DE GRUYTER

Studies in Nonlinear Dynamics & Econometrics

# Should you use GARCH models for forecasting volatility? A comparison to GRU neural networks



| | |
|---|---|
| Journal: | *Studies in Nonlinear Dynamics & Econometrics* |
| Manuscript ID | DGSNDE.2022.0025.R2 |
| Manuscript Type: | Research Article |
| JEL-Classification: | F37 - International Finance Forecasting and Simulation < F3 - International Finance < F - International Economics, G17 - Financial Forecasting < G1 - General Financial Markets < G - Financial Economics |
| Keywords: | Volatility forecasting, GARCH, Hidden Markov Models, Markow Switching GARCH, Gated Recurrent Unit |
| | |

De Gruyter Journal YYYY; aop

**Research Article**

Alberto Pallotta* and Ciciretti Vito

# Should you use GARCH models for forecasting volatility? A comparison to GRU neural networks

https://doi.org/10.1515/sample-YYYY-XXXX
Received March 11, 2022; revised November 30, 2022; accepted October 3, 2023

**Abstract:** The GARCH model is the most used technique for forecasting conditional volatility. However the nearly integrated behaviour of the conditional variance originates from structural changes which are not accounted for by standard GARCH models. We compare the forecasting performance of the GARCH model to three regime switching models: namely, the Markov Switching GARCH, the Hidden Markov Model, and the Gated Recurrent Unit neural network. We define the number of optimal states by means of three methods: piecewise linear regression, Baum-Welch algorithm and Markov Chain Monte Carlo. Since forecasting volatility models face the bias-variance trade-off, we compare their out-of-sample forecasting performance via a walk-forward methodology. Moreover, we provide a robustness check for the results by applying k-fold cross-validation to the original time series. The Gated Recurrent Unit network is the best suited for volatility forecast, while the Hidden Markov Model is the best at discerning the market regimes.

**Keywords:** Volatility forecasting, GARCH, Hidden Markov Models, Markow Switching GARCH, Gated Recurrent Unit, walk-forward.



## 1 Introduction

The S&P 500 index is by far the most known financial instrument in the financial industry. It is defined as the weighted average of the stock prices of the 500 largest US-listed companies, weighted by their market capitalization (the product of the number of shares outstanding and dollar value of each individual share) [3]. Being a gauge of the world's largest financial market, it has historically been one of the most traded and studied financial instruments. As such, forecasting its volatility is of vital importance with respect to many widespread applications. For instance, volatility is relevant to the process of estimating and covering the market risk exposures, to allocating the correct amount of capital in each investment, to pricing option contracts, to gauge the overall sentiment of the markets, to forecasting the optimal timing of investment strategies, etc.

The main idea of this paper is to forecast the daily realized volatility of the SP&500 using a dataset of daily open prices spanning over the period of January 1990 - September 2021. The financial volatility literature has often provided evidence of structural breaks [17]. As such, we employ two different models that account for structural breaks across different markets regimes, namely the Markov Switching GARCH (MS GARCH) [24] and the Hidden Markov Model (HMM) [18]. Moreover, we add to the comparison the Gated Recurrent Unit (GRU) model [13], a neural network based model whose hidden layers and data memory specification emulate structural breaks. For baseline comparison, we use the GARCH model [10], the most widespread model for volatility forecasting. While the latter does not account for regime switches, the MS GARCH and the HMM allow different model parameters in each latent state with a switch governed by a

***Corresponding author: Alberto Pallotta,** Middlesex University, Economics, The Burroughs, London NW4 4BT. e-mail: a.pallotta@mdx.ac.uk.
**Ciciretti Vito,** Independent Researcher, Berlin, BorsigStr. 3, Germany. e-mail: vciciretti8@gmail.com.

stochastic and a deterministic process respectively. To estimate the number of optimal regimes, we consider each model's specific characteristics. In fact, for the MS GARCH we employ the Markov Chain Monte Carlo algorithm (MCMC) [4] which samples the original distribution by means of the Metropolis-Hastings algorithm [27] and is capable to reproduce the Markov Chain property of the model. For the HMM we use the Baum-Welch algorithm, the most used algorithm for calibrating this kind of models [37]. Finally, since GRU networks do not possess an explicit notion of states, to find its optimal number of states we apply the piecewise linear regression (PLR) method [38], which allows to fit different regressions for each structural break in the underlying data.

The enhanced feature of allowing regimes switches results in more complex specifications, in turn resulting in lower bias but higher estimation variance. Hence, we face the classic bias-variance trade-off [21]. As such, we focus our analysis on out-of-sample performance to inquire whether the decrease in bias offsets the increase in variance. The out-of-sample dataset comprises 20% of randomly sampled and adjacent observations of the original dataset. Sampling only adjacent observation allows us to take into consideration the serially correlated nature of the time-series data [42]. Moreover, we base the out-of-sample forecasting analysis on a walk-forward algorithm [46]. To assess the robustness of the results, we sub-sample the dataset ten times by means of k-fold cross validation [22] and run the walk-forward algorithm on each sample. Also, this allows us to run an out-of-sample backtest on heterogeneous sets of market conditions, comprising both periods of crisis and periods of low volatility. We assess the out-of-sample performance in terms of mean squared error (MSE) and mean absolute error (MAE) - where realized volatility is used as the benchmark - as well as in terms of variance regression. While the MSE and MAE help us choosing the model that performs the best in generating the most accurate volatility forecast, the variance regression is the guidance in discerning the correct market regime in each period [14].

The MSE highlights the stronger capability of the GRU (best performing) and HMM models to outperform both the GARCH and MS GARCH in the out-of-sample setting. This means that despite the higher complexity, the models also generalize out-of-sample. Moreover, also in terms of variance regression, the HMM (best performing) proves to outperform as its intercept estimate is close to 0 – hinting at an unbiased estimator - and the coefficient close to 1 – proving its consistency - (both p-values <1%).

The rest of the paper is organized as follows. Section 2 reviews the relevant volatility forecasting literature. Section 3 describes the methodology of the four models, the algorithm used to select the optimal number of market regimes, the grid search approach used for fine-tuning the models and the walk forward methodology used for the out-of-sample backtesting. Section 4 introduces the dataset and analyses the in-sample performance. Section 5 deals with the out-of-sample performance evaluation and robustness check. Section 6 concludes.

## 2 Literature review

For an excellent review of volatility forecasting methods, see [2]. Moreover, for a general review of forecasting methods in Finance, see [44] who discusses the basic financial predictability problem with peculiar focus on volatility and density forecasting. [36] offer a review of volatility forecasting in financial markets and describe several evaluation metrics.

The most frequently used models for forecasting volatility are GARCH models (1,298 articles), followed by Markov-Switching GARCH (96 articles), Hidden Markov Models (27 articles) and GRU networks (18 articles) [1]. GARCH models exploit the high persistency of the conditional variance [10]. [26] compare the out-

1 Number of papers refers to November 2022. We used a combination of the following search strings in JSTOR: risk management, volatility, GARCH, Hidden Markov Models, Markov Switching GARCH

of-sample performance of 330 ARCH-type to forecast the conditional variance of developed markets exchange rates. As a result, they found no evidence that a GARCH(1,1) is outperformed by more sophisticated models. Nevertheless, [17] and [32], among others, argue that the nearly integrated behaviour of the conditional variance may originate from structural changes in the variance process which are not accounted for by standard GARCH models. These findings indicate a potential source of misspecification, to the extent that the form of the conditional variance is relatively inflexible and held fixed throughout the entire sample period. Hence the estimates of a GARCH model may suffer from a substantial upward bias in the persistence parameter. Therefore, models in which the parameters are allowed to change over time - such as in Markov Switching or Hidden Markov specifications or model with stacked data memory as in GRU networks - may be more appropriate for modelling and forecasting volatility.

Since the seminal contributions by [25], the financial and statistical literature has witnessed many applications of Markov Switching models (MSMs). Their use has pivoted on the capability of filtering from raw data the underlying but unobservable state governing the data generating process. Among the first to use MSMs were [19], trying to explain the 1980s dramatic US dollar rise vis-a-vis the Deutsche mark, the French franc, and the British pound. [23] compare MSMs and DCC-GARCH to an international equity portfolio construction in terms of in-sample fit and implications for the dynamics of higher-order moments, such as skewness and kurtosis. [39], using daily S&P 500 returns for sample 1928-1991, applied MSMs to ten equal consecutive sub-samples to find the most appropriate number of regimes. Their bootstrapping technique indicated that two regimes are always appropriate against a single one, yet in many sub-samples, two is rejected in favour of three. [1] applied MSMs to investigate the regime-switching nature of US corporate bonds to liquidity shocks of stocks and Treasury bonds. This is an important work that proved the intuitive idea that in a regression model the slope and intercept may follow an MS specification with important implications in the light of the 2008-2009 Great Financial Crisis.

Apart from MSMs, another way to model changes over time is Hidden Markov Models. These models are very popular in fields such as molecular biology, but they haven't been much applied in financial volatility forecasting yet. Some relevant applications include [40], who apply HHM to model the S&P 500 conditional volatility arguing that the advantage over MSM is the fewer number of parameters to estimate. [34] model both the conditional mean and volatility process by means of an HMM applied to daily gold prices. The results proved that the model is sufficient to describe the dynamics of the data, though its predictive ability does not hold in the long term. [47] apply a hybrid model to forecasting the volatility of crude oil, composed of three parts: an HMM for detecting regime changes, a GARCH model for filtering the data and for producing residuals and a least square support vector machine for forecasting the residuals. The results proved that the hybrid model can significantly improve the out-of-sample forecasting accuracy when compared to traditional volatility forecasting models.

The GRU, a type of recurrent neural network, is predominantly used for natural language processing, as demonstrated by [11]. GRU holds features such as data context, thus represent an efficient methodology for text classification. As such, these features can also be useful in terms of volatility forecasting with several states or breaks. [12] showed that GRU outperforms another deep learning algorithms in forecasting CSI 300 realized volatility. Moreover, other studies showed that GRU outperforms popular GARCH models in forecasting the volatility of Bitcoin [43].

However, satisfactory in-sample performance does not guarantee similar out-of-sample results.. The reason is mostly due to data snooping bias [33], hyperparameters overfitting and unforeseen structural changes in the data generating process [9], [16]. A study by [5] illustrates this point well. Using monthly data on 3-month and 5-year rates of zero-coupon bonds from the US, the UK and Germany, they find that while MSMs do not always outperform single-regime models in terms of in-sample diagnostics, they tend to outperform in the out-of-sample setting. In fact, major efforts are needed to re-focus the volatility forecasting literature on the out-of-sample performance of forecasting models [7].

Our work contributes to the existing literature in three ways. First, we showcase a comparison between MS-GARCH, HHM and GRU models against the classic GARCH model. Second, we stress the importance of out-of-sample forecasting performance of the different models based on a walk-forward approach, which allows to clearly avoid typical backtesting mistakes, such as data snooping and survivorship bias. Third, we contribute to the volatility forecasting literature by further studying the application of GRU models focusing on their out-of-sample performance.

# 3 Methodology

The methodology is based on a four-step univariate parametric approach. First, we build the statistical specification of the return generating process. Second, we choose the optimal number of market regimes. Third, we perform in-sample model calibration. Fourth, the one-day-ahead forecast distribution of returns is simulated through a walk-forward approach. Finally, the relevant evaluation metrics are computed on this set and on samples deriving from ten-folds cross validation.



## 3.1 Model specifications

### 3.1.1 Markov switching GARCH

We define $y_t \in \mathbb{R}$ as the log-return of the financial asset at time t. In specifying the return generating process, we allow for regime switching in the conditional variance. Denote by $\mathcal{F}_t$ a filtration of events observed up to time $t-1$.

The general Markov-Switching GARCH specification is:

$$y_t \mid (S_t = k,\ \mathcal{F}_{t-1}) \sim \mathfrak{O}_k(0,\ g_{k,t},\ \xi_k)$$

where $\mathfrak{O}_k\left(0,\ g_{k,t},\ \xi_k\right)$ is a continuous distribution with mean equal to zero, a time-varying conditional variance process $g_{k,t}$ and $\xi_k$ is an additional shape parameter vector. The distribution has subscript $\mathfrak{O}_k$ to allow different specifications across the $k \in \mathbb{N}$ regimes. The discrete stochastic variable $S_t \equiv \{s_t = k,\ \ k \in \mathbb{N}\}$ describes the GARCH regime switches in a Markov-Chain fashion. The standardized innovations are defined as:

$$\eta_{k,t} \equiv \frac{y_t}{g_{k,t}}^{\frac{1}{2}} \sim iid\ \mathfrak{O}_k(0,1,\ \xi_k).$$

The dynamic of the state variable $S_t$ is assumed to be described by a first-order ergodic homogeneous Markov chain which characterizes the Markov Switching GARCH model of [24]. In this setting, the unobserved first-order ergodic homogeneous Markov chain is equipped with a square $k$-dimensional transition matrix of probabilities $\mathbf{P}$, where its single elements $p_{i,j}$ represent the probability of a transition from state $s_{t-1}{=}i$ to state $s_t{=}j$. Obviously, $0 < p_{i,j} < 1\ \forall\ i,j \in \{1,\dots,K\}$ and $\sum_{j=1}^{K} p_{i,j} = 1,\ \forall\ i \in \{1,\dots,K\}$.

The variance of $y_t$ conditional on the discrete realization $s_t = k$ is $g_{k,t}$. Such conditional variance is assumed to be described by a GARCH model. Therefore, conditional on $s_t = k$, the conditional variance $g_{k,t}$ is $g_{k,t} \equiv g(y_{t-1},\ g_{k,t-1},\ \theta_k)$, where $g(\bullet)$ is a $\mathcal{F}_{t-1}$ Borel-measurable function, $y_{t-1}$ and $g_{k,t-1}$ represent the GARCH-*like* dependence upon past observations and past variances respectively and $\theta_k$ is a regime dependent vector of parameters and allows for different scedastic specifications in different regimes. When $k \equiv 1$, a traditional (single - regime) GARCH is obtained by the same $g(\bullet)$. Finally, the initial values of the variance recursion are set equal to the unconditional variance in each $k$-th regime.

Under the assumption that the error term is *i.i.d.*, that it has a continuous density on the whole real dominium and non-negativity conditions are respected, then the MS GARCH process is geometrically ergodic and if it is initiated from a stationary distribution then the process is strictly stationary and

absolutely regular. Moreover, its moments exist strictly finite. The ergodicity property not only ensures that a unique stationary probability measure exists and that the chain converges to it at a geometric pace with respect to the total variation norm. Ergodic Markov Chains satisfy conventional limit theorems such as the central limit theorem for any given starting value.

For model estimation, Markov Switching GARCH models are estimated using the Bayesian Markov chain Monte Carlo method.

The likelihood is combined with a diffuse truncated prior $f(\Psi)$ to build the kernel of the posterior distribution f($\Psi$, $\mathcal{F}_t$). The normalizing constant of the posterior is numerically intractable, hence its unknown form was drawn with the adaptive random-walk Metropolis-Hastings sampler [45]. Additional positivity and covariance-stationarity constraints are added during the estimation.

The loss function can be seen as the opposite of the likelihood [41]:

$$Loss = -\prod_{t=1}^{T} f(y_t|\Psi, \mathcal{F}_t - 1) = -\prod_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p_{i,j} z_{i,t-1} f_{\mathfrak{D}_k}(y_t|s_t = j, \Psi, \mathcal{F}_t - 1) \quad (1)$$

where $z_{i,t-1} \equiv P[s_{t-1} = i|\Psi, \mathcal{F}_{t-1}]$. The statistical robsustness of the Monte Carlo Markov Chain (MCMC) estimates for the Markov switching GARCH model is discussed in [45].

### 3.1.2 GARCH model

The density function of a univariate GARCH model is a function of the location and scale parameters normalized to yield zero mean and unit variance $\alpha_t = (\mu_t, \sigma_t, \omega)$, where $\mu_t$ is the conditional mean, $\sigma_t$ is the conditional variance and $\omega = (\theta, x_t)$ denotes the other parameters of the distribution. The innovations z(t) are scaled by the conditional means and variance and have conditional density $g\,(z \mid \omega) = \frac{d}{dz}\, P\,(z_t < z|\,\omega)$ which relates to $f\,(y|\,\alpha)$ via:

$$f\left(y_t \mid \mu_t, \sigma_t^2, \omega\right) = \frac{1}{\sigma_t} g(z_t|\omega)$$

Under the standard GARCH specification [10], the conditional variance is specified as:

$$\sigma_t^2 = \left(\omega + \sum_{j=1}^{m} \zeta_j v_{j,t}\right) + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

where $\omega$ represents the intercept, $\varepsilon^2$ the residuals from the filtration of the conditional mean process discussed above, $v_{j,t}$ are $m$ external regressors with sensitivity $\zeta$. The GARCH order is defined by the couple of lags ($q$, $p$) and up to $m$ external regressors $v_j$. Finally, the unconditional variance is given by: $\widehat{\sigma}^2 = \frac{\widehat{\omega}}{1-\hat{P}}$, with $\hat{P}$ being the estimated persistency (given by the sum of the $\alpha$ and $\beta$ coefficients of the GARCH specification). Over the years, a large number of GARCH model specification have been proposed. [35] proposed the exponential GARCH (eGARCH) for the conditional variance process to be described by:

$$log(\sigma_t^2) = \left(\omega + \sum_{j=1}^{m} \zeta_j v_{j,t}\right) + \sum_{j=1}^{q} \alpha_j z_{t-j} + \gamma_j(|z_{t-j|} - E|z_{t-j}|) + \sum_{j=1}^{p} \beta_j log(\sigma_{t-j}^2)$$

with $\alpha_j$ capturing the sign effect, $\gamma_j$ the size effect and with the expected value of the absolute innovation $|z_t|$ being $E|z_t| = \int_{-\infty}^{\infty} |z| f(z, 0, 1, ..)\, dz$ . Nelson's GARCH accounts the so-called leverage effect. This is an empirical feature observed in the Equity and Fixed income markets which are captured by a downward looking relationship between past returns and conditional volatility. Compared to the standard linear GARCH model, positiveness is automatically ensured by the presence of the natural logarithm in the model specification.

To fit the parameters of the GARCH model, we use the Maximum Likelihood Estimation (MLE) method. Once again, its loss function can be seen as the opposite of the likelihood [41]:

$$Loss = -\prod_{t=1}^{T} f(y_t|\mu_t, {\sigma_t}^2, \omega) = -\prod_{t=1}^{T} \frac{1}{\sigma_t} g(z_t|\omega) \tag{2}$$

which should be minimized with respect to $(z_t, \omega)$. For the GARCH model, the MLE method results in estimators that are asymptotically efficient, as they achieve the Cramer-Rao lower bound in the limit.

### 3.1.3 Hidden Markov models

Hidden Markov Models are regime dependent models where the regime $S_t$ is a discrete process. As such, they represent another useful financial tool for modelling a time series with structural breaks. The dynamics of the series and the change points are determined by latent discrete Markov chain $S_t$ transition probabilities which, once again, are: $p_{i,j} \equiv P[s_t = j|s_{t-1} = i]$. Consequently, $y_t$ follows a Markov chain process similar to the Markov switching specification with the only difference that the model estimation is carried out by using MLE. In this case, the loss function - seen as the opposite of the likelihood - is given by:

$$Loss = -p_1(s_1) \prod_{t=2}^{T} p(s_t|s_{t-1}) \prod_{t=2}^{T} p(y_t|s_t) \tag{3}$$

which should be minimized with respect to $(z_t, \omega)$. For the Hidden Markov Model, the MLE method results in estimators that are consistent and asymptotically normal, as shown by [8].

### 3.1.4 Gated Recurrent Unit Neural Network

The Gated Recurrent Unit [13] is a neural network based model derived from the LSTM model of [29]. This network architecture contains updates and reset gates which are used to update the hidden state via a two-steps approach. As such, they allow for a stateful representation of the time-series resembling a regime-switching mechanism. In the first step, the update and reset gates are computed as:

$$z_t = f_{activation}(W^z y_{t-1} + U^z h_{t-1})$$

$$r_t = f_{activation}(W^r y_{t-1} + U^r h_{t-1})$$

where $z_t$ is the update gate, $W^z$ is the weight of the previous output $y_{t-1}$ on the new one $y_t$ and $h_t - 1$ represents the previous hidden layer weighted by $U^z$.

The hidden layer is then updated via:

$$h_t^{/} = f_{tanh}(W y_{t-1} + r_t \bigodot U h_{t-1})$$

and finally the output is given by:

$$h_t = z_t \bigodot h_{t-1} + (1 - z_t) \bigodot h_t^{/}.$$

The update and reset gates are used to decide how much information from the previous time-step is carried over to the next one. In particular, the reset gate decides how much of the previous hidden state is carried over, while the update gate decides the relative strength of the retained information on the new hidden state. This allows to apply a recurrent connection between layers and in turn to emulate long memory networks. In other words, GRU is a stateful neural network representation. As such, market regimes can be fitted by these memory blocks and improve the forecasting of realized volatility.

To train the Gated Recurrent Unit neural network we use the cross-entropy loss function. The inputs and the outputs are normalized by means of the minimum-maximum range function. Hence, the loss function yields to:

$$Loss = -\sum_{t=1}^{T} \log(\hat{y}_t) y_t \tag{4}$$

which should be minimized.

We choose the GRU network over the LSTM due to its lower complexity and enhanced ability to generalize out-of-sample compared to the latter [15]. Moreover, several works in the literature highlighted better performances of the GRU models compared to LSTM [30].

## 3.2 Choosing the optimal number of market regimes

The selection of the optimal number of market regimes for the MS-GARCH and the HMM models can be either empirically based or data-driven. Empirically, financial practitioners distinguish between two market regimes: a high volatility regime and a calm one. In fact, in equity and fixed income markets there is an empirically observed inverse relationship between volatility and market returns [28]. In this work, we instead use a data-driven approach based on three statistical specifications to test whether the resulting optimal number of market regimes would deviate from the empirically observed two states.

The first approach is based on a piecewise linear regression (PLR) [38]. The reason why we choose this method is that it allows to better fit the data by using multiple linear regressions. In fact, this methodology fits the data with a different line by means of an additional regression, such as:

$$y_i = \eta_i + \beta_i (x - b_i) \; \forall b_i < x < b_i + 1 \tag{5}$$

where $b_i$ and $b_i + 1$ are the $x$ location in the first and second break points respectively. As such, PLR can identify regime switches based on hard tresholdings and linear connection between states.

As second approach, we consider the Baum-Welch (BW) algorithm [37], a special case of the Expectation Maximization algorithm which is widely used to calibrate the Hidden Markov models. As such, the Baum-Welch algorithm returns the internal number of states resulting from fitting an HMM to the underlying data.

As a third approach, we consider the Markov chain Monte Carlo (MCMC) algorithm [4], a class of algorithm that allows to sample from an underlying probability distribution. Using this algorithm, it is possible to build a Markov chain with a target distribution equal to the one of the underlying data. To construct the chain, we employ the Metropolis-Hastings algorithm [27].

For fitting the parameter search via the BW and MCMC algorithms, we use the gamma distribution for the realized volatility. The gamma distribution is quite general and can fit most distributions by tweaking its inner parameters. While one option would be taking the arithmetic average of the three algorithms, we decide to consider the inner specifications of each model. As such, we consider that the PLR method is based on a regression with an underlying normal distribution assumption and linear connections, the BW is derived from a lambda distribution and MCMC from a Gamma distribution. Thus, we couple the MS GARCH with internal skewed t-student distribution to the MCMC method. On the other hand, we couple HMM - with its internal rigid states matrix - to the BW method. Finally, as GRU does not posses a notion of states in its model specification, we couple it to the PLR algorithm as the number of hidden units influences the memory ability of the network.

## 3.3 Tuning the model parameters

We perform a model parameter search with reference to the in-sample logarithmic returns using the AIC and BIC information criteria as well as the log-likelihood. The in-sample train set is composed of 80% of adjacent

observations, while the remaining 20% is unused during the fine-tuning. Given that financial data exhibit strongly persistent serial correlations, the training set should consist of adjacent observations to preserve the serial correlation of the original time series. For the GARCH model, we test three model specifications for various $(p, q)$ lags: a standard linear GARCH, the non-linear GJR GARCH and the eGARCH. Note that the GARCH(1,1) model is linear in the sense that the effect of the lagged shock on the conditional variance is symmetric, as opposed to other non-linear models that comply with the so-called leverage effect. Moreover, we test for several distributions, specifically: normal, t-student, GED, generalized hyperbolic, normal-inverse Gaussian and Gamma distribution. Furthermore, we test whether introducing skewness into the unimodal standardized distributions improves model fitting. To do so, we followed the methodology introduced by [20] by using an additional parameter $\varepsilon > 0$, where $\varepsilon = 1$ implies a symmetric distribution. [6] derived the moments of the standardized skewed distributions which are particularly needed in the estimation of the eGARCH. Finally, we test whether to include a variance targeting, wherein the long-term unconditional variance is given by: $\omega = \widehat{\sigma}^2 \left(1 - \hat{P}\right) - \sum_{j=1}^{m} \zeta_j \overline{v}_j$, where $\sigma^2$ is the unconditional variance, $\overline{v}_j$ represents the sample mean of the $j^{-th}$ external regressor in the variance equation (assuming stationarity) and $\hat{P}$ is the persistency index.

It is worth noticing that GARCH, Markov Switching GARCH and Hidden Markov model use the loss functions to calibrate the model parameters, yielding consistency and asymptotical normality. On the other hand, Gated Recurrent Unit uses the loss function only to target the fitting accuracy on an estimated set of outputs, without providing a statistical interpretation of the parameters.

### 3.4 Out-of-sample backtesting

The fine-tuned and in-sample optimized models have been compared in terms of their out-of-sample performance. In particular, we run an out-of-sample walk-forward simulation of the model performance on the out-of-sample dataset. When implementing this procedure, we cared for eliminating any look-ahead-bias, which means not taking into account any information that is not public yet at the moment of the simulated price action. The walk-forward methodology involves first calibrating the model in-sample and then making a one-step-ahead forecast based on the previously obtained calibration. The latter procedure is repeated each day with the calibration window moving by effect of the additional observations being added to the in-sample dataset. This design allows to closely mirror the real-world forecasting exercise at the same time of being in a look-ahead bias-free setting. The same methodology was applied on ten different out-of-sample datasets sampled via ten-folds cross validation. The different performance measures have then been averaged through the 10 different samples to provide robustness to the results.

## 4 Data and in-sample analysis

The dataset used is a univariate time-series of S&P 500 daily open prices for the period spanning over January 1990 - September 2021. We calibrate the models on daily logarithmic returns and evaluated them against daily realised volatility. We calculate daily logarithmic returns $r_t$ as: $r_t = \ln(\frac{P_t}{P_{t-1}})$ where $P_t$ and $P_{t-1}$ indicate the current price and the price one period before respectively.

The null hypothesis of daily logarithmic returns having zero mean is rejected under the assumption of normal distribution. Therefore, we demeaned the returns by subtracting their arithmetic vintage mean[1]. Table 1 reports the main descriptive statistics of the logarithmic returns.

Daily logarithmic returns exhibit leptokurtosis. Nevertheless, almost no excess of kurtosis is left after removing the 20-most outliers. This points to a distribution with a slower exponential decay in the tails than a normal distribution. These outliers were not dropped from the estimation, rather we simply tested the

1 By vintage mean we intend the average logarithmic return in each calendar year.

**Tab. 1:** Unconditional descriptive statistics of the daily S&P logarithmic returns.

| | | | |
|---|---|---|---|
| **Mean** | **0.037%** | **Median** | **0.079%** |
| **Daily Standard Deviation** | **1.09%** | **Sample Daily Variance** | **0.01%** |
| **Kurtosis** | **10.597** | **Skewness** | **-0.418** |
| **Maximum** | **10.96%** | **Minimum** | **-9.47%** |

decay of the excess of kurtosis. In fact, these observations are likely falling within a specific volatility regime and dropping them could skew the results. A Jarque-Bera normality test rejects the null of normality with a p-value $< 1\%$. We consider several distributions during the estimation and the parameter tuning process. The augmented Dickey-Fuller test confirms that the logarithmic return process is stationary. A Portmanteau test of the Box-Pierce type rejects the null-hypothesis of identically and independently distributed daily logarithmic returns (p-value $< 1\%$ at different lag specifications) also for squared and absolute values. Serial autocorrelation is persistent above 20% also when tested by removing the 20-th outmost outliers, showing a clear presence of volatility clustering and conditional heteroskedasticity. Given that by demeaning we obtained a zero mean return process, the realized daily conditional volatility is simply computed as squared daily returns. The long-term unconditional volatility is 13.98% in annualized terms.

The next step is determining the optimal number of market regimes $k$ by applying the PLR, BW and MCMC algorithms described above. Our calculation resulted in different optimal numbers of regimes, with $k_{PLR} = 6$, $k_{BW} = 5$ and $k_{MCMC} = 2$. The selected MS GARCH model has an eGARCH(1,1) specification with a skewed t-student distribution, with two regimes (as defined by the MCMC method) and no variance targeting. Stationarity conditions are imposed through the estimation process. Not surprisingly, the search shows that a rather simple two regimes model performs the best even when compared to higher complexity specifications. This is particularly sustained by the financial markets' empirical evidence of the existence of a calm (low-volatility) and turbulent (high-volatility) market regimes. Nevertheless, the five regimes model is best suited for the Hidden Markov Model specification (as defined by BW method), indicating how the MS GARCH needs fewer regimes thanks to its stochastic switching and ARMA(1,1) embedded specification. The number of hidden units is the main parameter of the GRU network. To define this parameter we use a two-steps grid search. First, we train several networks with the number of hidden units ranging from 5 to 100, with a step size of 1. Then, based on the resulting MSE and MAE metrics, we choose the number of hidden units minimizing the errors. In the second step, we cross-check the result with a backward search starting from 100 hidden units and decreasing to 5, with the same step size of 1. However the grid search approach is computationally expensive. For this reason, we found that - as a rule of thumb - $k_{GRU} = -27 + 16 * K_{PLR}$ approximates the optimal number of hidden units avoiding any numerical complexity induced by the grid search.

Both the MS eGARCH(1,1) and the benchmark eGARCH(1,1) models mostly have significant coefficients that reject the null hypothesis of being equal to zero (p-values $<5\%$). MS eGARCH(1,1) coefficients are much more stable than those of eGARCH(1,1). This fact is a strong argument for using MS eGARCH models with several regimes, since the standard eGARCH model adapts its coefficients for each sample. Such adapting effect resembles the regimes switching property, yet with lower significance of parameters. Regarding the HMM model, almost all parameters are significant as well as the lambda values are mostly $<1$, indicating that the residuals are not leptokurtic. However, the HMM may suffer more from the bias-variance trade-off as its 5 regimes specification is more complex than the MS eGARCH one.

Coming to residual diagnostics, the first issue is that within the Markov switching class, the standardized residuals fail to show *iid*-ness since the innovations are serially uncorrelated only within a given regime [31]. The same issue also leads to departures from normality. Rejection of the null of zero predictive power would point to the misspecification of the conditional volatility function. In our case, residuals fail to display any autoregressive structure. The HMM model suffers of the same issue. In fact, both normality and *iid*-test fail. However applying the same hypothesis testing of zero predictive power yield the same result of no

autoregressive structure. A Portmanteau test of the Box-Pierce type rejected the null hypothesis of *iid*-ness. However the hypothesis is not rejected when squared standardized residuals are Portmanteau tested, as the Autocorrelation Function (ACF) strictly falls within the 98% Bartlett's interval.

As it is visible from the Figure 1 in Appendix, all models manage to track the in-sample volatility. While the state transition is smooth for eGARCH and MS-eGARCH, HMM and GRU network produce in-sample rigid transitions when the market regime is moving from one state to another. Looking at the HMM model specification, this behaviour is expected due to the internal structure of regime transitioning based on a rigid matrix. On the other hand, this behaviour is unexpected from GRU network which does not possess any internally modelled regime switching dependency. The figure shows that GRU regime transitions appear more rigid than HMM.

## 5 Out-of-sample performance

In this section, we focus our attention on the out-of-sample performance of the four models. The models are backtested on ten sub-samples spanning the entire dataset. This backtesting period spans two periods of crises - the dot.com bubble and the 2008 Global Financial crisis - as well as several other periods of medium-low volatility. As such, an heterogeneous set of market conditions is backtested. Table 2 summarizes the out-of-sample forecasting performances of all the models in terms of MSE and MAE against the realized volatility. In each validation sample, the out-of-sample dataset comprises of 20% randomly sampled and adjacent observations of the original dataset. Sampling only adjacent observations allows us to take into consideration the serially correlated nature of the data. All models have been simulated out-of-sample via the walk forward mechanism.

**Tab. 2:** MSE and MAE (in percentage terms) resulting from the deployment of the four models on the out-of-sample datasets via the walk forward methodology across the 10 sub-samples. In boldface, we highlight the best performing model according to each evaluation metric.

| Sample | GARCH | | MS-GARCH | | HMM | | GRU | |
|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| 1 | 59.45% | 39.92% | 102.86% | 63.82% | 70.43% | 39.04% | **21.45%** | **30.08%** |
| 2 | 19.41% | 29.04% | 15.67% | 34.67% | **13.74%** | **25.29%** | 65.56% | 55.42% |
| 3 | 40.12% | 32.44% | 31.65% | 46.23% | 22.08% | 27.35% | **15.39%** | **20.79%** |
| 4 | 23.08% | 33.00% | 25.97% | 41.99% | 12.70% | **23.09%** | **12.23%** | 30.62% |
| 5 | 11.96% | 23.59% | 28.98% | 48.00% | **8.05%** | **20.00%** | 14.79% | 30.23% |
| 6 | 38.59% | 39.02% | 31.06% | 37.64% | 31.64% | 35.55% | **16.22%** | **28.50%** |
| 7 | 32.49% | 34.90% | 17.49% | 30.43% | **28.72%** | **31.58%** | 53.40% | 48.48% |
| 8 | 38.32% | 44.24% | 48.40% | 50.44% | 22.49% | 33.57% | **18.04%** | **30.92%** |
| 9 | 27.83% | 35.50% | 30.14% | 39.44% | 15.46% | **25.04%** | **13.23%** | 31.62% |
| 10 | 157.73% | 87.93% | 50.65% | 51.06% | 179.76% | 89.80% | **19.20%** | **29.77%** |
| *Average* | 44.90% | 39.96% | 38.29% | 44.37% | 40.51% | 35.03% | **24.95%** | **33.64%** |

Both the Mean Squared Error (MSE) and Mean Absolute Error (MAE) highlight the stronger out-of-sample performance of the GRU model as it scores the lowest MSE and MAE across the sub-samples. However, the standard eGARCH model outperforms MS eGARCH in terms of MAE as it records broader parameter changes resembling regime switches. It is also worth noticing the relative performance of the three models. In fact, switching from the benchmark eGARCH to the MS eGARCH would reduce the MSE by 7%, while to the GRU network would result in a 20% reduction. This strong improvement points to the enhanced capability of the GRU network to achieve better forecasts.

Moreover, the forecasting quality is also assessed in terms of variance regression by regressing the realized volatility $rv$ on the corresponding forecast $\hat{h}_{t+1|t}$ by means of the following linear regression:

$$rv_{t+1} = b_0 + b_1 \hat{h}_{t+1|t} + u_{t+1}, \qquad u \sim N(0,1)$$

where $b_0$ should be equal to 0 for an unbiased estimator and $b_1$ equal to 1 for a consistent estimator. As discussed by [2], the R-squared of the variance forecast should not be evaluated since the proposed methodology for the realized volatility is a noisy estimator.

Table 3 below reports the results of the variance regression. In this case, the HMM model outperforms the other three. In fact, it has the closest $b_0 \approx 0$ and $b_1 \approx 1$ (both p-values <1%). This implies that the GRU network is the best for volatility prediction while HMM is the best at discerning the market regimes.

**Tab. 3:** Results of the variance regression for the three models. An intercept terms equal to zero implies an unbiased estimator, while a slope coefficient equal to one implies a consistent one. We also report the standard error of both the coefficients. In boldface, we highlight the best performing model according to each metric.

| | GARCH | MS GARCH | HMM | GRU |
|---|---|---|---|---|
| Intercept | -0.001 | -0.001 | **-0.0006** | 0.003 |
| *s.e.* | (0.0002) | (0.0003) | (0.0002) | (0.0001) |
| Slope | **1.054** | 0.998 | **1.054** | 0.677 |
| *s.e.* | (0.017) | (0.06) | (0.019) | (0.013) |



Finally, Figure 2 in Appendix offers visual evidence of how the forecast values provided by the four models compare to realized volatility in the out-of-sample setting. The eGARCH model produces forecasts which tend to follow one-step-behind values, reacting to new market news with a one-day lag. On the other hand, the Markov switching eGARCH and the Hidden Markov models appear to better react to market news, with the latter being less influenced by stronger market fluctuations but still exhibiting its characteristic rigidity. Finally, the forecasts of the Gated Recurrent Unit model seem rather rigid.

# 6 Conclusion and future work

The aim of this paper is to compare the out-of-sample realized volatility forecasting ability of four models. The input dataset is composed of daily, open prices for the SP&500 spanning 30 years. We tested the workhorse of volatility modelling - the GARCH model - against the Markov Switching GARCH (MS GARCH), the Hidden Markov Model (HMM) and the Gated Recurrent Unit (GRU) network. The model choice is driven by the usage of regime-switching specifications against the non-switching classic GARCH model. Moreover, the GRU is used to test the performance of modern artificial intellingence techniques. The out-of-sample performance of the four models has been compared in terms of MSE, MAE and variance regression. The same test was performed on ten validation sub-samples composed of 20% adjacent observations. Finally, the out-of-sample forecasting has been obtained by means of a walk-forward algorithm. In fact, not much literature has focused on the out-of-sample performance of forecasting models.

Both MSE and MAE demonstrate the superior efficiency of GRU networks in modelling realized volatility. In terms of variance regression, Hidden Markov Model is the best performing as its intercept was close to 0 (unbiased estimator) and the coefficient close to 1 (consistent estimator) (both p-values <1%). This allows us to conclude that, to maximize the out-of-sample performance in volatility forecasting, the GRU network should be the preferred model, while the HMM is the best suited for discerning market regimes.

Future work should focus on extending this study to other asset classes or markets, especially Emerging Markets which often exhibit liquidity driven jumps. Finally, future research could also focus on predicting the probability of landing in a market regime and hence create stateful portfolio management models for each predicted regime.

**Funding:** The authors of this manuscript report no conflict of interest with an entity - with a financial or non financial interest - in the subject matter or materials discussed in this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

# References

[1] V. V. Acharya, Y. Amihud, and S. T. Bharath, Liquidity risk of corporate bond returns: conditional approach, *Journal of financial economics* **110** (2013), 358–386.

[2] T. G. Andersen, T. Bollerslev, P. F. Christoffersen, and F. X. Diebold, Volatility and correlation forecasting, *Handbook of economic forecasting* **1** (2006), 777–878.

[3] R. C. Anderson and D. M. Reeb, Board composition: Balancing family influence in S&P 500 firms, *Administrative Science Quarterly* **49** (2004), 209–237.

[4] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, An introduction to MCMC for machine learning, *Machine learning* **50** (2003), 5–43.

[5] A. Ang and G. Bekaert, Regime switches in interest rates, *Journal of Business & Economic Statistics* **20** (2002), 163–182.

[6] D. Ardia, *Financial Risk Management with Bayesian Estimation of GARCH Models Theory and Applications*, Springer, 2008.

[7] D. H. Bailey, J. Borwein, M. Lopez de Prado, and Q. J. Zhu, Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance, *Notices of the American Mathematical Society* **61** (2014), 458–471.

[8] P. J. Bickel, Y. Ritov, and T. Ryden, Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models, *The Annals of Statistics* **26** (1998), 1614–1635.

[9] G. Boero and E. Marrocu, The performance of non-linear exchange rate models: a forecasting comparison, *Journal of Forecasting* **21** (2002), 513–542.

[10] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31** (1986), 307–327.

[11] S. Cascianelli, G. Costante, T. A. Ciarfuglia, P. Valigi, and M. L. Fravolini, Full-GRU natural language video description for service robotics applications, *IEEE Robotics and Automation Letters* **3** (2018), 841–848.

[12] X.-m. Chen, S.-f. Ji, Y.-h. Liu, X.-m. Xue, J. Xu, Z.-h. Gu, S.-l. Deng, C.-d. Liu, H. Wang, Y.-m. Chang, and others, Ginsenoside Rd ameliorates auditory cortex injury associated with military aviation noise-induced hearing loss by activating SIRT1/PGC-1$\alpha$ signaling pathway, *Frontiers in Physiology* **11** (2020), 788.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, *arXiv preprint arXiv:1406.1078* (2014).

[14] R. Christensen, *Analysis of variance, design, and regression: applied statistical methods*, CRC Press, 1996.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014).

[16] R. Dacco and S. Satchell, Why do regime-switching models forecast so badly?, *Journal of forecasting* **18** (1999), 1–16.

[17] F. X. Dieobold, Modeling the persistence of conditional variances: A comment, *Econometric Reviews* **5** (1986), 51–56.

[18] S. R. Eddy, What is a hidden Markov model?, *Nature biotechnology* **22** (2004), 1315–1316.

[19] C. Engel and J. D. Hamilton, Long swings in the dollar: Are they in the data and do markets know it?, *The American Economic Review* (1990), 689–713.

[20] C. Fernandez and M. FJ Steel, On Bayesian modeling of fat tails and skewness, *Journal of the American Statistical Association* **93** (1998), 359–371.

[21] Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Vol. 2. New York: springer, 2009.

[22] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, *Statistics and Computing* **21** (2011), 137–146.

[23] M. Guidolin and G. Nicodano, Small caps in international equity portfolios: the effects of variance risk, *Annals of Finance* **5** (2009), 15–48.

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

[25] J. D. Hamilton, Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates, *Journal of Economic Dynamics and Control* **12** (1988), 385–423.

[26] P. R. Hansen and A. Lunde, A forecast comparison of volatility models: does anything beat a GARCH(1, 1)?, *Journal of Applied Econometrics* **20** (2005), 873–889.

[27] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, (1970).

[28] R. A. Haugen, A. C. MacKinlay, and W. N. Torous, The effect of volatility changes on the level of stock prices and subsequent expected returns, *The Journal of Finance* **46** (1991), 985–1007.

[29] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation* **9** (1997), 1735–1780.

[30] R. Jozefowicz, W. Zaremba, and I. Sutskever, An empirical exploration of recurrent network architectures, *International conference on machine learning* (2015), 2342–2350.

[31] H.-M. Krolzig, The markov-switching vector autoregressive model, *Markov-Switching Vector Autoregressions* (1997), 6–28.

[32] C. G. Lamoureux and W. D. Lastrapes, Heteroskedasticity in stock return data: Volume versus GARCH effects, *The Journal of Finance* **45** (1990), 221–229.

[33] A. W. Lo and A. C. MacKinlay, Data-snooping biases in tests of financial asset pricing models, *The Review of Financial Studies* **3** (1990), 431–467.

[34] R. S. Mamon, C. Erlwein, and R. B. Gopaluni, Adaptive signal processing of asset price dynamics with predictability analysis, *Information Sciences* **178** (2008), 203–219.

[35] D. B. Nelson, Conditional heteroskedasticity in asset returns: A new approach, *Econometrica: Journal of the Econometric Society* (1991), 347–370.

[36] S.-h. Poon and C. WJ Granger, Forecasting volatility in financial markets: a review, *Journal of Economic Literature* (2003).

[37] L. Rabiner, First hand: the Hidden Markov Model, *IEEE Global History* (2013).

[38] H. P. Ritzema, Frequency and Regression Analysis, Publication 16, *International Institute for Land Reclamation and Improvement, ILRI, Wageningen* (1994), 175–224.

[39] T. Rydén, T. Teräsvirta, and S. Åsbrink, Stylized facts of daily return series and the hidden Markov model, *Journal of applied econometrics* **13** (1998), 217–244.

[40] A. Rossi and G. M. Gallo, Volatility estimation via Hidden Markov Models, *Journal of Empirical Finance* **13** (2006), 203–230.

[41] R. J. Rossi, *Mathematical statistics: an introduction to likelihood based inference*, John Wiley & Sons, 2018.

[42] M. Sewell, Characterization of financial time series, *Rn* **11** (2011), 01.

[43] Z. Shen, Q. Wan, and D. J. Leatham, Bitcoin Return Volatility Forecasting: A Comparative Study of GARCH Model and Machine Learning Model, (2019).

[44] A. Timmermann, Forecasting methods in finance, *Annual Review of Financial Economics* **10** (2018), 449–479.

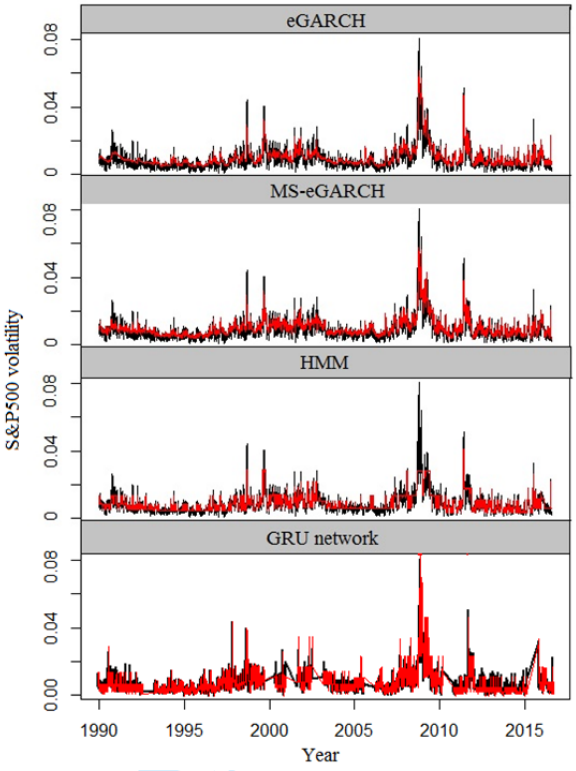
[45] M. Vihola, Robust adaptive Metropolis algorithm with coerced acceptance rate, *Statistics and Computing* **22** (2012), 997–1008.

[46] K. Żbikowski, Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy, *Expert Systems with Applications* **42** (2015), 1797–1805.

[47] Y.-J. Zhang and J.-L. Zhang, Volatility forecasting of crude oil market: A new hybrid method, *Journal of Forecasting* **37** (2018), 781–789.

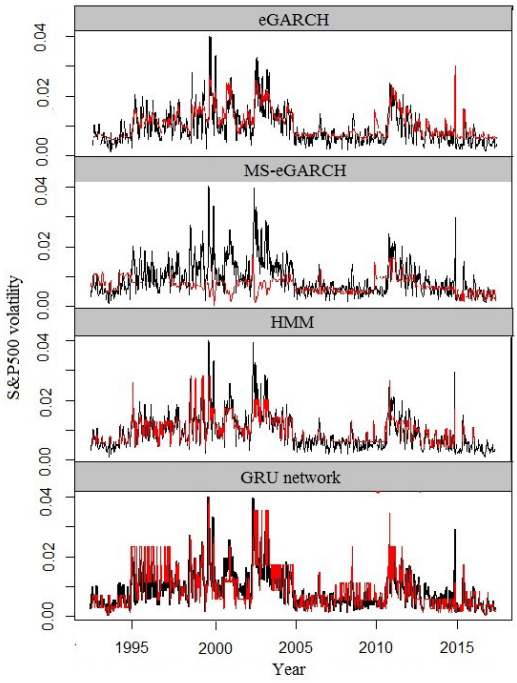
# A In-sample and out-of-sample volatility

**Fig. 1:** Realized volatility vs in-sample fitted conditional volatility for each model




Note: All the methodologies manage to track the in-sample volatility thanks to the regime switching specifications. However, the regime transition is smooth in the case of eGARCH and MS GARCH, while it appears rigid in the case of HMM (due to its rigid transition matrix specification) and GRU (less expected behaviour since regime switches are not directly modelled).

**Fig. 2:** Realized volatility vs out-of-sample forecast conditional volatility for each model




Note: All the methodologies manage to track the out-of-sample volatility. However, the GARCH model appears to react to market news with one-day lag. On the other hand, the Hidden Markov model produces well timed forecasts which are robust to stronger market fluctuations.

**For the Editor-in-Chief**

Dear Prof. Bruce Mizrach,

Thank you very much for conditionally accepting our manuscript and for the opportunity to submit a final version of our paper *"Should you use GARCH models for forecasting volatility? A comparison to GRU neural networks"*. We believe the suggestions and comments from the referees have significantly improved and clarified the final manuscript.

We have tried to incorporate all the actions concerning the final submission, including the usage of your proprietary LaTex template. We have also taken the opportunity to improve some other parts and correct some misprints. If there are aspects that we may not have fully understood, we are happy to revisit them.

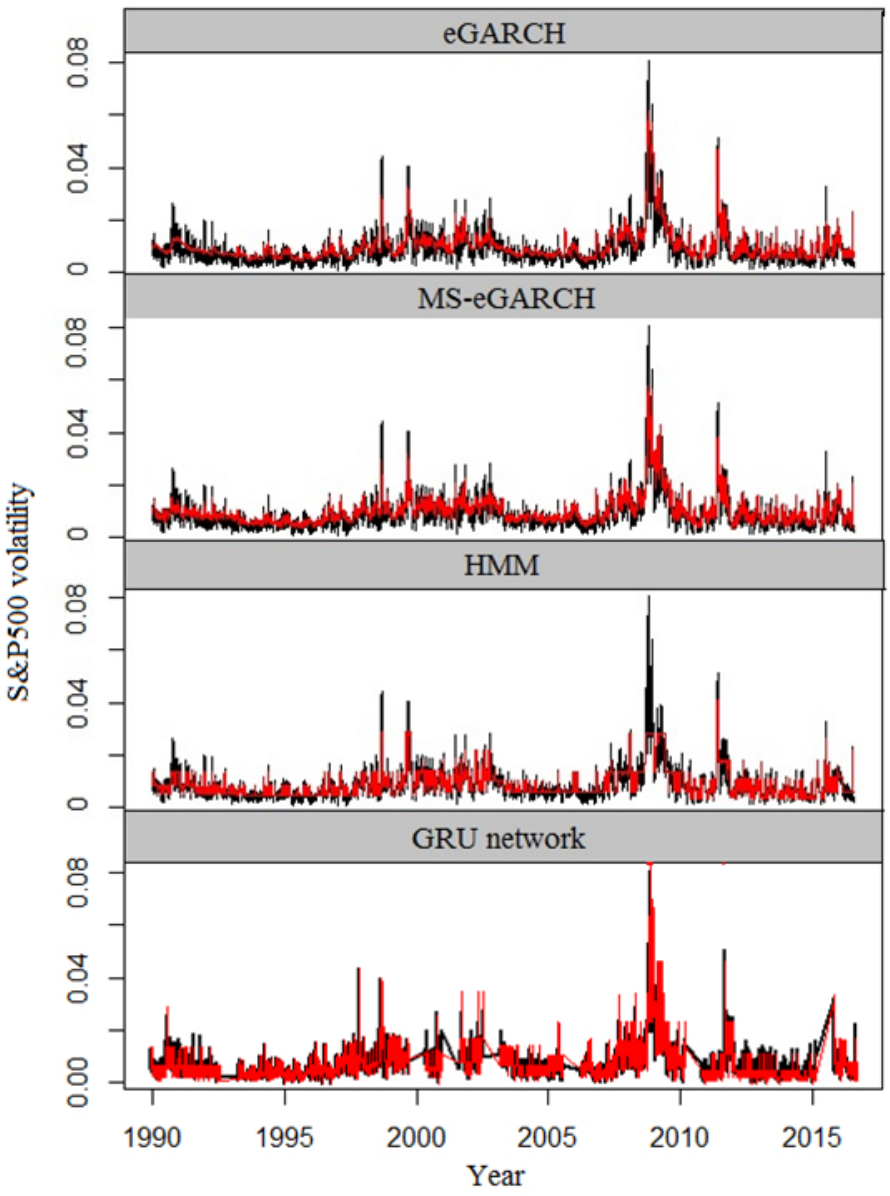
We have uploaded the following documents:

- A copy of the final version of the manuscript.
- The dataset and the code in a .zip folder, along with a readme.txt file explaining how to use them. The code also includes some tests using other datasets that we used as a control for our methodology.

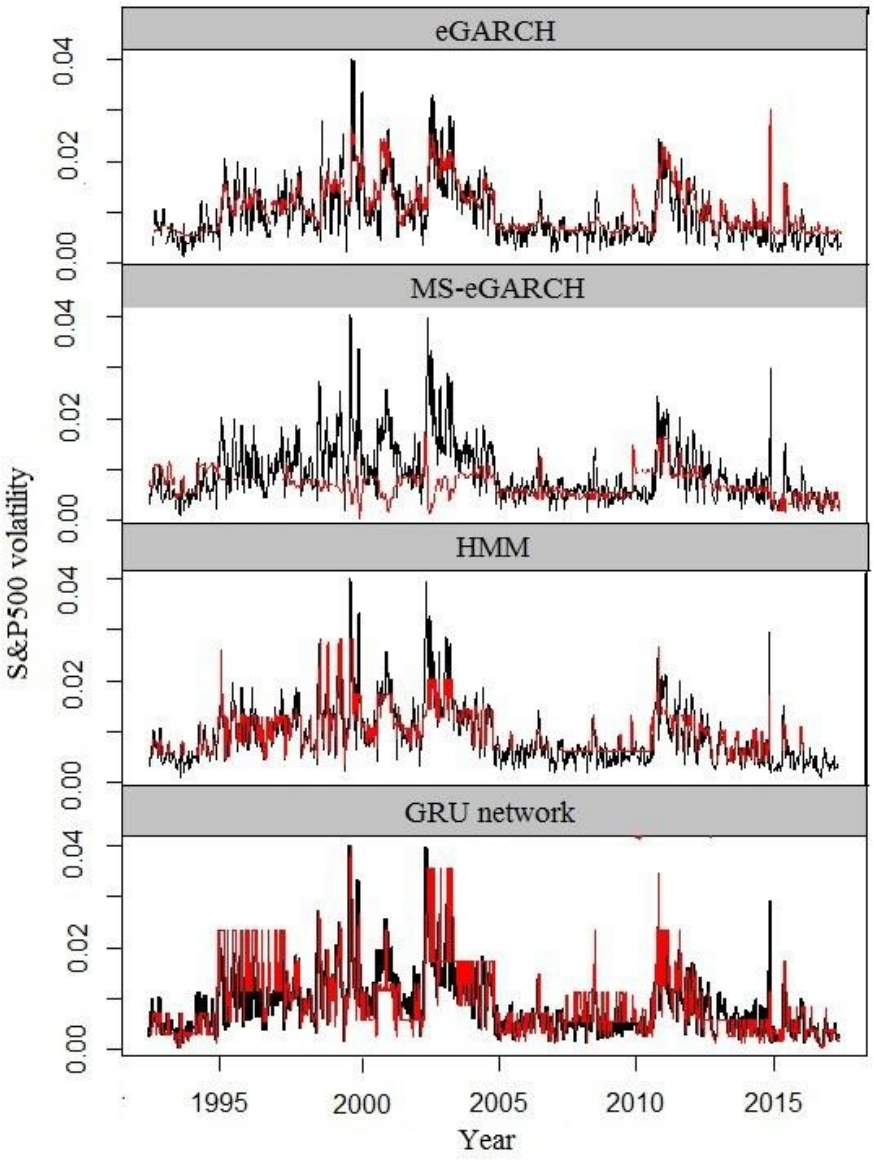
We look forward to hearing soon about the next step for publication.

# 1 Response to reviewer's comments

**Reviewer 1:** Ready for publication.
**Answer:** We sincerely thank you for your comments that have led to an improved final version of the manuscript.




192x254mm (72 x 72 DPI)




216x258mm (72 x 72 DPI)