APPLIED ECONOMICS

Forecasting Realized Volatility of the Oil Future Prices via Machine Learning*

Byung-June Kim[a], Taeyoon Kim[b], Myung-Jun Kim[c, †], and Bong-Gyu Jang[d,†]

[a]Department of Industrial and Management Engineering, POSTECH, Korea. Tel: +82-54-279-2864, Fax: +82-54-279-2870, kbj219@postech.ac.kr

[b]Department of Industrial and Management Engineering, POSTECH, Korea. Tel: +82-54-279-2864, Fax: +82-54-279-2870, taeyoonkim.ime2016@postech.ac.kr

[c,†]Corresponding Author, Department of Industrial and Management Engineering, POSTECH, Korea. Tel: +82-54-279-2864, Fax: +82-54-279-2870, kmj4720@postech.ac.kr

[d,†]Corresponding Author, Department of Industrial and Management Engineering, POSTECH, Korea. Tel: +82-54-279-2864, Fax: +82-54-279-2870, bonggyujang@postech.ac.kr

ABSTRACT

This paper explores the potential use of machine learning models in crude oil realized volatility forecasting through a variety of empirical analyses and robustness checks. Although the conventional Heterogeneous Autoregressive (HAR) model is widely accepted, the machine learning models with the HAR factors can significantly improve the forecasting performance. We also found that macroeconomic variables such as supply factors, implied volatility indices and uncertainty factors can be useful in forecasting oil volatility.

KEYWORDS

Volatility Forecasting; Oil Future; Machine Learning; Forecasting Model

JEL CLASSIFICATION

C5, C22, G1, Q4

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea.(NRF-2022S1A3A2A02089950)

1. Introduction

Volatility is a key characteristic of commodity markets, particularly evident in the oil sector. Historically, oil prices have been subject to considerable fluctuations, driven by a variety of factors including geopolitical unrest, imbalances in supply and demand, market speculation, and major economic changes(Hamilton, 2009). The advent of the COVID-19 pandemic has exerted a profound impact on the commodity market, leading to "unprecedented disruptions in worldwide supply chains, dramatic shifts in oil demand, and increased economic instability."

These events have triggered severe swings in oil prices, influencing a wide spectrum of market stakeholders from traders and investors to policymakers. The unparalleled drop in demand, severe conditions in the job market, and restrained consumer spending resulting from the COVID-19 crisis are unparalleled in living history. A stark example of this was when dwindling demand drove oil future prices into negative figures, reaching -$36.98 in May 2020.

The extraordinary plunge in oil prices is not to be underestimated due to the considerable economic repercussions it carries for industries reliant on oil(Apergis and Miller, 2009; and Kilian, 2009).[1] Even prior to the global ramifications of COVID-19, uncertainty around oil prices had been on a gradual rise over the past decades, posing a substantial risk to the global economy.

This paper explores the possibility of the potential usage of machine learning in the field of volatility forecasting. Comparison of various forecasting models with a rich set of data and various forecasting horizons guides us to the fact that a combination of conventional models and machine learning techniques can improve the forecasting performance in out-of-sample. Ma et al. (2017) and Ma et al. (2018) investigated autoregressive-type forecasting models for oil price volatility, including the heterogeneous autoregressive (HAR) models. However, they covered only autoregressive-type models and did not check the time consistency of the forecasting performance.Unfortunately, although

[1]The 1980s oil glut, for example, has caused a blow to the exports of the Soviet Union, and it may have been a factor in accelerating the collapse of the Soviet Union. On the other side, many crop prices also remained in the doldrums amid weak raw material prices, which resulted in a recession in the U.S. agricultural industry, leading to bankruptcy in many small banks. As another example, the increased U.S. shale supply caused a sharp fall in oil prices again in 2014, which has yet to recover since then. As a result, Russia fell into recession, and growth in the Middle East slowed.

HAR-X type models are successful in-sample prediction, the performance of HAR-X is not as effective as that of other machine learning models in out-of-sample. We show that the machine learning models with factors in the HAR model as input variables have the potential to enhance the out-of-sample forecasting performance with one-week, bi-week, and month ahead horizons.

We examine the out-of-sample forecasting performance of ten different models as well as their robustness with four different metrics: $\hat{R}^2_{oos}$ score, R^2_{oos} score, Model Confidence Set (MCS) test (Hansen et al. (2011)), and Diebold-Mariano (DM) test (Diebold and Mariano, 2002). We consider the period from April 2002 to April 2024 to include the Great Recession and COVID-19.

The statistic $\hat{R}^2_{oos}$ is employed to compare the forecasting residual with the forecasting residual of the benchmark model, designated as HAR. The application of machine learning models incorporating HAR factors, such as random forest regression (RFR) or ElasticNet, has the potential to achieve notably high values of R^2_{oos} in out-of-sample. We found that the longer forecast horizon tends to yield more accurate forecasts of models, which may be attributed to the presence of noise due to data frequency. Additionally, the performance orders can vary across different forecast horizons. The performance of conventional HAR and HAR-X models illustrates the distinct importance of lagged variables for different horizons. Exposure to varying levels of autocorrelation effects may contribute to the observed performance order differences across different horizons.

While $\hat{R}^2_{oos}$ is based on the benchmark model, R^2_{oos} compares the forecasting residual with the variance of true values in out-of-sample (Guo and Lin (2020)). It measures the predictability of the model by residuals between predicted values and true values. With R^2_{oos}, we can compare the forecasting performance of different lengths of the test period. In our study, the performance of the conventional HAR model decreases as the period becomes more recent. This might suggest that the importance of factors other than lagged variables, that is, memory or momentum variables, has increased recently. In contrast, some machine learning models such as the random forest regression and the recurrent neural network model, show very stable R^2_{oos} scores even during the period of COVID-19.

In addition, we construct a rich set of data including implied volatility indices and

uncertainty factors, following Degiannakis and Filis (2022), Delis et al. (2022), Delis et al. (2023), Miao et al. (2017), Wei et al. (2017), and Ma et al. (2018).[2] As expected in Ma et al. (2018), appending uncertainty factors help models improve the out-of-sample performance and its robustness. Furthermore, the number of selected features plays a crucial role in performance. Linear models like HAR-X, LASSO, and ElasticNet tend to perform better as more features are included. In contrast, nonlinear machine learning models such as RFR and GBR achieved high out-of-sample scores with just 20 features, suggesting that too many features could negatively impact their performance. Deep learning models, including ANN and RNN, show more variable results. This might be due to the nonlinear relationship between realized volatility of crude oil and uncertainty indices.

Our study is distinct from other research in some significant respects. First, we incorporated uncertainty factors, which permitted us to examine a more expansive spectrum of variables that influence oil volatility. Secondly, we employed feature selection techniques with diverse machine learning models to examine the extensive range of models and the vast number of features. Moreover, our models are all interpretable, enabling us to ascertain which feature plays a pivotal role in predicting oil prices. Finally, our methods produced favorable outcomes, demonstrating the efficacy of our approach. These concepts offer invaluable insights and a more efficient research methodology than previous studies.

The remaining sections of the paper are organized as follows. In section 2, we illustrate input data and data preprocessing procedure. In section 3, we explain the training procedure of various forecasting models. In section 4, we present the out-of-sample $\hat{R}^2_{oos}$ score, R^2_{oos} score, MCS test, and DM test, as the forecasting performance measures, and analyze the results of the forecasting models. We also present which feature is chosen in the models frequently.

Factor Group	Individual Variable	Frequency	Source
Prices	WTI (West Texas Intermediate) future prices	Daily	Energy Information Administration
	WTI spot prices	Daily	Energy Information Administration
	Brent oil spot prices	Daily	Energy Information Administration
	NGL (Natural Gas Liquids) future prices	Daily	Energy Information Administration
	NGL spot prices	Daily	Energy Information Administration
Supply Factors	Global crude oil production	Monthly	JODI-Oil Database
	Global crude oil stock	Monthly	JODI-Oil Database
	Global crude oil export	Monthly	JODI-Oil Database
	Total OPEC production capacity	Monthly	Energy Information Administration
	Capacity utilization rate	Weekly	Energy Information Administration
Demand Factors	Global crude oil import	Monthly	JODI-Oil Database
	Liquid fuels consumption in World	Monthly	Energy Information Administration
	PPI in China	Monthly	Federal Reserve Bank of St. Louis Economic Database
	PPI in US	Monthly	Federal Reserve Bank of St. Louis Economic Database
	PPI in EU	Monthly	Federal Reserve Bank of St. Louis Economic Database
Financial Factors	S&P 500 Adjusted Close	Daily	Yahoo Finance
	Japan / US Foreign Exchange Rate	Daily	Federal Reserve Bank of St. Louis Economic Database
	US / Euro Foreign Exchange Rate	Daily	Federal Reserve Bank of St. Louis Economic Database
	US / UK Foreign Exchange Rate	Daily	Federal Reserve Bank of St. Louis Economic Database
	China / US Foreign Exchange Rate	Daily	Federal Reserve Bank of St. Louis Economic Database
	Federal Funds Rate	Monthly	Federal Reserve Bank of St. Louis Economic Database
	MSCI World Standard (Large+Mid Cap)	Monthly	MSCI
Implied Volatility Indices	CBOE Volatility Index	Daily	Yahoo Finance
	CBOE Crude Oil Volatility Index	Daily	Yahoo Finance
	CBOE DJIA Volatility Index	Daily	Yahoo Finance
	CBOE Gold Volatility Index	Daily	Yahoo Finance
Uncertainty Factors	Global Economic Policy Uncertainty (current)	Monthly	Davis (2016), `https://www.policyuncertainty.com`
	Global Economic Policy Uncertainty (ppp)	Monthly	Davis (2016), `https://www.policyuncertainty.com`
	Daily Infectious Disease Equity Market Volatility Tracker	Daily	Baker et al. (2019) and Baker et al. (2020) `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Economic Policy Uncertainty	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Monetary Policy	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Fiscal Policy (Taxes or spending)	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Taxes	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Government Spending	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Healthcare	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in National Security	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Entitlement Programs	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Regulation	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Financial Regulation	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Trade Policy	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	US Economic Policy Uncertainty in Sovereign Debt and currency crises	Monthly	Baker et al. (2016), `https://www.policyuncertainty.com`
	World Uncertainty Index	Quarterly	Ahir et al. (2018), `https://worlduncertaintyindex.com/`
	World Pandemic Uncertainty Index	Quarterly	Ahir et al. (2018), `https://worlduncertaintyindex.com/`

Table 1. There are six groups of data - prices, supply factors, demand factors, financial factors, implied volatility indices, and uncertainty factors. The dependent variable is WTI future price, and the rest of the data are explanatory variables used in the models.

2. Data

In line with Miao et al. (2017), Wei et al. (2017), and Ma et al. (2018), we construct the realized volatility of oil future contracts on the West Texas Intermediate (WTI) crude oil as follows:

$$RV_t = \sqrt{\sum_{j=1}^{N_t} r_{t,j}^2}, \quad \text{for } r_{t,j} = 100 \times log(p_{t,j}/p_{t,j-1}),$$

where N_t indicates the number of business days in the t-th week, and $p_{t,j}$ represents the daily WTI future prices on j-th business day of the t-th week. We consider the period from April 2002 to April 2024 to include the Great Recession and COVID-19. Because each future contract must be exercised at maturity, to estimate continuous volatility, daily prices are calculated based on the roll-over rule (Schwager, 2017).[3] The set of

[2]Kilian and Hicks (2013) argued that repeated shocks of oil demand cause the oil price shock. Other plentiful researches also assert interdependence between the oil price uncertainty and the structural fluctuation of global economic activity (Herrera et al., 2019; and Caggiano et al., 2020).

[3]We follow 'Definitions, Sources, and Explanatory Notes' in the EIA webpage (`https://www.eia.gov/dnav/pet/TblDefs/pet_pri_fut_tbldef2.asp`) to calculate continuous WTI future prices. On each monthly roll-over

uncertainty indices is more closely associated with macroeconomic trends than with the specific behavior of the oil market. Consequently, a sudden change in these indices will require a period of adjustment in the oil market. We hypothesize that uncertainty indices do not exert an immediate (daily) influence on oil prices, and thus, we anticipate weekly price movements.

A total of 42 explanatory variables were employed in the analysis (see Table 1). Each explanatory variable is classified into one of six categories: prices, supply factors, demand factors, financial factors, implied volatility indices, and uncertainty factors. We apply a difference transformation to the data that does not satisfy stationarity criteria over the entire period. For higher-order differences (greater or equal to 3), we set the order of difference to 2 to preserve information. Additionally, we consider the publication lags of the explanatory variables to mitigate potential look-ahead bias in the forecast.

Prices West Texas Intermediate (WTI) has a long-run equilibrium relationship with Brent (Hammoudeh et al., 2008), and close interconnection with natural gas (Brown and Yucel, 2008). We adopted the future and spot prices of WTI (only spot prices), Brent, and natural gas liquid (NGL) from the US Energy Information Administration (EIA), as explanatory variables.

Supply Global crude oil data, including production, stock, and export, are from world-primary data in the JODI-Oil Database. Oil supply and demand are accepted factors to describe the oil price dynamics (Kilian, 2009; Ma et al., 2018; Wei et al., 2017). Hallock Jr et al. (2004) showed a strong relationship between oil production and exportation. Hamilton (2009) asserted the importance of oil inventory to prices. Total OPEC production capacity and Capacity utilization rate are from EIA.

Demand Demand factors have a significant influence on oil prices (Hamilton, 2009; and Kilian, 2009) The demand factors consist of the global crude oil import from the JODI-Oil Database, liquid fuel consumption in the World from the EIA, and the producer Price Index (PPI) in China, the US, and the EU from the Federal Reserve Bank (FRB) of St. Louis Economic Database.

---

day, the oil price is adjusted by the price difference between the future contract with the first-nearest-to-maturity and the future contract with the second-nearest-to-maturity. The weekly realized volatility is calculated by the squared sum of daily WTI returns on each last business day of the week. Each contract expires on the third business day before the 25th calendar day. When the 25th calendar day is not a business day, the contract expires on the third business day before the latest business day prior to the 25th calendar day. We excluded the date when the oil price became negative. More details about the roll-over rule are in Appendix A.

Financial Factors Following Miao et al. (2017), financial factors include the federal fund rate, the daily exchange rate of Japan-US, US-EU, US-UK, and China-US from the FRB, the S&P 500 Adjusted Close data from Yahoo Finance and MSCI World Standard (Large+Mid Cap) from MSCI.

Implied Volatility Indices As showed in the Delis et al. (2022) and Delis et al. (2023), the implied volatility indices have potential explanatory power in forecasting the oil implied volatility index. In this regard, four implied volatility indices are included: the CBOE Volatility Index (VIX), the CBOE Crude Oil Volatility Index (OVX), the CBOE DJIA Volatility Index (VXD), and the CBOE Gold Volatility Index (GVZ). The data were collected from Yahoo Finance.

Uncertainty Factors Ma et al. (2018) also considers uncertainty indices when handling the oil. Following Ma et al. (2018), Baker et al. (2016), Baker et al. (2019), Baker et al. (2020), Ahir et al. (2018), and Davis (2016) with their websites, we use various uncertainty indices as an input.[4]

3. Forecasting Models

3.1. *Cross Validation*

In the time-series model, preventing the use of future information is a pertinent first step. Similar to Gu et al. (2020), we construct the out-of-sample data and design cross-validation as shown in Fig 1. For each time, we stack s historical data, each of which has length-w training data. With $s \times w$ training input and corresponding $s \times 1$ output, standard cross-validation is performed to find the best hyperparameters.

Now one-step ahead length-w data (test data or the out-of-sample data) is chosen to be a test input. With the in-sample trained model and the one-step-ahead length-w input, the model predicts one-week volatility. After prediction, we reset the model and move our s set of historical data forward, and follow the same procedure. This process guarantees that models do not use future information as input. The number of samples

[4]Economic uncertainty data were retrieved from `https://www.policyuncertainty.com/` and `https://worlduncertaintyindex.com/`. This webpage includes data from numerous papers covering economic policy uncertainty data (EPU), categorical economic policy uncertainty data, world uncertainty index (WUI) to even infectious disease data.

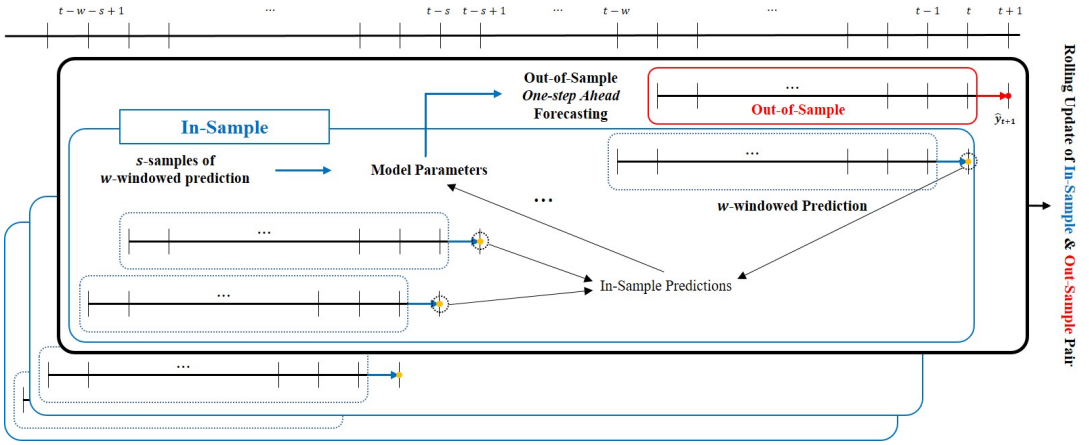

Figure 1. Walk-Forward Cross-Validation A single forecast on dependent variable y_{t+1} in time $t+1$ is based on explanatory variables in the time interval $[t-w+1, t]$. Regardless of the forecasting model choice, this single forecast needs parameters for each explanatory variable in the w-windowed time interval. The estimation of the corresponding parameters for each explanatory variable uses s-sample points by minimizing the objective function, e.g., the sum of the least square, of in-sample residuals. When w is large, a single forecast considers more past information. A large number of samples s would stabilize parameter estimation, but it would catch up with recent changes slowly.

(s) and windows (w) is chosen to contain enough information from the past but not overfit to the training data (Fig 1).[5]

3.2. *Forecasting Models*

Machine learning models are becoming more prevalent in the field of forecasting (Gu et al., 2020; Ghoddusi et al., 2019). We use both time series models - the heterogeneous autoregressive with exogenous variables (HAR-X), the time-varying parameter heterogeneous autoregressive (TV-HAR) - and various machine learning models.[6] The HAR-X model from Corsi (2009) is a benchmark. The reader can find the analysis on various machine learning models - the least absolute shrinkage and selection operator (LASSO), the elastic net (EN), the decision tree regression (DTR), the random forest regression (RFR), the gradient boosting regression (GBR), the artificial neural network (ANN), and the recurrent neural network (RNN). Detailed descriptions of how we train models are in Appendix B.

[5] In this study, all models use $w = 1, s = 52$.

[6] The reader should beware that depending on the operating system, version of packages, type of CPUs, and other factors may affect the results. We have tested in different environments, and in most cases, the differences were relatively insignificant. Additionally, alternative models, including ARIMA-X and principal component regression (PCR), were evaluated. However, the outcomes were unsatisfactory, and thus, they have been excluded from this paper.

Heterogeneous Autoregressive with Exogenous Variables

Corsi (2009) found the stylized facts about the long memory and multi-scaling behavior of the realized volatility. The HAR-X contains two explanatory variables: weekly and monthly realized volatility. These variables represent short-term and medium-term, respectively.[7] The HAR-X is defined by

$$\begin{aligned}
\widehat{RV}^{W}_{t+1} &= c_W + \beta_{W,w} RV^{W}_{t} + \beta_{W,m} RV^{W}_{t-4:t} + \beta_{W,q} RV^{W}_{t-13:t} \\
&\quad + \sum_{k=0}^{w-1} \sum_{m=1}^{n} \eta_{W,m,t-k} x_{m,t-k} + \epsilon_{W,t+1} \\
\widehat{RV}^{2W}_{t+1} &= c_{2W} + \beta_{2W,w} RV^{2W}_{t} + \beta_{2W,q} RV^{2W}_{t-6:t} + \beta_{2W,y} RV^{2W}_{t-26:t} \\
&\quad + \sum_{k=0}^{n-1} \sum_{m=1}^{n} \eta_{2W,m,t-k} x_{m,t-k} + \epsilon_{2W,t+1} \\
\widehat{RV}^{M}_{t+1} &= c_M + \beta_{M,W} RV^{M}_{t} + \beta_{M,q} RV^{M}_{t-3:t} + \beta_{M,y} RV^{M}_{t-12:t} \\
&\quad + \sum_{k=0}^{n} \eta^{m,t-k}_{m=1} x_{m,t-k} + \epsilon_{M,t+1}
\end{aligned}$$

where RV^{W}_{t}, RV^{2W}_{t}, and RV^{M}_{t} are the weekly, bi-weekly, and monthly realized volatility based on daily return, respectively. For the weekly realized volatility RV^{W}_{t}, the monthly average realized volatility $RV^{W}_{t-4:t}$ and the quarterly realized volatility $RV^{W}_{t-13:t}$, which are the average RV from $t-4$ to t and $t-13$ to t respectively, represent the short-term variable and long-term variable. Every model below also contains realized volatility factors as explanatory variables. The parameter estimation process is the same as the linear regression.[8]

Time-Varying Parameter Heterogeneous Autoregressive

The time-varying parameter heterogeneous autoregressive (TV-HAR) is an extension of the HAR, considering the parameter changes over time. This method is commonly used to enhance the forecasting power of the HAR-type models (Delis et al. (2022)). In

[7] The original HAR contains daily realized volatility to forecast daily realized volatility. However, our goal in this paper is to forecast the weekly realized volatility. So, we exclude the daily realized volatility factor.

[8] Therefore, machine learning models are the extended version of the HAR, in the aspect of variable usage. When the machine learning models do not consider realized volatility factors as explanatory variables, they perform poorly.

TV-HAR, the coefficients in the equation above change to time-varying coefficients as follows:

$$\begin{aligned}
\widehat{RV}^{W}_{t+1} &= c_W + \beta_{W,w,t} RV^{W}_{t} + \beta_{W,m,t} RV^{W}_{t-4:t} + \beta_{W,q,t} RV^{W}_{t-13:t} \\
&\quad + \sum_{k=0}^{w-1} \sum_{m=1}^{n} \eta_{W,m,t-k} x_{m,t-k} + \epsilon_{W,t+1} \\
\widehat{RV}^{2W}_{t+1} &= c_{2W} + \beta_{2W,w,t} RV^{2W}_{t} + \beta_{2W,q,t} RV^{2W}_{t-6:t} + \beta_{2W,y,t} RV^{2W}_{t-26:t} \\
&\quad + \sum_{k=0}^{n-1} \sum_{m=1}^{n} \eta_{2W,m,t-k} x_{m,t-k} + \epsilon_{2W,t+1} \\
\widehat{RV}^{M}_{t+1} &= c_M + \beta_{M,W,t} RV^{M}_{t} + \beta_{M,q,t} RV^{M}_{t-3:t} + \beta_{M,y,t} RV^{M}_{t-12:t} \\
&\quad + \sum_{k=0}^{n} \eta^{m,t-k}_{m=1} x_{m,t-k} + \epsilon_{M,t+1}
\end{aligned}$$

The time-varying coefficients can be modeled using various methodologies, such as the Markov Chain Monte Carlo (MCMC) method, the Kalman filter method, and the kernel smoothing method. In this research, we apply the kernel smoothing method to estimate the parameters.

Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) reduces the risk of over-fitting by adding a penalty term to the cost function, and it performs variable selection:

$$\begin{aligned}
&\min_{\phi,\eta} \left\{ \left(\mathrm{RV}_{t+1} - \Big(\sum_{i=0}^{w-1} \phi_i \mathrm{RV}_{t-i} + \sum_{k=0}^{w-1} \sum_{m=1}^{n} \eta_{m,t-k} x_{m,t-k} \Big) \right)^2 \right\}, \\
&\text{subject to } \sum |\phi| + \sum |\eta| < c \text{ (constant)}.
\end{aligned}$$

It also has a useful application in the field of energy forecasting, such as research on the oil price predictability in Miao et al. (2017) and Ma et al. (2018).

Elastic Net

Elastic Net is the combinatorial extension of the Ridge and the Lasso. Although the Ridge and the Lasso enhance the prediction accuracy with the bias-variance tradeoff, they have different performance advantages depending on the data. The Lasso is better when there are fewer significant variables, and the Ridge is better when there are vice versa. The Elastic Net considers both penalty terms of the Ridge and the Lasso. Because Elastic Net has both $L1$ and $L2$ norms, it can perform well with large data.

Decision Tree Regression

Decision Tree Regression (DTR) builds models in the form of a tree structure by dividing the data set into smaller subsets while gradually elaborating related decision trees. The predicted outcome becomes continuous real values.

Hastie et al. (2009) formally describes this concept as follows: when X is an input variable and defining half-planes as $R_1(j,s) = \{X|X_j \leq s\}$ and $R_2(j,s) = \{X|X_j > s\}$, DTR tries to find the splitting variable j and split point s that solve

$$\min_{j,s} \Big[\min_{c_1} \Sigma_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \Sigma_{x_i \in R_2(j,s)} (y_i - c_2)^2 \Big].$$

where y_i's are true values. After finding the best split, repeat this process for each split region. They further discuss some important questions like how large a should tree grow. See Hastie et al. (2009) for the full detailed explanation.

Random Forest Regression

Random Forest Regression (RFR) is the learning algorithm that uses an ensemble learning method for several decision trees. Even if trees can make an excellent predictive model, they can be very noisy. RFR generates multiple decision trees and averages them to reduce noise and the risk of overfitting. See Hastie et al. (2009) for detailed descriptions.

Gradient Boosting Regression

Gradient Boosting Regression (GBR) produces a model from an ensemble of weak predictive models. One obvious candidate for the weak predictor will be the tree. The

gradient boosting algorithm updates its predictor by calculating the negative gradient of the loss criteria, then regress a tree to those residuals. A detailed explanation for the algorithm is in Hastie et al. (2009).

Artificial Neural Network

Artificial Neural Network (ANN) is a computational model inspired by the human brain, designed to recognize patterns and make decisions. It consists of interconnected layers of nodes (neurons), including an input layer, one or more hidden layers, and an output layer. Each connection has a weight that is adjusted during training, allowing the network to learn from data. See Goodfellow (2016) for more details.

Recurrent Neural Network

Recurrent Neural Network (RNN) is a specific type of artificial neural network designed to process sequential data. Unlike feedforward networks, RNNs have connections that form directed cycles, allowing them to retain information from earlier inputs in the sequence. This makes RNNs particularly effective for tasks like time series prediction. For more details, see Goodfellow (2016).

Using all input data to train models often raises several caveats. The most prominent issue will be training time. As the number of explanatory variables rises, models suffer from a surge in training time. Also, unnecessary variables could hurt the out-of-sample performance. To prevent this, we select features for each sample window before we start training the model.[9] During training, we use the grid search method to find optimal hyperparameters.

For feature selection, we choose the f-regression method. It first computes the correlation between the dependent variable and each factor. Then it is converted to F-score, and p-value of F-score (Pedregosa et al. (2011)). The linearity of the f-regression method can quickly calculate the effect of each factor on dependent variables, which is adequate in this paper that uses multiple factors. After calculating F-scores and p-values, we

[9]Observing the high performance of the HAR-X model, we make an educated guess that the lagged volatility and mean of volatility have high explanatory power. These variables are exempt from the feature selection stage and always belong to the set of explanatory variables.

choose a defined number of factors from the highest score. We always include lagged variables as explanatory variables even if the result of the f-regression method rejects their efficiency to make the machine learning models more coherent with HAR-X.

4. Empirical Analysis

4.1. *The Measure of Evaluation and Robustness*

Although the R^2 measure is more widely accepted in linear models, it is one of the most intuitive measures for evaluating the performance of a model. In this paper, we examine two R^2 measures: $\hat{R}^2_{oos}$ score and our version of the out-of-sample R^2_{oos} score. Let y_t, $\bar{y}_t$, $\hat{y}_t$ and $\hat{y}_{Bench,t}$ denote true values, the average of true values, predicted values, and the predicted values of the benchmark model HAR, respectively. First, conventionally, R^2_{oos} score is defined as:

$$R^2_{oos} = 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y}_t)^2}$$

Numerous literature argue that $\bar{y}_t$ should be replaced by the benchmark forecast, $\hat{y}_{Bench,t}$ (e.g., Christiansen et al. (2012), Wang et al. (2018)). Although the historical average is powerful for forecasting asset returns, it fails to predict asset return volatility. However, the benchmark model HAR has shown a slight decline in its forecasting performance compared to other models recently. So, we introduce a more strict performance measure, $\hat{R}^2_{oos}$ score:

$$\hat{R}^2_{oos} = 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \hat{y}_{Bench,t})^2}$$

Second, we define $\bar{y}_t$ in $\hat{R}^2_{oos}$ score as the average value of future true values. Note that $\bar{y}_t$ is not just the historical value, but the average value of our future target variables. Predicted values are stacked outputs from rolling out-of-sample prediction described in Section 3. Therefore, R^2_{oos} can compare the real-time forecasting performances of various models. Furthermore, the subperiod R^2_{oos} score is calculated to check the robustness. R^2_{oos} in the recent 10, 5, 3, 2, and 1-year intervals are measured ($R^2_{oos,t}$, $t \in \{10, 5, 3, 2, 1\}$,

respectively). Furthermore, the stationarity of the residual $\hat{y}_t - y_t$ is guaranteed in most cases. It makes our model economically meaningful and safe to use performance metrics.[10] As additional metrics, we practice the Model Confidence Set (MCS) introduced by Hansen et al. (2011). For each time horizon - weekly, bi-weekly, and monthly - a set of forecasting models is constructed with a given level of confidence of 0.05. The MCS shows that the best accurate models are different for each time horizon.

In order to test the statistical significance, the Diebold-Mariano (DM) test is employed.[11] The results indicate that the majority of high-performing models exhibit statistically significant differences from the benchmark. We employ the stationarity of the residuals as a proxy for the statistical significance of machine learning models, as these models are unable to provide statistical significance in the form of $F-$ or $p-$ values. The majority of the models satisfy the 5% criterion, however, as the periods increase, the value tends to increase (although in most cases, it remains below 5%).

[10] We also tested the predictability of residuals $\hat{y}_t - y_t$ using our models. However, none of the models could predict the one-step-ahead residual with a positive R^2_{oos}.
[11] More details about DM test is in Appendix C

Forecasting Model	Number of Feature Selection	$\hat{R}^2_{oos}$					
		With Uncertainty			Without Uncertainty		
		M	2W	W	M	2W	W
HAR	-	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)
TV-HAR	-	-0.0203 (2.079E-13)	0.2508 (2.648E-19)	0.2392* (1.421E-19)	-0.0203 (2.079E-13)	0.2508 (2.648E-19)	0.2392* (1.421E-19)
HAR-X	0	-0.0203 (2.079E-13)	0.2508 (2.648E-19)	0.2392* (1.421E-19)	-0.0203 (2.079E-13)	0.2508 (2.648E-19)	0.2392* (1.421E-19)
	10	0.0958 (4.276E-11)	0.3445 (8.056E-16)	0.2528* (1.967E-17)	-0.2564 (2.145E-07)	0.2779 (1.239E-07)	0.1377* (3.584E-14)
	20	0.1260 (4.334E-12)	0.3085 (2.286E-15)	0.2505* (5.876E-17)	0.0739 (1.379E-15)	0.2377 (3.635E-17)	0.0311 (4.732E-15)
	all	**0.3708*** (4.671E-11)	**0.4154*** (3.443E-13)	**0.3236*** (3.356E-12)	0.0125 (6.184E-17)	0.2106 (1.366E-16)	-1.2143 (6.322E-06)
LASSO	0	0.0599 (5.319E-15)	0.3018 (4.384E-20)	0.2756* (1.552E-16)	0.0599 (5.319E-15)	0.3018* (4.384E-20)	0.2756* (1.552E-16)
	10	0.1964 (1.983E-14)	0.3622* (1.719E-19)	0.2766* (1.717E-16)	0.0452 (1.151E-16)	0.2335 (2.426E-15)	0.2605* (2.258E-17)
	20	0.2110 (1.528E-14)	0.3623* (1.723E-19)	0.2766* (1.718E-16)	0.0550 (6.858E-17)	0.2458 (3.340E-15)	0.2603* (2.308E-17)
	all	**0.4007** (2.232E-17)	**0.4512*** (8.323E-14)	**0.3547*** (5.456E-14)	0.2554* (7.102E-17)	0.3980* (5.951E-16)	0.2439* (1.809E-14)
ElasticNet	0	0.0613 (6.682E-15)	0.2985 (3.345E-20)	0.2698* (1.053E-17)	0.0613 (6.682E-15)	0.2985* (3.345E-20)	0.2698* (1.053E-17)
	10	0.2137 (2.232E-14)	0.3849* (4.921E-19)	0.2969* (2.775E-17)	0.1537* (3.724E-15)	0.3245* (3.286E-14)	0.2714* (3.825E-18)
	20	0.2237 (2.426E-14)	0.3807* (3.516E-19)	0.2971* (3.317E-17)	0.2023* (1.340E-20)	0.3319* (1.301E-20)	0.2726* (5.200E-18)
	all	**0.4043*** (3.434E-17)	**0.4397*** (3.443E-14)	**0.3533*** (5.409E-14)	0.2396* (6.246E-17)	0.3884* (8.830E-17)	0.2007* (1.019E-13)
DTR	0	0.0513 (1.209E-12)	-0.0220 (7.737E-08)	-0.2066 (1.258E-17)	0.0240 (2.446E-12)	-0.1064 (2.973E-08)	-0.1085 (7.020E-16)
	10	0.3294* (4.235E-14)	**0.3065** (3.638E-12)	-0.2111 (7.734E-15)	0.2264* (7.963E-16)	0.0022 (4.163E-15)	-0.2102 (1.461E-19)
	20	**0.3452*** (2.868E-15)	0.2293 (5.137E-13)	-0.2178 (4.559E-18)	0.2755* (3.774E-14)	-0.0087 (1.436E-13)	**0.0697** (1.118E-12)
	all	0.3308* (1.535E-15)	0.1281 (1.858E-11)	-0.2805 (7.418E-19)	0.0503 (5.726E-08)	0.1752 (2.422E-18)	-0.0473 (3.920E-13)
RFR	0	0.2827* (1.293E-12)	0.3737* (2.818E-16)	0.2297 (1.783E-17)	0.2823* (1.240E-12)	0.3720* (2.061E-16)	0.2428* (3.330E-17)
	10	0.4623* (3.638E-16)	**0.4884*** (1.350E-15)	0.3434* (3.149E-16)	0.4228* (1.145E-13)	0.4060* (2.940E-18)	0.2760* (2.131E-15)
	20	**0.4909*** (1.028E-15)	0.4678* (2.964E-17)	0.3743* (2.684E-17)	0.4216* (1.606E-13)	0.4604* (3.747E-17)	0.3352* (2.574E-16)
	all	0.4843* (3.575E-15)	0.4668* (2.657E-17)	**0.3751*** (1.246E-17)	0.4175* (8.148E-14)	0.4470* (3.617E-18)	0.3328* (9.301E-17)
GBR	0	0.2160 (5.184E-13)	0.3269 (3.493E-17)	0.2092 (6.703E-16)	0.2296* (6.920E-13)	0.3205* (2.413E-17)	0.2233* (3.080E-16)
	10	0.4292* (1.710E-15)	**0.4891*** (1.786E-13)	0.3221* (1.266E-15)	0.3102* (1.699E-06)	0.3305* (1.026E-16)	0.0601 (2.916E-19)
	20	**0.4783*** (2.345E-15)	0.4644* (1.528E-16)	**0.3601*** (3.625E-16)	0.3382* (2.420E-16)	0.3751* (4.168E-18)	0.2670* (2.752E-19)
	all	0.4641* (1.117E-15)	0.4709* (1.059E-16)	0.3591* (1.366E-16)	0.3575* (3.946E-17)	0.3919* (8.140E-19)	0.2863* (2.269E-17)
ANN	0	-0.0483 (1.238E-13)	0.2008 (8.166E-20)	0.2090 (6.319E-18)	-0.0554 (4.572E-14)	0.2214 (1.274E-19)	0.1738* (4.084E-18)
	10	0.2757* (9.829E-11)	0.4146* (5.968E-13)	**0.3610*** (3.832E-13)	0.1210 (4.540E-14)	0.4633* (1.464E-13)	0.2837* (8.791E-13)
	20	0.2485 (2.381E-12)	0.3708* (2.239E-15)	0.2801* (4.347E-14)	0.2626* (1.589E-13)	**0.5103*** (1.832E-12)	0.0482 (4.429E-13)
	all	**0.2923*** (5.360E-12)	0.3285 (4.946E-16)	0.2618* (1.763E-13)	-2.2483 (2.300E-18)	-0.0852 (3.332E-14)	-0.7148 (1.649E-11)
RNN	0	0.4398* (5.641E-10)	0.3868* (2.328E-15)	0.3121* (8.896E-15)	0.4372* (6.683E-10)	0.3878* (8.813E-16)	0.3132* (1.264E-14)
	10	0.4711* (1.647E-11)	**0.4777*** (1.101E-13)	0.3956* (7.267E-14)	0.4800* (5.836E-11)	0.4696* (1.699E-12)	**0.4070*** (6.483E-15)
	20	**0.4906*** (4.541E-10)	0.4549* (3.328E-15)	0.3841* (1.998E-15)	0.4803* (3.066E-10)	0.4457* (2.522E-07)	0.3488* (3.794E-14)
	all	0.4813* (6.191E-10)	0.4316* (2.055E-13)	0.3347* (4.252E-15)	0.4742* (8.149E-10)	0.4135* (5.288E-14)	0.3280* (4.572E-15)

Table 2. Forecasting Performance Compared to the Benchmark HAR. $\hat{R}^2_{oos}$ is the performance measurement compared to the benchmark HAR. For each model and each forecasting horizon, bold numbers are the best performance measurements with feature selection and data set. For each time horizon, models that pass the MCS test with a significant level of 0.05 have an asterisk with each performance measure. The numbers in parentheses indicate the p-values of residual stationarity.

Forecasting Model	Number of Feature Selection	DM Test					
		With Uncertainty			Without Uncertainty		
		M	2W	W	M	2W	W
HAR	-	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)	0.0000 (0.000E+00)
TV-HAR	-	-0.0453 (9.639E-01)	1.0316 (3.023E-01)	1.3210 (1.865E-01)	-0.0453 (9.639E-01)	1.0316 (3.023E-01)	1.3210 (1.865E-01)
HAR-X	0	-0.0453 (9.639E-01)	1.0316 (3.023E-01)	1.3210 (1.865E-01)	-0.0453 (9.639E-01)	1.0316 (3.023E-01)	1.3210 (1.865E-01)
	10	0.2264 (8.209E-01)	1.1966 (2.315E-01)	1.2413 (2.145E-01)	-0.4538 (6.500E-01)	0.8356 (4.034E-01)	0.6654 (5.058E-01)
	20	0.2949 (7.680E-01)	1.1187 (2.633E-01)	1.1772 (2.391E-01)	0.1583 (8.742E-01)	0.7380 (4.605E-01)	0.1904 (8.490E-01)
	all	**0.7735** (4.392E-01)	**1.3406** (1.801E-01)	**1.6385** (1.013E-01)	0.0251 (9.800E-01)	0.6020 (5.471E-01)	-2.1993 (2.786E-02)
LASSO	0	0.1370 (8.910E-01)	1.2562 (2.091E-01)	1.5992 (1.098E-01)	0.1370 (8.910E-01)	1.2562 (2.091E-01)	1.5992 (1.098E-01)
	10	0.4524 (6.510E-01)	1.4092 (1.588E-01)	1.6055 (1.084E-01)	0.0950 (9.243E-01)	0.9466 (3.439E-01)	1.5345 (1.249E-01)
	20	0.4886 (6.251E-01)	1.4095 (1.587E-01)	1.6054 (1.084E-01)	0.1163 (9.074E-01)	1.0021 (3.163E-01)	1.5334 (1.252E-01)
	all	**0.8641** (3.875E-01)	**1.5312** (1.257E-01)	**1.9618** (4.979E-02)	0.5400 (5.892E-01)	1.3853 (1.659E-01)	1.3523 (1.763E-01)
ElasticNet	0	0.1396 (8.890E-01)	1.2414 (2.145E-01)	1.5810 (1.139E-01)	0.1396 (8.890E-01)	1.2414 (2.145E-01)	1.5810 (1.139E-01)
	10	0.4955 (6.203E-01)	1.4632 (1.434E-01)	1.7104 (8.719E-02)	0.3447 (7.303E-01)	1.2239 (2.210E-01)	1.5775 (1.147E-01)
	20	0.5184 (6.042E-01)	1.4592 (1.445E-01)	1.7080 (8.764E-02)	0.4584 (6.466E-01)	1.2538 (2.099E-01)	1.5816 (1.137E-01)
	all	**0.8611** (3.892E-01)	**1.5033** (1.328E-01)	**1.9617** (4.980E-02)	0.5064 (6.126E-01)	1.3055 (1.917E-01)	1.0979 (2.723E-01)
DTR	0	0.1157 (9.079E-01)	-0.0534 (9.574E-01)	-0.7162 (4.739E-01)	0.0542 (9.568E-01)	-0.2489 (8.034E-01)	-0.4847 (6.279E-01)
	10	0.6236 (5.329E-01)	**0.9530** (3.406E-01)	-0.6466 (5.179E-01)	0.4396 (6.602E-01)	0.0071 (9.944E-01)	-0.5948 (5.520E-01)
	20	**0.6709** (5.023E-01)	0.7310 (4.648E-01)	-0.6852 (4.932E-01)	0.5042 (6.141E-01)	-0.0303 (9.759E-01)	**0.3664** (7.141E-01)
	all	0.6438 (5.197E-01)	0.3978 (6.908E-01)	-0.8483 (3.963E-01)	0.0800 (9.362E-01)	0.5657 (5.716E-01)	-0.2662 (7.901E-01)
RFR	0	0.5910 (5.545E-01)	1.2729 (2.031E-01)	1.2499 (2.113E-01)	0.5902 (5.551E-01)	1.2656 (2.057E-01)	1.3154 (1.884E-01)
	10	0.9039 (3.660E-01)	**1.6849** (9.201E-02)	1.9719 (4.862E-02)	0.8215 (4.113E-01)	1.5325 (1.254E-01)	1.4702 (1.415E-01)
	20	**0.9639** (3.351E-01)	1.6549 (9.795E-02)	2.0061 (4.484E-02)	0.8104 (4.177E-01)	1.6460 (9.976E-02)	1.7020 (8.875E-02)
	all	0.9434 (3.455E-01)	1.6520 (9.854E-02)	**2.0083** (4.461E-02)	0.8075 (4.194E-01)	1.6108 (1.072E-01)	1.7388 (8.207E-02)
GBR	0	0.4689 (6.391E-01)	1.0528 (2.924E-01)	1.1469 (2.514E-01)	0.4946 (6.209E-01)	1.0299 (3.031E-01)	1.2093 (2.265E-01)
	10	0.8215 (4.114E-01)	**1.6461** (9.975E-02)	1.8191 (6.889E-02)	0.5836 (5.595E-01)	1.2225 (2.215E-01)	0.2408 (8.097E-01)
	20	**0.9446** (3.449E-01)	1.5738 (1.155E-01)	**1.8751** (6.078E-02)	0.6352 (5.253E-01)	1.4954 (1.348E-01)	1.4490 (1.473E-01)
	all	0.9041 (3.659E-01)	1.6231 (1.046E-01)	**1.9438** (5.192E-02)	0.6768 (4.986E-01)	1.5035 (1.327E-01)	1.5062 (1.320E-01)
ANN	0	-0.1063 (9.153E-01)	0.8240 (4.100E-01)	1.2178 (2.233E-01)	-0.1211 (9.036E-01)	0.9147 (3.604E-01)	1.0322 (3.020E-01)
	10	0.6114 (5.409E-01)	1.2917 (1.965E-01)	**2.1024** (3.552E-02)	0.2777 (7.813E-01)	1.2259 (2.202E-01)	1.8050 (7.107E-02)
	20	0.5951 (5.518E-01)	1.2480 (2.120E-01)	1.3910 (1.642E-01)	0.5359 (5.920E-01)	**1.3297** (1.836E-01)	0.2697 (7.874E-01)
	all	**0.6462** (5.182E-01)	1.1663 (2.435E-01)	1.3365 (1.814E-01)	-0.9776 (3.283E-01)	-0.2081 (8.351E-01)	-1.1137 (2.654E-01)
RNN	0	0.8373 (4.024E-01)	1.3570 (1.748E-01)	1.8215 (6.852E-02)	0.8281 (4.076E-01)	1.3550 (1.754E-01)	1.8357 (6.641E-02)
	10	0.9064 (3.648E-01)	**1.6585** (9.721E-02)	**2.2714** (2.313E-02)	0.9412 (3.466E-01)	1.6314 (1.028E-01)	**2.2082** (2.723E-02)
	20	**0.9504** (3.419E-01)	1.5747 (1.153E-01)	2.2383 (2.520E-02)	0.9256 (3.547E-01)	1.5598 (1.188E-01)	2.1720 (2.985E-02)
	all	0.9364 (3.490E-01)	1.5226 (1.279E-01)	2.0474 (4.062E-02)	0.9155 (3.599E-01)	1.4484 (1.475E-01)	2.0146 (4.394E-02)

Table 3. Additional Metrics. The DM contain *p*- values for whether each model has significantly different accuracy compared to benchmark HAR. The numbers in parentheses indicate the *p*-values of residual stationarity.

4.2. *Empirical Results and Potential Explanation*

Overall Performance The out-of-sample test provides evidence that machine learning models are capable of achieving high levels of accuracy in forecasting oil volatility. This is evidenced by a notable enhancement in the adjusted R-squared (R^2) scores relative to the benchmark HAR model. The most commonly used time-series models, such as HAR-X and TV-HAR, demonstrate a less impressive performance than machine learning models, yet they also exhibit a superior performance compared to the HAR. The results inhibit that linear or simple machine learning models, including LASSO, ElasticNet, and DTR, demonstrate inconsistent performance. These models typically demonstrate superior performance when forecasting past periods, whereas their efficacy tends to diminish in current periods. Also, the number of feature selection plays an important role in performance. Nonlinear machine learning models with many small predictors such as the RFR and GBR can reach an impressive out-of-sample R^2 score.

Furthermore, the number of feature selections is a substantial determinant of performance. The performance of linear models, including HAR-X, LASSO, and ElasticNet, is enhanced as the number of selected features increases. Conversely, nonlinear machine learning models, such as RFR and GBR, demonstrated an exceptional out-of-sample R^2 score with 20 features. This suggests that an excess of features may negatively impact the performance of the models. Deep learning models, including ANN and RNN, yield varying results. Given the sensitivity of deep learning models to hyperparameters and their tendency to require more training, the outcomes of deep learning models may be inconsistent and subject to change over time. However, the performance of RNN is one of the most notable models in recent times.

As described in section 4.1, we provide a wide range of additional analysis and robustness checks. The MCS test and the Diebold-Mariano test verifies which models comparatively outperform the benchmark and other models. We also provide the robustness checks of the forecast horizon, the number of feature selections, the sub-period length of performance measurement, and an additional set of uncertainty indices.

The MCS test in Table 2 shows that the best-performing models vary over time horizons and a set of input data. For the monthly forecast, the random forest and RNN models work well in the out-of-sample. However, for the weekly forecast, some

of the models even underperform the TV-HAR, which implies the strong short-term momentum of the weekly realized volatility.

Table 3 provides additional metrics - the DM test. We present the statistic and p-values of the DM test with the benchmark model. Although the DM test is another comparative statistic about forecasting performance, the results can be different from those of the MCS test. This is because the MCS test is a statistical test for multiple models, excluding inferior models well, while the DM test is the one-to-one comparison with the benchmark. Most models from HAR-X to RNN show positive statistics, which means that the models are better than the benchmark HAR. Interestingly, some models such as DTR, GBR, and ANN are not differentiated statistically with HAR. One explanation could be lagged variables. The two most important factors in the HAR model are short-term and long-term lagged variables. In DTR and GBR, the lagged variables could affect DTR and GBR the most, and therefore, not statistically differentiated from the HAR model.

Forecasting Model	Number of Feature Selection	R^2_{OOS}	$R^2_{OOS,10}$	$R^2_{OOS,5}$	$R^2_{OOS,3}$	$R^2_{OOS,2}$	$R^2_{OOS,1}$
HAR	-	-0.8224	-0.8918	-1.0895	-0.1665	-0.1312	-1.0531
		(2.273E-03)	(4.362E-03)	(2.453E-01)	(9.286E-01)	(1.728E-02)	(7.422E-02)
TV-HAR	-	-0.1154	-0.1759	-0.2851	-0.1238	0.3608	0.1883
		(5.893E-21)	(4.605E-19)	(1.943E-05)	(4.493E-01)	(2.088E-01)	(3.401E-02)
HAR-X	0	-0.8224	-0.8918	-1.0895	-0.1665	-0.1312	-1.0531
		(2.273E-03)	(4.362E-03)	(2.453E-01)	(9.286E-01)	(1.728E-02)	(7.422E-02)
	10	-0.1838	-0.2409	-0.3486	-0.9011	-1.0257	-0.2328
		(5.802E-05)	(2.546E-04)	(3.216E-02)	(1.565E-02)	(3.521E-04)	(1.051E-02)
	20	0.2034	0.1826	0.1161	-1.4794	-1.6086	-0.8476
		(3.729E-07)	(1.876E-06)	(1.012E-01)	(6.567E-04)	(3.814E-03)	(1.001E-03)
	all	0.0591	0.0706	0.0697	-2.0926	-1.7339	-0.4224
		(9.227E-08)	(7.447E-07)	(7.134E-09)	(1.938E-01)	(2.477E-01)	(9.975E-01)
LASSO	0	-0.0256	-0.0788	-0.1795	-0.0518	0.3053	-0.0484
		(5.237E-11)	(6.249E-10)	(4.716E-05)	(5.148E-01)	(8.338E-01)	(7.259E-02)
	10	0.1452	0.1024	0.0301	-0.0758	0.3622	0.2709
		(2.770E-16)	(1.519E-14)	(2.777E-02)	(1.269E-10)	(1.406E-04)	(3.048E-02)
	20	0.1591	0.1170	0.0353	-0.0854	0.3333	0.2862
		(5.892E-16)	(3.173E-14)	(2.446E-05)	(5.776E-09)	(1.099E-04)	(3.157E-02)
	all	**0.4333**	**0.4293**	**0.3940**	-0.4985	-0.1757	0.2427
		(2.474E-17)	(2.084E-16)	(2.004E-09)	(2.058E-02)	(2.772E-01)	(5.396E-01)
ElasticNet	0	-0.0376	-0.0911	-0.1924	-0.0595	0.3015	-0.0452
		(1.311E-10)	(1.308E-09)	(8.332E-05)	(5.239E-01)	(7.996E-01)	(6.511E-02)
	10	0.2946	0.2612	0.2136	0.0662	0.4722	0.3081
		(1.414E-15)	(5.491E-14)	(1.169E-05)	(1.267E-01)	(5.872E-05)	(1.638E-02)
	20	0.3418	0.3113	0.2602	-0.0608	0.4135	0.4078
		(1.298E-13)	(2.513E-12)	(1.372E-06)	(3.984E-02)	(4.591E-01)	(5.725E-01)
	all	0.4092	0.4075	0.3714	-0.5334	-0.2036	0.1589
		(1.381E-16)	(9.441E-16)	(4.328E-03)	(1.420E-02)	(2.680E-01)	(6.463E-01)
DTR	0	-0.2242	-0.2852	-0.3480	-2.5173	-4.5398	-9.0861
		(5.670E-19)	(2.993E-10)	(8.087E-05)	(3.350E-06)	(9.969E-01)	(3.418E-02)
	10	0.2721	0.2786	0.3157	0.1458	-0.1037	-0.4031
		(4.367E-15)	(9.592E-14)	(6.653E-04)	(6.463E-07)	(1.000E+00)	(9.925E-01)
	20	-0.0420	-0.0536	-0.0797	0.0129	0.0686	0.1317
		(1.149E-15)	(5.663E-14)	(1.362E-09)	(2.201E-06)	(1.621E-03)	(1.744E-01)
	all	-0.3388	-0.3719	-0.4815	-0.1227	-0.0804	-0.4701
		(1.001E-14)	(5.967E-18)	(1.616E-05)	(3.519E-06)	(1.485E-01)	(9.981E-01)
RFR	0	0.1644	0.1218	0.0765	-0.5679	-0.2707	-1.2435
		(6.369E-17)	(2.867E-15)	(6.978E-04)	(6.444E-09)	(6.657E-03)	(9.910E-01)
	10	0.3904	0.3652	0.3311	**0.3023**	0.5334	0.4312
		(2.400E-16)	(8.450E-15)	(4.116E-02)	(3.036E-07)	(4.232E-04)	(1.389E-01)
	20	0.3509	0.3233	0.2802	0.1156	0.4165	0.4939
		(6.550E-17)	(2.254E-15)	(7.894E-02)	(5.473E-07)	(4.323E-01)	(9.963E-01)
	all	0.3366	0.3092	0.2620	0.1012	0.4069	0.4465
		(4.679E-17)	(1.436E-15)	(8.484E-02)	(1.527E-01)	(8.992E-01)	(9.864E-01)
GBR	0	0.1262	0.0821	0.0368	-0.5351	-0.3546	-1.1792
		(2.970E-17)	(1.418E-15)	(3.394E-04)	(6.945E-08)	(3.171E-02)	(1.000E+00)
	10	0.2712	0.2489	0.2077	0.1331	0.2344	0.2781
		(1.077E-14)	(2.654E-13)	(1.503E-02)	(3.247E-01)	(8.151E-05)	(2.180E-01)
	20	0.1082	0.0739	0.0023	-0.4301	-0.6828	0.3736
		(1.293E-19)	(6.170E-18)	(2.454E-02)	(2.570E-06)	(8.497E-01)	(9.956E-01)
	all	0.1630	0.1266	0.0598	-0.1553	-0.0990	0.3394
		(1.314E-19)	(4.881E-18)	(1.564E-02)	(2.266E-06)	(1.317E-08)	(5.420E-02)
ANN	0	-0.1792	-0.2224	-0.3356	-0.0549	0.2994	0.1402
		(5.113E-12)	(4.647E-12)	(5.583E-05)	(7.479E-01)	(5.361E-01)	(6.861E-02)
	10	0.3931	0.3763	0.3590	-0.1711	0.1225	-0.0981
		(4.965E-06)	(2.072E-05)	(1.444E-06)	(2.086E-01)	(9.991E-01)	(4.872E-01)
	20	0.4052	0.3891	0.3680	-0.6005	-0.6262	-0.8552
		(2.865E-16)	(2.758E-15)	(6.164E-06)	(7.019E-06)	(2.046E-03)	(1.657E-01)
	all	-3.7435	-3.9785	-4.6024	-0.9025	-0.3906	-0.8096
		(4.146E-18)	(1.864E-16)	(3.584E-09)	(5.994E-07)	(7.837E-02)	(1.790E-01)
RNN	0	0.0488	0.0102	-0.0347	-0.0255	0.1527	-0.4145
		(5.805E-10)	(4.372E-09)	(8.138E-04)	(7.119E-01)	(2.079E-01)	(3.593E-01)
	10	0.0892	0.0449	-0.0159	0.1064	0.5063	0.2701
		(1.141E-10)	(1.593E-09)	(3.448E-04)	(6.350E-08)	(6.999E-03)	(1.280E-01)
	20	0.1032	0.0676	0.0215	0.1578	**0.5888**	**0.6056**
		(6.659E-10)	(7.596E-09)	(1.424E-03)	(2.054E-01)	(5.465E-05)	(3.197E-02)
	all	0.0663	0.0294	0.0061	0.2718	0.5737	0.5002
		(6.894E-10)	(2.572E-09)	(2.506E-04)	(3.152E-07)	(2.702E-02)	(9.885E-01)

Table 4. Monthly Forecasting Performance. R^2_{oos} scores for entire period and subperiod ($t \in \{10, 5, 3, 2, 1\}$) are represented. The numbers in parentheses indicate the p-values of residual stationarity. The bold numbers are the best performance measurements for each subperiod.

Forecasting Model	Number of Feature Selection	R^2_{OOS}	$R^2_{OOS,10}$	$R^2_{OOS,5}$	$R^2_{OOS,3}$	$R^2_{OOS,2}$	$R^2_{OOS,1}$
HAR	-	-0.4331	-0.5165	-0.6729	0.0788	0.0702	-0.0623
		(1.471E-15)	(2.663E-12)	(2.883E-07)	(3.739E-12)	(2.547E-03)	(2.408E-06)
TV-HAR	-	0.1240	0.0891	0.0197	0.2160	0.2571	0.1236
		(8.207E-18)	(7.085E-14)	(1.216E-07)	(1.090E-12)	(1.267E-12)	(6.754E-06)
HAR-X	0	-0.4331	-0.5165	-0.6729	0.0788	0.0702	-0.0623
		(1.471E-15)	(2.663E-12)	(2.883E-07)	(3.739E-12)	(2.547E-03)	(2.408E-06)
	10	-0.0442	-0.0670	-0.1590	-0.1486	-0.1498	0.0735
		(1.502E-05)	(3.839E-10)	(2.935E-26)	(2.547E-13)	(2.798E-13)	(1.778E-09)
	20	-0.0568	-0.0768	-0.1621	-0.6356	-0.9367	-1.9959
		(9.822E-14)	(8.450E-11)	(1.794E-05)	(6.374E-14)	(2.520E-12)	(2.046E-04)
	all	-0.4293	-0.1344	-0.2318	-1.1232	-1.8694	-3.6012
		(1.761E-13)	(1.518E-09)	(1.534E-05)	(6.778E-13)	(6.827E-10)	(2.259E-03)
LASSO	0	0.1724	0.1455	0.0814	0.2501	0.2458	0.0157
		(3.072E-17)	(4.827E-14)	(1.725E-08)	(1.690E-12)	(8.294E-12)	(1.583E-05)
	10	0.0692	0.0312	-0.0522	0.2273	0.2668	0.0754
		(3.391E-17)	(6.652E-14)	(1.655E-08)	(1.077E-10)	(1.866E-10)	(5.149E-05)
	20	0.0871	0.0511	-0.0289	0.1833	0.2666	0.0814
		(1.555E-17)	(3.371E-14)	(1.464E-08)	(1.917E-11)	(2.988E-12)	(3.661E-05)
	all	0.0867	0.1744	0.0995	-0.5802	-1.2603	-1.6866
		(1.900E-12)	(5.250E-10)	(1.554E-05)	(2.665E-12)	(1.210E-11)	(9.654E-03)
ElasticNet	0	0.1713	0.1451	0.0810	0.2486	0.2446	0.0197
		(2.450E-17)	(3.507E-14)	(1.360E-08)	(1.721E-12)	(8.137E-12)	(1.713E-05)
	10	0.1450	0.1135	0.0423	0.2495	0.2417	0.2120
		(5.186E-16)	(3.881E-13)	(1.020E-07)	(1.478E-11)	(3.881E-12)	(9.338E-07)
	20	0.1700	0.1391	0.0715	0.1906	0.2595	0.2136
		(1.243E-16)	(1.359E-13)	(7.929E-08)	(2.680E-12)	(2.098E-14)	(4.261E-07)
	all	0.0228	0.1533	0.0768	-0.6958	-1.4011	-2.1249
		(6.375E-13)	(7.425E-11)	(2.863E-06)	(2.870E-12)	(4.179E-12)	(4.316E-02)
DTR	0	-0.4325	-0.4692	-0.5355	-3.7725	-1.4038	-2.4123
		(8.693E-16)	(2.541E-12)	(4.940E-07)	(3.658E-15)	(7.028E-12)	(1.147E-05)
	10	-0.2340	-0.2539	-0.3301	-1.8767	-0.7251	-1.3815
		(2.188E-17)	(1.437E-11)	(8.426E-09)	(1.045E-14)	(1.157E-12)	(1.490E-07)
	20	-0.0839	-0.0952	-0.1367	-1.9534	-0.2968	-0.7026
		(0.000E+00)	(2.079E-30)	(5.238E-25)	(2.002E-12)	(2.542E-11)	(1.418E-03)
	all	0.0647	0.0470	0.0123	-0.3377	-0.5570	-0.0361
		(9.164E-16)	(7.741E-13)	(1.856E-08)	(2.446E-19)	(6.756E-17)	(2.082E-07)
RFR	0	0.2153	0.2149	0.1811	-0.3692	0.0289	0.0245
		(5.245E-15)	(3.728E-12)	(9.815E-07)	(2.024E-15)	(1.631E-12)	(1.216E-01)
	10	0.3123	0.3205	0.2754	0.3384	0.0147	-0.0736
		(2.016E-16)	(9.289E-14)	(1.423E-08)	(1.029E-11)	(3.652E-03)	(3.997E-04)
	20	0.4017	0.4056	0.3755	0.4038	0.2424	0.0699
		(4.199E-15)	(4.871E-12)	(1.403E-07)	(1.540E-12)	(4.180E-04)	(1.067E-04)
	all	**0.4059**	**0.4121**	**0.3833**	0.3914	0.2325	0.1978
		(1.036E-16)	(3.783E-14)	(1.575E-08)	(1.180E-12)	(2.667E-12)	(1.191E-05)
GBR	0	0.1327	0.1364	0.1045	-0.6758	0.0271	-0.1844
		(3.405E-15)	(3.168E-12)	(7.101E-07)	(3.688E-16)	(1.539E-12)	(9.086E-02)
	10	0.2331	0.2320	0.1814	0.4332	0.2055	0.1296
		(5.805E-17)	(3.191E-15)	(3.698E-09)	(3.902E-10)	(4.199E-10)	(2.195E-05)
	20	0.3058	0.3084	0.2749	0.4065	0.2670	0.1989
		(1.199E-15)	(2.082E-13)	(2.461E-08)	(9.598E-11)	(1.084E-11)	(4.096E-05)
	all	0.3687	0.3785	0.3586	0.4026	**0.3182**	0.2105
		(4.186E-18)	(8.608E-16)	(1.349E-10)	(4.284E-10)	(2.447E-04)	(2.794E-01)
ANN	0	0.0898	0.0667	-0.0034	0.1598	0.0732	-0.0893
		(2.762E-17)	(5.953E-14)	(5.195E-09)	(6.739E-14)	(7.992E-13)	(5.639E-05)
	10	0.0432	0.0096	-0.0860	-0.3812	-1.1930	-3.1175
		(1.201E-09)	(1.558E-08)	(2.050E-04)	(4.485E-13)	(7.545E-13)	(8.074E-08)
	20	0.1891	0.1766	0.1195	-0.7372	-0.9783	-1.9248
		(9.951E-11)	(2.494E-30)	(8.753E-25)	(1.257E-04)	(8.250E-09)	(1.002E-06)
	all	-1.8634	-2.0457	-2.4374	-0.8875	-0.9708	-1.9250
		(4.023E-30)	(9.166E-28)	(4.068E-19)	(1.316E-11)	(3.112E-08)	(1.432E-10)
RNN	0	0.2298	0.2127	0.1725	0.1970	0.1550	0.0850
		(3.430E-15)	(2.047E-12)	(8.075E-07)	(2.705E-15)	(2.888E-13)	(7.678E-07)
	10	0.3321	0.3359	0.3025	0.4098	0.2218	0.1539
		(1.961E-11)	(2.146E-09)	(3.208E-05)	(8.352E-11)	(3.023E-11)	(3.771E-05)
	20	0.2806	0.2667	0.2188	0.2068	0.1776	0.0173
		(1.830E-11)	(3.291E-09)	(2.911E-05)	(1.743E-10)	(8.148E-13)	(2.713E-07)
	all	0.2578	0.2373	0.1996	**0.4725**	0.2964	**0.2947**
		(6.886E-08)	(6.112E-06)	(2.988E-06)	(1.072E-10)	(1.575E-02)	(9.153E-05)

Table 5. Bi-weekly Forecasting Performance. R^2_{oos} scores for entire period and subperiod ($t \in \{10, 5, 3, 2, 1\}$) are represented. The numbers in parentheses indicate the p-values of residual stationarity. The bold numbers are the best performance measurements for each subperiod.

Forecasting Model	Number of Feature Selection	R^2_{OOS}	$R^2_{OOS,10}$	$R^2_{OOS,5}$	$R^2_{OOS,3}$	$R^2_{OOS,2}$	$R^2_{OOS,1}$
HAR	-	-0.0714	-0.1217	-0.2127	-0.0059	0.0323	-0.1408
		(6.045E-17)	(9.597E-13)	(2.885E-07)	(6.728E-18)	(7.068E-17)	(5.059E-09)
TV-HAR	-	0.3117	0.2990	0.2749	0.1455	0.1225	0.0129
		(2.151E-19)	(6.913E-15)	(8.617E-09)	(6.334E-27)	(8.163E-11)	(1.361E-14)
HAR-X	0	-0.0714	-0.1217	-0.2127	-0.0059	0.0323	-0.1408
		(6.045E-17)	(9.597E-13)	(2.885E-07)	(6.728E-18)	(7.068E-17)	(5.059E-09)
	10	0.2495	0.2571	0.2279	-0.2038	-0.3675	-0.5944
		(1.013E-12)	(3.066E-10)	(2.523E-05)	(2.234E-24)	(1.808E-21)	(7.910E-16)
	20	0.2082	0.2708	0.2487	-0.4295	-0.7749	-1.7198
		(1.646E-14)	(1.611E-11)	(6.935E-06)	(6.239E-24)	(1.835E-20)	(3.046E-12)
	all	-1.1823	0.2932	0.2632	-0.4458	-0.9169	-1.8841
		(4.979E-04)	(1.317E-09)	(2.614E-19)	(2.301E-22)	(3.547E-17)	(1.661E-08)
LASSO	0	0.3367	0.3321	0.3114	0.1872	0.1673	0.0231
		(1.174E-14)	(1.205E-11)	(8.683E-07)	(1.618E-26)	(2.353E-22)	(1.272E-14)
	10	0.3296	0.3239	0.3016	0.1889	0.1700	0.0260
		(6.190E-16)	(1.045E-12)	(1.621E-07)	(2.387E-26)	(3.201E-22)	(1.472E-14)
	20	0.3299	0.3239	0.3016	0.1889	0.1700	0.0260
		(6.322E-16)	(1.045E-12)	(1.621E-07)	(2.387E-26)	(3.201E-22)	(1.472E-14)
	all	0.2928	0.3658	0.3338	-0.1595	-0.4780	-0.7846
		(2.227E-13)	(1.310E-10)	(3.810E-06)	(3.262E-21)	(7.187E-17)	(3.602E-10)
ElasticNet	0	0.3398	0.3332	0.3124	0.1852	0.1645	0.0248
		(1.618E-16)	(5.191E-13)	(1.502E-07)	(1.320E-26)	(1.758E-22)	(1.237E-14)
	10	0.3577	0.3509	0.3335	0.2113	0.1847	0.0572
		(1.943E-17)	(1.161E-13)	(4.603E-08)	(3.535E-26)	(2.769E-22)	(1.350E-14)
	20	0.3584	0.3521	0.3350	0.1975	0.1786	**0.0735**
		(2.920E-17)	(1.258E-13)	(4.836E-08)	(2.361E-26)	(5.986E-23)	(4.284E-15)
	all	0.2568	0.3645	0.3332	-0.2146	-0.5612	-1.0275
		(2.294E-12)	(4.002E-10)	(7.815E-06)	(5.368E-22)	(1.015E-17)	(4.570E-10)
DTR	0	-0.2206	-0.1725	-0.1309	-1.8657	-1.4720	-2.3459
		(2.153E-16)	(1.191E-11)	(7.171E-06)	(3.911E-22)	(9.982E-19)	(5.843E-09)
	10	-0.2822	-0.2305	-0.2959	-5.9182	-13.7236	-41.8268
		(1.195E-19)	(1.764E-15)	(1.667E-08)	(1.224E-25)	(1.383E-20)	(2.284E-12)
	20	-0.0511	0.0016	-0.0174	-0.9437	-2.4880	-5.3090
		(7.846E-20)	(2.562E-08)	(1.872E-18)	(2.031E-25)	(1.856E-20)	(1.641E-11)
	all	-0.1687	-0.1334	-0.1704	-1.9978	-4.7846	-12.5673
		(5.427E-14)	(1.110E-10)	(2.366E-07)	(2.579E-26)	(1.276E-20)	(1.110E-12)
RFR	0	0.2462	0.2516	0.2453	0.0239	0.0903	-0.0627
		(1.157E-15)	(3.219E-12)	(3.336E-06)	(6.389E-25)	(6.060E-10)	(4.391E-09)
	10	0.3319	0.3463	0.3125	-0.6355	-2.0051	-6.5592
		(2.497E-14)	(1.239E-15)	(2.334E-08)	(5.984E-25)	(9.341E-20)	(2.090E-12)
	20	0.3900	0.4037	0.3846	-0.3368	-1.3978	-5.0238
		(1.917E-15)	(4.770E-12)	(1.325E-06)	(4.639E-27)	(1.119E-14)	(9.976E-09)
	all	0.3877	0.3990	0.3768	-0.2965	-1.2702	-4.4404
		(4.186E-16)	(9.884E-13)	(2.392E-07)	(1.884E-26)	(1.573E-14)	(1.375E-13)
GBR	0	0.2203	0.2309	0.2075	-0.1693	0.1164	0.0024
		(2.421E-14)	(3.937E-11)	(3.826E-07)	(7.883E-15)	(8.966E-22)	(6.636E-12)
	10	0.1269	0.1225	0.0629	-2.1277	-5.3814	-17.8760
		(2.634E-20)	(5.655E-16)	(1.191E-08)	(5.354E-25)	(5.946E-20)	(8.667E-12)
	20	0.3544	0.3729	0.3600	-0.3647	-1.4314	-5.0739
		(1.580E-18)	(2.536E-14)	(1.228E-22)	(1.521E-26)	(8.279E-15)	(2.670E-13)
	all	0.3614	0.3745	0.3611	-0.4204	-1.4232	-4.9260
		(6.777E-16)	(2.961E-12)	(1.637E-11)	(9.155E-26)	(1.170E-14)	(2.181E-13)
ANN	0	0.2772	0.2693	0.2415	0.1924	0.1193	0.0177
		(1.110E-16)	(2.283E-14)	(1.089E-08)	(2.328E-25)	(1.425E-19)	(2.016E-10)
	10	0.3867	0.3861	0.3719	-0.0020	-0.3044	-1.2751
		(9.693E-23)	(2.699E-19)	(1.481E-11)	(1.479E-22)	(1.383E-13)	(1.518E-11)
	20	0.2431	0.2492	0.2340	-0.4173	-0.9081	-1.8306
		(7.053E-21)	(1.389E-16)	(5.689E-13)	(3.548E-04)	(6.624E-13)	(1.869E-11)
	all	-0.4727	-0.5301	-0.6294	-1.0350	-2.1418	-5.1300
		(2.819E-15)	(5.803E-28)	(5.788E-20)	(8.204E-22)	(2.608E-13)	(3.830E-11)
RNN	0	0.3254	0.3200	0.3020	0.1655	0.0944	0.0168
		(8.035E-14)	(1.731E-10)	(8.619E-06)	(1.473E-24)	(8.473E-20)	(2.083E-11)
	10	**0.4191**	**0.4254**	**0.4150**	0.3174	0.0222	-0.3291
		(2.713E-14)	(2.070E-11)	(1.021E-06)	(2.404E-25)	(1.345E-18)	(8.663E-12)
	20	0.3584	0.3606	0.3450	0.3407	0.1835	-0.0238
		(5.590E-14)	(5.548E-11)	(1.817E-05)	(1.316E-23)	(3.676E-18)	(1.233E-12)
	all	0.3210	0.3265	0.2966	**0.3722**	**0.1942**	-0.0373
		(2.498E-15)	(8.128E-13)	(2.166E-07)	(4.872E-25)	(2.693E-18)	(2.482E-10)

Table 6. Weekly Forecasting Performance. R^2_{oos} scores for entire period and subperiod ($t \in \{10, 5, 3, 2, 1\}$) are represented. The numbers in parentheses indicate the p-values of residual stationarity. The bold numbers are the best performance measurements for each subperiod.

Forecasting Horizon The longer the forecast horizon, the better the performances of most models tend to be (Table 4, 5, 6). This might be simply from the noise of the data, as data with shorter frequencies contain a smaller set of data for each realized volatility. However, on the contrary, the performance of the benchmark model decreases as the forecast horizon increases. Since the HAR model works well enough with the different horizons, it cannot be said that this is because memory is lost as the time horizon lengthens. Rather, it can be interpreted that the short-term variable and long-term variables of the HAR model do not sufficiently reflect the momentum of the realized volatility. We can see that the RFR and RNN models tend to make robust predictions over the forecast horizon compared to other models, and the results themselves are superior to others.

In addition to the forecast horizon, robustness to the sub-period length also has important implications. While it is very important to get it right over the entire time, it is also very important to get it right over the recent period. LASSO and ElasticNet show great performance for longer sub-periods, while RFR and RNN show better performance for recent times. The RFR and RNN show a very stable R^2 score even during the period of COVID-19. This is particularly meaningful because not only the result is statistically significant (the residual is stationary) but they are robust to the test period. Interestingly, the results for LASSO and ElasticNet using recent time periods are a bit unusual. While the performance of the models themselves decreases as the forecast horizon decreases, the ranking of LASSO and ElaticNet actually increases. This suggests that oil prices have been very volatile in recent periods, and the momentum played much important role than other features.

Forecasting Model	Number of Feature Selection	R^2_{OOS}	$R^2_{OOS,10}$	$R^2_{OOS,5}$	$R^2_{OOS,3}$	$R^2_{OOS,2}$	$R^2_{OOS,1}$
HAR	-	-0.8224	-0.8918	-1.0895	-0.1665	-0.1312	-1.0531
		(2.273E-03)	(4.362E-03)	(2.453E-01)	(9.286E-01)	(1.728E-02)	(7.422E-02)
TV-HAR	-	-0.1154	-0.1759	-0.2851	-0.1238	0.3608	0.1883
		(5.893E-21)	(4.605E-19)	(1.943E-05)	(4.493E-01)	(2.088E-01)	(3.401E-02)
HAR-X	0	-0.8224	-0.8918	-1.0895	-0.1665	-0.1312	-1.0531
		(2.273E-03)	(4.362E-03)	(2.453E-01)	(9.286E-01)	(1.728E-02)	(7.422E-02)
	10	0.1144	0.0777	0.0183	-0.9011	-1.0257	-0.2328
		(3.581E-17)	(1.620E-15)	(1.738E-08)	(1.565E-02)	(3.521E-04)	(1.051E-02)
	20	0.0387	0.0034	-0.0793	-1.3337	-2.2149	-0.3676
		(1.409E-20)	(1.080E-18)	(1.168E-10)	(4.948E-01)	(4.345E-06)	(4.611E-04)
	all	0.4685	0.4636	0.4581	-1.1800	-1.6785	-0.4692
		(3.248E-07)	(1.159E-13)	(1.734E-04)	(2.572E-02)	(2.566E-01)	(1.000E+00)
LASSO	0	-0.0256	-0.0788	-0.1795	-0.0518	0.3053	-0.0484
		(5.237E-11)	(6.249E-10)	(4.716E-05)	(5.148E-01)	(8.338E-01)	(7.259E-02)
	10	0.2071	0.1678	0.1058	-0.0758	0.3622	0.2709
		(9.274E-13)	(2.343E-11)	(7.440E-02)	(1.269E-10)	(1.406E-04)	(3.048E-02)
	20	0.2215	0.1830	0.1116	-0.0758	0.3622	0.2709
		(5.528E-13)	(1.470E-11)	(4.797E-02)	(1.269E-10)	(1.406E-04)	(3.048E-02)
	all	**0.5602**	**0.5474**	**0.5216**	-0.5571	-0.2091	0.0758
		(1.104E-17)	(1.177E-15)	(2.227E-08)	(9.731E-02)	(5.403E-05)	(7.254E-02)
ElasticNet	0	-0.0376	-0.0911	-0.1924	-0.0595	0.3015	-0.0452
		(1.311E-10)	(1.308E-09)	(8.332E-05)	(5.239E-01)	(7.996E-01)	(6.511E-02)
	10	0.2301	0.1929	0.1344	0.0662	0.4722	0.3081
		(1.160E-12)	(1.685E-11)	(4.118E-02)	(1.267E-01)	(5.872E-05)	(1.638E-02)
	20	0.2119	0.1740	0.1011	0.0111	0.4727	0.3075
		(8.128E-13)	(1.224E-11)	(4.506E-02)	(1.356E-01)	(5.603E-05)	(1.653E-02)
	all	0.5581	0.5463	0.5212	-0.5747	-0.2567	0.0769
		(2.712E-17)	(2.577E-15)	(3.835E-08)	(8.831E-02)	(8.186E-05)	(8.503E-02)
DTR	0	-0.1846	-0.2439	-0.3223	-2.5200	-4.5855	-9.4199
		(7.083E-12)	(1.508E-10)	(1.143E-05)	(6.053E-06)	(9.410E-01)	(4.531E-02)
	10	0.0772	0.0730	0.0865	0.2048	0.0403	-0.1204
		(1.938E-13)	(7.730E-12)	(9.661E-02)	(1.514E-06)	(1.000E+00)	(3.485E-02)
	20	0.0855	0.0826	0.0670	0.0598	0.2555	-0.2466
		(1.015E-13)	(3.509E-12)	(1.018E-01)	(2.612E-09)	(5.921E-01)	(9.447E-01)
	all	-0.0329	-0.0420	-0.0856	-0.0870	0.1783	-0.2413
		(4.458E-13)	(1.372E-11)	(1.110E-01)	(1.333E-12)	(2.495E-06)	(9.927E-01)
RFR	0	0.1619	0.1191	0.0741	-0.5734	-0.2769	-1.2444
		(7.406E-17)	(3.315E-15)	(7.310E-04)	(8.765E-09)	(8.249E-03)	(9.927E-01)
	10	0.3996	0.3747	0.3445	0.2765	**0.5252**	0.4003
		(8.176E-13)	(1.728E-11)	(9.641E-05)	(5.663E-07)	(4.202E-04)	(1.305E-01)
	20	0.3890	0.3628	0.3281	0.2058	0.4313	**0.4681**
		(2.925E-13)	(5.877E-12)	(7.337E-05)	(1.140E-07)	(1.715E-04)	(1.893E-01)
	all	0.3900	0.3633	0.3263	**0.2798**	0.4507	0.4183
		(3.774E-13)	(7.489E-12)	(1.052E-04)	(2.022E-01)	(7.107E-01)	(2.762E-01)
GBR	0	0.1129	0.0680	0.0209	-0.5325	-0.3486	-1.1792
		(1.800E-17)	(8.916E-16)	(2.536E-04)	(8.862E-08)	(4.154E-02)	(1.000E+00)
	10	0.3156	0.2944	0.2640	0.1661	0.2838	0.2808
		(6.920E-13)	(1.740E-11)	(6.580E-05)	(3.576E-01)	(1.597E-04)	(2.158E-01)
	20	0.3279	0.3078	0.2804	-0.6926	-1.2274	0.3230
		(1.122E-11)	(1.901E-10)	(2.462E-05)	(7.473E-08)	(2.856E-09)	(9.529E-02)
	all	0.3690	0.3447	0.3150	-0.4082	-0.3846	0.3471
		(2.175E-12)	(4.404E-11)	(9.949E-06)	(1.114E-07)	(2.762E-07)	(1.276E-01)
ANN	0	-0.1702	-0.2119	-0.3231	-0.0139	0.2969	0.0773
		(7.448E-12)	(1.009E-11)	(8.182E-05)	(6.271E-01)	(7.602E-01)	(1.347E-01)
	10	0.4105	0.3971	0.3884	-0.1129	0.1507	-0.1438
		(6.680E-14)	(4.604E-13)	(5.001E-05)	(2.673E-01)	(1.000E+00)	(6.024E-01)
	20	0.3738	0.3486	0.3458	-0.1116	-0.2108	-0.6445
		(1.097E-16)	(1.992E-15)	(1.470E-16)	(2.468E-07)	(7.774E-01)	(6.415E-01)
	all	0.3484	0.3289	0.3275	-0.7419	-0.9420	-2.6058
		(5.600E-21)	(1.795E-19)	(6.915E-04)	(2.500E-07)	(4.844E-01)	(7.730E-01)
RNN	0	0.0469	0.0066	-0.0369	-0.0439	0.1248	-0.4881
		(3.390E-10)	(2.627E-09)	(7.359E-04)	(6.892E-01)	(4.299E-01)	(3.271E-01)
	10	0.1347	0.0939	0.0426	0.1647	0.4339	0.2462
		(4.888E-10)	(6.055E-09)	(7.142E-04)	(3.782E-08)	(1.713E-02)	(1.278E-01)
	20	0.1524	0.1202	0.0652	0.0964	0.1485	-0.0698
		(1.322E-10)	(1.146E-09)	(1.569E-03)	(2.063E-01)	(1.321E-02)	(9.962E-01)
	all	0.0755	0.0408	0.0187	0.2765	0.5025	0.4409
		(7.702E-10)	(2.657E-09)	(2.060E-04)	(2.970E-07)	(1.557E-02)	(9.990E-01)

Table 7. Monthly Forecasting Performance with Uncertainty Indices. R^2_{oos} scores for entire period and subperiod ($t \in \{10, 5, 3, 2, 1\}$) are represented. The numbers in parentheses indicate the p-values of residual stationarity. The bold numbers are the best performance measurements for each subperiod.

Forecasting Model	Number of Feature Selection	R^2_{OOS}	$R^2_{OOS,10}$	$R^2_{OOS,5}$	$R^2_{OOS,3}$	$R^2_{OOS,2}$	$R^2_{OOS,1}$
HAR	-	-0.4331	-0.5165	-0.6729	0.0788	0.0702	-0.0623
		(1.471E-15)	(2.663E-12)	(2.883E-07)	(3.739E-12)	(2.547E-03)	(2.408E-06)
TV-HAR	-	0.1240	0.0891	0.0197	0.2160	0.2571	0.1236
		(8.207E-18)	(7.085E-14)	(1.216E-07)	(1.090E-12)	(1.267E-12)	(6.754E-06)
HAR-X	0	-0.4331	-0.5165	-0.6729	0.0788	0.0702	-0.0623
		(1.471E-15)	(2.663E-12)	(2.883E-07)	(3.739E-12)	(2.547E-03)	(2.408E-06)
	10	0.2601	0.2612	0.2202	-0.2401	-0.0644	-0.1814
		(4.170E-14)	(9.497E-12)	(2.150E-06)	(3.988E-13)	(4.247E-13)	(5.232E-01)
	20	0.2335	0.2304	0.1916	-0.5422	-0.1445	0.0414
		(1.692E-13)	(4.809E-11)	(3.602E-06)	(1.840E-12)	(7.702E-13)	(3.656E-07)
	all	0.2516	0.2458	0.1988	-0.7750	-0.9573	-0.5857
		(7.572E-10)	(4.139E-08)	(6.211E-05)	(1.112E-11)	(3.391E-09)	(5.791E-04)
LASSO	0	0.1724	0.1455	0.0814	0.2501	0.2458	0.0157
		(3.072E-17)	(4.827E-14)	(1.725E-08)	(1.690E-12)	(8.294E-12)	(1.583E-05)
	10	0.2412	0.2188	0.1670	0.1928	0.2668	0.0814
		(4.763E-17)	(4.405E-14)	(2.112E-08)	(1.998E-11)	(3.151E-12)	(3.661E-05)
	20	0.2409	0.2188	0.1670	0.1928	0.2668	0.0814
		(4.693E-17)	(4.405E-14)	(2.112E-08)	(1.998E-11)	(3.151E-12)	(3.661E-05)
	all	0.3177	0.3053	0.2517	0.1807	-0.0947	0.0516
		(3.398E-12)	(1.112E-09)	(2.688E-05)	(2.919E-14)	(3.681E-12)	(7.178E-05)
ElasticNet	0	0.1713	0.1451	0.0810	0.2486	0.2446	0.0197
		(2.450E-17)	(3.507E-14)	(1.360E-08)	(1.721E-12)	(8.137E-12)	(1.713E-05)
	10	0.2844	0.2651	0.2191	0.2263	0.2989	0.2097
		(1.273E-16)	(3.364E-14)	(2.301E-08)	(1.449E-11)	(5.154E-13)	(2.380E-05)
	20	0.2814	0.2602	0.2126	0.2216	0.2985	0.2097
		(7.886E-17)	(3.258E-14)	(2.291E-08)	(1.387E-11)	(5.136E-13)	(2.380E-05)
	all	0.3094	0.2958	0.2429	0.0689	-0.2175	-0.0307
		(1.226E-12)	(5.212E-10)	(1.578E-05)	(1.446E-13)	(2.824E-11)	(1.611E-04)
DTR	0	-0.4030	-0.4207	-0.4609	-1.0836	-1.4369	-2.4828
		(2.277E-15)	(7.154E-12)	(1.433E-07)	(3.983E-14)	(3.434E-12)	(9.017E-06)
	10	-0.0261	-0.0170	-0.0362	-1.8843	-0.7706	-1.0411
		(4.796E-17)	(5.623E-13)	(7.743E-07)	(1.928E-16)	(9.395E-17)	(6.755E-03)
	20	0.1224	0.1261	0.1072	-1.8057	-0.4107	-0.7176
		(3.113E-11)	(4.435E-09)	(3.469E-05)	(5.907E-15)	(1.873E-12)	(9.336E-05)
	all	0.0022	-0.0103	-0.0634	-0.0145	-0.1713	-0.0195
		(7.695E-11)	(4.156E-08)	(1.761E-04)	(1.140E-07)	(4.450E-16)	(1.961E-09)
RFR	0	0.2228	0.2229	0.1903	-0.3520	0.0430	0.0110
		(8.574E-15)	(5.530E-12)	(1.257E-06)	(1.421E-15)	(1.502E-12)	(1.322E-01)
	10	0.4157	0.4263	0.3989	0.4385	0.2313	0.0671
		(5.086E-13)	(1.075E-10)	(2.047E-06)	(1.168E-12)	(1.199E-08)	(3.233E-05)
	20	0.4332	0.4365	0.4103	0.4379	0.2529	0.1195
		(3.607E-15)	(1.372E-12)	(1.168E-07)	(3.055E-12)	(3.044E-12)	(8.426E-06)
	all	0.4220	0.4262	0.3991	0.4003	0.2742	0.2264
		(9.403E-16)	(3.742E-13)	(6.208E-08)	(1.501E-12)	(2.426E-12)	(9.950E-06)
GBR	0	0.1325	0.1324	0.0997	-0.6054	0.0271	-0.1844
		(4.129E-15)	(3.481E-12)	(6.767E-07)	(1.188E-15)	(1.539E-12)	(9.086E-02)
	10	0.4224	0.4355	0.4149	0.4191	0.3391	0.1340
		(4.113E-12)	(2.471E-10)	(2.917E-06)	(3.185E-10)	(3.119E-12)	(2.318E-05)
	20	**0.4382**	**0.4516**	**0.4395**	0.3489	0.2681	0.1326
		(1.572E-14)	(2.465E-12)	(8.803E-08)	(4.628E-08)	(1.491E-10)	(3.073E-06)
	all	0.4275	0.4409	0.4327	0.4118	0.2815	0.2174
		(4.274E-15)	(4.644E-13)	(2.171E-08)	(3.122E-11)	(2.002E-11)	(2.143E-05)
ANN	0	0.0682	0.0427	-0.0317	0.1667	0.0431	-0.1548
		(1.534E-17)	(4.212E-14)	(4.223E-09)	(7.337E-14)	(4.340E-13)	(1.125E-05)
	10	0.2614	0.2473	0.1919	0.1750	0.1655	-0.0554
		(2.180E-10)	(1.381E-09)	(1.600E-05)	(8.319E-09)	(8.225E-11)	(1.199E-03)
	20	0.3023	0.2989	0.2736	0.0370	-0.1053	-0.1325
		(2.594E-12)	(1.920E-11)	(2.379E-06)	(6.188E-12)	(7.267E-11)	(9.531E-06)
	all	0.2774	0.2871	0.2747	-0.3376	-0.4558	-0.3589
		(7.206E-14)	(2.639E-12)	(1.778E-07)	(2.387E-10)	(1.255E-03)	(1.872E-01)
RNN	0	0.2079	0.1921	0.1485	0.1730	0.0915	-0.0419
		(1.431E-14)	(8.350E-12)	(1.880E-06)	(6.609E-14)	(1.368E-12)	(2.223E-06)
	10	0.3486	0.3379	0.3027	0.4052	**0.3539**	0.1496
		(4.026E-13)	(1.103E-10)	(2.899E-06)	(9.886E-09)	(4.915E-11)	(3.805E-03)
	20	0.3111	0.3055	0.2787	0.3561	0.2412	0.2320
		(7.085E-14)	(1.848E-11)	(8.996E-07)	(1.002E-07)	(1.206E-01)	(1.697E-04)
	all	0.2516	0.2280	0.1828	**0.4650**	0.2926	**0.2460**
		(3.833E-13)	(8.594E-11)	(2.698E-06)	(3.253E-09)	(2.174E-03)	(2.352E-04)

Table 8. Bi-weekly Forecasting Performance with Uncertainty Indices. R^2_{oos} scores for entire period and subperiod ($t \in \{10, 5, 3, 2, 1\}$) are represented. The numbers in parentheses indicate the p-values of residual stationarity. The bold numbers are the best performance measurements for each subperiod.

Forecasting Model	Number of Feature Selection	R^2_{OOS}	$R^2_{OOS,10}$	$R^2_{OOS,5}$	$R^2_{OOS,3}$	$R^2_{OOS,2}$	$R^2_{OOS,1}$
HAR	-	-0.0714	-0.1217	-0.2127	-0.0059	0.0323	-0.1408
		(6.045E-17)	(9.597E-13)	(2.885E-07)	(6.728E-18)	(7.068E-17)	(5.059E-09)
TV-HAR	-	0.3117	0.2990	0.2749	0.1455	0.1225	0.0129
		(2.151E-19)	(6.913E-15)	(8.617E-09)	(6.334E-27)	(8.163E-11)	(1.361E-14)
HAR-X	0	-0.0714	-0.1217	-0.2127	-0.0059	0.0323	-0.1408
		(6.045E-17)	(9.597E-13)	(2.885E-07)	(6.728E-18)	(7.068E-17)	(5.059E-09)
	10	0.3059	0.3186	0.2995	-0.0861	0.0106	-0.1231
		(1.592E-16)	(5.208E-14)	(4.284E-08)	(1.004E-23)	(2.169E-20)	(1.450E-12)
	20	0.2824	0.3074	0.2823	-0.2541	-0.0957	0.0305
		(7.778E-16)	(4.292E-13)	(1.648E-07)	(8.955E-25)	(7.375E-22)	(7.469E-14)
	all	0.3474	0.3852	0.3628	-0.2546	-0.2978	-0.2114
		(4.481E-11)	(4.360E-08)	(8.877E-05)	(4.451E-23)	(2.499E-17)	(7.067E-12)
LASSO	0	0.3367	0.3321	0.3114	0.1872	0.1673	0.0231
		(1.174E-14)	(1.205E-11)	(8.683E-07)	(1.618E-26)	(2.353E-22)	(1.272E-14)
	10	0.3376	0.3331	0.3126	0.1889	0.1700	0.0260
		(1.282E-14)	(1.221E-11)	(8.546E-07)	(2.387E-26)	(3.201E-22)	(1.472E-14)
	20	0.3376	0.3331	0.3126	0.1889	0.1700	0.0260
		(1.283E-14)	(1.221E-11)	(8.546E-07)	(2.387E-26)	(3.201E-22)	(1.472E-14)
	all	0.3848	0.3969	0.3705	0.1992	0.0852	0.0585
		(9.367E-14)	(7.932E-11)	(3.422E-06)	(7.732E-25)	(2.885E-05)	(9.827E-14)
ElasticNet	0	0.3398	0.3332	0.3124	0.1852	0.1645	0.0248
		(1.618E-16)	(5.191E-13)	(1.502E-07)	(1.320E-26)	(1.758E-22)	(1.237E-14)
	10	0.3588	0.3524	0.3353	0.2128	0.2001	0.0822
		(7.259E-16)	(1.281E-12)	(1.967E-07)	(4.884E-26)	(5.583E-22)	(3.547E-14)
	20	0.3575	0.3513	0.3340	0.1981	0.1948	0.0977
		(6.743E-16)	(1.131E-12)	(1.810E-07)	(3.484E-26)	(1.171E-22)	(1.123E-14)
	all	0.3878	0.3984	0.3735	0.1769	0.0863	0.0564
		(9.756E-14)	(1.049E-10)	(3.877E-06)	(3.044E-25)	(1.489E-05)	(3.906E-14)
DTR	0	-0.2966	-0.2424	-0.2173	-0.7698	-1.2630	-1.5798
		(2.571E-18)	(9.006E-14)	(7.943E-08)	(1.270E-21)	(1.293E-10)	(1.565E-10)
	10	-0.2386	-0.2220	-0.2806	-0.1886	-0.7208	-1.1564
		(4.435E-18)	(1.048E-13)	(8.784E-08)	(6.482E-25)	(8.301E-17)	(5.378E-12)
	20	-0.2525	-0.2309	-0.2726	-0.3411	-0.4462	-0.7275
		(2.395E-12)	(6.108E-09)	(1.036E-04)	(3.523E-29)	(1.706E-10)	(1.472E-12)
	all	-0.3255	-0.3111	-0.3645	-0.5204	-0.7524	-0.9286
		(4.783E-14)	(1.652E-08)	(2.058E-08)	(1.362E-29)	(1.248E-03)	(3.870E-08)
RFR	0	0.2455	0.2510	0.2449	0.0316	0.0896	-0.0775
		(4.173E-16)	(1.282E-12)	(2.246E-07)	(5.965E-25)	(4.826E-10)	(5.790E-09)
	10	0.4214	**0.4407**	**0.4274**	0.3240	0.1997	0.0566
		(2.654E-15)	(1.983E-12)	(1.080E-07)	(4.968E-24)	(1.325E-18)	(1.695E-10)
	20	**0.4227**	0.4376	0.4243	0.3481	0.1747	**0.1215**
		(1.143E-16)	(2.685E-13)	(4.200E-08)	(7.566E-26)	(1.782E-05)	(1.674E-12)
	all	0.4206	0.4353	0.4196	0.3022	0.1179	0.0993
		(6.574E-17)	(1.226E-13)	(1.633E-08)	(5.679E-26)	(2.372E-05)	(1.249E-12)
GBR	0	0.2001	0.2083	0.1802	-0.1197	0.1183	0.0202
		(2.660E-14)	(1.094E-12)	(1.146E-06)	(1.383E-14)	(1.108E-21)	(1.710E-11)
	10	0.3083	0.3169	0.2931	0.2776	0.1671	0.0378
		(1.613E-16)	(2.175E-11)	(1.079E-06)	(8.415E-24)	(8.784E-20)	(1.140E-09)
	20	0.3561	0.3694	0.3557	0.2418	0.1512	0.0228
		(1.443E-16)	(6.048E-12)	(8.274E-07)	(2.766E-23)	(3.112E-05)	(1.756E-11)
	all	0.3497	0.3606	0.3395	0.2606	0.1617	0.0634
		(1.997E-15)	(3.194E-12)	(5.048E-07)	(2.245E-23)	(2.052E-05)	(4.427E-12)
ANN	0	0.2995	0.2950	0.2730	0.2002	0.1233	0.0065
		(2.356E-16)	(2.791E-14)	(1.293E-08)	(1.167E-25)	(4.722E-20)	(2.788E-11)
	10	0.3883	0.3957	0.3864	0.2196	0.1475	-0.1258
		(2.275E-15)	(4.064E-12)	(1.313E-06)	(6.690E-18)	(1.359E-16)	(2.137E-08)
	20	0.2902	0.2980	0.2911	-0.1979	-0.2956	-1.0640
		(1.299E-16)	(2.206E-13)	(3.362E-07)	(9.375E-23)	(4.379E-21)	(2.091E-12)
	all	0.2713	0.2888	0.3015	-0.1994	-0.6120	-1.4317
		(1.271E-15)	(3.977E-12)	(1.706E-06)	(1.129E-21)	(1.359E-03)	(3.721E-09)
RNN	0	0.3247	0.3203	0.3034	0.2132	0.1238	0.0993
		(6.578E-14)	(1.723E-10)	(8.985E-06)	(1.041E-24)	(5.747E-21)	(1.853E-11)
	10	0.4174	0.4222	0.4135	0.3297	0.1765	0.0408
		(3.292E-13)	(2.262E-10)	(4.210E-06)	(1.956E-23)	(4.850E-19)	(5.604E-09)
	20	0.3895	0.4044	0.3974	0.2838	0.0861	-0.1597
		(3.013E-15)	(2.065E-12)	(1.171E-07)	(1.568E-24)	(1.835E-14)	(5.615E-10)
	all	0.3473	0.3540	0.3236	**0.4095**	**0.2544**	-0.0095
		(5.470E-15)	(1.796E-12)	(4.216E-07)	(1.039E-23)	(1.597E-19)	(4.046E-11)

Table 9. Weekly Forecasting Performance with Uncertainty Indices. R^2_{oos} scores for entire period and subperiod ($t \in \{10, 5, 3, 2, 1\}$) are represented. The numbers in parentheses indicate the p-values of residual stationarity. The bold numbers are the best performance measurements for each subperiod.

Uncertainty Indices Furthermore, we find that the use of uncertainty indices can push the accuracy of models even higher (Table 7, 8, and 9). The results vary slightly by model, with ElasticNet, DTR, and RNN seeing performance improvements by adding uncertainty features in about half of the cases, while LASSO, RFR, and ANN improve in most cases.

There are also significant differences based on the forecasting horizon. For most models, the shorter the forecast horizon, the better the use of uncertainty indices. This can be explained by the fact that the uncertainty indices reduce overfitting and are more influential in dealing with noise in short-term forecasts.

In many cases, adding uncertainty data to the training data set is beneficial. However, we could not guarantee that all uncertainty data has high explanatory power because many of them are likely to be pruned during the feature selection process. Even so, the highest performer of each model shows a higher performance when uncertainty indices join the training data set.

Feature Selection

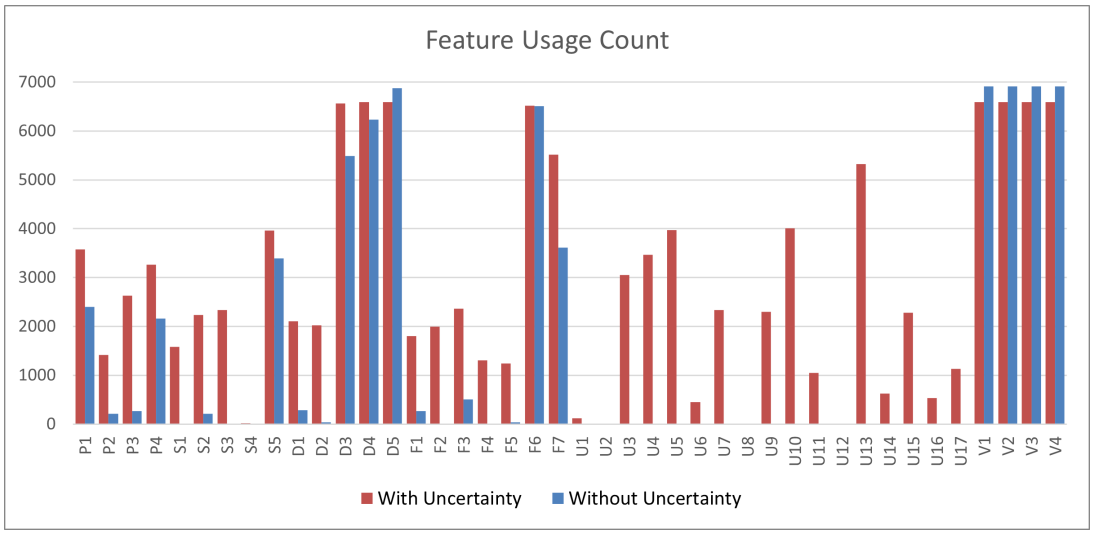

Figure 2. Feature Usage Count The feature usage was aggregated for the best performance models for each time horizon that selected 10 or 20 features, excluding cases where the entire set of features was selected. Each feature is assigned a number in Table 1 based on the first letter of each factor group (P, S, D, F, U, V) and in order from top to bottom (except for WTI future price, which is used as a dependent variable), e.g., WTI Spot price is equal to P2, PPI in US is equal to D4, etc.

We examine the feature usage in the models. First, we defined the best performance model as the feature selection methodology that performed the best in each model for

each time horizon being predicted. Then, we counted and summarized the usage of the features utilized by each best model when predicting each time point. Also, in order to clearly see the difference in the number of features selected, we excluded all cases where all features were selected and only kept cases where 10 or 20 features were selected. The result is presented in Figure 2. Each feature is assigned a number in Table 1 based on the first letter of each factor group (P, S, D, F, U, V) and in order from top to bottom (except for WTI future price, which is used as a dependent variable). For example, WTI Spot price is equal to P2, PPI in US is equal to D4.

Looking at the Figure 2, we can see that several features are highly selected, both with and without uncertainty features. First of all, the volatility indices(V1 to 4) are all highly selected, with PPI in China, US, and EU(D3 to 5) and federal funds rate and MSCI World Standard Index(F6,7) being the most selected. Next, WTI spot prices and NGL spot prices(P1,4) and capacity utilization rate(S5) are being selected. For the uncertainty features, US economic policy uncertainty in financial regulation, monetary policy, national security, and economic policy uncertainty(U13,5,10,4) are mostly selected. As in Degiannakis and Filis (2022) and Delis et al. (2023), the implied volatility indices are important in forecasting the volatility. Also, as the oil is one of the most important commodity in production, the PPI plays significant role. Federal funds rate and MSCI World Index represents the global economy, which is important for the volatility.

On the other hand, total OPEC production capacity(S4) and Brent oil spot prices(P2), US/UK, China/UK foreign exchange rate(F4,5) failed to being selected. Also, Global economic policy uncertainty, US economic policy uncertainty in fiscal policy, government spending, regulation, trade policy, and World uncertainty index(U1,2,6,8,12,14,16) are not selected.

Interestingly, there were features that were selected in the models with uncertainty but not in the models without uncertainty. These features are all price features(P1 to 4), global crude oil production, stock, and export(S1 to 3), MSCI World Standard Index(F7), and global crude oil import and liquid fuels consumption(D1,2). These differences are expected to have contributed to the difference in performance between the two models.

Potential Explanation Now, two main question rises. Why the performance orders are different for each forecast horizon? How do some machine learning models outperform the linear model? To answer the first question, we first see the performance of lagged value, in other words, the No-change model. The no-change model calculates the R^2 score of only the first lagged value of the target time series. Therefore, it calculates the effect of autocorrelation of lag 1. For many cases, the No-change model's R^2 is positive, which means the autocorrelation's effect is not negligible. Because the models with different forecast horizons are exposed to the different levels of autocorrelation effect, the performance orders could change.

The answer to the second question can come from many points, but the assumption of Gaussian distribution seems most restricted to the linear models. Our linear model assumes that the residual follows the normal distribution. However, because the model is linear, the real-world fat-tailed data may not fit the linear model. The nonlinear model, on the other hand, has the potential to explain such fat-tailed data. Another reason is the in-sample over-fitting. For example, the decision tree model is notorious for its tendency to over-fit. To overcome that issue, researchers develop a method to reduce the shortcoming and one way is the random forest, one of our best performers. Last but not least, the model may not converge well. Data is not always clean. Such ill-posed data can hurt the model during the training stage which finds the optimal weights to minimize the error.

4.3. *Discussion and Possible Extensions*

Volatility is one of the key variables for trading strategies, asset allocation, and risk management. Now we introduce the quantitative analysis suggested by previous studies on the usefulness of volatility forecasting. One way to express the quantitative advantage for accurate future volatility is the certainty equivalent return (CER). Yin and Yang (2016), Ma et al. (2018), Zhang et al. (2018), and Zhang et al. (2019) show that the mean-variance utility investor can benefit from forecasting future volatility in terms of CER. When the mean-variance utility investor considers the crude oil future and

risk-free asset, the portfolio return is

$$R_p = wl(r + r_f) + (1 - w)r_f$$

where w is the portfolio weight of crude oil futures, l is the leverage ratio, r is the excess return, and r_f is the risk-free rate. Then, CER of the investor is:

$$CER = \bar{R}_p - \frac{1}{2}\gamma\sigma_p^2$$

where $\bar{R}_p$ and σ_p^2 are the mean and the variance of the portfolio return R_p over the out-of-sample period. It is straightforward that a more accurate prediction yields a higher CER.

Another way to emphasize the importance of a volatility forecast is the Sharpe ratio of the simple directional trading strategy. Following Easley et al. (2020), consider n independent and identically distributed bets, where the outcome y of a bet with a profit $\pi > 0$ with probability $P[y = \pi] = p$ and a loss π with probability $1 - p$. The expected profit is $E[y] = \pi(2p - 1)$, and the variance is $Var[y] = 4\pi^p(1 - p)$. Then, the Sharpe ratio is:

$$\theta(p, n) = \frac{nE[y]}{\sqrt{nVar[y]}} = \frac{2p - 1}{2\sqrt{p(1 - p)}}\sqrt{n}$$

Easley et al. (2020) stated that repeated trading with sufficiently large enough trials ($n = 13{,}000$) can achieve the Sharpe ratio 2.04 with the small enhancement of directional forecast $p = 0.52$. However, our models achieve far more accurate predictability around 0.6 to 0.7. Therefore, even smaller trials $n = 100$ can get a higher Sharpe ratio of 4.36 with $p = 0.7$.

The bet above is to analyze the advantage of accurate directional forecasting. However, in practice, we can consider the investment in the volatility ETFs such as OVX or trading strategies based on volatility. Delis et al. (2022) introduces a simple strategy for volatility trading. In the case where the s-step forecasted oil price volatility of the model i at time t (denoted as $\widehat{RV}_{t+s|t}$) is higher than that of the actual volatility at time

t (denoted as RV_t), the trader will take a long position in an asset that exhibits similar performance to the oil realized volatility, such as OVX. Conversely, if the forecasted volatility of model i at time $t+s$ is lower than that of the actual volatility at time t, the trader will take a short position. Consequently, the cumulative returns of model i, r^i over the out-of-sample forecasting period 1 to t_N can be calculated as follows:

$$r^i = \sum_{t=1}^{t_{N-s}} \frac{(RV_{t+s} - RV_t)d_t^i}{RV_t}$$

where $d_t^i = 1$ if $\widehat{RV}_{t+s|t} > RV_t$, and $d_t^i = -1$ if $\widehat{RV}_{t+s|t} \leq RV_t$.

Over the past 11 years, HAR has generated cumulative returns of 1,589%, while LASSO and RNN have generated cumulative returns of 2,448% and 2,355% over the same period. It is clear that the machine learning models outperform the benchmark HAR, with LASSO demonstrating the most impressive performance overall. However, for RNN, which exhibited superior results in the recent period, the cumulative return increase exhibits a steeper slope than LASSO, and this period also encompasses the COVID-19 global pandemic, suggesting that RNN models are also significant. We showed the cumulative returns of LASSO, RNN, and benchmark HAR in Figure 3 below.

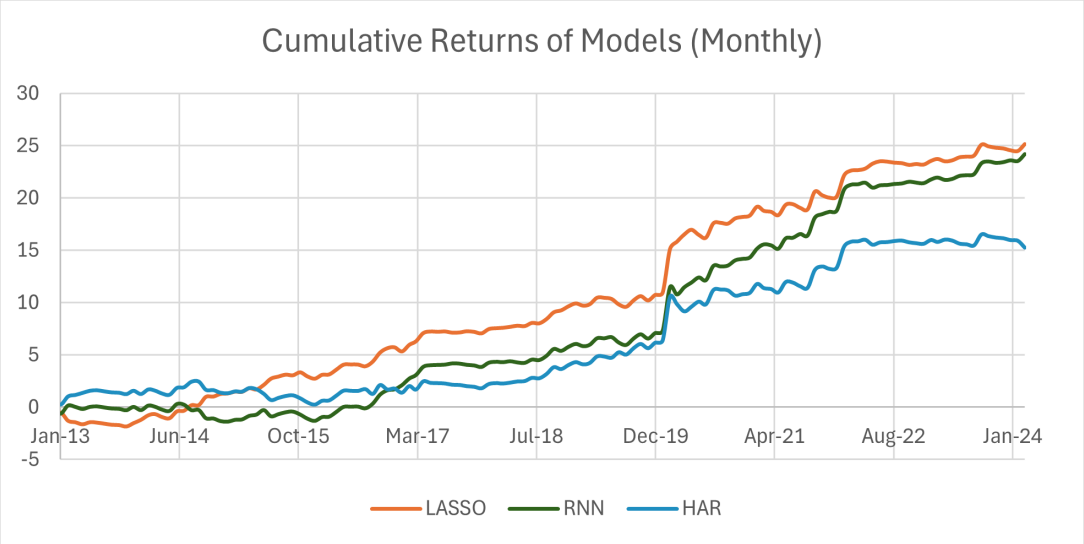

Figure 3. Cumulative Returns of LASSO, RNN, and HAR The forecasting horizon is monthly, and both models contain uncertainty factors. Furthermore, both models utilize all features selected through feature selection.

The preceding examples illustrate the effectiveness of a precise volatility forecast

in various contexts. This paper demonstrates that integrating representative machine learning models with the conventional HAR model can significantly enhance forecasting performance.

Going further, we might be able to improve performance through various detailed methodologies for time-series forecasting. There can be a variety of improvements - for instance, combining multiple forecasting models such as dynamic model averaging, dynamic model selection, or other ensemble methods (Wang et al. (2016); Audrino and Knaus (2016); Ding (2018); Zhang et al. (2018)), improving feature selection methods (Ghadimi et al. (2018); Karasu et al. (2020)), and reducing the noise of time series data using bilateral filters, wavelet denoising, autoencoder, etc (Uddin et al. (2019); Wang et al. (2017)).

5. Conclusion

This paper compares and analyzes the predictability of the realized volatility of crude oil future prices with various forecasting models from April 2002 to April 2024 including the Great Recession and COVID-19. With a wide variety of analyses and comparisons, we can explore the possibility of the potential usage of machine learning models in the field of volatility forecasting. A large set of explanatory variables is considered, and each explanatory variable belongs to one of six groups -prices, supply, demand, financial, implied volatility indices, and uncertainty factors. The walk-forward cross-validation shows the out-of-sample forecasting performance of ten types of forecasting models - HAR, TV-HAR, HAR-X, LASSO, ElasticNet, DTR, RFR, GBR, ANN, and RNN.

Although HAR is a popular conventional model in forecasting the realized volatility, the performance decreases in out-of-sample. Various machine learning models with the momentum factors of the HAR were tested, and some of them have significantly outperformed out-of-sample forecasts, such as the RFR and RNN.

Additionally, an investigation was conducted into the features selected in the models, which revealed that the implied volatility indices and uncertainty factors were the most frequently chosen. Furthermore, the PPI in China, the US, and the EU, the federal funds rate, and the MSCI World Standard Index were identified as the most frequently

selected variables.

Diverse previous literature has dealt with the advantages of accurate volatility forecasting. Volatility is one of the key variables for trading strategies, asset allocation, and risk management. Therefore, future works to enhance the predictive power are worthy with numerous potential extensions and applications.

Disclosure statement

There is no relevant interest to declare.

Funding

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea.(NRF-2022S1A3A2A02089950)

References

Ahir, H., N. Bloom, and D. Furceri (2018). The world uncertainty index. *Working Paper, Available at SSRN 3275033*.

Apergis, N. and S. M. Miller (2009). Do structural oil-market shocks affect stock prices? *Energy economics 31*(4), 569–575.

Audrino, F. and S. D. Knaus (2016). Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews 35*(8-10), 1485–1521.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

Baker, S. R., N. Bloom, S. J. Davis, and K. J. Kost (2019). Policy news and stock market volatility. Technical report, National Bureau of Economic Research.

Baker, S. R., N. Bloom, S. J. Davis, K. J. Kost, M. C. Sammon, and T. Viratyosin (2020). The unprecedented stock market impact of covid-19. Technical report, National Bureau of Economic Research.

Brown, S. P. and M. K. Yucel (2008). What drives natural gas prices? *The Energy Journal 29*(2).

Caggiano, G., E. Castelnuovo, and R. Kima (2020). The global effects of covid-19-induced uncertainty. *Economics Letters 194*(109392).

Christiansen, C., M. Schmeling, and A. Schrimpf (2012). A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics 27*(6), 956–977.

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

Davis, S. J. (2016). An index of global economic policy uncertainty. Technical report, National Bureau of Economic Research.

Degiannakis, S. and G. Filis (2022). Oil price volatility forecasts: What do investors need to know? *Journal of International Money and Finance 123*, 102594.

Delis, P., S. Degiannakis, and G. Filis (2022). What matters when developing oil price volatility forecasting frameworks? *Journal of Forecasting 41*(2), 361–382.

Delis, P., S. Degiannakis, and K. Giannopoulos (2023). What should be taken into consideration when forecasting oil implied volatility index? *The Energy Journal 44*(5), 231–250.

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

Ding, Y. (2018). A novel decompose-ensemble methodology with aic-ann approach for crude

oil forecasting. *Energy 154*, 328–336.

Easley, D., M. López de Prado, M. O'Hara, and Z. Zhang (2020, 07). Microstructure in the Machine Age. *The Review of Financial Studies 34*(7), 3316–3363.

Ghadimi, N., A. Akbarimajd, H. Shayeghi, and O. Abedinia (2018). Two stage forecast engine with feature selection technique and improved meta-heuristic algorithm for electricity load forecasting. *Energy 161*, 130–142.

Ghoddusi, H., G. G. Creamer, and N. Rafizadeh (2019). Machine learning in energy economics and finance: A review. *Energy Economics 81*, 709–727.

Goodfellow, I. (2016). *Deep learning*, Volume 196. MIT press.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Guo, B. and H. Lin (2020). Volatility and jump risk in option returns. *Journal of Futures Markets 40*(11), 1767–1792.

Hallock Jr, J. L., P. J. Tharakan, C. A. Hall, M. Jefferson, and W. Wu (2004). Forecasting the limits to the availability and diversity of global conventional oil supply. *Energy 29*(11), 1673–1696.

Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007-08. Technical report, National Bureau of Economic Research.

Hammoudeh, S. M., B. T. Ewing, and M. A. Thompson (2008). Threshold cointegration analysis of crude oil benchmarks. *The Energy Journal 29*(4).

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting 13*(2), 281–291.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Herrera, A. M., M. B. Karaki, and S. K. Rangaraju (2019). Oil price shocks and us economic activity. *Energy Policy 129*, 89–99.

Karasu, S., A. Altan, S. Bekiros, and W. Ahmad (2020). A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. *Energy 212*, 118750.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review 99*(3), 1053–69.

Kilian, L. and B. Hicks (2013). Did unexpectedly strong economic growth cause the oil price shock of 2003–2008? *Journal of Forecasting 32*(5), 385–394.

Ma, F., J. Liu, M. Wahab, and Y. Zhang (2018). Forecasting the aggregate oil price volatility in a data-rich environment. *Economic Modelling 72*, 320–332.

Ma, F., M. Wahab, D. Huang, and W. Xu (2017). Forecasting the realized volatility of the oil futures market: A regime switching approach. *Energy Economics 67*, 136–145.

Miao, H., S. Ramchander, T. Wang, and D. Yang (2017). Influential factors in crude oil price forecasting. *Energy Economics 68*, 77–88.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*(85), 2825–2830.

Schwager, J. D. (2017). *A Complete Guide to the Futures Market: Technical Analysis, Trading Systems, Fundamental Analysis, Options, Spreads, and Trading Principles*. John Wiley & Sons.

Smith, T. G. et al. (2017). pmdarima: Arima estimators for Python. [Online; accessed ¡today¿].

Uddin, G. S., R. Gençay, S. Bekiros, and M. Sahamkhadam (2019). Enhancing the predictability of crude oil markets with hybrid wavelet approaches. *Economics Letters 182*, 50–54.

Wang, L., Z. Zhang, and J. Chen (2017). Short-term electricity price forecasting with stacked denoising autoencoders. *IEEE Transactions on Power Systems 32*(4), 2673–2681.

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.

Wang, Y., Y. Wei, C. Wu, and L. Yin (2018). Oil and the short-term predictability of stock return volatility. *Journal of Empirical Finance 47*, 90–104.

Wei, Y., J. Liu, X. Lai, and Y. Hu (2017). Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? *Energy Economics 68*, 141–150.

Yin, L. and Q. Yang (2016). Predicting the oil prices: Do technical indicators help? *Energy Economics 56*, 338–350.

Zhang, Y., F. Ma, B. Shi, and D. Huang (2018). Forecasting the prices of crude oil: An iterated combination approach. *Energy Economics 70*, 472–483.

Zhang, Y., Y. Wei, Y. Zhang, and D. Jin (2019). Forecasting oil price volatility: Forecast combination versus shrinkage method. *Energy Economics 80*, 423–433.

Appendix A. Data Description

Rolling Return Strategy

We forecast the monthly volatility for futures contracts, nearest to maturity. Because each future contract must be exercised at the maturity date, to estimate continuous volatility, daily future prices are calculated based on the roll-over rule.[12] On each monthly roll-over day, the oil price is adjusted by the price difference between the future contract with the first nearest maturity and the future contract with the second nearest maturity. The monthly volatility of oil price is calculated by the standard deviation of daily WTI prices at each last business day of the month. A large body of literature investigated the forecasting oil price volatility using historical realized volatility. For example, Ma et al. (2017), Ma et al. (2018), and Wei et al. (2017) studied forecasting models of realized oil volatility.

Data Preprocessing

Data needs preprocessing procedure for distinct frequency, stationarity, and integration before analysis. If the date of the publication of infrequent data is not specified, we assume that it is announced at the last moment of the month (or the quarter).[13] For example, the Fed Funds rate or MSCI data are monthly, but their release dates are specified. Therefore, we use announced data from the exact date that they are announced and fill missing values by using them until the next announcement date. When the announced date is not specified (such as categorical EPU data), we use them from one day after its announced month (or quarter). Each categorical data, such as oil production, consumption, and export, are country-level data. They are summed to make a global value of each categorical data. For instance, country-specific oil production data are aggregated to one global oil production.

The stationarity of all data from April 2002 to April 2024 is checked using Python's pmdarima package (Smith et al., 2017) except the lagged volatility variable, the mean of lagged volatility, and the dependent variable itself. If we have to take the

[12]Each contract expires on the third business day before the 25th calendar day. When the 25th calendar day is not a business day, the contract expires on the third business day before the latest business day prior to the 25th calendar day.

[13]Capacity utilization rate is the only weekly disclosed data, and each date of the announcement is recorded.

difference of each sample more than twice to meet the criteria of stationarity (such as adjusted Dicky-Fuller test), we set them to be 2. Even if high order differentiation may ensure the stationarity, we believe those process will omit valuable information.

Appendix B. Training Models

Our forecasting models consist of two parts: feature selection and training. We first select features for each sample using the f-regression method in Python's scikit-learn package (Pedregosa et al., 2011). As different sample has different characteristics, there is no guarantee that all samples use the same features. Feature selection helps models train in a reasonable time and helps prevent the notorious issue of over-fitting. After the aforementioned transformation, we fit our data to various models. Because there could be randomness during training models, scikit-learn models' random-states are fixed to a specific number. Without this, the model will predict different values each time we train.

As the optimal predictive model may vary depending on the sample, the optimal hyperparameters may also vary. Therefore, for each training sample, different hyperparameters may be used for each model. Scikit-learn's GridSearchCV module is an effective tool for identifying these optimal hyperparameters, and we have adopted it for this purpose.

Appendix C. Methods of Empirical Analysis

Diebold-Marino Test

The Diebold-Marino (DM) test is a method of comparing the performance of two or more forecasting models, introduced by Diebold and Mariano (2002) and modified by Harvey et al. (1997). The DM test compares the forecasting errors of two models and statistically verifies whether one model **overpredicts or underpredicts** compared to other models. When $g(e_{i,t})$ is the squared-error loss or the absolute error loss of the

forecasting error $e_{i,t} = \hat{y}_{i,t} - y_{i,t}$ for model $i = 1, 2$, the loss differential is defined by $d_t = g(e_{1,t}) - g(e_{2,t})$. The null hypothesis indicates that prediction accuracy for two models are same. The DM statistics is defined as $DM = \frac{\bar{d}}{\sqrt{\hat{\gamma}_0/h}} \sim N(0,1)$, where $\bar{d} = \sum_{t=1}^{T} d_t/h$ is the sample mean with h-step forward forecast, $\bar{\gamma_0}$ is the consistent estimate of the variance for $h\bar{d}$.

$$H_0 : E[d_t] = 0, \quad H_1 : E[d_t] \neq 0, \quad \forall t$$