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

Forecast combinations for Value at Risk and Expected Shortfall

Wilco Koomen (545486)

Abstract

Value at Risk (VaR) and Expected Shortfall (ES) are widely used risk measures by investment and commercial banks for both internal risk management and regulatory purposes, and even required by the Basel Committee. This research aims to examine if combining multiple forecasts can lead to improved predictive ability over their individual components. This is done by comparing six different individual forecasting methods: GJR-GARCH, HAR range, CARE-AS, CAViAR-AS-EVT, Historical Simulation and Asymmtric Laplace EWMA. These methods are combined in three ways, a simple average, Minimum Score combination and Relative Score combination. For these scores and the evaluation of the individual and combined forecasts, different scoring functions are used. The data used for this research has been sourced from Bloomberg and consists of 6000 log returns, from different start dates in 1999 till the 28th of April 2023, of five stock indices, the CAC 40, DAX 30, FTSE 100, NIKKEI 225 and S&P500. We find that combining different forecasts does indeed lead to better predictive ability. Especially the Relative Score combination excluding the Historical Simulation method consistently outperforms the individual methods. The AL-EWMA model does not outperform the individual methods, but does help in the diversification and improvement of the combinations due to the different way the information is used.

Supervisor:	dr. Bram van Os
Second assessor:	dr. Anastasija Tetereva
Date final version:	2nd July 2023

1 Introduction

Value at Risk (VaR) is a widely used tool for risk management in the financial world as it reflects the potential loss in an asset portfolio (Holton, 2002), and is also in use by several large investment banks. VaR denotes the lowest quantile (1%, 5% etc.) in the return distribution of an investment portfolio. The VaR measure is of use for both internal risk management and regulatory purposes. Regarding the regulatory aspect, the VaR is prescribed by both the second and third Basel Accord (Basel Committee on Banking Supervision, 2019), meaning that banks have to justify their investments using this metric. A limitation of this VaR method is however that it provides no insight in the potential exceedances under this quantile. This is where the Expected Shortfall (ES) comes in. Where the VaR denotes the quantile, the ES denotes the expected value of the returns under this quantile. The importance of this metric has increased in recent years with the previously mentioned Basel Committee now also recommending the use of the ES as a risk measure. Forecasting these risk measures accurately is of upmost importance for banks, as an overestimation of the risk could lead to more liquidity and disappointing returns due to conservative investing behaviour. The consequence of an underestimation of the risk probably does not need an explanation as it can lead to high exposure leading to in the worst case defaults.

One possibility to increase the accuracy of the forecasts of these risk measures is to combine different individual forecasts. The idea behind the combinations of individual forecasts is that when combined the resulting forecasting could cancel out certain deviations which are present in individual forecasts (Atiya, 2020). Empirical support on the benefit of these combinations are available across multiple different applications for example in the case of inflation in Engle, Granger and Kraft (1984).

This research aims to investigate the added benefit of combining multiple forecasts of VaR and ES. Previous research has shown that combined forecasts can outperform their individual components in predicting for example VaR, this includes: McAleer, Jimenez-Martin and Perez-Amaral (2013a), McAleer, Jiménez-Martín and Pérez-Amaral (2013b), Halbleib and Pohlmeier (2012), Fuertes and Olmo (2013) and Jeon and Taylor (2013). We perform the forecasts on five different stock indices, the S&P 500, FTSE 100, NIKKEI 225, CAC 40 and DAX 30 on the period from 1999 till 2023.

The presented analysis is based on the paper by Taylor (2020) and consists of a replication as well as the addition of a (robust) Asymmetric Laplace Exponentially Weighted Moving Average (EWMA) model, which is compared to the methods used in the paper by Taylor (2020). Furthermore, the earlier mentioned paper is extended by adding newer data to also include the recent, turbulent period with the COVID-19 pandemic and Ukraine war. The research question is therefore:

Can combined forecasts for Value at Risk and expected shortfall outperform those of individual methods, including an asymmetric Laplace EWMA?

The hypothesis is that this is in fact the case, as the combination of different individual methods has previously been shown to improve forecasting ability (Pooter, Ravazzolo & Dijk, 2010).

The asymmetric Laplace EWMA (AL-EWMA) extension is an evolution of the standard EWMA, of which a specific version is known as the RiskMetrics method, introduced back in 1994 by JP Morgan (Longerstaey & Spencer, 1996). This is a widely used method to calculate market risk by large investment banks like JP Morgan. The RiskMetrics and the AL-EWMA models are basically restricted versions of respectively a standard or assymetric GARCH model. We expect this model to have added value for a few reasons. Firstly, it uses the absolute returns instead of the squared ones, making it more robust to large shocks, and potentially improving performance in more tumultuous times like the recent COVID pandemic. Secondly, this method allows for skewness in the returns, providing an advantage over the symmetric methods like the standard GARCH or RiskMetrics models. Together, these factors can lead to diversification benefits in the combinations of the individual methods, as this method uses the available information in a different way and is thus likely to generate alternative results that may sketch potential developments that might otherwise be missed.

It is interesting to see how this AL-EWMA method compares to the five already included models. This may also prove to be a useful addition to the current literature on the topic of forecasting VaR and ES which is discussed below in Section 2.

We have found that combined forecasts do in fact outperform their individual counterparts, confirming the hypothesis set before and showing the merits of combined forecasts. The second important finding is that AL-EWMA model does not outperform the other individual forecasting methods, but its added value can be found in the diversification of the combinations.

The further layout of the paper is as follows: firstly the relevant literature is discussed in Section 2, this is followed by the data used in Section 3 and the methodology in Section 4, the results of these methods are presented in Section 5 and to conclude the conclusion and discussion can be found in Section 6.

2 Literature

Interestingly, empirical studies have found that a simple average of individual forecasts is surprisingly competitive. When the mean has to be forecasted, least squares can be used for the optimisation of convex combining weights using the individual forecasts as regressors. As an extension to this, Granger (1989) and Granger, White and Kamstra (1989) suggest using quantile regression to combine quantile forecasts such as for VaR. Research by Taylor and Bunn (1998) restricts the parameters in this quantile regression by for example imposing a zero intercept convex combining weights, similar to the combinations of forecasts of the mean. Another way of calculating weights of individual weights in the combined forecasts is proposed by Shan and Yang (2009) who use the inverse of the quantile regression loss function, giving forecasts with a higher loss a smaller weight. To determine relative performances of different forecasting methods of the VaR, a scoring function is needed. An often used function for this can be found in Equation (1), this method has been proven to be consistent by Gneiting (2011). Here the Q_t indicates the predicted quantile and y_t the observed value,

$$S(Q_t, y_t) = (\alpha - \mathbb{1}[y_t < Q_t])(y_t - Q_t). \tag{1}$$

In this equation α denotes the probability level of the forecasted VaR quantile, this scoring function then measures how often the quantile is exceeded, this should be equal to the predetermined α.

The ES as a risk measure, however, is not elicitable (Fissler, Ziegel & Gneiting, 2015), meaning that there is no scoring function that can be used to evaluate and compare forecasting performance of the ES. A workaround for this problem, proposed by Fissler et al. (2015), would be to score ES jointly with VaR as this would be elicitable. The resulting equation proposed by them can be found in Equation (2):

$$\begin{aligned} S(Q_t, ES_t, y_t) = (\mathbb{1}[y_t \leq Q_t] - \alpha)G_1(Q_t) - \mathbb{1}[y_t \leq Q_t]G_1(y_t) + G_2(ES_t)* \\ (ES_t - Q_t + \mathbb{1}[y_t \leq Q_t]\frac{Q_t - y_t}{\alpha}) - \zeta_2(ES_t) + a(y_t). \end{aligned} \tag{2}$$

Varying the functions ζ_1, G_1, G_2 and a in the Equation (2) of this general joint scoring function, different functions can be obtained. An overview of this can be found in Table 1. These scoring functions will be used in this research as well to evaluate and compare the different forecasts.

Table 1: Different joint scoring functions

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

Note. In this table ζ_1, G_1, G_2 and a denote the different possible functions and α the quantile of the VaR and ES estimates and W a parameter set in order that $WQ_t < ES_t$ for all pairs of forecasts, this is set at 4 in this research.

3 Data

The data used for this research consists of the daily log returns of five different stock indices. These are the French CAC 40, German DAX 30, English FTSE 100, Japanese NIKKEI 225 and the American S&P 500, which are stock indices consisting of the 40, 30, 100, 225 or 500 largest companies on the countries stock exchange. The range of the series differs from the ones used in the research of Taylor (2020). In that paper 6000 observations were included with starting dates ranging from 26 October 1993, 27 September 1993, 1 September 1993, 4 January 1993 and 4 August 1993 due to different holiday periods in each country. The end dates were all the same at the 31st of May 2017. In this paper 6000 observations are used as well, but because more recent data are included the end date is now the 28th of April 2023. Starting dates range between 17 November 1999, 16 September 1999, 30 July 1999, 6 November 1998 and 25 June 1999 for the CAC, DAX, FTSE, NIKKEI and S&P respectively. The first 4000 observations will be used for the in-sample analysis, to train the models, whereas the final 2000 will be used as out-of-sample to evaluate the models.

The data for these series have been sourced from the Bloomberg Database. And the closing price levels together with possible dividend have been used to calculate the log returns of the

indices. The intraday ranges used in the HAR model have been calculated using the log of the ratio between the daily highs and lows. Some descriptive statistics can be found in Table 5 in the Appendix. From these decriptive statistics it can be seen that the log returns (in percentages) have fairly similar order of magnitude with means ranging from 0.0131 to 0.0267. On average returns of the FTSE 100 are the lowest and the ones from the S&P 500 are the highest. For the standard deviations of the series the same thing can be said: for all series the standard deviations have the same order of magnitude ranging from 1.246 to 1.584. Here the S&P 500 has the lowest variance and the DAX 30 the highest. Interestingly, all series are skewed negatively, indicating fatter tails on the negative side of the distribution of returns. This is especially interesting for the models that allow for asymmetry in the modelling of the conditional variance and estimated quantiles namely the GJR-GARCH, CAViaR-AS, CARE-AS and AL-EWMA models, as this could make them provide a better fit. For illustration purposes, a plot of the log returns of one of the series, in this case that of the CAC 40 can be seen in Figure 4 in the Appendix, the plots of the other series look very similar.

As was done by Taylor (2020), in this research the data series of the daily log returns have been run through an AR(1) filter. This means that firstly an AR(1) model is fitted on the series and then the resulting residuals are used for use in the actual VaR and ES prediction models. The use of this filter allows for the elimination of potential autoregressive patterns in the returns. The application of this filter was predominantly useful for the NIKKEI 225 and S&P 500, as these were the only series where the first order autoregressive coefficient was significant on a 1% level. For the other three indices this coefficient was not even significant on a 10% level.

4 Methodology

4.1 Different individual methods

As combining forecasts is said to work best when the different individual methods do not encompass eachother, i.e. they are very different and use different information (Giacomini & Komunjer, 2005). Therefore, the different methods that will be combined are chosen to be of different types (non parametric, parametric and semi parametric). The six different models are listed below. The first five are the same as used in Taylor (2020)[1] and the sixth and last one is an extension to check how the combined forecasts compare to AL-EWMA as a more advanced version of the widely used RiskMetrics method.

4.1.1 Historical simulation

As the simplest of the individual methods, this methods uses a certain amount of previous returns in order to calculate the cumulative distribution function of the log returns. In this case this amount will be 250, as research by Taylor (2020) has proven that larger samples do not provide added forecasting ability. The empirical CDF is then used to calculate the desired quantile of the distribution of returns.

[1]Replication code supplied by J. Taylor, this is used for the five methods, combining and evalution.

4.1.2 GJR-GARCH

The Glosten-Jagannathan-Runkle GARCH model, GJR-GARCH in short, extends the regular GARCH model by allowing for asymmetry in the update of the conditional variance. This can be seen in Equation (4). This property increases accuracy over a symmetric (GARCH) model. The specification of the conditional mean is the same as in the regualar GARCH model (Equation (3)). The exact model chosen is a GJR-GARCH(1,1). This model is closely related to a TGARCH model as introduced by Zakoian (1994), the difference being that the TGARCH model uses absolute residuals instead of squared residuals. In the Equations below the specification of the mean (3) and the variance (4) are shown:

$$y_t = \mu_t + u_t, \tag{3}$$

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma u_{t-1}^2 \mathbb{1}[u_{t-1} < 0]. \tag{4}$$

In these equations μ_t denotes the mean, σ_t the volatility and u_t the shocks, which are $i.i.d. \sim t(v)$ with v degrees of freedom. The estimated parameters are ω for the constant term, *alpha* the coefficient for the square of the lagged shocks, β the persistency coefficient, and γ the coefficient for the square of thelagged negative shocks.

4.1.3 CAViaR-AS-EVT

This Conditional Autoregressive Value at Risk - Asymmetric Slope - Extreme Value Theory model uses as a base a conditional autoregressive quantile model which is estimated using quantile regression as described by Engle and Manganelli (2004). This is then extended in a way proposed by Manganelli and Engle (2004) who use a peaks-over-threshold Extreme Value Theory on any value exceeding the quantile (threshold). The resulting distribution of extreme values is then used to estimate the VaR and ES. To allow for the same asymmetry as in the previously described GJR-GARCH model, an asymmetric slope is used for the CAViAR model (Equation (5)):

$$Q_t = \beta_0 + \beta_1 \mathbb{1}(y_{t-1} > 0)|y_{t-1}| + \beta_2 \mathbb{1}(y_{t-1} \leq 0)|y_{t-1}| + \beta_3 Q_{t-1}. \tag{5}$$

In this model, β_0 denotes a constant and β_3 the persistency parameter. β_1 and β_2 represent the coefficients for the positive and negative lagged returns respectively.

4.1.4 CARE-AS

As the third model, a conditional autoregressive expectile (CARE) model is used. Where expectiles are estimated using asymmetric least squares. This approach is introduced by Newey and Powell (1987). However, the use of these expectiles in a VaR and ES framework has been implemented first by Taylor (2008). The representation of this CARE model, using an asymmetric slope, can be seen in Equation (6) below:

$$\mu_t = \beta_0 + \beta_1 \mathbb{1}(y_{t-1} > 0)|y_{t-1}| + \beta_2 \mathbb{1}(y_{t-1} \leq 0)|y_{t-1}| + \beta_3 \mu_{t-1}. \tag{6}$$

In this model, β_0 denotes a constant and β_3 the persistency parameter. β_1 and β_2 represent the coefficients for the positive and negative lagged returns respectively. The expectile of interest is now μ_t. The expectile score used to estimate the parameters β to optimize estimations of quantiles is described in the following Equation (7):

$$S(\mu_t, y_t) = |\tau - \mathbb{1}(y_t \leq \mu_t)|(y_t - \mu_t)^2. \tag{7}$$

To convert the expectiles to the quantiles, which are of interest when estimating VaR, τ has to be chosen in a way that it approximates a quantile closest to α. This is done by re-estimating the CARE model, with a τ 0.0001 lower than in the previous iteration until the fitted expectile is exceeded close enough to $\alpha\%$ of observations.

4.1.5 HAR-range

The Heterogeneous AutoRegressive (HAR) model has been tied to volatility and subsequently VaR forecasting by Corsi, Audrino and Renó (2012). This model uses the realised volatility in earlier days to estimate volatility. The use of historic realised volatility is a commonly used tool to forecast daily volatility. However due to difficulties in obtaining data on realised volatility, this is replaced with the high-low intraday range in each day. This method has been proven to work by Alizadeh, Brandt and Diebold (2002). The implementation of this intraday range into a HAR model has been proposed by Brownlees and Gallo (2009), resulting in the following model as can be seen in Equation (8, 9 and 10). The parameters β are estimated using least squares and the conditional variance is a linear function of the ${Range_t}^2$, with the coefficients based on maximum likelihood. Using this model the variance can be forecasted, and when multiplied with the VaR and ES of a Student t distribution, the estimates of the VaR and ES are obtained. The resulting model is the following:

$$Range_t = \beta_1 + \beta_2 Range_{t-1} + \beta_3 Range_{t-1}^{week} + \beta_4 Range_{t-1}^{month} + \epsilon_t, \tag{8}$$

$$Range_{t-1}^{week} = \frac{1}{5}\sum_{i=1}^{5} Range_{t-i}, \tag{9}$$

$$Range_{t-1}^{month} = \frac{1}{22}\sum_{i=1}^{22} Range_{t-i}. \tag{10}$$

Where the $Range_t$ denotes the daily range at time t and the $Range_t^{week}$ and $Range_t^{month}$ denote the weekly and monthly average range respectively. The ϵ_t are i.i.d with a zero mean.

4.1.6 Asymmetric Laplace EWMA

As an addition to the GJR-GARCH model discussed in Section 4.1.2., a similar model is used in the form of a asymmetric Laplace EWMA model. The EWMA approach is a restricted version of GARCH in the way that both coefficients in the GARCH model are required to add up to one and the constant is restricted at 0.

And as also mentioned by Pooter et al. (2010) and Atiya (2020), the added benefit of combining methods increases when models are included that use different information. So this standard

EWMA approach has to be modified in a way that it incorporates different information or uses the information in a different way. This will be done in two ways. The first is to use the absolute returns instead of the squared ones. This makes the model less susceptible to large shocks and therefore more robust (Gerlach, Lu & Huang, 2013). The second way in which the model differences from the GJR-GARCH model (which is the closest in terms of specification) is by using a different distribution. The model mentioned before uses a t-distribution and the AL-EWMA model uses a (asymmetric) Laplace distribution for the estimation of the parameters as well as the calculation of the quantiles for the VaR.

The model uses two parameters which are both estimated by means of maximum likelihood estimation and using the asymmetric Laplace distribution. The first parameter is the persistency parameter λ, which determines to what degree the volatility in the next period is correlated to the current one. The second parameter is the skewness parameter p, which represents the degree to which the distribution deviates from a symmetrical distribution. The parameter p varies between 0 and 1 and when $p = 0.5$ the model reduces to the symmetric Laplace distribution. $p > 0.5$ indicates negative skewness and vice versa. The third parameter is k, however the value for k is dependent on p: $k = \sqrt{p^2 + (1-p)^2}$. The resulting model is the AL-EWMA approach as proposed by Gerlach et al. (2013), as displayed in Equation (11) below:

$$\sigma_{t+1} = \lambda\sigma_t + (1-\lambda)\left(\frac{k}{1-p}\mathbb{1}[r_t > 0] + \frac{k}{p}\mathbb{1}[r_t < 0]\right)|r_t|. \tag{11}$$

Here, λ and k are as specified earlier and σ_t r_t denote the volatility and return at time t respectively.

The estimates for the VaR and ES follow from the estimates for the volatility mentioned before, and the estimated values for p and k and using the asymmetric Laplace distribution. The resulting estimates for the VaR are then calculated as displayed in Equation (12). This is done on both a confidence level of 1% and 5% ($\alpha = 0.01$ and 0.05). The estimation of the α VaR quantiles at time t is as follows:

$$VaR_{\alpha t} = \begin{cases} \sigma_t \frac{p}{k} log(\frac{\alpha}{1-p}); & 0 \leq \alpha < p \\ -\sigma_t \frac{(1-p)}{k} log(\frac{1-\alpha}{1-p}); & p \leq \alpha < 1. \end{cases} \tag{12}$$

In this estimation, there is a distinction between two cases, the first being when the estimated skewness parameter p is higher than the set α and the second one when this one is lower. For the latter, the returns have to be highly positively skewed in order for the p to fall under the α of at most 0.05.

The AL-EWMA and the GJR-GARCH also have a common difference with regular GARCH models; both let go of the symmetry assumption in regular GARCH and EWMA approaches. Furthermore, this estimate of the AL-EWMA model in Equation (11) is a special case of the first-order threshold GARCH (TGARCH) model as introduced by Zakoian (1994), Which has the following form as can be seen in Equation (13):

$$\sigma_{t+1} = \alpha_0 + \alpha_1^+ r_t^+ + \alpha_1^- r_t^- + \beta_1 \sigma_t. \tag{13}$$

Where α_0 denotes the constant, α_1^+ and $\alpha_1^- r_t^-$ the different coefficients for the positive and negative returns respectively, and β_1 the persistency coefficient.

4.2 Combining methods

A set of three different combination methods will be used in order to optimally combine the individual methods. The first, and the most simple method is taking a simple average. Secondly, a set of two score combining methods are used: minimum and relative score combining. As the performances of the individual methods are very likely to differ, especially the historical simulation method most likely performs worse, it might be useful to make it possible for the weights of the individual methods in the combination to differ.

The first proposed approach with flexible weights involves the combination of forecasts of the spacing between expected shortfall (ES) and value at risk (VaR), rather than directly combining ES forecasts. This method is referred to as minimum score combining. Using this method, the combined forecasts are first constructed as functions of the weights and individual forecasts as can be seen in Equation (14 and 15) where w_m^Q indicates the weights in the quantile forecast combination, w_m^S the weight for the spacing between the estimated quantile and ES. Lastly, m indicates the individual method. These weights w are then optimised to minimize the scoring function of the combined forecast. Firstly the combined VaR is constructed as a function of the weights

$$\hat{Q}_{comb,t} = \sum_{m=1}^{M} w_m^Q \hat{Q}_{m,t}, \tag{14}$$

next, the combined ES follows from this combined VaR

$$\hat{ES}_{comb,t} = \hat{Q}_{comb,t} + \sum_{m=1}^{M} w_m^S (\hat{ES}_{m,t} - \hat{Q}_{m,t}). \tag{15}$$

The second approach is called relative score combining, which is a method of combining VaR and ES forecasts using convex combining weights that are inversely proportional to the mean squared error (MSE) (Bates & Granger, 1969). In this research, the joint scoring functions from Equation (2) and Table 1 are used to measure accuracy, resulting in a single set of weights for both VaR and ES prediction. These weights are calculated using the formula displayed in Equation (16). The parameter θ denotes the degree to which the weights are dependent on the scoring function, which a value close to zero making the weights close to the simple average and a value close to one resulting in a combination close to the best performing individual method. This parameter is estimated by minimizing the values of the scoring function in sample. The resulting combined forecasts are then very easily interpretable as a weighted sum of the individual forecasts (Equation (17, 18)) as listed below:

$$w_m = \frac{exp\left(-\theta \sum_{t=1}^{T-1} S\left(\hat{Q}_{m,t}, \hat{ES}_{m,t}, y_t\right)\right)}{\sum_{j=1}^{M}\left(-\theta \sum_{t=1}^{T-1} S\left(\hat{Q}_{j,t}, \hat{ES}_{j,t}, y_t\right)\right)}, \tag{16}$$

$$\hat{Q}_{comb,t} = \sum_{m=1}^{M} w_m \hat{Q}_{m,t}, \tag{17}$$

$$\hat{ES}_{comb,t} = \sum_{m=1}^{M} w_m \hat{ES}_{m,t}. \tag{18}$$

4.3 Evaluation methods

In order to determine the relative performances of all different predictive models for VaR and ES and to check the added benefit of the combinations of these individual methods, a selection of evaluation methods are used. These tests also provide insight in whether the addition of the AL-EWMA provides extra predictive ability. Firstly, the forecasts are evaluated using backtests. Secondly, the different individual forecasts and combinations are compared by means of model confidence sets.

These tests are performed over the 2000 out-of-sample observations as described in the data Section 3.

4.3.1 Backtesting

The forecasts of both the VaR and ES will be evaluated by backtesting using both calibration tests as proposed by Nolde and Ziegel (2017) and the scoring functions as described in Equation (2) and Table 1.

The more traditional way of backtesting is using calibration tests. In this case, the forecasted VaR quantile $\hat{Q}_t$ is calibrated if the expectation of $(Hit_t = \alpha - \mathbb{1}y_t \leq \hat{Q}_t)$ is equal to zero both conditionally and unconditionally. In the latter calibration, the Hit_t is tested to be significantly different zero using a binomally distribution based test. The conditional calibration is tested using a four lag dynamic quantile test as proposed by Engle and Manganelli (2004). The approach for performing backtests on the ES predictions comes from McNeil and Frey (2000). In this research the forecasted are tested for a zero mean in the discrepancy between the forecasted ES and the observed return in the periods where the return exceeds the forecasted VaR. In order to standardize, the discrepancies are divided by the VaR estimate. As the distribution of these discrepancies is unknown circular bootstrapping as described by Jalal and Rockinger (2008) is used.

Next to these calibration tests, the forecasts for VaR and ES are tested using the scoring functions as described in Equation (2) and Table 1 as well. For the VaR this has been done by calculating the quantile score of the different methods and comparing them to the one of the historical benchmark as a benchmark. This way we can calculate a "percentual increase in predictive ability of the method of interest, MoI" over the historical simulation benchmark in the following form: $100\%(1 - (QS_{MoI} - QS_{HS}))$. The same is done for the ES forecasts, but because of the elicitability problem discussed in Section 2, not the quantile score but the AL scoring function is used.

All these tests are then aggregated over the 5 different indices.

4.3.2 Model Confidence Set

In order to obtain more insight in what models and combinations of different models work best in forecasting VaR and ES, Model confidence Sets (MCS) in introduced. This follows the method as proposed by Hansen, Lunde and Nason (2011).

The resulting confidence set of models ensures that the best performing model is included with a certain probability. A not included model is therefore unlikely to be the best model. In each step of the composition of the set, one model is eliminated using the Diebold-Mariano based equivalence test and the one-sided elimination rule as introduced by Hansen et al. (2011) describing it as: $T_{max,M}$, indicating the maximum of the test statistic T over all available models M. From this same research, we also follow the proposition to use two different confidence levels for the MCS, 75% and 90%. This procedure is repeated for all 5 different scoring functions as decribed in Section 4.2.

5 Results

The discussion of the results are divided into different sections. Firstly the different combinations will be introduced which will then later be evaluated and compared to the individual methods. This will be done using the two different types of backtesting in the methodology and a Model Confidence Set framework.

The most important findings are that the combined forecasts do in fact consistently outperform their individual components in forecasting VaR and ES. In particular the combination made using the Relative Score combination method exclusing the Historical Simulation method shows the merits of the combined forecast. In addition, we find that the AL-EWMA method does not provide good individual forecasts compared to the other methods. This method does however help in improving the combined forecasts due to the diversification benefit.

5.1 Combinations of the individual models

As mentioned before, the individual methods are combined in three different ways, a simple average, the minimum score combining method and the relative score combining method. All these combinations are performed including and excluding the Historical Simulation method.
The weights of the simple average of the five or six models speak for themselves. For the minimum score combining method, the two different sets of weights have been estimated as displayed in Figure 1 and 2, these weights are for the DAX 30, this choice is random, the other indices show similar results. From the graphs of the weights a few interesting findings can be derived. Firstly, the weights vary quite a lot over time, indicating that certain models perform better in certain situations. A big drop in the quantile weight of the CAViAR model is visible around observations 5250, which is the second quarter of 2020 which coincides with the start of the COVID pandemic. This could indicate that the CAViAR model performs worse in times of high volatility or recession. A second interesting finding is the discrepancy between the weights for the quantiles and the ones for the spacing between the VaR and the ES. Apparently the models which are useful in the prediction of VaR are not neccesarily useful in forecasting ES. The dominant models in the minimum score combination for quantiles are the HAR range,

CARE and CAViAR models. The HAR range and CARE models als play a sizable role in the prediction of the spacing, although the CARE model sees a big drop towards the end of the sample period. In the combination of the spacing weights, the AL-EWMA model also plays a sizable role, possibly indicating that the addition of this model adds something to the prediction of this. In both sets of weights, the Historical Simulation has a very minor role, confirming the hypothesis that this model has inferior forecasting performance due to its simplicity.

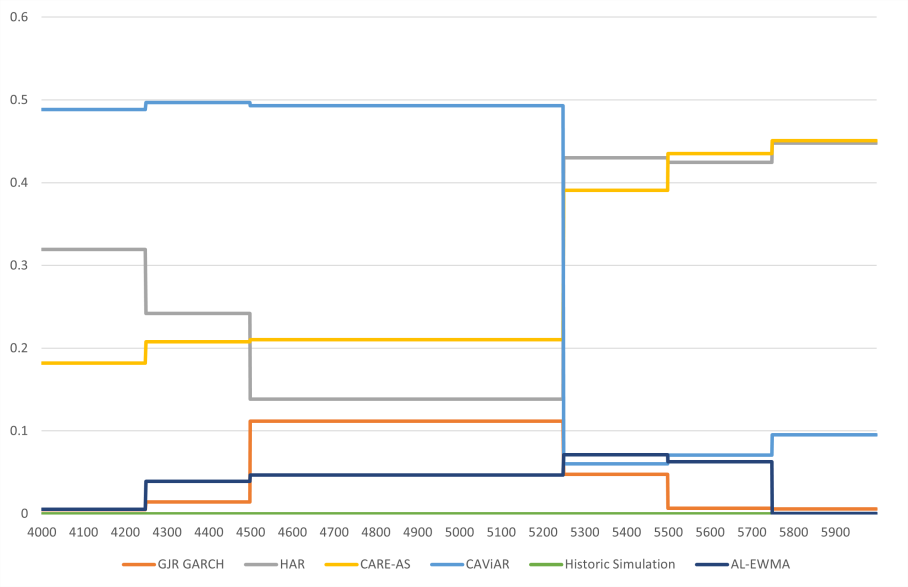

Figure 1: 1% quantile weights (Y-axis) of the Minimum Score combination of the six individual methods for the DAX 30

Note. This graph shows how the 1% quantile weights from the Minimum Score combination vary over the out of sample period (the last 2000 observations) which runs from 2015 till the end of the sample in april 2023. The weights for the German DAX 30 index are shown here, the other indices show similar patterns.

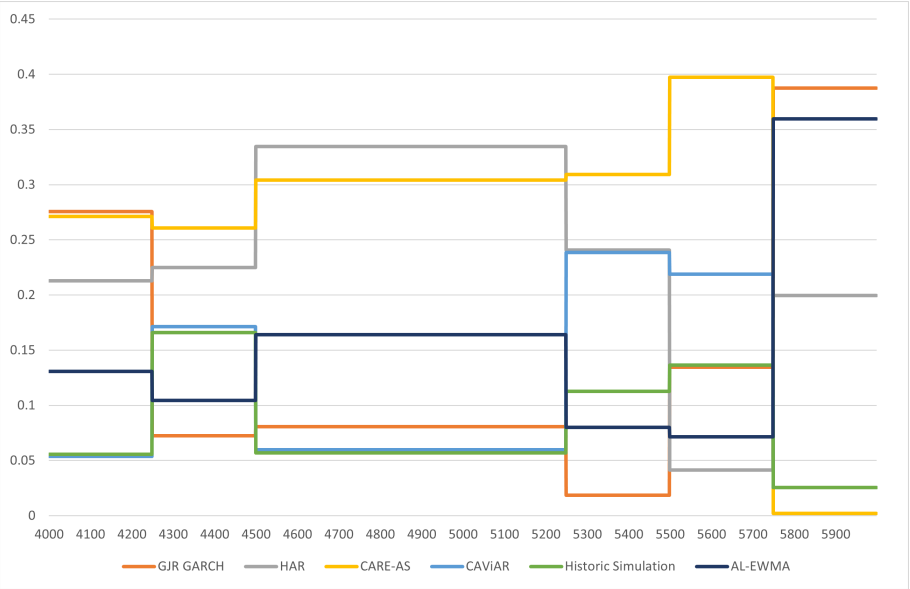

Figure 2: 1% spacing weights (Y-axis) of the Minimum Score combination of the six individual methods for the DAX 30

*Note.*This graph shows the weights for the spacing between the estimated VaR and ES (both on a 1% level), in the Minimum Score combination method, this spacing is weighted as well as the quantiles shown earlier. As was the case with the quantile weights, in this graph the weights for the German DAX 30 index are displayed.

The single set of weights for the relative score combination method are shown in Figure 3. From this graph we can make a few interesting remarks. Firstly, the GJR-GARCH model seems to be the dominant model, especially in the final part of the sample. As was the case with the earlier combinations, the share of the Historical Simulation model is minimal. And even though the share of the added AL-EWMA model is not very large, it is present, possibly indicating that adding this model that uses different information provides a useful addition.

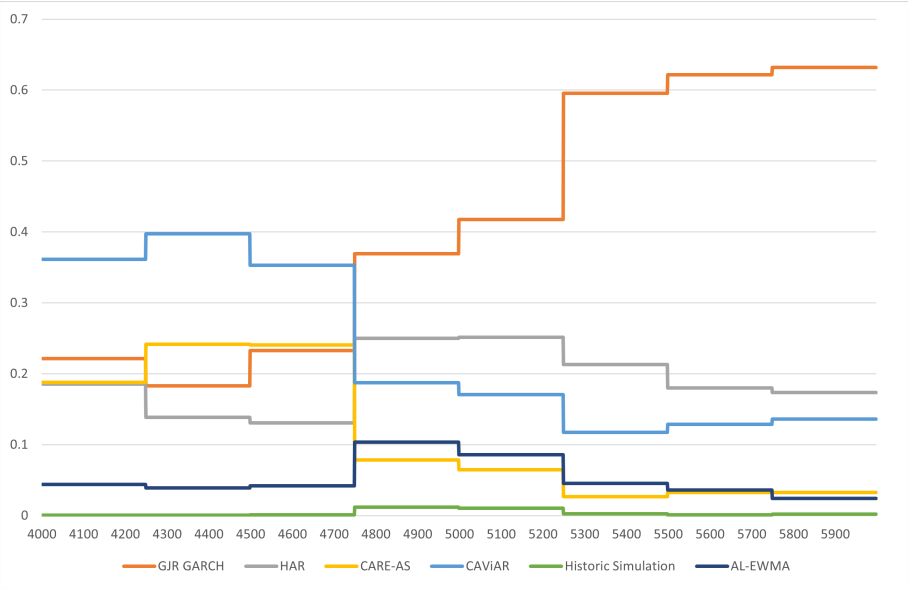

Figure 3: 5% weights (Y-axis) of the Relative Score combination of the six individual methods for the DAX 30

Note. As the Relative Score combination method only produces a single set of weights for both the VaR and ES, only one plot of these over the out of sample period is shown here. As opposed to the previously reported weights from the Minimum Score combination, here the 5% VaR and ES are used instead of the 1% VaR and ES, to provide a more diverse and complete image of the results.

5.2 Backtesting using calibration tests

The first part of the backtesting is done using three calibration tests, a VaR hit proportion test, dynamic quantile test and the ES bootstrap test as introduced by Jalal and Rockinger (2008). In Table 2 is counted for how many indices the null hypotheses of these tests is rejected at the 5% level. From this table it can be seen that the null hypotheses are rejected quite often, a lot more than in the research by Taylor (2020). This could be due to the more unpredictable and volatile data sample used in this research, which includes the COVID pandemic and the recent war in Ukraine. However, a possibly bigger factor in this inaccuracy is the estimation interval of the different models. This has been moved from a 1 day to a 250 day interval, due to computational time constraints. Interestingly, the combined methods do not seem to suffer as much from this, the null hypotheses get rejected less, indicating better performance of those. The CAViAR and AL-EWMA models perform especially poorly when using this evaluation method. The Historical Simulation and HAR range models seem to have the best performance of the individual methods.

Table 2: Results of the calibration tests aggregated over the five indices

	1% VaR and ES			5% VaR and ES		
	Hit %	Dynamic Quantile	ES Bootstrap (JR)	Hit %	Dynamic Quantile	ES Bootstrap (JR)
Individual Methods						
GJR-GARCH	4	4	2	1	5	0
HAR	1	2	2	0	1	4
CARE	4	4	0	4	4	4
CAViAR	5	5	5	5	5	4
HS	0	2	1	1	3	0
AL-EWMA	5	5	0	5	5	5
Combinations Including Historical Simulation						
Simple Average	3	4	1	2	1	5
Min. Score	4	5	2	0	1	4
Rel. Score	0	3	0	0	1	2
Combinations Excluding Historical Simulation						
Simple Average	3	3	1	2	1	5
Min. Score	3	3	0	1	1	5
Rel. Score	0	2	1	0	0	3

Note. This table shows the number of times the null hypotheses are rejected over the five indices for the three different calibration tests. A lower number indicates better performance of that model, as 0 indicates that the null hypothesis has not been rejected for any index, and 5 indicates that it has been rejected for all indices. These backtests are performed over the out of sample period as described in Section 3.

5.3 Backtesting using skill scores

The second part of the backtesting is done using quantile scoring function and the four different scoring functions as discussed in Section 2 and 4.3.1. The results of the ratios with the Historical Simulation method as benchmark have been reported in Table 3. These results are based on the AL scoring function, but the results for the other scoring functions show similar patterns. The interpretation for these is fairly straightforward, a positive value means the model outperforms the Historical Simulation. The value denotes a percentual increase in the scoring function compared to this benchmark. Subsequently a negative value indicates a percentual decrease in the scoring function, and worse performance. This type of backtesting yields similar results as the calibration tests. Of the individual methods, only the HAR range and CARE models outperform the benchmark, the CARE model only for the DAX, FTSE and NIKKEI indices. The model that stands out is the CAViAR model, for the 1% forecasts evaluated in Table 3 this model is dramatically outperformed by all models, including the Historical Simulation benchmark. This is due to the nature of the model, which only uses the fraction of the sample that falls in the tail of which the VaR estimates are made. For the 1% quantile as displayed in Table 3, this leaves only a small sample for the estimation. Changing the level of the VaR and

ES from 1% to 5% therefore has a big effect on the relative performance of this model, increasing the mean scoring ratio from -55.784 to -1.793.

The evaluation using the scoring functions yields the same conclusions regarding the combination of different forecasts, once again emphasizing the merit of combining different models. All combinations consistently outperform the benchmark and the other individual methods. The combination that provides the best forecasting performance seems to be the Relative Score combination excluding Historical Simulation, the scoring function values of this combination are the highest for nearly individual index and is the highest on average. This is the method that is constructed using weights that are inversely proportional to the scoring values of the individual methods. The exclusion of the Historical Simulation might lead to better overall performance because this model possibly does not incorporate the information different enough (which leads to the benefits of combining). This is also supported by the weights assigned to the Historical Simulation method in the combination of six models.

Table 3: Evaluation of 1% VaR and ES forecasts using the AL skill score

	CAC 40	DAX 30	FTSE 100	NIKKEI 225	S&P 500	Mean
Individual Methods						
GJR-GARCH	-7.45157	-4.30573	-6.59856	-8.33881	-7.07989	-6.75491
HAR	1.691999	5.599441	3.560607	0.125764	-0.26688	2.142187
CARE	-1.55214	2.411742	1.861259	1.99107	-1.52942	0.636502
CAViAR	-63.0872	-63.2812	-54.3823	-55.4355	-57.7356	-58.7844
HS	0	0	0	0	0	0
AL-EWMA	-6.36174	-3.18085	-3.96575	-7.60987	-6.47664	-5.51897
Combinations Including Historical Simulation						
Simple Average	1.858337	5.095319	2.872401	0.869095	0.209627	2.180956
Min. Score	1.895872	5.338747	3.289635	1.342824	-0.75163	2.223089
Rel. Score	3.007695	5.104094	4.30734	1.672401	0.198764	2.858059
Combinations Excluding Historical Simulation						
Simple Average	2.120282	5.619295	3.813234	1.311359	0.511155	2.675065
Min. Score	2.120282	5.619295	3.813234	0.912787	-0.6879	2.355541
Rel. Score	2.654037	5.253743	4.335773	1.477987	0.616607	2.867629

Note. In this table the ratio of the AL skill scores as discussed in Section 4.3.1 are reported for the 1% VaR and ES. A higher value means a higher percentual outperformance of the Historical Simulation method, so better forecasting ability. The other skill score ratios show fairly comparable results. These backtests are performed over the out of sample period as described in Section 3.

5.4 Model Confidence set

The final part of the evaluation of the methods and combinations introduced consists of a Model Confidence Set (MCS) framework. Table 4 displays the number of times a certain model is included in the model confidence set. This is done based on the five different scoring functions

and for both a 75% and 90% confidence level. This table shows the results for the forecasts of 5% VaR and ES. If a model is included in a Model Confidence Set, it means that with a confidence level of 75% or 90% respectively, the best model is included. If then a model is the only one left in the set, this indicates that this is the best performing. This is true for the Relative score combination excluding Historical Simulation in certain cases. This model is included in every single MCS, as seen by the 5 at the complete bottom of the table. This model was also indicated to be the best performing by the earlier evaluation methods. Overall the combined models outperform the individual ones, except the simple average combination of 6 models. Out of the individual models, only the HAR range and CARE models are included relatively often. The other individual models are not and especially GJR-GARCH and AL-EWMA models are never included, confirming that these are definitely not the best performing ones. The same holds to a lesser extent for the CAViAR and Historical Simulation models. The benefit of the addition of the AL-EWMA model should therefore not be sought in the improvement of the already included individual methods but in improving the combination as it incorporates the information in a different way than the other models, which adds to the diversification of the combination. Finally it can be remarked that in the MCS with 90% confidence level, naturally more models are included.

Table 4: Models in the different Model Confidence Sets for the 5% VaR and ES

	75% Confidence level					90% Confidence level				
	QU	AL	NZ	FZG	AS	QU	AL	NZ	FZG	AS
Individual Methods										
GJR-GARCH	0	0	0	0	0	0	0	0	0	0
HAR	2	1	2	2	3	3	3	3	3	3
CARE	2	1	2	2	2	3	3	3	3	3
CAViAR	0	0	0	0	1	0	0	0	0	3
HS	0	0	1	0	0	1	2	2	2	1
AL-EWMA	0	0	0	0	0	0	0	0	0	0
Combinations Including Historical Simulation										
Simple Average	1	0	1	1	2	1	2	2	2	3
Min. Score	4	4	4	4	4	5	4	5	5	4
Rel. Score	4	4	4	4	4	5	4	5	5	4
Combinations Excluding Historical Simulation										
Simple Average	2	0	2	2	2	4	3	4	4	3
Min. Score	4	3	4	4	4	5	4	5	5	4
Rel. Score	5	5	5	5	5	5	5	5	5	5

Note. This table shows the number of times a certain model is included in the both the 75% and 90% MCS for the five indices. This is reported for MCS based on the Quantile score and the four different joint scoring functions. A higher value indicates better performance as a 0 would for example indicate that with either a 75% or 90% confidence level the model of interest is not the best performing model for any of the indices. It can be noted that only the combinations have values close to 5 (best possible). This evaluation has been performed over

the out of sample forecasts of the 5% VaR and ES.

6 Conclusion

Looking at the results of this research, a few conclusions can be drawn in order to formulate an answer to the research question:

Can combined forecasts for Value at Risk and expected shortfall outperform those of individual methods, including an asymmetric Laplace EWMA?

The first conclusion already goes a long way in providing an answer: Combined forecasts for VaR and ES do in fact outperform those of individual methods in this research. This follows partly from the calibration tests performed, where the null hypotheses of correct specification of the forecasts is rejected less for the combined forecast. However, this conclusion is the most strongly supported by the skill score backtesting and Model Confidence Set framework. The average skill score ratio is 2.87 for the best combined method versus 2.14 for the (on average) best performing individual method, the CARE model. And for each of the six different combined methods it holds that they perform better than this individual method. Besides the increase in average performance, the combination also seems to be a lot more consistent, with a less fluctuation in the skill score values across the different indices. The MCS framework further supports this finding, with the combinations being included in the confidence sets a lot more than the individual methods. Out of the combined forecasts, the Relative Score combination method provides the best performing weighted combination of the individual methods. This is the method where the weights are proportional to the scoring function values of the individual methods. Furthermore, excluding the Historical Simulation method improves the performance of the combined forecasts with the Relative Score combination without the Historical Simulation providing the best forecasting ability. Confirming the merits of combinations shown in earlier research.

Regarding the addition of the asymmetric Laplace EWMA model, a few conclusions can be drawn as well. In terms of individual forecasting performance, the AL-EWMA model does not outperform the five already included models. This becomes clear from both the backtesting and the Model Confidence Sets. The model does not perform well in the calibration tests and the skill scores are lower than for example the HAR range and CARE models. Besides, the AL-EWMA is not included in the confidence sets. However, judging by the decently sized weights of this model in the combinations, the addition of this model does help in the overall performance and diversification of the combinations. This would be due to the fact that this model incorporates the available information in a different way by using absolute returns instead of squared ones. This would also be in line with earlier studies on the combinations of forecasts, for example in Atiya (2020), it was found that combinations of different types of forecasts boost the combination benefits. This research confirms this to be the case with the AL-EWMA model in the light of VaR and ES forecasts combined with the methods from Taylor (2020) as well.

Another interesting remark that can be made concerns the CAViAR model. This model is not suitable to predict 1% VaR and ES using this dataset. For the prediction of 5% VaR and

ES it does work, but due to the limited sample used for the lower quantile, the performance is severely limited relative to the other models.

A final conclusion from this research specifically is that the shift from a 1 day to a 250 day estimation interval for the individual methods has had negative impact on the accuracy of these models. This is the most prominent in the calibration tests. Especially the more advanced models suffer from this, the more simple Historical Simulation method does not seem to be affected as much. This could also be the reason why the Historical Simulation method performs relatively fine, which contradicts earlier research.

This bring us to the shortcomings of this research. Apart from shortening the estimation interval, which would increase computation time. There are more areas that could be improved in further research. One of which would be extending the asymmetric Laplace EWMA part with time varying parameters λ and p. Incorporating this could lead to better forecasting performance as it would increase the flexibility and adaptability of the model. Next to time varying parameters, another possibility would be to derestrict the parameters, this would result in a TGARCH model. And even though a less restricted model would possibly perform better individually than the AL-EWMA model, this model would come very close to the GJR-GARCH model, which could possibly impede the diversification benefits. Another possible improvement for future research would be the use of more observations of the series used, this could especially be an improvement for the CAViAR model. This method only uses a small subsample and did seem to struggle with forecasting the 1% VaR and ES in particular, using this data set. More data could increase the small subsample this method uses, improving the performance. Another extension of this research could be to compare the models and results of this research to those of Generalized Autoregressive Score (GAS) models as introduced by Creal, Koopman and Lucas (2013). In recent research, these GAS models have been proven to provide good forecasting ability in for example Liu, Semeyutin, Lau and Gozgor (2020). This GAS models are a very general form of Autoregressive models, which under specific restrictions turns into for example a GARCH model. However, due to the more general nature, these model are very free in the specification. This fairly different way of use of the available information could make it interesting in the combination of different methods due to the diversification benefits discussed earlier.

References

Alizadeh, S., Brandt, M. W. & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, *57*(3), 1047-1091. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6261.00454` doi: https://doi.org/10.1111/1540-6261.00454

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*(1), 197-200. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0169207019300779` (M4 Competition) doi: https://doi.org/10.1016/j.ijforecast.2019.03.010

Basel Committee on Banking Supervision. (2019, Jan). Minimum capital requirements for market risk. *The Bank for International Settlements*. Retrieved from https://www.bis.org/bcbs/publ/d457.htm

Bates, J. M. & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468.

Brownlees, C. T. & Gallo, G. M. (2009, 07). Comparison of Volatility Measures: a Risk Management Perspective. *Journal of Financial Econometrics*, *8*(1), 29-56. Retrieved from https://doi.org/10.1093/jjfinec/nbp009 doi: 10.1093/jjfinec/nbp009

Corsi, F., Audrino, F. & Renó, R. (2012). Har modeling for realized volatility forecasting.

Creal, D., Koopman, S. J. & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777-795. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1279 doi: https://doi.org/10.1002/jae.1279

Engle, R. F., Granger, C. W. & Kraft, D. (1984). Combining competing forecasts of inflation using a bivariate arch model. *Journal of economic dynamics and control*, *8*(2), 151–165.

Engle, R. F. & Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles [Article]. *Journal of Business and Economic Statistics*, *22*(4), 367 – 381. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-4444289240&doi=10.1198%2f073500104000000370&partnerID=40&md5=e0a534bb9072ee7500a88321de056de1 (Cited by: 1072) doi: 10.1198/073500104000000370

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

Fuertes, A.-M. & Olmo, J. (2013). Optimally harnessing inter-day and intra-day information for daily value-at-risk prediction. *International Journal of Forecasting*, *29*(1), 28-42. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207012000805 doi: https://doi.org/10.1016/j.ijforecast.2012.05.005

Gerlach, R., Lu, Z. & Huang, H. (2013). Exponentially smoothing the skewed laplace distribution for value-at-risk forecasting. *Journal of Forecasting*, *32*(6), 534–550.

Giacomini, R. & Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics*, *23*(4), 416-431. Retrieved from https://doi.org/10.1198/073500105000000018 doi: 10.1198/073500105000000018

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746-762. Retrieved from https://doi.org/10.1198/jasa.2011.r10138 doi: 10.1198/jasa.2011.r10138

Granger, C. W. J. (1989). Invited review combining forecasts—twenty years later. *Journal of Forecasting*, *8*(3), 167-173. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980080303 doi: https://doi.org/10.1002/for.3980080303

Granger, C. W. J., White, H. & Kamstra, M. (1989). Interval forecasting: An analysis based upon arch-quantile estimators. *Journal of Econometrics*, *40*(1), 87-96. Retrieved from https://www.sciencedirect.com/science/article/pii/0304407689900316 doi: https://doi.org/10.1016/0304-4076(89)90031-6

Halbleib, R. & Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, *36*(8), 1212-1228. Retrieved from https://www.sciencedirect.com/science/article/pii/S0165188912000887 (Quantifying and Understanding Dysfunctions in Financial Markets) doi: https://doi.org/10.1016/j.jedc.2011.10.005

Hansen, P. R., Lunde, A. & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. Retrieved 2023-05-30, from http://www.jstor.org/stable/41057463

Holton, G. A. (2002). *History of value-at-risk.* Citeseer.

Jalal, A. & 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. Retrieved from https://www.sciencedirect.com/science/article/pii/S0927539808000200 doi: https://doi.org/10.1016/j.jempfin.2008.02.004

Jeon, J. & Taylor, J. W. (2013). Using caviar models with implied volatility for value-at-risk estimation. *Journal of Forecasting*, *32*(1), 62-74. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/for.1251 doi: https://doi.org/10.1002/for.1251

Liu, W., Semeyutin, A., Lau, C. K. M. & Gozgor, G. (2020). Forecasting value-at-risk of cryptocurrencies with riskmetrics type models. *Research in International Business and Finance*, *54*, 101259. Retrieved from https://www.sciencedirect.com/science/article/pii/S0275531920301240 doi: https://doi.org/10.1016/j.ribaf.2020.101259

Longerstaey, J. & Spencer, M. (1996). Riskmetricstm—technical document. *Morgan Guaranty Trust Company of New York: New York*, *51*, 54.

Manganelli, S. & Engle, R. (2004). A comparison of value-at-risk models in finance [Article]. *Risk Measures for the 21st Century*, 123 – 144. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-7444222679&partnerID=40&md5=767f45ca295290516e6195d6ce9df142 (Cited by: 46)

McAleer, M., Jimenez-Martin, J.-A. & Perez-Amaral, T. (2013a). Has the basel accord improved risk management during the global financial crisis? *The North American Journal of Economics and Finance*, *26*, 250-265. Retrieved from https://www.sciencedirect.com/science/article/pii/S1062940813000193 doi: https://doi.org/10.1016/j.najef.2013.02.004

McAleer, M., Jiménez-Martín, J.- & Pérez-Amaral, T. (2013b). International evidence on gfc-robust forecasts for risk management under the basel accord. *Journal of Forecasting*, *32*(3), 267-288. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/for.1269 doi: https://doi.org/10.1002/for.1269

McNeil, A. J. & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, *7*(3), 271-300. Retrieved from https://www.sciencedirect.com/science/article/pii/S0927539800000128 (Special issue on Risk Management) doi: https://doi.org/10.1016/S0927-5398(00)00012-8

Newey, W. K. & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*(4), 819–847. Retrieved 2023-05-08, from http://www.jstor.org/stable/1911031

Nolde, N. & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, *11*(4), 1833 – 1874. Retrieved from https://doi.org/10.1214/17-AOAS1041 doi: 10.1214/17-AOAS1041

Pooter, M. D., Ravazzolo, F. & Dijk, D. J. V. (2010). Term structure forecasting using macro factors and forecast combination. *FRB International Finance Discussion Paper*(993).

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

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles [Article]. *Journal of Financial Econometrics*, *6*(2), 231 – 252. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-41049105103&doi=10.1093%2fjjfinec%2fnbn001&partnerID=40&md5=83247f273f7caffaa3138ba00c0ddb9c (Cited by: 151) doi: 10.1093/jjfinec/nbn001

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, *36*(2), 428-441. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207019301918 doi: https://doi.org/10.1016/j.ijforecast.2019.05.014

Taylor, J. W. & Bunn, D. W. (1998). Combining forecast quantiles using quantile regression: Investigating the derived weights, estimator bias and imposing constraints. *Journal of Applied Statistics*, *25*(2), 193-206. Retrieved from https://doi.org/10.1080/02664769823188 doi: 10.1080/02664769823188

Zakoian, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, *18*(5), 931-955. Retrieved from https://www.sciencedirect.com/science/article/pii/0165188994900396 doi: https://doi.org/10.1016/0165-1889(94)90039-6

A Appendix

Table 5: Descriptive statistics on the indices

	CAC40	DAX30	FTSE100	NIKKEI225	SP500
Mean	0.0188	0.0191	0.0131	0.0155	0.0267
Median	0.0663	0.0623	0.0664	0.0335	0.0656
Maximum	1.253	1.257	1.306	1.073	1.096
Minimum	−1.398	−1.394	−1.324	−1.127	−1.276
Std. Dev.	1.559	1.584	1.353	1.472	1.246
Skewness	−0.160642	−0.164103	−0.336711	−0.227275	−0.368247
Kurtosis	9.955209	8.705108	12.92317	6.930868	13.04769
Jarque-Bera	12119.54	8163.994	24730.69	3914.585	25374.62
Probability	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	112.504	114.309	78.691	93.172	160.042
Sum Sq. Dev.	14576.84	15057.19	10979.95	12989.70	9308.24
Observations	6000	6000	6000	6000	6000

Note. In this table, the descriptive statistics are given for the log returns (in percentages) of the five different indices. These statistics include the mean, maximum, minimum and standard deviation but also skewness and kurtosis.

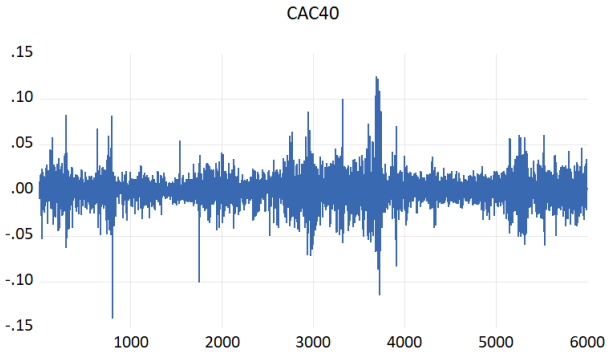

Figure 4: Plot of the log returns of the CAC 40

Note. In this figure, the log returns of the CAC 40 are plotted. The choice for this index is from alphabetical order, the other indices show similar patterns.

A.1 Code

The full code and datasets can be found in the zip file attached. The code consists of two parts, an R program and a GAUSS code. The R program implements the AL-EWMA model to produce VaR and ES forecasts. The produced forecasts are stored in txt files and loaded into the GAUSS program. This program first estimates the other five models, and produces forecasts.

These are then weighted in the three different ways. Lastly the forecasts are evaluated and the relavant metrics reported. This includes the weights, Model Confidence Sets and scoring function values. For the execution of the GAUSS program four different (paid) packages are needed. These are the following: co, cmlmt, pgraph, lpmt. The data used in this research is also provided, $Data\_van\_Bloomberg\_werkversie\_AR1filtered.xlsx$ is used for the AL-EWMA model in R, and the $Data\_van\_Bloomberg\_werkversie.txt$ file is used for the combined program in GAUSS.