Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics

A comparison of joint market risk estimation procedures at different holding periods

Koen Vleugels (616504)

Supervisor:	Andreas Pick
Second assessor:	Chen Zhou
Company supervisor:	Remco van der Molen
Date final version:	November 9, 2023

Abstract

Banks are mandated to gauge market risk using internal models for Value-at-Risk and Expected Shortfall. Broadly speaking, the models fall into three categories: unconditional, conditional and quantile. Using these categories and also including forecast combinations, we address two questions. First, we investigate which category of models provides 'adequate' joint risk estimation at both a 1- and 5-day holding period. Second, we test whether including time-varying components for the link between VaR and ES enhances quantile models. A novel score-driven extension is included to incorporate higher moments.

Using returns data from Morgan Stanley and HSBC in combination with rigorous testing and the Model Confidence Set, we aim to find an empirical answer. Findings highlight the underperformance of the unconditional models, even though it is used by roughly 75% of the banks under ECB supervision. Moreover, the potential of the forecast combination and quantile models is highlighted. The forecast combination using all models appears to be a stable risk estimator, not over- or underestimating the risk measures significantly. The time-varying component of the quantile models does not appear to provide improved results. Altogether, this thesis stresses the importance of more prudent model choices in risk evaluation.

Contents

1 Introduction **1**

2 Risk measures **2**
2.1 VaR and ES . . . 2

3 Models **4**
3.1 Unconditional models . . . 4
3.2 Conditional models . . . 6
3.2.1 Deterministic models . . . 7
3.2.2 Stochastic models . . . 9
3.2.3 EVT . . . 10
3.3 Quantile models . . . 11
3.3.1 CAViaR-ES . . . 11
3.3.2 MIDAS . . . 12
3.3.3 Time-varying extension . . . 13
3.4 Forecast combinations . . . 14
3.5 Model overview . . . 15

4 Model evaluation & selection **16**
4.1 Traditional backtesting . . . 16
4.1.1 Model testing - VaR . . . 16
4.1.2 Model testing - ES . . . 18
4.2 Model selection . . . 19

5 Data **20**

6 Results **23**
6.1 Per model category . . . 23
6.1.1 Unconditional . . . 23
6.1.2 Conditional . . . 27
6.1.3 Quantile . . . 29
6.1.4 Forecast combinations . . . 32
6.2 Performance comparison . . . 35

7 Sensitivity Analysis **37**
7.1 S&P500 . . . 38
7.2 Different unconditional size . . . 39

8 Conclusion **40**

A Data **47**

B Results **48**
B.1 Parameters 48
B.2 Conditional models volatility 50
B.3 Forecast combinations 51

1 Introduction

To measure Value-at-Risk (VaR) and Expected Shortfall (ES) for regulatory purposes, banks are allowed to create an internal model (BIS, 2019). In general, there are three categories to choose from, namely the unconditional, conditional and quantile models[1]. While VaR or ES models within a category have been compared previously, it may be more worthwhile from a regulatory perspective to include a variety of models from different categories in one paper to compare and contrast (Ardia et al., 2018). This can help with finding which (category of) models to exclude from usage due to undesirable characteristics such as consistent underestimation of risk. To our knowledge, a limited amount of papers have compared a wide variety of models producing VaR or ES forecasts, while even fewer have focused on comparing the models based on both risk measures (Righi and Ceretta, 2015; Steen et al., 2015; Trucíos and Taylor, 2022). Additionally, these papers focus on 1-day holding periods, while Basel rules require longer holding periods for estimations as banks may not be able to directly liquidate their positions, especially in distressed markets (BIS, 2019). Therefore, one of the questions this thesis will examine is:

RQ1 Looking at performance, which (group of) models are 'most capable'[2] of jointly estimating the 95% VaR and ES both for the 1-day and 5-day horizon, based on trading book data?

To assess the performance of the categories, it is valuable to include different kinds of models to capture the diversity of behaviour possible within a category. For example, one could use a different distribution within a model or incorporate time-varying elements. Focusing on time-varying elements, unconditional and conditional models either impose or simulate a distribution to estimate VaR and ES simultaneously. As a result, the relation between their VaR and ES estimates can be time-varying. Quantile models, on the other hand, let ES be dependent on the estimate of VaR. In a simple case, the ES is equal to the VaR estimate multiplied by a non-time-varying parameter (Le, 2020; Taylor, 2019). For such a case, the constant parameter implies ES will exactly be 'x' times as large as VaR irrespective of the period. Assuming this non-variability can be a significant assumption to pose. Therefore, we investigate the following question as well:

RQ2 Does applying time-varying elements to the relation between VaR and ES for quantile models increase explanatory power and improve results?

In addition to the models from each of the three aforementioned categories, forecast combinations are included in the performance comparison. One is added per category and one for all models. Regarding data, this thesis will use returns data from Morgan Stanley (MS) and HSBC Bank Plc (HSBC) to accurately reflect the patterns that occur in a trading book. Given the estimates of each of the models for both time horizons and datasets, several tests and a Model Confidence Set (MCS) using the joint elicitability

[1] See sections 3.1-3.3 for an explanation of the categories.
[2] Most capable referring to the joint elicitability approach.

approach will be used to come to a robust conclusion on the performance of the categories and individual models.

The results show the following patterns. First, the unconditional models appear to lack adaptability to changing market conditions. As a result, they are mostly excluded from the confidence set and perform generally poorly on the tests. This suggests that these models might not adequately capture VaR and ES, which has potential implications for the banking sector as a rough 75% of them employ unconditional models. The conditional models are underrepresented in the confidence set for lower confidence levels. This indicates that quantile and FC categories might be preferred to conditional models. A reason for this may be that conditional models rely on estimating the current volatility. If the volatility estimates are imprecise, this could lead to inaccurate forecasts. The quantile and FC categories perform well for most datasets. FC_{all} demonstrates strong performance, which is likely due to its ability to use a variety of models to capture current information. While the method appears to do well on this data, it uses a significant number of models. This might limit its practicality, especially for banks with small trading books. An alternative may be quantile models, which are frequently included in the confidence set. In particular, the non-time-varying setup for both the CAViaR-ES and MIDAS models is chosen more often as a top performer in the quantile category. Related to that, this research found the effect of adding a time-varying component to be limited, both performance-wise as well as when looking at the coefficients.

In conclusion, the study emphasizes the inadequacy of unconditional models and highlights the potential of FC, and possibly quantile, models. It underscores the importance of selecting appropriate models for accurate risk measure estimation in financial contexts. Further research is encouraged to refine these findings and explore their implications for risk management and regulatory decisions.

2 Risk measures

2.1 VaR and ES

Value-at-Risk can be mathematically defined as

$$VaR^{\alpha}(X) = \inf\{x \in \mathbb{R} : F_X(x) > \alpha\}, \tag{1}$$

where X defines a return distribution and α defines the respective quantile. At times, a minus is put in front to make the 'loss' a positive value. However, as we will see later, VaR can be positive. Hence, we refrain from using this common notation.

One of the flaws of VaR is that it is not a 'coherent' risk measure. A coherent risk measure satisfies four properties, namely monotonicity, sub-additivity, positive homogeneity, and translation invariance (Artzner et al., 1999). VaR violates the sub-additivity[3] property if the distribution of returns is not elliptically distributed. Thus, the measure is incoherent. As a result, portfolio diversification may be discouraged

[3]A risk measure is sub-additive when the risk of the total position is less than or equal to the sum of the risk of individual portfolios, implying that diversification leads to risk reduction

when only using the VaR measure. Next to that, VaR does not look at the tail risk itself. Rather, it looks at the boundary of tail risk. This results in two additional weaknesses. First, if a tail risk event occurs, an investor remains in the dark as to what the expected loss is. Second to that, two portfolios may have the same VaR, but different 'tail risks'. A fabricated example is shown in figure 1, where the average loss beyond the VaR is different for the two distributions. Expected Shortfall was introduced to the Basel legislature as a possible remedy by the Basel committee and can be defined as

$$ES^{\alpha}(X) = \mathbb{E}[X|X < VaR^{\alpha}(X)] = \alpha^{-1} \int_0^{\alpha} VaR^{s}(X)ds. \tag{2}$$

Next to being a coherent risk measure, ES may be able to take into account the distribution of the tail better than VaR as it measures the average loss beyond the VaR boundary. The method is, however, not a perfect solution. One of the caveats of ES is the larger estimation error compared to VaR (Yamai and Yoshiba, 2005). If the distribution of returns has fat tails, the probability of infrequent and large loss is high, resulting in larger errors, implying the approach is less robust. Additionally, ES is not elicitable[4], while VaR is (Gneiting, 2011). This means that models for ES cannot be directly compared, only approximately.

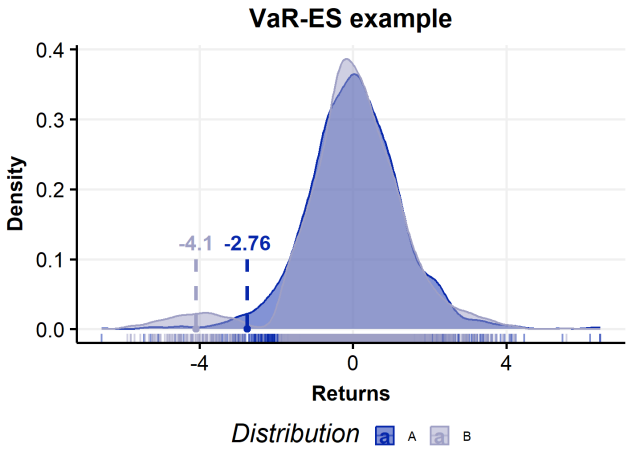


Figure 1: Two artificial distributions with a similar VaR, but different ES.

Overall, the previous arguments show that both risk measures have their strengths and weaknesses. Therefore, it may be wiser to assess the models through both metrics. A recent paper by Fissler and Ziegel (2016) showed that ES is jointly elicitable with VaR, working around the non-elicitability of ES[5]. Hence, we include both VaR and ES as risk measures.

[4]A risk measure that is elicitable means it can be acquired by minimizing the expectation of a forecasting objective function. This is considered advantageous since the forecasting objective function can be used to assess various methods

[5]More on this approach in section 4.1

3 Models

The following section introduces the models with which joint VaR-ES estimation will be done. For this paper, we take daily returns x_t from HSBC and MS, see section 5, to estimate the 1- and 5-day risk measures. Additionally, when we use VaR_t or ES_t, we imply the estimate of the risk measure at time $t-h$ for the period $t-h+1, ..., t$ for $\alpha = 0.05$.

3.1 Unconditional models

The first category is unconditional in the sense that no filter is put on observations. Based on historical observations of returns, a distribution is found. This can be done either in a parametric manner by imposing a distribution, using extreme-value-theory (EVT) or through a non-parametric take by using kernel density estimation. Based on the estimated distribution, VaR and ES can be calculated. A frequently used name for this method is historical simulation (HS). HS is used by a large share of financial institutions as it has the advantage of being simple to implement (Pérignon and Smith, 2010).

The unconditional models will be split up into a parametric, non-parametric, and EVT approach. We use the last $N = 252$ observations[6] of the returns, as at least one year of observations is required according to Basel standards and as it is recommended by Humphreys (2006). Starting with the parametric approach, we use historical data in combination with a pre-defined distribution to calculate risk measures. A stylized fact of financial time series is the leptokurtic behaviour. Therefore, next to the normal distribution, the student's t-distribution will be used. Where the normal distribution depends on μ and σ, the Student's t-distribution depends on the degrees of freedom ν as well. Thus, assuming the distribution F to be $\Phi_{\mu,\sigma}$ or $T_{\mu,\sigma,\nu}$, we define the parametric historical simulation estimator as

$$VaR_t = F^{-1}(\alpha), \tag{3}$$

$$ES_t = \alpha^{-1} \int_0^{\alpha} VaR_t ds. \tag{4}$$

However, given that the interest in risk management is the tail, imposing a structure based on all data may seem restrictive or unnecessary. EVT models may be the remedy, focusing solely on the tails. Assuming i.i.d. observations, Pickands III (1975) has shown that excesses $Y = X - u \sim GPD(\xi, \beta)$, where u is set as the threshold variable, β is a scale parameter and ξ is the tail parameter. Initially, the choice of u may seem important. Setting u too high can lead to a large variance due to limited observations, and setting it too low can lead to a bias as observations may not follow the GPD (Christoffersen, 2011). Additionally, the parameters fluctuate significantly depending on the chosen threshold

[6]The number of observations taken into account can be of great importance. Too few can lead to high uncertainty in the estimates. Too many may imply using observations that are not as realistic due to changes in financial markets. Therefore, we will also use $N = 504$ observations in the sensitivity analysis.

value. For example, Benito et al. (2023) find that the scale parameter increases by 2233% when moving the threshold from the 80th to 90th percentile for S&P500 data. However, changes in parameters or increased bias due to threshold value changes may not significantly change risk estimates. Benito et al. (2023) provide empirical support for this hypothesis by finding that VaR and ES estimations are practically equivalent using a wide variety of thresholds. Hence, we set the threshold u to the 85th percentile[7].

Returning to the estimation, the GPD is defined as

$$GPD_{\beta,\xi}(x) = \begin{cases} 1 - exp(1 + \frac{\xi x}{\beta})^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - exp(-\frac{x}{\beta}), & \text{if } \xi = 0. \end{cases} \tag{5}$$

For $\xi > 0$ the distribution has heavy tails. For $\xi = 0$ the tail decreases exponentially, like a normal distribution, and for $\xi < 0$ the distribution has a finite right endpoint. With the distribution defined, we can turn to the estimation of the risk measures. Given the threshold u, we can estimate VaR and ES through the following equations

$$VaR_t = u + \frac{\hat{\beta}}{\hat{\xi}}\Big(\big(\frac{N}{N_u}(1-\alpha)\big)^{-\hat{\xi}} - 1\Big), \tag{6}$$

$$ES_t = \frac{VaR_t}{1-\hat{\xi}} + \frac{\hat{\beta} - \hat{\xi}u}{1-\hat{\xi}}, \tag{7}$$

where N_u is the number of excesses beyond the threshold[8].

Last, it is possible to impose no structure on the returns through a non-parametric approach. There are several advantages the non-parametric approach has over the parametric approach and the EVT approach (Chen, 2008). First, there is the avoidance of misspecification. The focus of the risk metrics is on the tail of the distribution, which has a limited number of observations. Hence, it is hard to specify a proper parametric model. Moreover, it is well known that financial time series have shown tail dependence (Schmidt and Stadtmüller, 2006). The non-parametric approach has the advantage of being able to capture this, while an EVT approach assumes independence. Finally, a strength is that the probability of a more negative return may increase, such as in figure 1, which a non-parametric approach can identify, while e.g. a normal distribution cannot. The only requirement for the non-parametric method is the i.i.d. assumption.

Even though the non-parametric approach may have its theoretical advantages, research has shown that this effect is not found empirically. Using kernel smoothing, Chen (2008) shows that the non-parametric take introduced a bias, while not reducing the variance compared to a parametric approach, leading to a higher mean squared error (MSE). A possible reason for the unexpected result is that the unconditional ES functions as a mean parameter, which can be estimated with great accuracy through simple averaging. Nonetheless, the model is included due to its theoretical differences.

[7] Note that by setting this threshold, we have to transform the returns through $x_t^{new} = -x_t$ for the EVT approach. As a result, the negative of the VaR and ES estimates will need to be used.

[8] $\frac{N_u}{N}$ is the empirical estimate of $\overline{F}(u)$. By using the empirical estimator we are implicitly assuming that there is a sufficient proportion of sample values above the threshold u for reliable estimation.

Defining E' as the empirical, smoothed, distribution, we get

$$VaR_t = E'^{-1}(\alpha), \tag{8}$$

$$ES_t = (N\alpha)^{-1} \sum_{i=1}^{N} x_i \mathbb{I}\{x_i \leq VaR_t\}, \tag{9}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The focus of this research is on longer holding periods, with the minimum holding period being 5 trading days so that investors are able to liquidate their positions due to price changes. Unconditional models provide us with a distribution of 1-day returns, assuming we have daily data. A simple approximation to gain $h-$day forecasts for this is multiplying the VaR value by the square root of h days, e.g. the square root of 5 days. However, this is only valid if the data is i.i.d., which is a large assumption to make. Nevertheless, because regulatory bodies approve the measure, it is used in practice. In the case of ES, there is an approximation rule for log returns (Hull and White, 2014). However, given the data shown in section 5, this approximation cannot be constructed. Hence, we use the same square root rule to make sure the relation between the risk measures and the dataset is not altered significantly.

Having gone over all unconditional approaches, there are some broader issues that have not been covered yet. A large issue is the assumption of i.i.d. observations. Many stylized facts of financial time series, such as volatility clustering and asymmetric return behaviour, cause the assumption to break down. Next to that, as noted by Pritsker (2006), HS is highly sensitive to the estimation window, causing fluctuations in estimations. For example, the method is under-responsive to movements in conditional risk, which poses a serious problem. Finally, there is no way to extrapolate the 1-day forecast to a 5-day forecast rigorously. Thus, for this research, the unconditional methods will be used as the benchmark.

3.2 Conditional models

To combat the absence of a model and deal with volatility clustering, filtered historical simulation (FHS) was introduced (Barone-Adesi et al., 1999). Conditional models add an extra step by filtering the observations. Observations are filtered through volatility estimates, usually combined with an autoregressive model to include the conditional mean. The purpose of this additional step is to create an i.i.d. sample to draw from. To illustrate the filtering process, let us introduce a simple AR(1)-GARCH(1,1) model first

$$x_t = \nu + \mu x_{t-1} + \varepsilon_t, \tag{10}$$

$$\varepsilon_t = e_t \sigma_t, \quad e_t \sim N(0,1) \tag{11}$$

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{12}$$

We require α to be greater than zero and $\beta \geq 0$ to ensure positivity. Additionally, we set $\omega = 0$ to give more weight to recent data. For stationarity $\alpha + \beta < 1$ and $|\mu| < 1$. ν

is a constant, while σ_t and ε_t respectively denote the volatility and residual. Using the volatility estimates $\{\hat{\sigma}_t\}$ and residuals $\{\hat{\varepsilon}_t\}$ under the specified model, the first step of FHS is to standardize the residuals through

$$\hat{e_t} = \frac{\hat{\varepsilon}_t}{\hat{\sigma}_t}. \tag{13}$$

Assuming the model specification is correct, the standardized residuals are i.i.d. and thus suitable for HS, but in a different manner than seen in the previous section. Where in section 3.1 HS meant direct calculation of VaR and ES through an estimated distribution, FHS in this case takes a different route.

- From the standardized residuals $\{\hat{e}_1, ..., \hat{e}_s\}$ with length T, draw with replacement a random vector $\{e^*_{s+1}, ..., e^*_{s+h}\}$, where h is the number of days to look ahead.
- Set up initial values $\sigma^*_s = \hat{\sigma}_s$ and $\varepsilon^*_s = \hat{\varepsilon}_s$.
- For $t = s+1, s+2, ..., s+h$, plug the initial values into the following equations to obtain estimates for σ^2_t, ε_t and x_t,

$$\sigma^{2*}_t = \hat{\alpha}\varepsilon^{2*}_{t-1} + \hat{\beta}\sigma^{2*}_{t-1}, \tag{14}$$
$$\varepsilon^*_t = e^*_t \sigma^*_t, \tag{15}$$
$$x^*_t = \hat{\nu} + \hat{\mu} x^*_{t-1} + \varepsilon^*_t. \tag{16}$$

- Repeat this operation Q times, giving the bootstrapped returns $\{x^{1*}_t, ..., x^{Q*}_t\}$ for $t = s+1, s+2, ..., s+h$.
- Using these daily returns, we can get a vector of $h-$day returns $X^* = (x^{1*}, ..., x^{Q*})$ through

$$x^{j*} = \sum_{t=s+1}^{s+h} x^{j*}_t, \quad \text{for } j = 1, ..., Q. \tag{17}$$

Using the bootstrapped predicted returns for the simulated distribution E', we can calculate the VaR and ES through equation 8 and 9 respectively. For accurate estimates, we can use deterministic or stochastic modelling. Regarding the deterministic approach, a simple GARCH is a good starting point, but more elaborate models may be deemed more appropriate.

3.2.1 Deterministic models

An issue of the standard GARCH is that it may not be able to capture all stylized facts. There are plenty of GARCH variations, such as exponential GARCH (EGARCH) and Glosten-Jagannathan-Runkle GARCH (GJR-GARCH), each imposing a different structure to include more information. This, in combination with different error distributions, may solve issues of leverage, asymmetry and fat-tails in financial time series (Almeida

and Hotta, 2014; Komunjer, 2007). Zhu and Galbraith (2011), for example, compare more general models, where stylized facts may occur, to stringent symmetric models. Their findings indicate that the more general models outperform stricter models. Similar results are found by Komunjer (2007). To include fat-tails and asymmetry one may append equation 11 to

$$\varepsilon_t = e_t \sigma_t, \quad e_t \sim t(0, 1, \nu, \xi), \tag{18}$$

creating a t-GARCH model with skewness included through ξ.

Chen et al. (2012) show that next to accounting for fat-tails, adding a leverage effect into the equation may also be beneficial. The leverage effect occurs when negative returns have a greater effect on future volatility than positive returns. There are various GARCH alterations that take into account this effect. We opt to take the well-known GJR-GARCH approach,

$$\sigma_t^2 = \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 + \mathbb{I}\{\varepsilon_{t-1}\} \phi \varepsilon_{t-1}^2 \tag{19}$$

where $\mathbb{I}\{\varepsilon_{t-1}\} = 0$ if $\varepsilon_{t-1} \geq 0$ and $\mathbb{I}\{\varepsilon_{t-1}\} = 1$ if $\varepsilon_{t-1} < 0$. If $\phi > 0$ it is an indication that the leverage effect occurs. For positivity $\alpha > 0$ and $\beta, \phi \geq 0$. For covariance-stationarity $\alpha + \beta + \phi \mathbb{E}[\varepsilon_t^2 \mathbb{I}\{\varepsilon_t < 0\}] < 1$.

The previous method appends the regular GARCH structure but does not look into time-varying parameters, even though this might be beneficial. One could expect based on prior research that volatility behaves differently throughout time, for example in crisis versus non-crisis periods (Bauwens et al., 2014; Lamoureux and Lastrapes, 1990). One option to include time-varying parameters is through a regime-switching model such as Markov-switching GARCH (MS-GARCH). A thorough study done by Ardia et al. (2018) showed that MS-GARCH models acquire better results than single-regime GARCH models for daily, weekly, and ten-day equity returns.

Allowing the process to be regime-switching, the MS-GARCH process can be described as

$$\varepsilon_t | (s_t = k, \mathcal{I}_{t-1}) \sim \mathcal{D}_k(0, \sigma_{k,t}, \boldsymbol{\xi}_k), \tag{20}$$

where $\mathcal{I}_{t-1}$ denotes the information set up to $t-1$ and ε_t are the residuals of the AR(1) dynamics. $\mathcal{D}_k(0, \sigma_{k,t}^2, \xi_k)$ is a continuous distribution with zero mean, variance $\sigma_{k,t}^2$ and additional shape parameters $\boldsymbol{\xi}_k$. s_t is the integer value giving the current state on the discrete space $\{1, ..., K\}$. It evolves through a first-order Markov chain with $K \times K$ transition probability matrix P

$$\mathbf{P} \equiv \begin{bmatrix} p_{1,1} & \cdots & p_{1,K} \\ \vdots & \ddots & \vdots \\ p_{K,1} & \cdots & p_{K,K} \end{bmatrix}. \tag{21}$$

$p_{i,j}$ is defined as the probability of a transition from state $s_{t-1} = i$ to $s_t = j$, $P(s_t = j | s_{t-1} = i)$. The variance of the return x_t is thus conditional on the realization of the state s_t, $\mathbb{E}[\varepsilon_t^2 | s_t = k, \mathcal{I}_{t-1}] = \sigma_{k,t}^2$.

Taking K states, we can rewrite the GARCH models given in equation 12 to

$$\sigma^2_{k,t} = \alpha_k \varepsilon^2_{t-1} + \beta_k \sigma^2_{k,t-1}, \tag{22}$$

$$\tag{23}$$

for $k = 1, ..., K$. The same conditions as stated before hold, but now for each set of parameters per state k. See Ardia et al. (2019) for the estimation procedure, either through Bayesian Markov Chain Monte Carlo (MCMC) or Maximum Likelihood Estimation (MLE), and further details.

3.2.2 Stochastic models

A different approach to estimating volatility for FHS is using discrete time stochastic volatility (SV) models. Rather than imposing a structure, these models imply that volatility is randomly distributed. Looking at the theoretical properties, SV is able to capture stylized facts similar to the GARCH model and its variations (Pederzoli, 2006). A theoretical advantage is that SV contains a contemporaneous innovation in the variance equation, thus being able to include random new information. A disadvantage is the estimation procedure as the model is relatively more complex, also with respect to ES (Grabchak and Christou, 2021). Empirically, Carnero et al. (2001) find that SV models are better at capturing some aspects, such as excess kurtosis and high persistence of volatility, while also being less dependent on the choice for the distribution of the returns. With respect to risk measures specifically, both Grabchak and Christou (2021) and Chen et al. (2012), however, conclude that SV shows inferior outcomes compared to a GARCH approach. Note though, that most of the aforementioned papers use the canonical model autoregressive SV (ARSV).

Another option is to include fractionals. Mandelbrot and Taleb (2010) discuss that the use of fractals is essential to capture wild randomness[9], as opposed to mild randomness which can be captured by bell curves alike. It is argued that man-made variables such as returns exhibit the wild randomness property. Multifractality implies that the behaviour of volatility is a function of the time interval it operates on (Cont, 2001). GARCH models are not able to handle this, but multifractal models can. While not directly applied to ES, multifractal models showed promising results on VaR compared to GARCH models on financial data (Batten et al., 2014; Lee et al., 2016; Liu and Lux, 2008; Wei et al., 2013).

One such multifractal model is the Markov-Switching Multifractal (MSM). The model assumes a multiplicative and hierarchical structure of volatility with varying durations. It is able to create outliers, has a long memory, and can decompose volatility into components with differing decay rates (Calvet and Fisher, 2004). MSM has the

[9]Wild randomness can be seen as an environment in which a single observation can significantly influence calculations (Mandelbrot, 2013). See Mandelbrot and Taleb (2010) for examples.

following structure to decompose the volatility

$$\varepsilon_t = \sigma_t z_t, \quad \sigma_t = \sigma \left(\prod_{k=1}^{K} M_{k,t} \right)^{\frac{1}{2}}, \tag{24}$$

where σ is a positive constant and $\{z_t\}$ are i.i.d. with a standard normal distribution. Putting the K volatility components $M_{k,t}$ in a vector $M_t = (M_{1,t}, ..., M_{K,t})$, the time-varying volatility can be rewritten to $\sigma_t = \sigma[h(M_t)]^{\frac{1}{2}}$ with $h(M_t) = \prod_{k=1}^{K} M_{k,t}$. Thus, it is clear that the volatility σ_t is driven by changes in the vector M_t.

However, $M_{k,t}$ is not defined yet. Calvet and Fisher (2004) suggest the process to be defined as a first-order Markov process. Assuming the vector M_t is known, each component $M_{k,t+1}$ can be drawn in the following manner

$$M_{k,t+1} = \begin{cases} M_{k,t}, & \text{with probability } 1 - \gamma_k, \\ D, & \text{with probability } \gamma_k, \end{cases} \tag{25}$$

where each of the components $M_{k,t}$ adheres to the following conditions: $M_{k,t} \geq 0$ and $\mathbb{E}(M_{k,t}) = 1$. To satisfy these conditions, we impose a Bernoulli distribution on D. That is,

$$D = \begin{cases} m_0, & \text{with probability } \frac{1}{2}, \\ 2 - m_0, & \text{with probability } \frac{1}{2}, \end{cases} \tag{26}$$

where m_0 is an unrestricted parameter to be estimated. Additionally, the transition probabilities $\gamma \equiv (\gamma_1, ..., \gamma_K)$ are related through the following equation

$$\gamma_k = 1 - (1 - \gamma_1)^{b^{k-1}}, \tag{27}$$

where $\gamma_k \in (0, 1)$ and $b > 1$. The use of this equation is twofold. First, it makes sure the model is parsimonious as adding components does not result in more variables to be estimated. Additionally, the probability is increasing in k implying that components with a higher k switch relatively faster to a new value. Therefore, the higher k is, the lower the memory of the component $M_{k,t}$ is.

Altogether, the parameters to be estimated are $\theta = (m_0, \sigma, b, \gamma_1)$, highlighting the parsimony of the model. As it is a Markov-switching variant, the parameters can be estimated through MLE with Hamilton filtering (Calvet and Fisher, 2004). For this research, we set the number of components $K = 5$. For FHS, the process is identical to the GARCH models apart from equation 14, which should be replaced by equation 24.

3.2.3 EVT

The EVT model seen in the unconditional section assumes that extreme values are realized from i.i.d. samples. However, we have seen that the observations need to be made i.i.d. first through filtering. Adding this step has shown to be beneficial

(Allen et al., 2013; Chinhamu et al., 2015; Paul and Sharma, 2017). We follow the proposed algorithm of Danielsson and De Vries (2000) and McNeil and Frey (2000) to gain conditional EVT estimates:

- Take the standardized residuals found in equation 13 that exceed the threshold u and estimate a GPD distribution.
- Draw with replacement a random vector $\{e^*_{s+1}, ..., e^*_{s+h}\}$, where h is the number of days to look ahead.
- If one of the draws e^*_i exceeds the threshold u, replace it with a new observation $e^*_{i,new}$. Drawing an observation x from the estimated $G\hat{P}D(\hat{\xi}, \hat{\beta})$, $e^*_{i,new} = u - x$.
- Perform the same steps seen in equations 14-17 to obtain returns with which one may estimate VaR and ES.

While the EVT approach can use any of the aforementioned GARCH models, we opt to only use the EVT approach combined with the t-GARCH model to keep the number of models used in this paper reasonable.

3.3 Quantile models

3.3.1 CAViaR-ES

Opposite of the (un)conditional models, quantile models do not impose a distribution on the returns or the residuals. Rather, the dynamics of VaR are modelled directly. One example is the conditional autoregressive VaR (CAViaR) models (Engle and Manganelli, 2004). Christou and Grabchak (2022) and Taylor (2008) show it to be a competitor to benchmark models. With the benefit of not having to assume a distribution comes the cost of being unable to estimate the expected shortfall. To solve this Taylor (2019) proposes a semi-parametric approach to jointly estimate VaR and ES based on the Asymmetric Laplace distribution (ALD). Similar to CAViaR models, Meng and Taylor (2020) show that joint estimation of VaR and ES through CAViaR-ES provides competitive results. Rewriting the AL density to include the VaR and ES, we get the following density

$$f(x_t) = \frac{1-\alpha}{\mu_t - ES_t} exp\left(-\frac{(x_t - VaR_t)(\alpha - \mathbb{I}\{x_t \leq VaR_t\})}{\alpha(\mu_t - ES_t)}\right), \tag{28}$$

where VaR_t is the quantile at probability level α and ES_t is the conditional ES. μ_t is the conditional expectation of the returns, based on an AR(1) model $\mu_t = \varphi_0 + \varphi_1 x_{t-1}$. Even though Taylor (2019) avoids the use of the conditional mean as for daily returns it is approximately zero, for longer horizons this approximation may not hold. Also important to note is that no assumptions on a distribution for the returns are needed. Whilst $f(x_t)$ stems from the ALD, α is selected a priori and therefore only a quantile is estimated. ALD is merely used for its computational convenience to base the approach on.

The density can be used for MLE, but the dynamics of VaR_t and ES_t still need to be formalized. Starting with VaR, we can use the CAViaR model with symmetric effects introduced by Engle and Manganelli (2004)

$$VaR_t = \beta_0 + \beta_1|x_{t-1}| + \beta_2 VaR_{t-1}. \tag{29}$$

Having identified a relation for VaR, we turn to ES. An important restriction to take into account when modelling ES_t is that $ES_t \leq VaR_t$. Thus, a simple way to model ES is through

$$ES_t = \big(1 + exp(\gamma_0)\big)VaR_t. \tag{30}$$

The exponent makes sure that ES will be larger, in absolute terms, than VaR. γ_0 controls the joint dynamics of VaR and ES.

Similar to unconditional models, the CAViaR-ES approach gives us 1-day ahead forecasts but is unable to directly provide forecasts for longer horizons. Hence, the same transformation as for the unconditional models will be applied, even if the approach is questionable from a theoretical point of view.

For optimization, we start with 1000 randomly selected parameter sets, only restricted to adhere to the mentioned constraints. With these parameters, the likelihood can be estimated per set. Using the top three performers as starting values, the function is optimized by iterating through the Nelder-Mead and Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) until a convergence criterion is met. For simplicity, we set it such that the sum of absolute differences between the old and updated parameters needs to be larger than 0.01. Using the three optimized likelihoods, the parameter set with the best score is chosen.

3.3.2 MIDAS

Not all quantile methods struggle with direct multi-period forecasting. Le (2020) solves this by employing mixed data sampling (MIDAS). Using returns from the past D days, the following form is used

$$VaR_t = \beta_0 + \beta_1 \sum_{d=1}^{D} \phi_d(\kappa)|x_{t-d,1}|, \tag{31}$$

where $\phi_d(\cdot)$ is a polynomial that filters the returns linearly and κ defines the shape the filter has. β_0 is the constant for VaR, while β_1 contains the impact of the variations of past returns. $\theta = (\beta_0, \beta_1, \kappa)$ is the set of parameters that need to be estimated. Compared to the CAViaR-ES model, the MIDAS approach only depends on previous returns to estimate VaR, rather than also using the previous VaR estimate. Following Le (2020) and Ghysels et al. (2016), $\phi_d(\kappa)$ follows a Beta distribution defined as

$$\phi_d(\iota, \kappa) = \frac{f(\frac{d}{D}, \iota, \kappa)}{\sum_{d=1}^{D} f(\frac{d}{D}, \iota, \kappa)}, \tag{32}$$

where $f(\cdot)$ is a Beta distribution. Based on Ghysels et al. (2006) the lag length D is chosen to be 50 days, $\iota = 1$ and $\kappa > 1$. The ES can be modelled in the same way as in equation 30. Using the dynamics for VaR and ES, MLE can be applied using equation 28. For the optimization, $h-$day return time series are formed for $f(x_{t,h})$ through $x_{t,h} = \sum_{i=1}^{h} x_{t+i}$. Thus, the MIDAS approach does not require the square root rule. The optimization procedure of CAViaR-ES will be used for MIDAS as well.

3.3.3 Time-varying extension

As argued before, it can be restrictive to assume that VaR and ES have a non-changing relationship throughout time, implying that ES is always 'x' times as large as VaR. As the measures are jointly estimated, this may not only lead to biased results of ES but also of VaR. Thus, we include two time-varying options.

First, taking into account that ES and VaR do not need to have similar dynamics, the following relationship is put forward

$$ES_t = VaR_t - j_t, \tag{33}$$

$$j_t = \begin{cases} \gamma_0 + \gamma_1(VaR_{t-1} - x_{t-1}) + \gamma_2 j_{t-1}, & \text{if } x_{t-1} \leq VaR_{t-1} \\ j_{t-1}, & \text{else} \end{cases}, \tag{34}$$

where $\{\gamma_i\}$ are restricted to be non-negative to make sure that ES will be greater than VaR in absolute terms (Taylor, 2019). In this case, j_t is the parameter that relates ES to VaR, rather than γ_0 seen in equation 30. When an observation exceeds the VaR, j_t changes depending on the size of the observation. A larger deviation from VaR results in a larger j_t and therefore a larger ES.

While the prior option provides an extension that results in increased flexibility, we extend the proposed time-varying relationship by including a score-driven term. First proposed by Creal et al. (2013), score-driven models link the dynamics directly to the shape of the conditional observation density. This may be worthwhile to add when the form of the updating equation is not obvious (Artemova et al., 2022). In this case, we may suspect that an exceedance has an influence on the size of the difference between VaR and ES. Yet, it is not obvious how the influence takes shape. For example, squared terms or other forms of non-linearity could be required to correctly portray the process. Thus, we propose the following structure

$$ES_t = VaR_t - j_t, \tag{35}$$

$$j_t = \begin{cases} \gamma_0 + \gamma_1(VaR_{t-1} - x_{t-1}) + \gamma_2 j_{t-1} + \gamma_3 S_{t-1}\nabla_{t-1}, & \text{if } x_{t-1} \leq VaR_{t-1} \\ j_{t-1}. & \text{else} \end{cases}, \tag{36}$$

$$\text{where } \nabla_t = \frac{\partial log\big(f(x_t|j_t;\theta)\big)}{\partial j_t} \text{ and } S_t = \mathcal{I}_{t-1} = -\mathbb{E}\left[\frac{\partial^2 log\big(f(x_t|j_t;\theta)\big)}{\partial j_t^2}\right]^{-1} \tag{37}$$

∇_t provides the score function linking the distribution to j_t, while S_t scales the score. θ includes the other parameters. Plugging $ES_t = VaR_t - j_t$ into equation 28 and setting

$\varrho_t = j_t + \mu_t - VaR_t$, we find

$$\nabla_t = \alpha\big(\alpha - \mathbb{I}\{x_t \leq VaR_t\}\big)\frac{x_t - VaR_t}{(\alpha\varrho_t)^2} - \frac{1}{\varrho_t} \tag{38}$$

$$S_t = -\mathbb{E}\left[\frac{1}{\varrho_t^2} - 2\left(\alpha^3(\alpha - \mathbb{I}\{x_t \leq VaR_t\})\varrho_t\frac{\mathbb{I}\{x_t \leq VaR_t\}}{(\alpha\varrho_t)^4}\right)\right]^{-1} \tag{39}$$

Given the complexity of the derivatives, we opt to use the empirical information function, avoiding taking the expectation required for S_t.

3.4 Forecast combinations

While the individual models mentioned before can be useful, Taylor (2020) finds that combining forecasts is an improvement over individual forecasts. To create forecast combinations, individual forecasts are combined with weights as accuracy per forecast may differ. To estimate the weights for the FC, a score function is required. Fissler and Ziegel (2016) proved that the following score functions are consistent

$$\begin{aligned}\textbf{L1} : S\left(VaR_t, ES_t, x_t\right) = &\big(\mathbb{I}\{x_t \leq VaR_t\big) - \alpha\}G_1\left(VaR_t\right)\\ &- \mathbb{I}\{x_t \leq VaR_t\}G_1\left(x_t\right)\\ &+ G_2\left(ES_t\right)\left(ES_t - VaR_t + \mathbb{I}\{x_t \leq VaR_t\}\right.\\ &\left.\times\left(VaR_t - x_t\right)/\alpha\right) - \zeta_2\left(ES_t\right) + a\left(x_t\right),\end{aligned} \tag{40}$$

where G_1, G_2, ζ_2 and a are functions that need to have the following properties. $G_2 = \zeta_2'$, G_1 is increasing and ζ_2 is increasing and convex.

A wide variety of scoring functions can be based upon equation 40. Taylor (2019) tested these specifically for forecasting combinations, but no particular preference was found. Therefore, we opt to use $G_1(x) = 0$, $G_2(x) = -\frac{1}{x}$, $\zeta_2 = -log(-x)$ and $a = 1 - log(1-\alpha)$. Plugging the functions into equation 40 gives us the negative log-likelihood function of the ALD with time-varying location and scale. Thus, it is quite similar to the optimization function of the quantile models seen in section 3.3, which requires a constant scale instead of time-varying.

To create forecast combinations for VaR, a regular weighing option is used. To combine ES estimations, the difference between ES and VaR forecasts is combined. This is done because VaR and ES accuracy are closely related. Additionally, it makes sure that the ES forecast is larger in absolute terms (Taylor, 2019). The method, named minimum score combining, can be expressed as follows

$$\hat{Q}_{ct} = \sum_{i=1}^{M} w_i^Q \hat{Q}_{it}, \tag{41}$$

$$\hat{ES}_{ct} = \hat{Q}_{ct} + \sum_{i=1}^{M} w_i^S\left(\hat{ES}_{it} - \hat{Q}_{it}\right). \tag{42}$$

M is the number of models included. $\hat{Q}_{ct}$ and $\hat{ES}_{ct}$ are the combined forecasts for VaR and ES at time t, while $\hat{Q}_{it}$ and $\hat{ES}_{it}$ are the individual forecasts. w_i^Q and w_i^S are the combining weights for VaR and ES. Both are required to be non-negative and sum to 1. The weights can be estimated by minimizing the score function.

Next to the FC per category, we create one FC that includes all models. At the cost of being less interpretable directly, it may improve performance. Additionally, the weights may give a hint to which models are preferred, albeit less rigorously, as a high weight implies that a method is more relevant to minimizing the scoring function than other approaches.

3.5 Model overview

Including the FCs concludes the introduction of the models. For clarity, an overview of all models is given below. Each model is denoted as M_i, where M is the model and i is short for the method that is used to translate 1-day returns to $h-$day returns. In table 1 a summary of all the models is shown.

Table 1: All models

Category			
Unconditional	Conditional	Quantile	FC
$\text{Normal}_{\sqrt{h}}$	GARCH_{FHS}	$\text{CAViaR-ES}_{\sqrt{h},1}$	FC_u
$\text{Student's t}_{\sqrt{h}}$	t-GARCH_{FHS}	$\text{CAViaR-ES}_{\sqrt{h},2}$	FC_c
$\text{EVT}_{\sqrt{h}}$	GJR-GARCH_{FHS}	$\text{CAViaR-ES}_{\sqrt{h},3}$	FC_q
$\text{Non-parametric}_{\sqrt{h}}$	EVT-GARCH_{FHS^*}	$\text{MIDAS}_{M,1}$	FC_{all}
	MS-GARCH_{FHS}	$\text{MIDAS}_{M,2}$	
	MSM_{FHS^*}	$\text{MIDAS}_{M,3}$	

Note: The underscore notes which approach is used to gain $h-$day forecasts, where $h > 1$. It can be the square root rule $\mathrm{M}_{\sqrt{h}}$, FHS M_{FHS} or the MIDAS approach M_M. Because FHS is done slightly differently when using EVT or the MSM approach, it is denoted by M_{FHS^*}. The FCs require no translation, as they are optimized using the $h-$day forecasts.

Regarding the performance of the categories, Righi and Ceretta (2015) find that there is a predominance of conditional models, especially those that respect the stylized facts of financial returns. Additionally, they find that the VaR estimation is important for the ES estimation, as the models that obtain an incorrect violation rate also present low p-values for the ES. For cryptocurrency, Trucíos and Taylor (2022) observe that there are no clear superior models and that forecast combinations do not outperform individual models. However, it should be taken into account that cryptocurrency returns may not show similar behaviour to bank portfolio returns. Finally, for the commodity market, Steen et al. (2015) show that quantile regression outperforms historical simulation for VaR estimation. Altogether, we may expect conditional and quantile models to be a 'better' approach than an unconditional approach for a 1-day holding period. No research has

yet been done on how a longer time horizon will influence the performances per category. Therefore, we cannot define a hypothesis beforehand based on previous research.

4 Model evaluation & selection

4.1 Traditional backtesting

Given the variety of models, it is important to backtest to see which ones are appropriate and provide accurate results. To do so, there are two main backtesting options to include. First, the appropriateness of an individual model can be checked through model testing. Second, the models can be ranked through a loss function.

4.1.1 Model testing - VaR

Regarding VaR, one should expect an appropriate model to provide VaR estimates that are exceeded by the returns as often as defined by α. The exceedances should not occur too often or too few, as it implies underestimation or overestimation of risk respectively. Additionally, the occurrences should be independent over time. If occurrences show a pattern, for example, there are more exceedances around a period of higher volatility in the market, the model may not have been able to capture (changing) characteristics of the returns adequately.

The first test was originally introduced by Kupiec, Paul H and others (1995) and is named the unconditional coverage test. The focus is on testing if the number of exceedances is in line with the confidence level α. The test can be defined as follows

$$h_t^{\alpha} = \sum_{t=1}^{T} (x_t > VaR_t), \tag{43}$$

$$\hat{\alpha} = \frac{1}{T} h_t, \tag{44}$$

$$\mathbf{T_{unc}} : LR_{unc} = 2log\left(\left(\frac{1-\hat{\alpha}}{1-\alpha} \right)^{T-h_t} \left(\frac{\hat{\alpha}}{\alpha} \right)^{h_t} \right) \sim \chi_1^2, \tag{45}$$

where χ_1^2 denotes the chi-squared distribution with one degree of freedom. When the number of exceedances $\hat{\alpha}$ is equal to the expectation α, the test statistic has a value of 0. Increasing (or decreasing) the number of exceedances results in underestimation (or overestimation) of the risk in a portfolio. Taking the null hypothesis as

$$H_0 : \mathbb{E}(h_t) = \hat{\alpha}, \tag{46}$$

a rejection of the null implies that the model likely over- or underestimates the VaR given the data.

To test the independence of the exceedances, Christoffersen (1998) created a test based on the likelihood ratio. The intuition behind the proposed test is that if two

consecutive returns exhibit no correlation, then the probability of, for example, no exceedance the next day should be equal no matter if the previous observation was an exceedance or not. The test is based on Markov-chains. Thus, we introduce a 2x2 transition matrix based on equation 21

$$\Pi_1 = \begin{bmatrix} \pi_{0,0} & \pi_{0,1} \\ \pi_{1,0} & \pi_{1,1} \end{bmatrix}, \tag{47}$$

where $\pi_{i,j}$ is the probability of moving from state i to j. Additionally, $n_{i,j}$ is defined as the number of observations that have value i and previous value j, where $i, j = 1$ implies an exceedance and $i, j = 0$ no exceedance. Finally, we specify a hit sequence $\{I_t\}$, which indicates whether an exceedance occurred. The likelihood of the hit sequence then equals

$$L(\pi_1) := L(\pi_1; \{I_t\}) = \pi_{0,0}^{n_{0,0}} \pi_{0,1}^{n_{0,1}} \pi_{1,0}^{n_{1,0}} \pi_{1,1}^{n_{1,1}}. \tag{48}$$

This is the likelihood assuming the alternative is true. Under the null, the matrix transforms to

$$\Pi_2 = \begin{bmatrix} 1-\pi_2 & \pi_2 \\ 1-\pi_2 & \pi_2 \end{bmatrix}. \tag{49}$$

Thus, under the null, it is clear that the probability of a new observation landing in state 0 or 1 should not depend on the state of the previous observation. π_2 refers to the probability that an exceedance takes place. Therefore, $\pi_2 = \frac{n_{0,1}+n_{1,1}}{n_{0,0}+n_{0,1}+n_{1,0}+n_{1,1}}$ and the likelihood under the null equals

$$L(\pi_2) := L(\pi_2; \{I_t\}) = (1-\pi_2)^{(n_{0,0}+n_{1,0})} \pi_2^{(n_{0,1}+n_{1,1})}. \tag{50}$$

With both likelihoods specified, the LR test statistic for the independence of the observations is

$$\mathbf{T_{con}} : LR_{ind} = -2log\left(\frac{L(\Pi_1)}{L(\Pi_2)}\right) \sim \chi_1^2. \tag{51}$$

Assuming both the unconditional coverage test and the independence test find that the method is valid, the VaR estimations have a correct conditional coverage with a process defined as a martingale difference

$$\mathbb{E}[I_t | \mathcal{I}_{t-1}] = \alpha, \tag{52}$$

where $\mathcal{I}_{t-1}$ denotes the information set. While the aforementioned tests separately provide a test for unconditional coverage and independence, the dynamic quantile (DQ) approach is able to jointly test both properties. Additionally, it is able to include more lags, which $\mathbf{T_{con}}$ cannot.

Denoting the demeaned process of exceedances as $HIT_t = I_t - \alpha$, the conditional expectation given the information set at $t-1$ should be equal to zero. Next to that, it

must be uncorrelated with previous returns and other lagged variables such as previous VaR estimations. To test this, the following linear regression can be implemented

$$HIT_t = \delta + \sum_{k=1}^{K} \beta_k HIT_{t-k} + \sum_{k=1}^{K} \gamma_k g[HIT_{t-k}, HIT_{t-k-1,...,z_{t-k},z_{t-k-1},...}] + \varepsilon_t, \tag{53}$$

where ε is i.i.d. and $g(\cdot)$ is a function containing prior exceedances and other variables, such as squared returns. δ is a constant, β_k quantifies the relation to previous HIT and γ_k includes possible non-linear effects. We exclude squared returns or other forms and include 4 lags. To test for both properties, the null hypothesis is

$$H_0 : \delta = \beta_1 = ... = \beta_K = \gamma_1 = ... = \gamma_K = 0, \quad \forall k = 1, ..., K \tag{54}$$

All coefficients should be zero for unconditional coverage. For independence only, however, δ is allowed to be non-zero. Denoting $\Psi = (\delta, \beta_1, ..., \beta_K, \gamma_1, ..., \gamma_K)'$ as the vector of parameters, with Z containing the observations of equation 53, the test statistic is

$$\mathbf{T_{dq}} : DQ = \frac{\hat{\Psi}' Z' Z \hat{\Psi}}{\alpha(1-\alpha)} \sim \chi^2_{2K+1}. \tag{55}$$

4.1.2 Model testing - ES

There are several methods to compare models regarding ES, introduced by, for example, Acerbi and Szekely (2014) and Righi and Ceretta (2015). A flaw of these approaches, however, is that they require a return distribution to work with. This is not an issue for the (un)conditional models which use (part of) the distribution to compute VaR and ES. For the quantile methods, on the other hand, it poses a problem. This category specifically avoids introducing assumptions on the returns. Thus, we can only use tests in this thesis that avoid distributional assumptions.

A test to assess the quality of the ES estimates, while also avoiding distributional assumptions, has been introduced by McNeil and Frey (2000). Defining $(x, ES)^- = \{(x_t, ES_t) \mid x_t < VaR_t\}$ as the series of exceedances and corresponding ES with length N, we can find the exceedance residuals e_i as

$$\hat{e}_i = x_i^- - \hat{ES}_i^-, \quad \forall i = 1, ..., N \tag{56}$$

The exceedance residuals can be tested against the null hypothesis of a zero mean and i.i.d. behaviour, where we focus on the one-sided test to only look at the risk of underestimation as this is the likely direction of failure (McNeil and Frey, 2000). Using $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \hat{e}_i$ and $\sigma(\hat{\mu})$ as the standard error, we get the following test statistic

$$\mathbf{T_{er}} = \frac{\hat{\mu}}{\sigma(\hat{\mu})}. \tag{57}$$

To test the one-sided null hypothesis, the test statistic can be evaluated using a bootstrap test (Efron and Tibshirani, 1994).

Similar to the VaR tests, we would like to test the performance of ES not only through one test. Another backtest with a different approach, but also without distributional assumptions, is the regression-based test from Bayer and Dimitriadis (2022). Given estimates of ES, $\{\hat{ES}_t\}$, the returns can be regressed on the estimates and an intercept

$$x_t = \gamma_1 + \gamma_2 \hat{ES}_t + u_t^e, \tag{58}$$

where $ES(u_t^e|\mathcal{I}_{t-1}) = 0$ almost surely. As the forecasts are generated by the same information set, this can be rewritten to

$$ES(x_t|\mathcal{I}_{t-1}) = \gamma_1 + \gamma_2 \hat{ES}_t. \tag{59}$$

The hypothesis of correct specification of ES forecasts can be tested through

$$H_0 : (\gamma_1, \gamma_2) = (0, 1) \quad versus \quad H_A : (\gamma_1, \gamma_2) \neq (0, 1). \tag{60}$$

The parameters (γ_1, γ_2) cannot be directly estimated, as there is no strictly consistent loss function for ES. Bayer and Dimitriadis (2022) therefore employ a joint regression technique, using the joint loss function from Fissler and Ziegel (2016),

$$x_t = V_t'\beta + u_t^q, \quad \text{and} \quad x_t = W_t'\gamma + u_t^e, \tag{61}$$

where V_t and W_t are $k-$dimensional vectors. u_t^q and u_t^e denote the residuals of the regression using VaR and ES respectively. Also, $VaR(u_t^q|\mathcal{I}_{t-1}) = 0$ and $ES(u_t^e|\mathcal{I}_{t-1}) = 0$ almost surely. Using the joint regression, we can perform two backtests, the Auxiliary and Strict ESR Backtest. For the auxiliary backtest, $V_t = (1, \hat{VaR}_t)$ and $W_t = (1, \hat{ES}_t)$. For the strict backtest, $V_t = W_t = (1, \hat{ES}_t)$. In both cases the hypothesis specification seen in equation 60 is applied, giving the Wald-test statistic

$$\mathbf{T_{reg_{1,2}}} = T(\hat{\gamma} - (0, 1))\hat{\Omega}_\gamma(\hat{\gamma} - (0, 1))', \tag{62}$$

based on a covariance estimator $\hat{\Omega}_\gamma$ for the covariance of γ. $\hat{\gamma}$ contains the estimates $(\hat{\gamma}_1, \hat{\gamma}_2)$.

Note that the operations above pose no assumption on the distribution of returns. Thus, while the mathematics is written for the 1-day return case, the method will also be applicable for longer horizons. Even so, it is important to take into account that 5-day returns on day t will be closely related to 5-day returns on day $t+1$. Thus, for the 5-day holding period dataset, we create 5 separate datasets with a 5-day spacing between the observations. Because of the small dataset[10], we decided to do the tests with each separate dataset, after which the median of those results was taken for robustness.

4.2 Model selection

Having tested for the appropriateness of the models, no decision can yet be made about the 'best' category or model. Thanks to the work of Fissler and Ziegel (2016) we have

[10] See section 5.

a loss function with which models can be compared and ranked. One such loss function was mentioned earlier in equation 40. To add robustness to the model selection phase, an extra loss function is defined as

$$\begin{aligned}\textbf{L2} : S\left(VaR_t, ES_t, x_t\right) =& \alpha\left(\frac{ES_t^2}{2} + 2VaR_t^2 - VaR_t ES_t\right) + \mathbb{I}\{x_t \leq VaR_t\} \\ & \times \left(- ES_t(y_t - VaR_t) + 2(y_t^2 - VaR_t^2)\right).\end{aligned} \tag{63}$$

Using **L2**, the different frameworks will be contrasted on an individual and category level. This cannot be done rigorously by visually checking the differences in loss functions. Therefore, we will use the Model Confidence Set (MCS), proposed by Hansen et al. (2011), to find a superior set of models.

Starting with an initial set of models $\mathcal{M}_0$, the goal is a superior set $\mathcal{M}^*_{1-\alpha^*}$ at a specified confidence level. Using an equivalence test $\delta_{\mathcal{M}}$ and an elimination rule $e_{\mathcal{M}}$ at a low and high α^*, the number of initial models will be larger than the final superior set, $\mathcal{M}_0 > \mathcal{M}^*_{1-\alpha^*}$. Starting with the equivalence test, the following hypothesis is applied

$$\begin{aligned} H_{0,\mathbb{M}} &: \mathbb{E}(\Delta S_{i,j,t}) = 0, \quad \forall i, j \in \mathbb{M} \\ H_{A,\mathbb{M}} &: \mathbb{E}(\Delta S_{i,j,t}) \neq 0, \quad \text{for some } i, j \in \mathbb{M}, \end{aligned}$$

where $\Delta S_{i,j,t}$ denotes the loss difference of model i and j at time t, while $\mathbb{M} \subset \mathcal{M}_0$ holds the remaining models. The test statistic is

$$T_{\mathbb{M}} = \underset{i \in \mathbb{M}}{max} |t_{i,j}|, \quad \text{where} \tag{64}$$

$$t_{i,j} = \frac{\overline{\Delta S}_{i,j}}{\sqrt{\widehat{Var}\left(\overline{\Delta S}_{i,j}\right)}} \quad ; \quad \overline{\Delta S}_{i,j} = \frac{1}{T}\sum_{t=1}^{T} \Delta S_{i,j,t}. \tag{65}$$

$\widehat{Var}\left(\overline{\Delta S}_{i,j}\right)$ is estimated using a block-bootstrap with a block size of $l = 4$ and 5,000 trials. If the null is not rejected, the final set is $\mathcal{M}^*_{1-\alpha*}$. If it is rejected, a model will be removed from the set based on the elimination rule

$$e_{\mathcal{M}} = \underset{i \in \mathbb{M}}{\arg\max} \sup_{j \in \mathbb{M}} t_{i,j}, \tag{66}$$

where the model in $\mathbb{M}$ is eliminated if it has the highest value $t_{i,j}$. This procedure is repeated until the equivalence test is not rejected, resulting in one superior set of models $\mathcal{M}^*_{1-\alpha^*}$.

5 Data

Indices such as the S&P500 are frequently used for empirical financial papers. An advantage of using an index is that the data is readily accessible and of good quality.

There may, however, be a flaw to this approach. For example, the S&P500 aims to represent the (American) equity market, while a bank's portfolio includes debt securities, currencies and derivatives. Each of these assets may influence the day-to-day returns. Table 2 shows the risk-weighted assets (RWA) in 2022 under the standardized approach for three major European banks[11] and the largest Dutch bank, ING (BIS, 2022). While not extensive proof, it indicates that equity risk, as opposed to interest rate risk and foreign exchange risk, appears to play a smaller role in the trading book. Additionally, banks may incorporate hedges which can also result in a different behaviour. Therefore, taking an equity index may not accurately reflect the movements of a bank's portfolio.

Table 2: Risk-weighted assets per firm.

Bank	RWA		
	Interest rate risk	Equity risk	Foreign exchange risk
BNP Paribas	831		3,737
Crédit Agricole	70		497
HSBC	1,684	64	10,391
ING	10		5,332
Total	**2,595**	**64**	**19,957**

Under Pillar 3 MR4, banks are required to publish a comparison of VaR estimates with gains/losses if they use internal models to assess market risk (BIS, 2015). Using this data avoids having to mimic the portfolio of banks, which would take time and could lead to inacurrate representation, while giving accurate daily trading book profit and losses (PnL) figures.

For this research, we use the reports of Morgan Stanley and HSBC Bank Plc, giving returns from 2017-2022. There are three prime reasons for choosing these banks over others. First, the amount of the data and its quality differs per institution. Banks are only required to show a figure comparing their VaR estimates to their daily PnL. Hence, we need to extract the data from the figures[12]. Thus, some banks, such as ING, were dropped as the figures were of insufficient quality. Next to accurate data, some banks, such as ABN AMRO, have only recently started reporting this data, resulting in insufficient data points. HSBC and MS offered data of high quality and significant size. Second, the banks differ in their operations. MS is by and large known as an investment bank, not taking deposits and dealing as an intermediary between corporations and the financial markets. HSBC is, similar to the largest Dutch bank ING, a universal bank, offering both retail and investment services. Finally, whereas HSBC is under ECB supervision, the FED supervises MS. Thus, both banks have different regulations to adhere to. These differences may have an effect on their trading book and thus on their

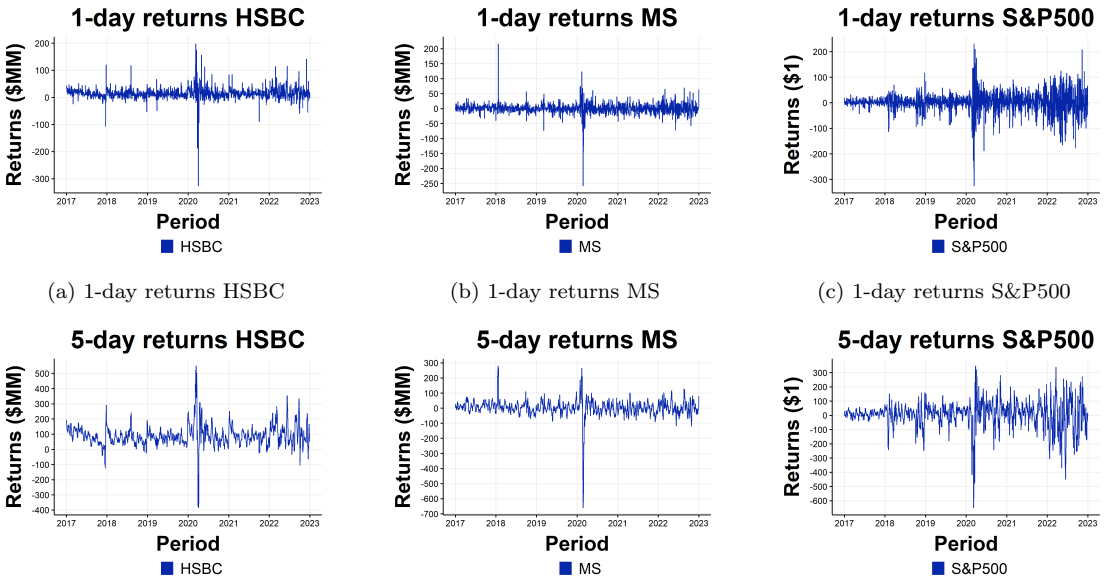
[11]Banks in the euro area are required to publicly disclose their RWA per risk category. Other, for example, American banks do not have to.

[12]We opted to use WebPlotDigitizer to extract the data.

PnL. Exposing the models to different returns helps to show the behaviour of the models under different circumstances.

Given the datasets, two choices need to be made: How long will the holding period be and what is the pre-defined probability of an exceedance? Basel Standards indicate a 10-day holding period to estimate VaR for the Fundamental Review of the Trading Book (FRTB) (EBA, 2023). Additionally, ES will be added soon for the 97.5% quantile (BIS, 2019). Thus, it seems logical to take 10 holding days with a probability of $\alpha = 0.025$. However, there is a large issue if we would proceed with this setting.

Knowing that we have a total of 6 years of data, of which one will be used to train the model initially, we are left with 5 years to backtest with. Since there are roughly 252 trading days per year, this results in around 1560 observations. Thus, using a set-up with a 10-day holding period at a probability of 0.025, we could expect roughly 4 exceedances. To accurately assess the models, this number is low. Especially for ES, as it is the average beyond VaR, more data is necessary. Therefore, we set the holding days to $h = 5$ with $\alpha = 0.05$. This results in 15-16 expected exceedances, which should be enough to roughly assess the performance, while still focusing on a longer than 1-day horizon and the tail. Given the limited exceedances, however, we should not look for one superior model for the 5-day horizon as a single model may coincidentally outperform the 'actual' superior model due to the limited size of the data. Rather, we should look at the group of superior models.



(a) 1-day returns HSBC (b) 1-day returns MS (c) 1-day returns S&P500

(d) 5-day returns HSBC (e) 5-day returns MS (f) 5-day returns S&P500

Figure 2: 1-day and 5-day returns figures of HSBC, MS and the S&P500 for the period 2017-2022.

Considering the returns in figure 2 and the time period, it is clear that each of the returns series was affected by the initial downturn of the markets that occurred around

the 'start' of the pandemic. In all cases, it is the largest downturn. Other than that, each series shows individual large negative returns that do not appear to be due to a widespread macroeconomic event. Additionally, the S&P500 appears to be increasingly volatile from 2022 onwards, more so than HSBC and MS. This may, in part, be because the price of the S&P500 has increased by 33% in that period. A larger size can result in larger absolute variations, while percentage-wise the variations remain stable. However, a transformation to log returns showed that around this period the volatility remains higher still.

Table 3: Summary statistics.

Data	Min.	Mean	Max.	Skewness	Kurtosis	% $x_t > 0$	$p_{\text{ARCH-test}}$
1-day							
HSBC	-326.16	15.10	197.51	-1.40	49.67	86.07	0
MS	-257.74	0.61	216.10	-1.01	45.86	53.55	0
S&P500	-324.89	1.02	230.38	-0.69	10.74	52.46	0
5-day							
HSBC	-384.81	75.60	549.56	0.77	13.07	96.67	0
MS	-660.03	3.38	280.10	-3.55	47.88	53.19	0
S&P500	-649.48	6.03	347.19	-1.18	8.75	62.12	0

Second, each of the time series exhibits stylized facts, such as non-normal behaviour due to fatter tails and heteroskedasticity, backed up by Engle's ARCH test seen in table 3. However, the behaviour seems to differ individually in a variety of ways. An obvious difference to notice is the percentage of positive returns. For the HSBC, this is significantly higher than that of MS and the S&P500, especially for the 5-day returns. Additionally, HSBC has higher average returns. Next to that, Appendix A shows that while the autocorrelation seen in figure 7 appears to be similar across the banks and index, the autocorrelation among the squared residuals differs per series. Compared to the banks, the index seems to still be dependent on larger lags. Finally, large negative or positive returns seem to appear less frequently for the trading book data of HSBC and MS compared to the index data of the S&P500.

6 Results

6.1 Per model category

6.1.1 Unconditional

Before we dive into the analysis of the figures and tables concerning the unconditional models, we first discuss the large amount of test results that are obtained. Table 4 shows a total of 144 tests. Thus, assuming all null hypotheses hold at a threshold of 0.05, roughly 7 are expected to be incorrectly rejected. This may lead to erroneous

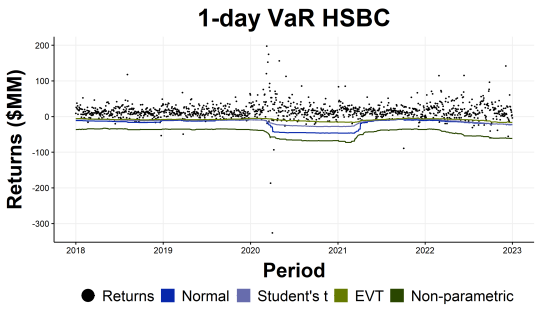
conclusions. To counter this approaches such as the Bonferroni correction or the Benjamini–Hochberg procedure can be applied (Benjamini and Hochberg, 1995). Before applying these methods to our results, we must first ask ourselves whether it is necessary to do so. We believe this is not the case for two reasons.

Let us set the following scenario. Currently, medicine X is in use. Researchers want to test if medicine Y is better, using a large range of illnesses. Even if medicine X is 'better' in reality for each case, the tests can wrongly indicate that medicine Y is better for a specific disease Z. Thus, the researchers highlight that medicine Y is better at treating disease Z. For such a scenario, applying a correction may be necessary to reduce type 1 errors. However, we will not draw conclusions from an individual test. For example, if we focus on a particular model and its performance with respect to ES forecasts, we use all 12 tests (3 tests per each of the 4 datasets) to form an opinion. While some of the tests and datasets may be related, the risk of multiple type 1 errors is lower than a single type 1 error. Hence, applying a correction may not be as worthwhile.

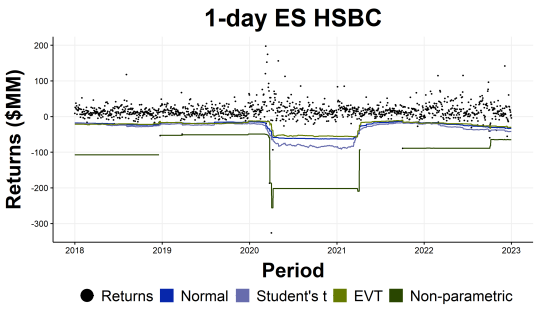
Additionally, while the corrections reduce the type 1 error, they may increase the type 2 error. When testing for a life-threatening disease, it is sensible to avoid a type 1 error as much as possible, as else the disease goes untreated. Thus, a more stringent p-value can be useful, even if it leads to diagnosing more healthy patients with a disease. In our case, a type 2 error implies that an inadequate model is not diagnosed as such. Hence, if we apply a correction, we run the risk of not finding as many deficient models. Next to that, we are in an early stage in this research. The (group of) models are not ranked definitively yet, and more broadly speaking this research provides an initial insight to build upon. Consequently, making a type 1 error compared to a type 2 error may not be as regrettable. Altogether, while we do not impose a correction, we do recognize the increased likelihood of having at least one type 1 error. Therefore, the next findings should be approached with more caution.

Moving onto the analysis, figure 3 displays the behaviour of each of the unconditional models concerning both risk measures. What can be noticed amongst all figures is the large losses around COVID-19 and the increased VaR and ES levels thereafter for about a year. The latter is likely because the unconditional models base their distribution on the last 252 observations[13]. The larger VaR and ES levels for a year can be seen as a drawback of the unconditional models, as for the period after the initial spike fewer to no observations exceed the VaR level even though we should expect around 12 to 13 exceedances ex-ante. This may contradict the independence of exceedances. The rejection of conditional coverage is partially backed up by test $\mathbf{T_{con}}$ seen in table 10. For most of the 1-day VaR estimates, the null of independence is rejected at a 5% level. For the 5-day period, the test rejects the null for all estimates of the HSBC dataset but does not reject a single null for the MS dataset. This appears to contradict the initial observations from the figures. However, it may also be due to the 5-day backtesting period resulting in a small sample with only around 50 observations the year after the COVID-19 peak. Comparing the models for the 1- and 5-day VaR of HSBC and MS, they all appear relatively stable except for the COVID-19 peak. Additionally, all models

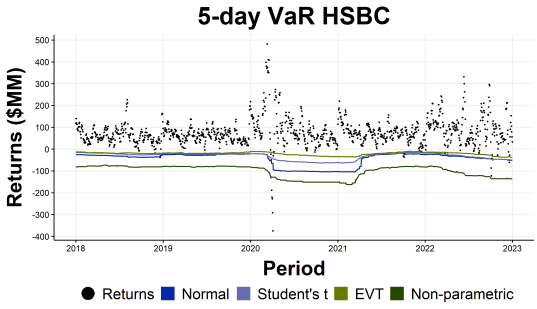
[13]See section 3.1.



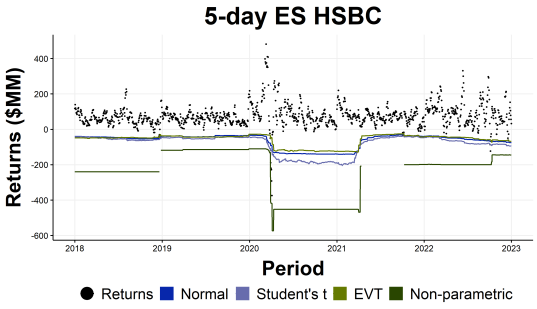
(a) 1-day VaR HSBC



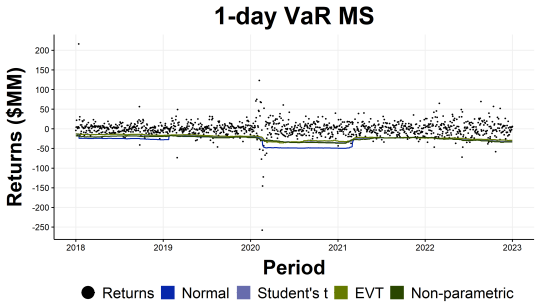
(b) 1-day ES HSBC



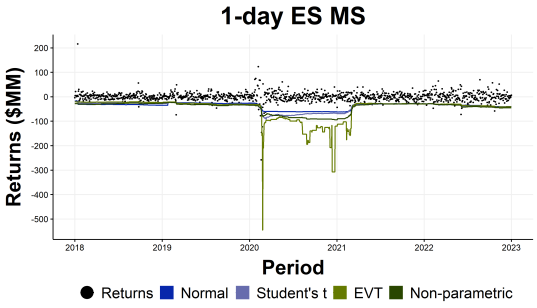
(c) 5-day VaR HSBC



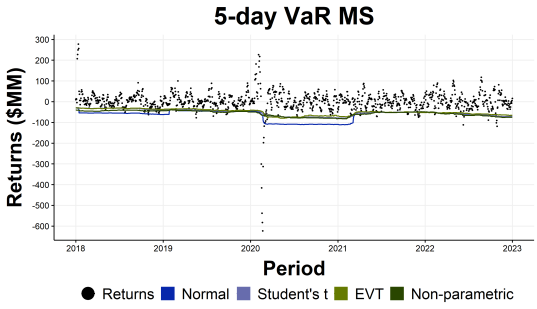
(d) 5-day ES HSBC



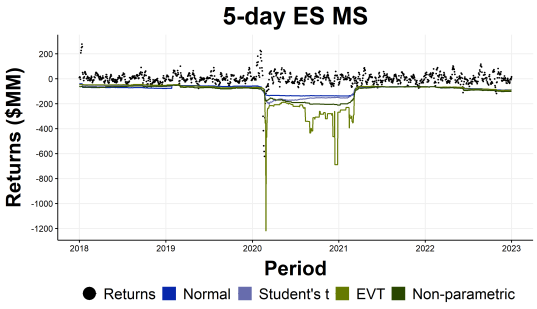
(e) 1-day VaR MS



(f) 1-day ES MS



(g) 5-day VaR MS



(h) 5-day ES MS

Figure 3: Unconditional VaR and ES for HSBC and MS.

show a gradual increase in the VaR level after 2022, with a more pronounced increase seen for HSBC. For HSBC, the EVT approach appears to have the closest hit rate for both holding periods. All other approaches overestimate the risk measure in the portfolio. For MS, there is no clear 'victor', with the Student's t being closest for the 1-day forecast, while the non-parametric approach has the closest hit rate for the 5-day forecast. Regarding the VaR tests, the results of tests $\mathbf{T_{unc,con}}$ appear to be in line with $\mathbf{T_{dq}}$ in about half of the cases. That is, when one or both of the tests reject the null, the joint null of unconditional coverage and independence is also rejected. For the 5-day forecasts, however, the tests consistently contradict each other. Again, this may be due to the limited number of exceedances. Given that $\mathbf{T_{unc,con}}$ are of low complexity, while $\mathbf{T_{dq}}$ requires a regression, we believe that the results of the former should be taken relatively more seriously, but they should definitely be interpreted with caution. The tests indicate that all the unconditional models appear inadequate for the HSBC data, which is similar to what the figures and the hit rates are showing as the models overestimate VaR and capture the peak inadequately. For MS, almost no tests $\mathbf{T_{unc,con}}$ are rejected at a 5% level. As HSBC has relatively few negative returns compared to MS, it may be more difficult for unconditional models to capture the 5% boundary of lowest returns that VaR tries to estimate for HSBC, hence the large differences.

Table 4: Median p-values of the VaR & ES tests for the unconditional models of HSBC and MS.

	HSBC							MS						
	Hit rate	VaR Tests			ES Tests			Hit rate	VaR Tests			ES Tests		
	$\alpha = \mathbf{0.05}$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$	$\alpha = \mathbf{0.05}$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$
1-day														
Normal	0.024	0	0	0	0.35	0.34	0.01	0.037	0.04	0.01	0	0.01	0.36	0.34
Student's t	0.030	0	0	0.08	0.17	0.99	0.89	0.051	0.82	0.15	0	0.28	0.84	0.61
EVT	0.051	0.87	0.01	0.63	0.34	0	0	0.057	0.26	0.01	0	0.91	0	0
Non-parametric	0.005	0	0	0	0.44	0	0	0.044	0.33	0.07	0	0.35	0.12	0.26
5-day														
Normal	0.012	0	0	0.56	0.15	0.29	0.29	0.037	0.33	0.22	0.01	0.14	0.83	0.83
T-dist	0.012	0	0	0.46	0.28	0.05	0.04	0.066	0.29	0.31	0.02	0.31	0.79	0.88
EVT	0.020	0.01	0.01	0.34	0.20	0.01	0.01	0.070	0.18	0.12	0	0.56	0.18	0.01
Non-parametric	0.004	0	0	0.21	-	0	0	0.045	0.72	0.38	0	0.26	0.78	0.78

Related to this, the overestimation of VaR for the non-parametric approach is worth mentioning as it has consequences for the estimation of ES. As can be seen in figure 3b and 3d, the ES estimates of the non-parametric approach are missing around the end of 2019 and halfway through 2021. This is because the ES is estimated by averaging all VaR exceedances, but there are no exceedances for this period. Figure 3a provides a visual backup for this statement. Note though, that the estimates are based upon the 1-day VaR and ES, as the 5-day VaR and ES are the 1-day estimates multiplied by $\sqrt{5}$. Hence, there are 5-day ES estimates (figure 3d) even though the 5-day VaR (figure 3c) shows no trespassers. In general, the ES tests seem to agree with each other, showing no rejection of the null hypothesis of correct specification. For HSBC, the correct specification of the

non-parametric ES is rejected. However, similar to the argumentation that we applied to the VaR tests, the ES test $\mathbf{T_{er}}$ for the 5-day period may provide more 'relevant' results as $\mathbf{T_{reg1,2}}$ are based on regression.

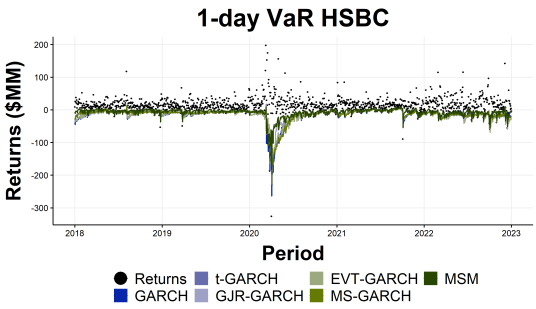
Finally, judging the figures, it appears that the EVT and non-parametric approaches provide a wider range of ES estimates with larger variations in small periods of time, while the normal and Student's t show smaller variations. This makes sense as the EVT and non-parametric take allow the tail to be estimated separately from the 'body' of the distribution, unlike the normal and student's t distribution. We also observe that while the student's t approach does not provide larger VaR estimates in general compared to the normal distribution, its ES is larger, especially for the HSBC returns. This is possibly due to the inclusion of fatter tails, which make sure that larger observations are still within a feasible realm according to the distribution.

Overall, the unconditional models seem to provide decent results, by not severely underestimating risk and providing sensible hit rates, but at the cost of inaccuracy and overestimation at times. A major point of improvement is the independence of the exceedances, which the unconditional models appear to mostly violate. The conditional and quantile models may provide a solution to this.

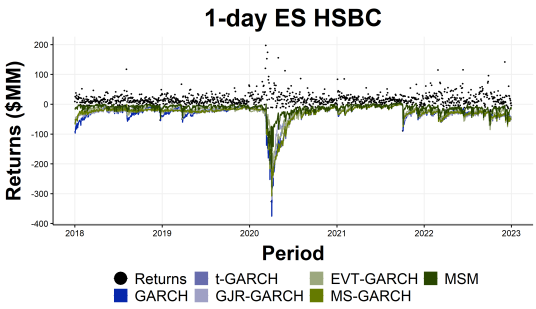
6.1.2 Conditional

At first glance, the conditional models appear to have a better adaptability to current circumstances. Figure 4 shows that all models have a considerable increase for both risk measures around the initial COVID-19 peak. However, this does not necessarily imply that the models capture the conditional coverage 'better'. More formally, the test $\mathbf{T_{con}}$ does not reject the null of conditional coverage in most of the cases. While for the unconditional models the null of $\mathbf{T_{con}}$ is rejected in approximately 38% of the cases at a threshold of 0.05, the conditional models have a rate of 21%. Secondly, we see that the hit rates are close to the expected hit rates in general, especially for the 1-day holding period. For the 5-day holding period, we find that the hit rates are higher than expected, which suggests an underestimation of risk. While the underestimation is limited for MS, it is not for HSBC. This is backed up by the rejection of test $\mathbf{T_{unc}}$ for most models. For the ES, almost all tests are not rejected, which gives the appearance that the volatility models do not underestimate ES.

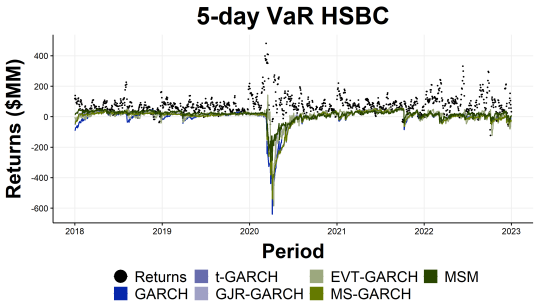
Interestingly, we remark that the VaR of HSBC for the 5-day holding period is at times positive, see figure 4c. This may appear counter-intuitive at first as VaR is sometimes informally described as a measure of potential loss, for example by Wikipedia (2023) and Investopedia (2023). However, by definition, it does not need to be as it merely looks at the lowest 5% of returns (see equation 1). Given the high mean returns for the 5-day period in table 3, it is not surprising to see positive values. Moving onto the individual differences, we notice that the MSM has a higher amount of test rejections among all datasets apart from the 5-day MS data. For all sets, the model appears to underestimate the risk. As the main difference between the stochastic and deterministic approach is the technique to estimate volatility, a reason for the difference might be found there. Based on figure 9, it appears that the volatility estimates vary



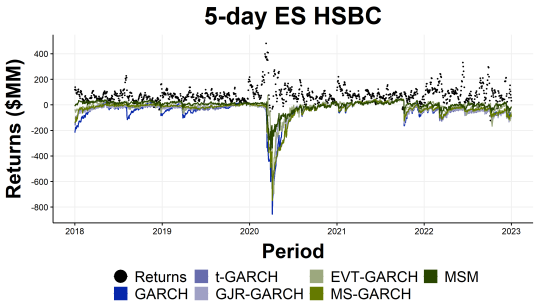
(a) 1-day VaR HSBC



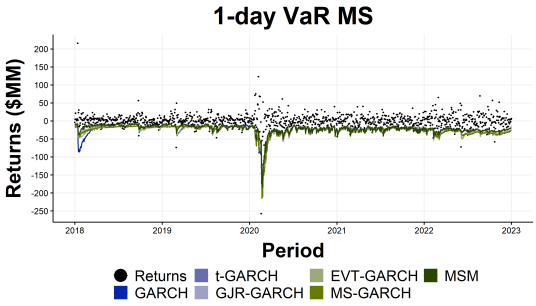
(b) 1-day ES HSBC



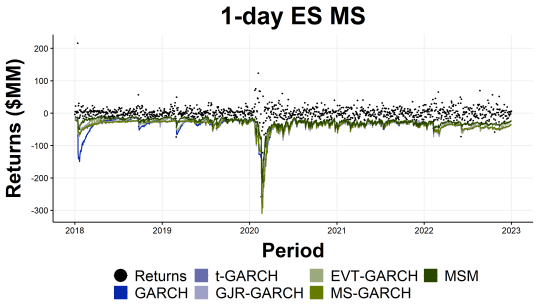
(c) 5-day VaR HSBC



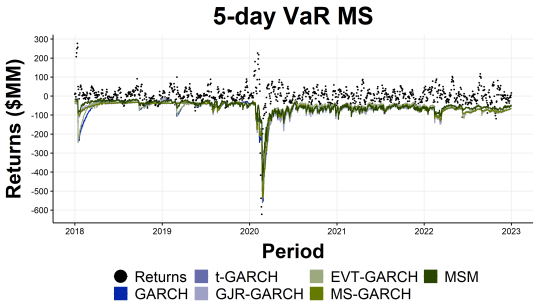
(d) 5-day ES HSBC



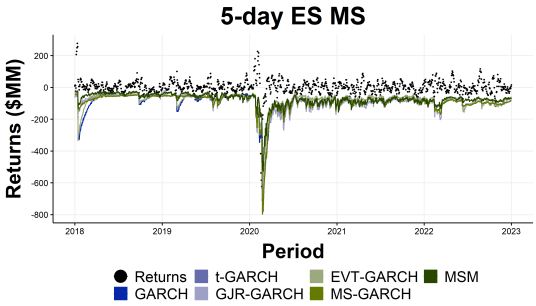
(e) 1-day VaR MS



(f) 1-day ES MS



(g) 5-day VaR MS



(h) 5-day ES MS

Figure 4: Conditional VaR and ES for HSBC and MS.

more than the others. That is, the day-to-day differences are larger than that of a GARCH-type model. As a result, the estimates of VaR are also more volatile, which can, for example, be seen in figure 4a. Second to that, we also notice that even though the volatilities have more variation, the estimates are relatively low compared to others around the COVID-19 peak for both banks, which can be observed in figure 9a. This results in lower VaR values. Next to underestimation, the 'unnecessary' spikes in VaR of the MSM method may also be undesirable as the approach is more susceptible to individual outliers, increasing risk due to a single observation rather than a 'period' of observations due to entering a new 'state'. This can negatively affect the independence of the exceedances.

Looking at the parameters of the models, we note that the leverage effect is small and negative for both banks. Thus, negative residuals induce a small reduction in the height of the volatility. The MS-GARCH parameters show limited changes for the MS data and larger changes for HSBC. Especially with respect to the degrees of freedom, the changes are large, with one state having $\nu = 31.100$. This implies that some of the residuals may better fit in a normal rather than student's t distribution.

Altogether, the conditional models are able to capture the variation over time better, which the unconditional models were lacking. Individually, the deterministic models have limited deviations, while the stochastic model underestimates the risk consistently. Finally, it appears that the risk measures for the 5-day returns of HSBC remain difficult to accurately estimate.

Table 5: Median p-values of the VaR & ES tests for the conditional models of HSBC and MS.

	HSBC							MS						
	Hit rate	VaR Tests			ES Tests			Hit rate	VaR Tests			ES Tests		
	$\alpha = 0.05$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$	$\alpha = 0.05$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$
1-day														
GARCH	0.046	0.54	0.81	0.83	0.33	0.31	0.36	0.055	0.45	0.46	0.32	0.42	0.86	0.88
t-GARCH	0.045	0.38	0.64	0.26	0.29	0.56	0.52	0.055	0.45	0.70	0.64	0.41	0.93	0.94
GJR-GARCH	0.044	0.31	0.55	0.61	0.16	0.59	0.57	0.056	0.32	0.60	0.00	0.48	0.81	0.86
EVT-GARCH	0.046	0.54	0.81	0.50	0.03	0.84	0.86	0.051	0.92	0.04	0.73	0.07	0.48	0.47
MS-GARCH	0.045	0.38	0.64	0.30	0.24	0.62	0.57	0.049	0.87	0.98	0.52	0.21	0.92	0.93
MSM	0.080	0.00	0.00	0.00	0.00	0.02	0.11	0.066	0.01	0.02	0.00	0.01	0.12	0.11
5-day														
GARCH	0.080	0.04	0.11	0.47	0.82	0.21	0.33	0.057	0.60	0.75	1.00	0.23	0.77	0.75
t-GARCH	0.072	0.13	0.12	0.75	0.92	0.37	0.45	0.053	0.50	0.73	1.00	0.26	0.70	0.67
GJR-GARCH	0.080	0.04	0.11	0.68	0.88	0.55	0.53	0.053	0.50	0.52	1.00	0.23	0.78	0.79
EVT-GARCH	0.092	0.01	0.02	0.12	0.25	0.42	0.39	0.066	0.29	0.56	1.00	0.20	0.78	0.87
MS-GARCH	0.072	0.13	0.31	0.77	0.66	0.59	0.62	0.053	0.60	0.85	0.99	0.25	0.67	0.70
MSM	0.112	0.00	0.00	0.00	0.04	0.18	0.17	0.066	0.29	0.38	1.00	0.17	0.78	0.87

6.1.3 Quantile

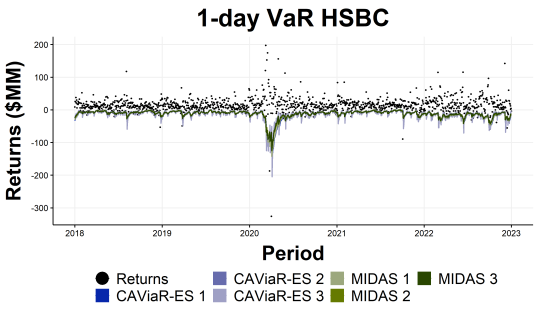
Similar to the conditional models, quantile models should, in theory, be able to capture the time-varying conditions better than the unconditional models. Based on figure 5 this

appears to be the case as both the CAViaR-ES and MIDAS models show increased VaR and ES levels around the extreme COVID-19 returns. This hypothesis is backed up in general by $\mathbf{T_{con}}$ in table 6. Most of the conditional coverage tests are not rejected. For the 5-day VaR of HSBC, though, all the tests are rejected. These rejections, however, may be due to another reason, which leads us to another interesting difference between the forecasts of the categories.

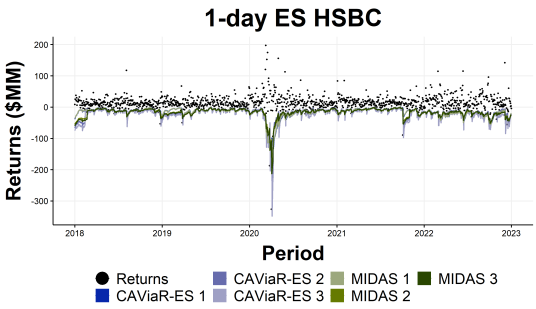
The unconditional models overestimated the 5-day VaR significantly, resulting in low hit rates, while the conditional models underestimated the 5-day VaR consistently. The quantile models appear to provide similar results to the unconditional models, at times overestimating the risk for the 5-day returns according to the hit rates in table 6. Figures 5c and 5g highlight this, showing limited exceedances and high VaR and ES levels for the whole period, especially for the 5-day returns of HSBC. These limited exceedances may hurt the quality of the conditional coverage test, which could be the cause of the rejections for the 5-day returns of HSBC. Overall though, the hit rates for the quantile models are quite decent, with some overestimation of the VaR in certain cases.

Regarding ES, the tests do not reject the hypothesis of no underestimation for the 1- and 5-day returns of MS and 5-day returns of HSBC, assuming $\mathbf{T_{er}}$ to be leading for the 5-day returns. However, the tests do show conflicting results for the 1-day holding period of HSBC as $\mathbf{T_{er}}$ rejects the hypothesis of no underestimation for each model, while tests $\mathbf{T_{reg1,2}}$ do not reject the null of correct specification for each model. As there are enough data points, no clear conclusion can be drawn on the accuracy of the ES estimations for the 5-day HSBC returns. Furthermore, the parameters provide additional insight, especially for the MIDAS models. For example, given the parameters in table 12, we see that for both banks β_1 is more negative for 5-day returns than for 1-day returns and for MS than for HSBC, implying a larger impact of the past returns for longer horizons, ceteris paribus. Additionally, κ is generally larger for MS than for HSBC, which implies that more weight is given to recent observations for MS returns. A higher β_1 and κ together should result in higher volatility of the VaR. Indeed we observe that for the 5-day returns of MS, see figure 5g, the VaR appears more volatile than that of HSBC, figure 5c. More formally, the standard deviation is 68.47 for MS and 26.81 for HSBC.

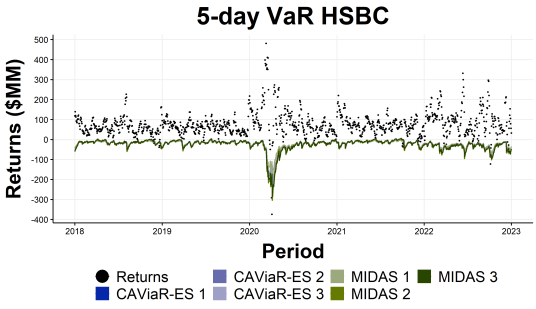
Diving into the comparison of the quantile models there are two main differences to discuss between the models, namely the model type and the specification for the relation between VaR and ES. Comparing the CAViaR-ES and MIDAS output, the 1-day forecasts appear to be relatively similar. The most notable difference can be observed for the risk measures of the 5-day returns of HSBC and MS. Whereas MIDAS is able to directly forecast $h-$days ahead, the CAViaR-ES approach requires one to multiply the results by $\sqrt{h}$. The latter is usually critiqued as it assumes i.i.d. observations. Looking at the figures 5c, 5d, 5g and 5h, we can see that in general the MIDAS estimations are more conservative than the CAViaR-ES estimations, both for the VaR and ES. While, as discussed previously, both models overestimate VaR for HSBC returns, it turns out that MIDAS (slightly) overestimates the VaR for MS returns as well. Concluding, while both models appear to provide decent results for the 1-day returns, the MIDAS approach



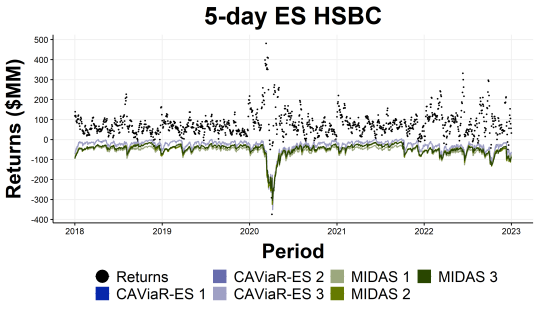
(a) 1-day VaR HSBC



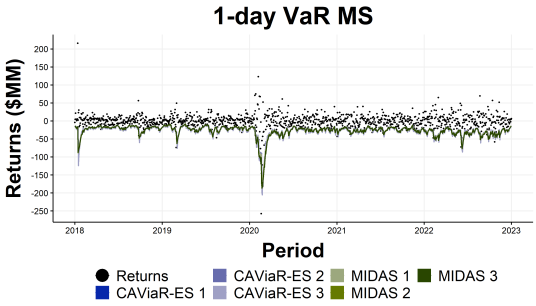
(b) 1-day ES HSBC



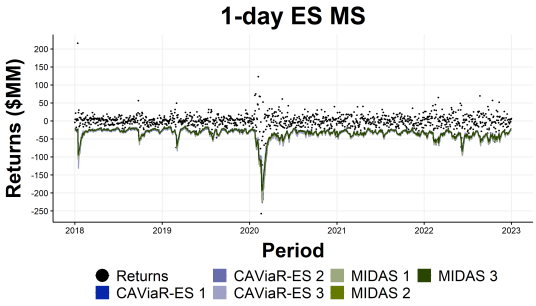
(c) 5-day VaR HSBC



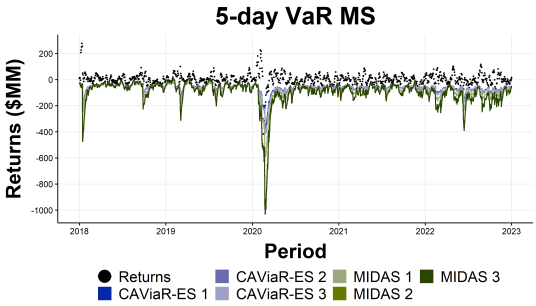
(d) 5-day ES HSBC



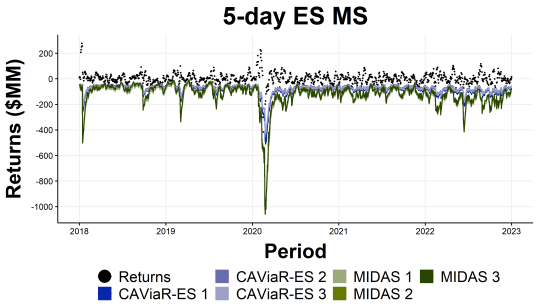
(e) 1-day VaR MS



(f) 1-day ES MS



(g) 5-day VaR MS



(h) 5-day ES MS

Figure 5: Quantile VaR and ES for HSBC and MS.

seems to overestimate the VaR for a 5-day holding period. This hypothesis, however, should be tested further using more data.

Looking at the specifications, recall that the first setup assumes that the relation between VaR and ES remains constant, while the second and third include the possibility of a time-varying relationship between the two risk measures. Based on table 12 it seems that, given setup 1, the ES is relatively larger than VaR for MS compared to HSBC, due to the higher γ_0. For setup 2 and 3, it seems that while HSBC mostly lets j_t depend on the difference between the exceedance and the VaR, the j_t for MS depends more on the previous value of j_t, thus it may fluctuate less. Finally, it is noticeable that the score-driven extension of γ_3 appears unimportant for HSBC, due to the minimal coefficient, and more useful for MS. Especially in the latter case it may be an interesting addition as it could be the reason for the improved hit rate of the MIDAS_3 for the 5-day returns of MS. Apart from that, it seems that the specification has a limited effect on the hit rate and ES tests. The MCS may, however, shine a more definitive light on the relative performance.

Table 6: Median p-values of the VaR & ES tests for the quantile models of HSBC and MS.

	HSBC							MS						
	Hit rate	VaR Tests			ES Tests			Hit rate	VaR Tests			ES Tests		
	$\alpha = \mathbf{0.05}$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$	$\alpha = \mathbf{0.05}$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$
1-day														
CAViaR-ES_1	0.053	0.67	0.03	0.88	0.00	0.08	0.08	0.039	0.07	0.15	0.06	0.18	0.41	0.42
CAViaR-ES_2	0.053	0.67	0.19	0.70	0.01	0.38	0.20	0.040	0.10	0.25	0.06	0.25	0.09	0.08
CAViaR-ES_3	0.043	0.25	0.14	0.83	0.02	0.88	0.84	0.035	0.01	0.03	0.01	0.10	0.38	0.27
MIDAS_1	0.049	0.83	0.98	0.95	0.00	0.20	0.19	0.039	0.07	0.03	0.85	0.29	0.49	0.48
MIDAS_2	0.047	0.63	0.78	0.98	0.01	0.27	0.22	0.037	0.03	0.07	0.67	0.15	0.34	0.34
MIDAS_3	0.046	0.45	0.38	0.97	0.01	0.45	0.38	0.038	0.03	0.09	0.73	0.16	0.36	0.38
5-day														
CAViaR-ES_1	0.016	0.00	0.02	0.96	0.11	0.29	0.32	0.049	0.61	0.85	1.00	0.23	0.61	0.62
CAViaR-ES_2	0.016	0.00	0.02	0.96	0.11	0.29	0.30	0.049	0.81	0.88	1.00	0.10	0.68	0.66
CAViaR-ES_3	0.016	0.00	0.02	0.96	0.11	0.14	0.16	0.049	0.81	0.88	1.00	0.16	0.56	0.54
MIDAS_1	0.016	0.00	0.02	0.95	0.93	0.01	0.01	0.025	0.04	0.11	0.97	1.00	0.00	0.00
MIDAS_2	0.016	0.00	0.02	0.90	0.74	0.12	0.00	0.037	0.32	0.21	0.73	0.98	0.02	0.02
MIDAS_3	0.016	0.00	0.02	0.96	0.69	0.01	0.01	0.041	0.51	0.32	0.83	1.00	0.00	0.00

6.1.4 Forecast combinations

As can be expected, the behaviour of the FC models is similar to what the individual models per category. For example, the unconditional FC seen in figure 11 shows a period of higher VaR and ES for a one-year duration after the COVID-19 peak, similar to the individual unconditional models. Or, when the individual models overestimate VaR, it leads to an overestimated FC VaR as well. Worthwhile to notice is that the FC_{all} appears to have a stable hit rate, hovering close to the expected hit rate, also for the 5-day returns of HSBC. The other FCs sometimes over- or underestimate the VaR significantly. Thus, FC_{all} appears to be a strong candidate model. This hypothesis will

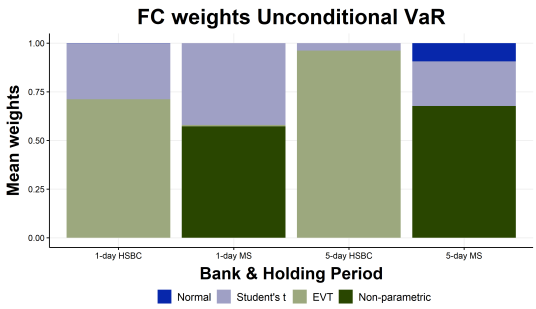
be tested more rigorously in the next section.

Aside from the individual behaviour and performance of the FCs, it may be interesting to look at how the weights of the FCs are built up. That way, we can see which models are preferred to build combinations with. Presented below are the average weights for each of the FCs over time. Before diving into the specifics, we notice globally that the weights for VaR and ES FCs may starkly differ. That is, a relevant model to include in the FC for VaR, may not necessarily be a relevant model to include in the FC for ES. This observation emphasizes the usage of a variety of models to find FCs.

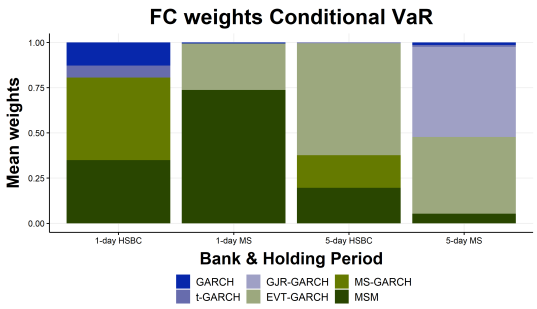
Starting with the unconditional weights, we see that for VaR the non-parametric and EVT approach have the largest weights, while for ES the normal and student's t take up the bulk of the weights. Thus, all models have their relevance for the unconditional FC. For the VaR of HSBC, the EVT has the largest weight. This shouldn't be too surprising, given the EVT has the closest hit rate as discussed before. For the VaR of MS, the non-parametric approach and the student's t are preferred. Again, both models have the closest hit rates so it may not be a large surprise. For ES, it is harder to discern why a model is preferred. In general, the normal and student's t have the least conservative ES, while the EVT and non-parametric approaches have a large ES for MS and HSBC respectively. Altogether, the hit rates of the unconditional FC are relatively close to the expected hit rate, apart from the HSBC 5-day returns as can be seen in table 13. This is no surprise given that all the individual models overestimate the VaR. Hence, the combination will too.

Moving onto the conditional FC weights, it is less apparent to see which models are generally preferred. This can be because the models, in general, produced similar forecasts. Hence, no clear 'victor' may exist. Regarding the VaR weights, the GARCH and t-GARCH models have relatively low weights, with the FC choosing the more complex methods instead. Additionally, the GJR-GARCH is only included for the 5-day returns of MS. In all cases, the MSM is included, even though it occasionally underestimates the risk by a large portion. As a result, the hit rates of the conditional FC are higher than 5%, indicating an underestimation of the VaR. Regarding the ES, the GARCH and especially the t-GARCH specifications are included more frequently. This again highlights that 'good' VaR models for the FC, do not necessarily have to be 'good' models for ES. Second to that, the MS-GARCH has low weights overall while the GARCH is only included for the 1-day forecasts. While the GJR-GARCH has a significant weight for the 1-day forecasts, it has little influence on the 5-day forecasts.

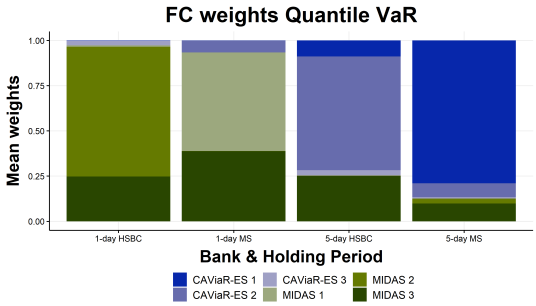
The quantile weights show that the MIDAS approach is preferred for the next-day forecasts of VaR, while the CAViaR-ES models are of higher relevance for the 5-day forecasts. This is interesting because the original theoretical advantage of the MIDAS setup was that it provided a solution to forecast multiple days ahead directly, without having to use an approximation rule such as the square root rule. It might be, however, that the advantage appears for longer horizons when the square root rule becomes more and more likely to have its assumptions violated. Finally, the MIDAS_3 method seems to be irrelevant for the VaR FC. The ES quantile weights show a preference for the CAViaR-ES models, with summed weights of more than 80% for all data aside from



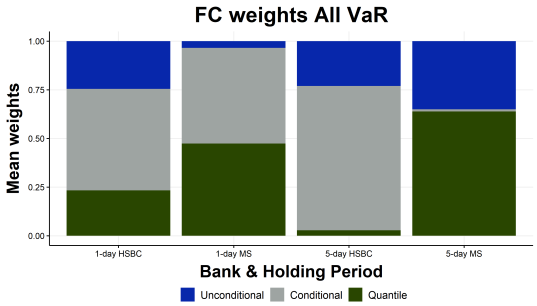
(a) Unc. VaR weights



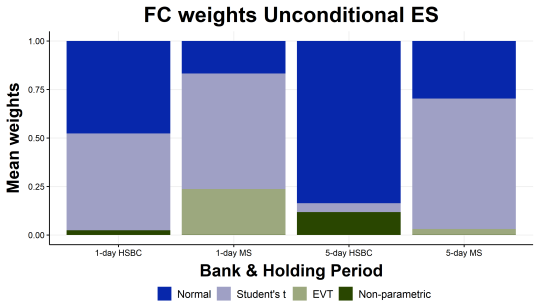
(b) Con. VaR weights



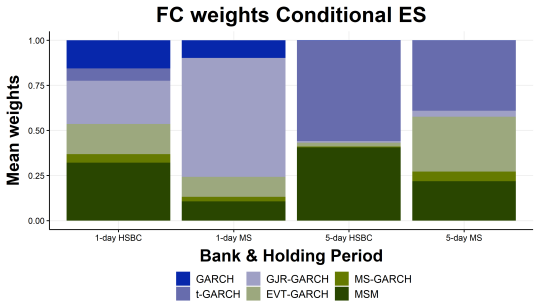
(c) Qua. VaR weights



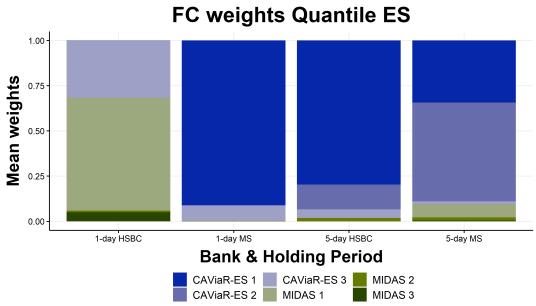
(d) All VaR weights



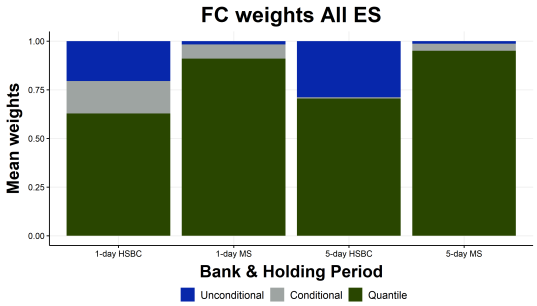
(e) Unc. ES weights



(f) Con. ES weights



(g) Qua. ES weights



(h) All ES weights

Figure 6: Weights of each of the forecast combinations.

the 1-day returns of HSBC. Judging the three setups, CAViaR-ES$_1$ or MIDAS$_1$ hold a relatively high weight for each of the datasets. Thus, the KISS rule appears to apply to the quantile approach.

Finally, all models were put together to create one FC. Regarding the weights of the FC for VaR, each of the categories of models is generally included with a significant weight. This may be a prime advantage of FC$_{all}$. For example, even though the unconditional and quantile models overestimate VaR for the 5-day HSBC returns, the conditional models underestimate it. Combining all three categories results in a more 'balanced' version with a closer hit rate. The ES weights also show an interesting perspective, as the quantile models take up the bulk of the weights, especially for the MS data. A reason for this inclusion may be that the quantile models directly estimate ES, whereas the unconditional and conditional models focus on the full or tail distribution and base the ES thereof. This could be less accurate.

6.2 Performance comparison

While the previous section described each of the models' behaviour, their test results and some possible advantages of the category, the MCS makes the choice for a (set of) models more rigorous. Shown in table 7, each of the models is included to form an informative decision on which (category of) models may be preferred for risk measure estimation.

Starting with the most evident, the unconditional models are mostly absent from each of the model confidence sets. Besides the EVT approach being chosen for the 5-day returns of HSBC and the student's t being included for the 1-day MS returns for the low α, it is clear that the MCS is quick to exclude the unconditional models. Note that not only the individual models are excluded, but the unconditional FC as well. A reason for the exclusion of the unconditional models could be the slow adaptability to new market situations. For example, if a new state occurs with a different mean and variance, the model would require a full new year of observations to include the new state accurately. Knowing the behaviour of financial markets can change quickly, this period may be too long. Given that at a low α almost all other models are within the set, we conclude that unconditional models are likely not able to capture VaR and ES adequately, given the loss function $\mathbf{L_2}$. This can have large consequences for lawmakers, as, during the 2022 Market Risk Benchmarking Exercise of the EBA, 29 out of the 40 banks applied the unconditional approach (EBA, 2022). This may be due to its ease of application and understanding, but, given that the other categories are also relatively easy to comprehend and tractable, simplicity should not be the main consideration if it proves to be an inadequate estimator. Therefore, we suggest policy makers to do further research on the comparison with more data and scenarios, to further test the hypothesis of the unconditional models being inadequate for risk estimation.

Next to the unconditional models being left out of most of the confidence sets, the conditional models are underrepresented compared to the quantile and FC categories for a high α. This contradicts the observations of Righi and Ceretta (2015). While we should be more wary to draw conclusions from this observation as the confidence

Table 7: The MCS for both the HSBC and MS at a low and high α.

	Low α				High α			
	HSBC		MS		HSBC		MS	
	1-day	5-day	1-day	5-day	1-day	5-day	1-day	5-day
Unconditional								
Normal								
Student's t			■					
EVT		■						
Non-parametric								
Conditional								
GARCH	■	■	■		■			
t-GARCH	■	■	■	■			■	
GJR-GARCH	■	■	■	■	■			
EVT-GARCH		■	■	■				■
MS-GARCH	■	■	■	■				
MSM	■	▨	■	■			■	■
Quantile								
CAViaR-ES$_1$	▨	■	▨	■	▨		■	■
CAViaR-ES$_2$	■	■	■	■	■		■	■
CAViaR-ES$_3$	■	■	■	■				■
MIDAS$_1$	■		▨	■	■		▨	▨
MIDAS$_2$	■	■	■	■	■		■	■
MIDAS$_3$	■	■	■	■	■		■	▨
FC								
FC$_u$								
FC$_c$	■	▨	■	■			■	
FC$_q$	▨	■	■	■	▨		▨	▨
FC$_{all}$	▨	▨	▨	▨	▨	▨	▨	■

Note: If a model is included in the MCS, it has a black square, ■. The top three models have a striped square with different colors. From rank 1 to 3, we use the squares ▨, ▨ and ▨. Similar to the tests, we use the median loss per model for robustness.

level has lowered, it nevertheless provides a possible indication that the quantile and FC categories may be preferred in general. As stated earlier, a possible reason for this may be that the conditional models find a simulated distribution from which to gather the VaR and ES. A major part of finding the simulated distribution depends upon the estimated current volatility. If the estimates are inaccurate, this can result in differences in the distribution, especially for the tail. The quantile and FC categories, on the other hand, directly estimate VaR and ES through optimization. Thus, their estimates may more accurately reflect the 'expected values', whereas the conditional approach may

better reflect how the general distribution looks. Finally, we note that, in general, it seems that the holding period does not matter for the conclusion on which category of models is (not) preferred.

Looking at the performance of the models, it is clear that the FC_q and FC_{all} perform well for most datasets. This challenges the conclusion of Trucíos and Taylor (2022), who found that forecast combinations provide no improvement, although it was on cryptocurrency. The FC_{all} is the only model selected for the 5-day returns of HSBC at a high α. An explanation could be that it is the only model that comes close to the expected hit rate, with a value of 5.6%. All other models notably over- or underestimate the VaR. Additionally, the FC_{all} has been selected in almost all cases as the top performer. The reason for the strong performance is possibly because the FC can use a diversity of approaches with their respective (dis)advantages to select a combination that captures the current information well. However, given that the method optimizes a function rather than looking at relationships, it can be interesting to do further research into the performance of FC for longer holding periods.

As banks have to choose a model, the MCS shows that the FC_{all} can be a good option to consider for both holding periods. A drawback is that it requires a significant amount of modelling to include a wide variety of models to draw from. For some institutions, this may be too capital-intensive, particularly if the trading book is small or has major hedges. For these banks, it may be more beneficial to apply a quantile approach. Markedly, of the quantile models, the simple setup 1 is most frequently chosen as a top performer. There is no clear preference for the MIDAS or CAViaR-ES approach.

7 Sensitivity Analysis

In this section, we apply two different scenarios to delve into how the performance of the models would have differed if the input changes. First, the well-known S&P500 returns are used to see how the results would have differed if we were to use an index, rather than trading book returns. While the latter is more realistic, the prior has easy access and is frequently used in analysis. Secondly, we change the number of observations included in the unconditional models from $N = 252$ to $N = 504$. Longer periods of data may provide more data for an accurate distribution, though older information may not portray the current environment as well.

Besides the aforementioned two input changes, we have tried and tested several other possible avenues. For example, a different distribution, such as the generalized error distribution, was applied for the residuals of the conditional models. Next to that, the loss function of the MCS was also altered, using the AL log score of Taylor (2019). Both of these applications led to small, insignificant changes to the prior results. Therefore, they are excluded from this section.

7.1 S&P500

In general, the outcomes in table 8 are in line with the findings of the previous section. For a low α, most of the unconditional models are taken out of the confidence set. For a high α most of the models are excluded, with mostly FC and quantile models remaining. Second to that, the FC category has the most 'top ranking' models. Finally, the conditional models are in general worse at estimating VaR and ES for the 5-day than the 1-day holding period, due to increased underestimation. Nevertheless, there are some minor differences.

Table 8: The MCS and some basic tests and statistics using the S&P500.

	Tests				MCS			
	Hit rate		$\mathbf{T_{er}}$		Low α		High α	
	1-day	5-day	1-day	5-day	1-day	5-day	1-day	5-day
Unconditional								
Normal	0.076	0.065	0.00	0.00		■		
Student's t	0.085	0.080	0.82	0.55				
EVT	0.069	0.061	0.06	0.09				
Non-parametric	0.077	0.072	0.03	0.14	■	■		
Conditional								
GARCH	0.057	0.061	0.96	0.23	■	■		
t-GARCH	0.055	0.069	0.98	0.18	■	■		
GJR-GARCH	0.060	0.076	0.98	0.07	■	■		
EVT-GARCH	0.073	0.088	0.00	0.02	■	■		
MS-GARCH	0.049	0.065	0.86	0.15	■	■		
MSM	0.093	0.118	0.00	0.00	■	■		
Quantile								
CAViaR-ES_1	0.050	0.072	0.83	0.53	■			
CAViaR-ES_2	0.050	0.065	0.06	0.02	■	■		
CAViaR-ES_3	0.050	0.065	0.57	0.17	■	■		
$MIDAS_1$	0.038	0.061	0.00	0.00	■	▩		▩
$MIDAS_2$	0.053	0.053	0.95	1.00	▩			
$MIDAS_3$	0.053	0.042	0.98	1.00	■	■		
FC								
FC_u	0.078	0.069	0.13	0.15		■		
FC_c	0.070	0.085	0.99	0.60	■	■		
FC_q	0.053	0.042	0.96	0.80	▩	▩		▩
FC_{all}	0.056	0.050	0.96	0.85	▩	▩	▩	▩

Note: If a model is included in the MCS, it has a black square, ■. The top three models have a striped square with different colors. From rank 1 to 3, we use the squares ▩, ▩ and ▩.

First, relatively more unconditional models are included, with 37.5% of the cases being included at a low α. For HSBC and MS, only 12.5% of the time an unconditional model was included. The inclusion may be due to a better performance of the unconditional models, a worse performance of the other categories, or both. It is not apparent what the 'true' reason for this discrepancy is. Altogether, more value should be given to the actual PnL data, but the sensitivity analysis is an indication that additional data may be necessary to further substantiate the findings.

Secondly, the FC_{all} is less frequently the top performer, particularly for the 5-day horizon. Nevertheless, it is still included in the confidence set for both a high and low α. Curiously, compared to the MCS of the HSBC and MS, far fewer models are included at a high α. This may indicate that there is a larger 'gap' in performance for the top models and the remainder for the S&P500.

Finally, at a low α, the amount of selected conditional and quantile models using 1-day returns are the same, but the number of quantile models selected for the 5-day returns is lower. This contradicts the findings using HSBC and MS returns. Therefore, to quantify a potential difference in performance between conditional and quantile models, more data is required. Of course, the actual trading book data should be preferred when making a decision. Drawing a conclusion on a possible preference is valuable, as FC may be too labour-intensive to implement for financial institutions.

7.2 Different unconditional size

Rather than using the past year as input, which is currently the minimum requirement and the preference of most banks in the European Monetary Union (EMU) according to supervisors, the past two years are used for the modelling.

From table 9 the main insight is that the unconditional models with 2 years of data are included more frequently than the models with 1 year of data. This holds for each of the banks, holding period and α's other than the 5-day returns of HSBC for a high α. Consequently, even if institutions find switching to more rigorous models not beneficial enough to implement, a possibly more fruitful solution can be quickly implemented by adding more data. An important note to make is that, as said frequently before, supplementary data is useful to gain a better understanding and draw a more 'definitive' conclusion to the question at hand.

Additionally, we see that the student's t and EVT for $N = 504$ are frequently selected at both α's. This is particularly interesting as the normal and non-parametric options are preferred for the S&P500 when an unconditional model is selected. For the HSBC and MS datasets, we have seen that the student's t and EVT approaches result in a VaR that is less conservative. Given the limited volatility of the returns of these banks in general, this may be more fitting. For the S&P500, which has more volatility in the returns, the models may insufficiently capture the VaR (and ES) because of underestimation. Altogether, though, it is hard to draw a strong conclusion on which of the unconditional models seems most fitting to implement. Applying a forecast combination instead of an individual model may be an option to consider. This can be looked into further if financial institutions decide not to deviate from using unconditional models.

Table 9: The MCS of different unconditional sizes for both the HSBC and MS at a low and high α.

	Low α				High α			
	HSBC		MS		HSBC		MS	
	1-day	5-day	1-day	5-day	1-day	5-day	1-day	5-day
N=252								
Normal								
Student's t			■	■				
EVT	■	■				■		
Non-parametric								
N=504								
Normal			■					
Student's t	■	■	■	■	■		■	■
EVT	■	■	■	■	■	■		
Non-parametric			■	■				

Note: If a model is included in the MCS, it has a black square, ■.

8 Conclusion

In conclusion, the analysis conducted on various categories of risk measurement models using actual PnL data at different holding periods from the HSBC and MS provides valuable insights into their performance. The MCS framework, in combination with visualizations and formal testing, has been instrumental in comparing the different model categories and their suitability for risk estimation. The key takeaways are the following:

- Unconditional Models: The MCS effectively excludes unconditional models, implying a lower performance. This exclusion can be attributed to their limited adaptability to rapidly changing market conditions, as the models require more data on the 'new state' than may be available, to accurately estimate VaR and ES. Given that roughly three-quarters of the banks under ECB supervision use the approach, it is suggested to further research the category to assess its adequacy for risk estimation in dynamic financial markets.

- Conditional Models: The MCS indicates that conditional models are represented more frequently than unconditional models. This is likely attributed to the better capturing of the dynamics of the market. The category is underrepresented in the MCS compared to quantile and forecasting combination categories, albeit at a low confidence. This could be due to the sensitivity of conditional models to volatility estimates and their potentially less accurate representation of the tails of the distribution.

- Quantile Models: The quantile models are frequently included in the MCS, possibly

showing their usefulness in risk estimation. Among the quantile models, the non-time-varying setup of the CAViaR-ES and MIDAS models most often emerged as a 'top performer' in risk estimation, which may be due to their simplicity and tractability. The other setups may provide similar results, but given a similar performance, simplicity may be a better option. There is no clear preference for the CAViaR-ES or MIDAS models.

- FC Models: The FCs showed a performance similar to their respective categories. The unconditional FC was excluded, for a similar reason as mentioned above. The other combinations were frequently chosen at a high confidence, but only FC_q and FC_{all} were consistently chosen at a low confidence. Additionally, the FC_{all} model stands out as a top performer across various datasets and holding periods. Its ability to optimize combinations of models is a likely contributor to its strong performance. However, it requires an extensive set of models, which may be too labour-intensive to implement for some institutions.

- Different holding periods: The different holding periods only led to marginal differences when looking at the performance of the models through the MCS. This may be due to using only a 5-day holding period, instead of a longer period, and could be investigated further.

The sensitivity analysis using the S&P500 backed up the aforementioned statements in general. Although most of the unconditional models were excluded, the frequency of inclusion in the confidence set increased compared to the HSBC and MS. Thus, if we had used the S&P500, the drawn conclusion would have been with lower confidence. We also observed a stronger difference between the number of models included at a high and low confidence. Finally, the quantile models were included at a similar rate as the conditional models at a high α. This is contrary to the HSBC and MS cases, where the quantile models were included more frequently. Next to the different dataset, we used a different number of observations for the unconditional models to test a possible preference. Two years of observations may provide better results than one year.

In light of these findings, we encourage policymakers to further study the behaviour and performance of each category, with a particular focus on the possible underperformance of the unconditional models. Given that banks have to choose a single model, we suggest the promising FC_{all} option for risk estimation as it has shown to be ranked highly consistently. For smaller institutions, a simpler conditional or quantile-based approach may be wiser. Overall, this study underscores the importance of rigorous model comparison and ongoing research to ensure accurate risk assessment in dynamic financial markets.

For further research, a study similar to ours using more PnL data, either through using other banks, more years or both, would be beneficial to further test the performance of each of the categories. Similar to the previous proposal, using longer time series can also lead to research on the performance for longer time horizons, which was unfeasible for this research. Finally, given the performance of the FC, it may be interesting to look into applying different weight structures to test for a preference.

References

Acerbi, C. and Szekely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11):76–81.

Allen, D. E., Singh, A. K., and Powell, R. J. (2013). EVT and tail-risk modelling: Evidence from market indices and volatility series. *The North American Journal of Economics and Finance*, 26:355–369.

Almeida, D. d. and Hotta, L. K. (2014). The leverage effect and the asymmetry of the error distribution in GARCH-based models: The case of Brazilian market related series. *Pesquisa Operacional*, 34:237–250.

Ardia, D., Bluteau, K., Boudt, K., and Catania, L. (2018). Forecasting risk with Markov-switching GARCH models: A large-scale performance study. *International Journal of Forecasting*, 34(4):733–747.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., and Trottier, D.-A. (2019). Markov-switching GARCH models in R: The MSGARCH package. *Journal of Statistical Software*, 91:1–38.

Artemova, M., Blasques, F., van Brummelen, J., and Koopman, S. J. (2022). Score-driven models: Methods and applications. In *Oxford Research Encyclopedia of Economics and Finance*.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

Barone-Adesi, G., Giannopoulos, K., and Vosper, L. (1999). VaR without correlations for portfolios of derivative securities. *Journal of Futures Markets*, 19(5):583–602.

Batten, J. A., Kinateder, H., and Wagner, N. (2014). Multifractality and value-at-risk forecasting of exchange rates. *Physica A: Statistical Mechanics and its Applications*, 401:71–81.

Bauwens, L., Dufays, A., and Rombouts, J. V. (2014). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*, 178:508–522.

Bayer, S. and Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3):437–471.

Benito, S., López-Martín, C., and Navarro, M. Á. (2023). Assessing the importance of the choice threshold in quantifying market risk under the pot approach (evt). *Risk Management*, 25(1):6.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300.

BIS (2015). Revised Pillar 3 disclosure requirements. https://www.bis.org/bcbs/publ/d309.pdf. Accessed: 2023-06-22.

BIS (2019). Explanatory note on the minimum capital requirements for market risk. https://www.bis.org/bcbs/publ/d457_note.pdf. Accessed: 2023-06-25.

BIS (2019). Minimum capital requirements for market risk. https://www.bis.org/bcbs/publ/d457.pdf. Accessed: 2023-07-20.

BIS (2022). Standardised approach: individual exposures. https://www.bis.org/basel_framework/chapter/CRE/20.htm#:~:text=Risk%20weighted%20assets%20are%20calculated,including%20partial%20write%2Doffs). Accessed: 2023-08-03.

Calvet, L. E. and Fisher, A. J. (2004). How to forecast long-run volatility: Regime switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, 2(1):49–83.

Carnero, M. Á., Peña, D., and Ruiz Ortega, E. (2001). Is Stochastic Volatility more flexible than GARCH?

Chen, Q., Gerlach, R., and Lu, Z. (2012). Bayesian Value-at-Risk and expected shortfall forecasting via the asymmetric Laplace distribution. *Computational Statistics & Data Analysis*, 56(11):3498–3516.

Chen, S. X. (2008). Nonparametric estimation of expected shortfall. *Journal of financial econometrics*, 6(1):87–107.

Chinhamu, K., Huang, C.-K., Huang, C.-S., Chikobvu, D., et al. (2015). Extreme risk, value-at-risk and expected shortfall in the gold market. *International Business & Economics Research Journal (IBER)*, 14(1):107–122.

Christoffersen, P. (2011). *Elements of financial risk management*. Academic press.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, pages 841–862.

Christou, E. and Grabchak, M. (2022). Estimation of expected shortfall using quantile regression: a comparison study. *Computational economics*, 60(2):725–753.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2):223.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Danielsson, J. and De Vries, C. G. (2000). Value-at-risk and extreme returns. *Annales d'Economie et de Statistique*, pages 239–270.

EBA (2022). EBA Report on Results from the 2022 Market Risk Benchmarking Exercise. https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2023/Credit%20and%20Market%20risk%20benchmarking/1052678/EBA%20Report%20results%20from%20the%202022%20Market%20Risk%20Benchmarking%20Exercise.pdf. Accessed: 2023-09-01.

EBA (2023). Article 365. https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/101466. Accessed: 2023-06-25.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of business & economic statistics*, 22(4):367–381.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle.

Ghysels, E., Plazzi, A., and Valkanov, R. (2016). Why invest in emerging markets? The role of conditional return asymmetry. *The Journal of Finance*, 71(5):2145–2192.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Grabchak, M. and Christou, E. (2021). A note on calculating expected shortfall for discrete time stochastic volatility models. *Financial Innovation*, 7(1):43.

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

Hull, J. and White, A. (2014). Hull and White on the pros and cons of expected shortfall. https://www.risk.net/risk-management/market-risk/2375185/hull-and-white-on-the-pros-and-cons-of-expected-shortfall. Accessed: 2023-06-13.

Humphreys, B. (2006). Various Assumptions Required: Historical Simulation VAR. https://www.risk.net/media/download/921841/download#:~:text=However%2C%20the%20historical%20VAR%20calculation,that%20this%20assumption%20is%20true. Accessed: 2023-06-03.

Investopedia (2023). Value at Risk. https://www.investopedia.com/terms/v/var.asp. Accessed: 2023-08-27.

Komunjer, I. (2007). Asymmetric power distribution: Theory and applications to risk measurement. *Journal of applied econometrics*, 22(5):891–921.

Kupiec, Paul H and others (1995). *Techniques for verifying the accuracy of risk measurement models*, volume 95. Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board.

Lamoureux, C. G. and Lastrapes, W. D. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business & Economic Statistics*, 8(2):225–234.

Le, T. H. (2020). Forecasting value at risk and expected shortfall with mixed data sampling. *International Journal of Forecasting*, 36(4):1362–1379.

Lee, H., Song, J. W., and Chang, W. (2016). Multifractal value at risk model. *Physica A: Statistical Mechanics and its Applications*, 451:113–122.

Liu, R. and Lux, T. (2008). Higher Dimensional Multi-fractal Processes: Filtering via Simulation.

Mandelbrot, B. B. (2013). *Fractals and scaling in finance: Discontinuity, concentration, risk. Selecta volume E*. Springer Science & Business Media.

Mandelbrot, B. B. and Taleb, N. N. (2010). Focusing on those risks that matter. *The known, the unknown, and the unknowable in financial risk management: Measurement and theory advancing practice*, 47.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3-4):271–300.

Meng, X. and Taylor, J. W. (2020). Estimating value-at-risk and expected shortfall using the intraday low and range data. *European Journal of Operational Research*, 280(1):191–202.

Paul, S. and Sharma, P. (2017). Improved VaR forecasts using extreme value theory with the Realized GARCH model. *Studies in Economics and Finance*, 34(2):238–259.

Pederzoli, C. (2006). Stochastic volatility and GARCH: A comparison based on UK stock data. *European Journal of Finance*, 12(1):41–59.

Pérignon, C. and Smith, D. R. (2010). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking & Finance*, 34(2):362–377.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking & Finance*, 30(2):561–582.

Righi, M. B. and Ceretta, P. S. (2015). A comparison of expected shortfall estimation models. *Journal of Economics and Business*, 78:14–47.

Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian journal of statistics*, 33(2):307–335.

Steen, M., Westgaard, S., and Gjølberg, O. (2015). Commodity value-at-risk modeling: comparing RiskMetrics, historic simulation and quantile regression.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2):231–252.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2):428–441.

Trucíos, C. and Taylor, J. W. (2022). A comparison of methods for forecasting value at risk and expected shortfall of cryptocurrencies. *Journal of Forecasting*.

Wei, Y., Chen, W., and Lin, Y. (2013). Measuring daily Value-at-Risk of SSEC index: A new approach based on multifractal analysis and extreme value theory. *Physica A: Statistical Mechanics and its Applications*, 392(9):2163–2174.

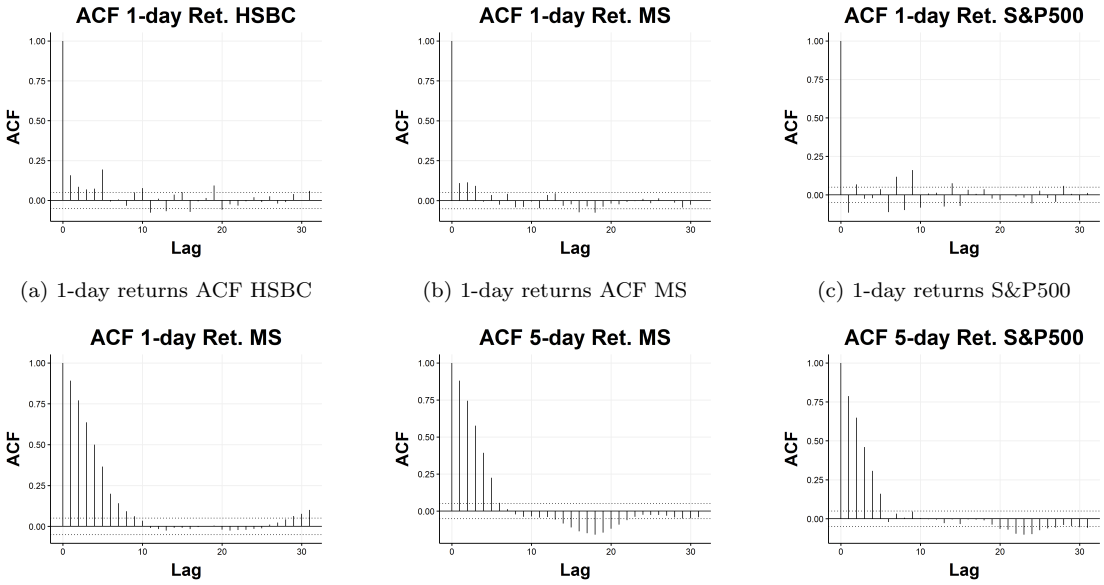
Wikipedia (2023). Value at Risk. https://en.wikipedia.org/wiki/Value_at_risk. Accessed: 2023-08-27.

Yamai, Y. and Yoshiba, T. (2005). Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking & Finance*, 29(4):997–1015.

Zhu, D. and Galbraith, J. W. (2011). Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions. *Journal of Empirical Finance*, 18(4):765–778.

A Data

Figures 7 and 8 show the autocorrelation function (ACF) of the returns and squared residuals for the HSBC, MS and S&P500 data. Starting with the ACF of the returns, the 1-day returns show low autocorrelation between the returns. On the other hand, the 5-day returns show significant autocorrelation. This is to be expected because the 5-day returns at time t are similar to the 5-day returns at time $t-1$.

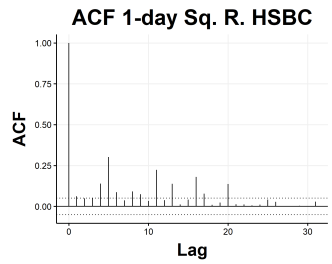


(a) 1-day returns ACF HSBC (b) 1-day returns ACF MS (c) 1-day returns S&P500

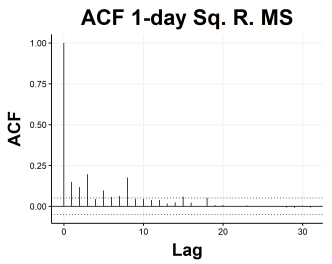
(d) 5-day returns ACF HSBC (e) 5-day returns ACF MS (f) 5-day returns ACF S&P500

Figure 7: ACF returns

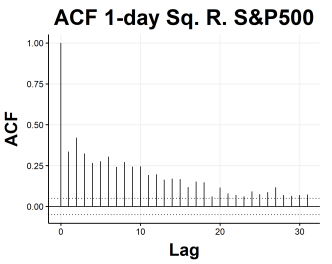
The ACF of the squared residuals should show some autocorrelation as it is a sign of heteroskedasticity of the volatility, a stylized fact of financial time series. Indeed, we observe that there is some autocorrelation for all figures. For the 5-day returns, however, the effect is more pronounced than the 1-day returns. Additionally, the ACF for the S&P500 appears to show a longer lag dependency than the ACF for the HSBC and MS data.



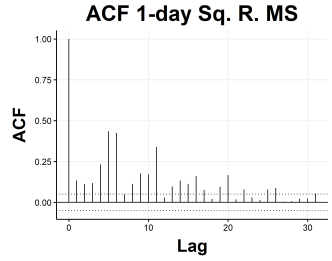
(a) 1-day squared residuals ACF HSBC



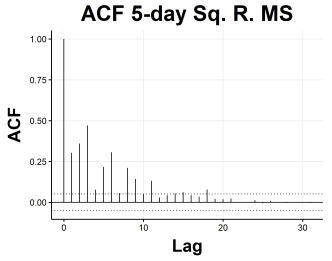
(b) 1-day squared residuals ACF MS



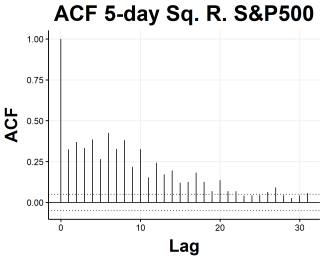
(c) 1-day squared residuals ACF S&P500



(d) 5-day squared residuals ACF HSBC



(e) 5-day squared residuals ACF MS



(f) 5-day squared residuals ACF S&P500

Figure 8: ACF Squared Residuals

B Results

B.1 Parameters

The following tables show the parameters of each of the models. For the unconditional models, the 1- and 5-day parameters are the same, as the 1-day risk measures are translated to 5-day risk measures through the square root rule. Unsurprisingly, the estimated mean is higher for HSBC. Next to that, the variance of the student's t is lower than that of the normal distribution. This is likely because the student's t is able to capture large returns by increasing the fatness of the tail through the degrees of freedom.

Table 10: Parameters of each unconditional model

Return series	Normal		Student's t			EVT	
	μ	σ	μ	σ	df	β	ξ
1-day & 5-day							
HSBC	18.20	25.01	16.19	17.50	3.71	-0.02	14.32
MS	0.98	20.01	0.50	16.28	5.71	0.048	10.37

The conditional parameters show that the current volatility is highly dependent on the previous volatility rather than on the squared residuals. The MSGARCH approach

shows a higher coefficient for the squared residuals compared to the other GARCH models. Comparing the time series, it appears that skewness has a larger influence on the HSBC returns than the MS returns. Finally, similar to the unconditional parameters, the degrees of freedom are higher for the MS data in general.

Table 11: Parameters of each conditional model

Model	HSBC									MS								
	α	β	ϕ	ν	ξ	m_0	σ	b	γ_1	α	β	ϕ	ν	ξ	m_0	σ	b	γ_1
1-day, 5-day																		
GARCH	0.01	0.99								0.02	0.98							
t-GARCH	0.03	0.97		3.95	1.23					0.02	0.97		8.66	1.06				
GJR-GARCH	0.03	0.97	-0.01	3.83	1.23					0.03	0.97	-0.01	8.73	1.06				
MS-GARCH	0.06	0.93		31.10	1.16					0.07	0.93		5.14	1.07				
	0.03	0.97		4.59	1.25					0.07	0.93		7.82	1.05				
MSM						1.46	4.87	1.00	19.45						1.29	4.80	1.00	17.16

Note: The EVT-GARCH is left out of the table given it uses the same parameters, but with a different simulation approach. Additionally, the 1-day and 5-day parameters are the same as FHS is used to gain $h-$day forecasts.

Table 12: Parameters of each quantile model

Model	HSBC							MS						
	β_0	β_1	β_2/κ	γ_0	γ_1	γ_2	γ_3	β_0	β_1	β_2/κ	γ_0	γ_1	γ_2	γ_3
1-day														
CAViaR-ES$_1$	-0.82	0.87	-0.22	-2.21				-0.84	0.82	-0.32	-1.02			
CAViaR-ES$_2$	-3.35	0.69	-0.39	0.15	0.42	0.00		-0.83	0.82	-0.32	8.48			
CAViaR-ES$_3$	-5.02	0.58	-0.53	0.51	0.51	0.07	0.00	-1.73	0.71	-0.51	0.03			
MIDAS$_1$	-6.38	-1.60	6.39	-2.07				-4.12	-1.81	8.67	-0.98			
MIDAS$_2$	-10.27	-1.23	6.57	0.78	0.28	0.00		-2.59	-2.02	9.30	3.71			
MIDAS$_3$	-10.61	-1.24	9.05	0.00	0.34	0.00	0.00	-2.24	-2.04	9.02	1.97			
5-day														
MIDAS$_1$	9.53	-4.59	5.31	-6.00				10.48	-8.97	9.04	-0.57			
MIDAS$_2$	9.99	-4.63	5.27	0.00	0.00	0.00		46.87	-11.76	8.71	1.63			
MIDAS$_3$	10.01	-4.63	5.20	0.00	0.00	0.11	0.00	43.39	-11.39	10.16	0.00			

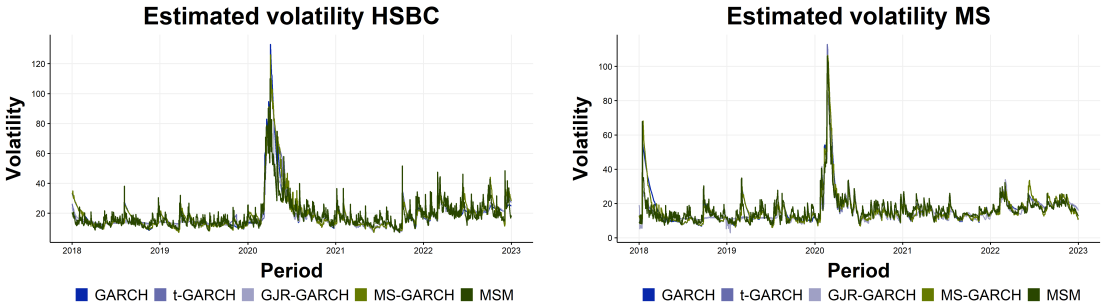
Note: Because CAViaR-ES uses the square root rule to translate 1- to $h-$day forecasts, the parameters of the 1-day returns are the same as the 5-day returns. Hence, they are excluded for the 5-day holding period. Additionally, keep in mind the multiplicative structure of the first setup, compared to the additive structure of the second and third setup. Hence, γ_0 has a different interpretation depending on the setup.

The quantile parameters have been discussed thoroughly in the results section. Some other interesting notes to add are that the constant, β_0, is much higher for the 5-day MS returns than the 1-day returns for the MIDAS models. Additionally, it appears that the ES is relatively close to the VaR estimates for the 5-day HSBC case as the $\gamma_{0:3}$

parameters are quite negative for setup 1 and close to zero for setup 2 and 3. Given that the former uses the translation $ES_t = (1 + exp(\gamma_0))VaR_t$, the estimates are practically equal for a large negative value of γ_0.

B.2 Conditional models volatility

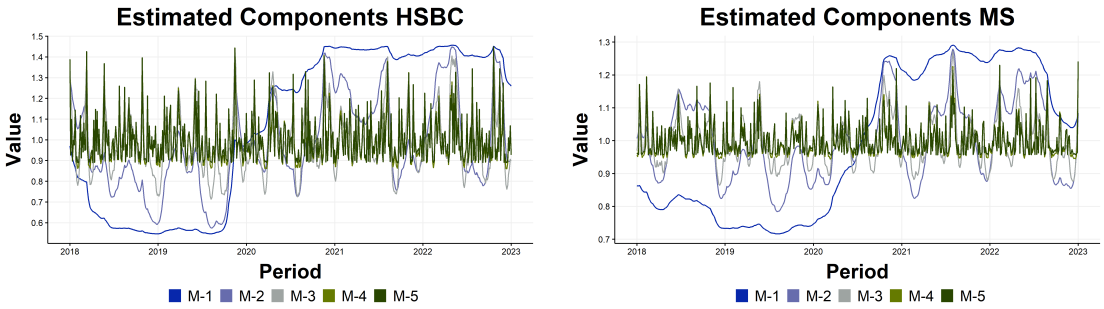
In figure 9, we see the volatilities over time of the different deterministic and stochastic models. In both cases, there was a significant peak around the start of the COVID-19 pandemic. Additionally, the returns show slightly larger volatility around 2022 and thereafter, with the HSBC data showing larger changes in volatility in this period as well. In general, the models appear to provide relatively similar volatility estimates, though the stochastic approach results in lower estimates in general.



(a) HSBC Volatilities

(b) MS Volatilities

Figure 9: Volatility estimations of the various models for HSBC and MS.



(a) HSBC Components

(b) MS Components

Figure 10: Components of the MSM approach for HSBC and MS.

The MSM consists of components with different states, going from longer-term states to short-term states. The components are shown over time in figure 10, where the first component has the least variability and the last, fifth, component the most. Around the COVID-19 peak, the first component for both datasets increased significantly. These components stayed at a high level for a while and dropped quickly, at least for the MS

case. For the HSBC case, the drop seems to occur at the final stages of the data. For the other components, it is harder to identify a specific pattern.

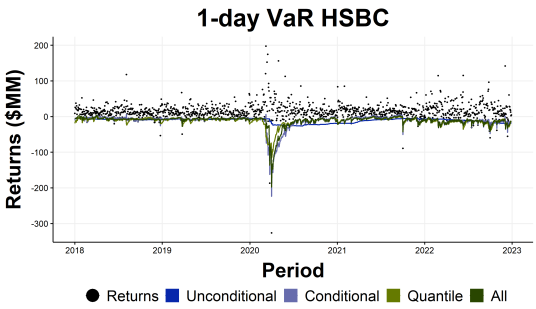
B.3 Forecast combinations

The forecast combination figures and test data are provided in figure 11 and table 13. Regarding the figures, the behaviour is as to be expected. For example, the unconditional FC is less volatile, with limited action around the COVID-19 peak. Surprising to see is the spike of the unconditional ES for the 1-day returns of MS in figure 11f, which is significantly larger than the other models and could be a significant overestimation.

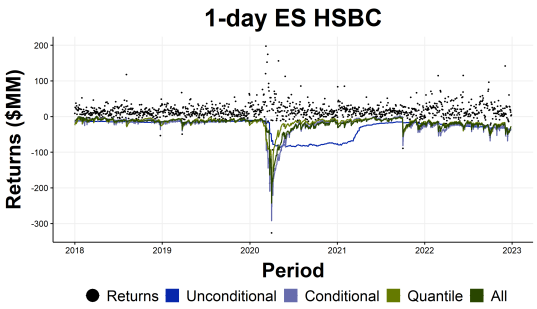
Table 13 provides the test results of each FC. The unconditional FC has the most tests rejected, while the FC including all models has only one test rejected. The latter also provides the most accurate hit rates on average, while the other FCs over- or underestimate VaR for a part of the data. Whereas the conditional FC consistently underestimates VaR, the quantile FC consistently overestimates it. Also relevant to note is that following $\mathbf{T_{er}}$, most of the models reject the null of no underestimation of the ES for the HSBC data at a 1-day holding period, while all models do not reject the null of underestimation of ES for the 5-day holding period.

Table 13: Median p-values of the VaR & ES tests for the forecast combination models of HSBC and MS.

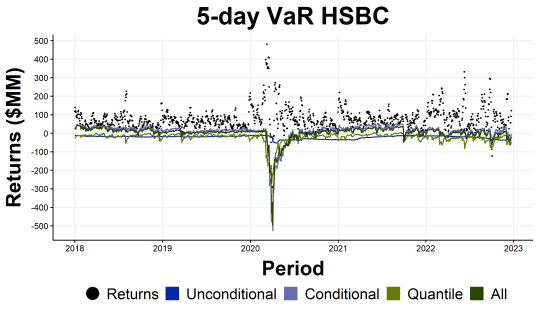
	HSBC							MS						
	Hit rate	VaR Tests			ES Tests			Hit rate	VaR Tests			ES Tests		
	$\alpha = \mathbf{0.05}$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$	$\alpha = \mathbf{0.05}$	$\mathbf{T_{unc}}$	$\mathbf{T_{con}}$	$\mathbf{T_{dq}}$	$\mathbf{T_{er}}$	$\mathbf{T_{reg1}}$	$\mathbf{T_{reg2}}$
1-day														
FC_u	0.039	0.07	0.07	0.15	0.10	0.01	0.03	0.049	0.88	0.10	0.00	0.77	0.49	0.47
FC_c	0.057	0.29	0.57	0.55	0.05	0.56	0.62	0.061	0.08	0.15	0.00	0.24	0.49	0.47
FC_q	0.046	0.46	0.70	0.98	0.01	0.43	0.22	0.038	0.05	0.02	0.83	0.28	0.51	0.58
FC_{all}	0.046	0.54	0.75	1.00	0.04	0.35	0.63	0.049	0.88	0.82	0.16	0.32	0.68	0.67
5-day														
FC_u	0.020	0.01	0.01	0.35	0.22	0.00	0.00	0.058	0.59	0.31	0.00	0.23	0.71	0.72
FC_c	0.100	0.00	0.01	0.17	0.76	0.56	0.47	0.062	0.42	0.53	1.00	0.21	0.87	0.94
FC_q	0.016	0.00	0.02	0.97	0.46	0.54	0.42	0.041	0.51	0.53	0.99	0.20	0.59	0.63
FC_{all}	0.056	0.66	0.39	1.00	0.55	0.66	0.79	0.045	0.51	0.56	1.00	0.20	0.70	0.68



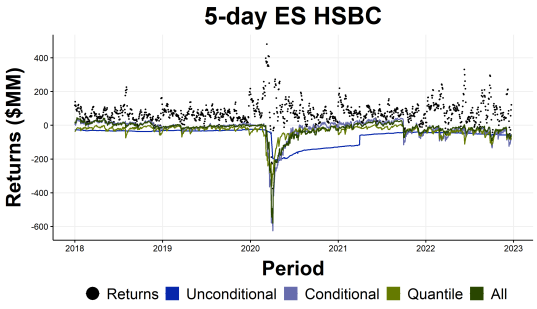
(a) 1-day VaR HSBC



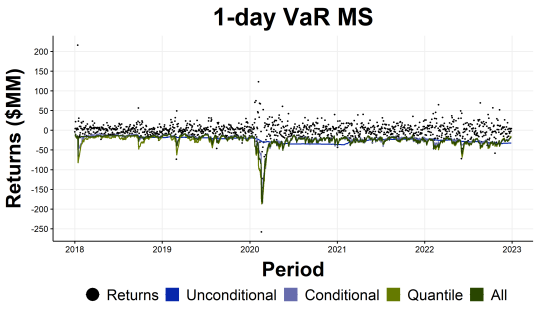
(b) 1-day ES HSBC



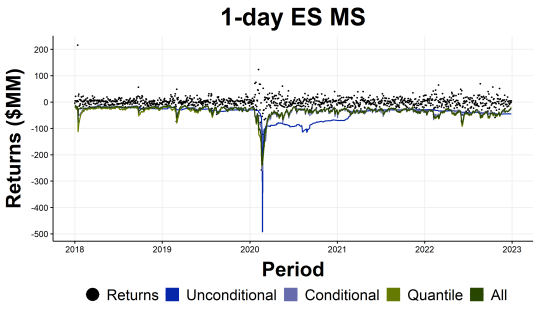
(c) 5-day VaR HSBC



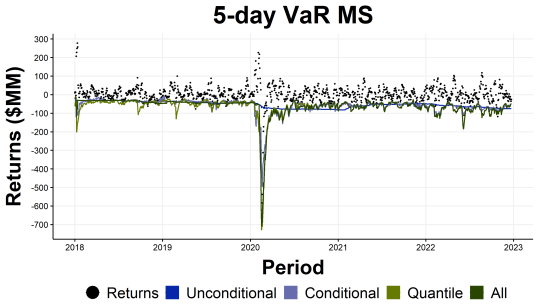
(d) 5-day ES HSBC



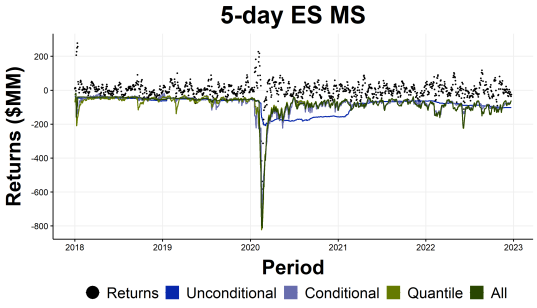
(e) 1-day VaR MS



(f) 1-day ES MS



(g) 5-day VaR MS



(h) 5-day ES MS

Figure 11: Forecast combinations VaR and ES for HSBC and MS.

Figure 12: Thanks for sitting through the thesis!