column.

Table 1 shows that the performance of $MAHAR_1$ and $MAHAR_2$ always beat HAR and LHAR. The superior performance of $MAHAR_1$ and $MAHAR_2$ is more significant in small sample size, and the performance of HAR and LHAR get closer to $MAHAR_1$ and $MAHAR_2$ as sample size increases. In all cases, we cannot distinguish $MAHAR_1$ from $MAHAR_2$. Not surprisingly, HAR is the most computationally efficient and $MAHAR_2$ is the least efficient since it creates candidate models constructed from the full permutation of all potential explanatory variables. Both LHAR and $MAHAR_1$ take more CPU time to execute than HAR. Taken together, we interpret these results as suggesting that proper model screening can dramatically reduce CPU time and be applied without jeopardizing forecast accuracy.

4 Data and Empirical Exercise

To examine the relative prediction efficiency of different HAR estimators with different ways of accounting for social media data, we conduct an h-step-ahead rolling window exercise of forecasting the VIX for different forecasting horizons. Table 2 lists each estimator analyzed in the exercise and for LHAR and LHARX* we choose the value for the tuning parameter (λ) that minimizes five-fold cross-validation. Contrasting forecasts between the second and third panel of the Table allows us to ascertain if social media data improves forecast accuracy and we use the $\star$ symbol to distinguish variants of the same estimator for specifications that additionally include social media data. When considering uncertainty in lag order specification for both LHARX and MAHARX (as well as LHARX* and MAHARX*), we let $\boldsymbol{l} = [1, 2, ..., 22]$ instead of $\boldsymbol{l}'$. Throughout the experiment, the window length is fixed at 600 observations (approximately half of the sample). We additionally varied window lengths and found the conclusions were robust to altering the size of the window.

Since social media data from Twitter is only recently available, we use data from January 7, 2013 to August 21, 2017 for this exercise. The evolution of the log of VIX index over this full sample period is presented in Figure 2(a). The VIX index has high degrees of persistence, a feature many empirical strategies account for by including lags of the VIX index as explanatory variables. To facilitate comparisons with Fernandes et al. (2014), we use an identical set of exogenous control variables. These variables include the logarithm

Table 2: List of Estimators

Panel A: Econometric Strategies Considered without Additional Controls		
(1)	RW	a random walk model
(2)	HAR	the conventional HAR model proposed in Corsi (2009) with $l = l'$
Panel B: Econometric Strategies Considered with Traditional Control Variables		
(3)	HARX	the conventional HAR model proposed in Corsi (2009) with $l = l'$
(4)	AHARX	the modified HAR model of Fernandes et al. (2014) with $l = l'$
(5)	LHARX	the Lasso-HAR method proposed in Audrino and Knaus (2016) with $l = [1, 2, ...22]$
(6)	MAHARX	the model averaging extension of the HAR estimator proposed in this paper with $l = [1, 2, ...22]$
Panel C: Econometric Strategies Considered with Traditional Control Variables and Social Media Sentiment		
(7)	HARX*	the conventional HAR model proposed in Corsi (2009) with $l = l'$
(8)	AHARX*	the modified HAR model of Fernandes et al. (2014) with $l = l'$
(9)	LHARX*	the Lasso-HAR method proposed in Audrino and Knaus (2016) with $l = [1, 2, ...22]$
(10)	MAHARX*	the model averaging extension of the HAR estimator proposed in this paper with $l = [1, 2, ...22]$

of the daily S&P500 index (SPX), volume of the S&P500 index (SPV), one-month crude oil futures contract (OIL), foreign exchange value of the US dollar index (USD), measures of the credit spread (CS), term spread (TS), and the difference between the effective and target Federal Funds rates (FFD).[16]

To measure social media sentiment, we use the deep learning algorithm developed in Felbo, Mislove, Søgaard, Rahwan, and Lehmann (2017), which is also used to create the Wall Street Journal - IHS U.S. Sentiment Index (USSI). In brief, for a random sample of 10% of all tweets every minute, the score is calculated as an equal tweet weight average of the sentiment values of the words within them.[17] Sentiment analysis is one of the main challenges in natural language processing and the algorithm we selected was trained on 124.6 million tweets containing emojis. The algorithm does not score individual emotion words in a Twitter message, but rather calculates a score based on the probability of each of 64 different emojis capturing the sentiment in the full Twitter message taking the

[16]The measure of credit spread is obtained from the excess yield of the Moody's seasoned Baa corporate bond over the Moody's seasoned Aaa corporate bond. The term spread (TS), is calculated as the difference between the 10-Year and 3-month treasury constant maturity rates.

[17]This is a 10% random sample of all tweets since the USSI was designed to measure the real time mood of the nation and the algorithm does not restrict the calculations to Twitter accounts that either mention any specific stock or are classified as being a market participant.

structure of the sentence into consideration. Thus, each emoji has a fixed score and the sentiment of a message is a weighted average of the type of mood being conveyed. Since the algorithm considers the ordering of all the words in a Twitter message, it should not be surprising that tests of the validity of the Felbo et al. (2017) algorithm with samples drawn from Amazon mechanical turk, have found it to be more accurate than competing sentiment algorithms. These competing algorithms range from less advanced methods that simply code messages using a binary indicator such as positive or not, to those based on scoring words via emotional valence.

Since the VIX index is measured at the daily level, we need to convert one-minute level USSI data to a daily measure. Prior work has shown that Mi(xed) Da(ta) S(ampling) regression (henceforth MIDAS regression) are useful in empirical practice where the relevant information is high frequency data, whereas the quantity of interest is a low frequency process. Lehrer, Xie, and Zeng (2018) propose the H-MIDAS strategy to weight observations from the high-frequency social media data to construct a measure at a lower frequency. The key feature is that the method allows for potential unsystematic heterogeneity in the weights used across time for the high frequency variable, thereby allowing for a more gradual depreciation relative to the common implementation of MIDAS.[18]

By using H-MIDAS, we transform 1,711,230 minute level observations for USSI variable into 1,206 daily observations for the USSI. The estimated weights used in the H-MIDAS transformation are presented in Figure 1. The weights display sharp variability in their importance throughout the day, particularly during time periods in which North American financial markets are closed. The weights are generally larger in the middle of

[18]This heterogeneity arises in part since social media users often revise their beliefs and comments in response to new information from the crowd. Further details of H-MIDAS are provided in Appendix C.

the graph that corresponds to the time when financial markets are operating.

Figure 1: Weights on the High Frequency Observations under Different Lag Indices

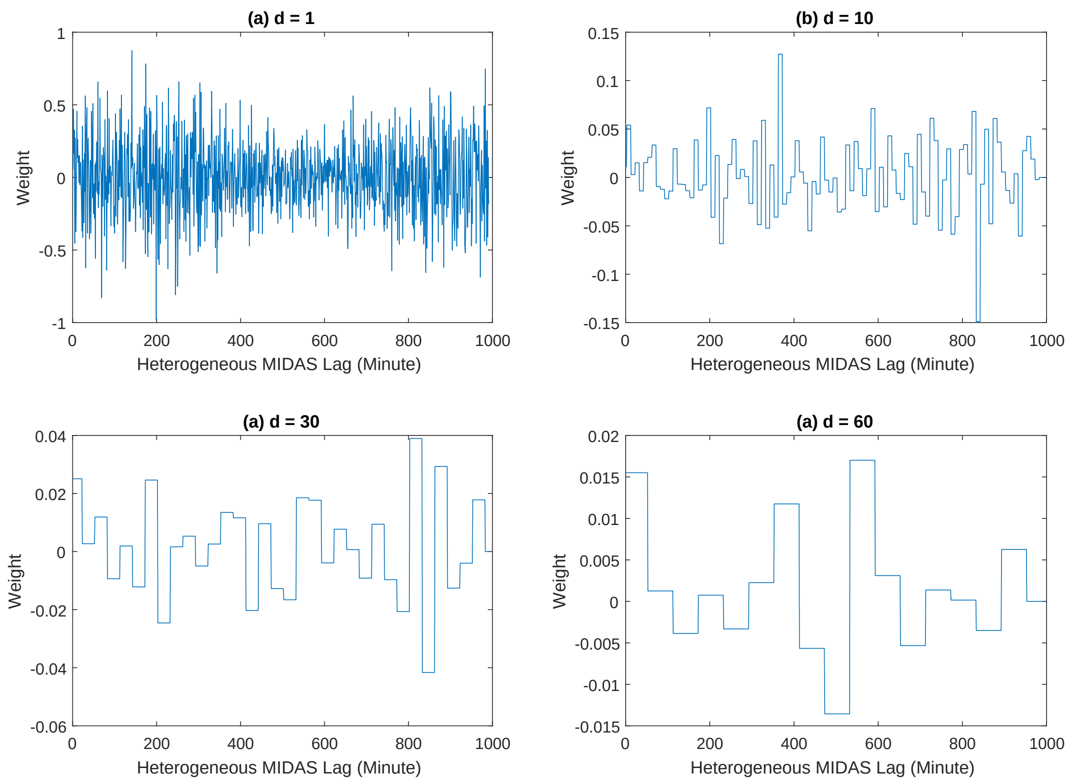
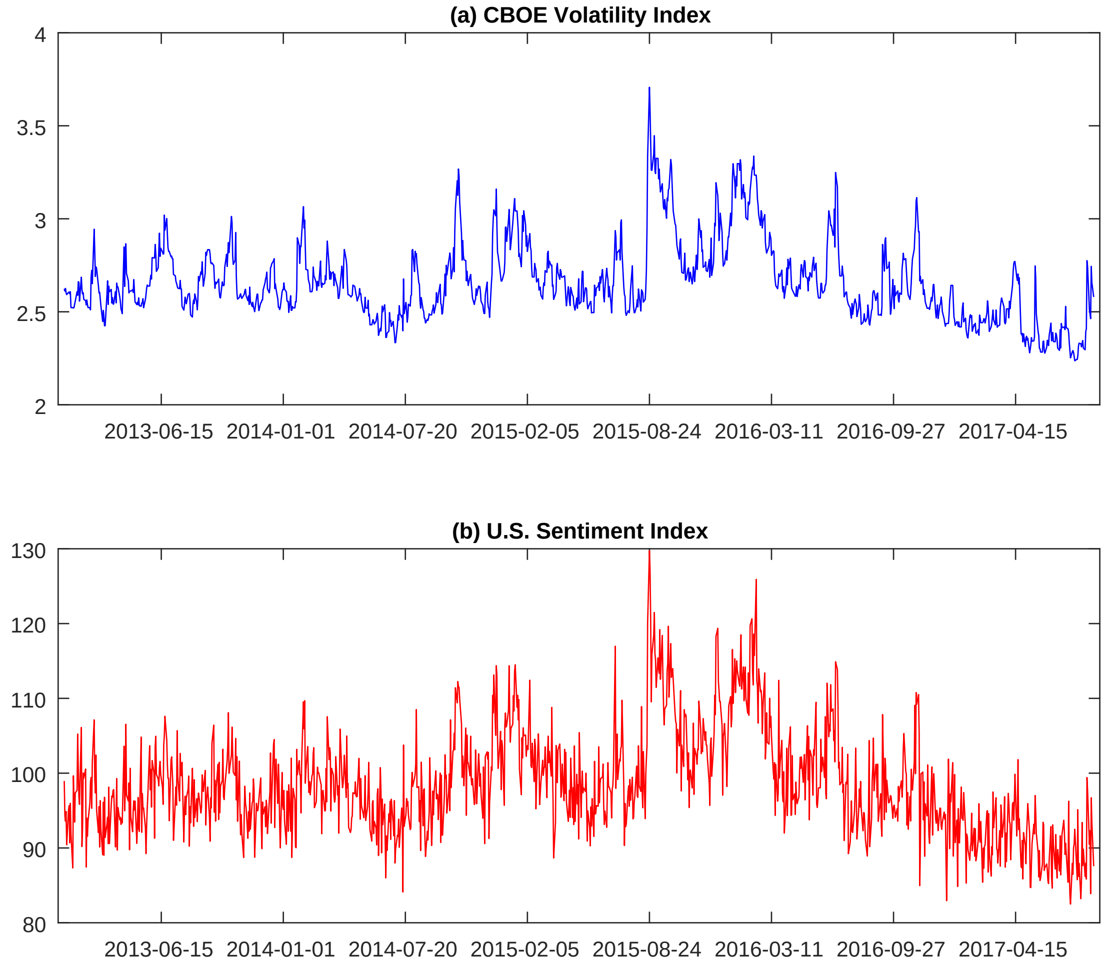


Figure 2(b) graphs the evolution of the one day lag of the H-MIDAS transformed USSI variable across the full sample period. The correlation between the VIX index and the one day lag of the USSI is 0.81. While the USSI series displays more volatility, ocular tests suggest that it does track trends in the VIX index presented in Figure 2(a).

Table 3 presents summary statistics for the data and *p*-values from both the Jarque-Bera test for normality and of the Augmented Dickey-Fuller (ADF) tests for unit root. With the exception of SPX, SPV and USD, each of the series exhibits tremendous variability and a large range across the sample period. Further, while none of the series are normally distributed, the ADF test concludes that the SPX, OIL, USD, CS, and TS series

Figure 2: The Daily Indices of VIX and USSI from January 7, 2013 to August 21, 2017



are nonstationary at 5% level. As such, we transformed each of these series to stationary by taking their first difference. ADF tests of first differences are presented in the last row of Table 3 and confirm that each transformed data series exhibits stationary.

4.1 Results of the Empirical Exercise

The results of the prediction experiment are presented in Table 4. The estimation strategy is listed in the first column and the remaining columns presents alternative criteria to

Table 3: Descriptive Statistics

	VIX	SPX	SPV	OIL	USD	CS	TS	FFD	USSI
Mean	2.6645	7.5913	21.9734	4.1445	4.4435	0.8979	2.0212	-0.0242	99.6868
Median	2.6254	7.6168	21.9662	3.9750	4.4885	0.8700	2.0300	-0.0150	98.5333
Minimum	3.7072	7.8164	22.7511	4.7026	4.5737	1.5400	2.9700	0.0350	133.7845
Maximum	2.2364	7.2842	20.1000	3.2869	4.2946	0.5300	1.0800	-0.3400	77.9024
Std. Dev.	0.2118	0.1232	0.1942	0.3808	0.0891	0.2329	0.4839	0.0409	9.3420
Skewness	1.0294	-0.4378	-0.9182	0.0892	-0.2776	0.9318	-0.0657	-1.0690	0.8104
Kurtosis	4.4941	2.6915	12.7020	1.5444	1.3478	3.2782	1.9709	6.3046	3.7692
Jarque-Bera	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ADF test	0.0000	0.0730	0.0000	0.8146	0.9606	0.9498	0.2269	0.0000	0.0000
*ADF test on Δ^1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

∗ We also perform ADF test on the first difference (Δ^1) of each data series.

evaluate the forecasting performance. The criteria include: (i) mean squared forecast error (MSFE), (ii) standard deviation of the forecast error (SDFE), (iii) mean absolute forecast error (MAFE), and (iv) the Mincer-Zarnowitz pseudo R^2. Each panel of Table 4 present result corresponding to scenarios of h-step-ahead forecasts of $[1, 5, 10, 22]$.

To ease interpretation, the results that identify the estimator with the best performance in each column of Table 4 is presented in bold. Across all h-step forecast lengths and forecasting performance criteria, MAHARX* method displays the best performance. For each panel, MAHARX demonstrates improved performance relative to HARX and MAHARX* similarly outperforms HARX*. The performance of Lasso based methods are unstable and both LHARX and LHARX* perform poorly in most forecast horizons. This set of results suggests that the uncertainty in the specification of the lag index cannot be fully accommodated by dimension reduction estimators.

Second, when comparing the panels of estimators across Table 4, there is clear evidence that those which incorporate the Twitter sentiment measure (denoted with $\star$) yield improved performance, irrespective of the forecast horizon. This result provides the first piece of suggestive evidence demonstrating the importance of using social media data in this VIX forecasting exercise.[19] The gains from including Twitter sentiment are stronger when the forecast horizon is shorter. Forecasting efficiency between MAHARX and MAHARX* improves by 30% in Panel A and the gains drop to 1.3% in Panel D.

To examine if the improvement in including social media sentiment data in forecasting exercise are statistically significant, we perform the modified Giacomini-White test (Giacomini and White, 2006) of the null hypothesis that the *column method* performs equally well as the *row method* in terms of MAFE. The corresponding p-values are presented in Table 5, in which each panel stands for a specific forecasting horizon varying between 1 to 22 days. To ease interpretation, the p-values smaller than 0.05 are presented in bold.

When the forecast horizon is a single day, the gains in forecast accuracy from MAHARX* relative to all other strategies are statistically significant. Further, all estimators that incorporate the Twitter sentiment exhibit significantly better performance relative to the same estimator that excludes social media data. This provides the strongest evidence of the benefits from incorporating social media data, although we note that the benefits decrease and become statistically insignificant with longer forecast horizons. The test statistics provide convincing evidence that accommodating model uncertainty with model averaging estimators is important in practice since irrespective of the forecast horizon, MAHARX significantly outperforms HARX; as MAHARX* outperforms HARX*.

To provide additional guidance to researchers, we examine the characteristics of the 5 models that were given the largest weight in the MAHARX* one-day ahead forecasting exercise with the full sample. Table 6 lists each of the potential explanatory variables in the first column and in columns 2 to 7 we denote which variables are included in the se-

[19] In Appendix B, we present additional illustrations of this benefit obtained from one-step-ahead forecasts calculated from OLS regressions that include the Twitter measure as an explanatory variable.

Table 4: Forecasting Performance at Different Horizons

	MSFE	SDFE	MAFE	Pseudo R^2
Panel A: 1 Days Ahead				
RW	0.0069	0.0830	0.0578	0.9036
HAR	0.0068	0.0828	0.0576	0.9041
HARX	0.0069	0.0830	0.0574	0.9036
AHARX	0.0069	0.0832	0.0576	0.9031
LHARX	0.0068	0.0826	0.0570	0.9040
MAHARX	0.0067	0.0817	0.0567	0.9064
HARX*	0.0049	0.0698	0.0514	0.9317
AHARX*	0.0049	0.0701	0.0516	0.9310
LHARX*	0.0051	0.0713	0.0544	0.9288
MAHARX*	**0.0047**	**0.0683**	**0.0503**	**0.9348**
Panel B: 5 Days Ahead				
RW	0.0294	0.1715	0.1202	0.5895
HAR	0.0280	0.1674	0.1206	0.6085
HARX	0.0264	0.1626	0.1195	0.6310
AHARX	0.0270	0.1645	0.1195	0.6225
LHARX	0.0272	0.1648	0.1153	0.6109
MAHARX	0.0256	0.1600	0.1166	0.6428
HARX*	0.0253	0.1592	0.1176	0.6465
AHARX*	0.0260	0.1611	0.1183	0.6377
LHARX*	0.0249	0.1577	0.1124	0.6499
MAHARX*	**0.0244**	**0.1563**	**0.1147**	**0.6589**
Panel C: 10 Days Ahead				
RW	0.0452	0.2127	0.1490	0.3705
HAR	0.0434	0.2084	0.1562	0.3950
HARX	0.0373	0.1930	0.1470	0.4801
AHARX	0.0394	0.1985	0.1491	0.4514
LHARX	0.0394	0.1986	0.1388	0.4354
MAHARX	0.0364	0.1907	0.1460	0.4923
HARX*	0.0360	0.1898	0.1434	0.4978
AHARX*	0.0406	0.2014	0.1516	0.4334
LHARX*	0.0365	0.1909	**0.1367**	0.4795
MAHARX*	**0.0353**	**0.1879**	0.1423	**0.5080**
Panel D: 22 Days Ahead				
RW	0.0589	0.2426	0.1758	0.1924
HAR	0.0595	0.2439	0.2007	0.1777
HARX	0.0397	0.1992	0.1543	0.4560
AHARX	0.0419	0.2047	0.1595	0.4266
LHARX	0.0464	0.2155	0.1596	0.3544
MAHARX	0.0383	0.1957	0.1510	0.4750
HARX*	0.0392	0.1980	0.1528	0.4629
AHARX*	0.0431	0.2076	0.1630	0.4086
LHARX*	0.0405	0.2013	0.1501	0.4347
MAHARX*	**0.0378**	**0.1943**	**0.1494**	**0.4830**

This table reports the out-of-sample results for predicting h-day future realized variation using the different predictor variables and risk models. The results are based on the CBOE VIX data spanning from January 7, 2013 to August 21, 2017 (a total of 1,206 observations). We use a rolling window of 600 observations to estimate the coefficients of the models, and evaluate the out-of-sample forecast performance at four horizons ($h = 1$, $h = 5$, $h = 10$ and $h = 22$). Each panel in the table corresponds to a specific forecast horizon, which varies from 1 day to 22 days. Bold numbers indicate the best performing model by each criterion at each forecast horizon.

Table 5: Giacomini-White Test Results

	RW	HAR	HARX	AHARX	LHARX	MAHARX	HARX*	AHARX*	LHARX*
Panel A: 1 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.8347	-	-	-	-	-	-	-	-
HARX	0.9996	0.7772	-	-	-	-	-	-	-
AHARX	0.8304	0.6840	0.7789	-	-	-	-	-	-
LHARX	0.8380	0.9806	0.8316	0.7193	-	-	-	-	-
MAHARX	0.2058	0.1042	**0.0096**	0.1229	0.1776	-	-	-	-
HARX*	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0001**	**0.0000**	-	-	-
AHARX*	**0.0000**	**0.0001**	**0.0001**	**0.0000**	**0.0003**	**0.0002**	0.5630	-	-
LHARX*	**0.0015**	**0.0021**	**0.0017**	**0.0009**	**0.0035**	**0.0042**	0.3073	0.5012	-
MAHARX*	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0014**	**0.0153**	**0.0229**
Panel B: 5 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.4343	-	-	-	-	-	-	-	-
HARX	0.2319	0.3853	-	-	-	-	-	-	-
AHARX	0.2230	0.5437	0.6841	-	-	-	-	-	-
LHARX	0.3241	0.9311	0.5630	0.6737	-	-	-	-	-
MAHARX	0.1268	0.1828	**0.0413**	0.3405	0.3692	-	-	-	-
HARX*	0.1018	0.1651	0.0055	0.2695	0.3104	0.6439	-	-	-
AHARX*	0.1012	0.2619	0.7310	0.1084	0.3928	0.7941	0.6384	-	-
LHARX*	**0.0057**	0.0917	0.4904	0.2435	**0.0429**	0.7960	0.8923	0.5978	-
MAHARX*	**0.0458**	0.0637	**0.0010**	0.0954	0.1783	**0.0070**	**0.0407**	0.2476	0.7183
Panel C: 10 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.6465	-	-	-	-	-	-	-	-
HARX	0.1902	0.1785	-	-	-	-	-	-	-
AHARX	0.1971	0.2367	0.5918	-	-	-	-	-	-
LHARX	0.2079	0.5727	0.6008	0.8332	-	-	-	-	-
MAHARX	0.1370	0.1215	0.1273	0.4254	0.5013	-	-	-	-
HARX*	0.1290	0.1132	**0.0120**	0.3850	0.4670	0.5237	-	-	-
AHARX*	0.4283	0.5571	0.4080	0.5396	0.9832	0.2715	0.2481	-	-
LHARX*	**0.0326**	0.1759	0.9929	0.6260	0.2505	0.8372	0.7674	0.5346	-
MAHARX*	0.1049	0.0894	**0.0087**	0.2864	0.4033	**0.0212**	0.1508	0.1731	0.6491
Panel D: 22 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.9119	-	-	-	-	-	-	-	-
HARX	0.0647	**0.0158**	-	-	-	-	-	-	-
AHARX	0.1078	**0.0378**	0.5248	-	-	-	-	-	-
LHARX	0.1451	0.2107	0.3053	0.4893	-	-	-	-	-
MAHARX	0.0510	**0.0104**	**0.0147**	0.3339	0.2365	-	-	-	-
HARX*	0.0584	**0.0143**	**0.0317**	0.4326	0.2722	0.1308	-	-	-
AHARX*	0.1513	0.0654	0.3425	0.3481	0.6298	0.2103	0.2807	-	-
LHARX*	**0.0458**	0.0684	0.7914	0.9170	0.1004	0.6307	0.7225	0.7651	-
MAHARX*	**0.0444**	**0.0089**	**0.0006**	0.2472	0.2043	**0.0314**	**0.0046**	0.1551	0.5563

The modified Giacomini-White test (Giacomini and White, 2006) is implemented to test the null hypothesis that the *row method* (in vertical headings) performs equally well as the *column method* (in horizontal headings) in terms of the absolute forecast error. Corresponding p values for a number of forecasting horizons ($h = 1, 5, 10, 22$) are reported in Panels A to D, respectively. Bold numbers indicate the null hypothesis can be rejected at 5% level of significance.

lected model with an "x". The models are arranged from that with the highest weight to the fifth highest weight and the overall MAHARX* model is presented in column 7. Notice first that consistent with other work using model averaging (e.g. Durlauf, Navarro, and Rivers, 2016, Lehrer and Xie (2017, 2018) among others), we observe that very few models account for the majority of the weights in the final MAHARX* estimate. The top three models account for 90% of the weight and four explanatory variables are common in these models: $\bar{y}_t^{(1)}$, $\bar{y}_t^{(21)}$, $\bar{y}_t^{(22)}$, and the USSI variable. The weekly lagged variable does not enter in any of the top models. The inclusion of the USSI in each of the top models reinforces the importance of incorporating the social media sentiment in volatility forecasts, even though the bottom row of the table shows that gain in the centered R^2 when models include the USSI is quite small and of the order of 3-5%.

In Appendix D, we first provide evidence of the robustness of our main findings to using an equal weighted measure of the USSI. We then conduct further analyses and statistical tests that indicates forecasts with the H-MIDAS transformed USSI are significantly more accurate than predictions using the equal weighted USSI. This latter finding is consistent with Nofer and Hinz (2015), who also show that one must properly account for how information on Twitter is transmitted over the course of the day when undertaking financial forecasts. Finally, adding confidence to the conclusions we draw in this section, is all of our initial analyses used a shorter time series for the forecasting exercise. We found each of our main findings were robust and strengthened as we added additional months of data from 2017, a period where the financial environment exhibited much lower volatility.

Table 6: Describing the 5 Highest Weight Models

	Model 1	Model 2	Model 3	Model 4	Model 5	MAHARX*
Weight	0.3086	0.2927	0.2263	0.1116	0.0546	
HAR 1-day Average	x	x	x	x	x	x
HAR 2-day Average						x
HAR 3-day Average						x
HAR 4-day Average						x
HAR 5-day Average						x
HAR 6-day Average						x
HAR 7-day Average						x
HAR 8-day Average						x
HAR 9-day Average						x
HAR 10-day Average						x
HAR 11-day Average					x	x
HAR 12-day Average						x
HAR 13-day Average						x
HAR 14-day Average		x		x		x
HAR 15-day Average		x		x	x	x
HAR 16-day Average		x		x		x
HAR 17-day Average						x
HAR 18-day Average						x
HAR 19-day Average						x
HAR 20-day Average				x	x	x
HAR 21-day Average	x	x	x			x
HAR 22-day Average	x	x	x	x	x	x
S&P500 1-day Return						x
S&P500 5-day Return						x
S&P500 10-day Return						x
S&P500 22-day Return						x
S&P500 Volume Change						x
Oil 1-day Return		x				x
Oil 5-day Return						x
Oil 10-day Return						x
Oil 22-day Return						x
USD Change						x
Credit Spread		x	x	x	x	x
Term Spread						x
FF Deviation	x	x	x	x	x	x
U.S. Sentiment Index	x	x	x	x	x	x
R^2 w/ SV.	0.9076	0.9084	0.9079	0.9083	0.9081	0.9092
R^2 w/o SV.	0.8737	0.8750	0.8743	0.8750	0.8746	0.8793

Note: Symbol "x" denotes that explanatory variable is included in the particular model, SV denotes social media data.

5 Conclusion

In this paper, we introduce a new model averaging estimator that allows for model uncertainty in the specification of the lag index when using the HAR to forecast volatility. We conduct h-step ahead prediction exercises for the VIX index to contrast the empirical performance of the MAHAR estimator to popular estimators to forecast volatility. Our empirical results find that the MAHAR estimator has significantly greater forecast accuracy than other HAR-type estimators. This result is consistent with the heterogeneous market hypothesis that suggests that financial market volatility arises since market participants interpret the same information in different ways according to their trading preferences and opportunities.

Further, our empirical analysis finds that when conducting forecasts in short time horizons, there are benefits from incorporating social media data. These results are robust to the inclusion of additional data that captures an extended period of low volatility which on the one hand is unsurprising since appropriate model specifications may vary across periods of different volatility. On the other hand, the continued importance of social media data demonstrates that this all user-aggregated data captures some relevant information related to investor sentiment. Extending the econometric methodology to both use less time-aggregated sources of social media data and forecast measures of realized volatility, present an agenda for future research.

References

ANDERSEN, T., T. BOLLERSLEV, F. DIEBOLD, AND H. EBENS (2001a): "The distribution of realized stock return volatility," *Journal of Financial Economics*, 61, 43–76.

ANDERSEN, T. G., T. BOLLERSLEV, F. X. DIEBOLD, AND P. LABYS (2001b): "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55.

AUDRINO, F. AND S. D. KNAUS (2016): "Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics," *Econometric Reviews*, 35, 1485–1521.

BARNDORFF-NEILSEN, O. E., S. KINNEBROUK, AND N. SHEPHARD (2010): "Measuring Downside Risk: Realised Semivariance," in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, ed. by T. Bollerslev, J. Russell, and M. Watson, Oxford University Press, 117–136.

BATES, J. M. AND C. W. J. GRANGER (1969): "The Combination of Forecasts," *OR*, 20, 451–468.

BELLONI, A. AND V. CHERNOZHUKOV (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19, 521–547.

BOLLEN, J., H. MAO, AND X. ZHENG (2011): "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2, 1–8.

BOLLERSLEV, T., J. LITVINOVA, AND G. TAUCHEN (2006): "Leverage and Volatility Feedback Effects in High-Frequency Data," *Journal of Financial Econometrics*, 4, 353–384.

CORSI, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7, 174–196.

CORSI, F., F. AUDRINO, AND R. RENÒ (2012): "HAR Modeling for Realized Volatility Forecasting," in *Handbook of Volatility Models and Their Applications*, John Wiley & Sons, Inc., 363–382.

CORSI, F., D. PIRINO, AND R. RENÒ (2010): "Threshold bipower variation and the impact of jumps on volatility forecasting," *Journal of Econometrics*, 159, 276 – 288.

CRAIOVEANU, M. AND E. HILLEBRAND (2012): "Why It Is OK to Use the HAR-RV $(1, 5, 21)$ Model," *Technical Report*, University of Central Missouri.

DACOROGNA, M. M., U. A. MŰLLER, R. J. NAGLER, R. B. OLSEN, AND O. V. PICTET (1993): "A geographical model for the daily and weekly seasonal volatility in the foreign exchange market," *Journal of International Money and Finance*, 12, 413 – 438.

DURLAUF, S. N., S. NAVARRO, AND D. A. RIVERS (2016): "Model uncertainty and the effect of shall-issue right-to-carry laws on crime," *European Economic Review*, 81, 32 – 67.

FELBO, B., A. MISLOVE, A. SØGAARD, I. RAHWAN, AND S. LEHMANN (2017): "Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm," *Machine Learning*, Accepted at EMNLP 2017.

FERNANDES, M., M. C. MEDEIROS, AND M. SCHARTH (2014): "Modeling and Predicting the CBOE Market Volatility Index," *Journal of Banking & Finance*, 40, 1–10.

GIACOMINI, R. AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

HANSEN, B. E. (2008): "Least-squares forecast averaging," *Journal of Econometrics*, 146, 342–350.

ING, C.-K. AND C.-Z. WEI (2003): "On same-realization prediction in an infinite-order autoregressive process," *Journal of Multivariate Analysis*, 85, 130 – 155.

KUERSTEINER, G. AND R. OKUI (2010): "Constructing Optimal Instruments by First-Stage Prediction Averaging," *Econometrica*, 78, 697–718.

LEHRER, S. F. AND T. XIE (2017): "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?" *The Review of Economics and Statistics*, 99, 749–755.

——— (2018): "The Bigger Picture: Combining Econometrics with Analytics Improve Forecasts of Movie Success," *Working Paper*.

LEHRER, S. F., T. XIE, AND T. ZENG (2018): "Does High Frequency Data Improve Our Confidence in Forecasts of Low Frequency Measures?" *Working Paper*.

MCALEER, M. AND M. C. MEDEIROS (2008): "A multiple regime smooth transition Heterogeneous Autoregressive model for long memory and asymmetries," *Journal of Econometrics*, 147, 104–119.

NOFER, M. AND O. HINZ (2015): "Using Twitter to Predict the Stock Market," *Business & Information Systems Engineering: The International Journal of WIRTSCHAFTSINFORMATIK*, 57, 229–242.

PATTON, A. J. AND K. SHEPPARD (2015): "Good Volatility, Bad Volatility: Signed Jumps and The Persistence of Volatility," *The Review of Economics and Statistics*, 97, 683–697.

SCHARTH, M. AND M. MEDEIROS (2009): "Asymmetric effects and long memory in the volatility of Dow Jones stocks," *International Journal of Forecasting*, 25, 304–327.

SHAO, J. (1997): "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–264.

Timmermann, A. (2006): "Chapter 4 Forecast Combinations," Elsevier, vol. 1 of *Handbook of Economic Forecasting*, 135–196.

Vortelinos, D. I. (2017): "Forecasting realized volatility: HAR against Principal Components Combining, neural networks and GARCH," *Research in International Business and Finance*, 39, 824 – 839.

Wang, Y., F. Ma, Y. Wei, and C. Wu (2016): "Forecasting Realized Volatility in A Changing World: A Dynamic Model Averaging Approach," *Journal of Banking & Finance*, 64, 136–149.

Whaley, R. E. (2000): "The Investor Fear Gauge," *The Journal of Portfolio Management*, 26, 12–17.

Xie, T. (2015): "Prediction Model Averaging Estimator," *Economics Letters*, 131, 5–8.

——— (2017): "Heteroscedasticity-robust Model Screening: A Useful Toolkit for Model Averaging in Big Data Analytics," *Economics Letters*, accepted.

Zhang, X., H. Fuehres, and P. A. Gloor (2011): "Predicting Stock Market Indicators Through Twitter, 'I hope it is not as bad as I fear'," *Procedia - Social and Behavioral Sciences*, 26, 55 – 62.

Zhang, X., A. T. Wan, and G. Zou (2013): "Model Averaging by Jackknife Criterion in Models with Dependent Data," *Journal of Econometrics*, 174, 82–94.

APPENDIX

A Proof

Proof of Theorem 1 Note that

$$\text{PMA}(\boldsymbol{w}) = L_T(\boldsymbol{w}) + \|\boldsymbol{e}\|^2 + 2\boldsymbol{e}^\top \boldsymbol{\mu} - 2\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu} + 2\sigma^2 \text{tr}\big(\boldsymbol{P}(\boldsymbol{w})\big) - 2\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}.$$

Since the terms $\|\boldsymbol{e}\|^2$ and $\boldsymbol{e}^\top \boldsymbol{\mu}$ are independent of the weight vector $\boldsymbol{w}$, Theorem 1 is valid if the following hold:

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\sigma^2 \text{tr}\big(\boldsymbol{P}(\boldsymbol{w})\big)\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0, \tag{A.1}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0, \tag{A.2}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0, \tag{A.3}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \left|\frac{L_T(\boldsymbol{w})}{R_T(\boldsymbol{w})} - 1\right| \xrightarrow{p} 0, \tag{A.4}$$

as $n \to \infty$.

Since for any $\delta > 0$,

$$\Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\sigma^2 \text{tr}\big(\boldsymbol{P}(\boldsymbol{w})\big)\right|}{R_T(\boldsymbol{w})} \geq \delta\right\} \leq \Pr\{r\xi_T^{-1} \geq \delta\sigma^{-2}\} \to 0$$

according to (C.1). We obtain (A.1).

Let $\boldsymbol{y}^{(-i)} = [y_{1-i}, ..., y_{n-i}]^\top$, $\boldsymbol{U} = [\boldsymbol{1}, \boldsymbol{y}^{(-1)}, \ldots, \boldsymbol{y}^{(-r)}, \boldsymbol{Z}]$. By conditions (C.2) – (C.4), the proof of (A.63) of Zhang et al. (2013), and Lemma 4 and Theorem 2 of Ing and Wei (2003), we have

$$T^{-1}r^{-1}\boldsymbol{e}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{e} = O_p(1) \tag{A.5}$$

and

$$T \cdot \mathcal{S}\left((\boldsymbol{U}^\top \boldsymbol{U})^{-1}\right) = O_p(1). \tag{A.6}$$

Let

$$\boldsymbol{B}^* = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1/2 & 1/3 & 1/4 & \cdots & 1/r \\ 0 & 0 & 1/2 & 1/3 & 1/4 & \cdots & 1/r \\ 0 & 0 & 0 & 1/3 & 1/4 & \cdots & 1/r \\ 0 & 0 & 0 & 0 & 1/4 & \cdots & 1/r \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1/r \end{pmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{pmatrix} B^* & 0 \\ 0 & I_q \end{pmatrix},$$

then $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{B}$. It it easily to see that

$$\mathcal{S}(\boldsymbol{B}\boldsymbol{B}^\top) = O_p(r). \tag{A.7}$$

From (A.5), (A.6) and (A.7), we have

$$\begin{aligned} T^{-1}r^{-2}\boldsymbol{e}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{e} &= T^{-1}r^{-2}\boldsymbol{e}^\top\boldsymbol{U}\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{U}^\top\boldsymbol{e} \\ &\leq r^{-1}\mathcal{S}(\boldsymbol{B}\boldsymbol{B}^\top)T^{-1}r^{-1}\boldsymbol{e}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{e} = O_p(1) \end{aligned} \tag{A.8}$$

and

$$\begin{aligned} & r^{-2}T\cdot\mathcal{S}\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right) \\ = \; & r^{-2}T\cdot\mathcal{S}\left((\boldsymbol{B}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{B})^{-1}\right) \\ = \; & r^{-2}T\cdot\mathcal{S}\left(\boldsymbol{B}^{\top^{-1}}(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\boldsymbol{B}^{-1}\right) \\ = \; & r^{-2}T\cdot\mathcal{S}\left(\boldsymbol{B}^{\top^{-1}}\right)\mathcal{S}\left((\boldsymbol{U}^\top\boldsymbol{U})^{-1}\right)\mathcal{S}\left(\boldsymbol{B}^{-1}\right) \\ = \; & O_p(1). \end{aligned} \tag{A.9}$$

By (A.8), (A.9), Markov inequality, and (C.1), for any $\delta > 0$,

$$\begin{aligned} \Pr\left\{\sup_{w\in\mathcal{H}_T}\frac{\left|\boldsymbol{e}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\right|}{R_T(\boldsymbol{w})} \geq \delta\right\} &\leq \Pr\{\xi_T^{-1}\boldsymbol{e}^\top\boldsymbol{P}\boldsymbol{e} \geq \delta\} = \Pr\left\{\xi_T^{-1}\boldsymbol{e}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{e} \geq \delta\right\} \\ &\leq \Pr\left\{\underline{r^4\xi_T^{-1}}\,\underline{r^{-2}T\cdot\mathcal{S}\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)}\,\underline{T^{-1}r^{-2}\boldsymbol{e}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{e}} \geq \delta\right\} \\ &\to 0. \end{aligned}$$

We obtained (A.2).[20]

To verify (A.3), we first show

$$\left|\boldsymbol{e}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right| = \left|\boldsymbol{e}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}\boldsymbol{\mu}\right| \leq \left(\boldsymbol{e}^\top\boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\,\boldsymbol{\mu}^\top\boldsymbol{P}\boldsymbol{\mu}\right)^{1/2} \leq \left(\boldsymbol{e}^\top\boldsymbol{P}\boldsymbol{e}\,\boldsymbol{\mu}^\top\boldsymbol{P}\boldsymbol{\mu}\right)^{1/2}.$$

[20]For convenience, we underlined the terms that converge to 0.

Moreover, according to (A.8) and (A.9), we have

$$r^{-4}\boldsymbol{e}^\top \boldsymbol{P}\boldsymbol{e} = O_p(1). \tag{A.10}$$

Then,

$$\begin{aligned}
\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right|}{R_T(\boldsymbol{w})} &\leq \xi_T^{-1}\left(\boldsymbol{e}^\top \boldsymbol{P}\boldsymbol{e}\,\boldsymbol{\mu}^\top \boldsymbol{P}\boldsymbol{\mu}\right)^{1/2} \\
&= \left(r^{-4}\boldsymbol{e}^\top \boldsymbol{P}\boldsymbol{e} r^4 \boldsymbol{\mu}^\top \boldsymbol{P}\boldsymbol{\mu}\xi_T^{-2}\right)^{1/2} \\
&\to 0,
\end{aligned}$$

that is (A.3).

Since

$$\left|\frac{L_T(\boldsymbol{w})}{R_T(\boldsymbol{w})} - 1\right| = \left|\frac{\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{e} - \sigma^2 \mathrm{tr}\left(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\right) - 2\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{A}(\boldsymbol{w})\boldsymbol{\mu}}{R_T(\boldsymbol{w})}\right|,$$

to prove (A.4), it suffices to show that

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\sigma^2 \mathrm{tr}\left(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\right)\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0, \tag{A.11}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0, \tag{A.12}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{A}(\boldsymbol{w})\boldsymbol{\mu}\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0, \tag{A.13}$$

Based on the proof of (A.1) and (A.2), it is straightforward to verify that, for any $\delta > 0$

$$\begin{aligned}
\Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\sigma^2 \mathrm{tr}\left(\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\right)\right|}{R_T(\boldsymbol{w})} \geq \delta\right\} &\leq \Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\sigma^2 \mathrm{tr}\left(\boldsymbol{P}(\boldsymbol{w})\right)\right|}{R_T(\boldsymbol{w})} \geq \delta\right\} \to 0, \\
\Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\right|}{R_T(\boldsymbol{w})} \geq \delta\right\} &\leq \Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\right|}{R_T(\boldsymbol{w})} \geq \delta\right\} \to 0.
\end{aligned}$$

We obtain (A.11) and (A.12). Since

$$\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{A}(\boldsymbol{w})\boldsymbol{\mu}\right| = \left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu} - \boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right|,$$

and we already proved (A.3), to prove (A.13), we only need to verify

$$\sup_{w\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right|}{R_T(\boldsymbol{w})} \xrightarrow{p} 0. \tag{A.14}$$

We see that

$$\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right| = \left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}\boldsymbol{\mu}\right| \leq \left(\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{e}\,\boldsymbol{\mu}^\top \boldsymbol{P}\boldsymbol{\mu}\right)^{1/2} \leq \left(\boldsymbol{e}^\top \boldsymbol{P}\boldsymbol{e}\,\boldsymbol{\mu}^\top \boldsymbol{P}\boldsymbol{\mu}\right)^{1/2}. \tag{A.15}$$

Following the proof of (A.3) and (A.15), we have

$$\begin{aligned}\sup_{w\in\mathcal{H}_T} \frac{\left|\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\boldsymbol{P}(\boldsymbol{w})\boldsymbol{\mu}\right|}{R_T(\boldsymbol{w})} &\leq \xi_T^{-1}\left(\boldsymbol{e}^\top \boldsymbol{P}\boldsymbol{e}\,\boldsymbol{\mu}^\top \boldsymbol{P}\boldsymbol{\mu}\right)^{1/2} \\ &= \left(r^{-4}\boldsymbol{e}^\top \boldsymbol{P}\boldsymbol{e} r^4 \boldsymbol{\mu}^\top \boldsymbol{P}\boldsymbol{\mu}\xi_T^{-2}\right)^{1/2} = o_p(1).\end{aligned}$$

This completes the proof. ∎

B OLS Estimation

We present OLS estimates from a one-step-ahead forecasting with full data sample in Table A1. The explanatory variables are listed in the first column of Table A1. Columns 2 to 5 respectively presents results of the following models: (i) conventional HAR model with no control variables; (ii) the HARX^1 model with all exogenous variables but USSI; (iii) the HARX^2 model with the USSI variable only as the sole exogenous explanatory variable; and (iv) the $\text{HARX}^\star$ model with all of the exogenous variables included. The last two columns of Table A1 show the centered R^2 and adjusted R^2 respectively. Numbers with symbol $*$, $**$, or $***$ represent significance level at 10%, 5%, or 1%, respectively.

We observe that the HAR 1-day average component is always highly significant among all four models, which implies that the VIX index is quite persistent. Interestingly, all other HAR component are not statistically significant at conventional levels, with the exception of the HAR 22-day average component in $\text{HARX}^\star$ (10% significance). The columns of Models HARX^1 and $\text{HARX}^\star$ indicate that, most of the exogenous explanatory variables do not have significant associations with the VIX index. In fact, only the S&P 500 22-day Return, the Credit Spread, and the FF Deviation are observed to have any significant forecasting power. On the other hand, the USSI variable does exhibit a strong association with the VIX. The strength of its forecasting power in similar in both the HARX^2 and $\text{HARX}^\star$ model, suggesting that it is capturing different information than the control variables. In summary, the results in Table A1 show that the USSI variable is a very useful predictor of the VIX index and there is strong uncertainty in both the HAR components and other exogenous variables.

Table A1: Estimation Results with USSI

Variable	HAR	HARX[1]	HARX[2] (USSI Only)	HARX*
Constant	0.1293***	0.2857***	0.2207***	0.0029
	(0.0342)	(0.0475)	(0.0415)	(0.0441)
HAR 1-day Average	0.9211***	0.9262***	0.9107***	0.7129***
	(0.0297)	(0.0450)	(0.0296)	(0.0409)
HAR 5-day Average	-0.0250	-0.0513	-0.0292	-0.0251
	(0.0566)	(0.0706)	(0.0563)	(0.0616)
HAR 22-day Average	0.0434	-0.0012	-0.0090	-0.0668*
	(0.0334)	(0.0439)	(0.0359)	(0.0385)
S&P500 1-day Return		0.5541		0.3878
		(0.4151)		(0.3624)
S&P500 5-day Return		-0.0441		-0.0140
		(0.2633)		(0.2298)
S&P500 10-day Return		0.1306		0.0786
		(0.2036)		(0.1777)
S&P500 22-day Return		-0.2414*		-0.1670
		(0.1246)		(0.1088)
S&P500 Volume Change		-0.1304		-0.1416
		(0.1237)		(0.1080)
Oil 1-day Return		0.0924		0.0824
		(0.0785)		(0.0685)
Oil 5-day Return		-0.0292		-0.0149
		(0.0615)		(0.0536)
Oil 10-day Return		-0.0230		0.0010
		(0.0316)		(0.0276)
Oil 22-day Return		-0.0146		-0.0053
		(0.0124)		(0.0108)
USD Change		-0.1746		-0.4682
		(0.5424)		(0.4737)
Credit Spread		0.0162***		0.0085***
		(0.0051)		(0.0045)
Term Spread		0.0880		-0.0026
		(0.0630)		(0.0552)
FF Deviation		0.0714***		0.0330**
		(0.0159)		(0.0140)
U.S. Sentiment Index			0.0532***	0.0092***
			(0.0139)	(0.0005)
Centered R^2	0.8752	0.8790	0.8768	0.9079
Ajusted R^2	0.8747	0.8772	0.8762	0.9064

* 10% level of significance.
** 5% level of significance.
*** 1% level of significance.

C The Heterogeneous MIDAS Approach

In this section, we briefly explain the H-MIDAS method introduced in Lehrer et al. (2018). Let Y_t be a low frequency variable that is sampled at periods denoted by a time index t for $t = 1, ..., n$. Consider a higher frequency (indicated by a superscript h throughout the paper) predictor $\boldsymbol{X}_t^h$ that are sampled m times within the period of t:

$$\boldsymbol{X}_t^h \equiv \left[X_t^h, X_{t-\frac{1}{m}}^h, ..., X_{t-\frac{m-1}{m}}^h\right]^\top. \tag{A.16}$$

A specific element among the high frequency observations in $\boldsymbol{X}_t^h$ is denoted by $X_{t-\frac{i}{m}}^h$ for $i = 0, ..., m-1$. Denoting $L^{i/m}$ as the lag operator, then $X_{t-\frac{i}{m}}^h$ can be reexpressed as $X_{t-\frac{i}{m}}^h = L^{i/m} X_t^h$ for $i = 0, ..., m-1$.

Since $\boldsymbol{X}_t^h$ on Y_t are measured at different frequencies, we need to convert the higher-frequency data to match the lower-frequency data. A simple average of the high frequency observations $\boldsymbol{X}_t^h$:

$$\bar{X}_t = \frac{1}{m} \sum_{i=0}^{m-1} L^{i/m} X_t^h,$$

where $\bar{X}_t$ is likely the easiest way to estimate a low frequency X_t that can match the frequency of Y_t. With the variables Y_t and $\bar{X}_t$ are measured in the same time domain, a regression approach is simply

$$Y_t = \alpha + \gamma \bar{X}_t + \epsilon_t = \alpha + \frac{\gamma}{m} \sum_{i=0}^{m-1} L^{i/m} X_t^h + \epsilon_t, \tag{A.17}$$

where α is the intercept, γ is the slope coefficient on the time-averaged $\bar{X}_t$. This approach assumes that each element in $\boldsymbol{X}_t^h$ has an identical effect on explaining Y_t since they all share the same coefficient γ.

These homogeneity assumptions may be quite strong in practice. One could assume that each of the slope coefficients for each element in $\boldsymbol{X}_t^h$ is unique. Extending Model (A.17) to allow for heterogeneous effects of the high frequency observations generates

$$Y_t = \alpha + \sum_{i=0}^{m-1} \gamma_i L^{i/m} X_t^h + \epsilon_t, \tag{A.18}$$

where γ_i represents a set of slope coefficients for all high frequency observations $X_{t-\frac{i}{m}}^h$.

Since γ_i is unknown, estimating these parameters can be problematic when m is a relatively large number.The H-MIDAS method uses a step function to allow for heterogeneous effects of different high frequency observations on the low frequency dependent variable. A low frequency $\bar{X}_t^{(l)}$ can be constructed following

$$\bar{X}_t^{(l)} \equiv \frac{1}{l} \sum_{i=0}^{l-1} L^{i/m} X_t^h = \frac{1}{l} \sum_{i=0}^{l-1} X_{t-\frac{i}{m}}^h, \tag{A.19}$$

where l is a pre-determined number and $l \leq m$. Equation (A.19) implies that we compute variable $\bar{X}_t^{(l)}$ by a simple average of the first l observations in $\boldsymbol{X}_t^h$ and ignored the remaining observations. We consider different values of l and group all $\bar{X}_t^{(l)}$ into $\tilde{\boldsymbol{X}}_t$ such that

$$\tilde{\boldsymbol{X}}_t = \left[\bar{X}_t^{(l_1)}, \bar{X}_t^{(l_2)}, \dots, \bar{X}_t^{(l_p)}\right],$$

where we set $l_1 < l_2 < \cdots < l_p$. Consider a weight vector $\boldsymbol{w} = \left[w_1, w_2, \dots, w_p\right]^\top$ with $\sum_{j=1}^p w_j = 1$, we can construct regressor X_t^{new} as $X_t^{new} = \tilde{\boldsymbol{X}}_t \boldsymbol{w}$. The regression based on the H-MIDAS estimator can be expressed as

$$Y_t = \beta X_t^{new} + \epsilon_t = \beta \sum_{s=1}^{p} \sum_{j=s}^{p} \frac{w_j}{l_j} \sum_{i=l_{s-1}}^{l_s-1} L^{i/m} X_t^h + \epsilon_t = \beta \sum_{s=1}^{p} \sum_{i=l_{s-1}}^{l_s-1} w_s^* L^{i/m} X_t^h + \epsilon_t, \tag{A.20}$$

where $l_0 = 0$ and $w_s^* = \sum_{j=s}^p \frac{w_j}{l_j}$ can be interpreted as the weights on the high frequency observations.

The weights $\boldsymbol{w}$ play a crucial role in this procedure. We first estimate $\widehat{\beta \boldsymbol{w}}$ following

$$\widehat{\beta \boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathcal{W}}{\arg\min} \left\| Y_t - \tilde{\boldsymbol{X}}_t \cdot \beta \boldsymbol{w} \right\|^2$$

by any appropriate econometric method necessary, where $\mathcal{W}$ is some predetermined weights set. Once $\widehat{\beta \boldsymbol{w}}$ is obtained, we estimate the weight vector $\hat{\boldsymbol{w}}$ by rescaling following

$$\hat{\boldsymbol{w}} = \frac{\widehat{\beta \boldsymbol{w}}}{\text{Mean}(\widehat{\beta \boldsymbol{w}})},$$

since the coefficient β is a scalar.

In practice, we need to select the lag index $\boldsymbol{l} = [l_1, ..., l_p]$ and determine the weight set $\mathcal{W}$ before the estimation. In this exercise, we apply the H-MIDAS method to the USSI data using four sets of lag indices: $\boldsymbol{l} = [1 : d : 1000]$, where $d = 1, 10, 30$, and 60 in the sense of capturing 1, 10, 30, and 60 minutes-level components of the USSI data. For the weight set $\mathcal{W}$, we follow Kuersteiner and Okui (2010) and set $\mathcal{W} \equiv \{\boldsymbol{w} \in [-1, 1]^p : \sum_{j=1}^p w_j = 1\}$. Then, we use OLS to estimate $\widehat{\beta \boldsymbol{w}}$. Estimates of the H-MIDAS weights for our application are presented in the main text in Figure 1.

D Results of Using Equal Weighted Average USSI

In this section, we explore the robustness of our results to a simple equal weighted average USSI. This measure is simple to calculate and in essence we are simply only replacing the H-MIDAS step. In the remainder of the section, we use the subscript a to denote specifications that incorporate the average weighted Twitter sentiment measure. The first main result we examine the robustness of to this alternative means of incorporating social media data is the prediction experiment that was presented in Tables 4 and 5.

Tables A2 and A3 show that all of the conclusions we drew from Tables 4 and 5 are robust to replacing the H-MIDAS transformed USSI to an the equal weighted USSI. Notice that, MAHARX^a demonstrates better forecasting accuracy than other estimators considered in Table A2 for short forecasting horizons. As the forecasting horizon increases, the results also find that the benefits from incorporating the USSI diminish sharply. These results parallel those presented in the main text.

The improvement in forecast accuracy between MAHARX^a relative to MAHARX appears to be smaller than the gap MAHAR exhibits with MAHARX*. Thus, we next formally compare the forecasting efficiency of using simple equal weighted average USSI with H-MIDAS weighted average USSI. The results of forecasting comparison are presented in Table A4. MAHARX* using H-MIDAS weighted average USSI has better performance since the equal weighted USSI imposes a homogeneity assumption that induces bias during the frequency matching procedure.

To examine if the improvement between incorporating the H-MIDAS transformed USSI over the equally weighted USSI is statistical significant, we perform the GW test and report the results in Table A5. Notice, that all of the p-values are smaller than 10%, irrespective of the forecast horizon signifying the importance of using the appropriate averaging method to convert high frequency data to lower frequency.

Table A2: Forecasting Performance at Different Horizons

	MSFE	SDFE	MAFE	Pseudo R^2
Panel A: 1 Days Ahead				
RW	0.0069	0.0830	0.0578	0.9036
HAR	0.0068	0.0828	0.0576	0.9041
HARX	0.0069	0.0830	0.0574	0.9036
AHARX	0.0069	0.0833	0.0576	0.9028
LHARX	0.0068	0.0826	0.0570	0.9040
MAHARX	0.0067	0.0817	0.0567	0.9064
$HARX^a$	0.0069	0.0831	0.0575	0.9032
$AHARX^a$	0.0070	0.0836	0.0576	0.9021
$LHARX^a$	0.0069	0.0829	0.0584	0.9037
$MAHARX^a$	**0.0067**	**0.0817**	**0.0567**	**0.9064**
Panel B: 5 Days Ahead				
RW	0.0294	0.1715	0.1202	0.5895
HAR	0.0280	0.1674	0.1206	0.6085
HARX	0.0264	0.1626	0.1195	0.6310
AHARX	0.0273	0.1652	0.1201	0.6191
LHARX	0.0272	0.1648	0.1153	0.6109
MAHARX	0.0256	0.1600	0.1166	0.6428
$HARX^a$	0.0266	0.1631	0.1202	0.6284
$AHARX^a$	0.0275	0.1658	0.1200	0.6160
$LHARX^a$	**0.0251**	**0.1585**	**0.1121**	**0.6420**
$MAHARX^a$	0.0258	0.1606	0.1173	0.6396
Panel C: 10 Days Ahead				
RW	0.0452	0.2127	0.1490	0.3705
HAR	0.0434	0.2084	0.1562	0.3950
HARX	0.0373	0.1930	0.1470	0.4801
AHARX	0.0423	0.2057	0.1548	0.4090
LHARX	0.0394	0.1986	**0.1388**	0.4354
MAHARX	0.0364	0.1907	0.1460	0.4923
$HARX^a$	0.0377	0.1943	0.1491	0.4714
$AHARX^a$	0.0430	0.2073	0.1553	0.3978
$LHARX^a$	**0.0367**	**0.1910**	0.1451	**0.4847**
$MAHARX^a$	0.0368	0.1919	0.1483	0.4844
Panel D: 22 Days Ahead				
RW	0.0589	0.2426	0.1758	0.1924
HAR	0.0595	0.2439	0.2007	0.1777
HARX	0.0397	0.1992	0.1543	0.4560
AHARX	0.0435	0.2085	0.1634	0.4032
LHARX	0.0464	0.2155	0.1596	0.3544
MAHARX	**0.0383**	**0.1957**	**0.1510**	**0.4750**
$HARX^a$	0.0418	0.2045	0.1606	0.4250
$AHARX^a$	0.0457	0.2139	0.1683	0.3704
$LHARX^a$	0.0400	0.1998	0.1540	0.4359
$MAHARX^a$	0.0404	0.2010	0.1573	0.4449

This table reports the out-of-sample results for predicting h-day future realized variation using the different predictor variables and risk models. The results are based on the CBOE VIX data spanning from January 7, 2013 to August 21, 2017 (a total of 1,206 observations). We use a rolling window of 600 observations to estimate the coefficients of the models, and evaluate the out-of-sample forecast performance at four horizons ($h = 1$, $h = 5$, $h = 10$ and $h = 22$). Each panel in the table corresponds to a specific forecast horizon, which varies from 1 day to 22 days. Bold numbers indicate the best performing model by each criterion at each forecast horizon.

Table A3: Giacomini-White Test Results

	RW	HAR	HARX	AHARX	LHARX	MAHARX	HARX*	AHARX*	LHARX*
Panel A: 1 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.8347	-	-	-	-	-	-	-	-
HARX	0.9996	0.7772	-	-	-	-	-	-	-
AHARX	0.7395	0.5903	0.6433	-	-	-	-	-	-
LHARX	0.8380	0.9806	0.8316	0.6294	-	-	-	-	-
MAHARX	0.2058	0.1042	**0.0096**	0.0798	0.1776	-	-	-	-
HARX^a	0.8616	0.5907	0.2733	0.8144	0.6746	**0.0044**	-	-	-
AHARX^a	0.5429	0.4147	0.4381	0.2018	0.4541	**0.0487**	0.5492	-	-
LHARX^a	0.9831	0.8338	0.9753	0.7885	0.8841	0.2482	0.8539	0.6522	-
MAHARX^a	0.2059	0.1043	**0.0096**	0.0798	0.1776	0.6435	**0.0044**	**0.0487**	0.2483
Panel B: 5 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.4343	-	-	-	-	-	-	-	-
HARX	0.2319	0.3853	-	-	-	-	-	-	-
AHARX	0.3108	0.6675	0.5563	-	-	-	-	-	-
LHARX	0.3241	0.9311	0.5630	0.7905	-	-	-	-	-
MAHARX	0.1268	0.1828	**0.0413**	0.2399	0.3692	-	-	-	-
HARX^a	0.2516	0.4273	0.4048	0.6403	0.6149	**0.0096**	-	-	-
AHARX^a	0.3449	0.7547	0.5042	0.5173	0.8670	0.2252	0.5666	-	-
LHARX^a	**0.0084**	**0.0156**	0.3666	0.1242	0.0817	0.7007	0.3010	0.1058	-
MAHARX^a	0.1533	0.2321	0.2331	0.3194	0.4297	0.2649	0.0563	0.2858	0.6126
Panel C: 10 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.6465	-	-	-	-	-	-	-	-
HARX	0.1902	0.1785	-	-	-	-	-	-	-
AHARX	0.6284	0.8279	0.2220	-	-	-	-	-	-
LHARX	0.2079	0.5727	0.6008	0.7827	-	-	-	-	-
MAHARX	0.1370	0.1215	0.1273	0.1345	0.5013	-	-	-	-
HARX^a	0.2185	0.2127	0.3356	0.2517	0.6773	**0.0346**	-	-	-
AHARX^a	0.7431	0.9679	0.2496	0.5254	0.7153	0.1672	0.2682	-	-
LHARX^a	**0.0065**	**0.0152**	0.4107	0.0815	0.0931	0.5310	0.3284	0.0954	-
MAHARX^a	0.1607	0.1510	0.7475	0.1490	0.5701	0.3441	0.1236	0.1745	0.4439
Panel D: 22 Days Ahead									
RW	-	-	-	-	-	-	-	-	-
HAR	0.9119	-	-	-	-	-	-	-	-
HARX	0.0647	**0.0158**	-	-	-	-	-	-	-
AHARX	0.1633	0.0729	0.3087	-	-	-	-	-	-
LHARX	0.1451	0.2107	0.3053	0.6677	-	-	-	-	-
MAHARX	0.0510	**0.0104**	**0.0147**	0.1911	0.2365	-	-	-	-
HARX^a	0.1154	**0.0387**	0.0799	0.6929	0.5118	**0.0110**	-	-	-
AHARX^a	0.2483	0.1372	0.1644	**0.0482**	0.8954	0.1041	0.3769	-	-
LHARX^a	**0.0166**	**0.0070**	0.4202	0.1483	0.0720	0.6269	0.2150	0.0929	-
MAHARX^a	0.0864	**0.0241**	0.5131	0.4593	0.4015	0.0586	**0.0169**	0.2359	0.3594

The modified Giacomini-White test (Giacomini and White, 2006) is implemented to test the null hypothesis that the *row method* (in vertical headings) performs equally well as the *column method* (in horizontal headings) in terms of the absolute forecast error. Corresponding p values for a number of forecasting horizons ($h = 1, 5, 10, 22$) are reported in Panels A to D, respectively. Bold numbers indicate the null hypothesis can be rejected at 5% level of significance.

Table A4: Forecasting Performance at Different Horizons

	MSFE	SDFE	MAFE	Pseudo R^2
Panel A: 1 Days Ahead				
MAHARX^a	0.0067	0.0817	0.0567	0.9064
MAHARX*	**0.0047**	**0.0683**	**0.0503**	**0.9348**
Panel B: 5 Days Ahead				
MAHARX^a	0.0258	0.1606	0.1173	0.6396
MAHARX*	**0.0244**	**0.1563**	**0.1147**	**0.6589**
Panel C: 10 Days Ahead				
MAHARX^a	0.0368	0.1919	0.1483	0.4844
MAHARX*	**0.0353**	**0.1879**	**0.1423**	**0.5080**
Panel D: 22 Days Ahead				
MAHARX^a	0.0404	0.2010	0.1573	0.4449
MAHARX*	**0.0378**	**0.1943**	**0.1494**	**0.4830**

This table reports the out-of-sample results for predicting h-day future realized variation between MAHARX^a and MAHARX*. The results are based on the CBOE VIX data spanning from January 7, 2013 to August 21, 2017 (a total of 1,206 observations). We use a rolling window of 600 observations to estimate the coefficients of the models, and evaluate the out-of-sample forecast performance at four horizons ($h = 1$, $h = 5$, $h = 10$ and $h = 22$). Each panel in the table corresponds to a specific forecast horizon, which varies from 1 day to 22 days. Bold numbers indicate the best performing model by each criterion at each forecast horizon.

Table A5: Giacomini-White Test Results

	$h = 1$	$h = 5$	$h = 10$	$h = 22$
p-value	**0.0001**	0.0841	**0.0014**	**0.0061**

The modified Giacomini-White test (Giacomini and White, 2006) is implemented to test the null hypothesis that MAHARX^a performs equally well as MAHARX* in terms of the absolute forecast error. Corresponding p values for a number of forecasting horizons ($h = 1, 5, 10, 22$) are reported in Panels A to D, respectively. Bold numbers indicate the null hypothesis can be rejected at 5% level of significance.