Combination volatility forecasts of duration-dependent Markov-switching models

Douglas Eduardo Turatti[a*], **Fernando Henrique de Paula e Silva Mendes**[b] **and João H. G. Mazzeu**[c]

Abstract

Duration-dependent Markov switching models (DDMS) require a user-specified threshold hyperparameter, for which there is currently no established procedure for estimation or testing. As a result, an ad-hoc duration choice must be heuristically justified. This paper proposes a methodology for handling duration selection in DDMS models, with a focus on volatility forecasting. The main novelty lies in generating forecasts through model combination techniques. The idea is that the combined forecasts will be more robust to misspecification in selecting the duration structure, thus yielding more accurate forecasts. Additionally, the paper contributes to the literature by evaluating the out-of-sample volatility forecasting performance of DDMS models compared to benchmark conditional volatility models. Empirical analysis involves returns from three distinct asset classes: a cryptocurrency, a stock market index, and a foreign currency exchange rate. Various volatility proxies and robust loss functions are incorporated into our analysis. The results indicate that combined forecasts outperform individual models and, in some cases, are more accurate than GARCH and MS-GARCH models. On the other hand, models with fixed duration typically underperform the simple GARCH model, often resulting in test rejections.

Keywords: Duration-dependent Markov-switching model; conditional volatility models; forecasting combination; forecasting volatility.

JEL: C22, C53, G10.

1 Introduction

Many authors have proposed various Markov-switching model variants since Hamilton (1989)'s paper. Durland and McCurdy (1994) introduced the duration-dependent Markov-switching (DDMS) model, using a higher-order Markov chain with duration-dependent state transition probabilities. Initially developed to investigate the business cycle's dependence on the duration of expansions or recessions, the DDMS model has found applications in identifying bull and bear markets in stock analysis (Maheu and McCurdy, 2000a) and estimating volatility in foreign exchange markets (Maheu and McCurdy, 2000b). The empirical literature has also explored generalizations of the DDMS model, expanding the duration dependence structure (see, for example, Lam, 2004; Pelagatti, 2007; Bejaoui and Karaa, 2016, among others).

A crucial aspect of DDMS models is the user-specified duration parameter, as it cannot be estimated. Current literature usually employs arbitrary duration values, assuming that the duration dependence diminishes with a "reasonably" large value. The challenge of selecting and justifying the duration remains the main criticism of DDMS models, often hindering their practical application. This is particularly relevant for conditional volatility models, where it is difficult to set the duration on theoretical economic grounds.

[*]Corresponding author.
[a]Aalborg University, Denmark
[b]Federal University of Rio Grande do Sul, Brazil
[c]Federal University of Juiz de Fora, Brazil
This version: March 2024.

In this context, the main contribution of the present paper is to develop and examine a strategy to forecast volatility using DDMS models that attenuates the problem of duration selection. Our approach is based on a forecasting combination of several DDMS models with different duration parameters instead of relying on a single fixed value. Our rationale is that the combined forecast will be more robust to misspecification in selecting the duration parameter, thus yielding more accurate forecasts. This robustness is achieved by assigning greater weights to more accurate models and lower weights to those with less accuracy, based on their past performance.

Given the potential misspecification or time instability of such parameter, this forecating scheme may lead to lower forecasting errors. It is unlikely that a single DDMS model would consistently outperform others across all time points. Combining forecasts from several DDMS models can thus be seen as a method to enhance the robustness of forecasts against potential biases resulting from such unknown misspecification issues. This becomes particularly relevant when considering the absence of a current procedure for estimating or testing whether a specific duration is suitable for a dataset. In the challenging context of selecting a correctly specified DDMS model, combining forecasts arises as a means to improve predictions in the face of specification issues.

A second contribution of this paper is to assess the out-of-sample volatility forecasting performance of DDMS models. Previous research has primarily considered the model within an in-sample framework (see, for example, Maheu and McCurdy, 2000b). Consequently, the predictive power of the model in an out-of-sample context, utilizing volatility proxies and robust loss functions, remains unknown. To address this gap, we compare the predictive power of the model with that of the standard GARCH and MS-GARCH models. This comparison is important, as the DDMS model possesses potentially advantageous modeling features in relation to standard volatility models (see Maheu and McCurdy, 2000b). Specifically, the conditional volatility reaction to previous shocks is a non-linear function in DDMS, whereas it is linear in GARCH models.

The study employs a model confidence set and pairwise Diebold-Mariano tests to evaluate individual and combined DDMS models, comparing them with GARCH-type models. For robustness, we consider three asset classes, different volatility proxies and loss functions commonly found in the literature. Results indicate the superiority of combining DDMS models, with the choice of duration parameter significantly impacting forecasting performance.

Subsample analysis reveals that the best DDMS model varies across forecasting samples, while the combination approach consistently outperforms individual models. These findings underscore the complexity of duration selection, as no individual DDMS model appears superior across all proxies and loss functions. Although combination methods may not outperform GARCH and MS-GARCH models in all cases, they frequently yield lower losses and achieve statistical significance in specific instances, showcasing their effectiveness in addressing DDMS specification challenges and enhancing forecasting performances. In contrast, models with fixed durations typically underperform the simple GARCH model, often leading to rejection in tests.

The remainder of this paper is structured as follows: Section 2 introduces the proposed model, estimation methods, and forecasting combination techniques. In Section 3, we conduct an empirical analysis of the proposed method using daily returns of Bitcoin, the S&P 500 index, and foreign exchange rates. Finally, Section 4 concludes the paper.

2 Methodology

2.1 Model

Based on Maheu and McCurdy (2000a,b), we start by considering a conditional volatility model given by

$$Y_t = \mu\big(S_t, D(S_t)\big) + \sigma\big(S_t, D(S_t)\big) Z_t, \tag{1}$$

where S_t denotes the state-mixing variable, $D(S_t)$ is the duration of the state S_t at time t, and $Z_t \sim N(0,1)$ is an i.i.d. error term. The duration $D(S_t)$ depicts the length of a run of realizations of a particular state and could, in principle, become very large, leading to estimation problems and numerical instability. To avoid such issues, we define

$$D(S_t) := \min\big\{D(S_{t-1}) I(S_t = S_{t-1}) + 1, \tau\big\}, \tag{2}$$

where I is the indicator function, and $\tau \in \mathbb{N}$ is a user-chosen threshold. We refer to τ as duration parameter. This choice ensures that the duration is accounted for up to time τ, preventing numerical instability. It is crucial to note that, while this hyper-parameter cannot be estimated, the user must still fix its value to prevent $D(S_t)$ from becoming excessively large. In the existing literature, the selection of τ is typically arbitrary, with the argument that the impact of duration dependence diminishes after choosing a "reasonably" large value.

The transition probabilities associated with the latent states S_t are parameterized using an approach similar to that in generalized linear models through a link function. The most commonly applied link is the logit, leading to

$$P\big(S_t = i | S_{t-1} = i, D(S_{t-1}) = d\big) = \frac{\exp\left(\gamma_1(i) + \gamma_2(i)(d \wedge \tau)\right)}{1 + \exp\left(\gamma_1(i) + \gamma_2(i)(d \wedge \tau)\right)}, \quad i = 0, 1, \tag{3}$$

where $d \wedge \tau = \min\{d, \tau\}$, and $\gamma_j(i)$, $i, j \in \{1, 2\}$, are parameters to be estimated. Following Maheu and McCurdy (2000b), the model for the conditional volatility is given by:

$$\sigma\big(S_t, D(S_t)\big) = \big(\omega(S_t) + \zeta(S_t) D(S_t)\big)^2, \tag{4}$$

where the latent state affects the level of volatility, $\omega(S_t) = \omega_0(1 - S_t) + \omega_1 S_t$. Additionally, the duration of the states, $D(S_t)$, influences the dynamics of volatility through $\zeta(S_t)D(S_t)$, where, $\zeta(S_t) = \zeta_0(1 - S_t) + \zeta_1 S_t$.

Unlike many conditional volatility models, the DDMS model incorporates discrete states. However, through conditioning on the duration $D(S_t)$ in the conditional variance, it allows for a smoother transition between volatility levels compared to a simple 2-state Markov-Switching (MS) model. As demonstrated in the Appendix A, DDMS models can be viewed as a multi-state MS model, introducing a much richer set of dynamics. Furthermore, in contrast to GARCH, DDMS incorporates the modeling feature of time-varying persistence in volatility levels. Unlike GARCH models, where a unit change in Z_t^2 has a linear effect on volatility forecasts, DDMS introduces a highly non-linear impact, as detailed by Maheu and McCurdy (2000b). Consequently, in this paper, we conduct a comparison of forecast accuracy between DDMS and GARCH models.

2.2 Estimation

The DDMS model can be estimated using maximum likelihood. It is considered an extension of Hamilton's model (Hamilton, 1989), where a new latent variable, $\mathcal{S}_t$, encompasses all possibilities of historical trajectories from S_t to τ:

$$\begin{aligned}
\mathcal{S}_t &= 1, \quad \text{if} \quad S_t = 1, S_{t-1} = 0, S_{t-2}, \cdots, D(S_t) = 1, \\
\mathcal{S}_t &= 2, \quad \text{if} \quad S_t = 1, S_{t-1} = 1, S_{t-2}, \cdots, D(S_t) = 2, \\
\mathcal{S}_t &= 3, \quad \text{if} \quad S_t = 0, S_{t-1} = 1, S_{t-2}, \cdots, D(S_t) = 1, \\
&\vdots \\
\mathcal{S}_t &= N, \quad \text{if} \quad S_t = 0, S_{t-1} = 0, S_{t-2} = 0, \cdots, D(S_t) = \tau,
\end{aligned}$$

where $N := 2 + 2(\tau - 1)$ are the extended states. Let $\mathcal{P}$ be the $N \times N$ transition matrix associated to $\mathcal{S}_t$, whose (i, j)-*th* entry is given by $[P]_{i,j} = P(\mathcal{S}_t = j|\mathcal{S}_{t-1} = i)$. State filtering and maximum likelihood estimation can be performed based on these principles, following Hamilton (1989). Additional details on the log-likelihood function for DDMS models are available in Appendix A.

However, maximum likelihood estimation of DDMS models can be challenging and prone to numerical issues. The likelihood function may have multiple local maxima, flat and spiky regions making numerical optimization difficult. Additionally, the estimation process in DDMS models with a large duration parameter can also pose computational difficulties. A large duration parameter often leads to a large sparse transition matrix in the extended representation of the DDMS model. This results in a sparse transition matrix that may approach singularity for several combinations of parameters. Singularity means that some states are given probabilities numerically close to zero, causing the unconditional probabilities to not exist. This results in the likelihood function being undefined at multiple points in the parameter space.

To address these challenges, we developed an optimization algorithm especially tailored to maximize the log-likelihood function of DDMS models. The proposed approach applies a combination of random draws for initial parameter values and constrained numerical optimization within defined bounds to tackle multimodality. We also impose nonlinear restrictions to ensure the invertibility of the transition matrix. The details of the algorithm are outlined in Appendix B.

2.3 Forecast combination

This paper presents a strategy for forecasting volatility using DDMS models, addressing challenges in duration selection by aggregating predictions from multiple models. We employ a model combination procedure to integrate forecasts from N individual models, each estimated using a different duration parameter, into a pooled modeling approach. Let $\hat{\sigma}^2_{i,t+1}$ represent the conditional variance forecast estimated by a DDMS model with duration τ_i. We define $\hat{\sigma}^2_{t+1}$ as the weighted average of the individual volatility forecasting models $\left\{\hat{\sigma}^2_{i,t+1}\right\}_{i=1}^{N}$:

$$\hat{\sigma}^2_{t+1} = \sum_{i=1}^{N} w_{i,t}\, \hat{\sigma}^2_{i,t+1}, \tag{5}$$

where $\{w_{i,t}\}_{i=1}^{N}$ are the combination weights at time t. Note that $\hat{\sigma}^2_{t+1}$ depends on various DDMS models with different duration parameters. The combination weights play a crucial role in correcting for potential misspecification or instabilities in τ, leading to potential improvements in forecasting accuracy. Models associated with higher forecasting errors due to a misspecified τ_i are expected to be assigned lower weights in the combination process. The idea is that the combined forecast $\hat{\sigma}^2_{t+1}$ will be more precise than each individual model

$\{\hat{\sigma}^2_{i,t+1}\}_{i=1}^N$. For a comprehensive review of the theory behind forecasting combination, refer to Timmermann (2006).

For the combining weights, we apply several different methods. Initially, we utilize simple approaches such as the sample mean and the sample median. Furthermore, we employ the discounted mean square error method (MSE) of Stock and Watson (2004). In particular, this method uses the historical forecasting performance of the individual models over a holdout out-of-sample period as weights:

$$w_{i,t} = \varphi_{i,t}^{-1} \Big/ \sum_{j=1}^{N} \varphi_{j,t}^{-1}, \tag{6}$$

where

$$\varphi_{i,t} = \sum_{s=n+1}^{t} \theta^{t-s} \left(\bar{\sigma}_s^2 - \hat{\sigma}_{i,s}^2\right)^2, \tag{7}$$

where θ is a discount factor, n is the holdout out-of-the sample period, and $\bar{\sigma}_s^2$ represents the volatility proxy. We set $\theta = 1$, i.e. there is no discounting and the individual forecasts are uncorrelated. This case corresponds to the Bates and Granger (1969) optimal weighting scheme.

Additionally, we employ the Granger and Ramanathan (1984) OLS estimator, expressed in terms of MSE as follows

$$\mathbf{w}_t = \underset{0 \le w_{i,t} \le 1, \sum_{i=1}^{N} w_{i,t}=1}{\arg\min} \quad \frac{1}{(t-s)} \sum_{s=n+1}^{t} \left(\bar{\sigma}_s^2 - \hat{\sigma}_s^2\right)^2. \tag{8}$$

Here, $\mathbf{w}_t = \{w_{i,t}\}_{i=1}^N$. Following Timmermann (2006), we impose convexity constraints $0 \le w_{i,t} \le 1$ and the additivity constraint $\sum_{i=1}^N w_{i,t} = 1$. The convexity constraints prevent the combined forecast from lying outside the range of the individual forecasts. The additivity constraint allows for the interpretability of the weights as a measure of the importance of each individual model.[5] Throughout this paper, we refer to this method as *optimal weights MSE*.

Finally, the last method we consider follows a similar approach to the Granger and Ramanathan (1984) OLS estimator, albeit within the context of a different loss function. This strategy was previously employed by Patton and Sheppard (2009), who utilized volatility forecast combinations based on the robust loss function proposed by Patton (2011). We incorporate this additional method due to the prevalent use of robust loss functions in evaluating volatility forecasts, especially with common volatility proxies. In this case, the weights are determined by solving the following minimization problem:

$$\mathbf{w}_t = \underset{0 \le w_{i,t} \le 1, \sum_{i=1}^{N} w_{i,t}=1}{\arg\min} \quad \frac{1}{(t-s)} \sum_{s=n+1}^{t} L(\bar{\sigma}_s^2, \hat{\sigma}_s^2), \tag{9}$$

where $L(\bar{\sigma}_s^2, \hat{\sigma}_s^2)$ represents a robust loss function. Similar to the previous method, we impose convexity and additivity constraints.[6] In this paper, we employ the QLIKE function as the

[5]The additivity constraint relies on the assumption of unbiased individual forecasts. However, this specification may not be efficient in the presence of any biases in the forecasts. On the other hand, Diebold (1988) argues that biased forecasts are not inherently undesirable. Without the additivity constraint, forecasting errors may exhibit their own predictability. For robustness we also studied unconstrained forecast combination, which led to higher forecasting errors. However, these adjustments did not alter the overall qualitative results. This additional set of results is available upon request.

[6]In this case, the additivity restrictions were not binding, and we observed virtually no difference in forecasting performance.

additional robust loss in the case of volatility proxies (see Patton, 2011). We refer to this method throughout the paper as *optimal weights QLIKE*.

While extensive research has delved into advanced model combination techniques, including time-varying weights and Bayesian model averaging methods (see for example Koop and Korobilis, 2012), our emphasis is on simpler, traditional methods. Given the numerous methods for forecasting combination in the literature, we choose to concentrate on traditional approaches. Our objective is not to identify the best combination approach for DDMS models, but rather to analyze whether combination methods can effectively address the challenge of duration selection.

3 Empirical results

For our study, we utilize daily log-returns from three distinct asset classes: a cryptocurrency (Bitcoin or BTC), a stock market index (S&P 500 or SP500), and a foreign currency exchange rate (EUR/USD or FX). The dataset spans from January 2015 to January 2020, encompassing a total of 1824, 1259, and 1291 observations for Bitcoin, SP500, and FX, respectively. Let P_t represent the daily closing value of the asset, then log-returns are defined as $r_t = \log(P_t) - \log(P_{t-1})$. Figure (1) presents a time series plot of the data.

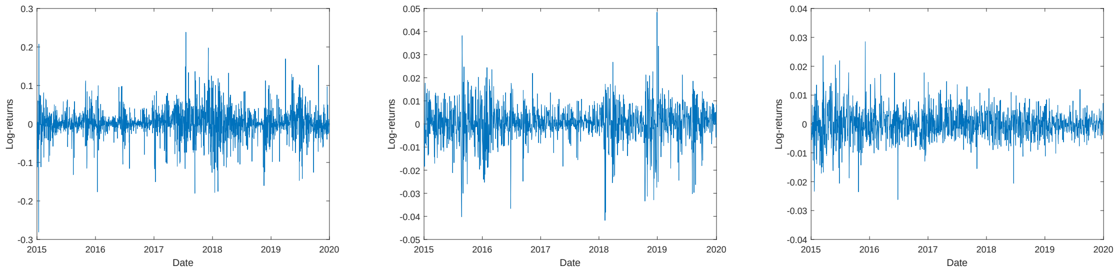
Table (1) provides descriptive statistics for the log-returns of the three assets. All returns exhibit mean values statistically close to zero, with Bitcoin demonstrating higher variability. Bitcoin's largest price increase is 0.2384, while the largest price decrease is -0.2175, resulting in a range of 0.5193. In contrast, SP500 and FX report smaller ranges of 0.0902 and 0.0547, respectively. Examining the distribution shapes, Bitcoin and SP500 display negative skewness, while FX exhibits positive skewness close to zero. Additionally, all returns have kurtosis values exceeding three. Consequently, Jarque-Bera (JB) statistics indicate a departure from normality for all three series. We also present the Lagrange Multiplier Test and Ljung-Box statistics, both highly significant, suggesting ARCH effects in all three series. Finally, the Augmented Dickey-Fuller tests reject the null hypothesis, indicating that all series are stationary.

Table 1: Descriptive statistics of the log-returns. JB refers to the Jarque-Bera test of normality. LM is the Lagrange Multiplier test for ARCH effects in the demeaned returns, while Q^2 is the corresponding Ljung-Box statistic on the squared demeaned returns, respectively. ADF refers to the p-value of the Augmented Dickey-Fuller test for stationarity.

	Obs.	Mean	Std. Dev.	Max.	Min.	Skewness	Kurtosis	JB	LM(8)	$Q^2(8)$	ADF
BTC	1824	0.0017	0.0394	0.2384	-0.2809	-0.2802	8.4721	2299.6	181.5	273.1	0.001
SP500	1259	0.0004	0.0085	0.0484	-0.0418	-0.5260	6.8360	829.9	177.3	373.8	0.001
FX	1291	-0.0001	0.0053	0.0285	-0.0262	0.0305	5.9067	454.7	58.7	90.3	0.001

The out-of-sample analysis is conducted using an expanding window estimation from April 2018 to January 2020. During this period, we have 641, 443, and 451 daily out-of-sample observations for Bitcoin, SP500, and FX, respectively. For computing combination weights, a holdout out-of-sample period of 30 days is employed. This forecasting sample period is intentionally chosen to be very volatile, aiming to test the forecasting ability of the DDMS model under challenging market conditions. Additionally, this period extends beyond other empirical applications that utilize Markov-switching models to forecast BTC volatility (see, for example Ardia et al., 2019; Segnon and Bekiros, 2020).

To evaluate the DDMS combination volatility forecast, we utilize intraday 5-minute quotes as an alternative source to proxy volatility, departing from the use of squared daily returns



(a) BTC log-returns (b) SP500 log-returns (c) FX log-returns

Figure 1: Time series plot of the log-returns of BTC, SP500, and FX.

(see Andersen and Bollerslev, 1998). We employ the realized variance (RV) and several other realized measures robust to microstructure noise and jumps.

3.1 Realized measures

A well-known problem in financial econometrics is that volatility is a latent variable and, therefore, cannot be directly observed, turning the evaluation of the predictive power of different approaches problematic. In this section, we consider the 5-minute intra-day quotes as an alternative source to proxy volatility. Specifically, in addition to the traditional realized variance (RV) (see Andersen et al., 2003), we explore alternative realized measures robust to jumps and microstructure noise to proxy the true volatility. These measures include bipower variation (BV) (see Barndorff-Nielsen and Shephard, 2004), MinRV, and MedRV (Andersen et al., 2012).

In general, these measures are regarded as better alternatives to proxy the true volatility compared to the squared daily return (see, for instance, McAleer and Medeiros, 2008; Alizadeh et al., 2002, among others.). Their definition is as follows:

$$
\begin{aligned}
\mathrm{RV}_t &:= \sum_{i=1}^{m} r_{i,t}^2, \\
\mathrm{BV}_t &:= \frac{\pi}{2}\left[\frac{m}{m-1}\right]\sum_{i=1}^{m-1} |r_{i,t} r_{i+1,t}|, \\
\mathrm{MinRV}_t &:= \frac{\pi}{\pi-2}\left[\frac{m}{m-1}\right]\sum_{i=1}^{m-1} \min\big\{|r_{i,t}|, |r_{i+1,t}|\big\}^2, \\
\mathrm{MedRV}_t &:= \frac{\pi}{6-4\sqrt{3}+\pi}\left[\frac{m}{m-2}\right]\sum_{i=2}^{m-1} \mathrm{med}\big\{|r_{i-1,t}|, |r_{i,t}|, |r_{i+1,t}|\big\}^2,
\end{aligned}
$$

where, $r_{i,t}$ is the ith high-frequency return of day t. The 5-minute intra-day quotes were obtained from First Rate Data website https://firstratedata.com.

3.2 Robust loss function

Typically, volatility proxies are used to evaluate volatility forecasts. However, these proxies are estimates of the integrated variance and, as such, they are imperfect. Patton (2011) defines a sense of robustness for loss functions in ranking volatility forecasts, and based on this concept, derived a general class of loss functions that are robust in that sense. Letting $\bar{\sigma}^2$ the volatility proxy and $\hat{\sigma}^2$ the volatility forecasts, we consider two loss functions to evaluate volatility forecasts, namely, the MSE and the QLIKE, given by

$$\text{MSE}(\bar{\sigma}^2, \hat{\sigma}^2) := \frac{(\bar{\sigma}^2 - \hat{\sigma}^2)^2}{2}, \quad \text{QLIKE}(\bar{\sigma}^2, \hat{\sigma}^2) := \frac{\bar{\sigma}^2}{\hat{\sigma}^2} - \log\left(\frac{\bar{\sigma}^2}{\hat{\sigma}^2}\right) - 1.$$

Observe that the MSE and QLIKE losses as defined by Patton (2011) differ from the usually applied loss functions of the same name.

3.3 Forecasting results

We are investigating the conditional volatility model described in equations (1)-(4) assuming a zero-mean process. To proceed, it is necessary to define the pool of individual DDMS models, which will subsequently be combined for forecasting. Previous research has often set $\tau = 25$, arguing its adequacy to capture duration effects in transition probabilities (See Maheu and McCurdy, 2000b). In this study, we construct a set of DDMS models centered around $\tau = 25$, encompassing both larger and smaller duration parameters. Specifically, we consider five DDMS models with $\tau \in \{5, 15, 25, 35, 45\}$. Importantly, in all cases, the duration parameter associated with the highest likelihood value at the beginning of the forecasting sample falls within this interval.

The forecasting study is structured employing a Model Confidence Set (MCS) to assess both individual and combined DDMS models. Additionally, a pairwise Diebold-Mariano test is employed for comparison with GARCH-type models.

3.3.1 Model confidence set

Tables (2)-(4) provide the average MSE and QLIKE obtained for the corresponding realized measures for BTC, SP500, and FX, respectively. Shaded in gray are the models with superior out-of-sample performance, as determined by the Model Confidence Set (MCS) approach outlined by Hansen et al. (2011), at a 75% significance level. The best models, i.e., those in the first position in the MCS rank, are indicated in bold.

DDMS models with optimal weights consistently emerge as the best-performing models, securing their position in the superior set across all cases. In contrast, individual models typically do not make it to the superior set. Notably, simpler methods like the median often outperform individual models in both MSE and QLIKE. Moreover, it is noteworthy that, although the model with $\tau = 25$ performs well in some cases, the $\tau = 5$ model closely competes, particularly for BTC, and the $\tau = 45$ model proves competitive for FX. Additionally, the best-performing individual model appears to vary depending on the loss function used, as observed in the case of SP500.

In summary, the MCS results suggest that combination methods are able to outperform individual models, and potentially mitigate the problem of duration selection in DDMS models.

Table 2: Average MSE and QLIKE values for log-returns of BTC. Gray cells indicate the set of models with the best out-of-sample performance obtained using the MCS approach at 75% significance level following (Hansen et al., 2011). 10000 bootstrapped samples were used to construct the test. In bold, the best model (first position in the MCS rank) in each case.

Models	MSE ($\times 10^{-5}$)				QLIKE			
	MedRV	MinRV	BV	RV	MedRV	MinRV	BV	RV
Mean	0.2228	0.2832	0.2451	0.2032	0.5363	0.5403	0.5136	0.4812
Median	0.2278	0.2832	0.2547	0.2180	0.5069	0.5082	0.4942	0.4809
Discounted MSE	0.2185	0.2788	0.2417	0.2004	0.5271	0.5316	0.5058	0.4748
Optimal weights MSE	**0.2007**	**0.2623**	**0.2304**	**0.1952**	**0.4123**	**0.4179**	**0.4254**	**0.4293**
Optimal weights QLIKE	0.2075	0.2716	0.2385	0.2020	0.4307	0.4372	0.4419	0.4502
$\tau = 5$	0.2544	0.3140	0.2784	0.2376	0.5388	0.5385	0.5292	0.5240
$\tau = 15$	0.2942	0.3560	0.3191	0.2786	0.5780	0.5772	0.5688	0.5641
$\tau = 25$	0.2476	0.3079	0.2762	0.2378	0.5230	0.5273	0.5253	0.5327
$\tau = 35$	0.3711	0.4300	0.3917	0.3542	0.6449	0.6453	0.6312	0.6245
$\tau = 45$	0.4248	0.4863	0.4383	0.3861	0.6658	0.6714	0.6440	0.6147

Table 3: Average MSE and QLIKE values for log-returns of SP500. Gray cells indicate the set of models with the best out-of-sample performance obtained using the MCS approach at 75% significance level following (Hansen et al., 2011). 10000 bootstrapped samples were used to construct the test. In bold, the best model (first position in the MCS rank) in each case.

Models	MSE ($\times 10^{-6}$)				QLIKE			
	MedRV	MinRV	BV	RV	MedRV	MinRV	BV	RV
Mean	0.0148	0.0153	0.0149	0.0145	0.7067	0.7291	0.6817	0.6285
Median	0.0028	0.0029	0.0028	0.0027	0.6277	0.6496	0.6043	0.5548
Discounted MSE	0.0107	0.0113	0.0108	0.0102	0.7135	0.7353	0.6883	0.6358
Optimal weights MSE	0.0026	0.0027	0.0024	0.0021	0.4343	0.4529	0.4232	0.3989
Optimal weights QLIKE	**0.0017**	**0.0020**	**0.0019**	**0.0019**	**0.4092**	**0.4237**	**0.4002**	**0.3856**
$\tau = 5$	0.0075	0.0075	0.0075	0.0074	0.7058	0.7256	0.6809	0.6319
$\tau = 15$	0.0309	0.0323	0.0313	0.0302	0.6570	0.6759	0.6369	0.5981
$\tau = 25$	0.0071	0.0072	0.0071	0.0070	0.6579	0.6791	0.6347	0.5871
$\tau = 35$	0.1912	0.1920	0.1913	0.1903	0.7020	0.7212	0.6784	0.6299
$\tau = 45$	0.0115	0.0115	0.0115	0.0114	0.7117	0.7352	0.6918	0.6459

Table 4: Average MSE and QLIKE values for log-returns of FX. Gray cells indicate the set of models with the best out-of-sample performance obtained using the MCS approach at 75% significance level following (Hansen et al., 2011). 10000 bootstrapped samples were used to construct the test. In bold, the best model (first position in the MCS rank) in each case.

Models	MSE ($\times 10^{-9}$)				QLIKE			
	MedRV	MinRV	BV	RV	MedRV	MinRV	BV	RV
Mean	0.2530	0.2550	0.2431	0.2242	0.4445	0.4497	0.4213	0.3789
Median	0.2208	0.2226	0.2116	0.1946	0.4205	0.4255	0.3983	0.3586
Discounted MSE	0.2350	0.2373	0.2246	0.2068	0.4326	0.4379	0.4096	0.3677
Optimal weights MSE	0.0454	0.0472	0.0491	0.0572	0.1585	0.1631	0.1609	0.1666
Optimal weights QLIKE	**0.0452**	**0.0463**	**0.0472**	**0.0585**	**0.1507**	**0.1551**	**0.1537**	**0.1630**
$\tau = 5$	0.7036	0.7050	0.6904	0.6635	0.5041	0.5090	0.4807	0.4380
$\tau = 15$	0.2548	0.2583	0.2463	0.2294	0.4282	0.4341	0.4074	0.3695
$\tau = 25$	0.2467	0.2486	0.2378	0.2222	0.3993	0.4048	0.3788	0.3436
$\tau = 35$	0.2504	0.2521	0.2406	0.2218	0.4309	0.4357	0.4086	0.3674
$\tau = 45$	0.2239	0.2258	0.2149	0.1987	0.4097	0.4148	0.3882	0.3502

3.3.2 Subsample analysis

To assess the robustness of our methodology, we conduct a subsample analysis by examining the values of the loss functions in the first and second halves of the forecasting sample. Our primary objectives are twofold: first, to explore the behavior of individual models and duration selection within these subsamples; second, to verify whether the combination methods are consistent in outperforming individual methods.

Table (5) presents the average MSE values for the first and second halves of the forecasting sample across the three assets. We chose to present the five individual models and the two best combination methods. The best-performing individual model in each subsample is highlighted in bold. Notably, the optimal individual model varies depending on the forecasting sample. For BTC and FX, the best duration does not even consist of contiguous models, while for SP500, the best full-sample model differs from the best models in the subsamples. These observations highlight the complex nature of duration selection, as no single duration model appears to consistently outperform in both samples. Additionally, combination methods show robustness in their results, consistently outperforming individual models in both subsamples.

In Table (6), we also present the average value of the QLIKE. A similar pattern emerges as observed with MSE. The optimal duration appears to change throughout the forecasting sample. Furthermore, it is noteworthy that the best duration according to QLIKE differs from that based on MSE for BTC and FX. Nevertheless, combination methods demonstrate consistency in their performance, outperforming individual models even when evaluated using different loss functions.

These findings suggest the complex nature of duration selection in forecasting. The choice of the duration parameter, and consequently, model, appears to vary depending on the forecasting sample and the employed loss function. However, combination methods are able to accommodate possible specification issues in the duration, thereby providing more accurate forecasts.

Table 5: Average MSE in the first and second halves of the forecasting sample. The five individual models and the two best combination methods are presented. In bold, the best individual model. The MSE scale for each asset corresponds to the one presented in the MCS results.

Asset	Models	First Half				Second Half			
		MedRV	MinRV	BV	RV	MedRV	MinRV	BV	RV
BTC	$\tau = 5$	0.1442	0.1413	0.1385	0.2032	**0.3651**	**0.4872**	**0.4188**	0.3378
	$\tau = 15$	0.2234	0.2184	0.2168	0.2216	0.3653	0.4940	0.4217	**0.3357**
	$\tau = 25$	**0.1160**	**0.1120**	**0.1186**	**0.1276**	0.3796	0.5045	0.4343	0.3483
	$\tau = 35$	0.1590	0.1562	0.1583	0.1678	0.5839	0.7048	0.6259	0.5412
	$\tau = 45$	0.2380	0.2342	0.2289	0.2317	0.6122	0.7392	0.6483	0.5411
	Optimal weights MSE	0.0646	0.0611	0.0718	0.0899	0.3373	0.4641	0.3895	0.3009
	Optimal weights QLIKE	0.0615	0.0579	0.0662	0.0848	0.3540	0.4860	0.4114	0.3196
SP500	$\tau = 5$	**0.0072**	**0.0071**	**0.0072**	**0.0073**	0.0079	0.0079	0.0078	0.0076
	$\tau = 15$	0.0592	0.0618	0.0599	0.0579	**0.0027**	**0.0027**	**0.0026**	**0.0026**
	$\tau = 25$	0.0082	0.0082	0.0081	0.0080	0.0061	0.0062	0.0061	0.0059
	$\tau = 35$	0.3775	0.3792	0.3779	0.3760	0.0049	0.0049	0.0048	0.0045
	$\tau = 45$	0.0148	0.0149	0.0148	0.0149	0.0081	0.0082	0.0081	0.0079
	Optimal weights MSE	0.0043	0.0043	0.0039	0.0032	0.0009	0.0010	0.0009	0.0010
	Optimal weights QLIKE	0.0027	0.0033	0.0030	0.0028	0.0008	0.0008	0.0008	0.0009
FX	$\tau = 5$	0.7122	0.7129	0.6972	0.6682	0.6950	0.6971	0.6836	0.6588
	$\tau = 15$	0.2326	0.2377	0.2218	0.2032	0.2772	0.2789	0.2709	0.2557
	$\tau = 25$	0.2896	0.2920	0.2770	0.2567	**0.2036**	**0.2051**	**0.1984**	**0.1876**
	$\tau = 35$	0.2522	0.2542	0.2394	0.2194	0.2487	0.2499	0.2419	0.2241
	$\tau = 45$	**0.2232**	**0.2257**	**0.2111**	**0.1911**	0.2246	0.2259	0.2187	0.2064
	Optimal weights MSE	0.0587	0.0615	0.0630	0.0706	0.0321	0.0328	0.0352	0.0437
	Optimal weights QLIKE	0.0640	0.0654	0.0650	0.0788	0.0264	0.0271	0.0294	0.0381

3.3.3 Comparison to GARCH-type models

To assess the accuracy of forecasts generated by both combined and individual DDMS models, our analysis also incorporates traditional GARCH-type models for pairwise evaluation. Following Haas et al. (2004), we employ a plain vanilla specification given by

$$\begin{aligned} Y_t &= \sigma_{k,t} Z_t, \\ \sigma^2_{k,t} &= \omega_k + \alpha_k Y^2_{t-1} + \beta_k \sigma^2_{k,t-1}, \end{aligned}$$

where k is the number of regimes. We considered $k = 1$ for the traditional model and $k = 2$ for the two-regime model. In both models Z_t are i.i.d. $N(0, 1)$. Despite the various variations of this modeling approach, our goal is to evaluate the performance of the DDMS models with that of simpler yet useful GARCH models.

Tables (7)-(9) present the t-statistics resulting from Diebold-Mariano tests, examining equal predictive accuracy between the benchmark GARCH and MS-GARCH models and the DDMS models. This analysis considers the same two losses and four realized measures as in

Table 6: Average QLIKE in the first and second halves of the forecasting sample. The five individual models and the two best combination methods are presented. In bold, the best individual model.

Asset	Models	First Half				Second Half			
		MedRV	MinRV	BV	RV	MedRV	MinRV	BV	RV
BTC	$\tau = 5$	0.5832	0.5667	**0.5423**	**0.5140**	0.4943	0.5102	0.5160	0.5340
	$\tau = 15$	0.6532	0.6305	0.6081	0.5876	0.5027	0.5238	0.5294	0.5405
	$\tau = 25$	**0.5573**	**0.5480**	0.5452	0.5515	**0.4886**	**0.5066**	**0.5053**	**0.5138**
	$\tau = 35$	0.6874	0.6744	0.6651	0.6548	0.6022	0.6162	0.5972	0.5942
	$\tau = 45$	0.7289	0.7276	0.6944	0.6583	0.6026	0.6151	0.5934	0.5709
	Optimal weights MSE	0.3846	0.3737	0.3968	0.4150	0.4400	0.4621	0.4541	0.4437
	Optimal weights QLIKE	0.3962	0.3808	0.3916	0.4127	0.4653	0.4938	0.4924	0.4877
SP500	$\tau = 5$	0.5403	0.5632	0.5245	0.4833	0.8713	0.8881	0.8373	0.7805
	$\tau = 15$	0.5513	0.5714	0.5384	0.5030	**0.7628**	**0.7803**	**0.7354**	**0.6932**
	$\tau = 25$	**0.4890**	**0.5115**	**0.4741**	**0.4351**	0.8268	0.8468	0.7954	0.7390
	$\tau = 35$	0.6085	0.6255	0.5938	0.5614	0.7954	0.8169	0.7631	0.6984
	$\tau = 45$	0.5976	0.6235	0.5894	0.5547	0.8257	0.8468	0.7941	0.7371
	Optimal weights MSE	0.3546	0.3741	0.3535	0.3304	0.5140	0.5317	0.4929	0.4673
	Optimal weights QLIKE	0.3653	0.3775	0.3572	0.3379	0.4532	0.4699	0.4432	0.4333
FX	$\tau = 5$	0.2869	0.2921	0.2655	0.2311	0.7224	0.7269	0.6969	0.5080
	$\tau = 15$	0.2810	0.2886	0.2623	0.2316	0.5761	0.5803	0.5532	0.2557
	$\tau = 25$	0.2789	0.2846	0.2583	0.2271	**0.5204**	**0.5256**	**0.4999**	**0.4606**
	$\tau = 35$	0.2835	0.2888	0.2628	0.2310	0.5790	0.5834	0.5551	0.5044
	$\tau = 45$	**0.2635**	**0.2691**	**0.2434**	**0.2121**	0.5566	0.5612	0.5337	0.4889
	Optimal weights MSE	0.1006	0.1068	0.1018	0.1016	0.2167	0.2197	0.2204	0.2318
	Optimal weights QLIKE	0.1053	0.1109	0.1060	0.1088	0.1963	0.1995	0.2017	0.2175

the MCS, specifically addressing Bitcoin, SP500, and FX, respectively. A negative t-statistic indicates that the DDMS models' forecasts produced a larger average loss than the GARCH models. A t-statistic greater than 1.65 and 1.96 in absolute value indicates a rejection of the null hypothesis of equal predictive accuracy at the 0.10 and 0.05 levels. Instances where the statistic is positive, denoting point superiority of DDMS models, are highlighted in bold. For ease of reference, we include p-values in parentheses.

In the BTC results presented in Table (7), the tests generally yield positive statistics, with some instances pointing towards rejection when optimal weights are employed, indicating the superiority of DDMS combinations in specific cases. This contrast is not observed in individual models. For the MS-GARCH model, the statistic tends to be negative, but with optimal weights, the test does not consistently achieve statistical significance, suggesting comparable accuracy. In terms of MSE, alternative methods also do not lead to null hypothesis rejection. Notably, individual models do not prove competitive against MS-GARCH, as evidenced by the rejection of all tests.

For the SP500, the test statistics for optimal weights are predominantly positive. In some of these cases, the test statistics are significant, suggesting that DDMS models provide more accurate forecasts compared to GARCH and MS-GARCH models. It is noteworthy that for individual models, none of the test statistics are positive, mostly leading to the rejection of the null hypothesis of equal forecasting accuracy.

Similar patterns are observed for FX. While the test statistics for comparisons with

GARCH models are mostly negative, they do not result in the rejection of the null hypothesis for optimal weights. On the other hand, these statistics are predominantly positive and significant for the comparison with MS-GARCH models.

In summary, the comparison with GARCH models highlights the advantages of combining DDMS models. Individual models often exhibit higher loss functions than GARCH models, leading to frequent rejection of their test statistics. In contrast, combination methods demonstrate the capacity to achieve lower loss functions and yield statistically significant results in certain cases. While the combination approach does not necessarily outperform GARCH and MS-GARCH in all cases, our findings suggest that this strategy, which addresses the issue of duration selection, has the potential to outperform benchmark models.

4 Conclusions

This paper investigates the volatility forecasting performance of duration-dependent Markov-switching model addressing the issue of duration selection. Previous research has neither consider the forecasting performance of DDMS volatility models nor the implications of using a constant duration parameter. The application of DDMS for forecasting is hindered by the absence of a procedure to estimate or test a specific duration for a given dataset. To tackle the issue of duration selection, we propose the use of forecasting combinations of DDMS models with different duration parameters. Our assumption is that the combined forecast will be more robust to duration misspecification, thus yielding more accurate forecasts. To achieve this, we rely on a model combination approach, aggregating forecasts from several individual models, each indexed by a fixed duration parameter.

The forecasting study employs a Model Confidence Set to evaluate both individual and combined DDMS models. Additionally, a pairwise Diebold-Mariano test is used for comparison with GARCH-type models. Model Confidence Set results indicate the superiority of combined DDMS models. The choice of the duration parameter significantly influences forecasting performance. Subsample analysis reveals that the best individual models vary across forecasting samples, while the combination approach consistently outperforms individual models. These findings underscore the complexity of duration selection, as no individual model appears superior across subsamples and loss functions. In conclusion, combination methods demonstrate promise in outperforming individual models and addressing duration specification challenges in DDMS models.

The comparison with GARCH and MS-GARCH models emphasizes that the combination approach improves forecasting performances. While individual models often incur higher loss functions than GARCH models, combination methods usually achieve lower losses and statistical significance in specific cases. Although not necessarily outperforming GARCH and MS-GARCH in all cases, our findings suggest that this strategy, addressing specification issues in duration, further highlights the forecasting power of DDMS models.

For future research, different extensions of this paper can be pursued. This includes assessing the performance of the combination approach in other economic applications, such as risk assessment through Value-at-Risk (VaR) estimation or constructing investment strategies based on the regime probabilities of these classes of models, as seen in the trading strategies literature.

Table 7: Results from the Diebold-Mariano test applied to the Garch model and MS-GARCH and BTC. Presented are the t-statistics from Diebold-Mariano test along with their p-value in parentheses. In bold, cases where the test statistic is positive.

Models		Loss	MedRV		MinRV		BV		RV	
GARCH	Mean	MSE	**0.43**	**(0.669)**	**0.53**	**(0.598)**	**0.60**	**(0.549)**	**0.73**	**(0.461)**
		QLIKE	**0.07**	**(0.947)**	**0.09**	**(0.926)**	-0.22	(0.824)	-0.74	(0.461)
	Median	MSE	**0.49**	**(0.627)**	**0.62**	**(0.539)**	**0.47**	**(0.642)**	**0.33**	**(0.743)**
		QLIKE	**0.40**	**(0.693)**	**0.43**	**(0.666)**	**0.58**	**(0.563)**	**0.88**	**(0.376)**
	Discounted MSE	MSE	**0.61**	**(0.539)**	**0.71**	**(0.480)**	**0.75**	**(0.456)**	**0.87**	**(0.385)**
		QLIKE	**0.06**	**(0.949)**	**0.03**	**(0.973)**	**0.33**	**(0.739)**	**0.83**	**(0.407)**
	Optimal weights MSE	MSE	**1.68**	**(0.092)**	**1.67**	**(0.095)**	**1.34**	**(0.180)**	**1.18**	**(0.238)**
		QLIKE	**2.02**	**(0.044)**	**2.04**	**(0.041)**	**1.62**	**(0.105)**	**1.51**	**(0.132)**
	Optimal weights QLIKE	MSE	**1.66**	**(0.098)**	**1.52**	**(0.129)**	**1.36**	**(0.175)**	**1.17**	**(0.243)**
		QLIKE	**2.02**	**(0.043)**	**1.98**	**(0.047)**	**1.77**	**(0.076)**	**1.64**	**(0.101)**
	$\tau = 5$	MSE	-1.09	(0.274)	-0.85	(0.393)	-0.93	(0.354)	-0.91	(0.361)
		QLIKE	-0.13	(0.898)	-0.08	(0.936)	**0.01**	**(0.990)**	**0.27**	**(0.786)**
	$\tau = 15$	MSE	-2.57	(0.01)	-2.51	(0.012)	-2.41	(0.016)	-2.41	(0.016)
		QLIKE	**0.13**	**(0.893)**	**0.11**	**(0.914)**	**0.07**	**(0.944)**	**0.10**	**(0.918)**
	$\tau = 25$	MSE	-1.14	(0.254)	-0.92	(0.360)	-1.19	(0.233)	-1.10	(0.270)
		QLIKE	**0.13**	**(0.893)**	**0.11**	**(0.914)**	**0.07**	**(0.944)**	**0.10**	**(0.918)**
	$\tau = 35$	MSE	-2.12	(0.034)	-2.06	(0.039)	-2.03	(0.043)	-2.03	(0.042)
		QLIKE	-1.46	(0.145)	-1.49	(0.135)	-1.23	(0.220)	-0.91	(0.363)
	$\tau = 45$	MSE	-2.12	(0.034)	-2.06	(0.039)	-2.03	(0.043)	-2.03	(0.042)
		QLIKE	-2.23	(0.026)	-2.22	(0.026)	-2.05	(0.040)	-2.03	(0.043)
MS-GARCH	Mean	MSE	-1.23	(0.218)	-1.03	(0.305)	-0.79	(0.432)	-0.41	(0.679)
		QLIKE	-5.25	(< 0.001)	-5.17	(< 0.001)	-4.49	(< 0.001)	-3.03	(0.002)
	Median	MSE	-2.98	(0.003)	-2.34	(0.019)	-2.57	(0.010)	-2.64	(0.008)
		QLIKE	-4.59	(< 0.001)	-4.57	(< 0.001)	-4.39	(< 0.001)	-3.51	(< 0.001)
	Discounted MSE	MSE	-1.02	(0.308)	-0.81	(0.420)	-0.63	(0.528)	-0.27	(0.786)
		QLIKE	-5.01	(< 0.001)	-4.95	(< 0.001)	-4.32	(< 0.001)	-2.91	(0.004)
	Optimal weights MSE	MSE	-0.01	(0.993)	**0.12**	**(0.904)**	-0.02	(0.984)	**0.03**	**(0.975)**
		QLIKE	-1.08	(0.279)	-1.22	(0.223)	-1.98	(0.048)	-1.63	(0.104)
	Optimal weights QLIKE	MSE	-0.98	(0.327)	-1.11	(0.265)	-1.10	(0.271)	-0.72	(0.469)
		QLIKE	-1.18	(0.239)	-1.18	(0.239)	-1.59	(0.113)	-1.99	(0.047)
	$\tau = 5$	MSE	-3.04	(0.002)	-2.53	(0.011)	-2.69	(0.007)	-2.82	(0.005)
		QLIKE	-3.41	(0.001)	-3.15	(0.002)	-3.03	(0.002)	-2.86	(0.004)
	$\tau = 15$	MSE	-4.49	(< 0.001)	-4.42	(< 0.001)	-4.18	(< 0.001)	-4.00	(< 0.001)
		QLIKE	-5.82	(< 0.001)	-5.51	(< 0.001)	-5.05	(< 0.001)	-4.41	(< 0.001)
	$\tau = 25$	MSE	-3.94	(< 0.001)	-3.46	(0.001)	-3.61	(< 0.001)	-3.53	(< 0.001)
		QLIKE	-3.91	(< 0.001)	-3.83	(< 0.001)	-3.62	(< 0.001)	-3.27	(0.001)
	$\tau = 35$	MSE	-2.58	(0.010)	-2.51	(0.012)	-2.47	(0.014)	-2.41	(0.016)
		QLIKE	-6.37	(< 0.001)	-6.54	(< 0.001)	-5.67	(< 0.001)	-4.72	(< 0.001)
	$\tau = 45$	MSE	-2.59	(0.010)	-2.56	(0.010)	-2.41	(0.016)	-2.40	(0.016)
		QLIKE	-6.47	(< 0.001)	-6.38	(< 0.001)	-5.91	(< 0.001)	-4.82	(< 0.001)

References

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, 57(3):1047–1091.

Table 8: Results from the Diebold-Mariano test applied to the Garch model and MS-GARCH and SP500. Presented are the t-statistics from Diebold-Mariano test along with their p-value in parentheses. In bold, cases where the test statistic is positive.

Models		Loss	MedRV		MinRV		BV		RV	
GARCH	Mean	MSE	-1.64	(0.101)	-1.65	(0.098)	-1.64	(0.101)	-1.62	(0.105)
		QLIKE	-10.52	(< 0.001)	-10.18	(< 0.001)	-10.25	(< 0.001)	-10.15	(< 0.001)
	Median	MSE	-0.75	(0.453)	-0.50	(0.616)	-0.67	(0.506)	-0.87	(0.384)
		QLIKE	-6.73	(< 0.001)	-6.45	(< 0.001)	-6.52	(< 0.001)	-6.59	(< 0.001)
	Discounted MSE	MSE	-1.72	(0.086)	-1.63	(0.103)	-1.69	(0.090)	-1.76	(0.078)
		QLIKE	-10.53	(< 0.001)	-10.12	(< 0.001)	-10.23	(< 0.001)	-10.24	(< 0.001)
	Optimal weights MSE	MSE	-0.35	(0.729)	-0.06	(0.951)	**0.00**	**(0.999)**	**0.42**	**(0.678)**
		QLIKE	**1.75**	**(0.080)**	**1.85**	**(0.065)**	**1.42**	**(0.157)**	**0.78**	**(0.434)**
	Optimal weights QLIKE	MSE	**1.85**	**(0.064)**	**1.72**	**(0.085)**	**1.59**	**(0.111)**	**1.29**	**(0.197)**
		QLIKE	**1.99**	**(0.046)**	**2.19**	**(0.029)**	**1.79**	**(0.073)**	**1.10**	**(0.273)**
	$\tau = 5$	MSE	-3.18	(0.001)	-2.98	(0.003)	-3.12	(0.002)	-3.24	(0.001)
		QLIKE	-8.78	(< 0.001)	-8.45	(< 0.001)	-8.50	(< 0.001)	-8.40	(< 0.001)
	$\tau = 15$	MSE	-1.11	(0.266)	-1.11	(0.269)	-1.11	(0.267)	-1.12	(0.265)
		QLIKE	-7.48	(< 0.001)	-7.51	(< 0.001)	-7.35	(< 0.001)	-7.19	(< 0.001)
	$\tau = 25$	MSE	-1.88	(0.061)	-1.83	(0.067)	-1.85	(0.064)	-1.87	(0.062)
		QLIKE	-7.18	(< 0.001)	-6.86	(< 0.001)	-7.00	(< 0.001)	-7.21	(0.001)
	$\tau = 35$	MSE	-1.16	(0.245)	-1.17	(0.243)	-1.16	(0.245)	-1.16	(0.247)
		QLIKE	-8.24	(< 0.001)	-8.19	(< 0.001)	-8.07	(< 0.001)	-7.66	(< 0.001)
	$\tau = 45$	MSE	-2.82	(0.005)	-2.79	(0.005)	-2.81	(0.005)	-2.84	(0.005)
		QLIKE	-9.00	(< 0.001)	-8.83	(< 0.001)	-8.90	(< 0.001)	-8.96	(< 0.001)
MS-GARCH	Mean	MSE	-1.63	(0.104)	-1.64	(0.101)	-1.63	(0.103)	-1.61	(0.108)
		QLIKE	-9.72	(< 0.001)	-9.40	(< 0.001)	-9.47	(< 0.001)	-9.60	(< 0.001)
	Median	MSE	-0.83	(0.407)	-0.49	(0.622)	-0.71	(0.475)	-1.00	(0.317)
		QLIKE	-6.19	(< 0.001)	-5.93	(< 0.001)	-6.00	(< 0.001)	-6.21	(< 0.001)
	Discounted MSE	MSE	-1.72	(0.085)	-1.65	(0.100)	-1.70	(0.089)	-1.76	(0.078)
		QLIKE	-9.91	(< 0.001)	-9.51	(< 0.001)	-9.64	(< 0.001)	-9.87	(< 0.001)
	Optimal weights MSE	MSE	-0.39	(0.694)	-0.06	(0.950)	-0.03	(0.980)	**0.41**	**(0.684)**
		QLIKE	**2.11**	**(0.035)**	**2.23**	**(0.026)**	**1.75**	**(0.081)**	**1.03**	**(0.303)**
	Optimal weights QLIKE	MSE	**1.96**	**(0.050)**	**1.71**	**(0.088)**	**1.64**	**(0.101)**	**1.36**	**(0.173)**
		QLIKE	**2.30**	**(0.021)**	**2.52**	**(0.012)**	**2.10**	**(0.036)**	**1.33**	**(0.183)**
	$\tau = 5$	MSE	-3.22	(0.001)	-2.97	(0.003)	-3.14	(0.002)	-3.29	(0.001)
		QLIKE	-8.42	(< 0.001)	-8.11	(< 0.001)	-8.19	(< 0.001)	-8.24	(< 0.001)
	$\tau = 15$	MSE	-1.11	(0.265)	-1.11	(0.268)	-1.11	(0.266)	-1.11	(0.264)
		QLIKE	-6.86	(< 0.001)	-6.92	(< 0.001)	-6.78	(< 0.001)	-6.69	(< 0.001)
	$\tau = 25$	MSE	-1.85	(0.065)	-1.79	(0.074)	-1.82	(0.069)	-1.85	(0.065)
		QLIKE	-6.79	(< 0.001)	-6.47	(< 0.001)	-6.62	(< 0.001)	-6.97	(0.001)
	$\tau = 35$	MSE	-1.16	(0.246)	-1.17	(0.243)	-1.16	(0.245)	-1.16	(0.247)
		QLIKE	-7.47	(< 0.001)	-7.42	(< 0.001)	-7.36	(< 0.001)	-7.17	(< 0.001)
	$\tau = 45$	MSE	-2.77	(0.010)	-2.73	(0.010)	-2.76	(0.016)	-2.80	(0.016)
		QLIKE	-8.61	(< 0.001)	-8.44	(< 0.001)	-8.52	(< 0.001)	-8.71	(< 0.001)

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2012). Jump-robust volatility estimation

Table 9: Results from the Diebold-Mariano test applied to the Garch model and MS-GARCH and FX. Presented are the t-statistics from Diebold-Mariano test along with their p-value in parentheses. In bold, cases where the test statistic is positive.

Models		Loss	MedRV		MinRV		BV		RV	
GARCH	Mean	MSE	-9.33	(< 0.001)	-9.36	(< 0.001)	-8.82	(< 0.001)	-8.11	(< 0.001)
		QLIKE	-13.33	(< 0.001)	-13.25	(< 0.001)	-12.63	(< 0.001)	-11.21	(< 0.001)
	Median	MSE	-16.25	(< 0.001)	-16.29	(< 0.001)	-15.49	(< 0.001)	-14.15	(< 0.001)
		QLIKE	-12.30	(< 0.001)	-12.26	(< 0.001)	-11.69	(< 0.001)	-10.40	(< 0.001)
	Discounted MSE	MSE	-10.01	(< 0.001)	-10.02	(< 0.001)	-9.76	(< 0.001)	-8.77	(< 0.001)
		QLIKE	-13.46	(< 0.001)	-13.39	(< 0.001)	-12.72	(< 0.001)	-11.19	(< 0.001)
	Optimal weights MSE	MSE	-1.04	(0.299)	-1.06	(0.288)	-1.42	(0.154)	-1.87	(0.061)
		QLIKE	-0.73	(0.465)	-0.81	(0.415)	-2.04	(0.041)	-3.59	(< 0.001)
	Optimal weights QLIKE	MSE	-0.77	(0.443)	-0.77	(0.444)	-1.16	(0.247)	-1.69	(0.091)
		QLIKE	**0.22**	**(0.828)**	**0.14**	**(0.890)**	-1.43	(0.152)	-3.99	(< 0.001)
	$\tau = 5$	MSE	-3.16	(0.002)	-3.16	(0.002)	-3.09	(0.002)	-2.98	(0.003)
		QLIKE	-11.84	(< 0.001)	-11.75	(< 0.001)	-11.40	(< 0.001)	-10.59	(< 0.001)
	$\tau = 15$	MSE	-13.14	(< 0.001)	-12.99	(< 0.001)	-12.62	(< 0.001)	-11.65	(< 0.001)
		QLIKE	-12.43	(< 0.001)	-12.43	(< 0.001)	-11.94	(< 0.001)	-10.85	(< 0.001)
	$\tau = 25$	MSE	-6.60	(< 0.001)	-6.63	(< 0.001)	-6.36	(< 0.001)	-5.95	(< 0.001)
		QLIKE	-12.49	(< 0.001)	-12.45	(< 0.001)	-11.82	(< 0.001)	-10.49	(< 0.001)
	$\tau = 35$	MSE	-18.52	(< 0.001)	-18.66	(< 0.001)	-17.71	(< 0.001)	-16.11	(< 0.001)
		QLIKE	-13.85	(< 0.001)	-13.80	(< 0.001)	-13.14	(< 0.001)	-11.58	(< 0.001)
	$\tau = 45$	MSE	-15.67	(< 0.001)	-15.63	(< 0.001)	-17.71	(< 0.001)	-16.11	(< 0.001)
		QLIKE	-12.92	(< 0.001)	-12.82	(< 0.001)	-12.24	(< 0.001)	-10.84	(< 0.001)
MS-GARCH	Mean	MSE	-8.56	(< 0.001)	-8.60	(< 0.001)	-8.14	(< 0.001)	-7.53	(< 0.001)
		QLIKE	-13.56	(< 0.001)	-13.48	(< 0.001)	-13.02	(< 0.001)	-11.92	(< 0.001)
	Median	MSE	-14.78	(< 0.001)	-14.85	(< 0.001)	-14.37	(< 0.001)	-13.47	(< 0.001)
		QLIKE	-12.08	(< 0.001)	-12.03	(< 0.001)	-11.65	(< 0.001)	-10.77	(< 0.001)
	Discounted MSE	MSE	-9.09	(0.085)	-9.13	(0.100)	-8.95	(0.089)	-8.07	(0.078)
		QLIKE	-13.90	(< 0.001)	-13.82	(< 0.001)	-13.31	(< 0.001)	-12.08	(< 0.001)
	Optimal weights MSE	MSE	**1.45**	**(0.148)**	**1.49**	**(0.136)**	**0.78**	**(0.437)**	-0.22	(0.829)
		QLIKE	**5.30**	**(< 0.001)**	**5.29**	**(< 0.001)**	**4.07**	**(< 0.001)**	**1.70**	**(0.089)**
	Optimal weights QLIKE	MSE	**1.15**	**(0.252)**	**1.28**	**(0.199)**	**0.91**	**(0.364)**	-0.28	(0.776)
		QLIKE	**6.28**	**(< 0.001)**	**6.26**	**(< 0.001)**	**5.27**	**(< 0.001)**	**2.48**	**(0.013)**
	$\tau = 5$	MSE	-3.05	(0.002)	-3.05	(0.002)	-2.99	(0.003)	-2.90	(0.004)
		QLIKE	-11.14	(< 0.001)	-11.06	(< 0.001)	-10.86	(< 0.001)	-10.36	(< 0.001)
	$\tau = 15$	MSE	-13.84	(< 0.001)	-13.74	(< 0.001)	-13.49	(< 0.001)	-12.79	(< 0.001)
		QLIKE	-12.10	(< 0.001)	-12.10	(< 0.001)	-11.80	(< 0.001)	-11.16	(< 0.001)
	$\tau = 25$	MSE	-5.91	(< 0.001)	-5.95	(< 0.001)	-5.74	(< 0.001)	-5.41	(< 0.001)
		QLIKE	-12.18	(< 0.001)	-12.13	(< 0.001)	-11.67	(< 0.001)	-10.72	(0.001)
	$\tau = 35$	MSE	-16.25	(< 0.001)	-16.38	(< 0.001)	-15.84	(< 0.001)	-14.73	(< 0.001)
		QLIKE	-13.87	(< 0.001)	-13.87	(< 0.001)	-13.38	(< 0.001)	-12.11	(< 0.001)
	$\tau = 45$	MSE	-14.11	(< 0.001)	-14.06	(< 0.001)	-13.72	(< 0.001)	-13.22	(< 0.001)
		QLIKE	-13.34	(< 0.001)	-13.19	(< 0.001)	-12.81	(< 0.001)	-11.92	(< 0.001)

using nearest neighbor truncation. *Journal of Econometrics*, 169(1):75–93.

Ardia, D., Bluteau, K., and Ruede, M. (2019). Regime changes in Bitcoin GARCH volatility dynamics. *Finance Research Letters*, 29:266–271.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *OR*, 20(4):451–468.

Bejaoui, A. and Karaa, A. (2016). Revisiting the bull and bear markets notions in the Tunisian stock market: New evidence from multi-state duration-dependence Markov-switching models. *Economic Modelling*, 59:529–545.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts. *Journal of Business & Economic Statistics*, 6(1):105–111.

Durland, J. M. and McCurdy, T. H. (1994). Duration-dependent transitions in a Markov model of U.S. GNP growth. *Journal of Business & Economic Statistics*, 12(3):279–288.

Granger, C. W. J. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204.

Haas, M., Mittnik, S., and Paolella, M. S. (2004). A new approach to Markov-Switching GARCH models. *Journal of Financial Econometrics*, 2(4):493–530.

Hamilton, J. (1994). *Time series analysis.* Princeton University Press, Princeton, New Jersey.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

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

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Lam, P.-s. (2004). A Markov-switching model of GNP growth with duration dependence. *International Economic Review*, 45(1):175–204.

Maheu, J. M. and McCurdy, T. H. (2000a). Identifying Bull and Bear markets in stock returns. *Journal of Business & Economic Statistics*, 18(1):100–112.

Maheu, J. M. and McCurdy, T. H. (2000b). Volatility dynamics under duration-dependent mixing. *Journal of Empirical Finance*, 7(3):345–372.

McAleer, M. and Medeiros, M. C. (2008). Realized volatility: A review. *Econometric Reviews*, 27(1-3):10–45.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Patton, A. J. and Sheppard, K. (2009). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, 25(2):218–238. Forecasting Returns and Risk in Financial Markets using Linear and Nonlinear Models.

Pelagatti, M. (2007). *Duration Dependent Markov-Switching Vector Autoregression: Properties, Bayesian Inference, Software and Application.* In: Business Fluctuations and Cycles, Science Publishers, Hauppauge NY.

Segnon, M. and Bekiros, S. (2020). Forecasting volatility in bitcoin market. *Ann Finance*, 16(3):435–462.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Timmermann, A. (2006). Chapter 4 forecast combinations. volume 1 of *Handbook of Economic Forecasting*, pages 135–196. Elsevier.

Appendix A: maximum likelihood estimation of DDMS models

This appendix briefly describes the estimation procedure of the DDMS model as outlined in Maheu and McCurdy (2000a). Consider the following specification for the DDMS model

$$Y_t = \mu_0(1 - S_t) + \mu_1 S_t + \sigma Z_t, \quad Z_t \sim N(0,1), \tag{10}$$

where σ is the standard deviation parameter and Z_t is assumed to follow an identically and independently normal distribution. Observe that in (10) we have $\mu(0) = \mu_0$ and $\mu(1) = \mu_1$. The DDMS model can be viewed as an extension of Hamilton's model (Hamilton, 1989), as a new latent variable, $\mathcal{S}_t$, encompasses all possibilities of historical trajectories from S_t to τ:

$$\begin{aligned}
\mathcal{S}_t = 1, &\quad \text{if} \quad S_t = 1, S_{t-1} = 0, S_{t-2}, \cdots, D(S_t) = 1,\\
\mathcal{S}_t = 2, &\quad \text{if} \quad S_t = 1, S_{t-1} = 1, S_{t-2}, \cdots, D(S_t) = 2,\\
\mathcal{S}_t = 3, &\quad \text{if} \quad S_t = 0, S_{t-1} = 1, S_{t-2}, \cdots, D(S_t) = 1,\\
&\quad \vdots\\
\mathcal{S}_t = N, &\quad \text{if} \quad S_t = 0, S_{t-1} = 0, S_{t-2} = 0, \cdots, D(S_t) = \tau,
\end{aligned}$$

where $N := 2+2(\tau-1)$ are the extended states. Let $\boldsymbol{\xi}_t := \big(I(\mathcal{S}_t = 1), \cdots, I(\mathcal{S}_t = N)\big)'$, where I denotes the indicator function. Let $\mathcal{P}$ be the $N \times N$ transition matrix associated to $\mathcal{S}_t$, whose (i,j)-th entry is given by $[P]_{i,j} = P(\mathcal{S}_t = j|\mathcal{S}_{t-1} = i)$. Let $\mathscr{F}_t := \sigma(Y_t, Y_{t-1}, \cdots, S_t, S_{t-1}, \cdots)$ denote the information available at time t and let $f(Y_t|\mathcal{S}_t = j, \mathscr{F}_{t-1})$ be the conditional density of Y_t. Let

$$\boldsymbol{\eta}_t := \big(f(Y_t|\mathcal{S}_t = 1, \mathscr{F}_{t-1}), \cdots, f(Y_t|\mathcal{S}_t = N, \mathscr{F}_{t-1})\big)',$$

and $\hat{\boldsymbol{\xi}}_{s|r}$ be a vector for which the j-th element is $P(\mathcal{S}_s = j|\mathscr{F}_r)$. Observe that we have

$$\hat{\boldsymbol{\xi}}_{t|t} = \frac{\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t}{\mathbf{1}'_N(\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t)}, \qquad \text{and} \qquad \hat{\boldsymbol{\xi}}_{t+1|t} = \mathcal{P}\hat{\boldsymbol{\xi}}_{t|t},$$

where $\mathbf{1}_N := (1, \cdots, 1)' \in \mathbb{R}^N$. Following Hamilton (1994), we start with

$$\boldsymbol{\xi}_{0|0} = (A'A)^{-1}A'E,$$

where $A = \begin{bmatrix} I_N - \mathcal{P} \\ \mathbf{1}'_N \end{bmatrix}$ e $E = \begin{bmatrix} \mathbf{0}_N \\ 1 \end{bmatrix}$, with $\mathbf{0}_N := (0, \cdots, 0)' \in \mathbb{R}^N$ and I_N denoting the $N \times N$ identity matrix. Finally, the conditional log-likelihood function is given by

$$L(\theta) = \sum_{t=1}^{n} \log\big(f(Y_t|\mathscr{F}_{t-1})\big),$$

where

$$f(Y_t|\mathscr{F}_{t-1}) = \mathbf{1}'_N(\hat{\boldsymbol{\xi}}_{t|t-1} \odot \boldsymbol{\eta}_t).$$

Appendix B: optimization algorithm

The details of the algorithm are outlined below:

1. Find the starting values using a combination of random and grid search: Let κ be the number of parameters in the DDMS model. Define a vector $\boldsymbol{b}$ consisting of 100 evenly spaced values between 0.1 and 10 for the parameter λ. For the remaining parameters, create a $(\kappa - 1) \times 100$ matrix C with random numbers, where each row is drawn from a continuous uniform distribution. The bounds of the uniform distributions are heuristically defined and may vary depending on the particular model and dataset. The resulting draws can then be represented as $\begin{bmatrix} C' & \boldsymbol{b}_i \otimes \mathbf{1}_{100} \end{bmatrix}'$, for all $i = 1, \ldots, 100$, where $\otimes$ denotes the Hadamard (elementwise) product and $\mathbf{1}_k \in \mathbb{R}^k$ denotes a vector of ones.

2. Evaluate the likelihood function at each of the points defined by $\begin{bmatrix} C' & \boldsymbol{b}_i \otimes \mathbf{1}_{100} \end{bmatrix}'$, for all $i = 1, \cdots, 100$. Sort the results in decreasing order and store the top s values. Then proceed with the first set of parameters.

3. Let $\boldsymbol{\theta}_0 := (\theta_{0,1}, \cdots, \theta_{0,\kappa})'$ be the current vector of starting values for optimization. We look for a local maxima in the domain $[\theta_{0,1} - r, \theta_{0,1} + r] \times \cdots \times [\theta_{0,\kappa} - r, \theta_{0,\kappa} + r]$, where r can either depend on the values of $\boldsymbol{\theta}_0$ or be fixed exogenously. For simplicity, we set $r = r_1$. However, it is important to note that some parameters may have restricted parameter spaces, such as λ, which must be positive. In such cases, the bounds must satisfy the restrictions on the parameter space.

4. To guarantee the existence of unconditional probabilities, which are defined by $\pi = (A'A)^{-1}A' \begin{bmatrix} \mathbf{0}_N \\ 1 \end{bmatrix}$, where $A = \begin{bmatrix} I_N - \mathcal{P} \\ \mathbf{1}_N' \end{bmatrix}$, with $\mathcal{P}$ denoting the transition matrix for the extended states, and $\mathbf{0}_N$ the null vector in $\mathbb{R}^N$, ensure that the reciprocal condition number of the matrix $(A'A)$ is above the machine precision. In practice, a small value such as 10^{-9} is sufficient to virtually eliminate numerical issues caused by a near singular transition matrix.

5. Use a derivative based numerical optimization method to find the maximum of a nonlinear, multivariate function with bounds, as specified in 3 and nonlinear constraints as defined in 4.

6. The optimization in step 5 is considered successful if the first-order optimality measure is close to zero and the proposed solution does not approach the boundary defined in step 3. We evaluate the proximity of the proposed solution to the bounds by computing the absolute percentage difference. This percentage difference should be above a specified threshold. More specifically, let

$$\ell_i^- := \frac{|\theta_{1,i} - (\theta_{0,i} - r)|}{|\theta_{1,i}|}, \quad \text{and} \quad \ell_i^+ := \frac{|\theta_{1,i} - (\theta_{0,i} + r)|}{|\theta_{1,i}|},$$

   for $i \in \{1, \cdots, \kappa\}$, where $\theta_{1,i}$ is the proposed value for the ith parameter. We say that the proximity criteria is met for a given threshold $\delta > 0$ if $\min\{\ell_1^-, \cdots, \ell_\kappa^-, \ell_1^+, \cdots, \ell_\kappa^+\} > \delta$. Note that some parameters have restricted spaces, such as $\lambda > 0$. In such cases, it is acceptable for the estimation to be close to the restrictions, and the percentage proximity criteria should not be calculated. We apply $\delta = 0.01$ in most cases.

7. If the first-order optimality measure or the proximity criteria are not satisfied:

7.1. If the first-order optimality measure is not close to zero, repeat the optimization process by returning to step 2 and choosing the next starting value. Continue this process until the first-order optimality criterion is met.

7.2. If the first-order optimality is close to zero but the proximity criteria is not satisfied, proceed to step 3 and adjust the value of r_1 to r_2, where, $r_2 > r_1$. If this second round of optimization still fails to satisfy the criteria, return to step 3 and use a much larger value of r, $r_3 > r_2$, only for those parameters that do not meet the criteria.

7.3. Report the optimization as unsuccessful if all s stored initial values fail to simultaneously meet both the first-order optimality and the proximity criteria.[7]

[7] In this paper we set r_1=1, r_2=2 and r_3=10, respectively.