Received: 19 October 2022 | Revised: 11 April 2023 | Accepted: 2 November 2023

DOI: 10.1002/for.3046

RESEARCH ARTICLE

WILEY

Out-of-sample volatility prediction: Rolling window, expanding window, or both?

Yuqing Feng | **Yaojie Zhang** | **Yudong Wang**

School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

Correspondence
Yaojie Zhang, School of Economics and Management, Nanjing University of Science and Technology, Xiaolingwei 200, Xuanwu District, Nanjing 210094, China.
Email: yaojie_zhang@126.com

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 72001110, 72371131

Abstract

Estimation windows, either rolling or expanding, are used for volatility forecasting. In this study, we propose a new approach relying on both estimation windows. Our method is based on how well these two windows performed in terms of prediction during a recent period of past time. We will continue to use whichever one has performed better in the past. Results show that our strategy significantly outperforms the individual and mean combination models. Whether the window is rolling or expanding, the relatively better performance is persistent. In other words, we document the existence of the momentum of predictability (MoP). A mean–variance investor can achieve the highest utility gains using our strategy for volatility forecasting. Moreover, the results pass a series of robustness tests.

KEYWORDS

expanding window, model switching, portfolio exercise, rolling window, volatility forecasting

1 | INTRODUCTION

Volatility forecasting plays a crucial role in asset pricing and risk management. Obtaining accurate volatility forecasts is important for regulators and financial market participants. The realized volatility (RV) proposed by Andersen and Bollerslev (1998) is the sum of the squared intraday (high-frequency) returns. RV uses intraday information and is much less noisy than traditional volatility measures, which is very helpful for improving the accuracy of volatility forecasting. So far, many literatures have proposed a series of volatility forecasting models based on RV, such as the heterogeneous autoregressive RV (HAR-RV) proposed by Corsi (2009), the HAR-RV-J and HAR-RV-CJ models formed by adding jump components that are proposed by Andersen et al. (2007) and so forth. However, one of the core issues with these models is the selection of estimation windows.

The rolling and expanding windows are popular for forecasting volatility. The problem of structural changes in the data can be efficiently handled by the rolling window, which uses the last observations to estimate parameters. In comparison, the expanding window effectively employs the information from all observations, using all available data. As each has advantages of its own, different papers have selected either of the two windows. For instance, the finance literature tends to construct forecasts using a rolling window (see, e.g., Bollerslev et al., 2016; Degiannakis & Filis, 2017; Ma et al., 2019; Zhang et al., 2020; He et al., 2021). Whereas the macroeconomics literature generally uses an expanding window for estimating parameters (see, e.g., Stock & Watson, 2003; Stock & Watson, 2007; Schrimpf & Qingwei, 2010; Gillitzer & McCarthy, 2019). Hence, which window is more appropriate seems to be an open question. To answer this important question, we propose a new model switch method based on the two estimation windows.

The switching behavior is based on how well each model performed in terms of prediction in the recent past. We will continue to choose a rolling window to predict the stock market's volatility in the near future if it performs better at predicting the recent past than an expanding window does. Otherwise, we will choose an expanding window to generate the realized variance (RV) forecasts. Our motivation is from the momentum of predictability (MoP), proposed by Wang et al. (2018), where the forecasting performance of some predictor regression is persistent in stock returns. Moreover, a stylized fact is that volatility is quite persistent. Therefore, the MoP strategy can be applied to volatility forecasts.

In our MoP strategy, with the purpose of evaluating the model's forecasting performance, we use three loss functions. Specifically, the loss functions are quasi-likelihood (QLIKE), mean square error (MSE), and mean absolute error (MAE) loss functions, respectively.

We use the equal-weight average combination as a competing model to rule out that the forecast combination is to blame for the success of our MoP strategy. With the intention of evaluating the out-of-sample performance, we apply the model confidence set (MCS) proposed by Hansen et al. (2011). We discover that, compared with the three competing models—with a rolling window, an expanding window, or a mean combination—our MoP strategy generates stronger predictive ability.

The MoP must exist for the MoP method to be successful. Using the Pesaran and Timmermann (2009) test, we demonstrate the existence of MoP in rolling and expanding windows of stock market volatility forecasts. The PT test's null hypothesis is that past and present performance are not dependent. Our results reject the null hypothesis. In other words, the MoP test shows a high degree of dependence between past and present performance. The existence of MoP proves that if a model using the rolling window can produce a more accurate RV forecast than the one using the expanding window over a recent past period, then this model can continue to produce a better RV forecast in the near future.

We test the economic significance of our MoP model through a portfolio exercise. Compared with the competing models, a mean–variance investor can apply our MoP strategy to allocate her assets between stocks and risk-free bills, which can help her obtain the highest utility gains.

We further provide a series of robustness tests. Our results are robust for four different look-back periods, alternative window sizes, an additional benchmark model, alternative volatility estimators, and the logarithmic HAR-RV model. We also divide the entire prediction period into different groups based on three grouping criteria, and the results still show that our MoP strategy is superior to the three competing models. The reason may be that, according to past performance, our strategy is able to retain better-performing forecasts in the corresponding period and discard the underperforming ones.

Finally, we extend our MoP model to the crude oil market. That is, we predict the crude oil market RV by employing the MoP model and the competing models and observe consistent results in the stock market. This case greatly reduces the risk of data mining.

Our paper closes to the related literature on window selection. If parameter breaks are believed to be either extremely rare or modest in margin, then the common approach is to use an expanding window. From another aspect, when the regression model's parameters are not considered constant over time, then predictions are usually constructed using a rolling window (Pesaran & Timmermann, 2002). When estimating linear regression forecasting models, Pesaran and Timmermann (2007) argue that the use of data before pre-break reduces the error in parameter estimation. Rapach et al. (2009) use the GJR-GARCH (1,1) model for portfolio forecasts of stock return volatility. They discover that portfolio forecasts with different estimation window sizes improve volatility forecasts under structural breaks. Inoue et al. (2017) propose a new method to choose the window size in the prediction of models with potential breaks, where the parameters of the model are set as a smooth function of time. All of the above literature considers predictions using only one estimation window. However, Clark and McCracken (2009) find that, compared with forecasts made using the rolling or expanding scheme, combining rolling and expanding forecasts improves forecast accuracy. By contrast, our paper contributes to their study by providing a model with a switching mechanism that enables the model to alternate between using rolling and expanding windows. Our model can make full use of the advantages of both windows, thereby improving the predictive ability of the model.

Additionally, our paper contributes to the relevant literature on the MoP (see, e.g., Wang et al., 2018; Zhang et al., 2019; Zhang et al., 2022). Wang et al. (2018) discover that if the performance of some predictor regression outperforms the historical average benchmark over a recent past period, it will continue to improve in the near future, a phenomenon they call MoP. Zhang et al. (2019) propose a new mixed-frequency model to forecast volatility that is based on the GARCH-class and the HAR-RV-type models. They prove the existence of MoP in terms of prediction frequency. Zhang et al. (2022) propose the "momentum of jumps." Their strategy enables the model to alternate between the HAR-RV model and the HAR-J model with a jump component. Our study presents a fresh analysis that focuses on the model's performance using two estimating windows in predicting stock market volatility. This is important and meaningful because determining which window is more appropriate for volatility forecasting seems to be a crucial question. Different window selections may lead to different out-of-sample evaluation results. In this research, we notice that using alternating rolling and expanding windows can considerably improve the accuracy of the volatility forecasting.

The remainder of the paper is organized as follows. Section 2 introduces the forecasting methodology. Section 3 describes the sources of data. Section 4 presents the out-of-sample analyses. Section 5 considers a series of robustness checks. Section 6 concludes.

2 | METHODOLOGY

2.1 | Realized volatility measure

According to Andersen and Bollerslev (1998), the daily RV can be calculated as the sum of squared intraday returns. Specifically, we calculate the RV of trading day t as follows:

$$RV_t = \sum_{i=1}^{N} r_{i,t}^2, \tag{1}$$

where $r_{i,t}$ is the ith intraday stock market return for day t. $N = 1/\Delta$, Δ is the sampling frequency.

2.2 | HAR-RV models

The HAR-RV model, pioneered by Corsi (2009), is the most popular volatility benchmark model. It takes into account some stylized facts in asset return volatility, like long memory and multi-scaling behavior. In addition, since there are only three predictors needed, it is easy to implement. The model can be shown as follows:

$$RV_{t+1:t+h} = \varphi_0 + \beta_d RV_t + \beta_w RV_{t-4:t} + \beta_m RV_{t-21:t} + \varepsilon_{t+1:t+h}, \tag{2}$$

where $RV_{t-h:t-1} = \left(\frac{1}{h}\right)(RV_{t-h} + \cdots + RV_{t-1})$. Particularly, RV_t, $RV_{t-4:t}$, and $RV_{t-21:t}$ denote the daily, weekly, and monthly RVs, respectively.

2.3 | MoP

Because of the structural change in daily data, a rolling window is typically used in the daily RV forecast, whereas the number of monthly data is usually small, and an expanding window is typically used for forecasting. However, as each has advantages of its own, we propose a model selection method between rolling and expanding windows.

Our MoP strategy switches between the models with the rolling and expanding windows based on how well they have performed historically. In our case, two RV forecasting strands from HAR-RV models will be provided, each with a rolling or an expanding window separately. Then, we consistently use the volatility model, which has had relatively good predictive performance in the past. Following Wang et al. (2018), Zhang et al. (2019), and Zhang et al. (2019), we evaluate the past performance of the HAR-RV model with a rolling window or one with an expanding window. The specific method is as follows:

$$pp_{t+1:t+h}(k) = I\left(\sum_{i=t-h-k+1}^{t-h}\left(RV_{i+1:i+h} - \widehat{RV}_{i+1:i+h}^{rolling}\right)^2 - \sum_{i=t-h-k+1}^{t-h}\left(RV_{i+1:i+h} - \widehat{RV}_{i+1:i+h}^{expanding}\right)^2 < 0\right), \tag{3}$$

where k denotes the length of the look-back period. We consider $k = 1, 5, 10, 22$. Moreover, $I(\cdot)$ represents an indicator function, and $RV_{i+1:i+h}$ is the true RV on days $i+1{:}i+h$. $\widehat{RV}_{i+1:i+h}^{rolling}$ and $\widehat{RV}_{i+1:i+h}^{expanding}$ are the HAR-RV with a rolling window forecast and the HAR-RV with an expanding window forecast, respectively, for $RV_{i+1:i+h}$.

On the basis of the recent past performance of $pp_{t+1:t+h}(k)$, we construct the MoP forecast in the following way:

$$\widehat{RV}_{\mathrm{t+1:t+h}}^{MoP}(k) = \begin{cases} \widehat{RV}_{t+1:t+h}^{rolling}, if\ pp_{t+1:t+h}(k) = 1 \\ \widehat{RV}_{t+1:t+h}^{expanding}, if\ pp_{t+1:t+h}(k) = 0 \end{cases}. \tag{4}$$

Meanwhile, we use the equal-weight combination forecast as a competing model, which can be computed as

$$\widehat{RV}_{t+1:t+h}^{AVG} = \frac{1}{2}\left(\widehat{RV}_{t+1:t+h}^{rolling} + \widehat{RV}_{t+1:t+h}^{expanding}\right) \tag{5}$$

Because of the well-known "prediction combination problem," which states that no complicated combination forecasts can outperform the mean combination forecast (see, e.g., Stock & Watson, 2004; Rapach et al., 2010; Zhang et al., 2019), we do not take into account any more complicated weighting schemes.

3 | DATA

Because the 5-min RV is frequently used and advised by numerous studies (see, e.g., Andersen et al., 2007; Haugom et al., 2014; Wang et al., 2016; Zhang et al., 2019; He et al., 2022), we choose this interval as our sampling frequency. Furthermore, Liu et al. (2015) document that the 5-min RV is hardly to be surpassed by volatility measures from any other financial assets or estimators.

We access the Oxford-Man Institute's Quantitative Finance Realized Library[1] to get the 5-min RV of the S&P 500 index. The entire sample period includes 5555 observations between February 2, 2000, and March 31, 2022. Moreover, we produce the most recent 2500 volatility forecasts for out-of-sample evaluation.

4 | OUT-OF-SAMPLE ANALYSES

4.1 | Evaluation framework

We quantitatively evaluate the out-of-sample predictive ability of several volatility forecasting models using three loss functions. Specifically, they are QLIKE, MSE, and MAE loss functions, and their statistical expressions are as follows:

$$QLIKE : L\left(\widehat{RV}_{t+1:t+h}, RV_{t+1:t+h}\right) = log\left(\widehat{RV}_{t+1:t+h}\right) + \frac{RV_{t+1:t+h}}{\widehat{RV}_{t+1:t+h}}, \tag{6}$$

$$MSE : L\left(\widehat{RV}_{t+1:t+h}, RV_{t+1:t+h}\right) = \left(\widehat{RV}_{t+1:t+h} - RV_{t+1:t+h}\right)^2, \tag{7}$$

$$MAE : L\left(\widehat{RV}_{t+1:t+h}, RV_{t+1:t+h}\right) = \left|\widehat{RV}_{t+1:t+h} - RV_{t+1:t+h}\right|, \tag{8}$$

where $RV_{t+1:t+h}$ is the actual RV for days $t+1:t+h$, $\widehat{RV}_{t+1:t+h}$ represents the forecast RV provided by a predictive model. Patton (2011) demonstrates that both QLIKE and MSE can withstand noise in the volatility proxy.

In addition to the loss functions, the MCS, proposed by Hansen et al. (2011), is frequently applied to evaluate how well various volatility forecasting models perform (see, e.g., Patton & Sheppard, 2009; Laurent et al., 2012; Liu et al., 2015; Wang et al., 2016; Wei et al., 2017; Ma et al., 2018; Zhang et al., 2019). Therefore, we assess different models' predictive power using the MCS test.

Given a confidence level, MCS is a collection of forecasting models that contains the optimal models. The predictive ability of the relevant model is thought to be stronger when the MCS p-value is higher. Generally, we take the confidence level as 90% (see, e.g., Hansen et al., 2011; Laurent et al., 2012; Wang et al., 2016; Gong & Lin, 2018; Ma et al., 2019; He et al., 2022). When the p-value is greater than 10%, the model will be contained in the MCS.

4.2 | Forecasting performance

Table 1 presents the results of the MCS test, which is conducted based on the case of a weekly look-back period (i.e., $k = 5$). Our MoP model always produces p-values greater than 10%. In other words, our MoP model is contained in the MCS all the time. By comparison, almost all three competing models fail to produce p-values greater than 10%, except for the 1-day forecast horizon, which indicates the inability of the associated models to meet the MCS at the 10% significance level. In summary, the MCS p-values show that the MoP model outperforms the competing HAR-RV models with a rolling or expanding window and mean combination in terms of forecasting performance.

4.3 | Testing the MoP

The existence of MoP is essential to the success of our MoP strategy. Therefore, we examine whether better past forecasting performance is usually followed by better future forecasting performance. In particular, the future predictive performance over days $t+1:t+h$ is given as follows:

$$fp_{t+1:t+h} = I\left(\left(RV_{t+1:t+h} - \widehat{RV}^{rolling}_{t+1:t+h}\right)^2 - \left(RV_{t+1:t+h} - \widehat{RV}^{expanding}_{t+1:t+h}\right)^2 < 0\right). \tag{9}$$

The dependence between $pp_{t+1:t+h}(k)$ and $fp_{t+1:t+h}(k)$ implies the existence of MoP. According to Wang et al. (2018), Zhang et al. (2019), and Zhang et al. (2022), we employ the chi-square statistic proposed by Pesaran and Timmermann (2009) to test for cross-dependence, in which the null hypothesis is that $pp_{t+1:t+h}(k)$ and

[1]https://realized.oxford-man.ox.ac.uk/

TABLE 1 Out-of-sample forecasting performance based on MCS test.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.002	**0.584**	**0.784**
Expanding	0.001	**1.000**	0.051
Mean	0.000	**0.867**	**0.395**
MoP	**1.000**	**0.560**	**1.000**
	Panel B: 5-day horizon		
Rolling	0.000	0.095	0.001
Expanding	0.000	0.095	0.001
Mean	0.000	0.095	0.001
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.125**	0.013
Expanding	0.002	**0.125**	0.013
Mean	0.002	**0.125**	0.013
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.046	**0.212**	0.015
Expanding	0.046	**0.212**	0.015
Mean	0.046	**0.212**	0.015
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS p-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k=5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.
The whole sample period includes 5555 observations between February 2, 2000, and March 31, 2022, whereas the out-of-sample forecast includes the most recent 2500 observations.

$fp_{t+1:t+h}(k)$ are cross-independence with self-dependence existing in each series. That is, if the null hypothesis is not accepted, we can statistically prove the existence of MoP.

Table 2 presents the p-values of the Pesaran and Timmermann (2009) statistics. We discover that all p-values are less than 0.01 in any of these cases. In other words, the assumption that $pp_{t+1:t+h}(k)$ and $fp_{t+1:t+h}(k)$ are independent is rejected as a null hypothesis at the 1% significance level, which proves the existence of MoP. This finding suggests that the model's past performance is always associated with its future performance.

To more intuitively represent the model switching between rolling and expanding windows, we draw the model selection results for different forecast horizons based on $k=5$ in Figure 1. First, we observe that our MoP model sometimes selects a rolling window and sometimes an expanding window, indicating that models using different windows cannot completely surpass each other. This evidence indicates that our MoP strategy selects relatively better models at different periods. Second, the choice of windows by the model is rather persistent. Our MoP strategy continues to select one of the models with a rolling or expanding window over a relatively long period of time. This evidence suggests that a model that outperformed competing models in the past tends to perform better in the future.

4.4 | Portfolio performance

After the MoP model passes the statistical test, we test its economic value through a portfolio exercise. Specifically, following Bollerslev et al. (2018), we suppose that a mean–variance investor will allocate her assets, with a fixed Sharpe ratio (SR hereafter), between a risky asset (i.e., stocks) and a risk-free asset (i.e., risk-free bills).

In the portfolio exercise, a mean–variance investor will invest a portion w_t of her present portfolio in stocks with a return of r_{t+1} and the remainder in risk-free bills, which will yield a return of r_t^f. She will thus receive the following returns on her portfolio:

$$r_{t+1}^p = w_t r_{t+1} + (1-w_t) r_t^f = w_t r_{t+1}^e + r_t^f, \tag{10}$$

where $r_{t+1}^e = r_{t+1} - r_t^f$.

The expected utility can be computed as follows:

$$U(w_t) = w_t E_t\left(r_{t+1}^e\right) - \frac{\gamma}{2} w_t^2 Var\left(r_{t+1}^e\right), \tag{11}$$

where γ is the mean–variance investor's risk aversion coefficient and $Var\left(r_{t+1}^e\right) = E_t(RV_{t+1})$. With the purpose of focusing on volatility forecasting, Bollerslev et al. (2018) propose the conditional SR, which can be measured as $SR = \frac{E_t\left(r_{t+1}^e\right)}{\sqrt{E_t(RV_{t+1})}}$ and be constant. Thus, the expected utility can be written as follows:

$$U(w_t) = w_t SR \sqrt{E_t(RV_{t+1})} - \frac{\gamma}{2} w_t^2 E_t(RV_{t+1}). \tag{12}$$

To reach the goal of maximum expected utility, the investor will distribute the stock weight to

TABLE 2 Testing results for the momentum of predictability.

Look-back periods	$h = 1$	$h = 5$	$h = 10$	$h = 22$
$k = 1$	0.000	0.000	0.000	0.000
$k = 5$	0.000	0.000	0.000	0.000
$k = 10$	0.000	0.000	0.000	0.000
$k = 22$	0.000	0.000	0.000	0.000

Note: The table displays the test results for the MoP. The MoP means that a model that outperforms competing models during a recent past period will also perform better in the near future. The future predictive performance over days $t+1{:}t+h$ is provided by

$$fp_{t+1:t+h} = I\left(\left(RV_{t+1:t+h} - \widehat{RV}^{rolling}_{t+1:t+h}\right)^2 - \left(RV_{t+1:t+h} - \widehat{RV}^{expanding}_{t+1:t+h}\right)^2 < 0\right)$$

where $I(\cdot)$ denotes an indicator function, $RV_{t+1:t+h}$ is the true RV on days $t+1{:}t+h$, and $\widehat{RV}^{rolling}_{t+1:t+h}$ and $\widehat{RV}^{expanding}_{t+1:t+h}$ are the HAR-RV with rolling window forecast and the HAR-RV with expanding window forecast, respectively, for $RV_{i+1:i+h}$. The past predictive performance is given by

$$pp_{t+1:t+h}(k) = I\left(\sum\nolimits_{i=t-h-k+1}^{t-h}\left(RV_{i+1:i+h} - \widehat{RV}^{rolling}_{i+1:i+h}\right)^2 - \sum\nolimits_{i=t-h-k+1}^{t-h}\left(RV_{i+1:i+h} - \widehat{RV}^{expanding}_{i+1:i+h}\right)^2 < 0\right)$$

where k denotes the length of the look-back period. The null hypothesis is tested using the chi-square statistic of Pesaran and Timmermann (2009), which states that $pp_{t+1:t+h}(k)$ and $fp_{t+1:t+h}(k)$ are cross-independence with self-dependence existing in each series. That is, if the null hypothesis is not accepted, we statistically prove the existence of MoP. Specifically, the relevant p-values are reported.

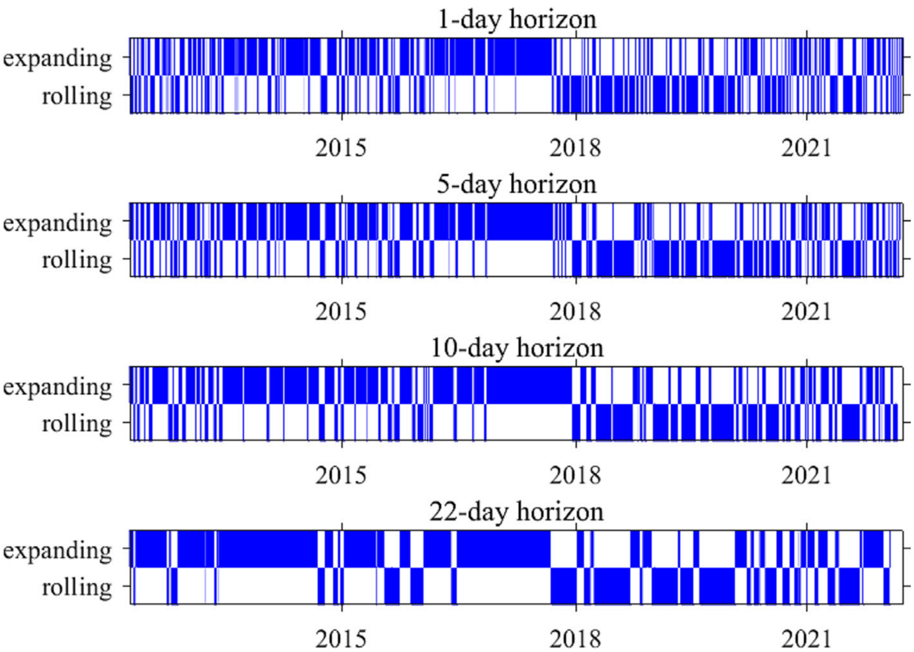

FIGURE 1 Model selection during out-of-sample forecasting period. The MoP strategy chooses the model with a rolling window or one with an expanding window when its past forecasting performance is better. Each prediction step's chosen (rejected) model is shown by a blue (white) bar.

$$w_t^* = \frac{SR/\gamma}{\sqrt{E_t(RV_{t+1})}}. \tag{13}$$

To put it simply, the optimal target for investors is the volatility of $\frac{SR}{\gamma}$ since $\sqrt{Var\left(w_t^* r_{t+1}^e\right)} = \frac{SR}{\gamma}$ is the conditional standard deviation of the portfolio's risky part. When the forecasted volatility risk of $\sqrt{E_t(RV_{t+1})}$ is larger than the optimal target of $\frac{SR}{\gamma}$ (i.e., $w_t^* < 1$), then the investor invests only a portion of her assets in the risky asset of stocks. Conversely, when $\sqrt{E_t(RV_{t+1})}$ is less than $\frac{SR}{\gamma}$ (i.e., $w_t^* > 1$), the investor needs to use leverage to reach her risk target.

This case in turn leads to an expected utility of

$$U\left(w_t^*\right) = \frac{SR^2}{2\gamma}. \tag{14}$$

However, $\sqrt{E_t(RV_{t+1})}$ is not available in real practice. By using the RV forecast of $\widehat{RV}_{t+1}$ for day $t+1$, we obtain the expected utility of

$$U\left(\widehat{RV}_{t+1}\right)=\frac{SR^2}{\gamma}\left(\frac{\sqrt{RV_{t+1}}}{\sqrt{\widehat{RV}_{t+1}}}-\frac{1}{2}\frac{RV_{t+1}}{\widehat{RV}_{t+1}}\right). \tag{15}$$

We estimate the average utility for the out-of-sample period, which is measured as follows:

$$\overline{U}\left(\widehat{RV}\right)=\frac{1}{q}\sum\nolimits_{t=R}^{R+p-1}\frac{SR^2}{\gamma}\left(\frac{\sqrt{RV_{t+1}}}{\sqrt{\widehat{RV}_{t+1}}}-\frac{1}{2}\frac{RV_{t+1}}{\widehat{RV}_{t+1}}\right), \tag{16}$$

where R and p stand for the length of in- and out-of-sample periods, respectively. According to Bollerslev et al. (2018), we choose $\gamma=2$ and $SR=0.4$, respectively. So as a result, $U\left(w_t^*\right)=4\%$. This result means that the investor is willing to pay 4% of her assets to acquire the w_t^* portfolio of risky assets instead of investing exclusively in risk-free bills.

Table 3 presents the portfolio performance as measured by the average realized utility. The profit (or return) of the portfolio, as adjusted for volatility risk, can be thought of as the realized utility. We find that our MoP model yields the greatest utility gains across four different forecast horizons: 3.489%, 3.754%, 3.734%, and 3.668%, respectively. This finding means that investors are more prepared to pay for the MoP model than they are for the other competing models. Our MoP model yields the greatest economic gains in an actual portfolio exercise.

TABLE 3 Portfolio performance.

Models	$h=1$	$h=5$	$h=10$	$h=22$
Rolling	3.477	3.470	3.400	3.199
Expanding	3.483	3.472	3.398	3.178
Mean	3.480	3.472	3.400	3.189
MoP	3.489	3.754	3.734	3.668

Note: The table reports portfolio performance as measured by the average realized utility. We suppose that a mean–variance investor will allocate her assets between stocks and risk-free bills through various RV forecasts, where the risk aversion coefficient is 2 and the constant Sharpe ratio is 0.4. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k=5$).

5 | ROBUSTNESS CHECKS

5.1 | Alternative look-back periods

Our MoP strategy depends on the performance of past forecasts, which is evaluated by the look-back period, namely, k. The results of previous reports are based on a weekly ($k=5$) look-back period. Considering this, we create an average MoP forecast while taking into account a few acceptable look-back periods. We take daily (1-day), weekly (5-day), biweekly (10-day), and monthly (22-day) look-back periods into consideration. The calculation method of an average MoP forecast ($\widehat{RV}_{t+1:t+h}^{MoP-AVG}$) is as follows:

$$\widehat{RV}_{t+1:t+h}^{MoP-AVG}=\frac{1}{4}\sum\nolimits_{k\in\{1,5,10,22\}}\widehat{RV}_{t+1:t+h}^{MoP}(k), \tag{17}$$

Table 4 shows the p-values of the MCS test, which include the average MoP strategy. Unsurprisingly, our MoP strategy continues to produce p-values larger than 0.1 throughout a variety of look-back periods and forecast horizons, which means that our MoP strategy is always contained in MCS at the 10% significance level. By contrast, a single model with a rolling or an expanding window and a mean combination model cannot always be included in MCS for various loss functions (except for the 1-day forecast horizon). In other words, when using acceptable look-back periods, our MoP model consistently outperforms the others. Tables A1, A2, and A3 display the MoP forecasting results for each look-back period (i.e., $k=1$, 10, and 22, respectively).

5.2 | Alternative window sizes

The size of the window has an impact on how well the rolling window forecasts (Rossi & Inoue, 2012). Additionally, a key of our MoP method is the combination of rolling and expanding windows, making window size crucial. We also take into account two alternative estimation window sizes for this.

In Tables 5 and 6, we produce the most recent 1500 and 2000 volatility forecasts for out-of-sample evaluation, respectively. Similar to the previous results, under different loss functions and forecast horizons, our MoP strategy is still able to produce p-values larger than 0.1. This case means that at the 10% significance level, our MoP strategy is always included in MCS. Moreover, the MoP model can generate the greatest p-values (i.e., 1) for all 12 cases in Tables 5 and 6. However, a single model with a rolling or expanding window and a mean combination model can only be contained in MCS in a

TABLE 4 MCS out-of-sample test based on different look-back period.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.013	**0.581**	**0.522**
Expanding	0.003	**0.895**	**0.141**
Mean	0.001	**0.793**	**0.263**
MoP	**1.000**	**0.640**	**0.955**
	Panel B: 5-day horizon		
Rolling	0.000	0.094	0.002
Expanding	0.000	0.094	0.002
Mean	0.000	0.094	0.002
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.125**	0.011
Expanding	0.002	**0.125**	0.011
Mean	0.002	**0.125**	0.011
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.052	**0.204**	0.016
Expanding	0.052	**0.204**	0.016
Mean	0.052	**0.204**	0.016
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by 1-, 5-, 10-, and 22-day look-back periods, respectively. The MoP forecast in this table is an average MoP forecast, which equals the simple mean of the four individual MoP forecasts according to the four distinct look-back periods. QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

TABLE 5 MCS out-of-sample test based on alternative evaluation sizes.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.039	**0.246**	**0.172**
Expanding	0.000	**0.246**	0.000
Mean	0.000	**0.228**	0.000
MoP	**1.000**	**1.000**	**1.000**
	Panel B: 5-day horizon		
Rolling	0.000	**0.120**	0.013
Expanding	0.000	**0.120**	0.012
Mean	0.000	**0.120**	0.013
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.007	**0.144**	0.029
Expanding	0.005	**0.144**	0.028
Mean	0.006	**0.144**	0.029
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.072	**0.172**	0.049
Expanding	0.069	**0.172**	0.048
Mean	0.072	**0.172**	0.049
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.
The whole sample period includes 5555 observations between February 02, 2000, and March 31, 2022, whereas the out-of-sample forecast includes the most recent 1500 observations.

few cases. The results of our prediction are therefore robust to various window sizes.

5.3 | HAR-RV-J models

The importance of jump is widely recognized in financial economics, and Corsi et al. (2010) further find that jumps are crucial for forecasting volatility. The HAR-RV-J model originated by Andersen et al. (2007) is one of the prevailing models for volatility forecasting using the jump component, which is described in the following way:

$$RV_{t+1:t+h} = \varphi_0 + \beta_d RV_t + \beta_w RV_{t-4:t} + \beta_m RV_{t-21:t} + \beta_J J_t + \varepsilon_{t+1:t+h}. \tag{18}$$

where the jump component $J_t = max\{RV_t - BPV_t, 0\}$, $BPV_t = u_1^{-2}\sum_{i=2}^{M}|r_{t,i}||r_{t,i-1}|$ is the realized bi-power variation (BPV), and $u_1 = \sqrt{\frac{2}{\pi}}$.

Thus, we generate a new MoP forecast by adding the jump model. Table 7 provides the predictive results when using the HAR-J model with rolling and expanding windows. Our MoP model is always able to produce *p*-values larger than 0.1. That means at the 10% significance level, our MoP strategy is always contained in MCS. Moreover,

TABLE 6 MCS out-of-sample test based on alternative evaluation sizes.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.055	**0.127**	0.002
Expanding	0.004	**0.153**	0.001
Mean	0.002	**0.127**	0.000
MoP	**1.000**	**1.000**	**1.000**
	Panel B: 5-day horizon		
Rolling	0.000	**0.113**	0.002
Expanding	0.000	**0.113**	0.002
Mean	0.000	**0.113**	0.002
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.001	**0.113**	0.012
Expanding	0.001	**0.113**	0.011
Mean	0.001	**0.113**	0.012
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.078	**0.185**	0.035
Expanding	0.073	**0.185**	0.033
Mean	0.078	**0.185**	0.035
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS p-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.
The whole sample period includes 5555 observations between February 02, 2000, and March 31, 2022, whereas the out-of-sample forecast includes the most recent 2000 observations.

TABLE 7 MCS out-of-sample test using the HAR-J model.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.000	0.095	0.000
Expanding	**0.204**	**0.582**	**1.000**
Mean	0.000	**0.201**	0.000
MoP	**1.000**	**1.000**	**0.217**
	Panel B: 5-day horizon		
Rolling	0.000	**0.161**	0.001
Expanding	0.000	**0.161**	0.004
Mean	0.000	**0.161**	0.003
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.125**	0.016
Expanding	0.002	**0.143**	0.016
Mean	0.002	**0.143**	0.016
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.050	**0.223**	0.018
Expanding	0.051	**0.223**	0.018
Mean	0.051	**0.223**	0.018
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS p-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-J model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-J model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

the MoP model can generate the largest p-values (i.e., 1) for 11 out of the 12 cases. However, the other models are rarely included in MCS and only generate the largest p-values when the forecast horizon is 1 day and the loss function is MAE. Therefore, the results of our prediction are robust to the jump model.

5.4 | Alternative volatility estimators

Considering that real market volatility is not observable, we use the realized kernel (RK), pioneered by Barndorff-Nielsen et al. (2008), another prevailing volatility estimator, to evaluate the forecasting performance of our MoP strategy. The RK is robust to market microstructure noise, which is described in the following way:

$$RK_t = \sum\nolimits_{h=-H}^{H} k\left(\frac{h}{H+1}\right)\gamma_h, \tag{19}$$

where

$$\gamma_h = \sum\nolimits_{j=|h|+1}^{N} r_{t,j} r_{t,j-|h|}, \tag{20}$$

and $k(x)$ is the Parzen kernel function and can be provided by

$$k(x)=\begin{cases}1-6x^2+6x^3 & 0\le x<\frac{1}{2}\\ 2(1-x)^3 & \frac{1}{2}\le x\le 1\\ 0 & x>1\end{cases}. \tag{21}$$

For a specific choice of H, please refer to Barndorff-Nielsen et al. (2009).

Table 8 shows the MCS results of using the RK to forecast future market volatility. For all 12 cases, the MoP model is always contained in MCS at the 10% significance level. Moreover, the MoP model can produce the greatest p-values for 11 out of the 12 cases. Nonetheless, all other models contained in MCS at different forecast horizons are only used when the loss function is MSE. Overall, when using RK as a volatility estimator, our MoP model continues to outperform the others. Therefore, the results of our prediction are robust to the alternative volatility estimator.

5.5 | Nonlinear HAR models

When forecasting volatility, nonlinear models known as the logarithmic HAR models are widely used (see, e.g., Andersen et al., 2007; Corsi et al., 2010; Prokopczuk et al., 2016; Liang et al., 2020). Specifically, the logarithmic HAR-RV model is given as follows:

$$ln(RV_{t+1:t+h})=\varphi_0+\beta_d ln(RV_t)+\beta_w ln(RV_{t-4:t})+\beta_m ln(RV_{t-21:t})+\varepsilon_{t+1:t+h}. \tag{22}$$

Table 9 provides the MCS results for the logarithmic HAR-RV models. At the 10% significance level, our MoP model is always contained in MCS for all 12 cases. Moreover, the MoP model is able to generate the greatest p-values for 11 of the 12 cases. Nevertheless, other competing models underperform the MoP model. Overall, when using the nonlinear HAR model, our MoP strategy continues to outperform the others. Therefore, the results of our prediction are robust to the nonlinear model.

5.6 | Business cycle

During the out-of-sample period, we distinguish between recession and expansion periods to investigate how the business cycle influences the forecasting abilities of various volatility models. Our basis for distinguishing business cycles is NBER-based recession indicators for the United States from the period following the peak through the trough. When the index is 1, it is a recession, and when the index is 0, it is an expansion. We can get corresponding data from FRED.[2] Here, the results are based on a weekly ($k=5$) look-back period and a weekly (5-day) forecast horizon. Numerous studies take the impact of the business cycle into consideration (see, e.g., Wang et al., 2018; Zhang et al., 2019; Dai et al., 2020; He et al., 2021; Zhang et al., 2021).

Table 10 provides the results of loss function values during different periods. First, the loss function values in the expansion period are evidently smaller than those in the recession period. Second, in the recession period, the model using the expanding window alone shows better results than the model using the rolling window alone.

TABLE 8 MCS out-of-sample test using the volatility measure of realized kernel.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.000	**0.126**	0.000
Expanding	0.070	**0.126**	**1.000**
Mean	0.000	**0.126**	0.000
MoP	**1.000**	**1.000**	**0.204**
	Panel B: 5-day horizon		
Rolling	0.000	**0.113**	0.000
Expanding	0.000	**0.113**	0.002
Mean	0.000	**0.113**	0.000
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.222**	0.000
Expanding	0.003	**0.222**	0.011
Mean	0.003	**0.222**	0.000
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.041	**0.216**	0.008
Expanding	0.046	**0.216**	0.015
Mean	0.046	**0.216**	0.011
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS p-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. In this table, the realized kernel (RK), not the RV, is used as the volatility estimator. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k=5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

[2]https://fred.stlouisfed.org/series/USREC

TABLE 9 MCS out-of-sample test using the logarithmic HAR-RV model.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.003	**0.591**	**0.786**
Expanding	0.000	**1.000**	0.053
Mean	0.000	**0.873**	**0.405**
MoP	**1.000**	**0.560**	**1.000**
	Panel B: 5-day horizon		
Rolling	0.000	0.089	0.001
Expanding	0.000	0.089	0.001
Mean	0.000	0.089	0.001
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.122**	0.012
Expanding	0.002	**0.122**	0.012
Mean	0.002	**0.122**	0.012
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.049	**0.199**	0.016
Expanding	0.049	**0.199**	0.016
Mean	0.049	**0.199**	0.016
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. This table displays the logarithmic HAR-RV model. The mean combination is an equally weighted average based on forecasts from the logarithmic HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the logarithmic HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

TABLE 10 Out-of-sample forecasting loss function values over business cycle.

Models	QLIKE	MSE	MAE
	Panel A: Recession period		
Rolling	2.956	58.717	5.213
Expanding	2.951	58.258	5.184
Mean	2.953	58.487	5.199
MoP	2.682	15.105	2.527
	Panel B: Expansion period		
Rolling	0.199	0.543	0.331
Expanding	0.199	0.541	0.332
Mean	0.199	0.542	0.332
MoP	0.069	0.146	0.196

Note: The table displays the loss function values for the four models. We specifically distinguish between expansion and recession periods during the out-of-sample period. Panels A and B report the relevant results for the recession and expansion periods, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration.

However, in the expansion period, the performances of the three competing models are not comparable. Third, our MoP model consistently outperforms the other three competing models, both in recession and expansion periods. In conclusion, our model can estimate volatility more accurately in a variety of economic environments and is robust in different business cycles.

5.7 | Different volatility levels

In order to explore the predictive power of our MoP strategy, we contrast the MoP model's predictive power with that of three competing models at different volatility levels. We divide the whole evaluation period into high- and low-volatility groups based on the median of the out-of-sample RV. We compute the loss function values for each model in different groups.

Table 11 reports the loss function values for various levels of volatility. In particular, for the high-volatility group in Panel A, the loss function values of four models are higher than those of the low-volatility group in Panel B. In addition, our MoP model consistently produces the lowest values of the loss function in both periods. In this instance, the results for both the low- and high-volatility periods show that our MoP strategy outperforms the others. Therefore, the results of our prediction are robust during different volatility periods.

5.8 | Different return levels

Does our MoP model produce significant predictive power under different return periods? For a plausible explanation, we further relate the predictability of RV to market returns. We divide the out-of-sample period into two groups according to positive and negative returns.

Table 12 reports the loss function values for the positive- and negative-return groups. We present the findings in two observations. First, all models have smaller loss function values in the positive-return group

TABLE 11 Out-of-sample forecasting loss function values between high- and low-volatility group.

Models	QLIKE	MSE	MAE
	Panel A: High-volatility group		
Rolling	0.881	2.988	0.627
Expanding	0.880	2.969	0.625
Mean	0.880	2.978	0.626
MoP	0.681	0.736	0.307
	Panel B: Low-volatility group		
Rolling	-0.388	0.110	0.205
Expanding	-0.386	0.111	0.207
Mean	-0.387	0.111	0.206
MoP	-0.453	0.073	0.166

Note: The table displays the loss function values for the four models. Based on the median of the out-of-sample RV, we specifically divide the entire evaluation period into high- and low-volatility groups. Panels A and B report the corresponding results. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration.

TABLE 12 Out-of-sample forecasting loss function values between positive and negative return group.

Models	QLIKE	MSE	MAE
	Panel A: Positive-return group		
Rolling	0.065	0.709	0.360
Expanding	0.067	0.707	0.361
Mean	0.066	0.708	0.360
MoP	-0.022	0.201	0.214
	Panel B: Negative-return group		
Rolling	0.464	2.557	0.483
Expanding	0.463	2.539	0.483
Mean	0.463	2.548	0.483
MoP	0.277	0.648	0.264

Note: The table displays the loss function values for the four models. According to the positive and negative returns, we sort the whole evaluation period into positive- and negative-return groups. Panels A and B report the relevant results for positive- and negative-return groups. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration.

TABLE 13 MCS out-of-sample test for crude oil market.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	**1.000**	**0.307**	**0.447**
Expanding	0.019	**1.000**	**1.000**
Mean	0.081	**0.332**	**0.520**
MoP	**0.436**	**0.332**	**0.634**
	Panel B: 5-day horizon		
Rolling	0.033	**0.149**	0.095
Expanding	0.027	**0.149**	0.090
Mean	0.033	**0.149**	0.095
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.044	**0.159**	**0.110**
Expanding	0.038	**0.159**	**0.107**
Mean	0.044	**0.159**	**0.110**
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.063	**0.268**	**0.151**
Expanding	0.059	**0.306**	**0.151**
Mean	0.063	**0.268**	**0.151**
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. In this table, we forecast the crude oil market (i.e., WTI) instead of the stock market. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a weekly look-back period ($k = 5$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

in Panel A than in the negative-return group in Panel B. Second, and more importantly, our MoP model always produces the smallest loss function values during positive- and negative-return periods. Our MoP model outperforms the other models for both positive- and negative-return periods. Therefore, the results of our prediction are robust for different return periods.

5.9 | Crude oil market

The oil market is also crucial for financial market predictability. Therefore, we consider the applicability of the

MoP strategy in the oil market. In particular, we extend the MoP strategy and the three competing models to the oil market to generate RV forecasts. The crude oil data we used are the daily 5-min high-frequency RV of WTI from the Oxford-Man Institute's Quantitative Finance Realized Library. Additionally, there are 5143 observations during the whole sample period, which is from January 31, 2002, to February 27, 2022. The first 2643 data are used as in-sample data, and we generate the most recent 2500 volatility forecasts for out-of-sample evaluation.

Table 13 provides p-values for the MCS test for the oil market. Specifically, the results are similar to those in the stock market. The MoP models all enter the MCS under the 10% significance level during different forecast horizons, and they outperform the other competing models. The results show that the MoP model is effective for forecasting oil market volatility and performs better in long-term forecasting.

6 | CONCLUSION

According to the MoP, we know that good past forecasting performance is always accompanied by good future forecasting performance. In light of this, we propose a model selection method that selects among models using rolling and expanding windows by observing their relative past forecasting performance.

Our empirical findings show that the MoP strategy can significantly improve the predictive power of the model, almost all of which are contained in MCS at the 10% significance level. Moreover, in portfolio performance, our MoP model can produce the highest utility gains among all four models. Additionally, our model passes a series of robustness tests. Therefore, our model passes the statistical and economic tests and thus exhibits strong predictive power.

The results of this study have some implications for policymakers and market investors. Policymakers can make more accurate volatility forecasts based on our strategy, which can help with more accurate pricing and more effective risk management. Meanwhile, according to the results of portfolio performance, investors can create more reliable and effective portfolio strategies to achieve higher investment returns.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (72001110, 72371131).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the Oxford-Man Institute's Quantitative Finance Realized Library and the Federal Reserve Bank of St Louis Economic Data (FRED).

ORCID

Yaojie Zhang https://orcid.org/0000-0002-4220-1623

REFERENCES

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*, 885–905. https://doi.org/10.2307/2527343

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, *89*, 701–720. https://doi.org/10.1162/rest.89.4.701

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, *76*, 1481–1536. https://doi.org/10.3982/ECTA6495

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, *12*, 1–32. https://doi.org/10.1111/j.1368-423X.2008.00275.x

Bollerslev, T., Hood, B., Huss, J., & Pedersen, L. H. (2018). Risk everywhere: Modeling and managing volatility. *Review of Financial Studies*, *31*, 2729–2773. https://doi.org/10.1093/rfs/hhy041

Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, *192*, 1–18. https://doi.org/10.1016/j.jeconom.2015.10.007

Clark, T. E., & McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, *50*, 363–395. https://doi.org/10.1111/j.1468-2354.2009.00533.x

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*, 174–196. https://doi.org/10.1093/jjfinec/nbp001

Corsi, F., Pirino, D., & Renò, R. (2010). Threshold bipower variation and the impact of jumps on volatility forecasting. *Journal of Econometrics*, *159*, 276–288. https://doi.org/10.1016/j.jeconom.2010.07.008

Dai, Z. F., Zhou, H. T., Dong, X. D., & Kang, J. (2020). Forecasting stock market volatility: A combination approach. *Discrete Dynamics in Nature and Society*, *2020*, 1–9. https://doi.org/10.1155/2020/1428628

Degiannakis, S., & Filis, G. (2017). Forecasting oil price realized volatility using information channels from other asset classes. *Journal of International Money and Finance*, *76*, 28–49. https://doi.org/10.1016/j.jimonfin.2017.05.006

Gillitzer, C., & McCarthy, M. (2019). Does global inflation help forecast inflation in industrialized countries? *Journal of Applied Econometrics*, *34*, 850–857. https://doi.org/10.1002/jae.2704

Gong, X., & Lin, B. (2018). Structural breaks and volatility forecasting in the copper futures market. *Journal of Futures Markets*, *38*, 290–339. https://doi.org/10.1002/fut.21867

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*, 453–497. https://doi.org/10.3982/ECTA5771

Haugom, E., Langeland, H., Molnár, P., & Westgaard, S. (2014). Forecasting volatility of the US oil market. *Journal of Banking & Finance*, *47*, 1–14. https://doi.org/10.1016/j.jbankfin.2014.05.026

He, M., Hao, X., Zhang, Y., & Meng, F. (2021). Forecasting stock return volatility using a robust regression model. *Journal of Forecasting*, *40*, 1463–1478. https://doi.org/10.1002/for.2779

He, M., Zhang, Y., Wen, D., & Wang, Y. (2022). Forecasting the Chinese stock market volatility: A regression approach with at-distributed error. *Applied Economics*, *54*, 1–16. https://doi.org/10.1080/00036846.2022.2053653

Inoue, A., Jin, L., & Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, *196*, 55–67. https://doi.org/10.1016/j.jeconom.2016.03.006

Laurent, S., Rombouts, J. V., & Violante, F. (2012). On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics*, *27*, 934–955. https://doi.org/10.1002/jae.1248

Liang, C., Wei, Y., & Zhang, Y. (2020). Is implied volatility more informative for forecasting realized volatility: An international perspective. *Journal of Forecasting*, *39*, 1253–1276. https://doi.org/10.1002/for.2686

Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, *187*, 293–311. https://doi.org/10.1016/j.jeconom.2015.02.008

Ma, F., Wahab, M. I. M., & Zhang, Y. (2019). Forecasting the U.S. stock volatility: An aligned jump index from G7 stock markets. *Pacific-Basin Finance Journal*, *54*, 132–146. https://doi.org/10.1016/j.pacfin.2019.02.006

Ma, F., Wei, Y., Liu, L., & Huang, D. (2018). Forecasting realized volatility of oil futures market: A new insight. *Journal of Forecasting*, *37*, 419–436. https://doi.org/10.1002/for.2511

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*, 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Patton, A. J., & Sheppard, K. (2009). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, *25*, 218–238. https://doi.org/10.1016/j.ijforecast.2009.01.011

Pesaran, M. H., & Timmermann, A. (2002). Market timing and return prediction under model instability. *Journal of Empirical Finance*, *9*, 495–510. https://doi.org/10.1016/S0927-5398(02)00007-5

Pesaran, M. H., & Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, *137*, 134–161. https://doi.org/10.1016/j.jeconom.2006.03.010

Pesaran, M. H., & Timmermann, A. (2009). Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association*, *104*, 325–337. https://doi.org/10.1198/jasa.2009.0113

Prokopczuk, M., Symeonidis, L., & Wese Simen, C. (2016). Do jumps matter for volatility forecasting? Evidence from energy markets. *Journal of Futures Markets*, *36*, 758–792. https://doi.org/10.1002/fut.21759

Rapach, D. E., Strauss, J. K., & Wohar, M. E. (2009). Forecasting stock return volatility in the presence of structural breaks. *Frontiers of Economics and Globalization*, *3*, 381–416. https://doi.org/10.1016/S1574-8715(07)00210-2

Rapach, D., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, *23*, 821–862. https://doi.org/10.1093/rfs/hhp063

Rossi, B., & Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, *30*, 432–453. https://doi.org/10.1080/07350015.2012.693850

Schrimpf, A., & Qingwei, W. (2010). A reappraisal of the leading indicator properties of the yield curve under structural instability. *International Journal of Forecasting*, *26*, 289–310. https://doi.org/10.1016/j.ijforecast.2009.08.005

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, *41*, 788–829. https://doi.org/10.1257/jel.41.3.788

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430. https://doi.org/10.1002/for.928

Stock, J. H., & Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*, 3–33. https://doi.org/10.1111/j.1538-4616.2007.00014.x

Wang, Y., Liu, L., Ma, F., & Diao, X. (2018). Momentum of return predictability. *Journal of Empirical Finance*, *45*, 141–156. https://doi.org/10.1016/j.jempfin.2017.11.003

Wang, Y., Ma, F., Wei, Y., & Wu, C. (2016). Forecasting realized volatility in a changing world: A dynamic model averaging approach. *Journal of Banking & Finance*, *64*, 136–149. https://doi.org/10.1016/j.jbankfin.2015.12.010

Wang, Y., Wu, C., & Yang, L. (2016). Forecasting crude oil market volatility: A Markov switching multifractal volatility approach. *International Journal of Forecasting*, *32*, 1–9. https://doi.org/10.1016/j.ijforecast.2015.02.006

Wei, Y., Liu, J., Lai, X., & Hu, Y. (2017). Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? *Energy Economics*, *68*, 141–150. https://doi.org/10.1016/j.eneco.2017.09.016

Zhang, Z., He, M., Zhang, Y., & Wang, Y. (2021). Realized skewness and the short-term predictability for aggregate stock market volatility. *Economic Modelling*, *103*, 105614. https://doi.org/10.1016/j.econmod.2021.105614

Zhang, Y., Ma, F., & Liao, Y. (2020). Forecasting global equity market volatilities. *International Journal of Forecasting*, *36*, 1454–1475. https://doi.org/10.1016/j.ijforecast.2020.02.007

Zhang, Y., Ma, F., Wang, T., & Liu, L. (2019). Out-of-sample volatility prediction: A new mixed-frequency approach. *Journal of Forecasting*, *38*, 669–680. https://doi.org/10.1002/for.2590

Zhang, Y., Wang, Y., Ma, F., & Wei, Y. (2022). To jump or not to jump: Momentum of jumps in crude oil price volatility prediction. *Financial Innovation*, *8*, 56. https://doi.org/10.1186/s40854-022-00360-7

Zhang, Y., Wei, Y., Ma, F., & Yi, Y. (2019). Economic constraints and stock return predictability: A new approach. *International Review of Financial Analysis*, *63*, 1–9. https://doi.org/10.1016/j.irfa.2019.02.007

Zhang, Y., Wei, Y., Zhang, Y., & Jin, D. (2019). Forecasting oil price volatility: Forecast combination versus shrinkage method. *Energy Economics*, *80*, 423–433. https://doi.org/10.1016/j.eneco.2019.01.010

AUTHOR BIOGRAPHIES

Yuqing Fengis a master at the School of Economics and Management, Nanjing University of Science and Technology. Her research interests include return predictability, volatility forecasting, energy forecasting, financial econometrics, and empirical finance. She has published a paper in Emerging Markets Finance and Trade.

Yaojie Zhang is an associate professor at the School of Economics and Management, Nanjing University of Science and Technology. His research interests include return predictability, volatility forecasting, energy forecasting, financial engineering, financial econometrics, and empirical finance. He has published many papers in the Journal of Empirical Finance, International Journal of Forecasting, Quantitative Finance, Energy Economics, Journal of Forecasting, Knowledge-Based Systems, International Review of Financial Analysis, Pacific-Basin Finance Journal, Economic Modelling, and Applied Economics.

Yudong Wang is a professor at the School of Economics and Management, Nanjing University of Science and Technology. His research interests include financial econometrics, empirical finance, and forecasting. He has published widely in the Management Science, Journal of Empirical Finance, Journal of Banking and Finance, and International Journal of Forecasting.

How to cite this article: Feng, Y., Zhang, Y., & Wang, Y. (2024). Out-of-sample volatility prediction: Rolling window, expanding window, or both? *Journal of Forecasting*, *43*(3), 567–582. https://doi.org/10.1002/for.3046

APPENDIX A

TABLE A1 MCS out-of-sample test based on 1-day look-back period.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.000	**0.549**	0.019
Expanding	0.000	**0.580**	0.002
Mean	0.000	**0.580**	0.000
MoP	**1.000**	**1.000**	**1.000**
	Panel B: 5-day horizon		
Rolling	0.000	**0.100**	0.003
Expanding	0.000	**0.100**	0.003
Mean	0.000	**0.100**	0.003
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.001	**0.123**	0.010
Expanding	0.001	**0.123**	0.009
Mean	0.001	**0.123**	0.010
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.051	**0.197**	0.020
Expanding	0.052	**0.197**	0.019
Mean	0.052	**0.197**	0.020
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a daily look-back period ($k = 1$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

TABLE A2 MCS out-of-sample test based on 10-day look-back period.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.005	**0.611**	**0.285**
Expanding	0.001	**1.000**	0.043
Mean	0.000	**0.862**	0.037
MoP	**1.000**	**0.611**	**1.000**
	Panel B: 5-day horizon		
Rolling	0.000	0.089	0.002
Expanding	0.000	0.089	0.002
Mean	0.000	0.089	0.002
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.120**	0.012
Expanding	0.002	**0.120**	0.012
Mean	0.002	**0.120**	0.012
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.052	**0.200**	0.013
Expanding	0.052	**0.200**	0.013
Mean	0.052	**0.200**	0.013
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a biweekly look-back period ($k = 10$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.

TABLE A3 MCS out-of-sample test based on 22-day look-back period.

Models	QLIKE	MSE	MAE
	Panel A: 1-day horizon		
Rolling	0.045	**0.581**	**1.000**
Expanding	0.010	**1.000**	**0.466**
Mean	0.006	**0.864**	**0.618**
MoP	**1.000**	**0.390**	**0.820**
	Panel B: 5-day horizon		
Rolling	0.000	0.091	0.002
Expanding	0.000	0.091	0.002
Mean	0.000	0.091	0.002
MoP	**1.000**	**1.000**	**1.000**
	Panel C: 10-day horizon		
Rolling	0.002	**0.133**	0.009
Expanding	0.002	**0.133**	0.009
Mean	0.002	**0.133**	0.009
MoP	**1.000**	**1.000**	**1.000**
	Panel D: 22-day horizon		
Rolling	0.057	**0.206**	0.016
Expanding	0.057	**0.206**	0.016
Mean	0.057	**0.206**	0.016
MoP	**1.000**	**1.000**	**1.000**

Note: The table displays the MCS *p*-values for the four models. The results corresponding to the 1-, 5-, 10-, and 22-day horizons are reported in Panels A, B, C, and D, respectively. The mean combination is an equally weighted average based on forecasts from the HAR-RV model with a rolling window or one with an expanding window. Our MoP model alternates between the HAR-RV model with a rolling window and one with an expanding window according to relative past predictions. The performance of past forecasts is assessed by a monthly look-back period ($k = 22$). QLIKE, MSE, and MAE are the three loss functions taken into consideration. Bold numbers indicate significant instances where the related models meet the 10% significance level of the MCS.