Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrics and Operations Research

Enhancing Value-at-Risk and Expected Shortfall Forecast Combinations: The Importance of Trimming

Akram Aykour (560202)

Supervisor:	B. van Os
Second assessor:	dr. A. Tetereva
Date final version:	02/07/2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

The value-at-risk (VaR) and expected shortfall (ES) are approved risk measures used by financial institutions around the globe. Hence, accurately forecasting these risk measures is of primary importance. Our empirical study consists of the returns and intraday ranges from five financial indices ranging from approximately 1999 to 2023. We evaluate the out-of-sample forecasting performance for the 1% and 5% one day ahead VaR and ES for a set of individual and combining methods such as the simple average, the minimum score, and relative score combinations. More importantly, we examine whether trimming the subpar methods from the forecast combinations based on the Model Confidence Set (MCS) of Hansen et al. (2011) improves forecast accuracy during periods of high volatility. The results reveal that trimming based on the MCS increases forecast accuracy often. Specifically, the simple average incorporating trimming is the most competitive forecasting method.

1 Introduction

Managing and evaluating risk is a crucial task in the financial industry, including banks and other financial institutions. The COVID-19 pandemic and the ongoing Russian-Ukrainian (2022) conflict have had notable effects on the volatility experienced by the financial world, see Baek et al. (2020) and Umar et al. (2022), respectively. Hence, the downward risk stemming from these events has bolded the importance of risk management. Accurate forecasts regarding risk in the financial markets can minimise the potential losses experienced by financial institutions. These accurate forecasts allow financial establishments to make justified decisions regarding determining and analyzing investment opportunities.

A popular measure of risk within the context of trading portfolios is the value-at-risk (VaR). The VaR is defined as the maximum loss a market portfolio can experience for a given confidence level over a specified time horizon. The VaR is used by bank regulators and pension plans to set capital requirements and measure the risk to which portfolios are exposed (Linsmeier & Pearson, 2000). However, the VaR is a non-subadditive measure that does not reveal any information regarding losses exceeding the quantile (Artzner et al., 1999). The expected shortfall (ES), a novel measure of risk, addresses these limitations as a subadditive measure conditional on the losses exceeding the VaR level (Yamai & Yoshiba, 2005). Accurate VaR and ES forecasts quantify the potential losses a financial institution can experience, hence allowing them to devise a mitigation strategy to minimise the downward risk.

A technique used to increase out-of-sample forecast accuracy is through employing combination methods. Specifically, even a simple combining approach like the equally weighted combination outperforms individual methods in accuracy (Lichtendahl Jr & Winkler, 2020).

Aiolfi et al. (2010) show that this naive combining method often outperforms linear and non-linear models in accurately forecasting a range of macroeconomic variables for several forecasting horizons. In financial risk management, forecast combinations also improve accuracy and are more robust to limitations faced by individual methods (Barrow & Kourentzes, 2016). There is not much literature regarding forecast combinations in the context of predicting both the VaR and ES. However, Taylor (2020) argues that the superiority of forecast combinations remains in this context.

The Model Confidence Set (MCS) proposed by (Hansen et al., 2011) statistically tests for the best forecasting methods. Hence, it allows us to perform forecast combinations with the methods with superior predictive ability. Incorporating trimming before combining models is common practice in the macroeconomic environment. Samuels & Sekkel (2017) show that utilising the MCS as a trimming method before averaging the forecasts results in a significant accuracy increase for forecasting macroeconomic indicators for the US. Utilising the MCS as a trimming technique to optimise the out-of-sample forecast combinations is new to the literature surrounding VaR and ES forecasting, to our knowledge.

This paper contributes to existing literature regarding forecast combinations in the scope of VaR and ES forecasting by extending the framework of Taylor (2020). Firstly, the data is extended to the start of 2023 to account for the recent crises experienced by the financial markets: the COVID-19 pandemic and the ongoing Russian-Ukrainian conflict. Furthermore, we include a global commodity index (S&P GSCI) alongside stock indices. We expand the data in this fashion to dictate the robustness of certain forecast combinations in periods of high volatility and other financial markets. In particular, the 2008 financial crisis caused commotion regarding the robustness of VaR forecasts (Halbleib & Pohlmeier, 2012). Finally, we employ the MCS as an initial trimming step before using the simple average, minimum score combining, and relative score combining methods. This additional step should reveal whether trimming based on the MCS should be common practice when investigating the out-of-sample performance of forecast combinations in risk management.

The results reveal that trimming based on the MCS before performing forecast combinations generally increases the forecast accuracy according to the out-of-sample scoring functions averaged over all five indices. In particular, the trimmed simple average produces the most accurate VaR and ES forecasts. The remainder of the paper is structured as follows. Section 2 presents the literature surrounding VaR and ES forecasts and corresponding forecast combinations. Section 3 shows the data utilised. The methodology is extensively presented in Section 4. The results and conclusions are in Sections 5 and 6, respectively.

2 Literature Review

Forecast combinations often increase forecasting accuracy compared to individual methods. Lichtendahl Jr & Winkler (2020) explain that the advantage of pooling forecasts is two-fold. Primarily, a diversity of predictions with low correlations amongst them increases accuracy. Secondly, forecast combinations for time series data are relatively robust, hence risk-reducing. Even simple forecast combinations like the equally-weighted technique produce precise results known as the 'equal weights puzzle' (Diebold & Shin, 2019). An extensive number of forecast combinations are present for the VaR metric. In particular, Bayer (2018) proposes a penalized quantile regression approach to combine VaR forecasts for stocks. The quantile regression approach involves regularization through shrinkage methods to reduce the multicollinearity amongst the predictions. This approach has superior forecasting ability relative to individual methods. Chiu et al. (2010) investigate the forecasting ability of linear combination techniques for the VaR for crude oil and Brent prices. The results show that linearly combining competing models often outmatch individual models at a VaR level of 1% and 5%. Finally, Halbleib & Pohlmeier (2012) improve the VaR forecasts by optimally combining forecasts for periods of high volatility. The first technique optimises weights based on different VaR evaluation schemes. The second technique imposes quantile regression on individual VaR forecasts. The results reveal that the proposed methods are robust and accurate for times of recession. Forecast combinations are not well-explored in the field of ES due to the ES' non-elicitable nature. Meaning that there is no appropriate loss function to evaluate the out-of-sample ES forecasts. Although, the elictability of the ES appears when considering the VaR and ES jointly. Fissler et al. (2015) introduce scoring functions to evaluate the VaR and ES forecasts together.

Taylor (2020) investigates the importance of combining techniques such as the equally-weighted and two different weight optimising techniques for accurately forecasting the VaR and the ES. The empirical study consists of five stock indices ranging from 1993 to 2017. Taylor (2020) uses a variety of scoring functions based on the form proposed by Fissler et al. (2015) to evaluate the out-of-sample forecasting performance of the methods. Additionally, different calibration tests are also employed to evaluate the forecasts. The individual models are uncompetitive relative to the forecast combinations. However, constructing forecast combinations faces a crucial consideration, namely whether we should include each method (Diebold & Shin, 2019). Including forecasts of models with inferior predictive accuracy in a forecast combination could degrade its out-of-sample performance (Lichtendahl Jr & Winkler, 2020).

The MCS proposed by Diebold & Shin (2019) statistically tests the out-of-sample performance of methods and constructs a confidence set involving the superior predictive models. The use of the MCS as an evaluation measure is well explored in the context of VaR and ES forecasts [see for example (Bernardi & Catania, 2016), (Taylor, 2020), (Luo et al., 2017)]. Trimming based on the MCS is well-versed in the context of macroeconomic predictions. Shang & Haberman (2018) average the forecasts of Japanese mortality rates selected by the MCS. Garcia et al. (2017) employ different Machine Learning (ML) methods to forecast Brazilian inflation for many different horizons. An equally weighted combination based on the MCS with a confidence level of 80% is the best out-of-sample forecasting method compared to the best individual model and a simple average of all models. Shang & Haberman (2018) average the model forecasts selected by the MCS at a 90% confidence level based on the out-of-sample root mean squared forecast error. This equally weighted combination is competitive in terms of out-of-sample accuracy. Finally, Amendola et al. (2020) show that the MCS averaged trimming technique achieves the optimal out-of-sample forecasts for multivariate volatility of U.S stocks.

Utilising the MCS as a trimming method before combining is not explored yet to our knowledge for VaR and ES forecasting. However, Happersberger (2021) utilises the partially egalitarian LASSO (peLASSO) proposed by Diebold & Shin (2019) as a combining technique of VaR and ES forecasts for 12 equity markets. The peLASSO method involves two steps that coincide with the simple average combination post trimming. Firstly, a shrinkage method such as LASSO selects the optimal forecasts based on an out-of-sample scoring function. In the second step, the weights of the surviving predictions are shrunk towards equality. The peLASSO method has accurate out-of-sample forecasting performance for the VaR and ES and outperforms other combining techniques. Additionally, Taylor (2020) also performs VaR and ES forecast combinations while discarding the historical simulation method due to its subpar performance. The results indicate that trimming the underperforming historical simulation from the forecast combinations increases the out-of-sample accuracy.

3 Data

In this empirical study, we focus on forecasting the one day ahead 1% and 5% VaR and ES. We consider the log-returns of the S&P 500 (US), CAC 40 (France), FTSE 100 (UK), DAX 30 (Germany), and NIKKEI 225 (Japan) stock indices, as well as a global commodity index (S&P GSCI). The S&P GSCI includes many diversified commodities such as precious metals and oils (McGlone & Gunzberg, 2011).

The starting dates are the 25th of June 1999, the 17th of November 1999, the 30th of July 1999, the 6th of November 1998, and the 6th of July 1999, respectively. The sample for the stock and commodity indices ends on the 28th of April 2023. This results in a sample size of

6000 observations for each index. The difference in starting dates among the indices is due to the different holidays in each respective country. We download the daily closing, maximum, and minimum prices from the Bloomberg terminal.

The daily log returns for period t are calculated as follows $r_t = ln(\frac{P_t}{P_{t-1}})$, where P_t denotes the price in period t. The intraday range IR_t is the log difference between the closed daily high P_t^{high} and daily low P_t^{low}. We use the initial 2000 observations for estimating the individual methods. The following 2000 out-of-sample forecasts determine the weights for the forecast combinations. We employ a rolling window approach with window size $RS = 2000$ days moving one trading year ahead (252 days) for the estimation procedures to reduce computation time. Finally, we evaluate the forecasts based on the last 2000 observations to determine how individual and combining methods perform during periods of high volatility e.g. COVID-19 and the Russian-Ukrainian conflict. After the evaluation of the initial (combining) methods, we reconstruct the combining methods based on the models selected by the MCS.

Figure 1: The log returns of the S&P GSCI commodity index over the last two decades.

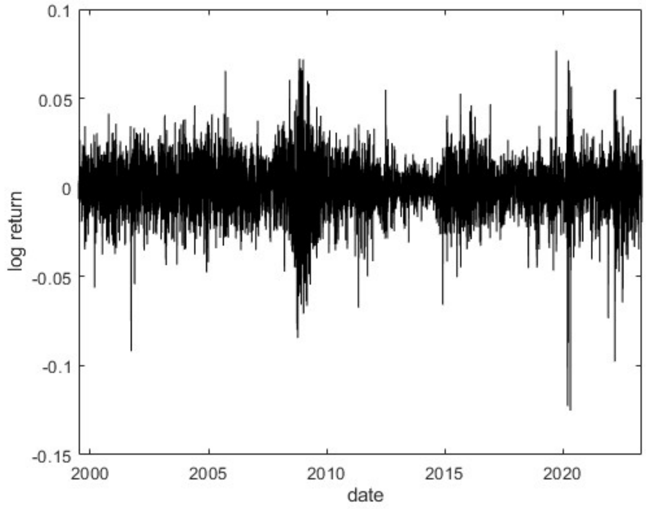

Figure 1 depicts the log returns of the S&P GSCI commodity index over the entire sample range. The Figure shows that the financial crisis of 2008 and the COVID-19 pandemic caused high volatility. These large volatility spikes remain visible in 2022 possibly due to the Russian-Ukrainian conflict. This ongoing conflict impacts numerous commodities and their respective markets (Alam et al., 2022). It is interesting to note that the COVID-19 pandemic caused larger negative returns relative to the financial market crisis of 2008 experienced by the commodity market.

Before the estimation procedure of the individual methods, we employ an AR(1) as a filter, for which we use a similar rolling window approach with RS 2000 as done in (Taylor, 2020).

4 Methodology

We follow the same methodological framework as in Taylor (2020). We first present the individual and combining methods. Subsequently, we give an overview of the evaluation criteria used to evaluate the VaR and ES forecasts.

4.1 Individual Methods

Methods forecasting the VaR are generally (semi)parametric or non-parametric (Engle & Manganelli, 2004). We use all three types in this study, with the addition of a method that incorporates intraday ranges for robustness. A variety of methods could assist the performance of forecast combinations as they process information differently (Happersberger, 2021).

The expression for the out-of-sample VaR and ES forecasts for $t+1$ for models $m = 1, ..., 5$ are shown in Equations (1) and (2), respectively. [1]

$$VaR_{m,t+1|t}(\alpha) = Q_{\alpha}(r_{t+1}|I_t), \tag{1}$$

where I_t represents the available information set containing information up till time t and Q_{α} denotes the quantile function with respect to the probability level α.

$$ES_{m,t+1|t}(\alpha) = \mathbb{E}(r_{t+1}|r_{t+1} \leq VaR_{t+1|t}, I_t). \tag{2}$$

We forecast the VaR and ES one day ahead, with probability levels 1% and 5%.

4.1.1 Historical Simulation

The historical simulation is a non-parametric method commonly used for forecasting quantiles. We use the historical simulation based on 250 observations as our benchmark model. We opt for 250 observations as Taylor (2020) reveals that a different number of observations does not lead to improved accuracy.

4.1.2 GJR-GARCH

The GARCH model and its respective adoptions are well known in the context of forecasting financial variables. Ergün & Jun (2010) investigate the predictive performance of numerous GARCH models for the VaR and ES for future indices of the S&P 500. The GARCH-based models have accurate out-of-sample VaR forecasts. The GJR-GARCH(1,1) model is an extension of the GARCH model often found to perform well in the context of VaR forecasting (Su et al.,

[1] We utilise the same notation for the VaR and ES forecasts as done in (Happersberger, 2021).

2011). The performance of the GJR-GARCH(1,1) model resonates with its ability to analyze the impact of negative and positive returns on conditional volatility. We opt for a GJR-GARCH(1,1) based on the student t-distribution. The expression of the GJR-GARCH(1,1) is in Appendix B.

4.1.3 CAViaR-AS-EVT

The conditional autoregressive value at risk (CAViaR) model proposed by Engle & Manganelli (2004) is an autoregressive model which utilises quantile regression for estimation procedures. The CAViaR model is commonly used in existing VaR forecasting literature due to its reputable predictive ability [see for example (Jeon & Taylor, 2013), (Huang et al., 2009), (Chen et al., 2012)]. To produce ES forecasts the CAViaR model is extended by incorporating the peaks-over-threshold extreme value theorem (EVT) to the standardised returns exceeding the VaR level. [2] Models incorporating EVT are found to produce accurate VaR forecasts (Ergün & Jun, 2010).

The CAViaR-AS-EVT is expressed as

$$Q_t = \beta_0 + \beta_1 1(r_{t-1} > 0)|r_{t-1}| + \beta_2 1(r_{t-1} \leq 0)|r_{t-1}| + \beta_3 Q_{t-1},$$

where Q_t denotes the quantile at time t. We extend the CAViaR model further by implementing the asymmetric slope (AS) to account for the leverage effect as done for the GJR-GARCH(1,1) model.

4.1.4 CARE-AS

Taylor (2008) introduces the conditional autoregressive expectile (CARE) to estimate the VaR and ES. The advantage of this method is its ability to avoid distributional assumptions. The estimation of τ is done as in the framework by Taylor (2008). The model is re-estimated continuously using different values of τ using step sizes of 0.0001 for the observations in the first rolling window. Based on experimentation performed by Taylor (2020), we set τ equal to 0.0018 and 0.0016 for α equal to 1% and 5%, respectively. The expression for the CARE-AS model is given in Equation (3).

$$\mu_t = \beta_0 + \beta_1 I(r_{t-1} > 0)|r_{t-1}| + \beta_2 (r_{t-1} \leq 0)|r_{t-1}| + \beta_3 \mu_{t-1}, \tag{3}$$

where μ_t is the expectile and where τ is used to estimate the quantile α. We employ the same estimation procedure as for the CAViaR-AS-EVT. However, we consider the following expectile scoring function:

[2] See (Taylor, 2020) for an extensive overview of the forecasting procedure using the CAViaR-EVT method.

$$S(\mu_t, r_t) = |\tau - I(r_t \leq \mu_t)|(r_t - \mu_t)^2.$$

4.1.5 HAR-RANGE

The heterogeneous autoregressive (HAR) is a dynamic model often used to forecast volatility in financial markets (Corsi & Reno, 2009). The desired VaR and ES forecasts are computed by multiplying the HAR-RANGE volatility forecasts by the standard deviation of the VaR and ES estimates based on the student t - distribution.

The daily range IR_t is then regressed on an intercept, the weekly range $IR^w_{t-1} = \frac{1}{5}\sum_{i=1}^{5} IR_{t-i}$, the monthly range $IR^m_{t-1} = \frac{1}{22}\sum_{i=1}^{22} IR_{t-i}$ and an error term ϵ_t i.i.d with zero mean as shown in Equation (4).

$$IR_t = \beta_1 + \beta_2 IR_{t-1} + \beta_3 IR^w_{t-1} + \beta_4 IR^m_{t-1} + \epsilon_t, \tag{4}$$

$$\sigma^2_t = \delta_1 + \delta_2 IR^2_t. \tag{5}$$

We estimate the parameters B_i in Equation (4) using least squares (LS). Furthermore, the coefficients for the conditional variance shown in Equation (5) are estimated using maximum likelihood estimation (MLE) based on the student t-distribution.

4.2 Combining Techniques

We use three different combining techniques to construct the forecast combinations. The first technique is the naive simple average that averages all method forecasts. The latter two allocate weights based on optimising in-sample scoring functions.

4.2.1 Minimum Score Combining

The proposed individual methods may have different forecasting qualities for the VaR and ES individually. However, allowing the weights of a combining method to differ is impractical because the ES is conditional on the VaR. Hence, we consider the minimum score combining method with difference spacing for the VaR and ES shown in the following lines.

$$VaR_{c,t+1|t} = \sum_{m=1}^{M} w^v_i VaR_{m,t+1|t}, \tag{6}$$

$$ES_{c,t+1|t} = VaR_{c,t+1|t} + \sum_{m=1}^{M} w_i^s (ES_{m,t+1|t} - VaR_{m,t+1|t}). \tag{7}$$

The weights w_i^v and w_i^s representing the VaR and ES weights, respectively, are both ensured to be non-negative and $\sum_{m=1}^{M} w_i = 1$. The optimal weights are determined by the in-sample minimum AL score. [3]

4.2.2 Relative Score Combining

Shan & Yang (2009) extend on literature regarding quantile forecasting combinations based on quantile loss functions. In which the weight of the candidate model i is shown in the following line,

$$w_i = \frac{exp(-\lambda \sum_{j=1}^{t-1} S(V_{ij}, E_{ij}, r_j)}{\sum_{k=1}^{M} exp(-\lambda \sum_{j=1}^{t-1} S(V_{kj}, E_{kj}, r_j)}. \tag{8}$$

The tuning parameter λ determines the correlation magnitude between the allocated weights and the chosen scoring function S. We tune the parameter λ by minimising the in-sample AL scoring function. The VaR and ES relative score combinations are presented in Equations (9) and (10), respectively.

$$VaR_{c,t+1|t} = \sum_{m=1}^{M} w_i VaR_{m,t+1|t}, \tag{9}$$

$$ES_{c,t+1|t} = \sum_{m=1}^{M} w_i ES_{m,t+1|t}. \tag{10}$$

4.3 Evaluation

We consider multiple scoring functions, calibration tests, and the MCS to evaluate the VaR and ES forecasts. Furthermore, we construct newly trimmed forecast combinations based on the MCS.

4.3.1 Scoring functions

The VaR is an elicitable risk measure for which the scoring functions take the following form,

$$S_V(v, r) = (\mathbf{1}[r \leq v] - \alpha)(G(v) - G(r)). \tag{11}$$

[3] The scoring functions are presented in Subsection 4.3.

In which v, r, and α denote the VaR estimate, the return, and the probability level, respectively. According to Taylor (2020), when the function G is set equal to the identity function (I), Equation 11 becomes the quantile score (QS). The quantile score evaluates the out-of-sample performance of the VaR forecasts at the 1% and 5% probability levels. As stated in Section 2, unlike the VaR, the ES is not an elicitable risk measure. Hence, we refer to the loss function form proposed by Fissler et al. (2015) in Equation 12 to jointly evaluate the VaR and ES forecasts.

$$S_{V,E}(v,e,r) = (\mathbf{1}[r \leq v] - \alpha)G_1(v) - \mathbf{1}[r \leq v]G_1(r) + G_2(e)(e - v + \mathbf{1}[r \leq v]\frac{(v-r)}{\alpha}) - \zeta_2(e) + a(r). \tag{12}$$

The considered G_1 and G_2 functions are strictly increasing and continuously differentiable. Additionally, ζ_2' is equal to G_2. Additional conditions that need to be satisfied can be found in Fissler et al. (2015). Finally, the ES estimate is denoted as (e). Based on Equation (12), several loss functions that satisfy the necessary properties have become available. These functions are in Table 1. [4]

	$G_1(x)$	$G_2(x)$	$\zeta_2(x)$	$a(r)$
AL	0	$-\frac{1}{x}$	-$ln(-x)$	1 - $ln(1-\alpha)$
NZ	0	$\frac{1}{2}(-x)^{-\frac{1}{2}}$	$-(-x)^{-\frac{1}{2}}$	0
FZG	x	$\frac{exp(x)}{(1+exp(x))}$	$ln(1+exp(x))$	$ln(2)$
AS	$-\frac{1}{2}Wx^2$	αx	$\frac{1}{2}\alpha x^2$	0

Table 1: The different joint scoring functions for the VaR and ES forecasts based on Equation (12).

The AS scoring function proposed by Acerbi & Szekely (2014) needs to satisfy $v \cdot W < e$. We set W = 4 as done in Taylor (2020) to ensure this requirement is satisfied.

4.3.2 Calibration tests

The VaR forecasts are generally assessed based on the conditional and unconditional calibration tests (Nolde & Ziegel, 2017). We perform these backtests at a 5% significance level. The VaR forecast of model i is unconditionally calibrated if Equation (13) is satisfied and conditionally calibrated if the conditional expectation of HIT_{t+1} is zero.

$$E(HIT_{t+1}) = E(\alpha - \mathbf{I}(r_{t+1} \leq VAR_{m,t+1}) = 0. \tag{13}$$

We use a test founded on the binomial distribution to test whether the mean of HIT_{t+1} is

[4] An overview of these individual scoring functions can be found in Taylor (2020)

significantly different from 0. Additionally, we employ the dynamic quantile (DQ) test proposed by Engle & Manganelli (2004) as done in Le (2020) to test whether the conditional mean of HIT_{t+1} is significantly different from 0. We include four lags in the DQ testing procedure.

The ES forecast of model i is unconditionally calibrated if the discrepancy between the return and the ES forecast for the period $t+1$ is zero, given that the return exceeds the corresponding VaR in that period (McNeil & Frey, 2000). We standardise the test by dividing the expected discrepancies by the VaR estimate. Similarly to (Taylor, 2020), we use a dependent circular block bootstrap to avoid any distributional assumptions of the standardised deviations.

4.3.3 Model Confidence Set

We use the MCS proposed by Hansen et al. (2011) as a final evaluation approach. This approach allows us to statistically detect the worst performing model and remove them from the original confidence set containing all models M^0. The set of superior models is defined as

$$M^* = \{i \in M^0 : \mu_{ij} \leq 0, \forall j \in M^0\}, \tag{14}$$

where μ_{ij} is defined as the expected loss difference between method i and j. We first apply the equivalence test δ_M based on the Diebold-Mariano test to M^0. If δ_M is rejected it indicates that the predictive performance of the models in the set is not equivalent. Subsequently, we utilise the elimination rule e_M to eliminate the worst-performing models within the set. We repeat this process until δ_M is accepted, and the set $M*$ consisting of the models with superior predictive ability is obtained.

The implementation of the MCS in this paper is two-fold. Firstly, we use it to statistically evaluate and compare the predictive ability of the individual and combining methods. Secondly, we construct forecast combinations with the superior individual models based on the MCS. We follow the same framework as Hansen et al. (2011), opting for a confidence level of 90% and 75%, and the number of bootstrap re-sampling B is 10,000. The MCS trimmed combinations are based on the out-of-sample AL score in order to stay consistent with the combining methods, for which the weights are optimised based on the AL score.

Taylor (2020) considers discarding the historical method forecast from the combining methods. The forecast combinations involving trimming are generally more accurate for the 1% and 5% VaR and the ES. This gain in accuracy is because including a subpar-performing model in a forecast combination is more likely to harm than contribute to its precision (Lichtendahl Jr & Winkler, 2020).

5 Results

The index and scoring function-specific results are presented only for the S&P GSCI commodity index and the AL score to save space. The results regarding the different stock indices and scoring functions are found in Appendix C.

Table 2: The model confidence set at a confidence level of 90% for all scoring functions for the S&P GSCI.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	1	1	1	1	1	1	0	1	1	1
HAR-RANGE	1	1	1	1	1	1	1	1	1	1
CARE-AS	0	0	0	0	0	0	0	0	0	0
CAViaR-AS-EVT	1	1	1	1	1	1	1	1	1	1
Combinations										
Simple Average	0	0	0	0	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	1	1	1	1	1	1

Note: A 1 indicates that the method is included in the MCS for that specific scoring function for the S&P GSCI index.

Table 2 reveals the models included in the MCS with a confidence level of 90% for the S&P GSCI index. The individual methods within the MCS for the 1% probability level are constant over the different scoring functions. Namely, the GJR-GARCH, HAR-RANGE, and CAViaR methods have the superior predictive ability (SPA). Furthermore, the relative and minimum score combinations methods in the MCS for all scoring functions over the different probability levels. The simple average does not forecast the VaR and ES well for the 1% probability level relative to the 5% level. This dissimilarity is because the 1% VaR and ES surround extreme events that correspond with more uncertainty, hence, it is more complicated to forecast accurately.

Table 3: The model confidence set at a confidence level of 90% for all scoring functions aggregated over all indices.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	1	0	0	0	0	0
GJR-GARCH	3	3	3	3	4	4	3	4	4	4
HAR-RANGE	5	5	5	5	4	5	4	4	4	5
CARE-AS	0	0	0	0	0	2	1	2	2	3
CAViaR-AS-EVT	4	4	4	4	4	4	4	4	4	4
Combinations										
Simple Average	3	3	3	3	4	4	4	4	4	4
Relative Score	5	5	5	5	4	5	5	5	4	5
Minimum Score	5	5	5	5	5	4	4	4	4	4

Note: The values depict the number of indices for which the method is included in the MCS at a 90% confidence level. Higher values indicate better performance, where 5 is the highest attainable value.

Table 3 reflects the MCS at a confidence level of 90% aggregated over all indices. The value reveals the number of times the method is in the MCS over the five indices. Hence, a large number suggests that the predictive ability of the method is robust and superior. In particular, the HAR-RANGE is included in the MCS for almost five indices for both probability levels, indicating accurate predictive performance. The accuracy of the HAR-RANGE forecasts is most likely due to the incorporation of high-frequency intraday data relative to solely using the daily returns such as the other individual models. Besides the performance of the HAR-RANGE, the CAViaR and GJR-GARCH also perform relatively well. The relative and minimum score combining methods also have SPA for generally all indices for both probability levels. The results of this Table are in line with Table 2 as both the CARE and historical simulation models are uncompetitive.

Table 4 shows the MCS at a 75% confidence level for all indices. The results do not change drastically compared to Table 3, as both tables show that the CARE-AS and historical simulation models are subpar at forecasting the VaR and ES at both probability levels.

Table 4: The model confidence set at a confidence level of 75% for all scoring functions aggregated over all indices.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	3	3	3	3	3	3	2	2	3	4
HAR-RANGE	4	5	5	4	3	4	4	4	4	5
CARE-AS	0	0	0	0	0	2	1	1	1	1
CAViaR-AS-EVT	3	4	4	3	3	4	4	4	4	4
Combinations										
Simple Average	3	3	3	3	3	4	4	4	4	4
Relative Score	3	4	4	3	3	4	4	5	4	4
Minimum Score	4	3	4	4	4	3	3	3	3	4

Note: The values depict the number of indices for which the method is included in the MCS at a 75% confidence level. Higher values indicate better performance, where 5 is the highest attainable value.

The trimmed forecast combinations include the models based on the 75% and 90% MCS for the AL scoring function for the 1% probability level. We regard the 1% VaR and ES as no methods are included in the 75% MCS for the S&P 500 index for the AL scoring function at the 5% probability level as seen in Table 14 in Appendix C. Furthermore, it is interesting to note that the MCS results are consistent for the AL score for the 1% probability level between the 75% and 90% confidence levels of the MCS. The trimmed forecast combinations of the S&P GSCI, CAC 40, and NIKKEI 225 include the GJR-GARCH, HAR-RANGE, and CAViaR methods. The HAR-RANGE and CAViaR models are in the trimmed forecast combinations for the FTSE 100 index. Finally, the trimmed combinations of the S&P 500 index are equal to that of the HAR-RANGE method. These results can be found in Appendix C.

The calibration test results for the S&P GCSI commodity index are placed in Table 5. A non-zero entry indicates that the corresponding calibration test is rejected at a 5% significance level. Hence, lower values are preferred. The results reveal that the forecasts of the CAViaR and HAR-RANGE methods are the most calibrated relative to the other methods. This finding is in line with the MCS results, in which the above-mentioned methods generally have SPA. Furthermore, the results reveal that the VaR and ES forecasts of the CARE-AS model are not (un)conditionally calibrated for either probability level.

Table 5: Results of the calibration tests for the S&P GCSI commodity index.

	1% probability level			5% probability level		
	VaR hit	VaR DQ	ES bootstrap	VaR hit	VaR DQ	ES bootstrap
Individual Methods						
Historical Simulation	0	1	0	0	1	0
GJR-GARCH	1	0	0	0	0	1
HAR-RANGE	1	0	0	0	0	0
CARE-AS	1	1	1	1	1	1
CAViaR-AS-EVT	0	0	0	0	0	1
Combinations						
Simple Average	1	0	0	0	0	1
Relative Score	1	0	1	0	0	1
Minimum Score	1	1	0	0	0	1
Trimmed Combinations						
Simple Average	1	0	0	0	0	1
Relative Score	1	0	0	1	0	1
Minimum Score	1	0	0	0	0	1

Note: A 1 depicts that the corresponding calibration test is rejected for the method at a 5% significance level.

Table 6 depicts the results of the calibration tests over all five indices. The VaR hit test at the 1% probability level for the trimmed forecast combinations is rejected only for the S&P GSCI index. Indicating that the 1% VaR forecasts for the trimmed combining methods are not calibrated unconditionally for this commodity index. Moreover, the same applies to the trimmed combinations for the 5% ES, where the ES bootstrap is rejected only for the S&P GSCI index. Meaning that the expected unconditional ES discrepancies to be zero is rejected. The trimmed and untrimmed combination forecasts do not differ tremendously in calibration. Taylor (2020) performs the calibration tests for both the 1% and 5% probability levels, for which the tests are rejected mainly for the historical simulation. In our case, rejection at either probability level occurs substantially more for all methods. This difference in calibrated forecasts could be due to the lack of estimation as we repeatedly move the rolling window one year ahead. Additionally, as seen in Figure 1, the data used to evaluate the forecasts is very volatile relative to the data used to estimate the methods and the combining weights.

Table 6: Results of the calibration tests aggregated over all five indices.

	1% probability level			5% probability level		
	VaR hit	VaR DQ	ES bootstrap	VaR hit	VaR DQ	ES bootstrap
Individual Methods						
Historical Simulation	3	5	2	1	5	2
GJR-GARCH	3	2	2	3	3	3
HAR-RANGE	2	2	0	1	2	2
CARE-AS	5	5	5	5	5	5
CAViaR-AS-EVT	0	1	1	0	1	1
Combinations						
Simple Average	2	1	1	0	2	1
Relative Score	1	1	3	1	2	1
Minimum Score	2	4	0	1	1	1
Trimmed Combinations						
Simple Average	1	1	0	1	2	1
Relative Score	1	2	1	2	2	1
Minimum Score	1	2	0	1	2	1

Note: The values depict the number of indices for which the calibration tests are rejected at a 5% significance level. Lower values are preferred.

The 1% VaR and ES evaluated using the AL skill score (%) is seen in Table 7. The AL and FZG skill scores are computed by first taking the ratio of the score of model i and that of the historical simulation forecast. We then subtract one from this ratio before multiplying it by 100. Moreover, the QS, NZ, and AZ scores take positive values, hence, we subtract the ratio from one before multiplying it by 100.

Table 7 shows that the most competitive methods are the HAR-RANGE and the (trimmed) combinations as they attain the highest AL skill score (bolded) for the 1% probability level. However, discarding forecasts before combining does not always lead to improvement. Specifically, the trimmed combinations perform worse than their untrimmed counterparts for the NIKKEI225 and FTSE 100 indices. This deterioration in performance could be because every method processes and captures information differently and can hence still contribute to the accuracy of the forecast combinations (Happersberger, 2021). Nevertheless, when regarding the geometrical mean across all indices, trimming before combining does increase predictive accuracy for both the simple average and minimum score combining. The forecast of the relative score combination being better before discarding inaccurate forecasts could be because the weights assigned to the surviving models depend on the in-sample AL score. This indicates a possible fluctuation in the in-sample and out-of-sample model predictive performance.

Table 7: 1% VaR and ES evaluated using the AL skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	10.486	9.691	12.194	12.825	6.716	10.382
HAR-RANGE	**30.104**	10.158	13.320	13.158	**9.688**	15.286
CARE-AS	-84.028	-70.388	-41.441	-38.974	-97.908	-66.548
CAViaR-AS-EVT	8.021	11.665	12.870	11.592	6.862	10.202
Combinations						
Simple Average	26.563	10.696	**15.187**	**14.224**	5.211	14.376
Relative Score	**30.104**	11.665	**15.187**	**14.224**	7.927	15.821
Minimum Score	29.861	11.630	14.607	12.258	6.752	15.022
Trimmed Combinations						
Simple Average	**30.104**	**11.809**	14.447	13.491	8.771	**15.724**
Relative Score	**30.104**	11.450	13.546	12.358	8.404	15.172
Minimum Score	**30.104**	11.342	13.610	12.625	8.294	15.195

Note: The largest values are bolded and correspond to the most accurate method(s) for each column (index).

Table 8: 5% VaR and ES evaluated using the AL skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	0.294	4.312	6.320	4.472	2.316	3.543
HAR-RANGE	**9.394**	3.613	6.348	4.028	**3.279**	5.332
CARE-AS	-3.042	2.535	5.221	3.750	-0.632	1.566
CAViaR-AS-EVT	-0.854	4.575	6.513	4.611	2.497	3.468
Combinations						
Simple Average	3.576	**5.041**	7.200	**5.000**	2.858	4.735
Relative Score	**9.394**	4.837	6.815	4.861	2.677	5.717
Minimum Score	8.700	4.254	6.843	4.833	2.738	5.474
Trimmed Combinations						
Simple Average	**9.394**	4.808	7.227	4.917	3.069	**5.883**
Relative Score	**9.394**	4.808	**7.255**	4.694	2.647	5.760
Minimum Score	**9.394**	4.662	7.172	4.861	2.677	5.753

Note: The largest values are bolded and correspond to the most accurate method(s) for each column (index).

The 5% VaR and ES forecasts evaluated by the AL skill score in Table 8 have similar results to the 1% probability level, as trimming does not always lead to improved predictive ability for all indices. However, according to the geometrical mean, the trimmed forecast combinations perform better than their untrimmed counterparts. In particular, the simple average with trimming has the most precise out-of-sample accuracy for both probability levels. Furthermore, it is interesting to note that the forecast of the CARE-AS method visibly underperforms relative to the benchmark HS method for the 1% VaR and ES. However, for the 5% VaR and ES, the CARE model produces more accurate forecasts than the benchmark method. Note, recall that the trimmed forecasts of the S&P 500 are the same as that of the HAR-RANGE method.

Table 9: VaR and ES evaluated using the skill scores (%) for all indices.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	21.033	10.382	12.530	0.840	33.498	12.210	3.543	7.573	0.325	21.168
HAR-RANGE	26.748	15.286	17.063	1.020	40.792	14.737	5.332	9.677	**0.414**	21.211
CARE-AS	-37.957	-66.548	-43.191	-1.625	-13.736	10.324	1.566	5.865	0.296	17.043
CAViaR-AS-EVT	21.780	10.202	2.864	0.512	35.885	12.513	3.468	7.883	0.325	19.693
Combinations										
Simple Average	23.003	14.376	14.884	0.900	36.127	13.713	4.735	8.802	0.355	20.266
Relative Score	27.189	**15.821**	**17.572**	1.020	40.275	15.930	5.717	10.270	**0.414**	23.150
Minimum Score	25.774	15.022	16.617	0.960	37.960	15.245	5.474	10.156	**0.414**	21.978
Trimmed Combinations										
Simple Average	27.361	15.724	15.444	**1.050**	**41.817**	**16.185**	**5.883**	**10.498**	**0.414**	**23.349**
Relative Score	26.569	15.172	16.816	0.990	40.797	16.023	5.760	10.392	**0.414**	23.061
Minimum Score	**27.736**	15.195	16.908	1.020	40.848	16.026	5.753	10.392	**0.414**	23.184

Note: The values are the geometrical mean skill scores over the five indices. The bolded values represent the most accurate method(s) for each column (skill score).

Finally, Table 9 reveals the geometric mean over each index for all scoring functions. The results indicate that the trimmed forecast combinations almost invariably achieve the highest out-of-sample scoring functions, except for the AL and NZ skill score for the 1% probability level. In particular, the MCS-trimmed simple average method performs the best. The equally weighted combination gains the most out-of-sample accuracy from trimming relative to the other combining methods. This difference in accuracy gained from trimming is because the relative score and minimum score methods assign weights to the methods based on the in-sample AL score. Hence, these combining methods already account for the performance of the individual models in the in-sample period. Therefore, the improvement in discarding the least accurate forecasts does not result in an enormous gain in predictive accuracy. Moreover, contrary to the simple average and minimum score combing method, the relative score combining deteriorates when incorporating trimming when considering the 1% VaR and ES.

Ultimately, the MCS reveals that the HAR-RANGE, GJR-GARCH, and CAViaR methods generally have SPA among the combinations. Furthermore, the CARE-AS method demonstrates significant limitations in forecasting the VaR and ES at both probability levels. This finding contradicts that of Taylor (2020), for which the CARE-AS model is the best-performing individual method for the 5% probability level. Furthermore, for both the 1% and 5% probability levels, the trimmed forecast combinations have the best predictive accuracy when considering the geometrical mean. Trimming forecasts based on the MCS before combining increases the out-of-sample forecast accuracy of the combining methods. However, the change in accuracy is dependent on the index and the probability level, indicating that trimming prior to combining does not always lead to better predictive ability.

6 Conclusion

This paper investigates the out-of-sample forecast accuracy gained when discarding non-competitive methods based on the MCS from forecast combinations. In particular, we evaluated the forecasting performance of individual models and (trimmed) forecast combinations during periods of high volatility caused by the COVID-19 pandemic. The MCS revealed that forecast combinations have superior predictive ability. However, the HAR-RANGE and CAViaR-AS-EVT methods are very competitive in producing accurate out-of-sample forecasts. The calibration tests showed that the CAViaR-AS-EVT method followed by the trimmed forecast combinations are rejected the least amount of times at a 5% significance level. The geometrical mean of the out-of-sample skill scores showed that discarding uncompetitive forecasts before combining methods increases the out-of-sample scoring functions. In particular, the equally-weighted forecast combination performs best. However, the advantages of trimming are not as apparent when regarding the indices individually.

The main limitation of this present study is the rolling window approach utilised to estimate the individual methods and the weights of the combining methods. Iteratively moving the rolling window one year ahead for the estimation procedure could lead to the method parameters not being optimised compared to moving the window by one day. This estimation procedure could explain why the CARE-AS method is subpar at predicting in our present study. Furthermore, the gains achieved by the trimmed forecast combinations could be due to look-ahead bias. To elude this bias, we can split the out-of-sample forecasting sample into two sub-samples as done in Garcia et al. (2017). The first sub-sample constructs the MCS at the two confidence levels. The second sub-sample evaluates and estimates the trimmed forecast combinations. Alternatively, we can base the MCS on the in-sample scoring functions.

With the accurate performance of the HAR-RANGE method and the increased availability of high-frequency data, we can include more models in our framework. For example, a novel approach that incorporates intraday returns proposed by Meng & Taylor (2020) can be used to extend this current paper. Including multiple competitive methods could further reveal how different MCS confidence levels could impact the forecast accuracy gained when trimming. Discarding many under-performing predictions from the combinations, known as aggressive trimming, substantially increases the out-of-sample accuracy (Timmermann, 2006). Moreover, it would be interesting to further investigate the forecasting performance during periods of high volatility. For example, we can employ the fluctuation test to evaluate the forecasting stability of the models. Finally, it would be engaging to compare the forecasting ability of the peLASSO against the trimmed simple average based on the MCS due to their similarity.

References

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

Aiolfi, M., Capistrán, C. & Timmermann, A. (2010). Forecast combinations. *CREATES research paper*(2010-21).

Alam, M. K., Tabash, M. I., Billah, M., Kumar, S. & Anagreh, S. (2022). The impacts of the russia–ukraine invasion on global markets and commodities: a dynamic connectedness among g7 and bric markets. *Journal of Risk and Financial Management*, *15*(8), 352.

Amendola, A., Braione, M., Candila, V. & Storti, G. (2020). A model confidence set approach to the combination of multivariate volatility forecasts. *International Journal of Forecasting*, *36*(3), 873–891.

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

Baek, S., Mohanty, S. K. & Glambosky, M. (2020). Covid-19 and stock market volatility: An industry level analysis. *Finance research letters*, *37*, 101748.

Barrow, D. K. & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: implications for inventory management. *International Journal of Production Economics*, *177*, 24–33.

Bayer, S. (2018). Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and statistics*, *8*, 56–77.

Bernardi, M. & Catania, L. (2016). Comparison of value-at-risk models using the mcs approach. *Computational Statistics*, *31*(2), 579–608.

Chen, C. W., Gerlach, R., Hwang, B. B. & McAleer, M. (2012). Forecasting value-at-risk using nonlinear regression quantiles and the intra-day range. *International Journal of Forecasting*, *28*(3), 557–574.

Chiu, Y.-C., Chuang, I.-Y. & Lai, J.-Y. (2010). The performance of composite forecast models of value-at-risk in the energy market. *Energy Economics*, *32*(2), 423–431.

Corsi, F. & Reno, R. (2009). Har volatility modelling with heterogeneous leverage and jumps. *Available at SSRN*, *1316953*.

Diebold, F. X. & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, *35*(4), 1679–1691.

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

Ergün, A. T. & Jun, J. (2010). Time-varying higher-order conditional moments and forecasting intraday var and expected shortfall. *The Quarterly Review of Economics and Finance*, *50*(3), 264–272.

Fissler, T., Ziegel, J. F. & Gneiting, T. (2015). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*.

Garcia, M. G., Medeiros, M. C. & Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, *33*(3), 679–693.

Halbleib, R. & Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, *36*(8), 1212–1228.

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

Happersberger, D. (2021). *Advancing systematic and factor investing strategies using alternative data and machine learning*. Lancaster University (United Kingdom).

Huang, D., Yu, B., Fabozzi, F. J. & Fukushima, M. (2009). Caviar-based forecast for oil price risk. *Energy Economics*, *31*(4), 511–518.

Jeon, J. & Taylor, J. W. (2013). Using caviar models with implied volatility for value-at-risk estimation. *Journal of Forecasting*, *32*(1), 62–74.

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

Lichtendahl Jr, K. C. & Winkler, R. L. (2020). Why do some combinations perform better than others? *International Journal of Forecasting*, *36*(1), 142–149.

Linsmeier, T. J. & Pearson, N. D. (2000). Value at risk. *Financial Analysts Journal*, *56*(2), 47–67.

Luo, L., Pairote, S. & Chatpatanasiri, R. (2017). Garch-type forecasting models for volatility of stock market and mcs test. *Communications in Statistics-Simulation and Computation*, *46*(7), 5303–5312.

McGlone, M. & Gunzberg, J. (2011). *Understanding commodities and the s&p gsci®* (Tech. Rep.). Working paper, Standard and Poors.

McNeil, A. J. & 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. & 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.

Nolde, N. & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation.

Nugroho, D., Kurniawati, D., Panjaitan, L., Kholil, Z., Susanto, B. & Sasongko, L. (2019). Empirical performance of garch, garch-m, gjr-garch and log-garch models for returns volatility. In *Journal of physics: Conference series* (Vol. 1307, p. 012003).

Samuels, J. D. & Sekkel, R. M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, *33*(1), 48–60.

Shan, K. & Yang, Y. (2009). Combining regression quantile estimators. *Statistica Sinica*, 1171–1191.

Shang, H. L. & Haberman, S. (2018). Model confidence sets and forecast combination: an application to age-specific mortality. *Genus*, *74*(1), 1–23.

Su, Y., Huang, H. & Lin, Y. (2011). Gjr-garch model in value-at-risk of financial holdings. *Applied Financial Economics*, *21*(24), 1819–1829.

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

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

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, *1*, 135–196.

Umar, Z., Polat, O., Choi, S.-Y. & Teplova, T. (2022). The impact of the russia-ukraine conflict on the connectedness of financial markets. *Finance Research Letters*, *48*, 102976.

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

Acknowledgements

I want to express my gratitude to Professor Taylor for sending the replication package corresponding to Taylor (2020). Furthermore, I would like to thank my supervisor Bram van Os for motivating and assisting me during my research.

7 Appendix

The raw results, code, and data used in this present study can be found in the supplementary materials attached to this file.

A Programming Packages

The code used to replicate the study by Taylor (2020) and our extension was done through the GAUSS software. The GAUSS software and the packages are downloaded from the APTECH store. [5] Both the replication code sent by Taylor (2020) and the extension code used for the MCS forecast combinations make use of the following packages: **CO(MT)**, **CML(MT)**, and **LP(MT)**.

The **CO** package stands for constrained optimisation and allows us to solve constrained non-linear programming problems. The **CML** depicts the constrained maximum likelihood package which enables us to estimate certain methods using maximum likelihood (ML) while adhering to constraints. Finally, the **LP** package is used to solve linear programming problems.

B Individual Methods

B.0.1 GJR-GARCH(1,1)

$$r_t = \sigma_t \epsilon_t,$$

$$\sigma_t^2 = \omega + (\alpha + \gamma \mathbf{1}_{t-1}) r_{t-1}^2 + \beta \sigma_{t-1}^2.$$

Where ϵ_t is the error term and σ_t^2 denotes the conditional variance of the returns on day t. We define the indicator function as

$$\mathbf{1}_t(r_t) := \begin{cases} 1 & \text{if } \ r_t < 0 \ , \\ 0 & \text{if } \ r_t \geq 0 \ . \end{cases}$$

to guarantee the positive nature of the conditional variance, the following constraints are kept in place: $\omega > 0 \ \beta, \alpha \geq 0$ and $\alpha + \gamma \geq 0$. Variance stationarity is ensured by $\alpha + \beta + 0.5\gamma < 1$ (Nugroho et al., 2019).

[5] The used packages can be found on this website: https://store.aptech.com/gauss-applications-category.html

C Additional Results

Table 10: The model confidence set at a confidence level of 90% for all scoring functions for the S&P 500.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	0	0	0	0	0	0	0	0	0	0
HAR-RANGE	1	1	1	1	0	1	1	1	1	1
CARE-AS	0	0	0	0	0	0	0	0	0	0
CAViaR-AS-EVT	0	0	0	0	0	0	0	0	0	0
Combinations										
Simple Average	0	0	0	0	0	0	0	0	0	0
Relative Score	1	1	1	1	0	1	1	1	0	1
Minimum Score	1	1	1	1	1	0	0	0	0	0

Table 11: The model confidence set at a confidence level of 90% for all scoring functions for the CAC 40.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	1	1	1	1	1	1	1	1	1	1
HAR-RANGE	1	1	1	1	1	1	0	0	0	1
CARE-AS	0	0	0	0	0	0	0	0	0	1
CAViaR-AS-EVT	1	1	1	1	1	1	1	1	1	1
Combinations										
Simple Average	1	1	1	1	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	1	1	1	1	1	1

Table 12: The model confidence set at a confidence level of 90% for all scoring functions for the FTSE 100.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	1	0	0	0	0	0
GJR-GARCH	0	0	0	0	1	1	1	1	1	1
HAR-RANGE	1	1	1	1	1	1	1	1	1	1
CARE-AS	0	0	0	0	0	1	0	1	1	1
CAViaR-AS-EVT	1	1	1	1	1	1	1	1	1	1
Combinations										
Simple Average	1	1	1	1	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	1	1	1	1	1	1

Table 13: The model confidence set at a confidence level of 90% for all scoring functions for the NIKKEI 225.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	1	1	1	1	1	1	1	1	1	1
HAR-RANGE	1	1	1	1	1	1	1	1	1	1
CARE-AS	0	0	0	0	0	1	1	1	1	1
CAViaR-AS-EVT	1	1	1	1	1	1	1	1	1	1
Combinations										
Simple Average	1	1	1	1	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	1	1	1	1	1	1

Table 14: The model confidence set at a confidence level of 75% for all scoring functions for the S&P 500.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	0	0	0	0	0	0	0	0	0	0
HAR-RANGE	0	1	1	0	0	1	1	1	1	1
CARE-AS	0	0	0	0	0	0	0	0	0	0
CAViaR-AS-EVT	0	0	0	0	0	0	0	0	0	0
Combinations										
Simple Average	0	0	0	0	0	0	0	0	0	0
Relative Score	0	1	1	0	0	0	1	1	0	0
Minimum Score	1	0	1	1	1	0	0	0	0	0

Table 15: The model confidence set at a confidence level of 75% for all scoring functions for the CAC 40.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	1	1	1	1	1	0	0	0	0	1
HAR-RANGE	1	1	1	1	1	0	0	0	0	1
CARE-AS	0	0	0	0	0	0	0	0	0	0
CAViaR-AS-EVT	1	1	1	1	1	1	0	1	1	1
Combinations										
Simple Average	1	1	1	1	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	0	1	1	1
Minimum Score	1	1	1	1	1	0	0	0	0	1

Table 16: The model confidence set at a confidence level of 75% for all scoring functions for the FTSE 100.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	0	0	0	0	0	1	1	1	1	1
HAR-RANGE	1	1	1	1	0	1	1	1	1	1
CARE-AS	0	0	0	0	0	1	0	0	1	1
CAViaR-AS-EVT	1	1	1	1	1	1	1	1	1	1
Combinations										
Simple Average	1	1	1	1	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	1	1	1	1	1	1

Table 17: The model confidence set at a confidence level of 75% for all scoring functions for the NIKKEI 225.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	1	1	1	1	1	1	1	1	1	1
HAR-RANGE	1	1	1	1	1	1	1	1	1	1
CARE-AS	0	0	0	0	0	1	1	1	1	1
CAViaR-AS-EVT	0	1	1	0	0	1	1	1	1	1
Combinations										
Simple Average	1	1	1	1	1	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	1	1	1	1	1	1

Table 18: The model confidence set at a confidence level of 75% for all scoring functions for the S&P GSCI.

	1% probability level					5% probability level				
	QS	AL	NZ	FZG	AS	QS	AL	NZ	FZG	AS
Individual Methods										
Historical Simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	1	1	1	1	1	1	0	0	1	1
HAR-RANGE	1	1	1	1	1	1	1	1	1	1
CARE-AS	0	0	0	0	0	0	0	0	0	0
CAViaR-AS-EVT	1	1	1	1	1	1	1	1	1	1
Combinations										
Simple Average	0	0	0	0	0	1	1	1	1	1
Relative Score	1	1	1	1	1	1	1	1	1	1
Minimum Score	1	1	1	1	0	1	1	1	1	1

Table 19: Results of the calibration tests for the S&P 500 stock index.

	1% probability level			5% probability level		
	VaR hit	VaR DQ	ES bootstrap	VaR hit	VaR DQ	ES bootstrap
Individual Methods						
Historical Simulation	1	1	1	0	1	1
GJR-GARCH	1	1	1	1	1	1
HAR-RANGE	0	1	0	1	1	0
CARE-AS	1	1	1	1	1	1
CAViaR-AS-EVT	0	0	0	0	0	0
Combinations						
Simple Average	0	1	0	0	1	0
Relative Score	0	1	0	1	1	0
Minimum Score	0	1	0	0	1	0
Trimmed Combinations						
Simple Average	0	1	0	1	1	0
Relative Score	0	1	0	1	1	0
Minimum Score	0	1	0	1	1	0

Table 20: Results of the calibration tests for the CAC 40 stock index.

	1% probability level			5% probability level		
	VaR hit	VaR DQ	ES bootstrap	VaR hit	VaR DQ	ES bootstrap
Individual Methods						
Historical Simulation	1	1	1	0	1	1
GJR-GARCH	1	0	1	0	0	1
HAR-RANGE	1	0	0	0	1	1
CARE-AS	1	1	1	1	1	1
CAViaR-AS-EVT	0	0	0	0	0	0
Combinations						
Simple Average	1	0	1	0	0	0
Relative Score	0	0	1	0	0	0
Minimum Score	0	0	0	0	0	0
Trimmed Combinations						
Simple Average	0	0	0	0	0	0
Relative Score	0	0	0	0	0	0
Minimum Score	0	0	0	0	0	0

Table 21: Results of the calibration tests for the FTSE 100 stock index.

	1% probability level			5% probability level		
	VaR hit	VaR DQ	ES bootstrap	VaR hit	VaR DQ	ES bootstrap
Individual Methods						
Historical Simulation	0	1	0	1	1	0
GJR-GARCH	0	1	0	1	1	0
HAR-RANGE	0	1	0	0	0	1
CARE-AS	1	1	1	1	1	1
CAViaR-AS-EVT	0	1	0	0	0	0
Combinations						
Simple Average	0	0	0	0	0	0
Relative Score	0	0	1	0	0	0
Minimum Score	0	1	0	1	0	0
Trimmed Combinations						
Simple Average	0	0	0	0	0	0
Relative Score	0	1	0	0	0	0
Minimum Score	0	1	0	0	0	0

Table 22: Results of the calibration tests for the NIKKEI 225 stock index.

	1% probability level			5% probability level		
	VaR hit	VaR DQ	ES bootstrap	VaR hit	VaR DQ	ES bootstrap
Individual Methods						
Historical Simulation	1	1	0	0	1	0
GJR-GARCH	0	0	0	1	1	0
HAR-RANGE	0	0	0	0	0	0
CARE-AS	1	1	1	1	1	1
CAViaR-AS-EVT	0	0	1	0	1	0
Combinations						
Simple Average	0	0	0	0	1	0
Relative Score	0	0	0	0	1	0
Minimum Score	1	1	0	0	0	0
Trimmed Combinations						
Simple Average	0	0	0	0	1	0
Relative Score	0	0	1	0	1	0
Minimum Score	0	0	0	0	1	0

Table 23: 1% VaR evaluated using the quantile skill score (%).

	S&P500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	28.362	15.714	23.033	24.501	13.555	21.033
HAR-RANGE	48.978	15.439	25.365	**26.605**	17.353	26.748
CARE-AS	-30.943	-42.246	-28.756	-26.606	-61.234	-37.957
CAViaR-AS-EVT	28.250	**18.221**	26.078	22.244	14.106	21.780
Combinations						
Simple Average	35.104	16.864	26.095	25.543	11.411	23.003
Relative Score	48.978	17.740	**28.615**	25.543	15.071	27.189
Minimum Score	**49.350**	17.311	26.189	23.720	12.299	25.774
Trimmed Combinations						
Simple Average	48.978	18.187	28.097	25.738	15.806	27.361
Relative Score	48.978	17.654	26.802	23.850	15.561	26.569
Minimum Score	48.978	17.448	26.896	24.327	**21.029**	**27.736**

Table 24: 1% VaR and ES evaluated using the NZ skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	14.592	10.656	14.423	15.138	7.843	12.530
HAR-RANGE	**31.760**	10.656	15.865	16.055	**10.980**	17.063
CARE-AS	-44.635	-45.902	-30.769	-27.982	-66.667	-43.191
CAViaR-AS-EVT	14.163	**12.295**	15.865	13.761	8.235	2.864
Combinations						
Simple Average	21.888	11.475	**18.269**	**16.514**	6.275	14.884
Relative Score	**31.760**	**12.295**	**18.269**	**16.514**	9.020	**17.572**
Minimum Score	**31.760**	11.885	17.308	14.679	7.451	16.617
Trimmed Combinations						
Simple Average	**31.760**	**12.295**	17.308	16.055	9.804	15.444
Relative Score	**31.760**	11.885	16.346	14.679	9.412	16.816
Minimum Score	**31.760**	11.885	16.346	15.138	9.412	16.908

Table 25: 1% VaR and ES evaluated using the FZG skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	1.201	**0.753**	0.744	**0.896**	0.605	0.840
HAR-RANGE	**1.952**	**0.753**	0.744	**0.896**	**0.756**	1.020
CARE-AS	-1.201	-1.958	-0.893	-0.896	-3.177	-1.625
CAViaR-AS-EVT	1.201	**0.753**	-0.744	0.746	0.605	0.512
Combinations						
Simple Average	1.502	**0.753**	**0.893**	**0.896**	0.454	0.900
Relative Score	**1.952**	**0.753**	**0.893**	**0.896**	0.605	1.020
Minimum Score	**1.952**	**0.753**	0.744	0.746	0.605	0.960
Trimmed Combinations						
Simple Average	**1.952**	**0.753**	**0.893**	**0.896**	**0.756**	**1.050**
Relative Score	**1.952**	**0.753**	0.744	0.746	**0.756**	0.990
Minimum Score	**1.952**	**0.753**	0.744	**0.896**	**0.756**	1.020

Table 26: 1% VaR and ES evaluated using the AS skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	49.962	25.071	36.840	43.338	21.278	33.498
HAR-RANGE	67.222	23.807	40.481	**46.787**	**25.664**	40.792
CARE-AS	-1.590	-18.337	-1.388	-11.460	-35.906	-13.736
CAViaR-AS-EVT	46.953	**28.043**	45.048	38.053	21.330	35.885
Combinations						
Simple Average	53.577	24.012	45.757	39.889	17.400	36.127
Relative Score	67.222	25.119	45.757	39.889	23.386	40.275
Minimum Score	**67.865**	24.723	40.019	39.583	17.608	37.960
Trimmed Combinations						
Simple Average	67.222	27.332	**46.529**	44.423	23.581	**41.817**
Relative Score	67.222	26.715	45.387	40.834	23.829	40.797
Minimum Score	67.222	26.257	45.449	41.419	23.894	40.848

Table 27: 5% VaR evaluated using the quantile skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	20.134	9.614	15.538	11.022	4.740	12.210
HAR-RANGE	**34.922**	8.376	14.937	9.606	**5.845**	14.737
CARE-AS	16.879	7.766	15.140	9.958	1.875	10.324
CAViaR-AS-EVT	19.079	10.376	16.936	11.169	5.004	12.513
Combinations						
Simple Average	23.318	10.723	17.642	11.374	5.510	13.713
Relative Score	**34.922**	10.571	17.267	11.602	5.290	15.930
Minimum Score	33.989	9.406	16.553	11.008	5.268	15.245
Trimmed Combinations						
Simple Average	**34.922**	10.717	17.522	**12.035**	5.730	**16.185**
Relative Score	**34.922**	**10.760**	**17.935**	11.184	5.312	16.023
Minimum Score	**34.922**	10.485	17.815	11.580	5.328	16.026

Table 28: 5% VaR and ES evaluated using the NZ skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	11.932	6.111	9.816	6.667	3.158	7.537
HAR-RANGE	**23.295**	5.000	9.816	6.061	**4.211**	9.677
CARE-AS	9.091	4.444	9.202	6.061	0.526	5.865
CAViaR-AS-EVT	11.364	**6.667**	10.429	**7.273**	3.684	7.883
Combinations						
Simple Average	15.341	**6.667**	**11.043**	**7.273**	3.684	8.802
Relative Score	**23.295**	**6.667**	10.429	**7.273**	3.684	10.270
Minimum Score	22.727	6.111	10.429	**7.273**	3.684	10.156
Trimmed Combinations						
Simple Average	**23.295**	**6.667**	**11.043**	**7.273**	**4.211**	**10.498**
Relative Score	**23.295**	**6.667**	**11.043**	**7.273**	3.684	10.392
Minimum Score	**23.295**	**6.667**	**11.043**	**7.273**	3.684	10.392

Table 29: 5% VaR and ES evaluated using the FZG skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	0.592	**0.296**	**0.295**	**0.295**	0.149	0.325
HAR-RANGE	**0.888**	**0.296**	**0.295**	**0.295**	**0.297**	**0.414**
CARE-AS	0.444	**0.296**	**0.295**	**0.295**	0.149	0.296
CAViaR-AS-EVT	0.592	**0.296**	**0.295**	**0.295**	0.149	0.325
Combinations						
Simple Average	0.592	**0.296**	**0.295**	**0.295**	**0.297**	0.355
Relative Score	**0.888**	**0.296**	**0.295**	**0.295**	**0.297**	**0.414**
Minimum Score	**0.888**	**0.296**	**0.295**	**0.295**	**0.297**	**0.414**
Trimmed Combinations						
Simple Average	**0.888**	**0.296**	**0.295**	**0.295**	**0.297**	**0.414**
Relative Score	**0.888**	**0.296**	**0.295**	**0.295**	**0.297**	**0.414**
Minimum Score	**0.888**	**0.296**	**0.295**	**0.295**	**0.297**	**0.414**

Table 30: 5% VaR and ES evaluated using the AS skill score (%).

	S&P 500	CAC 40	FTSE 100	NIKKEI 225	S&P GSCI	Geo. Mean
Individual Methods						
Historical Simulation	0	0	0	0	0	0
GJR-GARCH	30.531	14.138	23.373	20.447	8.040	19.306
HAR-RANGE	**43.814**	13.655	21.061	18.420	**9.104**	21.211
CARE-AS	24.828	13.246	25.365	16.657	5.122	17.043
CAViaR-AS-EVT	27.182	16.212	27.129	19.242	8.701	19.693
Combinations						
Simple Average	30.929	15.784	26.992	18.992	8.633	20.266
Relative Score	**43.814**	15.793	27.403	19.947	8.792	23.150
Minimum Score	42.872	13.928	25.806	18.537	8.747	21.978
Trimmed Combinations						
Simple Average	**43.814**	16.257	25.973	**21.607**	9.096	**23.349**
Relative Score	**43.814**	**16.303**	**27.418**	19.022	8.747	23.061
Minimum Score	**43.814**	16.130	27.281	19.903	8.792	23.184