ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A False Discovery Rate approach to optimal volatility forecasting model selection☆

Arman Hassanniakalager [a], Paul L. Baker [b,*], Emmanouil Platanakis [b]

[a] CMC Markets, United Kingdom
[b] School of Management, University of Bath, United Kingdom

ARTICLE INFO

Keywords:
Volatility forecasting
Multiple hypothesis testing
False discovery rate
Model selection
Bootstrapping

ABSTRACT

Estimating financial market volatility is integral to the study of investment decisions and behaviour. Previous literature has, therefore, attempted to identify an optimal volatility forecasting model. However, optimal volatility forecasting is dynamic. It depends on the asset being studied and financial market conditions. We propose a novel empirical methodology to account for this dynamism. Using our Multiple Hypothesis Testing with the False Discovery Rate (FDR) method, we identify buckets of superior-performing models relative to the literature's benchmark models. We present evidence that our proposed FDR bucket with GJR-GARCH has the lowest forecast error in predicting one-step-ahead realized volatility. We also compare our FDR method with two Family-Wise Error Rate model selection frameworks, and the evidence supports our proposed FDR methodology.

1. Introduction

Estimating risk is integral to investment decisions and the study of investment behaviour. Although volatility and risk have subtle differences, volatility is frequently used as a proxy for risk, and therefore, forecasting volatility is necessary. A large and growing literature on volatility forecasting attempts to either present new volatility models or examine the predictive ability of the traditional models, e.g. GARCH (1,1). However, the literature has not conclusively identified an optimal volatility forecasting model, defined here as the model with the minimum cross-sectional out-of-sample robust loss.[1] This is because optimal volatility forecasting is dynamic. The optimal volatility forecasting model depends on the specific forecasting scenario, which is a function of the studied asset and financial market conditions. Therefore, this paper addresses the issue by proposing a novel empirical methodology that enables the researcher and/or practitioner to evaluate a large pool of potential models and select the optimal scenario-specific volatility forecasting model from it. Using this methodology, we can identify significant models that outperform the literature's conventional benchmarks of GARCH, GJR-GARCH of Glosten, Jagannathan, and Runkle (1993) and HAR of Corsi (2009). To further establish the merit of our proposed method, we compare it against two Family-Wise Error Rate model selection frameworks. The empirical evidence supports

☆ This paper subsumes earlier versions under a similar title. All scripts used to generate the volatility forecasting models and the bucket formations are available through the GitHub repository at https://github.com/hkalager/FDR_buckets.

* Corresponding author.
E-mail address: P.L.Baker@bath.ac.uk (P.L. Baker).

[1] The definition of optimality can vary. For example, both in-sample and out-of-sample forecast errors can be considered as a proxy of accuracy and optimality for volatility forecasting models. For this study, we define an optimal volatility forecasting model as the model with the minimum cross-sectional out-of-sample robust loss as suggested in Patton (2011). This enables us to identify the most accurate forecasting model given the conditions of the respective financial markets.

https://doi.org/10.1016/j.ijforecast.2023.07.003

Please cite this article as: A. Hassanniakalager, P.L. Baker and E. Platanakis, A False Discovery Rate approach to optimal volatility forecasting model selection. International Journal of Forecasting (2023), https://doi.org/10.1016/j.ijforecast.2023.07.003.

our proposed methodology for selecting optimal volatility forecasting models over these alternative frameworks. Therefore, this paper's contribution provides researchers and practitioners with a replicable strategy to choose the most accurate volatility forecasting model for the given asset and financial market conditions.

Market volatility is a latent economic variable that is not directly observable and yet is of significant economic importance. Accurate estimates of market volatility are necessary for option pricing models, quantitative risk management, and asset pricing (Brooks & Persand, 2003; Harvey & Whaley, 1992; Poon & Granger, 2003). Christoffersen and Diebold (2000) assert that successful volatility forecasting improves such decisions. The challenge in optimal volatility forecasting is identifying and selecting the 'best' (most accurate) model. The likelihood of identifying the best model will increase with the size of the pool of candidate models being evaluated. However, as the number of candidate models increases, a pairwise comparison becomes computationally infeasible. Examples in the previous literature of a limited number of candidate models being evaluated include Kang, Kang, and Yoon (2009), who evaluate four potential models and Wei, Wang, and Huang (2010), who evaluate ten candidate models. The solution to this issue is using a Multiple Hypothesis Testing (MHT) model selection framework, which provides an unbiased simultaneous comparison of a large number of candidate models. This paper uses MHT to compare a pool of 325 models.

However, the challenge in comparing a large pool of models is the increasing probability – as related to the number of hypotheses being simultaneously tested – of false discovery (Type I error(s)). In this case, the findings may be due to chance rather than true performance. In a single hypothesis testing scenario, there is a level of uncertainty involved. Statistically, a false discovery occurs when a true null hypothesis is falsely rejected. In MHT, the probability of having at least one false discovery – the family-wise error rate (FWER) – increases with the number of hypothesis tests conducted. The MHT literature has sought to resolve this issue through one of two approaches. The FWER approach is an attempt to control the probability of false discoveries.[2] However, the FWER approach may lead to a lower power[3] since it focuses on the count of Type I error(s) rather than the proportion of Type I errors. The alternative approach to dealing with false discoveries is to estimate and control for the expected proportion of false discoveries within the set of rejections known as the False Discovery Rate (FDR).[4] Using the FDR approach, we control for and identify the false discoveries where the FDR approach enables us to provide a better balance between true and false rejections of a null hypothesis. To the best of our knowledge, no paper within the volatility forecasting literature uses an MHT strategy with the FDR solution to false discoveries.

Although the volatility forecasting literature has been growing, it has had limited success in identifying an optimal model, either as a function of a small pool of candidate models (e.g., Kang et al., 2009; Wei et al., 2010) and/or the use of the FWER approach to false discoveries. The closest paper to the work here is Hansen and Lunde (2005), who compare a pool of 330 candidate models.[5] However, as the authors use an extension of the FWER approach to false discoveries, they can only assert that – relative to their benchmarks – superior volatility forecasting models exist, as their approach cannot identify which models they are. Furthermore, as subsequently shown by Patton (2011), the loss functions used by Hansen and Lunde (2005) are not robust. Finally, the Hansen and Lunde (2005) study does not include prominent families of volatility models, including stochastic volatility and the Heterogeneous Autoregressive (HAR) model of Corsi (2009) and its extension.

Therefore, our paper contributes to the literature by proposing an econometric strategy which enables the researcher to identify and select optimal volatility forecasting models from any given pool of candidate models: a formal MHT evaluation framework with an $FDR^{+/-}$ (two-tailed false discovery rate test) solution with a Discrete Right Bound (DRB) treatment (to not limit the identification of false discoveries to a discrete or continuous space). We use this methodology to evaluate a large pool (325) of conditional volatility models using one-step-ahead forecasts for two stock and two commodity indices from 2014 to 2020. The candidate conditional volatility models are evaluated against the extant literature's benchmark models of GARCH, GJR-GARCH, and HAR using the five different robust loss functions of Patton (2011). Following Hurlin, Laurent, Quaedvlieg, and Smeekes (2017), we form buckets of volatility forecasting models that outperform the benchmark models. We replicate this process for two competing MHT model selection frameworks (described below) and compare performance across the three model selection frameworks as a function of forecast accuracy. In doing so, we find evidence of a significant performance improvement using our proposed $FDR^{+/-}$ with DRB strategy for model selection. We also investigate and show the robustness of our results subject to differences in sample size, financial market stress, and the effect of different time periods. These checks confirm our preferred approach.

The paper proceeds as follows. Section 2 reviews the literature on volatility forecasting models and model selection frameworks. Based on this review, Section 3 details our proposed volatility forecasting model selection methodology and our implementation of it. Section 4 takes our methodology to the data, discusses the results, and investigates their robustness. Section 5 concludes.

[2] See Bonferroni (1936), Holm (1979), White (2000) and Hansen (2005) for the literature developing the FWER approach along with its extensions.

[3] Type II or False Negative (FN) error corresponds to the cases where a false null hypothesis in not rejected in favor of the alternative hypothesis. Subsequently, the power of an individual hypothesis test is defined as $1 - FN$.

[4] See Benjamini and Hochberg (1995), Storey, Taylor, and Siegmund (2004), Barras, Scaillet, and Wermers (2010), Bajgrowicz and Scaillet (2012), Liang and Nettleton (2012), Bancroft, Du, and Nettleton (2013), and Liang (2016) for the literature developing the FDR approach along with its extensions.

[5] Hansen and Lunde (2005) use three conditional mean and two conditional variance specifications leading to a large proportion of structurally similar models. In our application we avoid such repetition and consider alternative categories of models.

2. Literature review of volatility forecasting models and a critique of model selection frameworks

In this section, we review the literature on modelling volatility, which forms the basis of the candidate pool of models to which we apply our model selection framework. We then review and critique the literature on model selection frameworks toward developing our proposed model selection methodology. Finally, we review the literature on evaluating volatility forecasting models and their lack of conclusiveness in identifying an optimal volatility forecasting model. This establishes our contribution to the literature of our proposed model selection methodology.

2.1. Modelling volatility

The literature has developed and advanced numerous statistical approaches toward modelling the variation of volatility across time. We draw from the four most common families of autoregressive models to build our pool of candidate volatility forecasting models: (i) generalized autoregressive conditional heteroskedasticity (GARCH); (ii) exponential weighting moving average (EWMA); (iii) stochastic volatility (SV); and (iv) heterogeneous autoregressive (HAR).

The GARCH models of Engle (1982) and Bollerslev (1986) are widely recognized for their accuracy and simplicity by researchers and practitioners. Various extensions of the GARCH models have been developed. The most common classes of GARCH models are I. the absolute value model of Taylor (1986) and Schwert (1989), II. the exponential GARCH (EGARCH) of Nelson (1991), III. the asymmetric GJR-GARCH model of Glosten et al. (1993), IV. the threshold GARCH (TGARCH) model of Zakoian (1994), V. the integrated GARCH (IGARCH) of Engle and Bollerslev (1986), VI. the fractionally integrated GARCH (FI-GARCH) of Baillie, Bollerslev, and Mikkelsen (1996), and VII. The exponential weighting moving average (EWMA) models of JP Morgan are effectively non-stationary extensions of IGARCH. See Poon and Granger (2003) for an extensive review of various classes of GARCH models.

The SV method of Taylor (1982) is an alternative technique for modelling time-varying volatility. The SV focuses on the latent feature of conditional volatility and models it by a stochastic process (see Sadorsky (2005) as an example, among many others). The SV literature is well-reviewed. See Ghysels, Harvey, and Renault (1996), Broto and Ruiz (2004), Sadorsky (2005), Asai, McAleer, and Yu (2006), and Chan and Grant (2016).

Corsi (2009) introduced the HAR model as a long-memory regression model to approximate daily realized volatility. The HAR involves daily, weekly, and monthly components. Despite its simple structure, HAR has emerged as a successful volatility forecasting method (Bollerslev, Patton, & Quaedvlieg, 2016; Corsi, 2009; Hansen & Lunde, 2011; Özbekler, Kontonikas, & Triantafyllou, 2021). Bollerslev et al. (2016) expanded on the HAR model and introduced the HAR Quarticity (HARQ) model.

The development of our pool of candidate models from the above four families of autoregressive models is discussed in detail in Section 3.2.

2.2. Model selection frameworks

We use MHT as our model selection framework to identify and select the optimal volatility forecasting model from the pool of candidates. In MHT, many hypotheses are tested simultaneously and compared to a benchmark. A pairwise comparison becomes computationally infeasible when the number of candidate models increases. The collective approach of MHT is the solution to address this. However, a significant issue arising from comparing many hypotheses is the false discovery (Type I error). The concern is that the results may be due to chance rather than model performance. We know that uncertainty is always involved, even in a single hypothesis testing scenario. Statistically, a false discovery occurs when a true null hypothesis is (falsely) rejected. In MHT, the probability of having at least one false discovery – the FWER – increases with the number of tests. However, the FWER tends to be overly conservative – for instance, having few Type I errors when testing hundreds of hypotheses may not drastically affect the set of discoveries. Another approach to deal with the Type I error is to estimate and control the expected proportion of false discoveries within the set of rejections, known as the False Discovery Rate (FDR). The FWER and FDR approaches to false discoveries have been widely used in the MHT literature.

One of the earliest efforts to deal with the FWER is the Bonferroni (1936) correction using a cut-off value of α/m to control the FWER for m hypothesis tests at significance level α. The Bonferroni correction is criticized for a high chance of Type II error and lack of power (Benjamini & Hochberg, 1995). A more lenient method to cope with the FWER is the sequential procedure of Holm (1979), where the null hypothesis i is rejected if $p_i \leq \alpha/(m - i + 1)$ for $i = 1, \ldots, m$. The procedure effectively allows for a larger number of rejections than the Bonferroni correction. However, it is still overly conservative for a large-scale MHT framework.

Continuing with the development of the FWER approach to false discoveries, White (2000) introduces the seminal Bootstrap Reality Check (BRC) to test if any candidate model can outperform a benchmark in terms of a performance metric. The BRC estimates the probability of having any discoveries while accounting for the FWER. Hansen (2005) then proposes the Superior Predictive Ability (SPA) test as an extension of the BRC by using the studentized test statistic and minimising the effect of irrelevant models on the test performance. However, despite their contribution to the MHT literature, the BRC and SPA tests focus on controlling the probability of having false discoveries at a certain level rather than finding the true rejections individually. Other critical limitations these tests share include structurally conservative p-values – due to the underlying FWER approach – and a lack of power when the MHT candidate models have a wide range of performance metrics (Romano, Shaikh, & Wolf, 2008a). Accordingly, Romano and Wolf (2005) (RW) propose a stepwise procedure to find the set of discoveries based on the BRC test while controlling the FWER at a target level. Romano et al. (2008a) (RSW) attempt to generalize the RW procedure by controlling the probability

of having k or more Type I errors (k-FWER) through their stepwise kStepM procedure. Despite this more considered approach to false discoveries, the k plays an important role as a hyperparameter, as the rejection set is quite sensitive to the practitioner's choice. As an alternative to manually setting k, RSW suggest a stepwise approach to estimating the expected false discovery proportion (FDP). Their kStepM − FDP algorithm starts by setting $k = 1$ and computes the expected FDP at target level γ. If the expected FDP within the rejection set is higher than γ the k is increased by a unit. The algorithm stops at the smallest k to satisfy $N_k \leq k/\gamma - 1$.

Hansen, Lunde, and Nason (2011) further develop the FWER literature and introduce the Model Confidence Set (MCS). MCS is a stepwise procedure toward an optimal subset of alternatives where the chance of a Type I error is controlled at significance level α.

In our analysis, we use the kStepM and the MCS (discussed above) FWER approaches as comparative model selection methods against our preferred FDR procedure (discussed next). In doing so, as shown in Section 4, the evidence supports our proposed FDR approach over the two FWER approaches.

The literature on FDR studies begins with Benjamini and Hochberg (1995). They define the FDR as the expected proportion of false discoveries with size F within the R instance of rejections. Storey (2002) then proposes the positive FDR (pFDR), where the FDR is conditional on having at least one discovery ($F > 0$). The pFDR is a direct procedure that simplifies the FDR estimation procedure of Benjamini and Hochberg (1995). The pFDR controls the FDR at confidence $\alpha / \Pr(F > 0)$. Storey et al. (2004) further provide for a sequential procedure to create a rejection region for p-values as $[0, \gamma]\gamma \geq 0$, while controlling both the FDR and pFDR at the target level α.

Barras et al. (2010) introduce $\text{FDR}^{+/-}$ to extend the concept of finding the true discoveries to both tails of a test and attempt to control the probability of falsely rejecting a candidate model as an underperformer. They offer FDR^{-} for *unlucky* candidate models and the FDR^{+} for the *lucky* ones. Bajgrowicz and Scaillet (2012) provide a sequential procedure for the $\text{FDR}^{+/-}$ based on the Storey et al. (2004) procedure for the pFDR (Storey, 2002). The recent FDR literature proposes adaptive approaches to optimize the tuning hyperparameter, which estimates the proportion of null hypotheses to measure the FDR. The adaptive FDR techniques allow for a higher power and number of rejections while controlling the FDR at a fixed target level (Bancroft et al., 2013; Liang, 2016; Liang & Nettleton, 2012).

2.3. Volatility forecasting literature

The literature on volatility forecasting is traced back to Dimson and Marsh (1990). The authors study stock market volatility in the UK over the period 1955–1989 and report that under an unbiased evaluation, "relatively sophisticated forecasting methods" underperform a random walk benchmark. In contrast, Andersen and Bollerslev (1998) study Deutschemark to U.S. dollar (DM/$) and Japanese yen to U.S. dollar spot exchange rates over 1987–1992 and report that ARCH and GARCH models provide accurate volatility forecasts. Kang et al. (2009) study four GARCH classes (GARCH, IGARCH, CGARCH, and FIGARCH) and use the forecast accuracy test (Diebold & Mariano, 1995) comparing one, two, and five-day-ahead forecasts for three crude oil time series. Their results show a significant difference in forecasting out-of-sample volatility between the competing oil price models. Engle and Figlewski (2015) studied 28 US stocks and developed a class of EGARCH models for analysing implied volatility while considering pairwise correlations between the assets. However, the respective candidate pools of models in these studies are small, and therefore, none of these papers uses a formal MHT error-controlling approach to evaluate the models.

Wei et al. (2010) expanded the pool of Kang et al. (2009) to nine GARCH classes of models (RiskMetrics, GARCH, IGARCH, GJR, EGARCH, APARCH, FIGARCH, FIAPARCH, and HYGARCH) and used the Hansen (2005) SPA test. Their results could not show that any candidate model outperformed its counterparts in forecasting the volatility of Brent and West Texas Intermediate (WTI) crude oil prices.

Hansen and Lunde (2005) combine a large pool of volatility models and compare them based on the SPA test. They report mixed results for the foreign exchange market and the stock market. They find no significant improvement in favour of exotic extensions of GARCH models compared to a GARCH(1,1) and an ARCH(1) benchmarks for DM/$ volatility. For IBM stock, the benchmark is outperformed by at least some models in the pool. However, Hansen and Lunde's (2005) approach is not a model selection framework and cannot individually identify superior forecasting models. Furthermore, Hansen and Lunde (2005) use a set of performance metrics – such as the mean absolute error – that are not robust or consistent (Patton, 2011).

The volatility forecasting literature continues with Bao, Lee, and Saltoglu (2006), who compare sixteen volatility models for Value at Risk (VaR) based on the BRC test. They study risk models for three periods in five Asian countries. Their findings do not identify superior models over the crisis periods, and the VaR models behave similarly. Esposito and Cummins (2016) expand the studied pool of Bao et al. (2006) and deploy a k-FWER approach to MHT by employing the kStepM procedure. They study sixteen out-of-sample periods for one-step-ahead and ten-step-ahead forecasts and report that superior models exist under a more powerful MHT procedure. Bollerslev et al. (2016) use the BRC test to compare HARQ against eight AR and HAR specifications. Their out-of-sample study for the daily S&P 500 index shows limited significant differences between the candidate models.

Notably, despite successful applications of the FDR procedure in other fields of study (see – among others – Anderson, 2017; Ardia & Hoogerheide, 2014; Bajgrowicz & Scaillet, 2012; Davis & Heller, 2020; Hurlin et al., 2017; Storey & Tibshirani, 2003; Sun, Reich, Cai, Guindani, & Schwartzman, 2015; Wu, 2008) this review of the literature on volatility forecasting finds no record of model

selection based on an FDR approach.[6] We address this gap in the literature and present a comparative analysis studying the role of MHT procedures in volatility forecasting.

3. Volatility forecasting model methodology, implementation and application

In this section, we detail our proposed methodology for optimal volatility forecasting model selection (MHT with FDR procedure), its implementation (candidate pool of models and performance metrics) and application (the data to which we bring our methodology). Briefly: We examine the accuracy of a large pool of volatility forecasting models through our MHT with the FDR approach. By considering alternative specifications for each class of volatility models and a range of plausible parameterizations for each specification, we construct a pool of 325 candidate volatility models. We then take this strategy to two stock and two commodity indices. We conduct the empirical analysis over seven calendar year periods. We follow the literature and the industry's practice of using one-step-ahead forecasting accuracy as the performance measure.[7]

In implementing the FDR model selection method, we compare the candidate models based on the stepwise FDR$^{+/-}$ procedure of Bajgrowicz and Scaillet (2012). We further adjust the FDR$^{+/-}$ method for structural dependence between studied models. The FDR$^{+/-}$ method is based on a homogenous continuous space p-value assumption. Such an assumption may not hold, given the clustered nature of market volatility Corsi and Renò (2012). We address this issue by using the Discrete Right Boundary (DRB) of Liang (2016). The DRB adjusts a false discovery controlling procedure for discrete p-values. We combine the FDR$^{+/-}$ with the DRB treatment in our application. The advantage of the DRB is that it is an adaptive process and does not limit the false discovery method to only a discrete or continuous space.

As will be shown in Section 4, we find a significant improvement in the performance of the volatility forecasting models selected by our proposed MHT with FDR$^{+/-}$ and DRB methodology relative to the conventional GARCH(1,1), GJR-GARCH(1,1), and HAR forecasting models. The models selected by our methodology have the lowest average error across all assets. We also find strong evidence in favour of using our FDR$^{+/-}$ model selection procedure relative to the two seminal FWER approaches – the *k*StepM and the MCS. We further explore the robustness of our methodology by considering different market conditions (levels of financial market stress), sample sizes and time periods. Our results offer a novel approach to improved performance by specifying volatility forecasting models.

3.1. FDR$^{+/-}$

The FDR$^{+/-}$ is a stepwise selection procedure aimed at finding the set of true rejections in an MHT set-up while controlling for false discoveries asymptotically. This procedure efficiently controls for Type I and Type II errors in model selection. It is built upon the premise of FDR; however, it is adjusted for adaptive hyperparameters and accounts for the discrete feature space.

Benjamini and Hochberg (1995) define FDR as the expected proportion of F false positives out of the R rejected hypotheses:

$$FDR = E\left(\frac{F}{R}\middle| R > 0\right) Pr(R > 0). \tag{1}$$

The authors further provide a sequential method to control the FDR based on the p-values. This approach is later improved for power and number of rejections while controlling the FDR at the same target level. Storey (2002) proposes a practical method to estimate the FDR for a given set of hypotheses. Following this approach and for a given set of p-values $p_i, i = 1, \ldots, m$, the size of discoveries R is given by counting the number of p-values p_i within the rejection region $[0, \gamma]$:

$$R(\gamma) = \#\{p_i \leq \gamma\}. \tag{2}$$

The p-values are estimated from alternative resampling approaches, including bootstrap, permutation tests, and randomization.[8] The estimate of false positives is given by:

$$\hat{F}(\gamma) = \hat{\pi}_0(\lambda)\,\gamma m, \tag{3}$$

where $\hat{\pi}_0(\lambda)$ is the estimated probability that the p-values come from the null distribution for a chosen λ. The null hypothesis proportion is given by:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)\,m} \tag{4}$$

Plugging $\hat{\pi}_0$ into Eq. (3), the estimated FDR can be computed as $\widehat{FDR}_\gamma = \hat{F}(\gamma)/R(\gamma)$. A stepwise procedure for rejecting multiple hypotheses while controlling the FDR is proposed by Storey et al. (2004). The authors propose sorting the p-values in ascending order $\hat{p}_{R_1} \leq \cdots \leq \hat{p}_{R_m}$ and reject the hypothesis $\hat{p}_i \leq \hat{p}_{R_1}, i = 1, \ldots, m$. The $\widehat{FDR}_{\gamma'}$ is then computed for $\gamma' = \hat{p}_{R_1}$ and compared with a target level of α. The procedure goes on to the second p-value, where $\gamma' = \hat{p}_{R_2}$, and the same steps are applied again. The algorithm completes when no further rejection

[6] The closest practice to using FDR-based approaches in volatility forecasting is in Hurlin et al. (2017). The authors form 'buckets' of firms with similar risk profiles (proxied by GARCH and/or VaR) by combining the MCS and Romano, Shaikh, and Wolf (2008b). The MHT procedure in Hurlin et al. (2017) is based on BRC and RW which are both later shown to be lacking power (Bajgrowicz & Scaillet, 2012). The merits of such an approach over the MCS have not been established. Hence, we use the MCS as a widely acknowledged benchmark to our FDR analysis and adopt the bucket forming practice as in Hurlin et al. (2017).

[7] One of the most common practices in volatility forecasting is to estimate a model based on historical data and project it forward (Bollerslev et al., 2016; Esposito & Cummins, 2016; Figlewski, 1997). Analysis of future predictions can reflect the model's strength. Investigating one-day ahead forecasts is an almost unanimous practice for trading and risk management purposes among practitioners (Christoffersen & Diebold, 2000).

[8] In this paper, the p-values for null hypothesis testing are calculated by the stationary bootstrap (Politis & Romano, 1994).

is possible. See Storey et al. (2004) for additional details on the model selection procedure.

The FDR approach of Storey (2002) is expanded by Barras et al. (2010) to include the probability of having false discoveries on both tails of a distribution. The key motive is to provide a framework to identify the truly significant hypotheses while controlling the FDR on the positive (FDR^{+}) and negative (FDR^{-}) sides. Consider an MHT framework for testing m individual hypotheses in the form of:

$$H_{0,i}: \varphi_i = 0 \quad \text{vs.} \quad H_{A,i}: \varphi_i \neq 0, i = 1, \ldots, m \tag{5}$$

where φ_i is the difference in forecast error for model i relative to a benchmark and $H_{0,i}$ and $H_{A,i}$ denote a null and an alternative hypothesis, respectively. A two-sided p-value is necessary to estimate the $FDR^{+/-}$. We follow Romano and Wolf (2016) and Leippold and Rueegg (2020) in using adjusted p-values for our MHT comparison. Based on a bootstrap resampling method with B replications, the estimated two-sided p-value is defined as:

$$\hat{p}_i = 2 \min \left(\frac{\#\left\{\varphi'_{b,i} \geq \hat{\varphi}_i\right\} + 1}{B+1}, \frac{\#\left\{\varphi'_{b,i} \leq \hat{\varphi}_i\right\} + 1}{B+1} \right), \quad b = 1, \ldots, B. \tag{6}$$

$\varphi'_{b,i}$ is the centred test statistic for hypothesis i over the replication b. The p-value is based on the number of times $\hat{\varphi}_i$ is exceeded by a bootstrap replication over a total number of replications B. Analogous to Eqs. (3) and (4), the number of rejections and false discoveries on the positive and negative sides are estimated by:

$$\hat{R}^{+}_{\gamma} = \#\left\{\hat{p}_i \leq \gamma, \varphi_i > 0\right\} \quad \text{and} \quad \hat{R}^{-}_{\gamma} = \#\left\{\hat{p}_i \leq \gamma, \varphi_i < 0\right\} \tag{7}$$

$$\hat{F}^{+/-}_{\gamma} = \frac{1}{2}\hat{\pi}_0(\lambda)\,\gamma m. \tag{8}$$

Finally, by plugging in $\hat{\pi}_0$ from Eq. (4) into Eq. (8), the $\widehat{FDR_{\gamma}}^{+} = \hat{F}^{+}_{\gamma}/\hat{R}^{+}_{\gamma}$ and $\widehat{FDR_{\gamma}}^{-} = \hat{F}^{-}_{\gamma}/\hat{R}^{-}_{\gamma}$.

The λ acts as a tuning parameter to adjust the estimation of the FDR in Eq. (8) and is crucial to the estimation. Under a continuous uniform distribution [0, 1] assumption for p-values, λ can be computed manually or automatically. The manual approach is to visually investigate the histogram of the p-values and select the λ as the cut-off point where the histogram becomes relatively flat. The automated approach chooses the λ based on the bootstrap performance in estimating the FDR (Storey, 2002). In this study, we follow the DRB approach of Liang (2016) to estimate λ. We consider n support points for λ between 0 and 1. The λ^* is set as the first element of $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ to reverse the decreasing pattern for the estimated null proportion, i.e., $\hat{\pi}_0(\lambda_l) \geq \hat{\pi}_0(\lambda_{l-1})$.

The $FDR^{+/-}$ has higher power and is more accurate than the simple FDR approach (Bajgrowicz & Scaillet, 2012; Barras et al., 2010; Sermpinis, Hassanniakalager, Stasinakis, & Psaradellis, 2021). However, Andrikogiannopoulou and Papakonstantinou (2019) recently questioned the applicability of the $FDR^{+/-}$ in performance evaluation and called for its use "with caution". Their study of mutual funds profitability finds that the FDR may lead to a wrong conclusion when the noise-to-signal ratio is high, and the analysis is made ignoring a discrete p-value space – as in the original $FDR^{+/-}$ of Barras et al. (2010). While the noise-to-signal concern is reduced in volatility forecasting compared to return analysis, a discrete p-value space in our application is plausible. Such a discrete space can originate from two sources: structural dependence between models in each class and family; the finite number of the bootstrap replications ($B = 1,000$ in Eq. (6)), which reduce the number of support points for the p-values. We acknowledge the findings in Andrikogiannopoulou and Papakonstantinou (2019) and, therefore, resolve the issue by using an adaptive approach to model selection: the $FDR^{+/-}$ framework with DRB treatment.[9] Furthermore, adaptiveness is of additional benefit when we study our pool of volatility models for different market conditions. The resulting method removes subjectivity from the practitioner, an issue that can lead to erratic results and unnecessary conservativeness (Liang, 2016; Liang & Nettleton, 2012). Our application of the MHT in volatility forecasting from an FDR perspective is comparable to Hurlin et al. (2017), forming sets or *buckets* of companies with the same risk profile by combining the MCS and the FDR procedure in Romano, Sheikh, and Wolf (2008b). See Appendix A for complete details on the bucket formation algorithm based on the FDR^{+} approach.

3.2. Pool of conditional volatility models

Given a price process x_t, intraday returns are calculated as $r_{t,i} = \log\left(x_{t,i}\right) - \log\left(x_{t,i-1}\right), i = 1, \ldots, M$ where M is the number of intraday prices used. Similarly, the daily return is given by $r_t = \log\left(x_{t,M}\right) - \log\left(x_{t,0}\right), t = 1, \ldots, K+1$ where $x_{t,M}$ and $x_{t,0}$ refer to the closing and opening price respectively for each trading day. The dataset is split into two sets of sub-samples: training (in-sample) and evaluation (testing). The first K observations are used as an in-sample set to estimate the return volatility model for the last observation (out-of-sample).

Given a σ-algebra $\mathcal{F}_{t-1}$, the conditional density function for $\{r_t\}$ based on information available at time $t-1$ is then given by $f(r|\mathcal{F}_{t-1})$. The conditional mean and variance are defined as $\mu_{r_t} = E(r_t|\mathcal{F}_{t-1})$ and $\sigma_t^2 = \text{var}(r_t|\mathcal{F}_{t-1})$ respectively.

Following Patton (2011) and Corsi and Renò (2012) (among others), we construct the volatility forecasting models based on the $\{r_t\}$ process. Financial time series are widely acknowledged to exhibit heavy tails (see Campbell and Hentschel (1992), Engle and Ng (1993)). Accordingly, we consider four alternative distributions for the innovations: (i) Gaussian, (ii) Student's t (Bollerslev, 1987), (iii) skewed student's t distribution (Hansen, 1994), and (iv) Generalized Error Distribution (GED) (Nelson, 1991).

[9] A similar combination of $FDR^{+/-}$ and DRB is proposed by Sermpinis et al. (2021). In the case of a continuous p-value space, the adjusted $FDR^{+/-}$ performs similar to the original method of Barras et al. (2010) asymptotically. For more detail, please refer to Liang (2016). Monte Carlo simulation results for the comparison of original and adjusted $FDR^{+/-}$ are available upon request.

Four common families of autoregressive models are used in this study – GARCH, SV, EWMA,[10] and HAR. The GARCH-MA and SV-MA are adopted from Chan and Grant (2016), where innovations are assumed to follow a first order MA process. The asymmetric SV with Leverage (SV-L) is adopted from Asai and McAleer (2011). The HAR model is adopted from Corsi (2009), and the HARQ is adopted from Bollerslev et al. (2016). We also consider a set of simple moving average (SMA) estimates of realized volatility, as in Patton (2011). In the SMA models, the daily realized volatility is estimated by the mean of squared daily open-to-close returns over a look-back period (between a week to six months). Similarly, we consider a HAR model with time-varying parameters (HAR-TVP) as an extension to HAR where the daily and weekly coefficients are adjusted for changes from monthly realized volatility. The HAR model with leverage (LHAR) is adopted from Corsi and Renò (2012). The HAR model with a jump component (HAR-J), along with its log and non-linear extensions, is adopted from Andersen, Bollerslev, and Diebold (2007). We further combine the HAR approach with a support vector regression (SVR) and artificial neural network (ANN) motivated by Wang, Athanasopoulos, Hyndman, and Wangs (2018). The latter forms nonlinear estimations of relationships between daily, weekly, and monthly components to estimate future realized volatility. As highlighted by Bucci (2020) and Kristjanpoller, Fadic, and Minutolo (2014), among others, simple ANN structures and their variants have been very widely used to forecast volatility since they can effectively approximate linear and nonlinear behaviours without prior knowledge of the data structure. Further, a large body of literature, such as Liu (2019) and others, has highlighted the effectiveness of SVRs for volatility forecasting. This is because combination forecasts can incorporate information from several individual forecasting models and reduce forecast volatility. We also consider a forecast combination (FC) approach proposed by Rapach, Strauss, and Zhou (2010) to generate a weighted average of all volatility forecasting models. Equations for each class of volatility models are given in Table 1.

In GARCH (p, q) models (Eqs. (10) to (20)), $p > 0$, $q > 0$, $\alpha_u \geq 0$, and $\beta_r \geq 0$. For $p = 0$, the GARCH process reduces to ARCH (q) as in Eq. (9). The ε_t is the zero-mean intraday (open-to-close) return residuals. In GJR-GARCH (Eq. (8)), $I_{\{\varepsilon_t<0\}}$ is a dummy variable accounting for leverage equal to the unit when the criterion $\varepsilon_t < 0$ is satisfied, or otherwise zero. In NGARCH and APARCH (Eqs. (19) and (20)), δ is an extra parameter estimated along with other parameters. In asymmetric models (Eqs. (13), (14), (15), (16), (18), (20), and (23)), the parameter η corresponds to the leverage effect. In EGARCH and SV-L (Eqs. (18) and (23)), $E\left|e_{t-u}\right| = (\pi/2)^{-1/2}$ given $e_t \sim \mathcal{N}(0, 1)$. This quantity should be computed accordingly for the other distributions (t-student, GED and skewed-t), as in Harvey and Sucarrat (2014). However, in this paper, the same quantity $(\pi/2)^{-1/2}$ is used as an approximation for other distributions by following Hansen and Lunde (2005).

In FIGARCH (Eq. (21)), $\alpha(L)$ and $\beta(L)$ are lag operators such that $\alpha(L) = \alpha_1 L^{(1)} + \cdots + \alpha_q L^{(q)}$ and $\beta(L) = \beta_1 L^{(1)} + \cdots + \beta_p L^{(p)}$. The $(1-L)^d$ corresponds to fractional lag with degree d in the interval $(0, 1)$. Further, $\zeta_t = \varepsilon_t^2 - \sigma_t^2$ is a zero-mean martingale process[11] representing innovations for the conditional variance. In SV models, $h_t = \log\left(\sigma_t^2\right)$ and μ_h is the unconditional mean of h_t. The considered lag for all GARCH and SV models is one and two. For EWMA models (Eq. (24)), a range of parameter υ - interpreted as the decay factor – is studied from the set $\{0.8, 0.82, \ldots, 0.98\}$. JP Morgan (1996) proposes a decay factor of 0.94 in their RM model for daily data. In the SMA models, as in Eq. (25), we consider a range of $p \in \{5, 10, 22, 63, 126\}$ for the number of days to estimate volatility. In HAR models (Eqs. (26) to (39)), RV_t is the sum squares of intraday returns. $RV_t^{(1)}$ and $RV_t^{(2)}$ are the weekly (past five working days) and monthly (past 22 working days) averages for RV_t. Our study uses 78 intraday returns to estimate RV_t. In HARQ models, $\beta_{1,t} = \beta_1 + \beta_{1Q} RQ_{t-1}$ where $RQ_t = \frac{M}{3}\sum_{i=1}^{M} r_{t,i}^4$, $M = 78$. In the HAR-TVP models, $\beta'_{1,t} = \beta_1 + \beta_{1M}\left|RV_{t-1} - RV_{t-1}^{(2)}\right|$ and $\beta'_{2,t} = \beta_2 + \beta_{2M}\left|RV_{t-1}^{(1)} - RV_{t-1}^{(2)}\right|$ as alternative time-varying parameters based on HARQ and SMA models. In the LHAR model, $\varepsilon_t^+ = \max(0, \varepsilon_t^+)$, $\varepsilon_t^- = \min(0, \varepsilon_t^-)$, $\tilde{\varepsilon}_t^+ = \max(\frac{1}{5}\sum_{u=1}^{5}\varepsilon_{t-u}, 0)$ and $\tilde{\varepsilon}_t^- = \min(\frac{1}{5}\sum_{u=1}^{5}\varepsilon_{t-u}, 0)$.[12] In HAR-J models (Eqs. (33) to (35)), the jump component is defined as $J_t = \max[RV_t(M) - BV_t(M), 0]$ where $BV_t(M)$ is the standardized realized bipower variation as $BV_t(M) = \mu_1^{-2}\Sigma_{i=1}^{M-1}\left|r_{t,i}\right|\left|r_{t,i+1}\right|$ and $\mu_1 = (\pi/2)^{-1/2}$. In SVR models (Eqs. (36) to (37)), we consider two alternative kernels (ϕ_{SVR}) of linear and Gaussian/radial basis function (RBF) transform of inputs (daily, weekly, monthly RVs) along with five choices for the width of the epsilon-insensitive band (ε_{SVR}) from the set $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. For ANN models (Eqs. (38) to (39)), we consider three activation functions (Θ_{ANN}) of none, sigmoid, and the hyperbolic tangent of inputs (daily, weekly, and monthly RVs) along with five choices for the number of neurons (N_{ANN}) from the set $\{1, 5, 10, 50, 100\}$.[13] Finally, for the FC models, we combine one-step-ahead forecasts from previous volatility forecasting models using a weighted average given by $w_{i,t} = \frac{\phi_{i,t}^{-1}}{\Sigma_{i=1}^{323}\left(\phi_{i,t}^{-1}\right)}$ where i $(i = 1, \ldots, 323)$ denotes the indice for each forecasting model considered, $\phi_{i,t} = \Sigma_{s=1}^{K}\theta_C^{K-s}\left|RV_{t-s+1} - \widehat{RV}_{t-s+1}\right|$, and θ_C is the exponential discounting factor of range $\{1, 0.9\}$.

Based on the above, $m = 325$ candidate models are generated by combining specifications for the respective

[10] RM is a specific case of the Exponential Weighted Moving Average (EWMA).

[11] Martingale corresponds to a sequence of random variables where the expected value for the next observation is equal to the present one or $E(\zeta_{t+1}|\zeta_1, \ldots, \zeta_t) = \zeta_t$.

[12] The original version of LHAR in Corsi and Renò (2012) has an added jump component. In our application, we simplify their model by dropping the jump component and use zero-mean returns for a better match with the rest of our volatility models.

[13] For all other choices of parameters and hyperparameters required to initiate and optimize SVR and ANN we adopt MATLAB R2022a default values. For instance for the ANN, we consider a feed-forward neural network model as in Glorot and Bengio (2010).

Table 1
List of GARCH, SV, EWMA, SMA, HAR, SVR, ANN, and FC models.

Model	Definition	Equation
ARCH	$\sigma_t^2 = \omega + \sum_{u=1}^{q} \mathfrak{a}_u \varepsilon_{t-u}^2 + e_t$	(9)
GARCH/GARCH-MA	$\sigma_t^2 = \omega + \sum_{u=1}^{q} \mathfrak{a}_u \varepsilon_{t-u}^2 + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r}^2 + e_t$	(10)
IGARCH	$\sigma_t^2 = \omega + \varepsilon_{t-1}^2 + \sum_{u=2}^{q} \mathfrak{a}_u(\varepsilon_{t-u}^2 - \varepsilon_{t-1}^2) + \sum_{r=1}^{p} \mathfrak{b}_r(\sigma_{t-r}^2 - \varepsilon_{t-1}^2) + e_t$	(11)
Taylor- Schwert	$\sigma_t = \omega + \sum_{u=1}^{q} \mathfrak{a}_u \lvert\varepsilon_{t-u}\rvert + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r} + e_t$	(12)
A-GARCH	$\sigma_t^2 = \omega + \sum_{u=1}^{q} [\mathfrak{a}_u \varepsilon_{t-u}^2 + \eta_u \varepsilon_{t-u}]^2 + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r}^2 + e_t$	(13)
NA-GARCH	$\sigma_t^2 = \omega + \sum_{u=1}^{q} \mathfrak{a}_u (\varepsilon_{t-u} + \eta_u \sigma_{t-u})^2 + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r}^2 + e_t$	(14)
TGARCH	$\sigma_t = \omega + \sum_{u=1}^{q} \mathfrak{a}_u \left[(1-\eta_u)\varepsilon_{t-u}^{+} + (1+\eta_u)\varepsilon_{t-u}^{-}\right]^2 + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r} + e_t$	(15)
GJR-GARCH	$\sigma_t^2 = \omega + \sum_{u=1}^{q} [\mathfrak{a}_u + \eta_u I_{\{\varepsilon_{t-u}<0\}}] \varepsilon_{t-u}^2 + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r}^2 + e_t$	(16)
log-GARCH	$\log \sigma_t = \omega + \sum_{u=1}^{q} \mathfrak{a}_u \lvert e_{t-u}\rvert + \sum_{r=1}^{p} \mathfrak{b}_r \log \sigma_{t-r} + e_t$	(17)
EGARCH	$\log \sigma_t^2 = \omega + \sum_{u=1}^{q} [\mathfrak{a}_u e_{t-u} + \eta_u(\lvert e_{t-u}\rvert - E\lvert e_{t-u}\rvert)] + \sum_{r=1}^{p} \mathfrak{b}_r \log \sigma_{t-r}^2 + e_t$	(18)
NGARCH	$\sigma_t^{\delta} = \omega + \sum_{u=1}^{q} \mathfrak{a}_u \lvert\varepsilon_{t-u}\rvert^{\delta} + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r}^{\delta} + e_t$	(19)
APARCH	$\sigma_t^{\delta} = \omega + \sum_{u=1}^{q} \mathfrak{a}_u [\lvert\varepsilon_{t-u}\rvert - \eta_u \varepsilon_{t-u}]^{\delta} + \sum_{r=1}^{p} \mathfrak{b}_r \sigma_{t-r}^{\delta} + e_t$	(20)
FI-GARCH	$[1 - \mathfrak{a}(L) - \mathfrak{b}(L)](1-L)^d \varepsilon_t^2 = \omega + [1 - \mathfrak{b}(L)]\zeta_t + e_t$	(21)
SV/SV-MA	$h_t = \mu_h + \sum_{r=1}^{p} \phi_r [h_{t-u} - \mu_h] + e_t$	(22)
SV-L	$h_t = \mu_h + \sum_{u=1}^{q} [\mathfrak{a}_u e_{t-u} + \eta_u(\lvert e_{t-u}\rvert - E\lvert e_{t-u}\rvert)] + \sum_{r=1}^{p} \mathfrak{b}_r [h_t - \mu_h] + e_t$	(23)
EWMA	$\sigma_t^2 = \upsilon \sigma_{t-1}^2 + (1-\upsilon)\varepsilon_{t-1}^2 + e_t$	(24)
SMA	$\sigma_t^2 = \frac{1}{p} \sum_{r=1}^{p} \mathfrak{r}_{t-r}^2 + e_t$	(25)
HAR	$RV_t = \omega + \mathfrak{b}_1 RV_{t-1} + \mathfrak{b}_2 RV_{t-1}^{(1)} + \mathfrak{b}_3 RV_{t-1}^{(2)} + e_t$	(26)
log-HAR	$\log RV_t = \omega + \mathfrak{b}_1 \log RV_{t-1} + \log \mathfrak{b}_2 RV_{t-1}^{(1)} + \log \mathfrak{b}_3 RV_{t-1}^{(2)} + e_t$	(27)
HARQ	$RV_t = \omega + \mathfrak{b}_{1,t} RV_{t-1} + \mathfrak{b}_2 RV_{t-1}^{(1)} + \mathfrak{b}_3 RV_{t-1}^{(2)} + e_t$	(28)
log-HARQ	$\log RV_t = \omega + \mathfrak{b}_{1,t} \log RV_{t-1} + \log \mathfrak{b}_2 RV_{t-1}^{(1)} + \log \mathfrak{b}_3 RV_{t-1}^{(2)} + e_t$	(29)
HAR-TVP	$RV_t = \omega + \mathfrak{b}'_{1,t} RV_{t-1} + \mathfrak{b}'_{2,t} RV_{t-1}^{(1)} + \mathfrak{b}_3 RV_{t-1}^{(2)} + e_t$	(30)
log-HAR-TVP	$\log RV_t = \omega + \mathfrak{b}'_{1,t} \log RV_{t-1} + \log \mathfrak{b}'_2 RV_{t-1}^{(1)} + \log \mathfrak{b}_3 RV_{t-1}^{(2)} + e_t$	(31)
LHAR	$\log RV_t = \omega + \mathfrak{b}_1 \log RV_{t-1} + \log \mathfrak{b}_2 RV_{t-1}^{(1)} + \log \mathfrak{b}_3 RV_{t-1}^{(2)} + \eta_1^{+} \varepsilon_{t-1}^{+} + \eta_1^{-} \varepsilon_{t-1}^{-} + \eta_2^{+} \tilde{\varepsilon}_{t-1}^{+} + \eta_2^{-} \tilde{\varepsilon}_{t-1}^{-} + e_t$	(32)
HAR-J	$RV_t = \omega + \mathfrak{b}_1 RV_{t-1} + \mathfrak{b}_2 RV_{t-1}^{(1)} + \mathfrak{b}_3 RV_{t-1}^{(2)} + \mathfrak{b}_4 J_{t-1} + e_t$	(33)
log-HAR-J	$\log RV_t = \omega + \mathfrak{b}_1 \log RV_{t-1} + \mathfrak{b}_2 \log RV_{t-1}^{(1)} + \mathfrak{b}_3 \log RV_{t-1}^{(2)} + \mathfrak{b}_4 \log(1 + J_{t-1}) + e_t$	(34)
Sqrt-HAR-J	$(RV_t)^{0.5} = \omega + \mathfrak{b}_1 (RV_{t-1})^{0.5} + \mathfrak{b}_2 \left(RV_{t-1}^{(1)}\right)^{0.5} + \mathfrak{b}_3 \left(RV_{t-1}^{(2)}\right)^{0.5} + \mathfrak{b}_4 (J_{t-1})^{0.5} + e_t e_t$	(35)
SVR-HAR	$RV_t = SVR(RV_{t-1}, RV_{t-1}^{(1)}, RV_{t-1}^{(2)}, \phi_{SVR}, \varepsilon_{SVR}) + e_t$	(36)
log-SVR-HAR	$\log RV_t = \mathrm{SVR}(\log RV_{t-1}, \log RV_{t-1}^{(1)}, \log RV_{t-1}^{(2)}, \phi_{SVR}, \varepsilon_{SVR}) + e_t$	(37)
ANN-HAR	$RV_t = \mathrm{ANN}(RV_{t-1}, RV_{t-1}^{(1)}, RV_{t-1}^{(2)}, N_{ANN}, \Theta_{ANN}) + e_t$	(38)
log-ANN-HAR	$\log RV_t = \mathrm{ANN}(\log RV_{t-1}, \log RV_{t-1}^{(1)}, \log RV_{t-1}^{(2)}, N_{ANN}, \Theta_{ANN}) + e_t$	(39)
FC	$\sigma_t^2 = \Sigma_{i=1}^{323} w_i(\theta_C) \hat{\sigma}_{i,t}^2 + + e_t$	(40)

Note: The table presents the 34 classes of volatility models considered in our pool of volatility models. The specifications of individual models are presented in *Appendix B*.

parameters. We consider three choices for the number of observations in the training sample to estimate volatility forecasting models $K \in \{252, 181, 91\}$ to predict out-of-sample one-step-ahead realized volatility. As a proxy for true realized volatility, we use $\sigma_t^2 = \sum_i \mathfrak{r}_{t,i}^2$, $i = 1, \ldots, M$. Following Patton (2011) and Bollerslev et al. (2016), we consider three choices of $M \in \{78, 26, 13\}$ to establish the role of intraday return frequency. Four innovation distributions are then considered for the 34 classes of the forecasting models. See Appendix B for additional details regarding the specifications and the characteristics of the models considered in this pool.

3.3. Data

We conduct our empirical analysis using four U.S. financial indices: two stocks (S&P 500, NASDAQ); and two commodities (gold, light oil). We benchmark the series by exchange-traded funds (ETFs) tracking indices in each category of assets. Rather than spot indices, ETFs ensure that the results are relevant to academics and practitioners. The ETFs are chosen to have the largest assets under management (AUM). All selected ETFs are traded on the New York Stock Exchange, and we use the prices reported over trading hours between 09:30 and 16:00 ET. The description of the studied ETFs and summary statistics for the daily range (open to close) logarithmic returns of the time series are presented in Table 2. The daily realized volatilities are constructed as the sum of $M = 78$ squared returns between the market open and close times following recent practice (see, for example, Andersen, Bollerslev, Diebold, & Ebens, 2001; Andersen, Thyrsgaard, & Todorov, 2019; Bollerslev et al., 2016; Corsi & Renò, 2012; Li & Xiu, 2016; Zhang, Ma, & Liao, 2020). The latter corresponds to using 5-minute log returns. We further consider $M = 26$ and $M = 13$ corresponding to using 15-minute and 30-minute returns as alternative proxies for realized volatility. The dataset includes price series from 1 January 2013 to 31 December 2020. The series are from the New York Stock Exchange Trade and Quotes (TAQ) and has been obtained via the Wharton Research Data Services (WRDS). The training period is set to one year ($K = 252$) and is used to predict the one-step-ahead conditional volatility as described in Section 3.2. We roll forward the estimation each day and repeat the process. Hence the training sample size remains $K = 252$ in all analyses.[14] The testing period spans over seven calendar

[14] The only exception is in Section 4.5 where we evaluate the role of K on the QLIKE performance of benchmarks and MHT buckets.

Table 2
Summary statistics of log daily range returns.

Ticker	SPY	QQQ	GLD	USO
Description	SPDR S&P 500 Trust	Invesco Nasdaq Trust	SPDR Gold Shares	U.S. Oil Fund
AUM (B)	$367.34	$151.52	$48.38	$2.13
Mean (%)	0.02	0.03	−0.01	−0.04
Median (%)	0.06	0.07	−0.01	0.00
Maximum (%)	4.65	4.95	5.25	11.62
Minimum (%)	−5.77	−6.27	−4.15	−14.44
Std. Dev. (%)	0.75	0.93	0.61	1.66
Skewness	−0.53	−0.55	0.29	−0.34
Excess Kurtosis	7.09	4.62	7.95	8.65
Jarque-Berra	4,312.46 ***	1,892.49 ***	5,339.98 ***	6,318.61 ***
Q (20)	74.68 ***	80.7 ***	40.34 **	31.4
ADF	−11.64 ***	−12.34 ***	−17.76 ***	−45.22 ***
P-P	−48.75 ***	−51.81 ***	−44.74 ***	−45.22 ***

Note: The AUM figures are denoted in billion dollars as of 31st October 2022 using Refinitiv estimates. The sample dataset used for this summary statistics table spans from 1st January 2013 to 31st December 2020. The Jarque-Berra statistic tests whether the skewness and kurtosis of a sample dataset match the normal distribution. Q (20) is the Ljung–Box statistic testing if the data is distributed independently. Serial correlation of order up to the 20th is considered. ADF and P-P are the statistics of the augmented Dickey-Fuller and Phillips-Perron unit root tests, respectively. The lag length for the unit root tests is set based on the lowest Akaike Information Criteria (AIC) value. *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels respectively.

years (2014 to 2020). We compare the volatility forecasting models over the entire testing period (Section 4.2). We then investigate the robustness of our results subject to M (Section 4.3), K (Section 4.4), horizon h for forecasting (Section 4.5), to periods of high and low financial market stress (Section 4.6), and the role of time (Section 4.7) to explore the dynamics in performance.

Table 2 presents summary statistics for the studied series. The mean and median returns are positive for all assets except for the oil index. The standard deviation is the highest for the oil ETF and lowest for the gold ETF. All securities are leptokurtic (except for GLD) with negative skewness. Accordingly, the Jarque and Bera (1980) test rejects the null hypothesis of a normal distribution for all series. Ljung and Box (1978) statistics reject serial independence for three out of four indices. The unit root test statistics of the Augmented Dickey and Fuller (1979) (ADF) and the Phillips and Perron (1988) (P-P) suggest that all return series are stationary.

3.4. Performance metrics

It is common in the volatility forecasting literature to use a loss function and benchmark to compare the cross-sectional performance of volatility models (Bollerslev et al., 2016; Brooks & Persand, 2003; Hansen & Lunde, 2005; Huang, Gong, Chen, & Wen, 2013; Wang, Pan, & Wu, 2018; Wei et al., 2010). Choosing the most appropriate loss function is challenging as no formal theory supports such a selection (Lopez, 2001). Patton (2011) introduces a series of robust loss functions to evaluate the performance of the volatility models we use in this study. For a chosen parameter β, an estimated realized volatility $\hat{\sigma}^2$ and a conditional realized variance σ^2, the robust loss (RL) is given by:

$$\mathrm{RL}\left(\hat{\sigma}^2, \sigma^2; \beta\right) = \begin{cases} \sigma^2 - \hat{\sigma}^2 + \hat{\sigma}^2 \log \frac{\hat{\sigma}^2}{\sigma^2}, & \text{for } \beta = -1 \\ \frac{\hat{\sigma}^2}{\sigma^2} - \log \frac{\hat{\sigma}^2}{\sigma^2} - 1, & \text{for } \beta = -2 \\ \frac{1}{(1+\beta)(2+\beta)}\left(\hat{\sigma}^{2(\beta+2)} - \sigma^{2(\beta+2)}\right) & \\ \quad -\frac{\sigma^{2(\beta+1)}}{(1+\beta)}\left(\hat{\sigma}^2 - \sigma^2\right), & \text{otherwise} \end{cases} \tag{41}$$

Following Patton (2011), we study a set of $\beta \in \{-5, -2, 0, 1\}$ to avoid biased findings. The choices of $\beta = -2$ and $\beta = 0$ correspond to the widely used quasi-like (QLIKE) and mean square error (MSE) loss functions, respectively.[15] We define our test statistic relative to a benchmark as:

$$\hat{\varphi}_i = \mathrm{RL}_0 - \mathrm{RL}_i, i = 1, \ldots, m \tag{42}$$

where RL_i is the calculated robust loss function for the candidate model i compared to the respective benchmark, RL_0.We compute the RL_is as in Eq. (41) by comparing forecasted realized volatility from each model $\hat{\sigma}_i^2, i = 1, \ldots, 325$ relative to the target realized volatility σ^2 – proxied by the sum of M squared log returns between daily open and close based on N out-of-sample observations.[16] We use the proposed setting in Eq. (42) for MHT approaches involving a benchmark, namely, FDR$^{+/-}$ and kStepM. The test statistic in Eq. (42) presents a gain

[15] The MSE and QLIKE can have alternative formulation with a multiplication or an additive element. In this study, we define the QLIKE as the RL with $\beta = -2$ and the MSE as the RL with $\beta = 0$ as in Eq. (41).

[16] This definition of realized volatility focuses on conditional variance as in Patton (2011) rather than the concept of 'integrated volatility' as in Zhang, Mykland, and Aït-Sahalia (2005) and Andersen et al. (2019).

function relative to the benchmark where a positive figure reflects a marginal advantage in using model *i*. The benchmark choice is crucial when using a MHT framework. We use three benchmarks to track the variations in the size of buckets. The benchmarks are GARCH(1,1) as in Eq. (9), GJR-GARCH(1,1) as in Eq. (16) and basic HAR as in Eq. (26). Both the GARCH and GJR-GARCH benchmarks are with Gaussian innovation. These choices allow for comparing our candidate models and the standard benchmarks in the previous literature (e.g., Hansen & Lunde, 2005). This enables us to assess for any significant differences between the top-performing models. The null hypothesis in Eq. (5) is then redefined to test whether a candidate can provide better accuracy compared to the alternative benchmarks of GARCH, GJR-GARCH, and HAR.

We use the robust loss and the test statistic in Eqs. (41) and (42), respectively, to compare volatility forecasting models. We form buckets of volatility forecasting models with similar statistical performance – statistically more accurate than the respective benchmark model – using the FDR^{+} as our testing framework. We then evaluate the characteristics of the buckets of volatility forecasting models with those of GARCH, GJR-GARCH, and HAR. To investigate the merits of the FDR^{+} relative to alternative MHT frameworks, we replicate the same bucket creation process using *k*StepM and MCS tests.

The selected MHT procedures (FDR^{+}, *k*StepM, MCS) are structurally different. The MCS does not involve a benchmark in its stepwise process and aims to form a confidence set of alternatives with similar performance. Accordingly, rather than rejecting the significantly superior models, the MCS drops (rejects) the models that are inferior to the set. We consider the set of models included in the confidence set as our MCS bucket. The *k*StepM procedure controls $\Pr(FWER \geq k)$ at a target confidence level. We use the FDP-controlling algorithm (the *k*StepM−FDP) in RSW at $\gamma = 0.1$ to match the FDR^{+} setting. In Appendix C, we investigate setting k at three different arbitrary levels relative to the FDP-based approach empirically used in Section 4.[17] Unlike the MCS, the *k*StepM is a relative test (like $FDR^{+/-}$) which allows testing relative to a benchmark. The first point of interest in choosing alternative MHTs is seeing the role of a benchmark, and the second is whether an FDR or FWER can outperform others.

4. Results

We begin by plotting the performance of our pool based on the robust loss functions QLIKE (Section 4.1). We then use three seminal MHT procedures – the MCS, the *k*StepM, and the FDR^{+} – to form buckets of similarly performing volatility forecasting models (Section 4.2). We use the MCS and the *k*StepM as our comparative FWER model selection frameworks. The MCS does not involve a benchmark and drops the significantly worse-performing alternatives, whereas the *k*StepM can involve a benchmark for comparison. We then compare the buckets of MCS and *k*StepM with those of the FDR^{+}. This comparison supports the role of MHT in volatility forecasting models and the performance improvement (accuracy) of our proposed FDR model selection framework relative to an FWER-based approach. In Section 4.3, we investigate how reducing the number of intraday returns (M) affects the performance of forecasting models. In Section 4.4, we investigate how reducing the number of observations (K) to estimate the 325 volatility forecasting models affects the models. Section 4.5 investigates how increasing the forecast horizon from one step to a week (5 periods) and a month (22 periods) impacts our findings. In Section 4.6, we investigate the robustness of our findings to disruption in the normal operations of financial markets (high versus low financial market stress.[18]) Section 4.7 investigates how time dynamics from 2014 to 2020 increase/decrease bucket sizes and QLIKE performances. In Appendix D, we analyze the robustness of the methodology's ability to select superior volatility forecasting models irrespective of the choice of the loss function. Finally, in Appendix E, we characterize models selected by MCS and FDR buckets by the family (GARCH, SV, EWMA, HAR, SVR, ANN, SMA, or FC).

4.1. Model performance

As the first step in evaluating the pool's performance, we depict a box plot of the volatility models using a robust loss function as in Eq. (41). The aim is to examine the characteristics of the distribution and identify clusters with increased density. Fig. 1 exhibits the QLIKE ($\beta = -2$) performance range for the 325 models over the testing period.

Fig. 1 shows that the population distribution is characterized by a wide range, sparsity and increased density around certain points. This is consistent with the clustered description of Corsi and Renò (2012). The wide range for equity and commodity ETFs is due to a proportion of the under-/out-performing models. The sparsity and increased density are due to the structural dependence in the volatility pool and the modest changes between similar models. The sparse pattern of the sample loss provides empirical evidence for choosing the DRB treatment for the $FDR^{+/-}$ approach, as discussed in Section 3.1.

Based on Fig. 1, there is preliminary evidence of a range of performances across volatility models. The question to be answered remains whether these differences are statistically significant. The following section uses the selected MHT procedures to find volatility buckets of forecasting models with similar performance profiles.

4.2. Model testing and bucket formation

The current literature investigating volatility forecasting models is confined to conventional FWER-based

[17] In Appendix C we show the variations in the *k*StepM bucket size for $k = 1, 5$, and 10 as three arbitrary choices. In the same appendix, we further compare the *k*StepM buckets to those of an SPA test. The findings in Appendix C shows how reducing the k can make the test conservative and lead to detecting no model appearing to outperform any of the benchmarks.

[18] We use the Office of Financial Research's Financial Stress Index for volatility as a measure of stress.

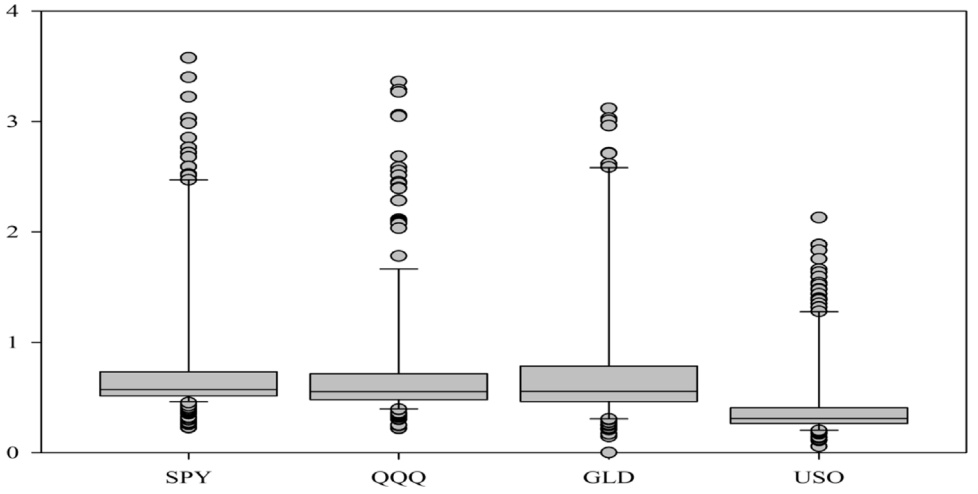

Fig. 1. Box plot for QLIKE.
Note: The figure presents variability in average robust loss across the 325 volatility forecasting models. The QLIKE for each model is calculated over the entire testing sample (2014–2020) as in Eq. (41).

procedures. As discussed in Section 2, we propose using the FDR$^+$ model selection framework as superior to FWER-based procedures. This section provides evidence for applying the MHT in selecting volatility forecasting models. For this purpose, we form buckets of volatility models with similar performance profiles – greater forecast accuracy than the benchmark models – based on three MHT procedures: FDR procedure, *k*StepM, and MCS.

To compare the buckets from the different MHT procedures, we use a stationary bootstrap approach with 1,000 replications and a block size of 10 working days. We use four levels of β in Eq. (41) to create our test statistic and set $\alpha = 0.1$. For the FDR$^+$, if no model is selected, the bucket consists of only the respective benchmark (GARCH, GJR-GARCH, or HAR) as the top-performing model.[19] Table 3 presents the average and standard deviations for the bucket size (the number of models with a similar performance profile) by the MCS, the *k*StepM and the FDR$^+$. In Appendix D, we exhibit the variations of results based on specific choices of β.

Table 3 shows that the bucket size varies by the MHT procedure. The FDP-controlling procedure of *k*StepM does not detect any model to perform significantly better than the three benchmarks despite aiming to control FWE and FDP. In contrast, both the MCS and the FDR$^+$ find tens of true discoveries. The results for *k*StepM reflect the restrictive nature of this procedure for empirical analyses. We also observe large bucket size variations across the MCS and the FDR$^+$ approaches. The size of the MCS bucket is consistently higher than the FDR$^+$. However, the wide range of bucket sizes in Table 3 raises the question of which testing framework selects the most accurate models. We investigate this in Table 4 by computing the average QLIKE for the buckets of superior-performing models selected by the FDR and MCS methods.

[19] The MCS does not require such consideration since it rejects the underperforming models. In case all alternatives are similar, the confidence set includes all models.

The results in Table 4 show that the lowest QLIKE for all ETFs are for the $FDR^+_{\text{GJR}-\text{GARCH}}$ and FDR^+_{GARCH}, respectively. While all buckets have, on average, lower QLIKE compared to GARCH, the MCS does not provide significantly different performance. We also note that the FDR$^+$ method has the lowest average forecast error across all benchmarks (GARCH, GJR-GARCH, HAR) compared to the MCS (see column *Average* in Table 4). Hence the first conclusion from this section is the significant evidence to support using FDR$^+$ in identifying volatility forecasting models with superior performance.

Our second finding is the large noise-to-signal ratio for the MHT buckets. For instance, the standard deviation is 59% (40%) of the mean bucket size for the MCS ($FDR^+_{\text{GJR}-\text{GARCH}}$) in Table 4. The same observation is made by Andrikogiannopoulou and Papakonstantinou (2019) regarding the use of FDR$^+$. The high noise-to-signal ratio is plausibly attributable to the dynamic nature of financial markets, which leads to significant variation in model performance over time. This potential explanation further speaks to the study of the effect of financial market stress and time on buckets of volatility forecasting models as in Section 4.6 and Section 4.7, respectively. Alternatively, the high ratio could be due to using a conservative testing framework underestimating the proportion of significantly better models or buckets of models, as suggested by Andrikogiannopoulou and Papakonstantinou (2019).

The findings in this section also enable us to show that both the MCS and FDR$^+$ frameworks can identify buckets of superior-performing volatility models relative to GARCH and HAR benchmarks for all assets. However, compared to the GJR-GARCH, the MCS does not offer an advantage in reducing forecast error for USO. The evidence also shows that the $FDR^+_{\text{GJR}-\text{GARCH}}$ performs superior to alternative models, with the least error in forecasting one-step-ahead realized volatility for equity index ETFs (SPY and QQQ). More generally, the results in this section establish the benefits of using an MHT approach in

Table 3
The bucket size under different testing frameworks ($N = 1,763$).

Method	*MCS*	FDR^{+}			*kStepM*		
Benchmark	–	GARCH	GJR-GARCH	HAR	GARCH	GJR-GARCH	HAR
SPY	149 (164)	25 (30.99)	18 (19.9)	108 (141.71)	0 (0)	0 (0)	0 (0)
QQQ	161.75 (149.46)	19 (35.34)	11.5 (20.34)	125 (149.12)	0 (0)	0 (0)	0 (0)
GLD	116 (131.21)	23.75 (28.18)	23.5 (28.45)	24 (28.31)	0 (0)	0 (0)	0 (0)
USO	211.5 (141.59)	15.25 (28.5)	9.75 (17.5)	131 (151.49)	0 (0)	0 (0)	0 (0)
Average	159.56 (136.22)	20.75 (27.92)	15.69 (20.42)	97 (123.22)	0 (0)	0 (0)	0 (0)

Note: The table presents the mean and standard deviation of true discoveries for FDR^{+}, MCS, and *k*StepM. For instance, the top left figure of 149 shows the average number of discoveries for SPY by the MCS procedure over the study period 2014–2020. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations. See *Appendix E*, for families (GARCH, SV, EWMA, SMA, HAR, SVR, ANN, or FC) of selected volatility forecasting models by the MHT procedures.

Table 4
The QLIKE performance under different testing frameworks ($N = 1,763$).

Method	SPY	QQQ	GLD	USO	Average
MCS	0.44 (0.27)	0.45 (0.2)	0.3 (0.25)	0.28 (0.16)	0.37 (0.22)
FDR^{+}_{GARCH}	0.27 (0.08)	0.24 (0.08)	0.37 (0.12)	0.2 (0.04)	0.27*** (0.1)
GARCH	0.57 (0)	0.52 (0)	0.49 (0)	0.3 (0)	0.47 (0.11)
$FDR^{+}_{\text{GJR-GARCH}}$	0.24 (0.04)	0.21 (0.05)	0.33 (0.17)	0.21 (0.06)	0.25*** (0.1)
GJR − GARCH	0.51 (0)	0.47 (0)	0.47 (0)	0.27 (0)	0.43 (0.1)
FDR^{+}_{HAR}	0.43 (0.31)	0.4 (0.27)	0.37 (0.12)	0.26 (0.12)	0.36* (0.21)
HAR	0.59 (0)	0.62 (0)	0.47 (0)	0.38 (0)	0.52 (0.1)

Note: The table presents the mean and standard deviation for the performance of the selected models by different methods. For instance, the top left figure of 0.44 shows the average QLIKE for SPY by the MCS over the study period 2014–2020. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations. The *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels, respectively, for a one-sided paired t-test with the null hypothesis that the GARCH(1,1) has the equal or less average QLIKE as alternative forecasting models across all ETFs.

forecasting volatility. Therefore, we call for further investigation into using MHT procedures in other econometric applications.

4.3. Construction of realized volatility

Section 3.2 establishes how M intraday returns are used to form target realized volatility for each day between market open and close using the sum of squared five-minute log returns ($M = 78$). Patton (2011) argues that using the robust loss functions as in Eq. (41) leads to an unbiased estimation of the forecast error with respect to true realized volatility for all levels of M. To further establish the robustness of our proposed methodology, in this section, we investigate how the performance of the volatility forecasting models varies by reducing the number of intraday returns to $M = 26$ and $M = 13$ corresponding to 15- and 30-minute returns, respectively.

Table 5 presents QLIKE for the MHT buckets and three benchmarks when lower levels of M are considered. The table shows that the performance of all volatility forecasting models is adversely affected when M is reduced. For instance, in comparison to the results in Table 4, on average, across the four assets with $M = 26$, QLIKE is increased by 43% for MCS and 48% for $FDR^{+}_{\text{GJR-GARCH}}$. For $M = 26$, the FDR-based buckets continue to outperform the MCS for all cases. When $M = 13$, the performance further deteriorates across all forecasting models and buckets. Yet the cross-sectional performance (superior performance of FDR-based approaches) remains unchanged.

Findings in Table 5 are comparable to those in Lyócsa, Molnár, and Výrost (2021), arguing that high-frequency volatility models outperform low-frequency volatility models for short-term forecasts (i.e., one day ahead in our case).

4.4. Training sample size

Section 3.2 establishes how K training observations form 325 forecasting models to forecast 1 out-of-sample realized volatility. In this section, we investigate whether volatility forecasting models and buckets are affected by decreasing sample size. In Table 6, we replicate the previous analysis for two random choices of $K \in \{91, 182\}$.

Table 6 shows that forecast error increases when the number of training observations is reduced. It is worth

Table 5
The QLIKE performance under different testing frameworks ($N = 1,763$).

Setting	Method	SPY	QQQ	GLD	USO	Average
$M = 26$	*MCS*	0.58 (0.32)	0.57 (0.26)	0.47 (0.36)	0.5 (0.33)	0.53* (0.29)
	FDR^{+}_{GARCH}	0.4 (0.16)	0.37 (0.15)	0.46 (0.22)	0.32 (0.08)	0.39*** (0.15)
	GARCH	0.77 (0)	0.69 (0)	0.73 (0)	0.51 (0)	0.67 (0.1)
	$FDR^{+}_{GJR-GARCH}$	0.35 (0.1)	0.35 (0.14)	0.46 (0.22)	0.35 (0.13)	0.38*** (0.15)
	GJR − GARCH	0.7 (0)	0.62 (0)	0.71 (0)	0.48 (0)	0.63 (0.1)
	FDR^{+}_{HAR}	0.49 (0.41)	0.43 (0.31)	0.46 (0.22)	0.47 (0.19)	0.46** (0.26)
	HAR	0.61 (0)	0.56 (0)	0.71 (0)	0.58 (0)	0.61* (0.06)
$M = 13$	*MCS*	0.77 (0.41)	0.81 (0.31)	0.72 (0.42)	0.68 (0.42)	0.74* (0.35)
	FDR^{+}_{GARCH}	0.48 (0.2)	0.51 (0.22)	0.75 (0.24)	0.52 (0.15)	0.57*** (0.22)
	GARCH	1.03 (0)	0.93 (0)	1.04 (0)	0.74 (0)	0.94 (0.12)
	$FDR^{+}_{GJR-GARCH}$	0.41 (0.11)	0.48 (0.2)	0.75 (0.24)	0.52 (0.16)	0.54*** (0.21)
	GJR − GARCH	0.94 (0)	0.85 (0)	1.03 (0)	0.71 (0)	0.88 (0.12)
	FDR^{+}_{HAR}	0.64 (0.52)	0.58 (0.4)	0.75 (0.24)	0.57 (0.14)	0.63*** (0.33)
	HAR	0.82 (0)	0.78 (0)	1.08 (0)	0.8 (0)	0.87 (0.12)

Note: The table presents the mean and standard deviation for the performance of selected models by different methods. For instance, the top left figure of 0.58 shows the average QLIKE for SPY by the MCS when the target realized volatility is constructed using 26 intraday returns. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations. The *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels, respectively, for a one-sided paired t-test with the null hypothesis that the GARCH(1,1) has equal or less average QLIKE than alternative forecasting models across all ETFs.

Table 6
The QLIKE performance under different numbers of training observations ($N = 1,763$).

Setting	Method	SPY	QQQ	GLD	USO	Average
$K = 182$	*MCS*	0.47 (0.25)	0.49 (0.2)	0.31 (0.26)	0.29 (0.15)	0.39 (0.22)
	FDR^{+}_{GARCH}	0.3 (0.1)	0.28 (0.08)	0.3 (0.1)	0.23 (0.1)	0.28*** (0.09)
	GARCH	0.58 (0)	0.56 (0)	0.46 (0)	0.33 (0)	0.48 (0.1)
	$FDR^{+}_{GJR-GARCH}$	0.24 (0.03)	0.24 (0.03)	0.35 (0.11)	0.23 (0.11)	0.26*** (0.09)
	GJR − GARCH	0.52 (0)	0.49 (0)	0.47 (0)	0.29 (0)	0.44 (0.09)
	FDR^{+}_{HAR}	0.44 (0.33)	0.44 (0.27)	0.36 (0.11)	0.32 (0.14)	0.39 (0.21)
	HAR	0.59 (0)	0.65 (0)	0.47 (0)	0.4 (0)	0.53 (0.1)
$K = 91$	*MCS*	0.49 (0.31)	0.56 (0.23)	0.36 (0.17)	0.72 (0.79)	0.53 (0.42)
	FDR^{+}_{GARCH}	0.23 (0.04)	0.32 (0.12)	0.34 (0.14)	0.31 (0.28)	0.3** (0.15)
	GARCH	0.56 (0)	0.56 (0)	0.45 (0)	0.27 (0)	0.46 (0.12)
	$FDR^{+}_{GJR-GARCH}$	0.23 (0.03)	0.31 (0.12)	0.34 (0.14)	0.31 (0.28)	0.3** (0.16)
	GJR − GARCH	0.5 (0)	0.49 (0)	0.43 (0)	0.26 (0)	0.42 (0.1)
	FDR^{+}_{HAR}	2.8 (1.24)	2.06 (0.8)	5.14 (2.59)	0.74 (0.2)	2.68 (2.12)
	HAR	326.22 (0)	193.51 (0)	363.82 (0)	50.83 (0)	233.6 (127.07)

Note: The table presents the mean and standard deviation for the performance of selected models by different methods when the training sample size is reduced. For instance, the top left figure of 0.47 shows the average QLIKE for SPY by the MCS over the study period 2014–2020 when $K = 182$. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. The *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels, respectively, for a one-sided paired t-test with the null hypothesis that the GARCH(1,1) has an equal or less average QLIKE than alternative forecasting models across all ETFs.

noting that reducing K from 252 (in Table 4) to 182 (as in Table 6) marginally worsens the performance for all forecasting models and buckets. For instance, the average QLIKE across all four assets for MCS (FDR^{+}_{GARCH}) in Table 6 is 0.39 (0.28) compared to 0.37 (0.27) in Table 4. Yet $K = 91$ leads to a larger deterioration of QLIKE with the largest change in performance for the HAR model (233.6 on average in Table 6 compared to 0.52 in Table 4). The latter observation can be attributed to the construct of the HAR model involving an annual component which may require a proportionally higher number of observations to specify regression coefficients correctly. As a final observation, our preferred FDR method outperforms the MCS approach.

4.5. Multi-step forecasting

The pool of 325 volatility forecasting models has thus far used K observations constructed by using M intraday returns to estimate one step ahead of daily realized volatility. In this section, we follow Lyócsa et al. (2021) to consider one-step ahead forecasts as estimated arithmetic averages across h next daily realized volatility and evaluate the performance of alternative volatility forecasting models and buckets for making multiple-step-ahead forecasts.

Accordingly, we start by using $M = 78$ and $K = 252$ to make a terminal forward-looking forecast for realized volatility. We then compare this forecast with the average actual realized volatility over $h \in \{5, 22\}$ periods. Next,

Table 7
The QLIKE performance under different numbers of training observations.

Setting	Method	SPY	QQQ	GLD	USO	Average
$h = 5$	*MCS*	0.4 (0.18)	0.29 (0.17)	0.18 (0.13)	0.31 (0.15)	0.29 (0.16)
	FDR^{+}_{GARCH}	0.19 (0.04)	0.25 (0.14)	0.13 (0.1)	0.08 (0.08)	0.16*** (0.11)
	GARCH	0.4 (0)	0.34 (0)	0.25 (0)	0.19 (0)	0.29 (0.08)
	$FDR^{+}_{\text{GJR-GARCH}}$	0.16 (0.03)	0.26 (0.13)	0.17 (0.15)	0.07 (0.08)	0.16*** (0.12)
	GJR − GARCH	0.35 (0)	0.31 (0)	0.25 (0)	0.16 (0)	0.27 (0.08)
	FDR^{+}_{HAR}	0.17 (0.02)	0.37 (0.13)	0.08 (0.09)	0.31 (0.36)	0.24 (0.21)
	HAR	0.38 (0)	0.39 (0)	0.23 (0)	0.25 (0)	0.31 (0.08)
$h = 22$	*MCS*	0.3 (0.16)	0.28 (0.16)	0.23 (0.27)	0.29 (0.18)	0.28 (0.18)
	FDR^{+}_{GARCH}	0.14 (0.08)	0.23 (0.09)	0.15 (0.08)	0.13 (0.07)	0.16** (0.08)
	GARCH	0.35 (0)	0.27 (0)	0.21 (0)	0.15 (0)	0.25 (0.08)
	$FDR^{+}_{\text{GJR-GARCH}}$	0.18 (0.14)	0.56 (0.87)	0.15 (0.16)	0.13 (0.07)	0.25 (0.44)
	GJR − GARCH	0.42 (0)	0.32 (0)	0.18 (0)	0.15 (0)	0.27 (0.11)
	FDR^{+}_{HAR}	0.3 (0.22)	0.32 (0.14)	0.18 (0.09)	0.15 (0.04)	0.24 (0.15)
	HAR	0.33 (0)	0.33 (0)	0.24 (0)	0.23 (0)	0.28 (0.05)

Note: The table presents the mean and standard deviation for the performance of selected models by different methods. For instance, the top left figure of 0.4 shows the average QLIKE for SPY by the MCS over the study period 2014–2020 when estimated realized volatility is compared to the average actual realized volatility over the h following period. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations. The *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels, respectively, for a one-sided paired t-test with the null hypothesis that the GARCH(1,1) has an equal or less average QLIKE than alternative forecasting models across all ETFs.

the panel data of K observations is rolled forward by h days, and new estimates are calculated for all volatility forecasting models. Table 7 presents the QLIKE performance for benchmarks (GARCH, GJR-GARCH, and HAR) and considered MHT buckets.

Table 7 shows that when longer-term forecasts are considered (i.e., $h = 5$ and $h = 22$), QLIKE improves for all benchmarks and buckets consistently over results in Section 4.2.

Also, differences among alternative approaches relative to a GARCH(1,1) are no longer significant for MCS or FDR^{+}_{HAR} (as in Table 4). Finally, for both cases of $h = 5$ and $h = 22$, the FDR^{+}_{GARCH} offers the lowest average QLIKE error compared to its alternatives across four ETFs. The results in this section corroborate the use of FDR^{+} bucket with a relevant benchmark for different assets, especially when longer-term forecasts are of interest.

4.6. Buckets under stress

The previous results reflect a high level of variation in the size of the volatility model buckets. This section investigates whether stress in financial markets can explain this high noise ratio. We use the Office of Financial Research (2021) (OFR) Financial Stress Index (FSI) for volatility to evaluate the variation in our results under alternative market conditions. The FSI volatility index monitors the systemic financial stress for implied and realized volatility in different market categories, including equity and commodity. The FSI index is zero when the financial markets are functioning normally, and high (low)-stress periods are identified with FSI indices above (below) zero, as exhibited in Fig. 2.

In Table 8, we replicate our previous bucket formation practice ($M = 78$) for high and low stress periods. We form buckets of volatility models and tracking the performances of the buckets and the benchmark models (GARCH, GJR-GARCH, and HAR) for both stress profiles.

From the table, we note that all MHT buckets (MCS and all FDR^{+} with different benchmarks) significantly forecast realized volatility better than GARCH(1,1) in periods of high stress in financial markets. Furthermore, the FDR^{+}_{GARCH} and $FDR^{+}_{GJR-GARCH}$ have the lowest QLIKE on average across all assets in periods of high stress. Yet the latter pattern is reversed for periods of low stress and $FDR^{+}_{GJR-GARCH}$ has the lowest average QLIKE error across four assets.

Regarding individual assets, the $FDR^{+}_{GJR-GARCH}$ bucket offers the best performance for equity ETFs (SPY and QQQ) in high and low stress periods. For GLD, $FDR^{+}_{GJR-GARCH}$ (MCS) is the best predictor in periods of high (low) stress. Finally, for USO, the FDR^{+}_{GARCH} and $FDR^{+}_{GJR-GARCH}$ are the best predictors for times of high and low stress, respectively. The latter shows the necessity of considering alternative benchmarks when markets experience periods of high stress.

4.7. Buckets over time

Our findings so far establish significant performance enhancement by using an MHT procedure. Yet there are relatively high variations in the number of discoveries in Table 3 and the performances in Table 4. These variations are potentially attributable to variations in the performance of models across periods of high/low stress (as in Section 4.6) or dynamics over time. Accordingly, in this section, we investigate whether the size of MHT buckets is time-dependent. Table 9 presents the variations in the number of discoveries over the seven (2014 to 2020) subperiods. Table 10 presents the variations in QLIKE for the considered buckets and benchmarks.

From Table 9, we make two observations. First, the bucket size is time dynamic. For instance, the average bucket size for MCS fell from 153 in 2014 to 86 in 2017 and rose again to 185 in 2020. Second and, more importantly, the calendar year for which the largest buckets

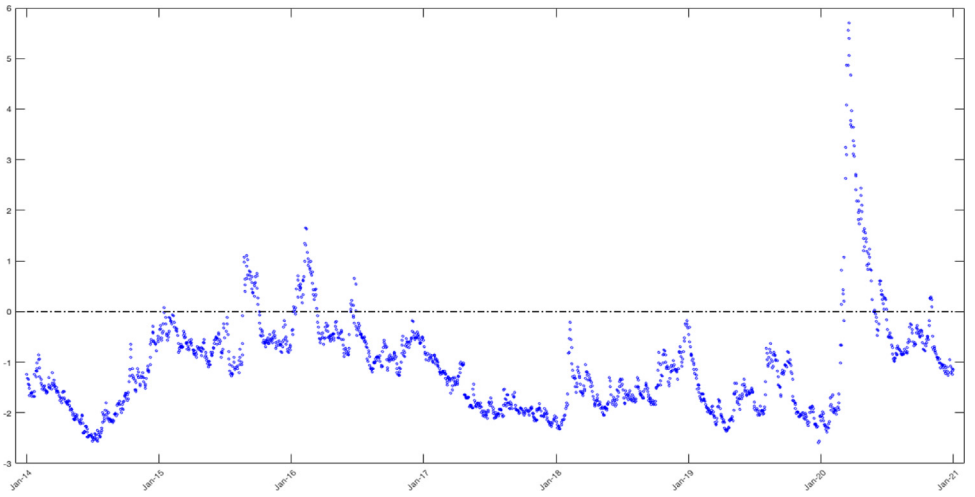

Fig. 2. OFR volatility FSI
Note: The figures present variations in the volatility FSI across our study period (2014–2020). The horizontal dotted line identifies the threshold for defining high and low stress levels.

Table 8
The QLIKE performance under different testing frameworks.

Stress level	Method	SPY	QQQ	GLD	USO	Average
High ($N = 164$)	*MCS*	0.22 (0.11)	0.25 (0.14)	0.2 (0.12)	0.23 (0.12)	0.23 (0.11)
	FDR^{+}_{GARCH}	0.18 (0.06)	0.16 (0.05)	0.11 (0.01)	0.18 (0.08)	0.16*** (0.06)
	GARCH	0.28 (0)	0.27 (0)	0.22 (0)	0.24 (0)	0.25 (0.03)
	$FDR^{+}_{\text{GJR}-\text{GARCH}}$	0.15 (0.04)	0.15 (0.06)	0.1 (0.06)	0.26 (0.19)	0.17** (0.11)
	$GJR - GARCH$	0.28 (0)	0.22 (0)	0.12 (0)	0.27 (0)	0.22 (0.07)
	FDR^{+}_{HAR}	0.19 (0.08)	0.17 (0.04)	0.12 (0.01)	0.24 (0.02)	0.18*** (0.06)
	HAR	0.32 (0)	0.35 (0)	0.28 (0)	0.72 (0)	0.42 (0.18)
Low ($N = 1, 599$)	*MCS*	0.42 (0.25)	0.43 (0.19)	0.26 (0.21)	0.22 (0.12)	0.33** (0.2)
	FDR^{+}_{GARCH}	0.35 (0.14)	0.34 (0.18)	0.29 (0.07)	0.17 (0.05)	0.29*** (0.13)
	GARCH	0.6 (0)	0.55 (0)	0.51 (0)	0.31 (0)	0.49 (0.11)
	$FDR^{+}_{\text{GJR}-\text{GARCH}}$	0.33 (0.14)	0.33 (0.18)	0.29 (0.07)	0.16 (0.05)	0.28*** (0.13)
	$GJR - GARCH$	0.53 (0)	0.49 (0)	0.51 (0)	0.27 (0)	0.45 (0.11)
	FDR^{+}_{HAR}	0.65 (0.18)	0.52 (0.23)	0.29 (0.07)	0.18 (0.06)	0.41 (0.24)
	HAR	0.62 (0)	0.65 (0)	0.49 (0)	0.35 (0)	0.53 (0.12)

Note: The table presents the mean and standard deviation for the performance of selected models by different methods. For instance, the top left figure of 0.21 shows the average QLIKE for SPY by the MCS over the study period 2014–2020. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations. The *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels, respectively, for a one-sided paired t-test with the null hypothesis that the GARCH(1,1) has equal or less average QLIKE than alternative forecasting models across all ETFs.

are identified varies across MHT procedures. For MCS, the largest difference in performance is in 2020, while for FDR^{+} procedures form the largest buckets in 2019. This observation can be linked to the construct of considered MHT procedures. The FDR^{+} requires the practitioner to identify a benchmark while *MCS* does not. A functional benchmark (e.g., GARCH/GJR-GARCH) can filter out time-linked fluctuations in volatility forecasting models. From Fig. 2 and Table 8, we recall 82 out of 154 days when markets underwent stress were in 2020. The results in Table 9 further support our findings in Section 4.6 that in periods of high stress (as in 2020), FDR^{+} approaches outperform *MCS* and can be seen as the more robust choice.

From Table 10, we note that the QLIKE performance varies over time. We also observe that the performance of MHT buckets worsens compared to Table 4. For example, the average annual QLIKE for MCS (FDR^{+}_{GARCH}) is 1.07 (0.37) as compared to 0.37 (0.27) in Table 4. The latter can be explained by the loss of power (i.e. the QLIKE increase) in MHT methods where the number of considered alternatives (m) is larger than the number of observations (N). The loss of power is noticeably greater for MCS, where the QLIKE is double relative to the FDR^{+} procedures. Accordingly, the pattern of the FDR^{+} method as being more accurate relative to the benchmarks continues to hold. To conclude this robustness check, we make the additional observation that when $N < m$ the MCS approach is no

Table 9
Bucket size variations over the study period ($N = 252$).

Method	Asset	2014	2015	2016	2017	2018	2019	2020	Average
MCS	SPY	188.8 (166.27)	138 (146.5)	163 (148.49)	127.2 (123.29)	180.2 (159.7)	137.8 (116.27)	193.6 (165.92)	161.23 (136.53)
	QQQ	185.8 (163.86)	167.2 (150.55)	116 (124.82)	80.8 (124.71)	151.4 (143.12)	133 (131.24)	171 (150.74)	143.6 (133.3)
	GLD	57 (82.53)	108.2 (106.71)	95.6 (111.16)	62.2 (107.21)	112.8 (147.62)	117 (105.07)	181 (156.36)	104.83 (114.84)
	USO	179.4 (161.95)	166.6 (139.3)	178.6 (158.64)	73.2 (104.33)	138.4 (160.1)	177.6 (160.06)	194.6 (174.93)	158.34 (144.04)
	Average	152.75 (147.16)	145 (128.04)	138.3 (130.46)	85.85 (108.77)	145.7 (142.43)	141.35 (121.3)	185.05 (149.22)	142 (133.1)
FDR^{+}_{GARCH}	SPY	65.2 (101.46)	27.2 (51.95)	42.4 (40.49)	62.2 (75.16)	31.6 (63.42)	109.6 (102.93)	9.6 (16.52)	49.69 (71.35)
	QQQ	53.8 (71.08)	23.4 (48.43)	73.6 (66.45)	52.4 (79.4)	33.8 (65.02)	111 (110.19)	8.8 (16.35)	50.97 (71.5)
	GLD	66.6 (51.01)	30.6 (51.54)	84.8 (74.24)	97 (58.06)	74.4 (70.17)	121.2 (136.23)	18.4 (31.74)	70.43 (75.46)
	USO	82.8 (77.54)	56 (84.14)	36.6 (74.06)	73.4 (107.63)	57.8 (83.91)	77.4 (84.21)	10.6 (21.47)	56.37 (76.43)
	Average	67.1 (71.81)	34.3 (57.3)	59.35 (63.45)	71.25 (77.18)	49.4 (67.72)	104.8 (102.33)	11.85 (20.93)	56.86 (73.39)
$FDR^{+}_{\text{GJR-GARCH}}$	SPY	20.2 (25.72)	33.8 (65.01)	83.8 (71.57)	82.4 (77.68)	6 (7.42)	62.8 (67.8)	5.4 (7.23)	42.06 (58.9)
	QQQ	21.8 (43.19)	17 (34.12)	117.2 (80.4)	63.2 (65.99)	12.6 (21.57)	61.6 (75.6)	5.4 (7.67)	42.69 (61.59)
	GLD	94.8 (55.34)	16.6 (20.57)	88.8 (85.33)	60 (40.39)	96.4 (105.24)	132.2 (141.96)	30.2 (60.3)	74.14 (83.81)
	USO	44 (68.22)	93 (85.73)	25.4 (53.45)	49.6 (96.13)	26.4 (49.1)	32 (52.87)	6.6 (12.52)	39.57 (64.24)
	Average	45.2 (55.79)	40.1 (61.67)	78.8 (75.8)	63.8 (68)	35.35 (65.68)	72.15 (91.8)	11.9 (30.65)	49.61 (68.61)
FDR^{+}_{HAR}	SPY	13 (10.46)	87 (131.73)	23.2 (14.65)	113.6 (101.7)	7.6 (10.99)	77.8 (105.38)	39.8 (83.42)	51.71 (83.24)
	QQQ	13.2 (15.11)	66 (139.23)	69.6 (89.23)	44.6 (87.1)	2.4 (1.52)	101.8 (110.73)	4 (5.2)	43.09 (82.89)
	GLD	72.6 (38.39)	42 (88.35)	14.8 (27.54)	66.8 (71.53)	51.6 (86.06)	75.2 (83.54)	40.2 (80.97)	51.89 (68.26)
	USO	126 (104.79)	43.4 (33.85)	7.4 (12.7)	40.6 (52.41)	17.4 (22.24)	63.8 (66.03)	32.2 (69.77)	47.26 (65.36)
	Average	56.2 (70.88)	59.6 (99.89)	28.75 (50.33)	66.4 (79.48)	19.75 (45.56)	79.65 (86.65)	29.05 (64.08)	48.49 (74.65)

Note: The table presents the mean and standard deviation for the size of the MHT buckets for each calendar year. For instance, the top left figure of 188.8 shows the average size of the MCS bucket for SPY over the calendar year 2014. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations.

longer able to perform better (have a lower forecast error) than the GARCH(1,1) model.

5. Conclusion

Financial market volatility underlies the study and practice of asset and risk management (Andersen et al., 2001). Correspondingly, a large and growing literature on volatility forecasting attempts to either present new volatility models or examine the predictive ability of traditional models. However, optimal volatility forecasting depends on the studied asset and respective market conditions. Therefore, this paper proposes an alternative solution: a novel empirical methodology that enables a researcher or practitioner to evaluate a large pool of potential models and select the optimal (most accurate) scenario-specific volatility forecasting model. The evidence shows that our methodology identifies significant models that outperform the previous literature's suggested benchmarks of GARCH(1,1), GJR-GARCH(1,1), and HAR when applying our methodology to a large pool of 325 candidate models for a dataset consisting of four indices (two stocks and two commodities).

Our proposed methodology consists of an MHT set-up with an $\text{FDR}^{+/-}$ treated for a discrete test statistic space. The MHT set-up enables the researcher to consider a large pool of candidate models to identify buckets of superior-performing volatility forecasting models. However, using

Table 10
Bucket QLIKE variations over the study period ($N = 252$).

Asset	Year	MCS	FDR^{+}_{GARCH}	*GARCH*	$FDR^{+}_{GJR-GARCH}$	GJR-GARCH	FDR_{HAR}	*HAR*
SPY	2014	1.72 (2.73)	0.45 (0.22)	0.64 (0)	0.44 (0.26)	0.5 (0)	0.44 (0.27)	0.38 (0)
	2015	1.05 (1.58)	0.37 (0.17)	0.49 (0)	0.39 (0.19)	0.45 (0)	0.39 (0.21)	0.43 (0)
	2016	1.76 (2.54)	0.33 (0.26)	0.62 (0)	0.35 (0.27)	0.51 (0)	0.41 (0.38)	0.73 (0)
	2017	2.25 (4)	0.66 (0.69)	0.57 (0)	0.64 (0.5)	0.58 (0)	0.63 (0.46)	0.71 (0)
	2018	1.08 (1.61)	0.28 (0.21)	0.42 (0)	0.25 (0.24)	0.35 (0)	0.34 (0.24)	0.33 (0)
	2019	1.97 (3.04)	0.54 (0.35)	0.67 (0)	0.62 (0.56)	0.61 (0)	0.56 (0.26)	0.81 (0)
	2020	1.28 (1.53)	0.4 (0.21)	0.56 (0)	0.46 (0.2)	0.58 (0)	0.46 (0.21)	0.74 (0)
QQQ	2014	1.03 (1.47)	0.31 (0.17)	0.46 (0)	0.31 (0.18)	0.4 (0)	0.32 (0.19)	0.34 (0)
	2015	0.83 (1.14)	0.3 (0.13)	0.35 (0)	0.29 (0.12)	0.33 (0)	0.38 (0.03)	0.57 (0)
	2016	0.98 (1.27)	0.33 (0.24)	0.61 (0)	0.32 (0.25)	0.49 (0)	0.41 (0.34)	0.79 (0)
	2017	1.19 (2.02)	0.41 (0.25)	0.49 (0)	0.41 (0.25)	0.47 (0)	0.37 (0.22)	0.55 (0)
	2018	0.77 (0.88)	0.29 (0.18)	0.43 (0)	0.29 (0.17)	0.35 (0)	0.43 (0.19)	0.43 (0)
	2019	1.27 (0.99)	0.66 (0.26)	0.79 (0)	0.65 (0.28)	0.7 (0)	0.76 (0.16)	1.12 (0)
	2020	0.85 (0.91)	0.32 (0.1)	0.54 (0)	0.44 (0.16)	0.52 (0)	0.35 (0.12)	0.58 (0)
GLD	2014	1.14 (1.71)	0.56 (0.23)	0.83 (0)	0.58 (0.22)	0.85 (0)	0.54 (0.24)	0.74 (0)
	2015	0.94 (1.5)	0.34 (0.18)	0.31 (0)	0.37 (0.21)	0.31 (0)	0.36 (0.2)	0.4 (0)
	2016	1.08 (1.56)	0.39 (0.3)	0.59 (0)	0.32 (0.23)	0.53 (0)	0.29 (0.23)	0.41 (0)
	2017	1.67 (2.92)	0.51 (0.27)	0.74 (0)	0.51 (0.28)	0.72 (0)	0.53 (0.32)	0.56 (0)
	2018	1.47 (2.87)	0.46 (0.43)	0.23 (0)	0.57 (0.55)	0.19 (0)	0.32 (0.18)	0.35 (0)
	2019	1.5 (2.7)	0.4 (0.35)	0.35 (0)	0.37 (0.27)	0.37 (0)	0.39 (0.29)	0.42 (0)
	2020	0.82 (1.11)	0.29 (0.13)	0.34 (0)	0.3 (0.17)	0.35 (0)	0.33 (0.13)	0.43 (0)
USO	2014	0.71 (0.95)	0.44 (0.69)	0.08 (0)	0.44 (0.68)	0.1 (0)	0.16 (0.11)	0.17 (0)
	2015	0.32 (0.25)	0.19 (0.08)	0.29 (0)	0.19 (0.09)	0.23 (0)	0.17 (0.12)	0.24 (0)
	2016	0.44 (0.38)	0.2 (0.12)	0.37 (0)	0.2 (0.11)	0.32 (0)	0.15 (0.09)	0.24 (0)
	2017	0.45 (0.44)	0.33 (0.14)	0.54 (0)	0.33 (0.14)	0.53 (0)	0.29 (0.14)	0.36 (0)
	2018	0.36 (0.45)	0.14 (0.06)	0.19 (0)	0.14 (0.06)	0.17 (0)	0.14 (0.06)	0.16 (0)
	2019	0.48 (0.37)	0.33 (0.11)	0.37 (0)	0.3 (0.11)	0.3 (0)	0.31 (0.11)	0.32 (0)
	2020	0.59 (0.37)	0.23 (0.19)	0.26 (0)	0.23 (0.19)	0.26 (0)	0.38 (0.22)	1.18 (0)
Average		1.07 (1.71)	0.37*** (0.28)	0.47 (0.18)	0.38** (0.29)	0.43* (0.17)	0.38*** (0.25)	0.52 (0.25)

Note: The table presents the mean and standard deviation of QLIKE for the MHT buckets and benchmarks for each calendar year. For instance, the top left figure of 1.72 shows the average QLIKE of the MCS bucket for SPY over the calendar year 2014. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations. The *, **, and *** indicate rejection at the 5%, 1%, and 0.1% significance levels, respectively, for a one-sided paired t-test with the null hypothesis that the GARCH(1,1) has the equal or less average QLIKE as alternative forecasting models across all ETFs.

the MHT set-up increases the challenge of false discoveries (Type I errors). Two solutions to this issue have been discussed in the MHT literature: the FDR and the FWER approaches. Our proposed $FDR^{+/-}$ selection framework – refined by the DRB – performs better than the literature's reliance on FWER selection frameworks by comparing its performance against the MCS and *k*StepM.

To establish the value of an MHT selection framework in identifying optimal volatility forecasting models, we first test three methods – the MCS, our proposed FDR^{+}, and the *k*StepM - against the literature benchmark models of GARCH, GJR-GARCH and HAR. We do this across four U.S. financial indices. In doing so, we find that both the MCS and the FDR^{+} methods identify buckets of superior-performing models. We then evaluate the forecast error performance of the MCS and FDR^{+} methods using the robust loss function, QLIKE. The evidence shows that our proposed FDR^{+} method consistently outperforms both the MCS method and literature benchmark models. To establish the robustness of the FDR^{+} we replicate the QLIKE performance tests using variation in the target realized volatility by reducing the number of intraday returns to $M = 26$ and $M = 13$ (corresponding to 15-minute and 30-minute returns, respectively). In both cases, the FDR^{+} continues to outperform the MCS and benchmark models. We further investigate the robustness of our proposed methodology to the number of observations K, the number of steps a forecast is made for h, and variation in financial market stress and time periods. The evidence again supports the FDR^{+} model selection method.

Our model selection strategy has not previously been used in the volatility forecasting model literature. This paper presents the methodology not as a means to identifying a unique 'best' model. Instead, we propose the methodology as a tool for the researcher and practitioner to use to identify buckets of 'best' models for the given set of market characteristics toward optimal volatility forecasting and risk estimation. The evidence supports our proposed FDR^{+} with DRB model selection strategy for optimal volatility forecasting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the associate editor, two anonymous reviewers, Michael Adams, David Newton, Georgios Sermpinis, Charalampos Stasinakis and Ania Zalewska, and the participants of the Global Finance Conference 2021 for their valuable comments.

Appendix A. Bucket formation algorithm for FDR^+

The following procedure is followed to find the superior models compared to the benchmark. The steps are the same for all five loss functions.

1. Specify $m = 325$ volatility forecasting models based on M intraday returns and K observations to predict one-step-ahead realized volatility.
2. Calculate the centred test statistics given the loss functions in Eq. (41) and the benchmarks (GARCH, GJR-GARCH, and HAR) based on N out-of-sample observations.
3. Generate $B = 1,000$ stationary bootstrap to generate the set of centred null test statistics $\varphi'_{b,i}$, for $b = 1, \ldots, B$.
4. Calculate the p-values $(\hat{p}_i)$ for all m models.
5. Estimate the optimal tuning parameter λ^* and estimated null proportion $\hat{\pi}_0(\lambda^*)$ as:
   a. Select the number of λ support points ($n = 20$).
   b. Define the $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$
   c. Compute the estimated null proportion as $\hat{\pi}_0(\lambda_l) = \frac{\#\hat{p}_i > \lambda_l}{(1-\lambda_l)m}$ for $i = 1, \ldots, m$ and $l = 1, \ldots, n$
   d. Find the first l where $\hat{\pi}_0(\lambda_l) \geq \hat{\pi}_0(\lambda_{l-1})$
6. Choose a target false discovery controlling level (α), e.g. 10%
7. Sort the p-values in ascending order where $\hat{p}_{R_1} \leq \cdots \leq \hat{p}_{R_m}$
8. Define step by $j = 1$ & corresponding significance region with $\gamma'_j = \hat{p}_{R_j}$
9. Compute the $\widehat{FDR}^+_{\gamma'_j}(\hat{\pi}_0)$ based on Eqs. (7) and (8).
   a. If the $\widehat{FDR}^+_{\gamma'_j} < \alpha$
      i. Reject all $P_i \leq P_{R_j}$
      ii. Set $j = j + 1$, $\gamma'_j = \hat{p}_{R_j}$ and go back to 8.
   b. Otherwise if the $\widehat{FDR}^+_{\gamma'_j} \geq \alpha$
      i. Reject all $\hat{p}_i \leq \hat{p}_{R_j}$
      ii. Terminate the process

Appendix B. Specification of the pool

Table B.1 below provides the characteristics of the models under study. We study twenty classes of volatility models from eight families (GARCH, SV, EWMA, HAR, SVR, MLP, SMA, and FC), and four distributions for innovation distribution fitted for the return process.

Appendix C. Alternative settings for *k*Step

Table C.1 explores using three pre-selected levels of $k = 1, 5, 10$ in kStepM procedure compared to the FDR-based approach as in Section 4. The table also presents results from using the SPA test. The figures show the average size for buckets of selected models by kStepM and SPA

Table B.1
Specification of the volatility forecasting pool.

Class	Family	Error Distribution	Parameters	Count
1. ARCH	GARCH	Gaussian, t, skewed t & GED	$q = 1, 2$	8
2. GARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
3. IGARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
4. Taylor/ Schwert	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
5. A-GARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
6. NA-GARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
7. TGARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
8. GJR-GARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
9. log-GARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
10. EGARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
11. NGARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
12. APARCH	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
13. FI-GARCH	GARCH	Gaussian, t, skewed t & GED	$p = 0, 1; q = 0, 1$	16
14. GARCH-MA	GARCH	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
15. SV	SV	Gaussian, t, skewed t & GED	$q = 1, 2$	8
16. SV-MA	SV	Gaussian, t, skewed t & GED	$p = 1, 2; q = 1, 2$	16
17. SV-L	SV	Gaussian, t, skewed t & GED	$q = 1, 2$	8
18. RM	EWMA	–	$\upsilon \in \{0.8, 0.82, \ldots, 0.98\}$	10
19. HAR	HAR	–	RV and log-RV	2
20. HARQ	HAR	–	RV and log-RV	2
21. TV-HAR	HAR	–	RV and log-RV	2
22. HAR-L	HAR	–	log-RV	1
23. HAR-J	HAR	–	RV, $RV^{0.5}$, and log-RV	3
24. SVR-HAR	SVR	–	$\varepsilon_{SVR} \in \{10^{-6}, \ldots, 10^{-2}\}$, $\phi_{SVR} \in \{linear, RBF\}$, RV and log-RV	20
25. ANN-HAR	MLP	–	$\Theta_{ANN} \in \{none, Tanh, RBF\}$, $N_{ANN} \in \{1, 5, 10, 50, 100\}$ RV and log-RV	30
26. SMA	SMA	–	$\mathfrak{p} \in \{5, 10, 22, 63, 126\}$	5

(continued on next page)

Table B.1 (*continued*).

Class	Family	Error Distribution	Parameters	Count
27. FC	FC	–	$\theta_C \in \{0.9, 1\}$	2
Total				**325**

Note: The table shows the details of individual models in the volatility forecasting pool. The formulas for each class of models are presented in Table 1.

Table C.1
The bucket size under different testing frameworks ($N = 1,763$).

Benchmark	Index	$kStepM_{\gamma=0.1}$	$kStepM(1)$	$kStepM(5)$	$kStepM(10)$	*SPA*
GARCH	SPY	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	QQQ	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	GLD	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	USO	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Average	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
GJR-GARCH	SPY	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	QQQ	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	GLD	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	USO	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Average	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
HAR	SPY	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	QQQ	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	GLD	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	USO	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Average	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Note: The table presents the mean and standard deviation for the size of buckets formed of true discoveries using alternative settings of *k*StepM. For instance, the top left figure of 0 shows the average number of discoveries for SPY by the *k*StepM − FDP procedure with $\gamma = 0.1$ over the study period 2014–2020. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations.

Table D.1
Number of discoveries for different levels of β ($N = 1,763$).

Method	*MCS*	FDR^+			*kStepM*		
Benchmark	–	GARCH	GJR-GARCH	HAR	GARCH	GJR-GARCH	HAR
$\beta = -5$	83.5 (118.62)	17.25 (18.21)	17 (17.94)	230.25 (130.33)	0 (0)	0 (0)	0 (0)
$\beta = -2$	4.5 (4.12)	63.75 (6.55)	44 (10.3)	155.75 (77.99)	0 (0)	0 (0)	0 (0)
$\beta = 0$	262.75 (51.41)	1 (0)	1 (0)	1 (0)	0 (0)	0 (0)	0 (0)
$\beta = +1$	287.5 (15.97)	1 (0)	0.75 (0.5)	1 (0)	0 (0)	0 (0)	0 (0)
Average	159.56 (136.22)	20.75 (27.92)	15.69 (20.42)	97 (123.22)	0 (0)	0 (0)	0 (0)

Note: The table presents the mean and standard deviation of true discoveries for FDR^+, MCS, and *k*StepM for different levels of robust β. For instance, the top left figure of 83.5 shows the average number of discoveries for $\beta = -5$ by the MCS procedure over the study period 2014–2020. The standard deviation in parentheses presents variations across all assets (SPY, QQQ, GLD, and USO). The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations.

procedures with GARCH, GJR-GARCH, and HAR benchmarks. The details of the respective MHT frameworks are discussed in Section 4.

The results in this appendix show that neither *k*StepM nor SPA can form buckets of significantly more accurate volatility forecasting models.

Appendix D. Testing framework performance for other loss functions

Table D.1 presents the average performance of selected models by the MCS procedure compared to the FDR^+ with the benchmarks of GARCH, GJR-GARCH, and HAR. The details of the respective MHT frameworks are discussed in Section 4.2.

Appendix E. Buckets specification

In Section 4.2, we present the main results of our analysis. In this appendix, we identify the volatility forecasting models selected models by each MHT procedure.

Table E.1
Number of models by type selected into MHT buckets ($N = 1,763$).

Asset	Bucket	GARCH	SV	EWMA	HAR	SVR	ANN	SMA	FC
SPY	*MCS*	98.25 (113.49)	16 (18.48)	5 (5.77)	6 (4.62)	11.25 (7.23)	9.75 (11.32)	2.5 (2.89)	0 (0)
	FDR^{+}_{GARCH}	8.5 (17)	2 (4)	3 (3.46)	3.5 (2.89)	5 (5.77)	6.25 (7.5)	0.5 (1)	0 (0)
	$FDR^{+}_{GJR-GARCH}$	1 (2)	0 (0)	2.25 (2.63)	3.5 (2.89)	5 (5.77)	5.75 (6.75)	0.25 (0.5)	0 (0)
	FDR^{+}_{HAR}	72 (99.08)	11 (15.1)	4 (4.9)	3.75 (3.2)	7.75 (9.67)	7.5 (8.66)	1.75 (2.36)	0.5 (1)
QQQ	*MCS*	105.25 (108.66)	17 (17.4)	7 (4.76)	6.75 (3.95)	13 (5.35)	10.75 (8.3)	3 (2.45)	0 (0)
	FDR^{+}_{GARCH}	8.25 (16.5)	2 (4)	1.25 (2.5)	2.5 (2.38)	4.25 (5.68)	3.5 (7)	0.25 (0.5)	0 (0)
	$FDR^{+}_{GJR-GARCH}$	1.25 (2.5)	1 (2)	0.5 (1)	2.5 (2.38)	2.75 (5.5)	3.25 (6.5)	0.25 (0.5)	0 (0)
	FDR^{+}_{HAR}	86.75 (104.51)	12 (15.32)	4.5 (5.26)	4 (3.56)	8 (9.8)	7.25 (8.38)	2 (2.45)	0.5 (1)
GLD	*MCS*	78.25 (91.61)	12 (15.32)	5.25 (5.5)	4.5 (4.8)	9.5 (7.37)	5.5 (8.02)	1.25 (2.5)	0 (0)
	FDR^{+}_{GARCH}	0.75 (0.5)	4 (8)	4.25 (4.92)	2.75 (3.2)	5.25 (6.08)	5.75 (6.75)	1 (1.41)	0 (0)
	$FDR^{+}_{GJR-GARCH}$	0.25 (0.5)	4 (8)	4 (4.69)	3 (2.94)	5.25 (6.08)	5.75 (6.75)	1 (1.41)	0 (0)
	FDR^{+}_{HAR}	0.75 (0.5)	4 (8)	4.25 (4.92)	2.75 (3.2)	5.25 (6.08)	5.75 (6.75)	1 (1.41)	0 (0)
USO	*MCS*	143.25 (95.59)	24 (16)	7.5 (5)	6.5 (4.36)	11.5 (7.94)	17.75 (13.72)	3.5 (2.38)	0 (0)
	FDR^{+}_{GARCH}	3.25 (6.5)	2 (4)	1.75 (3.5)	1.5 (3)	3 (6)	3.75 (6.18)	1 (1.41)	0 (0)
	$FDR^{+}_{GJR-GARCH}$	0 (0)	0 (0)	1.5 (3)	1.75 (2.87)	3 (6)	3.25 (5.19)	0.5 (0.58)	0 (0)
	FDR^{+}_{HAR}	92 (106.68)	12 (15.32)	4.75 (5.5)	3.5 (4.12)	8.25 (9.95)	8 (8.08)	2 (2.45)	0.5 (1)

Note: The table presents the mean and standard deviation for the number of selected models by different buckets. For instance, the top left figure of 98.25 shows the average number of GARCH models selected by MCS for SPY over the study period 2014–2020. The standard deviation in parentheses presents variations across different levels of $\beta \in \{-5, -2, 0, 1\}$. The target realized volatility is formed by using $M = 78$ intraday returns. Each one-step-ahead volatility forecast is made using $K = 252$ previous rolling observations.

Table E.1 specifies the eight families (GARCH, SV, EWMA, HAR, SVR, ANN, SMA, or FC) of the selected volatility forecasting models. We consider 216 GARCH, 32 SV, 10 EWMA, 10 HAR, 20 SVR, 30 ANN, 5 SMA, and 2 FC models. The table shows that the number of individual volatility forecasting models from each family varies depending on the considered combination of the underlying asset and the MHT framework. Hence, rather than looking for an individual model/class of models, our results suggest focusing on buckets over single benchmarks as the preferred approach.

References

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *88*, 5–905.

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*(4), 701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, *61*(1), 43–76.

Andersen, T. G., Thyrsgaard, M., & Todorov, V. (2019). Time-varying periodicity in intraday volatility. *Journal of the American Statistical Association*.

Anderson, M. L. (2017). The benefits of college athletic success: An application of the propensity score design. *The Review of Economics and Statistics*, *99*(1), 119–134.

Andrikogiannopoulou, A., & Papakonstantinou, F. (2019). Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *The Journal of Finance*, *74*(5), 2667–2688.

Ardia, D., & Hoogerheide, L. F. (2014). GARCH models for daily stock returns: Impact of estimation frequency on value-at-risk and expected shortfall forecasts. *Economics Letters*, *123*(2), 187–190.

Asai, M., & McAleer, M. (2011). Alternative asymmetric stochastic volatility models. *Econometric Reviews*, *30*(5), 548–564.

Asai, M., McAleer, M., & Yu, J. (2006). Multivariate stochastic volatility: Review. *Econometric Reviews*, *25*(2–3), 145–175.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*(1), 3–30.
Bajgrowicz, P., & Scaillet, O. (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, *106*(3), 473–491.
Bancroft, T., Du, C., & Nettleton, D. (2013). Estimation of false discovery rate using sequential permutation p-values. *Biometrics*, *69*(1), 1–7.
Bao, Y., Lee, T. H., & Saltoglu, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: A reality check. *Journal of Forecasting*, *25*(2), 101–128.
Barras, L., Scaillet, O., & Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, *65*(1), 179–216.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *28*, 9–300.
Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.
Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, *54*, 2–547.
Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, *192*(1), 1–18.
Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3–62.
Brooks, C., & Persand, G. (2003). Volatility forecasting for risk management. *Journal of Forecasting*, *22*(1), 1–22.
Broto, C., & Ruiz, E. (2004). Estimation methods for stochastic volatility models: A survey. *Journal of Economic Surveys*, *18*(5), 613–649.
Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, *18*(3), 502–531.
Campbell, J. Y., & Hentschel, L. (1992). No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics*, *31*(3), 281–318.
Chan, J. C., & Grant, A. L. (2016). Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Economics*, *54*, 182–189.
Christoffersen, P. F., & Diebold, F. X. (2000). How relevant is volatility forecasting for financial risk management? *The Review of Economics and Statistics*, *82*(1), 12–22.
Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196.
Corsi, F., & Renò, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics*, *30*(3), 368–380.
Davis, J. M., & Heller, S. B. (2020). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *The Review of Economics and Statistics*, *102*(4), 664–677.
Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431.
Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.
Dimson, E., & Marsh, P. (1990). Volatility forecasting without data-snooping. *Journal of Banking & Finance*, *14*(2–3), 399–421.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *98*, 7–1007.
Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, *5*(1), 1–50.
Engle, R. F., & Figlewski, S. (2015). Modeling the dynamics of correlations among implied volatilities. *Review of Finance*, *19*(3), 991–1018.
Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance*, *48*(5), 1749–1778.
Esposito, F. P., & Cummins, M. (2016). Multiple hypothesis testing of market risk forecasting models. *Journal of Forecasting*, *35*(5), 381–399.
Figlewski, S. (1997). Forecasting volatility. *Financial markets, institutions & instruments*, *6*(1), 1–88.
Ghysels, E., Harvey, A. C., & Renault, E. (1996). 5 stochastic volatility. *Handbook of Statistics*, *14*, 119–191.
Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256). JMLR Workshop and Conference Proceedings.
Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801.
Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, *70*, 5–730.
Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365–380.
Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH (1, 1)? *Journal of Applied Econometrics*, *20*(7), 873–889.
Hansen, P., & Lunde, A. (2011). Forecasting volatility using high frequency data. In *The oxford handbook of economic forecasting* (pp. 525–556). Blackwell: Oxford.
Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.
Harvey, A., & Sucarrat, G. (2014). EGARCH models with fat tails, skewness and leverage. *Computational Statistics & Data Analysis*, *76*, 320–338.
Harvey, C. R., & Whaley, R. E. (1992). Market volatility prediction and the efficiency of the S & P 100 index option market. *Journal of Financial Economics*, *31*(1), 43–73.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 5–70.
Huang, C., Gong, X., Chen, X., & Wen, F. (2013). Measuring and forecasting volatility in Chinese stock market using HAR-CJ-M model. In *Abstract and applied analysis, Vol. 2013*. Hindawi.
Hurlin, C., Laurent, S., Quaedvlieg, R., & Smeekes, S. (2017). Risk measure inference. *Journal of Business & Economic Statistics*, *35*(4), 499–512.
Jarque, C., & Bera, A. (1980). Efficient tests for normality homoscedasticity and serial independence of regression residuals. *Econometric Letters*, *6*, 255–259.
Kang, S. H., Kang, S. M., & Yoon, S. M. (2009). Forecasting volatility of crude oil markets. *Energy Economics*, *31*(1), 119–125.
Kristjanpoller, W., Fadic, A., & Minutolo, M. C. (2014). Volatility forecast using hybrid neural network models. *Expert Systems with Applications*, *41*(5), 2437–2442.
Leippold, M., & Rueegg, R. (2020). How rational and competitive is the market for mutual funds? *Review of Finance*, *24*(3), 579–613.
Li, J., & Xiu, D. (2016). Generalized method of integrated moments for high-frequency data. *Econometrica*, *84*(4), 1613–1633.
Liang, K. (2016). False discovery rate estimation for large-scale homogeneous discrete p-values. *Biometrics*, *72*(2), 639–648.
Liang, K., & Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *74*(1), 163–182.
Liu, Y. (2019). Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. *Expert Systems with Applications*, *132*, 99–109.
Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.
Lopez, J. A. (2001). Evaluating the predictive accuracy of volatility models. *Journal of Forecasting*, *20*(2), 87–109.
Lyócsa, Š., Molnár, P., & Výrost, T. (2021). Stock market volatility forecasting: Do we need high-frequency data? *International Journal of Forecasting*, *37*(3), 1092–1110.
Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *34*, 7–370.
Office of Financial Research (2021). *OFR financial stress index*. OFR, updated daily. https://www.financialresearch.gov/financial-stress-index/ (Accessed 07 Jul 2021).
Özbekler, A. G., Kontonikas, A., & Triantafyllou, A. (2021). Volatility forecasting in European government bond markets. *International Journal of Forecasting*, *37*(4), 1691–1709.
Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256.
Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313.
Poon, S. H., & Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, *41*(2), 478–539.
Rapach, D. E., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, *23*(2), 821–862.
Romano, J. P., Shaikh, A. M., & Wolf, M. (2008a). Formalized data snooping based on generalized error rates. *Economic Theory*, *24*(2), 404–447.
Romano, J. P., Shaikh, A. M., & Wolf, M. (2008b). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, *17*(3), 417–442.
Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.
Romano, J. P., & Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, *113*, 38–40.
Sadorsky, P. (2005). Stochastic volatility forecasting and risk management. *Applied Financial Economics*, *15*(2), 121–135.
Schwert, G. W. (1989). Why does stock market volatility change over time? *The Journal of Finance*, *44*(5), 1115–1153.
Sermpinis, G., Hassanniakalager, A., Stasinakis, C., & Psaradellis, I. (2021). Technical analysis profitability and persistence: A discrete false discovery approach on MSCI indices. *Journal of International Financial Markets, Institutions and Money*, *73*, Article 101353.
Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*(3), 479–498.
Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *66*(1), 187–205.
Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445.
Sun, W., Reich, B. J., Cai, T. T., Guindani, M., & Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *77*(1), 59.
Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: theory and practice*, *1*, 203–226.
Taylor, S. J. (1986). *Modelling financial time series*. Wiley.
Wang, J., Athanasopoulos, G., Hyndman, R. J., & Wang, S. (2018). Crude oil price forecasting based on internet concern using an extreme learning machine. *International Journal of Forecasting*, *34*(4), 665–677.
Wang, Y., Pan, Z., & Wu, C. (2018). Volatility spillover from the US to international stock markets: A heterogeneous volatility spillover GARCH model. *Journal of Forecasting*, *37*(3), 385–400.
Wei, Y., Wang, Y., & Huang, D. (2010). Forecasting crude oil market volatility: Further evidence using GARCH-class models. *Energy Economics*, *32*(6), 1477–1484.
White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.
Wu, W. B. (2008). On false discovery control under dependence. *The Annals of Statistics*, *36*(1), 364–380.
Zakoian, J. M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics & Control*, *18*(5), 931–955.
Zhang, Y., Ma, F., & Liao, Y. (2020). Forecasting global equity market volatilities. *International Journal of Forecasting*, *36*(4), 1454–1475.
Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, *100*(472), 1394–1411.