Bachelor Thesis Econometrie en Operationele Research

The Benefits of Combining Forecasts for Value-at-Risk and Expected Shortfall

Name student:
Foppe Coenen
Student ID number:
572689fc

Supervisor:
B. (Bram) van Os
Second Assessor:
dr. A. (Anastasija) Tetereva

Date final version: July 2, 2023

Abstract

Having accurate forecasts of the Value-at-Risk (VaR) and Expected Shortfall (ES) of any financial asset is crucial for risk management, as these provide information about potential losses and can enable efficient capital allocation. This research extends Taylor (2020)'s influential work on combining individual Value-at-Risk and Expected Shortfall forecasts to improve forecast accuracy by considering additional methods, applied to a diversified portfolio. Joint scoring functions for the VaR and ES are applied, as the ES is not elicitable independently from the VaR. The CAViaR model with Asymmetric Slope and Extreme Value Theory is found to be the most accurate individual method. The empirical analysis shows that combining methods, which incorporate information from different individual methods, produce more accurate and more robust forecasts at the 1% and 5% probability levels, demonstrating the benefits of forecast combinations for the VaR and ES. In particular, the winsorized mean emerges as the superior method, based on all evaluation criteria.

Keywords: Value-at-Risk, Expected Shortfall, forecast combinations, elicitability, joint scoring functions

Bachelor Thesis Econometrie en Operationele Research
Erasmus University Rotterdam
Erasmus School of Economics

Contents

1 Introduction

Value-at-Risk (VaR) and Expected Shortfall (ES) are essential risk measures widely employed by financial institutions, regulators, and market participants to evaluate the risks associated with their investments. These measures have gained significant importance, particularly since the market turbulence of the 1990s due to the 1992 Exchange Rate Mechanism (ERM) crisis, and the 1994 global bond market crash (Acerbi and Tasche, 2002), as they quantify potential losses that can arise from such extreme market events (Alexander, 2009). The VaR indicates the maximum loss that can be expected over a given time horizon at a specific confidence level, and is generally more interpretable than ES. However, Alexander (2009) explains that the ES measures the average loss beyond the VaR threshold, which makes it more informative in terms of capturing tail risk, and has led to a rising popularity of the ES in recent years. Given the central importance of the VaR and ES in risk management, the ability to accurately forecast these measures is crucial, not only for effective decision-making, but also for optimal capital allocation. A promising approach is the combination of individual forecasts. This provides a practical procedure to incorporate different insights gained from various models, generally resulting in a more accurate, combined forecast (Timmermann, 2006). This is especially important in forecasting the VaR and ES, as the unpredictability inherent in financial markets makes it improbable for any single technique to persistently outperform others (Zikovic and Filer, 2012). Accordingly, our research seeks to answer the following central research question:

How can combined Value-at-Risk and Expected Shortfall forecasts enhance the downside risk forecasts of a diversified portfolio?

This paper serves partially as a robustness check for the influential paper of Taylor (2020), who was the first to apply the concept of forecast combinations to Expected Shortfall forecasting. His analysis provides an extensive framework that can be used for continuing this line of work. We contribute to the literature in a variety of ways. First, our research investigates whether the findings presented by Taylor (2020) can be extrapolated to a wider variety of asset classes. More series and recent data are considered, and the analysis is applied to a diversified portfolio. Additionally, our study explores the ways in which combining methods can benefit from a larger set of competitive models. To this extent, the Generalized Autoregressive Scoring (GAS) model and the Cornish Fisher Expansion (CFE) of Cornish and Fisher (1938) are added to the set of indiviudal methods of Taylor (2020). Furthermore, to adequately examine the potential benefits of forecast combination methods, additional combining techniques are implemented. These are the median, the mode (with Gaussian kernel estimation), the winsorized mean, and the trimmed mean. Each of these techniques is relatively straightforward to implement, but provides an interesting trade-off between more freedom (like the mean), and more robustness (like the median or mode). While these combination techniques are well-known in the econometric literature, only few studies seem to have applied these techniques to Value-at-Risk and Expected Shortfall forecasting, making our analysis quite novel in the field of downside risk forecasting.

The main findings of this paper can be summarized as follows. The CAViAR-AS-EVT model produces the most accurate forecasts out of any considered individual method. However, indisputably, the combining methods produce more reliable and forecasts than all individual

methods, with a preference being visible for the winsorized mean forecast combination. This demonstrates that using forecast combinations of the Value-at-Risk and Expected Shortfall can considerably enhance the accuracy of downside risk forecasts for a portfolio that is diversified across multiple asset classes.

Given the importance of accurate VaR and ES forecasts, these downside risk measures have been the subject of extensive research. Our research aligns closely with the following literature. First, a fundamental obstacle in backtesting ES forecasts is that the measure is not elicitable, meaning that the optimal ES forecast is not the unique minimizer of any loss function. This issue is addressed by Fissler and Ziegel (2016), who propose a range of joint scoring functions for VaR and ES, rendering the two measures *jointly* elicitable. These scoring functions enable the joint evaluation of different VaR and ES forecasts.

Second, various models are proposed and considered for forecasting the VaR and ES. A common benchmark is the simple historical simulation, which is generally considered to be uncompetitive (Chen et al., 2012). The Generalized AutoRegressive Conditional Heteroscedasticity (GARCH) model introduced by Bollerslev (1986), along with its variants, are amongst the most popular models in this context (Gao and Song, 2008). GARCH models assume a fixed conditional distribution, which can lead to inaccurate forecasts if the shape of the actual distribution varies over time. This assumption is relaxed by the Conditional Autoregressive Value-at-Risk (CAViaR) model of Engle and Manganelli (2004), which models the conditional quantiles directly. Another well-known model is the Generalized Autoregressive Scoring (GAS) model, proposed by Creal et al. (2013). Its dynamic parameter updating procedure gives this model the ability to react quickly to changes in volatility, as opposed to for example the GARCH model. Furthermore, its structure enables information from the entire distribution to be incorporated, instead of only considering the second-order moments. Additionally, the Cornish-Fisher Expansion (CFE) of Cornish and Fisher (1938) provides direct equations for estimating the VaR and ES, incorporating higher order moments like Skewness and Kurtosis that are frequently overlooked in standard parametric methods. This technique is gaining more popularity since the works of Maillard (2018) and Amédée-Manesme et al. (2019), as it allows for a more nuanced representation of the distribution of returns, such that one can obtain a more comprehensive view of potential extreme losses.

However, due to the inherent complexity of financial markets, it is often found to be unlikely that any individual model consistently outperforms other models (Zikovic and Filer, 2012). This has led to the technique of *forecast combinations*, dating back to the seminal work of Bates and Granger (1969), becoming more popular. The literature suggests that combining different models can lead to more accurate predictions, with much of the idiosyncratic noise of the individual forecasts being substantially reduced (Timmermann, 2006). In particular, the literature motivates combining forecasts from a diverse set of models that are competitive and use different information or incorporate the same information in different ways, as combining those forecasts can provide a more comprehensive understanding of the underlying patterns in the data (Batchelor and Dua, 1995). Over time, a variety of combination techniques have received attention. The simple average is one of the most commonly used forecast combination techniques. It receives theoretical support by the works of Claeskens et al. (2016), and while

it is a straightforward technique, it still manages to outperform even sophisticated combination methods that are theoretically optimal under certain conditions, a finding that is often referred to as the 'Forecast Combination Puzzle' (Smith and Wallis, 2009). In addition to the mean, an increased consideration has been directed towards alternative simple combination techniques, like the median and mode (Kourentzes et al., 2014), as well as winsorized and trimmed means (Jose and Winkler, 2008), as these alternatives are less affected by aberrant observations in the set of forecasts (Wang et al., 2022). As Sinova et al. (2012) mention, by construction, the median less affected by extreme values than the mean, making it a more robust measure of central tendency. Kourentzes et al. (2014) find the median to be superior to the mean, while the more robust mode[1] outperforms both the mean and median forecasts. The mode is robust to asymmetric distributions (Silverman, 1986), and entirely insensitive to outliers (Tay and Wallis, 2000), unlike the median. According to Kourentzes et al. (2014), the mode requires the smallest number of individual forecasters to produce reliable forecasts, a finding which is especially useful when the number of individual models considered is restricted by computational limits. However, the literature does not seem to hold a decisive preference for any of these combination techniques. The winsorized and trimmed means are simple averages that limit the effect of larger forecasts on the mean. They are motivated by Genre et al. (2013) and Jose and Winkler (2008), whose findings imply that these combination techniques are beneficial when there is considerable fluctuation among the individual forecasts.

The literature for combining forecasts for the VaR and ES is still limited, presumably because the ES is not elicitable. The research of Taylor (2020) is one of the first comprehensive studies in this context. He uses the joint scoring functions proposed by Fissler and Ziegel (2016) to estimate combining weights for the VaR and ES, and finds a superiority of the combinations over any individual method. Trucíos and Taylor (2022) consider the mean, median, minimum, and maximum combination techniques for VaR and ES, but do not find these combinations to consistently outperform their individual models. To our best knowledge, the remaining combination techniques mentioned are not yet extensively researched in the context of VaR and ES, such that our research could provide novel insights.

The remainder of this paper is structured as follows. Section 2 introduces the data used for our empirical analysis, and displays some characteristics of this data. Section 3 formally introduces the two risk measures, and presents the techniques used to construct, combine, and evaluate forecasts of these measures. The main findings are discussed in Section 4. Finally, Section 5 concludes and provides suggestions for future research.

2 Data

In this research, methods are considered to construct and combine forecasts of one-day-ahead Value-at-Risk and Expected Shortfall of daily log returns of financial assets. This research extends the data of Taylor (2020), by considering more recent data, and by using a diverse range of assets. By considering a diverse range of asset classes, we ensure generalizability and more robustness across different financial markets. This not only improves the external validity of our

[1]For continuous data, usually kernel density estimation is applied, such that the data need not be discretized.

analysis, but also aligns our research with the practical needs of investors and risk managers, ensuring both academic and practical value. By incorporating these different asset classes, each with its own unique risk-return dynamics, our findings can provide valuable insights into forecasting the downside risks of both individual assets, as well as diversified portfolios. This is particularly helpful for the mean-investors, whose risk management strategies rely on diversifying their portfolio (Marling and Emanuelsson, 2012).

This research uses data from the equity market, fixed-income market, commodity market and foreign exchange market. Respectively, the indices considered for these classes are the S&P 500 and FTSE 100 stock indices, the iShares Core U.S. Aggregate Bond ETF (AGG), the S&P Goldman Sachs Commodity Index (GSCI), and EURUSD BGN Curncy (the exchange rate between the Euro (EUR) and the United States Dollar (USD)). Data on these series is available on Bloomberg. Information is collected on daily closing prices and closing exchange rates for a total of 6000 observations per series, all ending[2] on April 30, 2023, thereby extending the dataset used by Taylor (2020), whose research considered data up to 2017. Crypto currencies are not considered in this research, due to the small number of trading days since their release[3], but this could be considered in further research.

Additionally, in order to answer the research question more adequately, a diversified portfolio is constructed, where the investor invests 60% of his wealth into the equity market, 30% into the bond market, 5% into commodities and 5% in the Foreign Exchange Market. This kind of portfolio can often be found in pension funds, and tries to balance risk and return characteristics: while stocks are known to deliver opportunities for growth, the relatively large fraction of bonds adds stability (reduce risk), with additional diversification from the commodities, and exposure to global trends in the economy as a result of the foreign exchange market. Returns for this portfolio are calculated using the procedure described in Section A.2 in the Appendix.

The dataset is split into an in-sample part of the first 4000 observations, and an out-of-sample part of the final 2000 observations. The middle 2000 observations are used for in-sample validation and combining weight estimation, as explained in Section 3, and the last 2000 days of the samples are used for the comparative analysis of forecast accuracies of all considered techniques.

3 Methodology

This Section presents the different techniques for constructing, combining, and evaluating forecasts of the two risk measures of interest, the Value-at-Risk and the Expected Shortfall, building upon the foundational work of Taylor (2020). In particular, our analysis implements his forecasting methods[4], as well as the proposed scoring functions. His framework is extended in several ways. First, two additional individual forecasting techniques are implemented, which are the Cornish-Fisher Expansion (Section 3.2.5) and the Generalized Autoregressive Scoring (Section 3.2.6) model. Second, as the main focus of this research is combining forecasts, a wider variety of (simple) combination techniques are added. In particular, the median, mode, winsorized mean

[2]The starting dates of the series are different, due to the variation of holiday periods in different countries.
[3]Bitcoin first created in January 2009.
[4]As intra-day ranges are not available for some series, the HAR-range method of Taylor (2020) is discarded.

and trimmed mean are implemented. The larger set of individual methods should enhance the performance of the combining methods (Batchelor and Dua, 1995). Finally, the backtesting techniques considered by Taylor (2020) are extended by incorporating a test for zero autocorrelation (AC) in the Expected Shortfall forecasts.

Throughout this research, the daily log-returns of an asset j are denoted as $y_{j,t} = \log\left(\frac{P_{j,t}}{P_{j,t-1}}\right)$, with $P_{j,t}$ the price of asset j at time (day) t, with $j \in \mathbb{N}_J$, such that there are a total of J series[5]. Log returns of the portfolio ($j = 6$) are obtained from its Simple Returns[6]. In order to reduce the noise in the returns series, and to get a clearer indication of the patterns in these returns, an AR(1) filter is applied. That is, instead of the returns, the empirical analysis uses the residuals of an AutoRegressive (AR) model of order 1 for these returns. Hence, for the remainder of this research, $r_{j,t}$ denotes the *residual* return of an AR(1) model for series j, or $r_{j,t} := y_{j,t} - \hat{y}_{j,t}$, with $\hat{y}_{j,t}$ the fitted return of the AR(1) model[7].

3.1 Risk Measures: Value-at-Risk, Expected Shortfall

The Value-at-Risk (VaR) and Expected Shortfall (ES) are popular risk metrics, both aiming to quantify the downside risk of a (portfolio of) investment(s). This research uses the formulation with returns (rather than profits/losses), such that the Value-at-Risk is the minimum return over a given time period with a prespecified probability. Only 1-day-ahead forecasts of the VaR and ES are considered, but the techniques presented here can be extended to multihorizon settings. Formally, with a specified probability $0 < \alpha < 1$, the $100 \cdot \alpha\%$ Value-at-Risk over the next day is the level of returns for which lower returns are expected with probability α. For series j, this quantile, denoted $VaR^{\alpha}_{j,t+1}$, is the value of x which satisfies the following equation:

$$VaR^{\alpha}_{j,t+1} = \{x \in \mathbb{R} : \ P(r_{j,t+1} \leq x) = \alpha\}. \tag{1}$$

While the VaR is informative about the potential loss, it fixates on the exact value of the quantile, and is uninformative on the distribution of returns or losses beyond this value. Instead, the Expected Shortfall measures the *expected* loss, *given* that the Value-at-Risk is violated. This conditional aspect of the ES has led to the name Conditional Value-at-Risk (CVaR) being synonymous with the ES. It provides more insights into the shape of the loss distribution in the tail beyond the VaR quantile than the VaR itself. Formally, the Expected Shortfall corresponding to the $VaR^{\alpha}_{j,t+1}$, denoted $ES^{\alpha}_{j,t+1}$, is given by the following conditional expectation:

$$ES^{\alpha}_{j,t+1} = E\big(r_{j,t+1} \mid r_{j,t+1} \leq VaR^{\alpha}_{j,t+1}\big). \tag{2}$$

3.2 Forecast Construction: Individual Methods

This Subsection proposes the individual techniques used to construct VaR and ES forecasts. As mentioned, forecast combinations tend to benefit from a set of different competitive models, especially those that use information differently (Batchelor and Dua, 1995). For this reason, our analysis considers parametric, semi-parametric, and non-parametric individual methods, to obtain a set of diverse methods that use the available time series information in different ways.

[5] In our empirical analysis, a total of $J \equiv 6$ series are considered, which are the ones mentioned in Section 2.

[6] For completeness, Section A.2 in the Appendix explains in detail how these log returns are calculated.

[7] Formally, $\hat{y}_{j,t} := \hat{\theta}_1 y_{j,t}$, with $\hat{\theta}_1$ the OLS estimate of the AR parmamter θ_1 in the regression $y_{j,t} = \theta_1 y_{j,t-1} + \epsilon_{j,t}$

A rolling window approach is used for parameter estimation and out-of-sample forecasting. In particular, the parameters of the individual methods of Section 3.2 are estimated using a 2000-day window. They are re-estimated every 250 days, constituting about one trading year[8]. This procedure produces out-of-sample forecasts for the final 4000 days of each time series. For this period of 4000 days, the weights and percentiles to be used for the combined forecasts (as explained in Section 3.3), are again estimated using a 2000-day rolling window, with the weights being re-estimated every 250 days.

3.2.1 Historical Simulation

The first individual method is the historical simulation, using the 250 most recent observations[9]. This non-parametric is often used as a benchmark, and is known to be inaccurate, largely due to its assumption that all returns have the same distribution. This assumption is (in most cases) inappropriate, due to the prominent presence of volatility clustering in most series of returns.

3.2.2 GJR-GARCH

The second individual technique is the GJR-GARCH(1,1) model, with a Student's t-distribution[10]. This model is fully parametric, and extends the popular GARCH(1,1) model by allowing for asymmetric properties in the conditional variances. If $\sigma_{j,t}^2$ is the conditional variance of the returns $r_{j,t}$, the GJR-GARCH model for series j can be represented by the following equations;

$$r_{j,t} = \mu_{j,t} + \epsilon_{j,t}, \tag{3}$$

$$\sigma_{j,t}^2 = \omega_j + \left(\alpha_j + \gamma_j \cdot \mathbb{I}_{\{\epsilon_{j,t-1}<0\}}\right)\epsilon_{j,t-1}^2 + \beta_j \sigma_{j,t-1}^2, \tag{4}$$

$$\epsilon_{j,t} = \sigma_{j,t} \cdot z_{j,t}, \tag{5}$$

where $z_{j,t}$ is assumed to follow a Student's t-distribution, such that $\epsilon_{j,t}$ follows a scaled zero-mean t-distribution. The indicator function $\mathbb{I}_{\{\epsilon_{j,t-1}<0\}}$ allows for asymmetric properties, as a positive γ_j implies that negative shocks have a larger effect on the conditional volatility than equally-sized positive shocks[11], and vice versa. Parameters are estimated using Maximum Likelihood.

3.2.3 CAViaR-AS-EVT

The third model is referred to as the Asymmetric-Slope (AS) Conditional Autoregressive Value-at-Risk (CAViaR) model, which, for a series j, can be expressed by the following equation:

$$Q_{j,t} = \beta_{j,0} + \left(\beta_{j,1} \cdot \mathbb{I}_{\{r_{j,t-1}>0\}} + \beta_{j,2} \cdot (1 - \mathbb{I}_{\{r_{j,t-1}>0\}})\right) \cdot |r_{j,t-1}| + \beta_{j,3} \cdot Q_{j,t-1}, \tag{6}$$

where $Q_{j,t}$ denotes Value-at-Risk of series j at time t. Estimation of the parameters is achieved by using a quantile regression minimization[12], as explained by Engle and Manganelli (2004), and

[8] We alo experimented with refitting every 1,5,21 periods, but the resulting computational burden was too large. Conversely, refitting every 1000 or 2000 periods resulted in noticeably decreased forecasting accuracy.

[9] Using the 100, 500, 1000, or 2000 most recent observations instead did not improve the forecast accuracy.

[10] A GARCH model with skew-t distribution, filtered historical simulation, and the Extreme Value Theory (EVT) approach of McNeil and Frey (2000) were also considered, but these did not lead to more accurate forecsts.

[11] Enforcing $\gamma_j := 0$ in Equation (4) results in the standard GARCH(1,1) model.

[12] Before applying this regression, parameter vectors are sampled from a uniform distribution with bounds set by some experimentation, and the previously optimized vector is also included as a candidate. The final parameter vector is the one with the lowest score following a quasi-Newton algorithm with the three best vectors, based on the quantile score from Section 3.4.

introduced by Koenker and Bassett Jr (1978). The Asymmetric Slope (AS) allows for different effects of positive and negative returns. After estimating the parameters of Equation (6), the same procedure from Engle and Manganelli (2004) is implemented. That is, the exceedances are standardized by the corresponding estimated quantiles, and Peaks-over-Threshold Extreme Value Theory (EVT) is applied to these standardized exceedances. Finally, the Value-at-Risk and Expected Shortfall can be forecasted using the fitted Extreme Value distribution[13].

3.2.4 CARE-AS

While using quantiles is often the first choice for modeling tail risk, it is well known that *expectiles* are more robust to fat-tailed distributions and potential aberrant observations, resulting in a more robust framework for analyzing the tail properties of a distribution (Chen, 2018). Furthermore, according to Newey and Powell (1987), expectiles can capture the distribution in the tails more adequately than quantiles, as they are more sensitive to its tail behavior. These expectiles, denoted $\mu_{j,t}$ for asset j, minimize the weighted sum of asymmetrically scaled squared residuals (Asymmetric Least Squares), as explained by Newey and Powell (1987). A popular model using the expectiles is the Asymmetric Slope (AS) Conditional AutoRegressive Expectile (CARE), proposed by Taylor (2008), and is implemented in our research as well. This model can be expressed as follows;

$$\mu_{j,t} = \beta_{j,0} + \left(\beta_{j,1} \cdot \mathbb{I}_{\{r_{j,t-1}>0\}} + \beta_{j,2} \cdot (1 - \mathbb{I}_{\{r_{j,t-1}>0\}})\right) \cdot |r_{j,t-1}| + \beta_{j,3} \cdot \mu_{j,t-1}. \tag{7}$$

The parameters are estimated using a similar technique to that of the CAViaR model, but with the *expectile score* instead of the quantile score[14]. The τ expectile approximates the α quantile, and the ES can be expressed as a function of the expectile. The τ parameter is estimated using the optimization procedure of Taylor (2008)[15], which repeatedly re-estimates CARE-AS models.

3.2.5 Cornish-Fisher Expansion

Another method implemented in the empirical analysis is the Cornish-Fisher Expansion (CFE), as introduced by Cornish and Fisher (1938). Essentially, the CFE transforms a standard normal random variable into a non-normal random variable, by means of expanding the quantiles of a standard normal distribution into a series in terms of the standardized moments. This method accounts for the skewness and kurtosis, in contrast to standard parametric methods, which often overlook these higher-order moments. This is particularly important in the context of financial returns, which often exhibit significant skewness (asymmetry) and kurtosis (fat tails) as displayed by one of the stylized facts of daily asset returns, implying the assumption of normality may lead to the underestimation of potential risk (Sheikh and Qiao, 2010). Notwithstanding, the estimation of higher-order moments can be sensitive to aberrant observations, such that having a large enough sample size is crucial. The standardized Cornish-Fisher quantile, denoted z^{α}_{CF},

[13] This research uses The Generalized Pareto Distribution, as explained by Taylor and Yu (2016).

[14] The expectile score is defined as $S(\mu_{j,t},\ r_{j,t}) := |\tau - \mathbb{I}_{\{r_{j,t} \leq \mu_{j,t}\}}| \cdot (r_{j,t} - \mu_{j,t})^2$

[15] In particular, this optimization procedure involves iteratively re-estimating CARE models, decreasing τ by 0.0001 until the ratio of in-sample exceedances surpassing the fitted expectile is closer to α than a pre-specified tolerance. Initial τ values of 0.0018 and 0.0167 are used for the 1% and 5% probability levels, respectively.

and CVAR, denoted y^{α}_{CF}, for a probability level α are constructed as follows:

$$z^{\alpha}_{CF,t+1} = z_\alpha + \left(z_\alpha^2 - 1\right)\frac{S_{c,t}}{6} + \left(5z_\alpha - 2z_\alpha^3\right)\frac{S_{c,t}^2}{36} + \left(z_\alpha^3 - 3z_\alpha\right)\frac{K_{c,t}}{24}, \tag{8}$$

$$y^{\alpha}_{CF,t+1} = y_\alpha\left(1 + z_\alpha \cdot \frac{S_{c,t}}{6} + \left(1 - 2z_\alpha^2\right)\frac{S_{c,t}^2}{36} + \left(z_\alpha^2 - 1\right)\frac{K_{c,t}}{24}\right), \tag{9}$$

where $S_{c,t}$ and $K_{c,t}$ denote the Skewness and Kurtosis *parameters*, respectively, and z_α and y_α being the VaR and ES values for a Gaussian distribution respectively[16]. The index t enables time-varying forecasts, where a moving window of 2000 observations is again employed to calculate the sample Skewness and Kurtosis, and this window is moved forward one day at a time, to allow for time-varying VaR and ES forecast[17].

However, the Skewness and Kurtosis parameters do not correspond to the observed (sample) Skewness and Kurtosis. Maillard (2018) and Amédée-Manesme et al. (2019) explain that using the sample Skewness and Kurtosis in Equations (8) and (9) results in significant mis-estimation of the quantiles, and hence leads to poor forecasts. Maillard (2018) presents the Skewness and Kurtosis parameters as non-explicit functions of the observed moments, while Amédée-Manesme et al. (2019) applies Response Surface Methodology (RSM) to allow for a direct computation of these parameters. The latter approach is implemented in this research, as it alleviates the difficulty of working with non-explicit functions. The RSM involves estimating polynomial models for the Skewness and Kurtosis parameters, with the explanatory variables being the observed Skewness and Kurtosis, and transformations thereof. For completeness, an elaborate explanation about this methodology, as well as concrete details about its implementation, is given in Section A.3 in the Appendix. The standardized quantiles in Equations (8) and (9) are then linearly transformed, to obtain the following VaR and ES forecasts:

$$\widehat{VaR}^{\alpha}_{t+1} = \mu_t + z^{\alpha}_{CF,t+1} \cdot \sigma_t, \qquad \widehat{ES}^{\alpha}_{t+1} = \mu_t + y^{\alpha}_{CF,t+1} \cdot \sigma_t, \tag{10}$$

where μ_t and σ_t denote the sample mean and standard deviation, respectively, and the index for the assets (j) is dropped for notation convenience. The corrected Cornish-Fisher standard deviation σ_{CF} of Amédée-Manesme et al. (2019) was also considered[18], but this did not improve the forecasting accuracy, and so we simply take the sample standard deviation. Alternatively, as the volatility is known to be time-varying and partially predictable, one could implement dynamic volatility models[19], and use their volatility forecasts $(\hat{\sigma}^2_{t+1})$ as a substitute to the sample variance σ^2_t. While this technique could prove to be a valuable addition to the analysis, it was not feasible for us to implement it, due to time constraints and the scope of our research.

3.2.6 GAS

As a final individual technique for forecasting VaR and ES, the Generalized Autoregressive Scoring (GAS) model is considered, as proposed by Creal et al. (2013). GAS models allow

[16] These are calculated using the theoretical results of Khokhlov (2016). For completeness, the Gaussian VaR and ES are given by $z_\alpha = \Phi^{-1}(\alpha)$, and $y_\alpha = -\frac{1}{\alpha}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z_\alpha^2}{2}\right\}$, respectively.

[17] Forecasts were also constructed using windows of 250, 500, and 1000 observations. However, these did not lead to improved accuracy, so the 2000-observation window is used, aligning this technique with the other methods.

[18] Amédée-Manesme et al. (2019) follow Maillard (2018) by using $\sigma_{CF} := \frac{\sigma_t}{\sqrt{1+\frac{1}{96}K_{c,t}^2+\frac{25}{1296}S_{c,t}^2-\frac{1}{36}K_{c,t}\cdot S_{c,t}^2}}$.

[19] Potential candidates are Exponentially Weighted Moving Average (EWMA) models, GARCH models (of Section 3.2.2), or GAS models (of Section 3.2.6).

for time-varying parameters, and assume they evolve according to the score of the conditional distribution. Their structure allows for the incorporation of information from the entire distribution, instead of exclusively considering the second-order moments. Where the GARCH models of Section 3.2.2 use squared returns in the volatility updating Equation (4), the GAS model uses the score of the conditional distribution, as $\sigma^2_{j,t+1} = \omega_j + \beta_j \sigma^2_{j,t} + \delta s_{j,t}$. Here, the function $s_{j,t} := S_{j,t}(\sigma^2_{j,t}) \cdot \nabla_{j,t}(r_{j,t}, \sigma^2_t)$ uses a positive scaling function $S_{j,t}$ for the score $\nabla_{j,t}$, with the score being the gradient of the log-likelihood function of the conditional distribution, that is:

$$\nabla_{j,t}(r_{j,t}, \sigma^2_t) := \left(\frac{\partial \ln \left(f(r_{j,t}|\sigma^2_t) \right)}{\partial \sigma^2_t} \right). \tag{11}$$

Here, $f(\cdot)$ is the probability density function (pdf) of the assumed distribution for the disturbances. In order to account for the variance of the scoring function $\nabla_{j,t}$, the scaling function $S_{j,t}$ is commonly set to be a power $\rho \in \{0, \frac{1}{2}, 1\}$ of the inverse Fisher information matrix[20] of σ^2_t, as $S_{j,t}(\sigma^2_t) := \mathbf{I}_{j,t}(\sigma^2_{j,t})^{-\rho}$. The value $\rho := 1$ corresponds to Identity Scaling, and a value of $\rho = 1$ ($\rho = \frac{1}{2}$) uses scaling by premultiplying the scoring function by the inverse of the (square root of) the Fisher Information matrix, and the inverse of the Fisher information matrix, respectively.

Estimation of the parameters of the model is repeated every 250 days, using maximum likelihood with two distributions: a student's t distribution and a skew-t distribution, both being beneficial in providing robustness against extreme observations, as mentioned by Bell and Huang (2006) and Azzalini and Genton (2008). These GAS models with time-varying scale produce one-step-ahead *density* forecasts of the returns $r_{j,t}$, meaning the Value-at-Risk and Expected Shortfall forecasts at a probability level α can be computed numerically, as follows;

$$\widehat{VaR}^{\alpha}_{j,t+1} = \hat{F}^{-1}_{t+1}(\alpha), \qquad \widehat{ES}^{\alpha}_{j,t+1} = \frac{1}{\alpha} \int_{-\infty}^{\widehat{VaR}^{\alpha}_{j,t+1}} x \cdot f_{t+1}(x) dx. \tag{12}$$

Here, $\hat{F}^{-1}(\alpha)$ is the α-quantile of the estimated conditional distribution of returns $r_{j,t+1}$, computed by inverting the predicted cumulative density function (CDF) numerically, and the Expected Shortfall forecasts are computed by using Numerical Adaptive Integration of the forecasted density (Narasimhan et al., 2023), as used by Ardia et al. (2018). The specifications of the density functions $f_{t+1}(\cdot)$ and $F_{t+1}(\cdot)$ depend on the chosen distribution.

Our analysis considers GAS models with skew-t distribution and identity scaling, as well as the variants with student's t distribution and the three different scaling types. In-sample validation is used to select the best variant, which is the one that minimizes a scoring function[21] of Section 3.4. Validation using the FZG score[22] shows a preference for the GAS model with student's t-distribution and identity scaling is preferred and hence this variant is used in the empirical analysis. This makes the GAS model align closely with the GARCH model (of Section 3.2.2) that uses the same distribution, such that potential changes in forecast accuracy can be attributed to the more adequate updating of the volatility in the GAS model.

[20] For the univariate case, the Fisher Information matrix for $\sigma^2_{j,t}$ is defined as $\mathbf{I}_{j,t} := E\left[\left(\nabla_{j,t}(r_{j,t}, \sigma^2_t) \right)^2 | \mathcal{I}_{t-1} \right]$, where $\mathcal{I}_{t-1}$ is the Information Set at time $t-1$.

[21] Alternatively, one could compare the *likelihoods*. However, given our focus on the scoring functions of Section 3.4, it seems more appropriate to use these scores for the in-sample validation.

[22] The rankings are similar for the other scoring functions of Table 1, albeit slightly less decisive.

3.3 Forecast Combinations

This Subsection discusses several simple forecast combination techniques, like the simple average, median, mode, winsorized and trimmed means, as well as more sophisticated avaeraging techniques that involve estimation of combining weights. A distinction is made between two sets of individual methods. First, the set of all individual forecasters of Section 3.2 is considered, denoted by $\mathcal{M}_0$, containing $M := |\mathcal{M}_0|$ forecasts[23]. However, as mentioned, the historical simulation method is often found to be uncompetitive. Therefore, the set of all individual methods excluding historical simulation is also considered, which is denoted by $\mathcal{M}_H := \mathcal{M}_0 \setminus \{\text{HS}\}$.

3.3.1 Simple Combination Techniques

As a first approach to combining forecasts, the *simple average* is implemented, which is one of the most commonly used combination techniques. Although it provides equal weights to each method, thereby assuming that all the individual forecasts are equally accurate, its empirical support greatly motivates its inclusion in our empirical analysis. While the mean is often found to produce competitive forecasts, it is not robust to large observations in the set of forecasts. Hence, the more robust *median* and *mode* are also implemented. If the number of individual forecasts, $|\mathcal{M}|$, is odd, the median is the middle value of the sorted forecasts, otherwise it is the average of the two middle values. The median could prove useful if there are potential outliers in the set of forecasters, or if the forecasters are skewed.

The mode is defined as the most frequent value in the set of data, and is completely insensitive to outliers (Tay and Wallis, 2000). As the individual forecasts are continuous, the mode is calculated using Kernel Density Estimation (KDE) of Parzen (1962) and Rosenblatt (1956). KDE estimates the probability density function of a random variable, which in our case is the set of individual forecasts. Following Tay and Wallis (2000), the kernel is chosen to be the Gaussian kernel, and the rule of thumb proposed by Terrell and Scott (1992) is used for the bandwidth. The mode ensemble forecast is then calculated as the value that corresponds to the maximum estimated density (Kourentzes et al., 2014). For completeness, an elaborate explanation of the Gaussian KDE is given in Section A.4.

The *winsorized mean* can be another suitable choice when extreme values are present in the individual forecasts. This technique reduces the impact of potential outliers in the set of forecasters by replacing the most extreme values by observations at a certain percentile, called the winsorizing percentile. An equal-weighted average is then computed over the resulting set of values. By construction, the winsorized mean can be interpreted as a balance between the previously proposed mean and median, in the sense that, like the median, it reduces its sensitivity to potential outliers, while allowing for the resulting forecasts to be sensitive to changes in the data, like the mean. The winsorizing percentile is determined using in-sample validation based on the FZG score of Section 3.4, where the percentile is fixed across probability levels, but can be different across series[24]. This percentile is optimized once[25], using the full validation sample.

[23] Our empirical research considers the $M \equiv 6$ individual forecasters of Section 3.2.

[24] Experimenting with equal percentiles across the series, as well as different percentiles across probability levels, did not improve forecast accuracy, and consequently not considered in the empirical analysis.

[25] Re-estimating the optimal percentile every 250 periods, similarly as the combining methods in Sections 3.3.2 and 3.3.3, did not lead to more accurate forecasts, and is thus not considered, reducing computation times.

Finally, the last simple combination technique considered[26] is the *trimmed mean*. This technique is similar to the winsorized mean, in that it adjusts the most extreme values at some percentile, called the trimming percentile. However, these values are simply discarded, instead of being replaced by less extreme values. The trimming percentiles are calculated similarly as the winsorizing percentiles. The distinction between winsorizing and trimming, and the choice thereof, is subtle but important. A preference for either one should be determined mostly by the nature/type of the extreme values present. If these larger values are believed to be errors, the trimmed mean may be more applicable, while the winsorized mean could be preferred if these values, while extreme, are still assumed to provide information Jose and Winkler (2008). By winsorizing the most extreme values, part of the information within these values is still maintained, while trimming these observations eliminates their information entirely.

3.3.2 Minimum Score Combining

The first sophisticated combination technique involves the estimation of two different sets of weights to combine the VaR and ES forecasts, which seems appropriate given a method may produce VaR and ES forecasts that are of different quality. Taylor (2020) combines forecasts of the *difference* between the Value-at-Risk and Expected Shortfall, resulting in a method he referred to as *minimum score combining*. Letting $Q^c_{j,t}$ and $ES^c_{j,t}$ denote the combined VaR and ES forecasts respectively, this method can be presented by the following equations;

$$\hat{Q}^c_{j,t} = \sum_{i=1}^{|\mathcal{M}|} q_j^{(i)} \cdot \hat{Q}^{(i)}_{j,t}, \qquad q_j^{(i)} \in \left\{ \mathbb{R} : \min\left\{ q_j^{(i)} \right\} \geq 0, \sum_{i=1}^{|\mathcal{M}|} q_j^{(i)} = 1 \right\}, \tag{13}$$

$$\widehat{ES}^c_{j,t} = \hat{Q}^c_{j,t} + \sum_{i=1}^{|\mathcal{M}|} e_j^{(i)} \cdot \left(\widehat{ES}^{(i)}_{j,t} - \hat{Q}^{(i)}_{j,t} \right), \qquad e_j^{(i)} \in \left\{ \mathbb{R} : \min\left\{ e_j^{(i)} \right\} \geq 0, \sum_{i=1}^{|\mathcal{M}|} e_j^{(i)} = 1 \right\}. \tag{14}$$

Here, $\hat{Q}^{(i)}_{j,t}$ and $\widehat{ES}^{(i)}_{j,t}$ respectively denote the VaR and ES forecast of a single model[27] $i \in \mathcal{M}$ for series j, and $q_j^{(i)}$ and $e_j^{(i)}$ are the convex weights for the i-th individual VaR and ES forecasts, respectively[28].These weights ensure the combined ES forecasts exceed the combined VaR forecasts. The weights are estimated simultaneously, being the arguments that minimize the value of the AL scoring function[29], introduced in Section 3.4.

3.3.3 Relative Score Combining

A final way to combine forecasts is a technique which Taylor (2020) referred to as *relative score combining*. In contrast to the different combining weights for the VaR and ES forecasts in the minimum score combining, now the same set of weights is used, denoted $\omega_j^{(i)}$ for asset j using model i. This technique uses weights inversely proportional to the relevant measure of accuracy. Our research applies the same approach as Shan and Yang (2009), but uses the Scoring functions

[26] The minimum and maximum ensemble forecasts were also implemented, but these lead to poor forecasts (as perhaps expected), and are hence omitted.

[27] The set $\mathcal{M}$ either refers to $\mathcal{M}_0$ or to $\mathcal{M}_H$, depending on whether historical simulation is to be included.

[28] Enforcing $q_j^{(i)} := e_j^{(i)}$ resulted in similar findings. Therefore, these results are omitted, in order to save space.

[29] Using any of these functions resulted in similar findings. Hence, only the results using AL score is reported.

from Section 3.4 instead of the quantile score. This method is expressed as follows;

$$\hat{Q}^{c}_{j,t} = \sum_{i=1}^{|\mathcal{M}|} \omega_j^{(i)} \cdot Q^{(i)}_{j,t}, \tag{15}$$

$$\widehat{\text{ES}}^{c}_{j,t} = \hat{Q}^{c}_{j,t} + \sum_{i=1}^{|\mathcal{M}|} \omega_j^{(i)} \cdot \widehat{ES}^{(i)}_{j,t}, \tag{16}$$

$$\omega_j^{(i)} := \frac{\exp\left(-\lambda_j \sum_{d=1}^{t-1} S\left(\hat{Q}^{(i)}_{j,d},\ \widehat{ES}^{(i)}_{j,d},\ r_{j,d}\right)\right)}{\sum_{k=1}^{|\mathcal{M}|} \exp\left(-\lambda_j \sum_{d=1}^{t-1} S\left(\hat{Q}^{(k)}_{j,d},\ \widehat{ES}^{(k)}_{j,d},\ r_{j,d}\right)\right)}, \quad \lambda_j > 0 \tag{17}$$

where $S(\cdot)$ is the chosen scoring function from Section 3.4. The tuning parameter λ_j can be interpreted as the degree to which the weights $\omega_j^{(i)}$ depend on the given scoring function $S(\cdot)$, and are chosen/optimized by minimizing the in-sample values of this chosen scoring function.

3.4 Forecast Evaluation: Scoring Functions

This Subsection introduces the set of considered *scoring functions*, which are loss functions used to assess the quality/accuracy of forecasts of stochastic parameters. Evaluating VaR and ES forecasts is not straightforward: not only are the *actual* VaR and ES not observed, but the Expected Shortfall is not *elicitable*, meaning the actual correct forecast of the Expected Shortfall does *not* uniquely minimize any (expected) loss function. Especially this latter point complicates the forecast evaluation process, a process which is essential in order to determine a method's utility in accurately forecasting the VaR and ES. However, Fissler and Ziegel (2016) show that the Value-at-Risk and Expected Shortfall are elicitable *jointly*, and propose a set of joint VaR and ES loss functions for which these measures are indeed jointly elicitable. In this research, these joint loss functions are used, with the focus being on combining VaR and ES forecasts from the individual models/methods of Section 3.2.

3.4.1 Scoring Function for the VaR: the Quantile Score

First, the Value-at-Risk is considered separately, as this measure *is* elicitable, meaning there exists at least one loss function for which the correct VaR forecast is the unique minimizer of this expected scoring function. In this case, this scoring function is called *strictly consistent* for the VaR, for which Gneiting and Raftery (2007) provide a variety of forms. One of these is the widely used *quantile score*, which is used as the loss function for quantile regressions. This strictly consistent function has the following form (Gneiting, 2011);

$$S\left(\hat{Q}^{(i)}_{j,t},\ r_{j,t}\right) = \left(\alpha - \mathbb{I}_{\left\{r_{j,t} \leq \hat{Q}^{(i)}_{j,t}\right\}}\right) \cdot \left(r_{j,t} - \hat{Q}^{(i)}_{j,t}\right), \tag{18}$$

where $\hat{Q}^{(i)}_{j,t}$ is the Value-at-Risk forecast of method i at a probability level of α, and $\mathbb{I}_{\{r_{j,t} < \hat{Q}^{(i)}_{j,t}\}}$ is an indicator function taking on the value 1 when the VaR is violated/exceeded.

3.4.2 Joint Scoring Functions for the VaR and ES

As mentioned, Fissler and Ziegel (2016) show that the VaR and ES are jointly elicitable, and propose consistent scoring functions with the following general form;

$$S(Q_{j,t},\ ES_{j,t},\ r_{j,t}) = -\left(\alpha - \mathbb{I}_{\{r_{j,t}\leq Q_{j,t}\}}\right)\cdot G_1(Q_{j,t}) - \mathbb{I}_{\{r_{j,t}\leq Q_{j,t}\}}\cdot G_1(r_{j,t}) \\ + G_2(ES_{j,t})\cdot\left(ES_{j,t} - Q_{j,t} + \mathbb{I}_{\{r_{j,t}\leq Q_{j,t}\}}\cdot\frac{Q_{j,t}-r_{j,t}}{\alpha}\right) - \zeta_2(ES_t) + a(r_{j,t}), \tag{19}$$

where $G_1(\cdot)$, $G_2(\cdot)$, $\zeta_2(\cdot)$ and $a(\cdot)$ are functions to be specified, satisfying (at least) the constraints $\frac{d}{dx}G_1(x)\geq 0$, $\frac{d}{dx}\zeta_2(x)\equiv G_2(x)\geq 0$, and $\zeta_2(\cdot)$ convex. The scoring function in Equation (19) is *strictly* consistent if $\zeta_2(\cdot)$ is strictly increasing $\left(\frac{d}{dx}\zeta_2(x)>0\right)$ and strictly concave. The exact specifications of the functions $G_1(x)$, $G_2(x)$, $\zeta_2(x)$ and $a(r_j)$ for these scoring functions are shown in Table 1. The AL, NZ and FZG scores are used to estimate the weights[30] for the minimum score $\left(q_j^{(i)},\ e_j^{(i)}\right)$ and relative score $\left(\omega_j^{(i)}\right)$ combining methods.

First, Taylor (2019) proposes a scoring function being the negative of the log-likelihood function of an Asymmetric Laplace density with time varying parameters, which he refers to as the *AL score*. This function is supported by the works of Patton et al. (2019). Second, the research of Nolde and Ziegel (2017), which involves comparative backtests for several risk measures, proposes another scoring function, which is referred to in this paper as the *NZ score*. Third, the scoring function proposed by Fissler et al. (2015) is considered. Taylor (2020) adjusts this score to better distinguish between the different forecasts by setting $a(\cdot) := ln(2)$ in Equation (19) instead of $a(\cdot)=0$. This adjusted scoring function is implemented in this research as well, and is referred to as the *FZG score*. Finally, the fourth score considered is proposed by Acerbi and Szekely (2014), and is referred to as the *AS score*. This score uses $G_1(x) = -\frac{1}{2}Wx^2$, for a parameter $W\in\mathbb{R}$, and is strictly consistent if this W satisfies $WQ_{j,t} < ES_{j,t}$ Fissler and Ziegel (2016). The smallest integer that guarantees this condition is satisfied for every pair of VaR and ES forecasts is $W^*\equiv 4$, such that this value is used for all AS scores in this research.

Table 1: Specifications of the functions used in the joint scoring function for the VaR and ES in Equation (19). These correspond to four different scores; the AL, NZ, FZG and AS scores. Also included in the Table are the literature from which these functions originated.

	$G_1(x)$	$G_2(x)$	$\zeta_2(x)$	$a(r_j)$	Literature
AL	0	$-\frac{1}{x}$	$-\ln(-x)$	$1-\ln(1-\alpha)$	Taylor (2019)
NZ	0	$\frac{1}{2}\ (-x)^{-\frac{1}{2}}$	$-(-x)^{\frac{1}{2}}$	0	Nolde and Ziegel (2017)
FZG	x	$\frac{\exp(x)}{1+exp(x)}$	$\ln(1+\exp(x))$	$\ln(2)$	Fissler et al. (2015)
AS	$-\frac{1}{2}Wx^2$	αx	$\frac{1}{2}\alpha x^2$	0	Acerbi and Szekely (2014)

3.5 Backtesting VaR and ES Forecasts

The accuracy of the Value-at-Risk and Expected Shortfall forecasts of all proposed methods is evaluated by means of calibration tests (Nolde and Ziegel, 2017), the scoring functions from Section 3.4, and the Model Confidence Set procedure of Hansen et al. (2011).

[30] The AS score is not used for estimation, as it cannot be guaranteed that this function is strictly consistent for all VaR and ES forecasts, given our chosen value of W.

3.5.1 Backtesting using Calibration Tests

The calibration tests are implemented in both their unconditional and conditional variants. The unconditional calibration test examines whether the fraction of observations that violate the VaR is equal to the nominal coverage probability, or $E\left(\mathbb{I}_{\{r_{j,t}<\hat{Q}_{j,t}^{(i)}\}}\right) = \alpha$. Defining the variable $\mathbb{H}_{j,t}^{(i)} := \alpha - \mathbb{I}_{\left\{r_{j,t}<\hat{Q}_{j,t}^{(i)}\right\}}$, this is equivalent to testing $E\left(\mathbb{H}_{j,t}^{(i)}\right) = 0$, which is achieved by means of binomial test[31]. The conditional calibration test examines whether the *conditional* expectation of the $\mathbb{H}_{j,t}^{(i)}$ variable is 0, or $E\left(\mathbb{H}_{j,t}^{(i)}|\mathcal{I}_{j,t}\right) = 0$, with $\mathcal{I}_{j,t}$ the information set at time t for series j. Conditional calibration is tested by implementing the Dynamic Quantile (DQ) test of Engle and Manganelli (2004) with four lags.

Then, the ES forecasts are evaluated using tests for zero mean and zero autocorrelation (AC). First, the procedure of McNeil and Frey (2000) tests whether, in periods where actual returns exceed the VaR, the ES forecasts are in expectation equal to these observed returns. The differences between these returns and ES forecasts are standardized by dividing by the corresponding VaR estimates, and are then tested for a zero unconditional expectation. Moreover, the dependent circular block bootstrap procedure of Jalal and Rockinger (2008) is also used to test for zero mean, as this eliminates the need for assumptions about the distribution of these standardized discrepancies. Finally, the ES forecasts are evaluated using a test for zero *autocorrelation* (AC), using the procedure of McNeil and Frey (2000).

3.5.2 Backtesting using Scoring Functions

The accuracy of the different VaR and ES forecasts can also be evaluated based on the scoring functions of Section 3.4. To this end, the values of each of the scoring functions for the different methods are expressed relative to the historical simulation benchmark, and the resulting values are referred to as *skill scores* in the remainder of this research. First, to evaluate VaR forecasts, the *quantile skill score* is a transformation of the quantile score of Equation (18), denoted $QS_{(i,j)} := S(\hat{Q}_{j,t}^{(i)},\ r_{j,t})$ for a method i, and is defined as $QSS_{j,t}^{(i)} := 100 \cdot \left(1 - \frac{QS_{(i,j)}}{QS_{(HS,j)}}\right)$, for all methods $i \in \mathcal{M}_0 \setminus \{HS\}$ excluding historical simulation[32]. As lower values of the quantile score are optimal, the construction of the corresponding skill score implies that higher skill score values are preferred. To obtain a more accurate and robust representation of a method's forecast accuracy, the performance of each method across is summarized across the different series. To this extent, the geometric mean of the ratios is computed across the different series, as $GM_t^{(i)} := \sqrt[J]{\prod_{j=1}^{J} R_{j,t}(i, HS)}$, where the ratio is defined as $R_{j,t}(i, HS) := \frac{QS_{(i,j)}}{QS_{HS,j}}$. The same transformation as before is then applied to these geometric means to obtain the *Geometric Mean Quantile Skill Score* of a method $i \in \mathcal{M}_0 \setminus \{HS\}$, as $GMQSS_t^{(i)} := 100 \cdot \left(1 - GM_t^{(i)}\right)$.

Then, as the scoring functions shown in Table 1 allow for the ES forecasts to be *jointly* evaluated with the VaR forecasts, skill scores are computed for these functions as well, using an analogous approach to that of the quantile skill score. For example, if $S_{(i,j)} := S(\hat{Q}_{j,t}^{(i)},\ \widehat{ES}_{j,t}^{(i)},\ r_{j,t})$ denotes the AL score in Equation 19, then the *AL Skill Score* is computed as $ALSS_{j,t}^{(i)} := 100 \cdot \left(\frac{S_{(i,j)}}{S_{(HS,j)}} - 1\right)$ for a method $i \in \mathcal{M}_0 \setminus \{HS\}$ simulation. Then, the geometric mean of the

[31] Potential errors in estimation are not incorporated, but this could be done (Escanciano and Olmo, 2010)

[32] By construction, the skill scores for the historical simulation method are zero.

ratio of AL scores is calculated, subtracted by 1 and multiplied by 100, which results in the *Geometric Mean AL skill scores*[33] for each method $i \in \mathcal{M}_0 \setminus \{HS\}$. Skill scores are also constructed for the NZ, FZG and AS scores. As the NZ and AS scores take positive values, their skill scores can be computed similarly to the quantile skill score, while the FZG score is negatively valued, and its skill score is computed like the AL skill score.

3.5.3 Backtesting using Model Confidence Sets

The Model Confidence Set (MCS) procedure proposed by Hansen et al. (2011) serves as the final evaluation technique in this research. This procedure aims to iteratively reduce the size of the set of all methods to a smaller set having a predetermined probability of containing the best forecasting method, as evaluated by a given loss function. In alignment with Hansen et al. (2011), confidence levels of 75% and 90% are considered. An *equivalence test* is used to test for significant differences in forecast accuracy within the set of methods, for which the Diebold-Mariano test (Diebold and Mariano, 2002) is utilized. In each iteration, an *elimination rule* determines which method, if any, should be removed from the current set. This research adopts the $T_{max,M}$ statistic from Hansen et al. (2011) as the elimination rule, which is the maximum value of the standardized means of loss function differences. The forecast accuracies are measured using the scoring functions from Section 3.4. Given that VaR and ES forecasts are constructed for multiple indices, the accuracy of each method is quantified by the number of series for which each method is included in the MCS, such that higher numbers are more desirable.

4 Results

This Section presents and discusses the main results of this research, which are results of the VaR and ES forecast evaluation procedures explained in Section 3.5. The results of the skill scores of Section 3.5.2 are discussed first. Then, the results involving the hypothesis are examined, which are the Calibration tests of Section 3.5.1, and the Model Confidence Sets of Section 3.5.3. Finally, a summary of the most important findings is provided.

4.1 Results of the Scoring Functions

Table 2 presents the results of the skill scores of the different scoring functions of Section 3.5.2 for the forecasts at both probability levels. For completeness, Tables A.4 and A.5 in the Appendix provide a series-by-series breakdown, using the quantile score and AL score, respectively.

As a first observation, the individual methods display quite polarizing results, in the sense that there are noticeable differences in forecast accuracies across the different methods. For both probability levels and across all scoring functions, the CAViaR-AS-EVT model consistently outperforms all other individual methods, with the GJR-GARCH and GAS models also producing reliable forecasts. However, the Cornish-Fisher Expansion and especially the CARE-AS model

[33] As the AL score is negative-valued, its skill score is constructed slightly differently from the quantile score. Consequently, larger skill score values are again preferred, such all skill scores are easily comparable.

Table 2: For each (combination of) asset(s), the evaluation of the VaR according to the quantile skill score (in %) of Section 3.5.1, and the joint evaluation of the VaR and ES using the skill scores (in %) of the four scoring functions of Table 1, as defined in Section 3.5.2.

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
Individual methods										
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	19.3	8.2	11.3	0.6	31.9	9.6	3.9	6.2	0.2	15.2
Cornish-Fisher	−11.0	−3.0	−5.2	−0.4	−19.8	−1.6	−0.8	−1.3	0.0	−2.9
CARE-AS	−44.0	−72.1	−47.7	−1.3	−21.2	7.3	2.1	4.2	0.1	12.8
CAViaR-AS-EVT	21.0	8.5	12.3	0.6	33.9	9.9	4.0	6.4	0.2	15.7
GAS model	16.5	4.3	8.7	0.5	29.5	8.0	2.9	5.1	0.2	13.2
Combining all										
Simple average	18.7	9.8	12.0	0.6	27.5	9.4	4.3	6.4	0.2	14.2
Median	17.9	8.7	11.1	0.5	27.0	9.8	4.2	6.5	0.2	15.5
Mode	15.5	6.7	9.2	0.4	24.7	9.5	3.9	6.2	0.2	15.2
Winsorizing	20.6	10.5	13.1	0.6	32.0	10.1	4.5	6.7	0.2	15.9
Trimming	18.3	9.2	11.6	0.5	27.3	9.5	4.3	6.4	0.2	14.7
Relative score	20.4	8.7	12.1	0.6	32.6	10.1	4.2	6.6	0.2	16.0
Minimum score	18.9	8.9	11.6	0.6	29.7	9.7	4.3	6.5	0.2	15.5
Combining all except historical simulation										
Simple average	19.8	10.2	12.6	0.6	29.8	9.8	4.4	6.6	0.2	15.3
Median	18.6	8.0	11.0	0.5	30.1	9.8	4.0	6.3	0.2	15.7
Mode	18.2	6.8	10.3	0.5	30.5	9.7	3.9	6.2	0.2	15.6
Winsorizing	20.5	10.2	12.9	0.6	32.3	10.1	4.4	6.7	0.2	16.2
Trimming	18.6	8.1	11.1	0.5	30.0	9.9	4.2	6.5	0.2	15.8
Relative score	20.7	9.1	12.4	0.6	32.9	10.0	4.2	6.5	0.2	15.9
Minimum score	19.3	8.6	11.6	0.6	30.5	9.8	4.2	6.5	0.2	15.6

Notes: The values in this Table represent the average of each skill across all 6 assets (or combinations of assets in case of the portfolio). The colored cells indicate the best method for each scoring function.

substantially underperform in 1% forecasts of the VaR and ES, being outperformed by the Historical Simulation benchmark. While the CARE-AS model produces decent (but uncompetitive) forecasts at the 5% probability level, its decrease in performance at the lower probability level of 1% implies an inadequacy in its ability to forecast more extreme events. Potentially, Extreme Value Theory could be used to increase the forecasting utility of the CARE model, as it did for the CAViaR model. The poor forecast of the CFE can be explained by the slow and inadequate updating of the volatility. The sample variance is used to transform the standardized Cornish-Fisher statistics, even though the volatility is known to be time-varying and partially predictable. Future research could investigate the use of combining the volatility updating of the (GJR-)GARCH or GAS model, with the incorporation of the higher order moments as accomplished by the CFE.

Second, considering the combining methods, the clear winner is the winsorized mean, for both sets of methods, where a preference for any of the two sets is not apparent. Even if not all individual methods are equally competitive, including a sufficient number of them seems to improve, or at least not undermine, the accuracy of the resulting winsorized mean forecast. This makes sense if each individual method contains useful information about the true underlying data, as information from all the forecasts is still incorporated, but the effect of the more extreme observations on the mean is reduced. Furthermore, the relative score combining and the simple average, for both sets of methods, seem to produce accurate forecasts as well, proving to be strong alternatives for the winsorized mean. Comparing the individual and combining methods, it is noticeable that the winsorized mean (for both sets of methods) is the best choice, based

on all scoring functions and both probability levels, The CAViaR-AS-EVT follows as a strong competitor that requires only one model to be estimated, and the simple average and relative score combining (for both sets) are reliable/robust alternatives.

Furthermore, the results are relatively stable across the proposed scoring functions, demonstrating their competence and usefulness in backtesting Value-at-Risk and Expected Shortfall forecasts. Then, with the exception of the CARE-AS model, the results are comparable across the two probability levels. Future research could investigate the sensitivity of CARE-AS model to changes in the probability level and the parameter estimation process, as updating the parameters more frequently might enhance the forecast accuracy. Finally, for completeness, Tables A.4 and A.5 in the Appendix respectively display the results of the quantile and AL scoring function on a series-by-series basis. The results are consistent across the different series as well, which suggests that our findings can be generalizable to different assets, and therefore to different portfolio compositions as well. Finally, Figures A.2 and A.3 in the Appendix show the combining weights for the Minimum Score and Relative Score combining methods, respectively. These weights reflect the contributions of the different individual methods, and are in line with the results of Table 2. A more detailed discussion about these weights is provided in Section A.5 in the Appendix.

4.2 Results of the Statistical Tests

This Subsection discusses the results of the Calibration tests of Section 3.5.1 and the Model Confidence Sets of Section 3.5.3. Table 3 presents the results of the Calibration tests for the 1% and 5% VaR and ES, where the values indicate the number of series for which the (type of) calibration of each method is rejected. Table 4 presents the results of the Model Confidence Sets (MCS) at a 90% confidence level, showing for how many series each method was in the Model Confidence Set. For completeness, the results of the MCS using a confidence level of 75% are presented in Table A.6 in the Appendix, which shows comparable findings. In general, evaluating the VaR and ES forecasts based on the calibration tests and MCS procedures results in very similar findings to those of Section 4.1. Thus, this Subsection mainly examines the differences in the results, and the additional insights that are gained from these statistical testing procedures.

First, and most importantly, the results of the Model Confidence Sets in Table 4 display unmistakably the robustness and accuracy of the combined forecasts, with each type of forecast combination being in the Model Confidence Set for *all* series. While the best individual methods, the CAViaR and GARCH models, are able to keep up with combining methods as a whole, the benefits of combining forecasts cannot be overlooked. The MCS at 75%, as shown in Table A.6, supports these findings, and shows a slight superiority of most of the combining methods over even the best individual methods. Furthermore, the MCS results provide more support for the GAS model, showing it provides fairly reliable forecasts, especially at the 1% probability level. Then, Table 3 shows that the combining methods have slightly better calibrated 5% forecasts than 1% forecasts, perhaps implying that combinations are more useful in less extreme scenarios. Finally, as the skill scores of Section 3.5.2 are computed as ratios to the score of the historical simulation, no conclusions could be drawn from Table 2 about the performance of the historical simulation. Considering the Model Confidence Sets of Table 4, the historical simulation forecasts

seem to be quite reasonable, although Table 3 shows they are not correctly calibrated. The tests for zero autocorrelation in the relative discrepancies of the ES show that only the historical simulation and Cornish-Fisher Expansion are inadequate, in the sense that their ES forecasts seem to have more autocorrelation that the ES forecasts of all other methods. The remainder of the conclusions to be drawn from the calibration tests and Model Confidence Sets are similar to the results of Section 4.1, and are left out in order to avoid uninformative repetition of findings.

Table 3: Results of the calibration tests of Section 3.5 for the VaR and ES forecasts at both probability levels.

	1% Probability Level					5% Probability Level				
	VaR hit %	VaR DQ	ES mean*	ES mean**	ES AC	VaR hit %	VaR DQ	ES mean*	ES mean**	ES AC
Individual methods										
Historical simulation	5	6	4	4	1	1	6	4	2	2
GJR-GARCH	5	2	3	3	0	3	2	6	5	0
Cornish-Fisher	4	6	5	4	1	2	6	6	1	4
CARE-AS	6	6	6	6	0	6	6	6	5	0
CAViaR-AS-EVT	2	3	0	0	0	2	1	0	2	0
GAS model	5	5	6	5	0	3	3	6	5	0
Combining all										
Simple Average	1	4	5	4	1	1	3	2	1	1
Median	5	3	3	3	0	3	1	2	3	0
Mode	5	3	3	3	0	3	1	2	3	0
Winsorizing	0	1	0	0	0	1	2	0	0	0
Trimming	4	3	2	4	0	2	1	1	1	0
Relative score	2	2	3	2	0	2	0	4	4	0
Minimum score	3	3	4	4	0	2	1	0	0	0
Combining all except historical simulation										
Simple Average	1	3	6	5	0	1	0	2	1	1
Median	5	3	4	4	0	5	0	6	5	0
Mode	5	3	4	4	1	5	1	6	5	0
Winsorizing	2	2	1	1	0	1	0	2	2	0
Trimming	5	4	4	4	0	4	0	5	4	0
Relative score	2	2	2	1	0	3	1	3	2	0
Minimum score	1	3	6	5	0	2	1	2	1	0

Notes: The values in this Table represent the number of series for which the given test was significant at the 5% significance level. For any forecaster, lower values are preferred, as they imply fewer rejections of calibration.
(∗) This refers to test of McNeil and Frey (2000) for a zero mean in the relative discrepancies of the ES.
(∗∗) This test incorporates the dependent circular block bootstrap procedure of Jalal and Rockinger (2008).

4.3 Summary of Results

The results of this Section are now summarized. First, the CAViaR-AS-EVT is found to produce the most accurate VaR and ES forecasts out of any individual method, with the GJR-GARCH and GAS model being adequate alternatives. The similarities in the results of the GAS and GARCH models are not too unexpected, as the the same distribution (student's t) and scaling (identity) is used, and the differences thus stem from the volatility updating, as in Equation (4). The Cornish-Fisher Expansion and CARE-AS model both performed quite poorly. The inadequacy of the CFE could perhaps partially be explained by its slow and imprecise updating of the volatility, which uses a large window to repeatedly recalculate the sample standard deviation. Perhaps using a GARCH type model (EWMA, GAS are possible variants) for this volatility, and

Table 4: For each (combination of) asset(s), the Model Confidence Set (MCS) procedure of Hansen et al. (2011) is applied to evaluate the VaR separately based on the quantile score, and the VaR and ES jointly based on the four scoring functions in Table 1, being the AL, NZ, FZG and AS, respectively. A confidence level of 90% is used.

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
Individual methods										
Historical simulation	4	5	5	4	4	3	3	3	3	3
GJR-GARCH	6	6	6	6	6	6	6	6	6	6
Cornish-Fisher	1	2	1	1	1	1	1	1	1	1
CARE-AS	0	0	0	0	0	2	1	1	2	4
CAViaR-AS-EVT	6	6	6	6	6	6	6	6	6	6
GAS model	6	6	6	6	6	5	4	4	5	5
Combining all										
Simple average	6	6	6	6	6	6	6	6	6	6
Median	6	6	6	6	6	6	6	6	6	6
Mode	6	6	6	6	6	6	6	6	6	6
Winsorizing	6	6	6	6	6	6	6	6	6	6
Trimming	6	6	6	6	6	6	6	6	6	6
Relative score	6	6	6	6	6	6	6	6	6	6
Minimum score	6	6	6	6	6	6	6	6	6	6
Combining all except historical simulation										
Simple average	6	6	6	6	6	6	6	6	6	6
Median	6	6	6	6	6	6	6	6	6	6
Mode	6	6	6	6	6	6	6	6	6	6
Winsorizing	6	6	6	6	6	6	6	6	6	6
Trimming	6	6	6	6	6	6	6	6	6	6
Relative score	6	6	6	6	6	6	6	6	6	6
Minimum score	6	6	6	6	6	6	6	6	6	6

Notes: The quantile score is defined in Equation (18), and the four scoring functions are defined in Table 1, with brief descriptions given in 3.4. The values in this Table are the number of assets for which each method was located in its Model Confidence Set (of level 90%), such that higher values are preferred.

plugging this estimated volatility into the CFE, could result in more accurate forecasts. This remains to be researched. The CARE-AS model might simply require a much more frequent update of parameters, as Taylor (2020) found promising using this model when re-estimating parameters every period. However, due to computational constraints, this was not feasible in our research. This is also possible reason for future research.

Regarding the large set of combination methods, the winsorized mean stood out as the clear winner, in particular using the set of all individual methods. It is followed by the simple average and the relative score combining, although the difference between the different combination methods is less clear-cut than for the individual methods. These combining methods display superiority over the best individual methods, with all remaining forecast combination techniques being at least competitive with the individual methods. This clearly shows the benefit of using forecast combinations for accurately forecasting the Value-at-Risk and Expected Shortfall. Finally, the consistent results across the different proposed scoring functions imply an adequacy of these scoring functions to accurately evaluate different forecasts.

5 Conclusion

This research aims to explore how combined Value-at-Risk and Expected Shortfall forecasts can improve the accuracy of downside risk forecasts for a diversified portfolio. Extending the influential work of Taylor (2020), a variety of individual methods and forecast combination techniques are applied to financial asset returns from different asset classes, as well as a diversified portfolio. Evaluating the VaR and ES forecasts is not straightforward, as the actual VaR and ES are not observed, and the ES is not elicitable independently from the VaR. Consequently, joint scoring functions are employed that are strictly consistent for the VaR and ES jointly, providing a robust framework for the evaluation of different downside risk forecasts.

Among the considered individual methods, the Asymmetric Slope CAViaR model with Extreme Value Theory (CAViaR-AS-EVT) produces the most accurate forecasts. Nevertheless, the forecast combination techniques unequivocally outperform the individual methods. Out of all methods considered in this research, the winsorized mean, especially when all individual methods are included, proves to be the most accurate. Remarkably, the majority of these combinations are on par with the most accurate individual method. This underscores the benefits of combining individual Value-at-Risk and Expected Shortfall forecast to improve downside risk forecasts. Moreover, our research adds another piece to the forecast combination puzzle by demonstrating superior performance of the simple averaging techniques compared to the sophisticated combination procedures such as the Relative and Minimum Score combining methods.

In conclusion, our empirical analysis demonstrates that combining individual Value-at-Risk and Expected Shortfall forecasts can substantially improve the accuracy of the resulting downside risk forecasts for a diversified portfolio. This accentuates the unique ability of combining methods to incorporate the different insights gained from a set of individual forecasting methods.

Despite the valuable insights this research provides, there remain areas that can benefit from future research. First, additional individual methods and different sets thereof could be considered, as this could enable a clearer distinction between the various combining methods. Second, additional combination techniques could be implemented, like Machine Learning (ML) or Bayesian techniques. The use of ML techniques in creating Value-at-Risk and Expected Shortfall forecasts is still relatively novel, and warrants further exploration. Including covariates that are not directly related to past returns could potentially enhance forecast accuracy, by allowing different information to be incorporated. Another interesting area is the multi-horizon setting, where combining weights could either be estimated jointly over different horizons, or separately for each horizon. In this case, the multi-horizon MCS (MH-MCS), based on tests for Superior Predictive Ability (SPA), could be used to evaluate these forecasts. In line with convention, considering different and larger datasets are a possibility, with intra-day data being an interesting option. Finally, one could consider performing multivariate analysis, on portfolios that are constructed by mean-variance optimization, for example using Factor Models. These suggestions hopefully aid in extending and continuing this line of work.

Acknowledgments

We express our gratitude to Professor James W. Taylor for generously sharing replication packages and granting permission to build upon his influential work in Taylor (2020). This laid the foundation of our research, and influenced its development. Our appreciation extends to Mr. Bram van Os, whose insights, suggestions, and constructive feedback greatly improved the quality of this paper. Furthermore, we acknowledge the Erasmus University of Rotterdam for providing necessary resources, in particular access to the Bloomberg terminal, which enabled us to collect the data used in the empirical analysis. The availability of this high-quality data greatly facilitated the research process, and it enriched our empirical findings. Finally, while we have been fortunate to receive support from various individuals and institutions, the authors take responsibility for any remaining errors or shortcomings present in this paper.

References

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

Acerbi, C. and Tasche, D. (2002). Expected shortfall: a natural coherent alternative to value at risk. *Economic notes*, 31(2):379–388.

Alexander, C. (2009). *Market Risk Analysis, Value at Risk Models*. John Wiley & Sons.

Amédée-Manesme, C.-O., Barthélémy, F., and Maillard, D. (2019). Computation of the corrected Cornish–Fisher expansion using the response surface methodology: application to VaR and CVaR. *Annals of Operations Research*, 281:423–453.

Ardia, D., Boudt, K., and Catania, L. (2016). Generalized autoregressive score models in R: The GAS package. *arXiv preprint arXiv:1609.02354*.

Ardia, D., Boudt, K., and Catania, L. (2018). Downside Risk Evaluation with the R Package GAS. *The R Journal*, 10(2):410–421.

Azzalini, A. and Genton, M. G. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76(1):106–129.

Batchelor, R. and Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41(1):68–75.

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the operational research society*, 20(4):451–468.

Bell, W. R. and Huang, E. T. (2006). Using the t-distribution to deal with outliers in small area estimation. In *Proceedings of Statistics Canada Symposium*.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

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

Chen, J. M. (2018). On exactitude in financial regulation: Value-at-risk, expected shortfall, and expectiles. *Risks*, 6(2):61.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

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

Cornish, E. A. and Fisher, R. A. (1938). Moments and cumulants in the specification of distributions. *Revue de l'Institut international de Statistique*, pages 307–320.

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

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

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.

Escanciano, J. C. and Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business & Economic Statistics*, 28(1):36–51.

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

Fissler, T., Ziegel, J. F., and Gneiting, T. (2015). Expected Shortfall is jointly elicitable with Value at Risk-Implications for backtesting. *arXiv preprint arXiv:1507.00244*.

Gao, F. and Song, F. (2008). Estimation risk in GARCH VaR and ES estimates. *Econometric Theory*, 24(5):1404–1424.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

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

Jalal, A. and Rockinger, M. (2008). Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. *Journal of Empirical Finance*, 15(5):868–877.

Jones, M., Linton, O., and Nielsen, J. (1995). A simple bias reduction method for density estimation. *Biometrika*, 82(2):327–338.

Jose, V. R. R. and Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International journal of forecasting*, 24(1):163–169.

Khokhlov, V. (2016). Conditional value-at-risk for elliptical distributions. *Evropskỳ časopis ekonomiky a managementu*, 2(6):70–79.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kourentzes, N., Barrow, D. K., and Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, 41(9):4235–4244.

Maillard, D. (2018). A user's guide to the Cornish Fisher expansion. *Available at SSRN 1997178*.

Marling, H. and Emanuelsson, S. (2012). The markowitz portfolio theory. *November*, 25:2012.

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.

Narasimhan, B., Koller, M., Johnson, S. G., Hahn, T., Bouvier, A., Kiêu, K., Gaure, S., and Narasimhan, M. B. (2023). Package 'cubature'.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Nolde, N. and Ziegel, J. F. (2017). REJOINDER: "ELICITABILITY AND BACKTESTING: PERSPECTIVES FOR BANKING REGULATION". *The annals of applied statistics*, 11(4):1901–1911.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of econometrics*, 211(2):388–413.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pages 832–837.

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

Sheikh, A. Z. and Qiao, H. (2010). Non-normality of market returns: A framework for asset allocation decision making. *The Journal of Alternative Investments*, 12(3):8.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sinova, B., Gil, M. Á., Colubi, A., and Van Aelst, S. (2012). The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets and Systems*, 200:99–115.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford bulletin of economics and statistics*, 71(3):331–355.

Tay, A. S. and Wallis, K. F. (2000). Density forecasting: a survey. *Journal of forecasting*, 19(4):235–254.

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.

Taylor, J. W. and Yu, K. (2016). Using auto-regressive logit models to forecast the exceedance probability for financial risk management. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 1069–1092.

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265.

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

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

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2022). Forecast combinations: an over 50-year review. *International Journal of Forecasting*.

Zikovic, S. and Filer, R. K. (2012). Ranking of var and es models: performance in developed and emerging markets.

A Appendix

A.1 Overview of abbreviations

For completeness, Table A.1 provides a non-exhaustive overview of commonly used abbreviations in this research, along with their meanings.

Table A.1: A non-exhaustive list of the most frequently used abbreviations, and their meanings.

Abbreviation	**Meaning**
VaR	Value-at-Risk
ES	Expected Shortfall
AR(1)	AutoRegressive model of order 1
H.S.	Historical Simulation
GARCH	Generalized AutoRegressive Conditional Heteroscedasticity
GJR	Glosten-Jagannathan-Runkle, refers to the asymmetric variant of the GARCH model
CFE	Cornish-Fisher Expansion
CARE	Conditional AutoRegressive Expectiles
AS	Asymmetric Slope
CAViaR	Conditional Autoregressive Value at Risk
EVT	Extreme Value Theory
GAS	Generalized Autoregressive Score
MCS	Model Confidence Set
AL score	Joint scoring function of Taylor (2019)
NZ score	Joint scoring function of Nolde and Ziegel (2017)
FZG score	Joint scoring function of Fissler et al. (2015)
AS score	Joint scoring function of Acerbi and Szekely (2014)
DQ	Dynamic Quantile test of Engle and Manganelli (2004)
AC	Autocorrelation
ML	Machine Learning
RSM	Response Surface Methodology
KDE	Kernel Density Estimation S&P 500
Standard & Poor's 500 stock index	
FTSE 100	Financial Times Stock Exchange 100 index
AGG	iShares Core U.S. Aggregate Bond ETF
GSCI	(Standard and Poor's) Goldman Sachs Commodity Index
EURUSD BGN curncy	Exchange rate of the Euro against the United States Dollar

A.2 Calculating Portfolio Returns

This Section explains how to retrieve the log returns of a portfolio, denoted $y_{p,t}$, from the log-returns of the individual assets, denoted $y_{j,t},\ j = 1, ..., J$. Assuming that the investor invests a fraction $w_{j,t}$ into asset j at time t, such that these fractions form a convex set of weights, i.e. $\min_{j=1,...,J}\{w_{j,t}\} \geq 0,\ \sum_{j=1}^{J} = 1,\ \forall t$. The simple return of an asset j, denoted $R_{j,t}$, is defined as $R_{j,t} := \frac{P_{j,t}-P_{j,t-1}}{P_{j,t-1}}$, with $P_{j,t}$ the price of asset j at time t. This implies the relations $y_{j,t} \equiv \ln(1 + R_{j,t})$ and $R_{j,t} \equiv \exp\{y_{j,t}\} - 1$ between the log returns and simple returns.

The simple returns $R_{j,t}$ have the desirable characteristic that they can be linearly aggregated across assets (whereas log returns cannot), in the sense that the simple return of the portfolio is the weighted average of the returns of the individual assets it consists of, with the weights being the fractions of wealth invested into each asset. Mathematically, $R_{p,t} \equiv \sum_{j=1}^{J} w_{j,t} \cdot R_{j,t}$, where $R_{p,t}$ is the simple return of the portfolio, and $w_{j,t}$ is the weight of/fraction of wealth invested into asset j at time t. Therefore, to acquire the log returns of a portfolio based on the log returns

of the individual assets j, one can perform the following procedure:

1. Calculate the simple returns of the individual assets j, being $R_{j,t}$, from their log returns, as follows: $R_{j,t} = e^{y_{j,t}} - 1,\ j = 1, ..., J$.

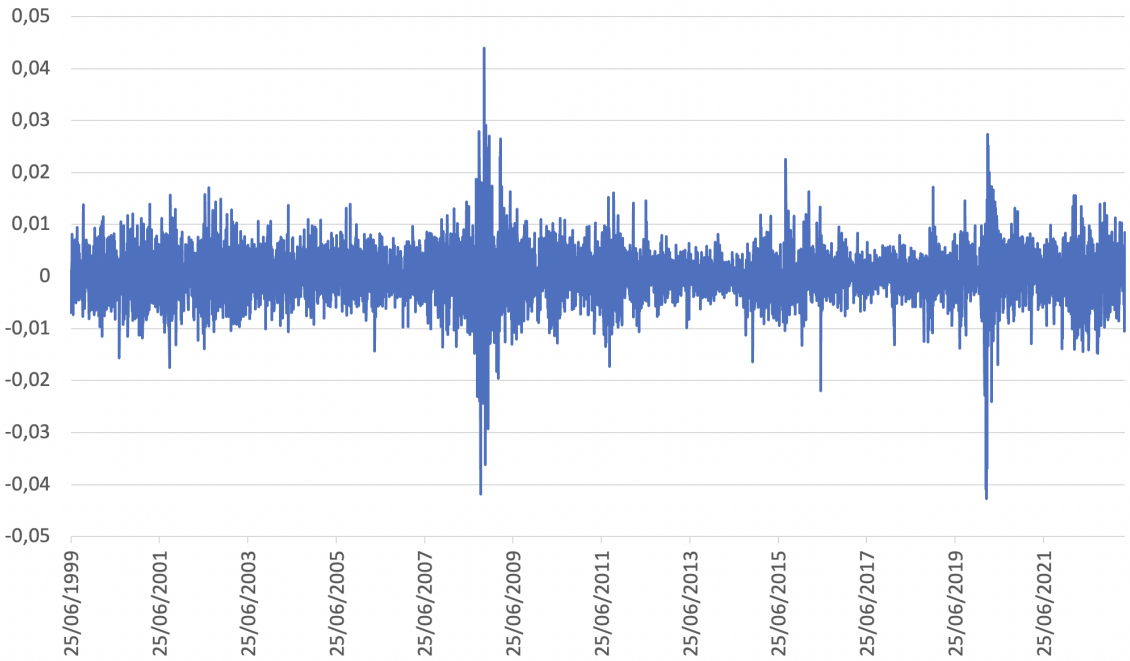
2. For each period t, retrieve the Simple Return of the Portfolio, by calculating the weighted average of simple returns on the different assets; $R_{p,t} := \sum_{j=1}^{J} w_{j,t} \cdot R_{j,t}$, with the weights $w_{j,t}$ being the fractions of wealth invested into the different assets.

3. For each time period t, convert the Simple Return of the Portfolio, $R_{p,t}$, into the log return of the portfolio, denoted $y_{p,t}$, as follows: $y_{p,t} = \ln(1 + R_{p,t})$.

4. Use the series $\{y_{p,t} :\ t = 1, ..., T\}$ (or, more specifically, the residuals of the AR(1) model for $y_{p,t}$) as part of the empirical analysis.

In our research, the series of log returns of the portfolio ($y_{p,t}$) corresponds to the sixth series, with a total of $J = 6$ series. The portfolio considered uses weights $\boldsymbol{w}_t := [w_{1,t}, w_{2,t}, ..., w_{6,t}]' \equiv [0.30, 0.30, 0.30, 0.05, 0.05]'$, meaning 60% of wealth is invested into the equity market, 30% into the fixed income market, 5% into commodities and 5% into the Foreign Exchange Market. These weights are kept constant over time (i.e. $\boldsymbol{w}_t \equiv \boldsymbol{w}, \forall t$). The resulting series of 6000 log returns for this portfolio are depicted in Figure A.1. This Figure clearly displays several stylized facts of asset returns, like volatility clustering and intermittency (Cont, 2001). Furthermore, the volatility is substantially larger during the periods around 2008 and 2020 is, corresponding to the Great Recession and the start of the COVID-19 virus, respectively.

Figure A.1: The log returns for the portfolio of Section 2, ending on April 28th, 2023.



Notes: These log returns are constructed using the log returns of the five series from Section 2, using the procedure explained in this Section. Formally, the series depicted corresponds to the set of 6000 returns $\{y_{p,t} : t = 1, ..., T\}$ constructed using the procedure described earlier in this Section.

A.3 Response Surface Methodology

This Section provides a more complete overview of the Response Surface Methodology (RSM), which is used to estimate the Skewness and Kurtosis parameters for the Cornish-Fisher Expansion of Section 3.2.5.

Response Surface Methodology refers to a selection of statistical techniques applied to models or problems where several variables influence a certain response variable of interest. A mathematical model is constructed in order to explore the relation between the chosen response variable and the set of predictor variables, with the approximated relation being straightforward to estimate. A function is fitted to the data, and optimization techniques are applied to find the optimal parameters, resulting in a model that is often less time-consuming than other modeling techniques Amédée-Manesme et al. (2019). Therefore, the benefits of RSM are noticed most in large-scale applications that involve time-consuming modeling tasks. Sequentially, a heuristic is used to explore local subareas of the global area of validity.

In case of the Cornish-Fisher Expansion, this RSM is used to estimate the Skewness and Kurtosis parameters that are used to construct the VaR and ES forecasts (as in Equations (8) and (9)) as functions of the observed Skewness and Kurtosis. Hence, the Skewness and Kurtosis parameters are the response variables), and (transformations of) the observed Skewness and Kurtosis are the explanatory variables. We follow the approach of Amédée-Manesme et al. (2019), which involves fitting first-order polynomials in the (transformations of) the observed Skewness and Kurtosis, per local area. Ordinary Least Squares (OLS) is applied to estimate the parameters, and Analysis Of Variance (ANOVA) is used to assess the significance of observed differences in variation. Following experimentation, Amédée-Manesme et al. (2019) split the domain into 5 subareas, based on the obseverd Skewness and Kurtosis. In each subset, the following polynomial models for the response variables S_c and K_c are estimated:

$$\begin{aligned} E(S_c|S,K) = {} & \delta + \gamma_1 S^{1/2} + \gamma_2 K^{1/2} + \gamma_3 S + \gamma_4 K + \gamma_5 S^{1/2}K^{1/2} + \gamma_6 S^{3/2} + \gamma_7 K^{3/2} \\ & + \gamma_8 S^{1/2}K + \gamma_9 SK^{1/2} + \gamma_{10} S^2 + \gamma_{11} K^2 + \gamma_{12} SK + \gamma_{13} S^{3/2}K^{1/2} \\ & + \gamma_{14} S^{3/2}K^{1/2} + \gamma_{15} SK^2 + \gamma_{16} S^2 K + \gamma_{17} S^{3/2}K^{3/2} + \gamma_{18} \ln(S)\ln(K) \\ & + \gamma_{19} \ln(S)K + \gamma_{20} S\ln(K) + \gamma_{21} S^{-1} + \gamma_{22} K^{-1}. \end{aligned} \tag{20}$$

$$\begin{aligned} E(K_c|S,K) = {} & \alpha + \beta_1 S^{1/2} + \beta_2 K^{1/2} + \beta_3 S + \beta_4 K + \beta_5 S^{1/2}K^{1/2} + \beta_6 S^{3/2} + \beta_7 K^{3/2} \\ & + \beta_8 S^{1/2}K + \beta_9 SK^{1/2} + \beta_{10} S^2 + \beta_{11} K^2 + \beta_{12} SK + \beta_{13} S^{3/2}K^{1/2} \\ & + \beta_{14} S^{3/2}K^{1/2} + \beta_{15} SK^2 + \beta_{16} S^2 K + \beta_{17} S^{3/2}K^{3/2} + \beta_{18} \ln(S)\ln(K) \\ & + \beta_{19} \ln(S)K + \beta_{20} S\ln(K) + \beta_{21} S^{-1} + \beta_{22} K^{-1}. \end{aligned} \tag{21}$$

Tables A.2 and A.3 are retrieved from Amédée-Manesme et al. (2019), and present the parameter estimates of δ and γ_i ($i \in \mathbb{N}_{22}$) of Equation (20), and α and β_i ($i \in \mathbb{N}_{22}$) of Equation (21), respectively, for the 5 considered subcases. These cases are defined based on the values of the observed Skewness and Kurtosis, and are also shown in the Tables. Finally, the Skewness and Kurtosis parameters are estimated as $\hat{S}_c$ and $\hat{K}_c$, using the observed Skewness and Kurtosis and the estimates of Tables A.2 and A.3.

Table A.2: Response Surface estimator of the Skewness parameter according to the 5 subsets.

	Case 1	Case 2	Case 3	Case 4	Case 5
	$0.5 \leq S \leq 2.2$ $5 \leq K \leq 40$	$0 < S \leq 0.5$ $5 \leq K \leq 40$	$S \geq 0.5$ $K \leq 5$	$0.25 \leq S < 0.5$ $K \leq 5$	$0 < S < 0.25$ $K \leq 5$
Constant	-1.816	-0.0189	2.111	0.172	0.00512
$S^{1/2}$	6.812	0.161	-	0.132	-0.0240
$K^{1/2}$	-0.577	0.0215	-3.498	-0.296	-0.00778
S	-8.636	0.453	-2.870	-	1.277
K	0.508	0.00139	-0.123	-0.0415	0.00499
$S^{1/2}K^{1/2}$	-	-0.0862	3.836	0.346	0.0386
$S^{3/2}$	4.235	0.326	2.956	1.491	-0.114
$K^{3/2}$	-0.00685	-0.00000851	-0.162	-0.0327	-0.000479
$S^{1/2}K$	-0.848	-0.00168	-	-	-0.0336
$SK^{1/2}$	2.671	0.230	-	-	-0.483
S^2	-0.0969	-0.0136	2.008	0.134	0.265
K^2	-0.000304	0.00000232	0.0370	0.00278	-0.0000520
$S^{3/2}K^{1/2}$	-1.259	-0.129	-4.884	-1.330	-0.0857
SK	0.226	-0.000326	1.720	0.249	0.109
$S^{1/2}K^{3/2}$	0.0191	-0.000151	-0.153	0.0333	0.00708
SK^2	0.0249	0.00662	0.239	0.205	-0.0332
$S^{3/2}K^{3/2}$	-0.00666	-0.000649	-0.0883	-0.0597	0.0161
$\ln(S)\ln(K)$	-0.105	0.00396	-0.227	-0.0109	-0.000270
$\ln(S)K$	0.0987	0.000457	-0.436	-0.0507	0.000262
$S\ln(K)$	-0.845	-0.221	0.700	0.114	0.0513
S^{-1}	0.135	0.000228	-0.0739	-0.00419	0.000000429
K^{-1}	-0.416	-0.0250	0.0414	0.00152	0.000110

Notes: The values in this Table are the parameter estimates of Equation (20), which is the polynomial model for the Skewness parameter, S_c. This Table is retrieved from Amédée-Manesme et al. (2019).

A.4 Kernel Density Estimation

This Section provides a more elaborate explanation about Kernel Density Estimation (KDE), in particular with a Gaussian kernel, that is applied to construct the mode ensemble forecasts of Section 3.3.1. The KDE is a non-parametric approach to estimate the probability density function (PDF) of a given stochastic variable, in our case the forecasts of the individual methods from Section 3.2. It is often applied in data science and Machine Learning, being used for classification problems and regressions. KDE enables us to approximate the distributions of the individual forecasts without making assumptions, by applying a kernel function to smooth the data. With forecasts of an unknown density function f, the following kernel density estimator can be used to approximate the shape of f:

$$\hat{f}_{t,h}^{(i)}(x) = \frac{1}{|\mathcal{M}| \cdot h} \sum_{i=1}^{|\mathcal{M}|} K\left(\frac{x - \hat{y}_{j,t}^{(i)}}{h}\right), \tag{22}$$

where $\hat{y}_{j,t}^{(i)}$ is the forecast of model $i \in \mathcal{M}$, and $K(\cdot)$ is the kernel with bandwidth h. A popular choice for this kernel is the Gaussian one (Kourentzes et al., 2014), which assumes the data is normally distributed around each point. Although this assumption of normality may seem unrealistic[34], the Gaussian kernel has some attractive computational features, and the resulting mode operator is still robust to deviations from normality (Tay and Wallis, 2000). For these

[34]In reality, the specification chosen for the kernel is often found to affect the outcomes only minimally (Jones et al., 1995).

Table A.3: Response Surface estimator of the Kurtosis parameter according to the 5 subsets.

	Case 1	Case 2	Case 3	Case 4	Case 5
	$0.5 \leq S \leq 2.2$ $5 \leq K \leq 40$	$0 < S \leq 0.5$ $5 \leq K \leq 40$	$S \geq 0.5$ $K \leq 5$	$0.25 \leq S < 0.5$ $K \leq 5$	$0 < S < 0.25$ $K \leq 5$
Constant	−5.962	0.0832	1.749	−1.612	−0.304
$S^{1/2}$	21.53	0.0451	–	1.894	0.743
$K^{1/2}$	−1.548	0.732	−6.604	1.938	0.597
S	−26.52	−0.601	3.425	–	−1.662
K	1.820	0.124	1.313	0.273	0.676
$S^{1/2}K^{1/2}$	–	0.396	7.491	−1.018	−1.073
$S^{3/2}$	11.08	1.261	−11.83	−4.220	0.226**
$K^{3/2}$	−0.0443	−0.0195	−0.858	−0.141	−0.299
$S^{1/2}K$	−2.564	−0.0704	–	–	0.490
$SK^{1/2}$	5.739	−0.528	–	–	2.314
S^2	0.342	−0.198	9.011	2.164	0.463
K^2	0.00162	0.00181	0.141	0.0247	0.0432
$S^{3/2}K^{1/2}$	−3.773	−0.122	−3.346	2.786	−0.234
SK	0.880	0.0836	0.638	−0.454	−0.891
$S^{1/2}K^{3/2}$	0.0328	0.000231	0.110	0.0381	−0.0254
SK^2	0.000901	0.0000956	−0.124	−0.0392	−0.00616
S^2K	0.0717	0.0133	−0.642	−0.862	−0.272
$S^{3/2}K^{3/2}$	−0.0216	−0.00373	0.499	0.307	0.205
$\ln(S)\ln(K)$	−0.721	−0.0305	−0.517	0.103	0.00942
$\ln(S)K$	0.349	0.00290	−0.650	0.0341	−0.00642
$S\ln(K)$	0.0928	0.240	0.834	−0.481	−0.164
S^{-1}	0.366	−0.000296	0.136	0.0164	−0.0000209
K^{-1}	−0.555	−0.444	0.0989	−0.00817	0.00151

Notes: The values in this Table are the parameter estimates of Equation (21), which is the polynomial model for the Kurtosis parameter, K_c. This Table is retrieved from Amédée-Manesme et al. (2019).

reasons, our research implements the Gaussian kernel as well. This kernel is defined as follows:

$$\phi_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2h^2}\right\}. \tag{23}$$

The bandwidth h is an important parameter in KDE, controlling the degree of smoothing in the kernel. This parameter needs to chosen carefully, as Silverman (1981) shows that an incorrect amount of smoothing results in an unrepresentative estimation of the density function. In particular, a large value of h may excessively smooth the data, such that important features are concealed, while using a value of h that is too small results in an estimate that is highly sensitive to noise in the data. To account for this trade-off, the rule of thumb proposed by Terrell and Scott (1992) is implemented. Finally, the mode ensemble forecast is set to be the mode of the underlying distribution of our individual forecasts, which is the value that corresponds to the maximum estimated density.

A.5 Minimum and Relative Score Combining Weights

This Section provides a deeper analysis of the Minimum Score and Relative Score combining weights, which should reflect the performance of the individual methods, as shown in Table 2.

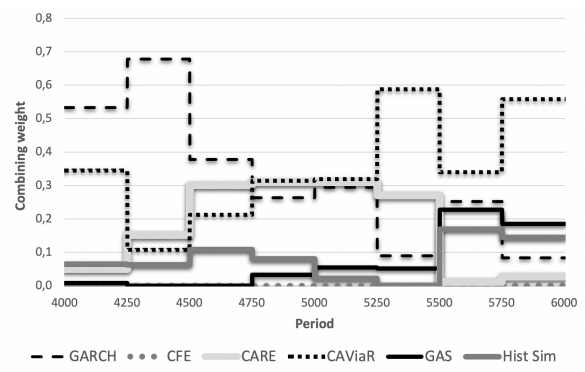
Figures A.2a and A.2b display the Minimum Score combining weights for the individual 5% VaR and ES forecasts of the portfolio, respectively[35]. Indeed, the combining weights for the VaR

[35]The Figures for the other series and the 1% probability level provide similar results, and are omitted. The

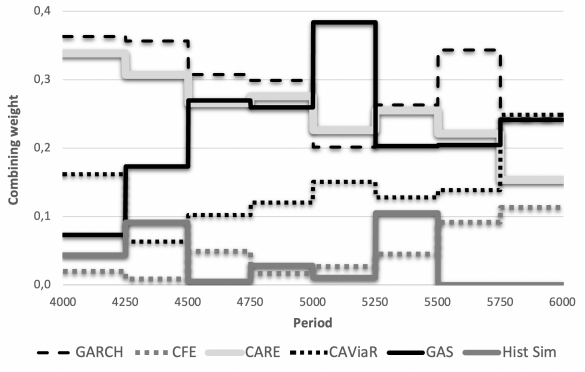
align well with the results from Table 2. In particular, the CAViaR and GARCH models receive the largest weights, as expected from their accurate VaR forecasts. Interestingly, over time, the weight of the CAViaR model increases while that of the GARCH model decreases, implying a shift in the relative importance or forecast accuracy of these models. The weights for the GAS and CARE models are substantially smaller than those of the GARCH and CAViaR, and the weights for the historical simulation and the CFE are even smaller, often indistinguishable from zero, signifying their insignificant contribution to the combined forecasts. Focusing on the

Figure A.2: For the Minimum Score Combining method of Section 3.3.2, the combining weights for the 5% forecasts of all individual methods are displayed for the portfolio, during the out-of-sample period. The VaR and ES forecasts are assigned two different sets of weights, shown in Figures A.2a and A.2b, respectively.

(a) Combining weights for the VaR



(b) Combining weights for the ES



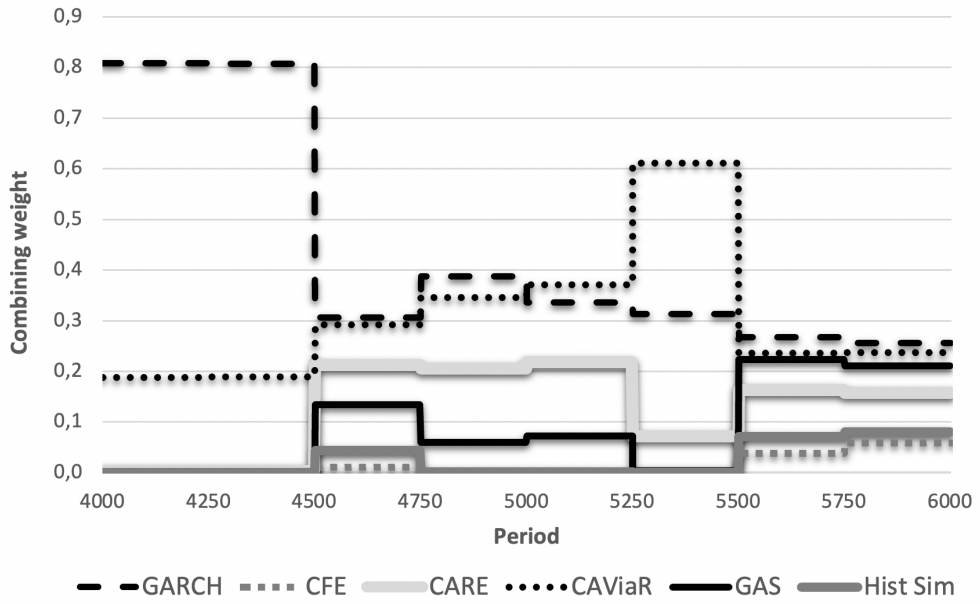
Notes: The 5% VaR combining weights of Figure A.2a correspond to the weights $q_j^{(i)}$ of Equation (13), and the 5% ES combining weights of Figure A.2b correspond to the weights $e_j^{(i)}$ of Equation (14), using the set of all individual methods (denoted $\mathcal{M}_0$ in Section 3.3) for the portfolio.

combining weights for the 5% ES forecasts in Figure A.2b, similar patterns are observed, with a notable difference for the GAS and CARE models, which receive much larger weights for their ES forecasts compared to their VaR forecasts. This implies these models are potentially more competent at accurately forecasting the Expected Shortfall over the Value-at-Risk. Moreover, the combining weights for the ES display less variation compared to those for the VaR, as evident from the smaller range in the ES weights. Thus, using equal weights (as done in the simple averaging techniques from Section 3.3.1) for the ES appears to be more appropriate than doing so for the VaR. The scattering of the combination weights for the VaR suggests that different weights for the different methods should be used, as the methods' contributions to the combined forecasts are different. This finding is partially supported by Table 2, albeit not conclusively. In particular, the quantile score for the VaR shows a slight inclination towards the sophisticated combining techniques using different weights for the different methods (as explained in Sections 3.3.2 and 3.3.3), while a preference for the simple averaging techniques is observed based on the joint scoring functions. Notwithstanding, the differences are not conclusive, and further research is required to make decisive statements about potential differences between the VaR and ES forecasts.

Figure A.3 shows the 5% Relative Score combining weights (of Section 3.3.3) for the portfolio,

analysis focuses on the portfolio, which is most appropriate for answering the central research question.

Figure A.3: For the Relative Score Combining method of Section 3.3.3, the combining weights for the 5% forecasts of all individual methods are displayed for the portfolio, during the out-of-sample period.



Notes: These 5% combining weights correspond to the weights $\omega_j^{(i)}$ of Equation (17), using the set of all individual methods (denoted $\mathcal{M}_0$ in Section 3.3) for the portfolio.

which are identical for the VaR and ES. These weights show similar patterns to those of the Minimum Score combining technique. In particular, these weights seem slightly more in line with the VaR weights of Figure A.2a, with the CAViaR and GARCH consistently being the two models with the largest contribution to the combined forecast. This motivates the relevance of the Relative Score combining method, as its estimation process is more straightforward than the Minimum Score combining[36],

[36]The Relative Score method requires the optimization of only one tuning parameter per series (λ_j of Equation (17).

A.6 Additional Evaluation Results

Table A.4: 1% VaR evaluated using the quantile skill score (in %) and its geometric mean, as defined in Section 3.5.1.

	S&P	FTSE	AGG	GSCI	EURUSD	Portfolio	Geo. mean
Individual methods							
Historical sim	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	28.4	16.7	10.9	13.5	9.7	33.3	19.3
Cornish-Fisher	−27.1	−22.4	4.1	−2.4	−4.2	−17.3	−11.0
CARE-AS	−29.9	−38.4	−55.0	−57.8	−45.6	−39.4	−44.0
CAViaR-AS-EVT	27.7	20.7	17.9	15.0	8.2	33.6	21.0
GAS model	25.1	10.9	12.1	10.5	9.1	28.8	16.5
Combining all							
Simple average	28.4	20.2	11.7	13.5	7.6	28.0	18.7
Median	28.9	15.7	10.9	11.9	8.9	28.6	17.9
Mode	26.4	12.0	8.5	9.1	8.9	25.6	15.5
Winsorizing	31.9	19.8	15.0	13.0	8.6	32.4	20.6
Trimming	29.3	18.3	12.0	11.9	7.6	28.0	18.3
Relative score	29.7	19.1	13.8	14.4	8.5	33.5	20.4
Minimum score	29.4	17.2	12.4	12.9	8.4	30.6	18.9
Combining all except historical simulation							
Simple average	29.4	20.4	12.4	15.0	8.5	30.6	19.8
Median	28.2	15.6	11.5	13.1	9.4	31.2	18.6
Mode	26.3	15.0	12.2	13.1	9.3	30.9	18.2
Winsorizing	30.0	19.8	12.6	15.9	9.6	32.6	20.5
Trimming	28.2	15.6	12.4	13.1	8.5	31.2	18.6
Relative score	29.0	19.2	15.3	16.0	8.2	33.8	20.7
Minimum score	28.5	19.7	15.1	13.8	6.4	29.8	19.3

Notes: The quantile score is defined in Equation (18), for which higher values are preferred. Bold values indicate the best method(s) for each (combination of) asset(s).

Table A.5: 1% VaR and ES evaluated using the AL skill score (in %) and its geometric mean, as defined in Section 3.5.2.

	S&P	FTSE	AGG	GSCI	EURUSD	Portfolio	Geo. mean
Individual methods							
Historical sim	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	10.6	9.8	3.7	6.6	4.6	14.2	8.2
Cornish-Fisher	−8.9	−4.3	0.2	−1.0	−0.2	−3.6	−3.0
CARE-AS	−86.3	−69.3	−40.4	−92.7	−28.2	−64.1	−72.1
CAViaR-AS-EVT	7.0	12.3	5.9	7.9	3.9	14.3	8.5
GAS model	5.2	1.8	2.9	0.6	4.1	11.7	4.3
Combining all							
Simple average	16.2	14.2	4.1	7.2	3.9	13.7	9.8
Median	14.7	10.2	3.7	5.9	4.3	13.8	8.7
Mode	10.7	6.6	2.6	4.1	4.2	12.1	6.7
Winsorizing	17.6	13.9	5.2	7.9	4.1	15.1	10.5
Trimming	16.0	12.3	4.1	5.9	3.9	13.7	9.2
Relative score	10.0	11.7	5.0	7.0	4.2	14.8	8.7
Minimum score	14.4	10.6	4.2	6.7	4.0	14.0	8.9
Combining all except historical simulation							
Simple average	15.8	14.5	4.4	7.9	4.2	15.0	10.2
Median	10.9	9.3	3.7	6.1	4.5	14.0	8.0
Mode	6.8	7.3	3.9	5.6	4.5	12.8	6.8
Winsorizing	15.5	13.2	4.2	9.5	4.5	15.0	10.2
Trimming	10.9	9.3	4.4	6.1	4.2	14.0	8.1
Relative score	9.7	12.1	5.4	8.5	4.3	14.9	9.1
Minimum score	10.0	12.4	5.2	7.0	3.6	13.4	8.6

Notes: The AL scoring function is defined in Table 1, and explained in Section 3.4. It is one possible specification of the joint VaR and ES score of Equation (19). The AL *skill* score is defined in Section 3.5.2, for which higher values are preferred. Bold values indicate the best method(s) for each (combination of) asset(s).

Table A.6: For each (combination of) asset(s), the Model Confidence Set (MCS) procedure of Hansen et al. (2011) is applied to evaluate the VaR separately based on the quantile score, and the VaR and ES jointly based on the four scoring functions in Table 1, being the AL, NZ, FZG and AS, respectively. A confidence level of 75% is used.

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
Individual methods										
Historical simulation	1	2	2	1	2	1	1	1	1	2
GJR-GARCH	6	5	5	6	6	6	4	5	6	6
Cornish-Fisher	1	1	1	1	1	1	0	0	1	0
CARE-AS	0	0	0	0	0	2	0	1	2	2
CAViaR-AS-EVT	6	6	6	6	6	5	5	5	5	6
GAS model	5	4	5	5	5	4	4	4	4	5
Combining all										
Simple average	6	6	6	6	6	6	6	6	6	6
Median	6	5	5	6	6	6	6	6	6	6
Mode	5	5	5	5	6	6	6	6	6	6
Winsorizing	6	6	6	6	6	6	6	6	6	6
Trimming	6	6	6	6	6	6	6	6	6	6
Relative score	6	6	6	6	6	6	6	6	6	6
Minimum score	6	5	5	6	6	6	6	6	6	6
Combining all except historical simulation										
Simple average	6	6	6	6	6	6	6	6	6	6
Median	6	6	6	6	6	6	5	6	6	6
Mode	5	5	5	5	6	6	5	6	6	6
Winsorizing	6	6	6	6	6	6	6	6	6	6
Trimming	5	5	5	5	5	6	6	6	6	6
Relative score	6	6	6	6	6	6	6	6	6	6
Minimum score	6	6	6	6	6	6	6	6	6	6

Notes: The quantile score is defined in Equation (18), and the four scoring functions are defined in Table 1, with brief descriptions given in 3.4. The values in this Table are the number of assets for which each method was located in its Model Confidence Set (of level 75%), such that higher values are preferred.

A.7 Comments on the Programming Code

The majority of our methods, in particular the techniques used by Taylor (2020), are implemented using the GAUSS software, as its constrained optimization libraries (`CMLMT`, `COMT`) greatly facilitate the estimation procedures. The before mentioned extensions are mostly implemented in R, benefitting from the use of the `GAS` package[37] and the `PerformanceAnalytics` package. For reproducibility purposes, the seed is set to 100 for all optimization and estimation procedures requiring simulation.

[37]This package is developed by Ardia et al. (2016), and can be found using the following link: https://CRAN.R-project.org/package=GAS.