Title: Forecasting Realized Volatility: HAR against PrincipalComponents Combining, Neural Networks and GARCH

Author: Dimitrios I. Vortelinos

PII: S0275-5319(15)00005-7
DOI: http://dx.doi.org/doi:10.1016/j.ribaf.2015.01.004
Reference: RIBAF 346

To appear in: *Research in International Business and Finance*

Received date: 27-1-2015
Accepted date: 29-1-2015

Please cite this article as: Dimitrios I. Vortelinos, Forecasting Realized Volatility: HAR against PrincipalComponents Combining, Neural Networks and GARCH, *Research in International Business and Finance* (2015), http://dx.doi.org/10.1016/j.ribaf.2015.01.004

Forecasting Realized Volatility: HAR against Principal Components Combining, Neural Networks and GARCH

Dimitrios I. Vortelinos*

January 27, 2015

Abstract

This paper examines whether nonlinear models, like Principal Component Combining, Neural Networks and GARCH are more accurate on realized volatility forecasting than the Heterogeneous (HAR) model. The answer is no. The realized volatility property of persistence is too important to leave out of a realized volatility forecasting model. However, the Principal Components Combining model is ranked very close to HAR. Analysis is implemented in seven US financial markets: spot equity, spot foreign exchange rates, exchange traded funds, equity index futures, US Treasury bonds futures, energy futures, and commodities options.

Keywords: HAR; Principal Components Combining; Neural networks; GARCH; Forecasting.

*Corresponding author. Lincoln Business School, University of Lincoln, UK. Contact at dvortelinos@lincoln.ac.uk, 0044-1522-835634 (tel.)

Forecasting Realized Volatility: HAR against Principal Components Combining, Neural Networks and GARCH

January 27, 2015

Abstract

This paper examines whether nonlinear models, like Principal Component Combining, Neural Networks and GARCH are more accurate on realized volatility forecasting than the Heterogeneous (HAR) model. The answer is no. The realized volatility property of persistence is too important to leave out of a realized volatility forecasting model. However, the Principal Components Combining model is ranked very close to HAR. Analysis is implemented in seven US financial markets: spot equity, spot foreign exchange rates, exchange traded funds, equity index futures, US Treasury bonds futures, energy futures, and commodities options.

Keywords: HAR; Principal Components Combining; Neural networks; GARCH; Forecasting.

1 Introduction

Apart from academics, financial markets participants, investment bankers, regulators and government agencies need more accurate volatility forecasts. Moreover, signifying the out-of-sample significance of forecasting models reveals the importance of the corresponding (to models) properties of realized volatility estimators. This opens new research areas in the realized volatility literature. The dependence of the option price, Value at Risk (VaR), Sharpe ratio and other investment indices/indicators on volatility also reveals the importance of accurate volatility forecasts. Volatility forecasting is also important to constructing equity portfolios, as indicated in Amenc et al. (2012). Find the most accurate model in volatility forecasting is an active research area in the literature. Moreover, there is no evidence enough whether simple or more complicated models better forecast volatility. Shephard and Sheppard (2010) provided evidence in favor of not heavily parameterized models in high frequency volatility forecasting. A more recent forecasting competition paper is Santos and Ziegelmann (2014), where they compare the MIDAS and HAR models in realized volatility forecasting for the IBOVESPA index is provided in.

The present paper examines whether the forecasting performance of the principal components combining (PCC) model and neural networks (NN) model can produce substantially accurate realized volatility forecasts or not. Daily volatility is non-parametrically estimated via the two-scale realized volatility estimator. Secondly, examine whether the PCC and NN models better forecast realized volatility series than the Generalized Autoregressive Conditional Heteroskedastic ($GARCH$) and the Heterogeneous Autoregressive (HAR) models which are the best models in the realized volatility forecasting literature. The forecasting performance of all four models is assessed relatively to a $AR(p_{AIC})$ benchmark model.[1] The relative evaluation criteria are ratios of either RMSE, MAE, R^2 or Direction criterion for a model or method to the corresponding evaluation criterion (either RMSE, MAE, R^2 or Direction) of the benchmark AR model. R^2 is the adjusted R^2-value of the Mincer and Zarnowitz (1969) predictive regression. The directional criterion reveals the times (as a %) that the direction of actual volatiity series is in average the same as the one of forecasts. The forecasting performance is compared across models via a Diebold Mariano test. Notation for the data used, the forecasting models as well as the evaluation criteria is depicted in Table 1.

The rest of the paper is organized as follows. In Section 2 literature is reviewed. In Section 3 a short description of data used is provided. In section 4 are presented the volatility estimation employed, the forecasting models used, as well as the forecasting performance evaluation criteria. In Section 5 empirical results are reported, and in Section 6 concluding remarks are retrieved.

[1] The analysis is in line with Nachane and Clavel (2008); where, they compare the forecasting performance of a combined wavelet artificial neural network (WANN) to a mixed spectrum analysis method and nonlinear ARMA models with Fourier coefficient (FNLARMA).

2 Literature review

The forecasting literature is vast. It includes forecasting models and methods from many different categories. de Gooijer and Hyndman (2006) reviewed the 25 years (from 1982 up to 2005) of research into time series forecasting. According to the literature, the classical econometric models, like Autoregressive models, have limitations. In order to allow for flexibility, more customized models and methods have to be employed. The forecasting literature of realized volatility is more limited. One of the first papers tried to examine realized volatility forecasting is Fung and Hsieh (1999). They reported that the forecasting power is higher for the naï¿œve realized volatility estimator than the implied volatility estimator and the daily Parkinson range-based volatility estimator. Andersen et al. (2003) empirically prove a forecasting-performance improvement of realized volatility compared to GARCH volatility under various forecasting models. The realized volatility forecasting was thoroughly examined in terms of microstructure noise by Andersen et al. (2011).

In the present paper, daily volatility is non-parametrically estimated via the two-scale realized volatility estimator. The paper that set up the foundations of modelling and forecasting realized volatility was Andersen et al. (2003). Ait-Sahalia and Mancini (2008) showed that using the two-scales estimator instead of standard realized volatility measures yields significant gains in volatility forecasting. Studies with empirical evidence on the two-scale realized volatility estimator are Andersen et al. (2011), Bandi and Russell (2011) and Ghysels and Sinko (2011). More recent literature researching such an estimator contains: Louzis et al. (2013) and Mancini et al. (2013), among others. The daily two-scale realized volatility series in US financial markets are forecasted using Principal Components Combining (PCC), Neural Networks (NN), Generalized AutoRegressive (GARCH), Heterogeneous AutoRegressive (HAR) and Autoregressive (AR) models.

The principal-components (PC) forecasting literature is really vast; however, not well-reviewd especially by economists and financiers. Ludvigson and Ng (2009) estimate predictive regressions of Principal Components (PC) on stock and bond returns. Neely et al. (2014) employed principal components of macroeconomic variables to forecast equity risk premium. Nomikos and Poullasis (2014) have recently used Principal Components to derive latent factors for forecasting via a multi-regime error correction factor model. Steeley (2014) used principal components to forecast the UK government bond yield curve. The most known principal components forecasting methods are: (i) principal components regression (see, de Mol, et al., 2008; and, Heij, et al., 2008); (ii) introduced matched principal component regression (MPCR); (iii) principal covariate regression (PCovR) (see, Heij, Grenen and van Dijk, 2007); (iv) dynamic factor model (DFM) forecasting (see, Bai and Ng, 2008); (v) targeted principal components (see, Boivin and Ng, 2006). Another, not so heavily studied principal components forecasting model is Principal Components Combining (PCC). Stock and Watson (2004) were the first that formed and the ones that analytically examined combination forecasts, utilizing the first m Principal Components of individual forecasts. Their parametirization is used in the present paper. A recent study on many combining methods is den Butter and Jansen (2013). Combinations can use principal components to estimate the common factors from the panel of forecasts, and regress a subset of these on a target variable. For a recent application, see Genre et al. (2013). The principal components combining (PCC) model is also simple in its perception and construction. It takes into account the advantages of the principal component analysis and the combining method of forecasts as well. However, there is no direct link between the PCC specification and the realized volatility properties.

The development of artificial neural network (NN) models depends on: selection of the learning algorithm, choice of the error and transfer functions, specification of the architecture, preparation of the data to match the architecture, and training of the network. As analyzed by Coakley and Brown (2000), neural network forecasting requires to follow the following process: (1) determine the type of research question and comparable parametric model, (2) select the learning algorithm (back propagation or recurrent; the latter is more appropriate for prediction), (3) select error and transfer functions (SSE error function and hyperbolic tangent transfer function for prediction), (4) select architecture (number of nodes in the input and output layers), (5) determine the number of layers and the number of nodes per layer (expansion and pruning algorithms can be used to overcome the 'trial and error' approach), (6) scale of the input data and generate the initial training weights (data-preparation software packages), (7) present data sets in order and adjust the weights after each data set, and (8) obtain a sufficient sample size (jacknife and cross validation methods in case of limited data sets). Artificial neural networks (NNs) have been widely applied to finance and economic forecasting as a powerful modeling method. The Neural Networks (NN) model provides accurate volatility forecasts, based on literature. Hamid and Iqbal (2004) found that Neural Networks perform well enough in volatility prediction. Another paper on the use of neural networks improving 'traditional' volatility forecasts is Aragones et al. (2007). The volatility forecasting literature provides strong evidence in favor of neural networks (NN). NN have mostly been employed in GARCH volatility series (Hajizadeh et al., 2012 and Kristjanpoller et al., 2014, among others). Recently, Fernandes et al. (2014) suggested a neural-network heterogeneous autoregression with exogenous variables (NNHARX) that better forecasts the implied volatility index VIX, compared to random walk, HAR, HAR with exogenous variables and HAR with exogenous variables and asymmetric effects. Moreover, NN seem to provide accurate realized volatility forecasts (Rosa et al., 2014). However, there is no direct link between each parameter selection and the properties of realized volatilities. Moreover, the NN specification is not easily selected and constructed.

The GARCH forecasting is very popular in academia. Some of the first influential papers on volatility GARCH forecasting is Andersen and Bollerslev (1998) and Andersen et al. (1999). A recent paper on dynamic conditional correlation GARCH forecasting is Boudt et al. (2013). GARCH models accurately forecast volatility; but, GARCH models do not provide as accurate realized volatility forecasts. However just recently, two papers enhanced the forecasting ability of GARCH process via mixed models. Yang and Chen (2014) suggested a fractionally integrated generalized autoregressive conditional heteroskedasticity error (HAR-D-FIGARCH) model for forecasting realized volatility. Recently, Yang et al. (2015) suggested ARFIMAX-FIGARCH models for the realized volatility forecast by utilizing a variety of estimation window sizes designed to accommodate potential structural breaks.

Long memory (persistence) is the most important realized volatility property. EWMA is one of the model that model persistence. However, the EWMA model hasn't been widely employed in the realized volatility forecasting literature. Long memory linear models, such as ARFIMA and HAR models, are considered as the best models to forecast realized volatility, however. Andersen et al. (2007) found that the Heterogeneous Auto Regressive (HAR) model showed better forecasting performance compared to the Auto Regressive Fractionally Integrated Moving Average (ARFIMA) model. Corsi (2009) and Corsi and Reno (2012), among others, signify the importance of long memory (persistence) in realized volatility forecasting. In a HAR model, long memory is incorporated via using lagged realized variances as predictors. Recently, HAR models have been employed for realized volatility forecasting in Bekaert and Hoerova (2014) and Santos and Ziegelmann (2014). Celik and Ergin (2014) found that the HAR models provided more accurate realized volatility forecasts than GARCH models. In Seo and Kim (2015), HAR models have also been employed in depicting the information content of option-implied information for realized volatility forecasting.

AR model has been widely used in the literature as a benchmark model. The AR model has superior forecasting accuracy and simplicity in its specification. However, not many properties of realized volatility estimators are incorporated in a AR model. Recently, Ferrara et al. (2014) employed an AR(1) model as a benchmark to an extended MIDAS model allowing exogenous variables sampled in different frequencies. Reeves and Xie (2014) not only used the AR model as a benchmark, but they also demonstrated superior forecasting performance from a simple autoregressive model with one lag of quarterly realized volatility AR(1) that dominates the MIDAS approach. Clark and Ravazzolo (2014) also used AR model as benchmark; AR was from the best forecasting models.

Forecasting evaluation is depicted as a ratio between a model's evaluation to the evaluation of an AR model as benchmark. The forecasting performance of the PCC, NN, GARCH and HAR models is assessed relatively to AR. Lopez (2001) studied the properties of the employed here (RMSE and MAE) as well as other loss functions. RMSE and MAE loss functions have been employed in a volatility forecasting competition in Hansen and Lunde (2005). The forecasting performance is compared between the five models via a Diebold Mariano test. The Diebold Mariano test is implemented as suggested in Giacomini and White (2006), with a QLIKE loss function as in Patton (2011b).

3 Data

The sampling frequency of data is 1 minute. Realized volatility series are estimated daily. Week-ends and a set of fixed and irregular holidays, as well as the days with too many missing values are removed. Sample starts on October 21, 2002 and ends on October 14, 2011. The out-of-sample evaluation for either criteria, coorresponds to a sample from February 1, 2009 up to October 14, 2011 (30% of observations as the out-of-sample period).[2] Forecasts are provided for 1-step(day)-ahead. Analysis concerns data series from seven US financial markets:

[2] For consistency reason, two more lengthy out-of-sample evaluation periods have been employed as well. The first started from January 3, 2010 up to October 14, 2011 (20% of observations as the out-of-sample period) and the second started from March 2, 2008 (40% of observations as the out-of-sample period). Moreover, these out-of-sample periods are not directly affected by the 2008 financial crisis.

(1) SE: spot equity market (Dow Jones Industrial Average: $INDU$), (2) SFE: spot foreign exchange rates ($EURUSD$ exchange rate), (3) ETF: exchange traded funds (PowerShares QQQ: QQQ), (4) EIF: equity index futures (E-Mini Dow futures continuous contract: YM), (5) TBF: US Treasury bonds futures (30-year US Treasury yield: TYX), (6) EF: energy futures market (Crude oil miNY futures continuous contract: QM), and (7) CO: Commodities options (CBOE gold index options: GOX).

4 Methodology

4.1 Volatility Estimation

Volatility is latent. Volatility is best estimated via integrated volatility. The best measure of integrated volatility is quadratic variation. Realized volatility is one of the best nonparametric estimators of quadratic variation. Andersen et al. (2001) introduced the realized volatility estimator, which simply is the sum of the observable intraday squared returns. In the present paper, the daily realized volatility estimates equal to the summation of intraday squared intraday (1-minute) returns as:

$$RV_t^{(m)}=\sum_{i=1}^{m} r_{i,m}^2 \tag{1}$$

In the absence of noise, this estimator is a consistent estimator of V_t as the sampling frequency increases. An intermediate sampling frequency that uses the highest possible sampling frequency and smooths the effect of microstructure noise is 5 minutes as suggested by Andersen et al. (2001). This estimator is denoted as:

$$RV_t^{(m/5)}=\sum_{i=1}^{m/5} r_{i,m}^2 \tag{2}$$

According to Zhang et al. (2005), the 5-minute realized volatility estimator is not ranked as the best estimator. The second-best estimator in their categorization is the sub-sampling estimator. Concerning this estimator, suppose the full grid with all observations within the day is defined as G and m is number of observations (the size of G). G is then partitioned into k non-overlapping sub-grids $G^{(k)}$ of size m_k. If a sparse sampling (e.g. every 5 or 15-minutes) approach is used, only one portion of the data set will be used. The sparse sampling realized volatility is defined over sub-grid i as:

$$RV_{sparse}^{(k)}=\sum_{t_i,t_{i+}\in G^{(k)}}^{m_k} (p_{t_{i+}}-p_{t_i})^2 \tag{3}$$

where $p_{t_{j+}}$ is the next observation within grid k, we can define the subsample average estimator as the average of all of the possible grids, or sub-samples:

$$RV_t^{(Avg)}=\frac{1}{k}\sum_{i=1}^{k} RV_{sparse}^{(k)} \tag{4}$$

This estimator, however, is still biased at high frequencies. The first best estimator in Zhang et al. (2005), known as the two time-scales estimator, uses $RV^{(Avg)}$ together with realized volatility calculated at the highest possible frequency possible m, $RV_t^{(m)}$:

$$RV_t^{(TS)}= RV_t^{(Avg)} - \frac{\bar{m}}{m} RV_t^{(m)} \tag{5}$$

where $\bar{m} = (m - k + 1)/k$. The asymptotically optimal number of subsamples, k_{opt} can be chosen as:

$$k_{opt}= \left(\frac{3\widehat{\sigma}_e^4}{\widehat{Q}_t}\right)^{\frac{1}{3}} m^{2/3} \tag{6}$$

where $\widehat{\sigma}_e^2$ and $\widehat{Q}_t$ are estimated as the optimal sampling frequency estimation in the optimally-sampled realized volatility estimator) by Bandi and Russell (2008).

The $RV_t^{(TS)}$ estimator, used in the present paper, is further improved by selecting the number of subsamples based on finite sample considerations, see Bandi and Russell (2008). $RV_t^{(TS)}$ can be rewritten as a modified Bartlett-type kernel estimator with the modification term deriving from subsampling (see, Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2008 and 2011). This modified two-scale estimator, select the following number of subsamples:

$$k^{fs}_{opt}= \left(1.5\frac{\left[RV_t^{(m/15)}/m\right]^2}{\widehat{Q}_t}\right)^{1/3} m \tag{7}$$

The resulting estimator employed, is denoted as $RV_t^{(TS,fs)}$.[3]

4.2 Forecasting Models

The present paper uses the maximum likelihood method for forecasting, which is mostly used in the related literature. One step (day) ahead forecasts are produced. Apart from the accuracy of different forecasting models, a specification of the forecasting method is examined. The rolling loop is compared to the recursive one. The selected specification of each model was chosen based firstly on literature, and secondly on general criteria for model selection. The selected specification is the one that minimizes most of the following four criteria: (i) Akaike Information Criterion (AIC), (ii) Finite Prediction Error (FPE), (iii) Swartz Bayesian Criterion (SBC), and (iv) Hannan and Qui (1979) (HQC). The forecasting performance is examined for the PCC, NN, $GARCH$, and HAR models relative to the $AR(p_{AIC})$ benchmark. Moreover, a direct comparison of those models is implemented via a Diebold Mariano test.

The AR model has superior forecasting accuracy and simplicity in its specification. However, not many properties of realized volatility estimators are incorporated in a AR model. The principal components combining (PCC) model is also simple in its perception and construction. It takes into account the advantages of the principal component analysis and the combining method of forecasts as well. However, there is no direct

[3] A useful summary upon the further modifications for bias corrections to the Zhang et al. (2005) class of estimators, in terms of prediction analysis, is provided by Bandi, Russell and Yang (2008).

link between the PCC specification and the realized volatility properties. The Neural Network (NN) model provides accurate volatility forecasts, based on literature. Moreover, NN seem to provide accurate realized volatility forecasts (Rosa et al., 2014). However, there is no direct link between each parameter selection and the properties of realized volatilities. Moreover, the NN specification is not easily selected and constructed. GARCH models accurately forecast volatility. However, the GARCH models do not provide as accurate forecasts for realized volatility series.

4.2.1 Principal components combining

Chan, Stock and Watson (1999) and Stock and Watson (2004) considered forming combination forecasts utilizing the first m principal components of the individual forecasts. Let m represent the first m estimated principal components of the un-centered second-moment matrix of the individual forecasts, $\hat{y}^h_{i,s+h|s}$, i $=$ 1, …, n, s $=$ R, …, t. To form a combination forecast of y^h_{t+h}, based on the fitted principal components, they estimated the regression model

$$y^h_{s+h} = \varphi_1 \hat{F}^h_{1,s+h|s} + ... + \varphi_m \hat{F}^h_{m,s+h|s} + \nu^h_{s+h} \tag{8}$$

where s $=$ R, …, t-h, R is the last observation of the in-sample period (just before the out-of-sample period) and h $=$ 1 step-ahead (forecasting horizon) for the daily volatility forecasts. The combination forecast is given by $y^h_{c,t+h|t} = \hat{\varphi}_1 \hat{F}^h_{1,t+h|t} + ... + \hat{\varphi}_m \hat{F}^h_{m,t+h|t}$ where $\hat{\varphi}_1, ..., \hat{\varphi}_m$ are the OLS estimates of $\varphi_1, ..., \varphi_m$ respectively in the above equation. Their results indicated either m $=$ 1 or m $=$ 2. The present paper employs the IC_{p^3} information criterion recently developed by Bai and Ng (2002) in order to select m (with four the maximum value of m) in computing forecasts using the Principal Component combining (PCC) method.

4.2.2 Neural networks

Neural Networks have been used in various fields to approximate complex nonlinear structures. Their name comes from the fact that they may be thought of as a network of neurons similar to (but of course much simpler as) the brain. The related computations may be extremely complex. Therefore, Neural Network analysis nowadays represents a subfield of computer science or, more precisely, of artificial intelligence.

According to Zhang (2007), any NN research should provide at minimum: (1) data spliting and processing, (2) architecture, (3) training settings, and (4) algorithm used. In terms of training settings, the researcher must juggle the number of input variable combinations to be trained, the interval of hidden neurons over which each network is to be tested, the number of randomly selected starting weights and the maximum number of runs. The product of all is the number of iterations.[4]

For the present paper, the following information covers these minimum requirements. Statistical normalization $x_n = (x_0 - \bar{x})/s$ of data series is selected for standardizing the inputs and outputs (Weigend et al.,

[4] See, Kaastra and Boyd (1996).

1992). All in-sample data series are selected as inputs to the NN.[5] The number of input nodes is selected according to the Akaike Information Criterion (AIC), as suggested in Cromwell et al. (1994). This number corresponds to the number of input nodes corresponds to the number of lagged observations used to discover the underlying pattern. The total number of nodes is the product of the AIC-selected input nodes. Common practice is to devide the time series into three sets: training, testing and validation sets[6]. The training set is used in order to learn the patterns present in data (first 65% of observations). Testing set is used to evaluate the generalization ability of a trained network (the next 15% of observations). Validation set consists of the most recent observations (the letest 20% of observations).

Across the NN finance literature, the selection of layers, hidden neurons tend to be 4: one input, one output and two hidden. Increasing the number of hidden layers, increases the risk of overfitting as well as computation time. The number of hidden nodes equals to the number of input nodes (Tang and Fishwick, 1993). According to Tang and Fishwick (1993), the number of hidden nodes does not have a significant effect on forecast performance, however. Because 1-minute quotes are used, the maximum number of weights is 90, which corresponds to 20 hidden neurons. The 20 hidden neurons have to be followed by 30 input neurons, and 5 output neurons.[7] The networks are fully connected in that all nodes in one layer are only fully connected to all nodes in the next higher layer except for the output layer (Tang and Fishwick, 1993).

The number of output nodes is 1 because of forecasting 1-step (day) ahead (Zhang, 1994). The activation (transfer) function (f) is the hyperbolic tangent transfer functions in both hidden and output layers (Zhang and Hutchinson, 1993). As training algorithm is selected the Levenberg-Marquardt nonlinear optimization method is employed (de Groot and Wurtz, 1991). The learning rate concerns how fast the network learns during training. This is set to 0.5 for all data series.

$$\hat{RV}_{t+1|n} = f_1\left(RV_{t|n}, RV_{t-1|n}, \ldots, RV_{t-p|n}\right) \tag{9}$$

Indicatively, a single hidden one-layer feed-forward network is considered, which may be best thought as a class of flexible nonlinear function of the form

$$f(X_{t-1}, \ldots, X_{t-p}) = \beta_0 + \sum_{j=1}^{q} G(\gamma_{0,j} + Y_t^T \gamma_j)\beta_j \tag{10}$$

where $Y_t = (X_{t-1}, \ldots, X_{t-p})^T$, Y_t is in place of $RV_{t+1|n}$, X_{t-p} is in place of $RV_{t-p|n}$, and the $\gamma_j = (\gamma_{1,j}, \ldots, \gamma_{p,j})^T$ are (px1) vectors for $j = 1, ..., q$, and $\beta_0, \beta_1, ..., \beta_q$ are scalar coefficients. p is the number of orders and q is the number of neurons. The function $G : \mathbb{R} \to [0, 1]$ is a pre-specified cumulative distribution function. Based on the NN literature, the hyperbolic function is $G(x) = tan\, h(x) = (e^{-x} - e^{-x})/(e^{x} + e^{-x})$. Functions of the above type can approximate broad classes of functions if is sufficiently large. Thus, if q

[5] For a detailed discussion, see Qi and Zhang (2001).

[6] For a detailed discussion, see Walczak (2001).

[7] For a detailed examination of the selection criteria for all NN paramets, see Qi and Zhang (2001).

increases with the sample size n, a good approximation of $f(X_{t-1},...,X_{t-p})$ will eventually result. The function in the above equation may also be estimated without specifying $G(.)$ by using the projection pursuit regression of Hutchinson et al. (1994).

4.2.3 GARCH

Engle (1982) introduced the Generalized Autoregressive Conditional Heteroscedastic (GARCH) model. GARCH is an AutoRegressive Conditional Heteroscedasticity model. Heteroskedasticity means non-constant variance. Dependence of the conditional variance on the past is the reason the process is not independent. Independence of the conditional mean on the past is the reason that the process is uncorrelated. The GARCH(p; q) model can be depicted as:

$$x_t = \epsilon_t \sigma_t^2 \tag{11}$$

$$\sigma_t^2 = \sqrt{a_0 + \sum_{i=1}^{q} a_i a_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2} \tag{12}$$

, where a_0, a_i and β_i are positive and $\epsilon_t \sim iid(0,1)$ (Gaussian white noise with unit variance). The present paper employs a GARCH (1, 1) forecasting model, based on the information set with number n of all in-sample observations.

$$\sigma_{t+1|n}^2 = E(\sigma_{t+1|n}^2) = a_0 + a x_t^2 + \beta \sigma_t^2 \tag{13}$$

The GARCH(1,1) has been chosen as the best among other GARCH-family models. As alternative to GARCH(1,1) model, the following GARCH models have been employed: GARCH-M, TGARCH, and EGARCH[8].

4.2.4 HAR

The Heterogeneous AutoRegressive (HAR) model uses lagged realized variance for a day, a week and a month as regressors. As researched in Corsi (2009) and Corsi and Reno (2012), there are used 5 and 22 days as the number of lags for the weekly and monthly regressors respectively.

$$\hat{RV}_{t+1|n} = \beta_0 + \beta_D RV_{(t+1)-1,t} + \beta_W RV_{(t+1)-5,t} + \beta_M RV_{(t+1)-20,t} \tag{14}$$

with $RV_{t,t+q} = q^{-1} \sum_{s=1}^{q} RV_{t+s}$ being the mean q-period lagged volatility. As explanatory variables, there are the daily (for q=1), the weekly (for q=5) and the monthly (for q=20) volatilities. The regression error

[8] Asymmetries are present in realized volatility series. However, the GARCH model that incorporates the asymmetric effects (EGARCH) did not provide more accurate forecasts than the GARCH(1,1). This is why such models have not been employed in the present forecasting competition.

term ϵ_t is allowed to have heteroskedasticity and serial correlation.

4.2.5 AR

The $AR(p_{AIC})$ forecasting model is the simplest linear forecasting model; yet it has a very good forecasting performance with regard to different data and sampling frequencies of high-frequency intraday data. The $AR(p_{AIC})$ process is

$$RV_{t+1|n} = a + b \cdot RV_{n+1-j|n}, \qquad \widehat{X}_{n+1-j|n} = X_{n+1-j} \text{ for } 1 \leq j \leq n \tag{15}$$

with j proposed by the AIC criterion. AR models are often used as benchmark models in the forecasting literature. In the present paper, the forecasting performance of the three nonlinear models is evaluated upon their performance relative to $AR(p_{AIC})$ forecasting model.

4.3 Evaluation criteria

Forecasts are evaluated via two loss functions, the Mincer Zarnowitz regression, a Directional criterion and the Diebold Mariano test.

4.3.1 Loss functions

Once the one-step ahead variance forecasts are available, their statistical performance can be assessed. To assess the statistical quality of the forecasts, two loss functions are used as well as a post-forecasting diagnostic regression. The mostly well-known loss functions in forecasting literature are the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), used by Patton (2011a) in order to evaluate the forecasting performance of realized volatility forecasts.

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2} \tag{16}$$

and

$$MAE = \frac{1}{n}\sum |e_t| \tag{17}$$

4.3.2 Forecast unbiasedness test

The well-known forecast unbiasedness Mincer Zarnowitz regression, introduced by Mincer and Zarnowitz (1969), where one regresses the forecasts on the actual values, is depicted as:

$$RV_{t+1} = \alpha + \beta \cdot \hat{RV}_{t+1|n} + u_{t+1|n} \tag{18}$$

where $\hat{RV}_{t+1|n}$ is the 1-step-ahead realized volatility forecast and RV_{t+1} is the actual realized volatility at time $t + 1$. Not a second model is incorporated in the regression. This is because the forecasting performance of the four models is compared through relative criteria to the AR benchmark model. The fit of this regression, through adjusted R^2, provides a measure of the relative performance of the forecasts in explaining the variability of the dependent variable: higher values indicate a better forecast. R^2 is selected as the best criterion from the Mincer Zarnowitz test because the t-test statistical significance of the regression coefficient is evident for almost all forecasts across all forecasting models. Haugom et al. (2014) have recently employed of the Mincer Zarnowitz regression for evaluating performance of models in realized volatility forecasting.

4.3.3 Direction

A directional criterion is used as well. This criterion indicates the times (as percentage) for which, the direction (first difference) of actual volatiity series is the same as the one (first difference) of forecasts.

$$\begin{array}{ll} if\ I\left(\Delta RV_{t+1}\right) = I\left(\Delta \hat{RV}_{t+1|n}\right) & \rightarrow\ Direction_t = 1 \\ otherwise & \rightarrow\ Direction_t = 0 \end{array} \tag{19}$$

4.3.4 Diebold Mariano test

Diebold and Mariano (1995) introduced the Diebold Mariano test for comparing the predictive ability of various forecasting models.[9] Giacomini and White (2006) suggested tests of the forecast accuracy conditional on some lagged variables, as lagged information set. For comparing the performance of models in realized volatility forecasting, Haugom et al. (2014) employed the Diebold Mariano (1995) test. Giacomini and White (2006) suggested imposing conditions to an equal accuracy test to enhance the test's strength. Such a test has been researched in the present paper.[10] However, there are no significant coefficients in the lagged conditional variables (lagged realized volatilities) for most of the comparisons across the forecasting models. This is evidence that the incorporated forecasting models can successfully capture the persistence property of realized volatility. So, there is no need for imposing any restriction (or assumption or condition) to the evaluation regressions. That is why, the Diebold Mariano (DM) test is employed for further comparing the forecasting accuracy of the competing models.

$$L\left(RV_{t+1}, \hat{RV}^{i}_{t+1|n}\right) - L\left(RV_{t+1}, \hat{RV}^{j}_{t+1|n}\right) = \beta_0 + e_t \tag{20}$$

where $\hat{RV}^{i}_{t+1|n}$ and $\hat{RV}^{i}_{t+1|n}$ are the 1-step-ahead realized volatility forecast of model i or j $(PCC, NN, GARCH, HAR$ and RV_{t+1} is the actual realized volatility at time $t + 1$. It is tested whether competing models have equal average predictive accuracy in the popular (pseudo-) distance measure (loss function) of QLIKE (see Patton, 2011b).

[9] Forecast evaluation tests were further studied in Diebold and Mariano (2002).

[10] Recently, Patton (2011b) employed such tests in the predictive ability of a series of realized volatility estimators.

$$QLIKE\ \ L\left(RV_{t+1}, \hat{RV}^{i}_{t+1|n}\right) = \frac{RV_{t+1}}{\hat{RV}^{i}_{t+1|n}} - log\left(\frac{RV_{t+1}}{\hat{RV}^{i}_{t+1|n}}\right) - 1 \tag{21}$$

The definition of QLIKE above has been normalised to yield a distance of zero when $\left(RV_{t+1} = \hat{RV}^{i}_{t+1|n}\right)$.

5 Empirical results

This section analyzes the forecasting performance of the models and methods alike across US financial markets. Notation is explained in Table 1. Table 2 indicates the relative forecasting performance of PCC, NN, $GARCH$ and HAR models as a ratio of the $RMSE$ criterion of each of these models to the $RMSE$ of the $AR(p)$ model. The $RMSE$ criterion concerns the Root Mean Square Error loss function. A ratio lower than 1 indicates that the model better forecasts volatility than the $AR(p_{AIC})$ benchmark. The PCC and HAR models are the only models that forecast better volatility than the $AR(p_{AIC})$ benchmark across all financial markets. Comparing the relative forecasting performance of the four models, HAR is the first, with PCC second, NN third and $GARCH$ last. The results of the RMSE evaluation are consistent between the rolling and recursive forecasting methods.

Table 3 presents the relative forecasting performance of PCC, NN, $GARCH$ and HAR models as a ratio of the MAE criterion of each of these models to the MAE of the $AR(p)$ model. The MAE criterion concerns the Mean Absolute Error loss function. A ratio lower than 1 indicates that the model better forecasts volatility than the $AR(p_{AIC})$ benchmark. The PCC and HAR models are the only models that forecast better volatility than the $AR(p_{AIC})$ benchmark across all financial markets. The first two rankings are occupied again by HAR and PCC. In the rolling forecasts, HAR is first with PCC second; whereas, in the recursive forecasts, PCC is first with HAR second. In both forecasting methods, NN comes third and $GARCH$ last.

Table 4 indicates the relative forecasting performance of PCC, NN, $GARCH$ and HAR models as a ratio of the R^2 criterion of each of these models to the R^2 of the $AR(p)$ model. R^2 criterion concerns the adjusted-R^2 of the Mincer zarnowitz regresion. A ratio higher than 1 indicates that the model better forecasts volatility than the $AR(p_{AIC})$ benchmark. There is no model with better forecasting performance than $AR(p_{AIC})$, in rolling forecasts. In recursive forecasts, only the HAR model has better performance than $AR(p_{AIC})$; second comes PCC, third NN, and last $GARCH$.

The relative forecasting performance of PCC, NN, $GARCH$ and HAR models as a ratio of the $Direction$ criterion of each of these models to the $Direction$ of the $AR(p)$ model is presented in Table 5. $Direction$ indicates the times (as percentage) for which, the direction (first difference) of actual volatiity series is the same as the one (first difference) of forecasts. A ratio higher than 1 indicates the model better forecasts volatility than the benchmark $AR(p_{AIC})$. In rolling forecasts, PCC is the only model that better forecasts volatility than the $AR(p_{AIC})$ benchmark across all financial markets. Second in order comes NN, third $GARCH$ and last HAR. In recursive forecasts, PCC is still first; however, HAR is now second with NN

third and $GARCH$ last.

The forecasting performance of the five models is assessed via the *Diebold Mariano* comperative predictive accuracy test. To strengthen the findings, a robustness check is implemented. Lagged realized volatility series are added as imposed conditions to the Diebold Mariano test, as suggested in Giacomini and White (2006). The comparative outcome (ranking) of the forecasting competition does not change because of imposing conditions to the Diebold Mariano test. Tables 6A and 6B report the estimated difference in average distance (t-statistic in brackets) between each i and j model (NN, $GARCH$, PCC or HAR).[11] Table 6A concerns the SE, SFE, ETF and EIF markets; and Table 6B concerns the TBF, EF and CO markets. A negative and statistically significant difference means that the model in the column is better than the one in the respective line. In most of comparisons in pairs, the statistically significant differences were positive. A positive and statistically significant difference means that the model in the line is better than the one in the respective column. For the spot equity market (SE; Table 6A, panel A), the order of models ranking is: (1st) HAR, (2nd) PCC, (3rd) $GARCH$, (4th) AR, and (5th) NN. For the spot foreign exchange market (SFE; Table 6A, panel B), the order of models ranking is: (1st) HAR, (2nd) AR, (3rd) $GARCH$, (4th) PCC, and (5th) NN. For the exchange traded funds market (ETF; Table 6A, panel C), the order of models ranking is: (1st) HAR, (2nd) $GARCH$, (3rd) PCC, (4th) AR, and (5th) NN. For the equity index futures market (EIF; Table 6A, panel D), the order of models ranking is: (1st) AR, (2nd) HAR, (3rd) $GARCH$, (4th) NN, and (5th) PCC. For the US treasury bonds futures market (TBF; Table 6B, panel A), the order of models ranking is: (1st) PCC, (2nd) NN, (3rd) HAR, (4th) AR, and (5th) $GARCH$. For the energy futures market (EF; Table 6B, panel B), the order of models ranking is: (1st) HAR, (2nd) $GARCH$, (3rd) PCC, (4th) AR, and (5th) $GARCH$. For the commodities options market (CO; Table 6B, panel C), the order of models ranking is: (1st) AR, (2nd) HAR, (3rd) $GARCH$, (4th) PCC, and (5th) NN. In four out of seven markets, HAR is the best model. Second best model may be considered PCC. In four out of seven markets, $GARCH$ is the third best model. In four out of seven markets, AR is the fourth best and NN is the last.

In Panel D, it is indicated in how many out of the total of seven financial markets there is a statistically significant and negative in sign difference. Across all markets, negative and statistically significant differences were: (i) in five out of seven markets, where $GARCH$ was better than PCC; (ii) in four out of seven markets, where PCC was better than AR; and (iii) in four out of seven markets, where HAR was better than AR. Positive and statistically significant differences were: (i) in four out of seven markets, where HAR was better than NN; (ii) in seven out of seven markets, where HAR was better than $GARCH$; (iii) in four out of seven markets, where AR was better than $GARCH$; and (iv) in six out of seven markets, where HAR was better than PCC. Combibing the results from both negative and positive statistically significant differences across all markets (Table 6B, panel D), HAR is clearly the best model. In the second place comes PCC with $GARCH$; next, comes AR and last NN.

Results are summarized in Table 7. The best forecasting model is reported for each financial market and

[11] Evaluation only on rolling forecasts is reported; results do not significantly change for recursive forecasts.

evaluation method. The Heterogeneous Autoregressive (HAR) and the Principal Components Combining (PCC) models are the best modes. However, HAR is slightly better than PCC. Results are consistent between rolling and recursive forecasts.

6 Concluding remarks

An interesting outcome is that the performance of forecasting models with entirely different specifications converge to a similar forecasting performance; this result is consistent across different evaluation criteria and financial time series. Long memory seems to matter a lot in realized volatility forecasting. The HAR model produced accourate enough forecasts to compete with the second best model, the PCC. This evidence is valid for all evaluation methods and consistent in both rolling and recursive forecasts.

According to all $RMSE$, MAE, R^2 and $Direction$ criteria, the HAR model is the best; second comes PCC; third is NN; and last $GARCH$. By looking at each market's $Diebold\ Mariano$ comparative evaluation (Tables 6A-6B), HAR is the best model (in four out of seven markets), PCC the second best model, $GARCH$ is the third best model (in four out of seven markets), AR is the fourth best (in four out of seven markets) and NN is the last (in four out of seven markets). Exactly, the same results are retrieved by the combibing the results from both negative and positive statistically significant differences across all markets (Table 6B, panel D).

Across all evaluation methods ($RMSE$, MAE, R^2 and $Direction$ criteria and $Diebold\ Mariano$ test), the Heterogeneous Autoregressive (HAR) model produces most accurate forecasts, with the Principal Components Combining (PCC) model second. Moreover, there is no clear cut evidence about the order of significance for the Neural Networks (NN), Generalized AutoRegressive Conditional Heteroscedastic ($GARCH$) and Autoregressive $AR(p_{AIC})$ models. The PCC and HAR models are the only models that forecast better volatility than the $AR(p_{AIC})$ benchmark across all financial markets, for both $RMSE$ and MAE loss functions, the R^2 and the $Direction$ criteria as well. This is also evident in the $Diebold\ Mariano$ test results.

Moreover, the recursive method provided more clear evidence compared to the rolling; especially, for the R^2 and $Direction$ criteria. So, it is suggested to be followed in realized volatility forecasting either with linear or nonlinear models. Results also revealed the superior forecasting ability of a linear model (HAR) compared to nonlinear (PCC, NN and $GARCH$) models and a linear AR benchmark. Moreover, there is evidence that a different GARCH model should have been employed, and the GARCH models tested weren't enough to produce accurate realized volatility forecasts and encapturing the realized volatility properties. The $GARCH$ forecasting performance here, contradicts to the results of Andersen et al. (2011) and Boudt et al. (2013).

References

[1] Ait-Sahalia, Y. & Mancini, L. (2008). Out of sample forecasts of quadratic variation. Journal of Econometrics, 147(1), 17-33.

[2] Andersen, T. & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39, 885-905.

[3] Andersen, T.G., Bollerslev, T. & Diebold, F.X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. Review of Economics and Statistics, 89 (4), 701-720.

[4] Andersen, T., Bollerslev, T., Diebold, F. & Labys, P. (2001). The distribution of exchange rate volatility. Journal of the American Statistical Association, 96, 42-55.

[5] Andersen, T., Bollerslev, T., Diebold, F. & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71, 579-625.

[6] Andersen, T., Bollerslev, T. & Lange, S. (1999). Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon. Journal of Empirical Finance, 6, 457-477.

[7] Andersen, T., Bollerslev, T. & Meddahi, N. (2011). Realized volatility forecasting and market microstructure noise. Journal of Econometrics, 160, 220-234.

[8] Aragonï¿œs, J.R., Blanco, C. & Estï¿œvez, P.G. (2007). Neural network volatility forecasts. Intelligent Systems in Accounting, Finance and Management, 15, 107-121.

[9] Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70, 191-221.

[10] Bai, J. & Ng, S. (2008). Forecasting economic time series using targeted preditors. Journal of Economics, 146, 304-317.

[11] Bandi, F. & Russell, J. (2008). Microstructure noise, realized variance, and optimal sampling. Review of Economic Studies, 75, 339-69.

[12] Bandi, F. & Russell, J. (2011). Market microstructure noise, integrated variance estimators, and the accuracy of asymptotic approximations. Journal of Econometrics, 160, 145-159.

[13] Bandi, F., Russell, J. & Yang, C. (2008). Realized volatility forecasting and option pricing. Journal of Econometrics, 147, 34-46.

[14] Barndorff-Nielsen, O., Hansen, P., 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.

[15] Barndorff-Nielsen, O., Hansen, P., Lunde, A. & Shephard, N. (2011). Subsampling realised kernels. Journal of Econometrics, 160, 204-219.

[16] Bekaert, G. & Hoerova, M. (2014). The VIX, the variance premium and stock market volatility. Journal of Econometrics, 183, 181-192.

[17] Boivin, J. & Ng, S. (2006). Are more data always better for factor analysis? Journal of Econometrics, 132, 169–194.

[18] Boudt, K., Danielsson, J. & Laurent, S. (2013). Robust forecasting of dynamic conditional correlation GARCH models. International Journal of Forecasting, 29(2), 244-257.

[19] den Butter, F. & Jansen, P. (2013). Beating the random walk: A performance assessment of long-term interest rate forecasts. Applied Financial Economics, 23, 749-765.

[20] Celik, S. & Ergin, H. (2014). Volatility forecasting using high frequency data: Evidence from stock markets. Economic Modelling, 36, 176-190.

[21] Chan, Y.L., Stock, J.H. & Watson, M.W. (1999). A dynamic factor model framework for forecast combination. Spanish Economic Review, 1, 21 – 121.

[22] Clark, T.E. & Ravazzolo, F. (2014). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. Journal of Applied Econometrics, forthcoming.

[23] Coakley, J. & Brown, C. (2000). Artificial neural networks in accounting and finance: Modeling issues. International Journal of Intelligent Systems in Accounting, Finance & Management, 9, 119-144.

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

[25] Corsi, F. & Reno, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. Journal of Business and Economic Statistics, 30(3), 368-380.

[26] Cromwell, J.B., Labys, W.C. & Terraza, M. (1994). Univariate Tests for Time Series Models. Saga Publications, Thousand Oaks.

[27] Diebold, F.X. & Mariano, R.S. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics, 13(3), 253-263.

[28] Diebold, F.X. & Mariano, R.S. (2002). Comparing predictive accuracy. Journal of Business and Economic Statistics, 20(1), 134-144.

[29] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50, 987-1007.

[30] Fernandes, M., Medeiros, M.C. & Scharth, M. (2014). Modeling and predicting the CBOE market volatility index. Journal of Banking and Finance, 40, 1-10.

[31] Ferrara, L., Marsilli, C. & Ortega, J.-P. (2014). Forecasting growth during the Great Recession: Is financial volatility the missing ingredient? Economic Modelling 36, 44-50.

[32] Fung, W. & Hsieh, D. (1999). A premier on hedge funds. Journal of Empirical Finance, 6, 309-331.

[33] Genre, V., Kenny, G., Meyler, A. & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting, 29, 108-121.

[34] Ghysels, E. & Sinko, A. (2011). Volatility forecasting and microstructure noise. Journal of Econometrics, 160, 257-271.

[35] Giacomini, R. & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74, 1545-1578.

[36] de Gooijer, J. & Hyndman, R. (2006). 25 years of time series forecasting. International Journal of Forecasting, 22, 443-473.

[37] de Groot, C. & Wurtz, D. (1991). Analysis of univariate time series with connectionist nets: a case study of two classical examples. Neurocomputing, 3, 177–192.

[38] Hajizadeh, E., Seifi, A., Zarandi, H. & Turksen, B. (2012). A hybrid modeling for forecasting the volatility of S&P 500 index return. Expert Systems with Applications, 39, 431-436.

[39] Hamid, A. & Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P500 index future prices. Journal of Business Research, 57(10), 1116-1125.

[40] Hansen, P.R. & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics, 20(7), 873-889.

[41] Haugom, E., Langeland, H., Molnar, P. & Westgaard, S. (2014). Forecasting volatility of the U.S. oil market. Journal of Banking and Finance, 47, 1-14.

[42] Heij, C., van Dijk, D. & Groenen, P. (2008). Macroeconomic forecasting with matched principal components. International Journal of Forecasting, 24, 87-100.

[43] Heij, C., Grenen, P. & van Dijk, D. (2007). Forecast comparison of principal component regression and principal covariance regression. Computational Statistics & Data Analysis, 51, 3612-3625.

[44] Hutchinson, J. M., Lo, A. W. & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. Journal of Finance, 49, 851 − 889.

[45] Kaastra, I. & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. Neurocomputing, 10, 215-236.

[46] Kristjanpoller, W., Fadic, A. & Minutolo, M. (2014). Volatility forecast using hybrid Neural Network models. Expert Systems with Applications, 41, 2437-2442.

[47] Lopez, J.A. (2001). Evaluation of predictive accuracy of volatility models. Journal of Forecasting, 20(1), 87-109.

[48] Louzis, D.P., Xanthopoulos-Sisinis, S. & Refenes, A.P. (2013). The role of high-frequency intra-daily data, daily range and implied volatility in multi-period value-at-risk forecasting. Journal of Forecasting 32 (6), 561-576.

[49] Ludvigson, S.C. & Ng, S. (2009). Macro factors in bond risk premia. Review of Financial Studies, 22, 5027-5067.

[50] Mancini, L., Ranaldo, A. & Wrampelmeyer, J. (2013). Liquidity in the foreign exchange market: Measurement, commonality, and risk premiums. Journal of Finance, 68 (5), 1805-1841.

[51] Mincer, J. & Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. Columbia University Press.

[52] de Mol, C., Giannone, D. & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? Journal of Econometrics, 146, 318-328.

[53] Nachane, D. & Clavel, J.G. (2008). Forecasting interest rates: a comparative assessment of some second-generation nonlinear models. Journal of Applied Statistics, 35, 493-514.

[54] Neely, C., Rapach, D., Tu, J. & Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators. Management Science, 60(7), 1772-1791.

[55] Nomikos, N.K. & Poullasis, P.K. (2014). Petroleum term structure dynamics and the role of regimes. Journal of Futures Markets, forthcoming.

[56] Patton, J. (2011a). Volatility forecast evaluation and comparison using imperfect volatility proxies. Journal of Econometrics, 160, 246–256.

[57] Patton, A.J. (2011b). Data-based ranking of realized volatility estimators. Journal of Econometrics 161, 284-303.

[58] Qi, M. & Zhang, G.P. (2001). An investigation of model selection criteria for neural network time series forecasting. European Journal of Operational Research, 132, 666-680.

[59] Reeves, J.J. & Xie, X. (2014). Forecasting stock return volatility at the quarterly frequency: An evaluation of time series approaches. Applied Financial Economics, 24(5), 347-356.

[60] Rosa, R., Maciel, L., Gomibe, F. & Ballini, R. (2014). Evolving hybrid neural fuzzy network for realized volatility forecasting with jumps. Computational Intelligence for Financial Engineering and Economics (CIFEr), 2014 IEEE Conference on, 481-488.

[61] Santos, D.G. & Ziegelmann, F.A. (2014). Volatility forecasting via MIDAS, HAR and their combination: An empirical comparative study for IBOVESPA. Journal of Forecasting, 33(4), 284-299.

[62] Seo, S.W. & Kim, J.S. (2015). The information content of option-implied information for volatility forecasting with investor sentiment. Journal of Banking and Finance, 50, 106-120.

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

[64] Steeley, J.M. (2014). Forecasting the term structure when short-term rates are near zero. Journal of Forecasting, 33(5), 350-363.

[65] Stock, J. H. & Watson, M. W. (2004). Combining forecasts of output growth in a seven-country data set. Journal of Forecasting, 23, 405 – 430.

[66] Tang, Z. & Fishwick, P.A. (1993). Feedforward neural nets as models for time series forecasting. ORSA Journal on Computing, 5, 374–385.

[67] Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. Journal of Management Information Systems, 17, 203-222.

[68] Weigend, A.S., Huberman, B.A. & Rumelhart, D.E. (1992). Predicting sunspots and exchange rates with connectionist networks. In: Casdagli, M., Eubank, S. (Eds.), Nonlinear Modeling and Forecasting. Addison-Wesley, Redwood City, CA, pp. 395–432.

[69] Yang, K. & Chen, L. (2014). Realized volatility forecast: Structural breaks, long memory, asymmetry, and day-of-the-week effect. International Review of Finance, 14(3), 345-392.

[70] Yang, K., Chen, L. & Tian, F. (2015). Realized volatility forecast of stock index under structural breaks. Journal of Forecasting, 34(1), 57-82.

[71] Zhang, X. (1994). Time series analysis and prediction by neural networks. Optimization Methods and Software, 4, 151–170.

[72] Zhang, G.P. (2007). Avoiding pitfalls in neural network research. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 37, 3-16.

[73] Zhang, L., Mykland, P. & Ait-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.

[74] Zhang, G.P. & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, 160, 501-514.

Tables

Table 1. Notation

Description	*Symbol*
Panel A. US financial markets	
Spot equity: Dow Jones Industrial Average ($INDU$)	SE
Spot foreign exchange rates: EUR/USD exchange rate	SFE
Exchange traded funds: PowerShares QQQ	ETF
Equity index futures: E-Mini Dow futures continuous contract (YM)	EIF
US Treasury bonds futures: 30-year US Treasury yield (TYX)	TBF
Energy futures: Crude oil miNY futures continuous contract (QM)	EF
Commodities options: CBOE gold index options (GOX)	CO
Average figures across the board of US financial markets	$Markets$
Panel B. Forecasting models	
Neural networks	NN
GARCH (1, 1)	$GARCH$
Principal components combined	PCC
Heterogeneous AutoRegressive	HAR
AR with AIC selected level of order	$AR(p_{AIC})$
Panel C. Evaluation criteria	
Root mean square error	$RMSE$
Mean absolute error	MAE
R-square of the Mincer-Zarnowitz regression	R^2
Directional criterion	$Direction$
Diebold Mariano test	$Diebold\ Mariano$

Notes. Table 1 lists the description with the corresponding symbol for US financial markets, forecasting models, and evaluation criteria.

Table 2. Forecast evaluation - $RMSE$

	SE	SFE	ETF	EIF	TBF	EF	CO	$Markets$
				Panel A. Rolling				
NN	1.4966	7.3427	1.1185	1.1016	1.3790	1.4345	1.2844	2.1653
$GARCH$	3.0790	2.0072	2.1454	2.2803	2.5994	*0.8500*	9.7688	3.2472
PCC	*0.8186*	1.3780	*0.0017*	*0.0191*	*0.0011*	*0.0269*	1.0183	*0.4662*
HAR	*0.1142*	*0.1034*	*0.1504*	*0.1786*	*0.2220*	*0.2220*	*0.0679*	*0.1648*
				Panel B. Recursive				
NN	1.5295	6.9244	1.1612	1.0766	1.3816	1.4045	1.2862	2.1091
$GARCH$	3.0793	2.0058	2.1472	2.2672	2.6092	*0.8569*	9.8841	3.2642
PCC	*0.8191*	1.3784	*0.0018*	*0.0225*	*0.0005*	*0.0276*	1.0238	*0.4677*
HAR	*0.8285*	1.3046	*0.7632*	*0.9496*	1.4244	*0.9385*	*0.9613*	*0.9773*

Notes. Table 2 presents the ratio of the $RMSE$ forecast evaluation of each forecasting model (NN, $GARCH$ or PCC) to the respective evaluation of the $AR(p_{AIC})$ model. A ratio lower than 1 indicates that the model better forecasts volatility than the $AR(p_{AIC})$ benchmark. Significance is indicated in *Italics*. $Markets$ deploits the average across the board of US financial markets.

Table 3. Forecast evaluation - MAE

	SE	SFE	ETF	EIF	TBF	EF	CO	$Markets$
				Panel A. Rolling				
NN	2.7678	8.8287	2.2331	1.4223	1.7857	1.7455	1.4863	2.8956
$GARCH$	3.6371	2.4171	4.7195	2.8387	3.5909	1.2104	9.0985	3.9303
PCC	*0.8474*	1.4198	*0.0002*	*0.0018*	*0.0001*	*0.0020*	*0.9621*	*0.4619*
HAR	*0.0958*	*0.0892*	*0.0960*	*0.1710*	*0.1528*	*0.2137*	*0.0709*	*0.1271*
				Panel B. Recursive				
NN	2.8919	8.0192	2.1673	1.3789	1.7517	1.7084	1.5171	2.7764
$GARCH$	3.6426	2.4157	4.7243	2.8361	3.5983	1.2194	9.1346	3.9387
PCC	*0.8317*	1.4208	*0.0002*	*0.0021*	*4.22e-5*	*0.0020*	*0.9660*	*0.4604*
HAR	*0.8868*	1.2590	*0.8384*	1.0759	1.1452	1.0680	*0.9631*	1.0338

Notes. Table 3 presents the ratio of the MAE forecast evaluation of each forecasting model (NN, $GARCH$, PCC or HAR) to the respective evaluation of the $AR(p_{AIC})$ model. A ratio lower than 1 indicates that the model better forecasts volatility than the $AR(p_{AIC})$ benchmark. Significance is indicated in *Italics*. $Markets$ deploits the average across the board of US financial markets.

Table 4. Forecast evaluation - R^2

	SE	SFE	ETF	EIF	TBF	EF	CO	$Markets$
				Panel A. Rolling				
NN	0.6740	0.1110	0.0879	0.8779	0.2768	0.0390	0.4530	0.3599
$GARCH$	0.5359	0.6413	0.8529	0.7311	*1.1357*	0.8838	0.2885	0.7242
PCC	0.3311	0.8183	*2.0890*	0.4474	0.8840	0.4215	0.4209	0.7732
HAR	0.3966	0.2313	0.2883	0.2608	0.3341	0.2937	0.3726	0.3111
				Panel B. Recursive				
NN	0.2423	0.4311	0.4429	*1.0854*	0.6552	0.1600	0.1687	0.4551
$GARCH$	0.5384	0.6419	0.8556	0.7345	*1.1326*	0.8732	0.2874	0.7234
PCC	0.2963	0.8372	*2.0854*	0.5659	*3.2212*	0.4358	0.3464	*1.1126*
HAR	*1.3649*	0.8758	*1.7755*	*1.1948*	*3.0149*	*1.0266*	*1.0705*	*1.4747*

Notes. Table 4 presents the ratio of the R^2 forecast evaluation of each forecasting model (NN, $GARCH$, PCC or HAR) to the respective evaluation of the $AR(p_{AIC})$ model. A ratio higher than 1 indicates that the model better forecasts volatility than the $AR(p_{AIC})$ benchmark. Significance is indicated in *Italics*. $Markets$ deploits the average across the board of US financial markets.

Table 5. Forecast evaluation - *Direction*

	SE	SFE	ETF	EIF	TBF	EF	CO	$Markets$
				Panel A. Rolling				
NN	0.9242	0.7217	0.8676	0.8246	0.9088	0.8615	0.9224	0.8615
$GARCH$	0.8265	0.7280	0.8132	0.7231	0.7703	0.6780	0.7557	0.7564
PCC	*1.0051*	0.8113	*1.6618*	*1.3835*	*1.5169*	*1.3641*	*1.0259*	*1.2526*
HAR	0.2156	0.2148	0.1876	0.1912	0.1943	0.1880	0.2164	0.2011
				Panel B. Recursive				
NN	0.8674	0.7476	0.8934	0.8371	0.8649	0.8385	0.9106	0.8514
$GARCH$	0.8250	0.7287	0.8107	0.7198	0.7682	0.6698	0.7527	0.7536
PCC	1.0240	0.8223	*1.6618*	*1.3835*	*1.5169*	*1.3641*	0.9835	*1.2464*
HAR	*1.0126*	0.9025	*1.0083*	0.9312	*1.0007*	0.8963	*1.0071*	0.9655

Notes. Table 5 presents the ratio of the *Direction* forecast evaluation of each forecasting model (NN, $GARCH$, PCC or HAR) to the respective evaluation of the $AR(p_{AIC})$ model. This criterion indicates the times (as percentage) for which, the direction of actual volatiity series is the same as the one of forecasts. A ratio higher than 1 indicates the model better forecasts volatility than the benchmark $AR(p_{AIC})$. Significance is indicated in *Italics*. $Markets$ deploits the average across the board of US financial markets.

Table 6A. Forecast evaluation - *Diebold Mariano*

	NN	*GARCH*	*PCC*	*HAR*
		Panel A. *SE*		
GARCH	14.5914 (1.5333)	-		
PCC	-0.3507 (-0.5826)	*4.2917 (7.7872)*	-	
HAR	34.3719 (1.9623)	*19.7805 (3.4515)*	*15.4888 (4.6498)*	-
AR	12.0448 (1.0648)	*-2.5465 (-10.0482)*	*-6.8382 (-3.5843)*	*-22.3270 (-3.0654)*
		Panel B. *SFE*		
GARCH	10.0784 (1.1790)	-		
PCC	-3.5293 (-0.4852)	*-2.1022 (-6.3757)*	-	
HAR	*98.4242 (3.7314)*	*88.3457 (4.1601)*	*90.4480 (4.0617)*	-
AR	12.7982 (1.2912)	*2.7198 (5.1112)*	*4.8221 (3.6626)*	*-85.6259 (-4.2948)*
		Panel C. *ETF*		
GARCH	*10.2137 (3.1992)*	-		
PCC	-14.6125 (-0.8686)	*-7.1497 (-4.5242)*	-	
HAR	*23.6996 (2.4275)*	*13.4859 (4.8378)*	*20.6356 (2.7851)*	-
AR	*1.4764 (6.3269)*	*-8.7373 (-3.6586)*	*-1.5876 (-5.3039)*	*-22.2232 (-2.5745)*
		Panel D. *EIF*		
GARCH	*1.8372 (2.2576)*	-		
PCC	*-4.1677 (-2.5497)*	*-0.4059 (-12.4972)*	-	
HAR	3.1554 (1.1686)	*1.3183 (3.7993)*	*1.7242 (2.7901)*	-
AR	*15.9864 (4.4780)*	*14.1493 (5.0410)*	*14.5552 (4.9229)*	*12.8310 (5.5845)*

Notes. Table 6A presents the *Diebold Mariano* forecast evaluation (the estimated difference in average distance with the t-statistic in brackets) between each i and j model (NN, $GARCH$, PCC or HAR). Table 6A concerns the SE, SFE, ETF and EIF markets. Significance is indicated in *Italics*. Evaluation only on rolling forecasts is reported; results do not significantly change for recursive forecasts.

Table 6B. Forecast evaluation - *Diebold Mariano*

	NN	*GARCH*	*PCC*	*HAR*
		Panel A. *TBF*		
GARCH	*-2.0415 (-8.7500)*	-		
PCC	-3.7435 (-0.8712)	*3.7507 (3.1199)*	-	
HAR	*-1.2293 (-14.5733)*	*0.8122 (7.7918)*	*-2.9386 (-4.0520)*	-
AR	*-0.9305 (-20.5506)*	*1.1109 (8.2328)*	*-2.6398 (-4.9717)*	*0.2988 (31.1233)*
		Panel B. *EF*		
GARCH	5.7458 (1.3381)	-		
PCC	*0.9250 (2.0049)*	*-0.4045 (-23.7153)*	-	
HAR	12.4859 (1.4351)	*6.7401 (2.8087)*	*7.1447 (2.6097)*	-
AR	4.1261 (1.3997)	*-1.6196 (-5.0388)*	*-1.2151 (-6.8472)*	*-8.3598 (-2.1797)*
		Panel C. *CO*		
GARCH	*5.2271 (6.2169)*	-		
PCC	-0.0974 (-1.1615)	*-1.8156 (-20.6807)*	-	
HAR	*17.1216 (4.5439)*	*11.8945 (6.5173)*	*13.7100 (5.8212)*	-
AR	*24.1488 (4.4518)*	*18.9217 (5.6802)*	*20.7373 (5.2612)*	*7.0272 (17.4114)*
		Panel D. All markets		
GARCH	1 / 7 (3 / 7)			
PCC	1 / 7 (1 / 7)	5 / 7 (2 / 7)		
HAR	1 / 7 (4 / 7)	0 / 7 (7 / 7)	1 / 7 (6 / 7)	
AR	1 / 7 (3 / 7)	3 / 7 (4 / 7)	4 / 7 (3 / 7)	4 / 7 (3 / 7)

Notes. Table 6B presents the *Diebold Mariano* forecast evaluation (the estimated difference in average distance with the t-statistic in brackets) between each i and j model (*NN*, *GARCH*, *PCC* or *HAR*). Table 6B concerns the *TBF*, *EF* and *CO* markets. In Panel D, it is indicated in how many out of the total of seven financial markets there is a statistically significant and negative in sign difference. In brackets of Panel D, is indicated in how many out of the total of seven financial markets there is a statistically significant and positive in sign difference. Significance is indicated in *Italics*. Evaluation only on rolling forecasts is reported; results do not significantly change for recursive forecasts.

Table 7. Summarized Results

	SE	*SFE*	*ETF*	*EIF*	*TBF*	*EF*	*CO*	*Markets*
				Panel A. Rolling				
RMSE	*HAR*	*HAR*	*PCC*	*PCC*	*PCC*	*PCC*	*HAR*	*HAR*
MAE	*HAR*	*HAR*	*PCC*	*PCC*	*PCC*	*PCC*	*HAR*	*HAR*
R^2	*NN*	*PCC*	*PCC*	*NN*	*GARCH*	*GARCH*	*NN*	*PCC*
Direction	*PCC*	*PCC*	*PCC*	*PCC*	*PCC*	*PCC*	*PCC*	*PCC*
				Panel B. Recursive				
RMSE	*PCC*	*HAR*	*PCC*	*PCC*	*PCC*	*PCC*	*HAR*	*PCC*
MAE	*PCC*	*PCC*	*PCC*	*PCC*	*PCC*	*PCC*	*HAR*	*PCC*
R^2	*HAR*	*HAR*	*PCC*	*HAR*	*PCC*	*HAR*	*HAR*	*HAR*
Direction	*PCC*	*HAR*	*PCC*	*PCC*	*PCC*	*PCC*	*HAR*	*PCC*
Diebold Mariano	*HAR*	*HAR*	*HAR*	*AR*	*PCC*	*HAR*	*AR*	*HAR*

Notes. Table 7 reports the best forecasting model for each financial market and for each evaluation method.