

Forecast Combinations for Value at Risk and Expected Shortfall

James W. Taylor
Saïd Business School
University of Oxford

International Journal of Forecasting, 2020, 36(2), 428-441.

Address for Correspondence:

James W. Taylor
Saïd Business School
University of Oxford
Park End Street
Oxford OX1 1HP, UK
Tel: +44 (0)1865 288927
Email: james.taylor@sbs.ox.ac.uk

Forecast Combinations for Value at Risk and Expected Shortfall

Abstract

Combining provides a pragmatic way of synthesising the information provided by individual forecasting methods. In the context of forecasting the mean, numerous studies have shown that combining often leads to improvements in accuracy. Despite the importance of the value at risk (VaR), though, few papers have considered quantile forecast combinations. One risk measure that is receiving an increasing amount of attention is the expected shortfall (ES), which is the expectation of the exceedances beyond the VaR. There have been no previous studies on combining ES predictions, presumably due to there being no suitable loss function for ES. However, it has been shown recently that a set of scoring functions exist for the joint estimation or backtesting of VaR and ES forecasts. We use such scoring functions to estimate combining weights for VaR and ES prediction. The results from five stock indices show that combining outperforms the individual methods for the 1% and 5% probability levels.

Keywords: Value at risk; expected shortfall; combining; elicibility; scoring functions.

1. Introduction

The value at risk (VaR) has been used widely as a measure of financial market risk for both regulatory purposes and internal risk management. VaR is a conditional quantile in the lower tail of the distribution of the return on a portfolio. While straightforward to interpret, VaR has the limitation that it provides no information regarding potential exceedances beyond the quantile. Recently, the expected shortfall (ES) has been receiving increasing attention as an alternative risk measure, and it is now recommended as a risk measure by the Basel Committee on Banking Supervision (Basel Committee, 2016). ES is the conditional expectation of exceedances beyond the VaR. Artzner, Delbaen, Eber, and Heath (1999) point out that, in contrast to the VaR, ES has the appealing property of subadditivity, which means that the measure for a portfolio cannot be greater than the sum of the measures for the constituent parts of the portfolio. One apparent disadvantage of ES is that it is not *elicitable*, which means that the correct ES forecast is not the unique minimiser of the expectation of any loss function. This presents a challenge for estimating and backtesting ES. Fissler and Ziegel (2016) address this by providing a set of joint loss functions for VaR and ES for which these two measures are jointly elicitable. The present paper uses these loss functions in the context of forecast combinations.

The essential motivation for combining forecasts is that, when competing forecasts are available, a combination can enable a pragmatic synthesis of the information that is inherent in the individual predictions. Another perspective is that the combination provides a potentially diversified portfolio of the different forecasts. Since the seminal work of Bates and Granger (1969), a large body of literature has developed on combining forecasts of the conditional mean, with empirical support being available across a variety of applications. An interesting empirical finding is that typically a simple average is very competitive. For forecasting the mean, least squares provides a natural approach to optimizing convex combining weights, or perhaps unconstrained weights in a model where individual forecasts

can be viewed as regressors. Building on this, Granger (1989) and Granger, White, and Kamstra (1989) suggest that quantile forecasts could be combined using quantile regression. Taylor and Bunn (1998) consider the appeal of constraining the quantile regression parameters. They consider a zero intercept term and convex combining weights, as has been common for combinations of forecasts of the mean. Giacomini and Komunjer (2005) use the quantile regression framework to enable tests of quantile forecast encompassing, which provides a theoretical justification for combining in cases where one forecast does not encompass another. Shan and Yang (2009) calculate weights based on the inverse of the quantile regression loss function.

In the VaR context, it is perhaps surprising that there has not been more consideration of forecast combinations, given the variety of different quantile forecasting methods available. In their recent review of the VaR literature, Nieto and Ruiz (2016) report just a handful of studies on combining. McAleer, Jiménez-Martín, and Pérez-Amaral (2013a,b) look at selecting the maximum, minimum or median of a set of forecasts. Halbleib and Pohlmeier (2012) derive combining weights by maximizing the conditional coverage, as well as by quantile regression. Jeon and Taylor (2013) and Fuertes and Olmo (2013) also use quantile regression. They combine individual forecasts constructed from different information sources, including historical daily returns, the option-implied volatility, the realized volatility and the intraday range.

Although there has been an increased interest in forecasting ES in recent years, we are not aware of any studies that have looked at combining ES forecasts, presumably due to ES not being elicitable. This paper proposes the use of Fissler and Ziegel's (2016) joint VaR and ES loss functions for estimating combining weights for VaR and ES prediction. Elliott and Timmermann (2004) show that the question of whether combining weights should be equal depends on the loss function, implying that empirical evidence from the literature on

forecasting the mean, where a squared loss function is appropriate, may not transfer to VaR and ES forecasting. Our paper presents some empirical results in this regard.

If VaR and ES predictions are obtained from density forecasts, an alternative to combining VaR and ES predictions would be to combine the density forecasts (see for example Hall & Mitchell, 2007; Jore, Mitchell, & Vahey, 2010). Opschoor, Van Dijk, & van der Wel (2017) describe how the combining method can be adapted to focus on a particular part of the density, such as the left tail when VaR and ES are of interest. However, this combining approach is of no use when combining VaR and ES forecasting methods that are not based on density forecasts. This is our interest in the present paper. For example, VaR and ES forecasts could be produced by autoregressive quantile or expectile models, as indeed is the case in our empirical analysis. We feel that it is important to consider this more general case because forecast combinations are particularly useful when the forecasts are produced by methods that are based on different information or use the information in notably different ways.

Section 2 briefly reviews loss functions for VaR and ES prediction. Section 3 describes the two combining formulations that we propose. Section 4 presents an empirical study based on daily stock indices. Section 5 provides a simulation study. Finally, Section 6 summarises and concludes the paper.

2. Scoring functions for VaR and ES

Scoring function is the term used in decision theory to describe a loss function that is used to evaluate a forecast of some measure of a probability distribution, such as a quantile. As we explained in the previous section, a measure is described as *elicitable* if the correct forecast of the measure is the unique minimiser of the expectation of at least one scoring function. Such scoring functions are called *strictly consistent* for the measure (Fissler & Ziegel, 2016). A strictly consistent scoring function can be used as the loss function in model

estimation (Gneiting & Raftery, 2007). We now describe strictly consistent scoring functions that we propose to use for estimating forecast combining weights for VaR and ES.

2.1. Scoring functions for VaR

VaR is an elicitable risk measure. Consistent scoring functions for VaR are of the following form (Gneiting & Raftery, 2007):

$$S(Q_t, y_t) = (\alpha - I(y_t \leq Q_t))(G(y_t) - G(Q_t)),$$

where y_t is the variable of interest; Q_t is the quantile with probability level α ; I is the indicator function; and G is a weakly increasing function. If G is strictly increasing, the scoring function is strictly consistent (Gneiting, 2011). Selecting G to be the identity function leads to the *quantile score* of the following expression:

$$S(Q_t, y_t) = (\alpha - I(y_t \leq Q_t))(y_t - Q_t). \quad (1)$$

This score is used widely in the VaR literature due to both its simplicity and its familiarity as the quantile regression loss function. Averaging the score across a sample gives a measure for evaluating quantile forecasts.

2.2. Joint scoring functions for VaR and ES

The ES is not an elicitable risk measure (Gneiting, 2011), meaning that no suitable scoring function exists for the sole purpose of estimating or evaluating ES forecasts. However, a measure that is not elicitable individually may be elicitable jointly with another measure. This is the case for the variance, which is only elicitable jointly (with the mean). With regard to the ES, Fissler and Ziegel (2016) prove that it is elicitable jointly with the VaR. They show that consistent scoring functions, for evaluating VaR and ES forecasts jointly, are of the following form:

$$S(Q_t, ES_t, y_t) = (I(y_t \leq Q_t) - \alpha)G_1(Q_t) - I(y_t \leq Q_t)G_1(y_t) + G_2(ES_t)(ES_t - Q_t + I(y_t \leq Q_t)(Q_t - y_t)/\alpha) - \zeta_2(ES_t) + a(y_t), \quad (2)$$

where ES_t is the ES; and G_1 , G_2 , ζ_2 and a are functions that satisfy a number of conditions, including the properties that $G_2 = \zeta_2'$, G_1 is increasing, and ζ_2 is increasing and convex. The scoring function is strictly consistent if ζ_2 is strictly increasing and strictly concave. (The domain of ζ_2 contains only negative values, because we are considering $\alpha < 50\%$, which implies that ES_t is negative.) Fissler and Ziegel (2016) note that the scoring function in Eq. (2) remains strictly consistent for the case of $G_1 = 0$. In Eq. (2), the terms involving G_1 collectively form a consistent scoring function for a quantile, with the other terms assessing both the quantile and ES (Fissler, Ziegel, & Gneiting, 2016). Therefore, one can reduce the emphasis on the quantile accuracy by setting $G_1 = 0$, as indeed has been the choice in several studies. Table 1 presents four scoring functions, of the form of Eq. (2), that have been proposed. We discuss them in the remainder of this section.

Table 1

Functions used within the joint VaR and ES scoring function of Eq. (2) to give four different versions of the score: the AL, NZ, FZG and AS scores.

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

Taylor (2019) points out that, if $G_1 = 0$, $G_2 = -1/x$, $\zeta_2(x) = -\ln(-x)$ and $a = 1 - \ln(1 - \alpha)$, the scoring function is equal to the negative of the log-likelihood function of an asymmetric Laplace (AL) density with time-varying location and scale parameters. The use of this scoring function for model estimation has some appeal because it can be viewed as a relatively minor extension of quantile regression, which is equivalent to maximizing an AL likelihood with a time-varying location and constant scale. We refer to Taylor's (2019)

proposed score as the *AL score*. Taylor (2019) uses the score to estimate dynamic joint models for VaR and ES, and this proposal is given theoretical support by the recent work of Patton, Ziegel, and Chen (2019).

Nolde and Ziegel (2017) consider comparative backtests for risk measures. Their numerical study essentially uses the AL score, as well as the score that results from setting $G_1 = 0$, $G_2 = \frac{1}{2}(-x)^{-\frac{1}{2}}$, $\zeta_2(x) = -(-x)^{\frac{1}{2}}$ and $a = 0$ in Eq. (2), which we refer to as the *NZ score*.

In their empirical analysis, Fissler et al. (2016) use the scoring function produced by using the following functions in Eq. (2): $G_1(x) = x$, $G_2 = \exp(x)/(1 + \exp(x))$, $\zeta_2(x) = \ln(1 + \exp(x))$ and $a = 0$. In our empirical work, we found that the first three significant figures of the values of this score did not differ between forecasting methods. This meant that it was difficult to distinguish between the methods when comparing relative measures, which we computed in order to average the performances across a set of stock indices, as we describe in detail in Section 4.3.2. To make the relative measures easier to compare, we set $a = \ln(2)$ in Eq. (2). We refer to this as the *FZG score*.

Another example of a joint scoring function is proposed by Acerbi and Székely (2014), and we refer to it as the *AS score*. It is produced by setting $G_1(x) = -\frac{1}{2}Wx^2$, $G_2(x) = \alpha x$, $\zeta_2(x) = \frac{1}{2}\alpha x^2$ and $a = 0$ in Eq. (2). Fissler and Ziegel (2016) explain that the score is strictly consistent, provided that the parameter W is chosen such that $WQ_t < ES_t$. (Recall that $ES_t < 0$ and $Q_t < 0$ because $\alpha < 50\%$.) In our empirical analysis, we used $W = 4$, as this was the smallest integer that ensured $WQ_t < ES_t$ for all pairs of forecasts of ES_t and Q_t from all methods in our study. We did not use the AS score for estimation because we could not guarantee that our chosen value of W would lead to $WQ_t < ES_t$ for all resulting pairs of forecasts of ES_t and Q_t .

We present the AL, NZ, FZG and AS scoring functions in Table 1. Our proposal is to use the first three of these to estimate forecast combining weights for the prediction of VaR and ES. In using such joint scoring functions for estimation, our work has similarities to that

of Taylor (2019) and Patton et al. (2019), who use the AL score to estimate dynamic models, and Dimitriadis and Bayer (2017), who present a regression framework for VaR and ES.

3. Methods for combining forecasts

3.1. Minimum score combining

This paper addresses the situation where we have a set of individual methods that each produces a forecast for the VaR and ES. As the quality of a method's VaR and ES forecasts may differ, it seems desirable to allow the combining weights for the VaR and ES to differ. However, it is not possible to distinguish the VaR accuracy from the ES accuracy, as the ES is equal to the sum of the VaR and the mean of the exceedances beyond the VaR. In view of this, our proposal is a formulation that does not combine ES forecasts, but instead combines forecasts of the difference between ES and VaR. We call this difference *spacing*. We refer to the method as *minimum score combining*, and express it as follows:

$$\hat{Q}_{ct} = \sum_{i=1}^M w_i^Q \hat{Q}_{it}, \quad (3)$$

$$\hat{ES}_{ct} = \hat{Q}_{ct} + \sum_{i=1}^M w_i^S \left(\hat{ES}_{it} - \hat{Q}_{it} \right), \quad (4)$$

where M is the number of individual methods; \hat{Q}_{it} is the quantile forecast and \hat{ES}_{it} the ES forecast produced by the i th individual method; \hat{Q}_{ct} is the combined quantile forecast; \hat{ES}_{ct} is the combined ES forecast; w_i^Q is the combining weight for the quantile forecast from the i th method; and w_i^S is the combining weight for the spacing between the ES and quantile forecasts from the i th method. We constrain the w_i^Q to be non-negative and to sum to 1, and impose the same constraints on the w_i^S . In addition to convex weights being common and

intuitively appealing, they ensure that \hat{ES}_{ct} will exceed \hat{Q}_{ct} , which is not easy to ensure if \hat{ES}_{ct} is constructed as a convex combination of the individual ES forecasts.

We estimated the two sets of combining weights, w_i^Q and w_i^S , in a single step by minimising a chosen scoring function. The optimal combining weights are those that lead to in-sample estimates for \hat{Q}_{ct} and \hat{ES}_{ct} that minimise the scoring function. We describe the minimisation further in Section 4.2.

3.2. Relative score combining

A simple method that is used to combine forecasts of the mean is to set convex combining weights to be inversely proportional to the mean squared error (MSE) (see Bates & Granger, 1969). This has the appeal of robustness when the estimation sample is small or there are many predictors (see for example Stock & Watson, 2001). Shan and Yang (2009) use the approach to combine quantile forecasts, but they use the quantile score to measure the accuracy instead of the MSE. We apply this idea to our VaR and ES context by using the joint scoring functions of Eq. (2) to measure the accuracy. The method leads to a single set of weights w_i for both VaR and ES prediction. We refer to the method as *relative score combining*, and present it as follows:

$$\hat{Q}_{ct} = \sum_{i=1}^M w_i \hat{Q}_{it}, \quad (5)$$

$$\hat{ES}_{ct} = \sum_{i=1}^M w_i \hat{ES}_{it}, \quad (6)$$

$$w_i = \frac{\exp\left(-\lambda \sum_{j=1}^{t-1} S\left(\hat{Q}_{ij}, \hat{ES}_{ij}, y_j\right)\right)}{\sum_{k=1}^M \exp\left(-\lambda \sum_{j=1}^{t-1} S\left(\hat{Q}_{kj}, \hat{ES}_{kj}, y_j\right)\right)}, \quad (7)$$

where S is the chosen joint scoring function, which is computed in each period j for each method i and then summed for all $t - 1$ in-sample observations; and $\lambda > 0$ is a tuning parameter that is included in the combining formulations of Shan and Yang (2009) and Stock and Watson (2001) for controlling how much the combining weights depend on the scoring function. A value of λ that is close to zero reduces the method to the simple average, while a high value of λ results in the selection of the individual method with the best historical accuracy. In our work, we optimised the value of λ by minimising the in-sample values of a chosen scoring function. We describe the optimisation further in Section 4.2.

4. Empirical analysis

Our empirical study considered the day-ahead forecasting of the 1% and 5% VaR and ES for daily log-returns of the following five stock indices: CAC 40, DAX 30, FTSE 100, NIKKEI 225 and S&P 500. We downloaded the data from Bloomberg. Each series consisted of the 6,000 daily observations, ending on 31 May 2017. The start dates for the five indices differed due to different holiday periods in each country, being 26 October 1993, 27 September 1993, 1 September 1993, 4 January 1993 and 4 August 1993 for the CAC 40, DAX 30, FTSE 100, NIKKEI 225 and S&P 500, respectively. Figure 1 shows the FTSE 100 returns, with the financial crisis being evident around 2008.¹

¹ Our decision to plot the FTSE 100 was made arbitrarily, as the time series of the other four indices showed similar features.

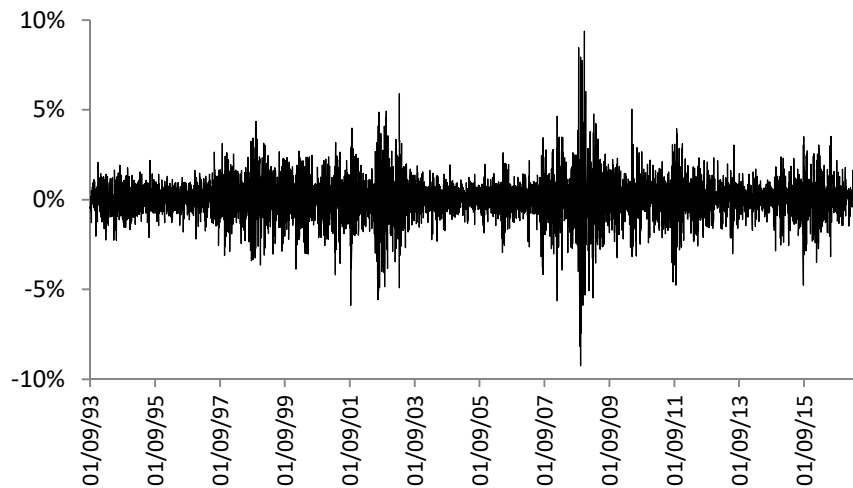


Figure 1. The series of 6,000 FTSE 100 returns ending on 31 May 2017.

We used a rolling window of 2,000 days, which we moved forward by one day at a time, for repeated re-estimation of the parameters of the individual forecasting methods. This enabled us to produce out-of-sample forecasts from each of these methods for the final 4,000 days in each series. Our combining methods focused on this period of 4,000 days, with a rolling window of 2,000 days being used for repeated re-estimation of the combining weights. The final 2,000 days were used to compare the out-of-sample forecast accuracies of the various methods. Prior to applying the VaR and ES estimation methods, we applied an autoregressive model of order 1 as an initial filter. The parameters of this filter were estimated using each rolling window of 2,000 returns.

4.1. Individual methods

As combining has the greatest potential when the individual methods use different information or use information in different ways, we implemented a diverse set of individual methods, including nonparametric, parametric and semiparametric time series methods, as well as a method based on intraday data. We now describe these methods.

4.1.1. Historical simulation

As a simple nonparametric method, we used historical simulation based on the 250 observations up to and including the forecast origin. We also considered the use of 100, 500 or 2,000 observations, but these did not lead to overall improvements in the forecast accuracy.

4.1.2. GJR-GARCH

As a common parametric method, we implemented a GJR-GARCH(1,1) model based on a Student t distribution. This asymmetric model was notably more accurate than a GARCH(1,1) model. We also considered filtered historical simulation, which applied historical simulation to the standardised residuals, as well as the method of McNeil and Frey (2000), which applies peaks-over-threshold extreme value theory (EVT) to the standardised residuals. However, these methods did not deliver substantial improvements, and so we used the Student t distribution, as this allowed us to have a fully parametric approach in our study.

4.1.3 CAViaR-AS-EVT

Conditional autoregressive value at risk (CAViaR) models are autoregressive quantile models that are estimated using quantile regression (see Engle & Manganelli, 2004). Although modelling VaR directly is appealing, it provides no insights regarding the ES. This limitation is addressed by Manganelli and Engle (2004), who estimate a CAViaR model for the 7.5% quantile and then apply peaks-over-threshold EVT to the exceedances after standardising by the corresponding quantile estimates. The fitted extreme value distribution is then used to obtain the VaR and ES estimates. We implemented this approach, and, in view of the superior performance of the asymmetric GJR-GARCH model relative to the GARCH

model, used the asymmetric slope (AS) CAViaR model, which we present in the following expression:²

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

4.1.4. CARE-AS

Expectiles are estimated by asymmetric least squares, and generalise the mean just as quantiles generalise the median (Nolde & Ziegel, 2017). They were introduced by Newey and Powell (1987), who explained that “expectiles are determined by tail expectations in the same way that quantiles are determined by the distribution”. The computational convenience of expectiles motivated Efron (1991) to suggest that they could be used to approximate quantiles. Drawing on this, Taylor (2008) proposed the use of a conditional autoregressive expectile (CARE) model for estimating VaR, and as the ES can be expressed as a simple function of the expectile, the CARE model can also be used to deliver a forecast for the ES. We implemented this approach in our study. An important issue is the choice of the τ expectile to use to approximate the α quantile. We optimised τ by following the approach of Taylor (2008), which involves re-estimating CARE models repeatedly, reducing the τ by 0.0001 each time, until the proportion of in-sample exceedances beyond the fitted expectile is closer to α than a predefined tolerance. Following initial experimentation, we started this procedure with values of $\tau = 0.0018$ and $\tau = 0.0167$ for the 1% and 5% probability levels, respectively. In view of our choice of the asymmetric GARCH and CAViaR models, we used the following asymmetric slope (AS) CARE model:

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

² The estimation of the parameters β_i proceeded by first sampling 10^4 candidate parameter vectors from uniform distributions with lower and upper bounds based on initial experimentation. As an additional candidate, we also included the parameter vector that had been optimised for the previous window of observations. From the set of candidate vectors, each of the three that gave the lowest values of the quantile score was used, in turn, as the initial vector in a quasi-Newton algorithm. The resulting vector with the lowest score was chosen as the final parameter vector.

where μ_t is the expectile. We estimated the parameters β_i using the same approach that we described in Section 4.1.3 for the CAViaR model, with the one difference being that we replaced the quantile score with the following expectile score:

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

4.1.5. HAR-Range

Intraday data have been found to be useful in estimating features of the distribution of daily returns. For example, the realized volatility has been used widely as a basis for forecasting the daily volatility. The heterogeneous autoregressive (HAR) model of the realized volatility is a simple and pragmatic approach, where a volatility forecast is constructed from the realized volatility over different time horizons (Corsi, 2009). However, intraday data can be expensive, and resources are required for pre-processing. Given the ready availability of the daily high and low prices, an alternative way of capturing the intraday volatility is to use the intraday range (see for example Alizadeh, Brandt, & Diebold, 2002; Gerlach & Chen, 2014). We take this approach, and follow Brownlees and Gallo (2010) by implementing the HAR model with the realized volatility replaced by the intraday range, as in the following expressions:

$$Range_t = \beta_1 + \beta_2 Range_{t-1} + \beta_3 Range_{t-1}^w + \beta_4 Range_{t-1}^m + \varepsilon_t,$$

$$Range_{t-1}^w = \frac{1}{5} \sum_{i=1}^5 Range_{t-i},$$

$$Range_{t-1}^m = \frac{1}{22} \sum_{i=1}^{22} Range_{t-i},$$

where $Range_t$ is the difference between the highest and lowest log prices on day t ; $Range_{t-1}^w$ and $Range_{t-1}^m$ are averages of $Range_t$ over a week and month, respectively; ε_t is an i.i.d. error term with zero mean; and the β_i are parameters that are estimated using least squares. The conditional variance is then expressed as a linear function of the square of $Range_t$, where the

intercept and the coefficient are estimated using maximum likelihoods based on a Student t distribution. This model is then used to produce a variance forecast, and VaR and ES forecasts are obtained by multiplying the forecast of the standard deviation by the VaR and ES of the Student t distribution.

4.2. Combining methods

We combined forecasts using the minimum score and relative score methods of Section 3, as well as simple averaging. We combined two different sets of forecasts. The first set included all five individual methods in Section 4.1. Even though the historical simulation method is known to be uncompetitive (see for example Chen, Gerlach, Hwang, & McAlee, 2012), we included it in the combination in order to check that the relative score and minimum score combining methods would produce sets of weights that would account for its weakness. The outcome should be that these combining methods outperform the simple average. Nevertheless, including a poor method in a combination increases the parameter estimation error unnecessarily, which can have a detrimental impact on the accuracy. Indeed, in practice, a method that is known to be poor is unlikely to be included. In view of this, we also applied the combining methods to a second set of individual methods that included just the four sophisticated methods, excluding historical simulation.

We estimated the combining weights using four different approaches, corresponding to the minimisation of the quantile score and the AL, NZ and FZG joint scoring functions. We obtained similar results when optimising with each of the joint scoring functions, and these results were better, or at least no worse, than those obtained by optimising the quantile score. In view of this, we report only the results produced using the AL score, to save

space.^{3,4}

In the minimum score combining method of Eqs. (3) and (4), we experimented with setting $w_i^s = w_i^q$ for each i , which implies the use of the same set of combining weights for the VaR and ES. The results were quite similar to those without the constraint, and so we report the results only for the unconstrained minimisation, to save space.

Figures 2 and 3 present the minimum score VaR and ES combining weights, respectively, for a combination of the five individual methods for the 5% probability level of the FTSE 100.⁵ For each of the 2,000 out-of-sample periods, the figures show the weights estimated by minimising the AL score using the 2,000 observations up to and including the forecast origin. Figure 2 also shows the minimised in-sample AL score plotted against the secondary y-axis. Although the ES weights are reasonably volatile over the out-of-sample period, the minimised AL score evolves smoothly, providing reassurance that the volatile ES weights are not due to a faulty optimisation procedure. For most of the out-of-sample period, CAViaR-AS-EVT and HAR-Range have the largest combining weights in Figures 2 and 3.

The corresponding combining weights for the relative score method are shown in Figure 4. Recall that, for this method, the set of weights for VaR combining is the same as that for ES combining. The figure shows GJR-GARCH and HAR-Range as having the largest weights for the first half of the plot. Interestingly, historical simulation has non-zero weights in Figures 2 to 4, even though it is the least accurate of the individual methods, as we show in Section 4.3.

³ For the combining methods, we used an optimisation approach similar to that described in Section 4.1.3 for the CAViaR model. For minimum score combining, we used 10^5 candidate parameter vectors, with entries sampled from uniform distributions between 0 and 1. We also included, as an additional candidate, the parameter vector that had been optimised for the previous window of observations. The 10 candidate vectors with the lowest values of the scoring function were then each used as the initial vector in a quasi-Newton algorithm. The resulting vector with the lowest scoring function, was chosen as the final parameter vector. For relative score combining, which has only the one parameter λ , we used 10^4 candidate values.

⁴ An online appendix contains out-of-sample results for the combining methods with weights estimated using the quantile score, and the NZ and FZG scoring functions.

⁵ Our decision to focus here on the FTSE 100 was made arbitrarily. We do not present the corresponding figures for the other four indices, as this would take considerable additional space, without providing significant useful additional insight.

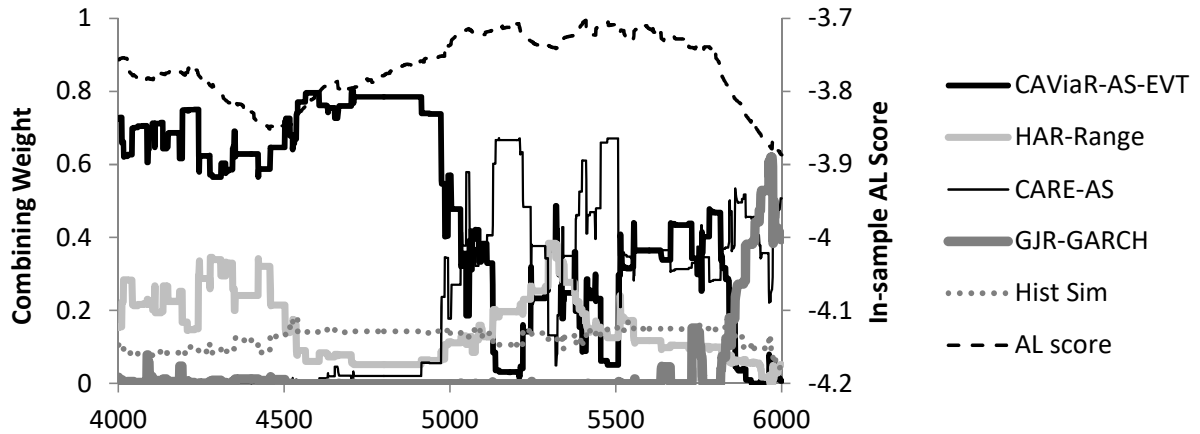


Figure 2. Minimum score combining weights w_i^o for the VaR combining of Eq. (3) used to combine five methods for the 5% probability level of the FTSE 100. The weights are optimised by minimising the in-sample AL score, which is also shown in the plot.

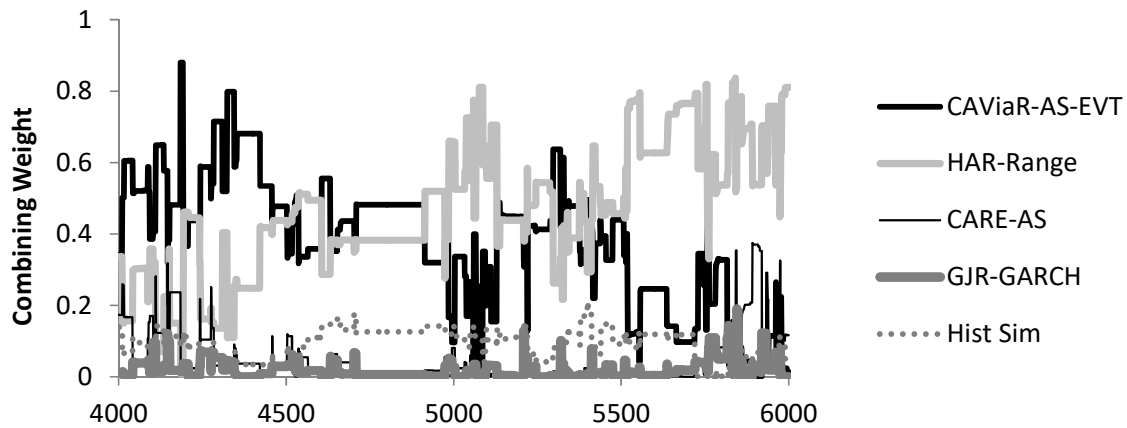


Figure 3. Minimum score combining weights w_i^s for the spacings combination of Eq. (4) used to combine five methods for the 5% probability level of the FTSE 100. The weights are optimised by minimising the in-sample AL score.

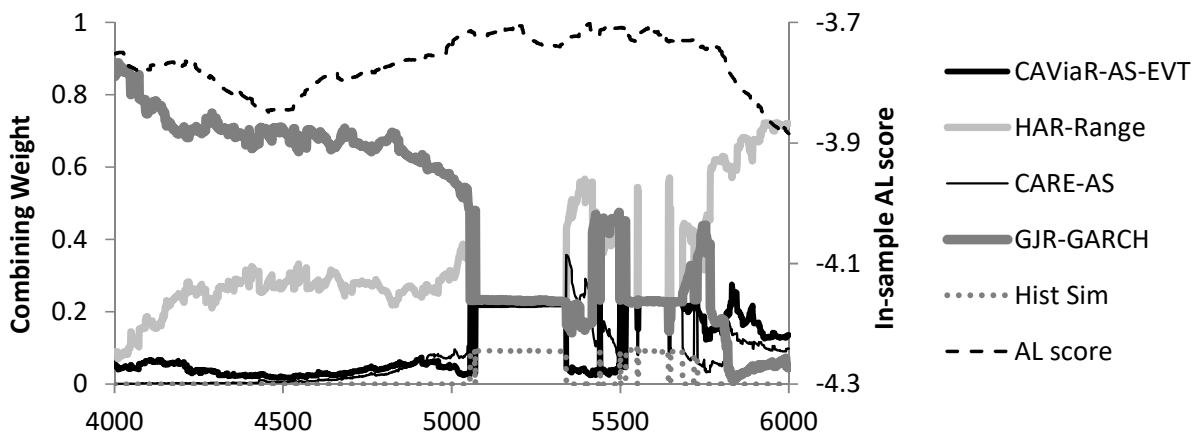


Figure 4. Relative score combining weights of Eqs. (5) to (7), used to combine five methods for the 5% probability level of the FTSE 100. The method is optimised by minimising the in-sample AL score, which is also shown in the plot.

4.3. Backtesting VaR and ES forecasts

Traditionally, VaR and ES forecasts have tended to be evaluated using tests that Nolde and Ziegel (2017) describe as unconditional and conditional *calibration* tests. We use these in Section 4.3.1, and then consider scoring functions in Sections 4.3.2. Our out-of-sample evaluation focuses on the final 2,000 periods of each series, as we have out-of-sample forecasts for each of these periods from all of the individual and combining methods.

4.3.1. Backtesting VaR and ES forecasts with calibration tests

Typically, VaR forecasts are evaluated using calibration tests. For probability level α , a quantile forecast \hat{Q}_t is unconditionally calibrated if the variable $Hit_t = \alpha - I(y_t \leq \hat{Q}_t)$ has zero unconditional expectation, and is conditionally calibrated if Hit_t has zero conditional expectation. We evaluated the unconditional calibration using a test based on the binomial distribution to test whether the mean of Hit_t was significantly different from zero. Following common practice, our calibration testing did not attempt to incorporate parameter estimation error (see Escanciano & Olmo, 2010). For unconditional calibration, Table 2 summarises the test results for the five indices at the 1% and 5% probability levels in the columns labelled ‘VaR hit %’. The values reported are the numbers of indices for which the test was significant at the 5% significance level. Throughout Table 2, lower values are better. The table shows that, in the columns headed VaR hit %, there are non-zero entries only for the historical simulation, GJR-GARCH and HAR-Range methods.

We tested for VaR conditional calibration by implementing Engle and Manganelli’s (2004) dynamic quantile test with four lags in the test’s regression, and summarise the results for the five indices in Table 2. For the 1% probability level, the historical simulation and HAR-Range methods are the worst-performing methods, while for the 5% VaR, the poorest results are for historical simulation.

Table 2
Results of calibration tests for the five stock indices.

	1% probability level			5% probability level		
	VaR hit %	VaR dynamic quantile	ES bootstrap test	VaR hit %	VaR dynamic quantile	ES bootstrap test
<i>Individual methods</i>						
Historical simulation	3	5	3	0	5	1
GJR-GARCH	1	0	1	0	0	3
HAR-Range	2	2	0	1	0	1
CARE-AS	0	0	0	0	1	0
CAViaR-AS-EVT	0	0	0	0	0	0
<i>Combining all</i>						
Simple average	0	1	0	0	0	0
Relative score	0	0	0	0	0	1
Minimum score	0	0	0	0	0	1
<i>Combining all except historical simulation</i>						
Simple average	0	0	1	0	0	0
Relative score	0	0	0	0	0	1
Minimum score	0	0	0	0	0	1

Notes: The values presented are the numbers of indices for which the test was significant at the 5% significance level. Lower values are better, in that they indicate smaller numbers of tests for which calibration was rejected. The tests are described in Section 4.3.1.

Following the approach of McNeil and Frey (2000), we evaluated the ES forecasts by testing for a zero mean in the discrepancy between the observed return and the ES forecast for the periods in which the return exceeds the VaR forecast. We standardised by dividing each discrepancy by the corresponding VaR estimate. The test examines whether the standardised discrepancies have a zero unconditional expectation. We avoided any assumptions about the distributions of the standardized discrepancies by using the dependent circular block bootstrap used by Jalal and Rockinger (2008). The results are presented in Table 2. The table shows that historical simulation performs relatively poorly for the 1% probability level, while the GJR-GARCH model performed the worst for the 5% probability level.

4.3.2. Backtesting VaR and ES forecasts with scoring functions

We evaluated the VaR forecast accuracy by calculating the quantile score, then calculated the *quantile skill score* for each method as the ratio of the score to that of the historical simulation method. We then subtracted this ratio from 1, and multiplied the result by 100. We report the skill scores in Table 3 for the 1% probability level. For all skill scores, higher values are preferable. We summarised performances across the five stock indices by calculating the geometric mean of the ratios of the score of each method to the score for the historical simulation reference method, then subtracting this from one and multiplying the result by 100. The resulting values are presented in the final column of the table.

Table 3
1% VaR evaluated using the quantile skill score (%).

	CAC	DAX	FTSE	NIKKEI	S&P	Geo. mean
<i>Individual methods</i>						
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	16.4	16.5	21.1	22.0	24.4	20.0
HAR-Range	18.3	20.6	20.2	18.8	26.4	20.8
CARE-AS	17.8	18.7	21.9	21.0	25.4	20.9
CAViaR-AS-EVT	17.0	17.3	21.3	21.5	23.7	20.1
<i>Combining all</i>						
Simple average	18.9	20.6	21.5	21.6	24.3	21.4
Relative score	18.9	20.9	21.3	22.6	25.9	21.9
Minimum score	18.8	21.2	22.2	22.2	24.5	21.8
<i>Combining all except historical simulation</i>						
Simple average	18.7	20.4	22.4	23.3	26.2	22.2
Relative score	18.9	21.2	21.1	22.6	26.6	22.1
Minimum score	19.3	20.9	22.1	22.8	26.8	22.4

Notes: The quantile score is presented in Eq. (1). Higher skill score values are better. Bold indicates the best method(s) in each column.

As we discussed in Sections 1 and 2, although ES is not elicitable by itself, it is elicitable jointly with VaR. Therefore, in terms of scoring functions, we must evaluate ES jointly with VaR. We do this using the AL score. However, as the AL score takes negative values, we computed its skill score slightly differently from the quantile skill score. That is, we calculated the *AL skill score* of each method as the ratio of the score to that of the

historical simulation method, then subtracted 1 from this ratio and multiplied the result by 100. Table 4 presents the AL skill scores for the 1% probability level.

Table 4
1% VaR and ES evaluated using AL skill score (%).

	CAC	DAX	FTSE	NIKKEI	S&P	Geo. mean
<i>Individual methods</i>						
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	6.6	6.4	8.6	12.9	8.9	8.7
HAR-Range	9.0	9.6	8.9	10.7	10.8	9.8
CARE-AS	7.6	7.6	8.9	13.2	10.6	9.5
CAViaR-AS-EVT	7.3	7.4	8.7	12.4	9.2	9.0
<i>Combining all</i>						
Simple average	8.7	9.0	9.2	13.0	10.2	10.0
Relative score	8.7	9.3	9.0	13.7	10.3	10.2
Minimum score	8.8	9.3	9.6	13.5	10.0	10.2
<i>Combining all except historical simulation</i>						
Simple average	8.3	9.1	9.5	14.1	10.5	10.3
Relative score	8.7	9.6	8.9	13.7	10.6	10.3
Minimum score	9.1	9.3	9.6	13.8	10.6	10.5

Notes: Table 1 defines the AL score, which is a version of the joint VaR and ES score of Eq. (2). Higher skill score values are better. Bold indicates the best method(s) in each column.

Table 5 summarises the out-of-sample results for the quantile score and the four joint scoring functions that we presented in Table 1, which evaluate the VaR and ES forecast accuracies jointly. For each scoring function, the table shows the skill scores averaged across the five indices. For the 1% probability level, the results for the quantile score and the AL score were also reported in the final columns of Tables 3 and 4. As the NZ, FZG and AS scores are all positive-valued, we computed their skill scores using the same approach that we used for the quantile score. An alternative to our use of a finite set of joint scoring functions is proposed by Ziegel, Krüger, Jordan, and Fernando (2017), who build on the work of Ehm, Gneiting, Jordan, and Krüger (2016) by using Murphy diagrams to establish whether one method dominates another in terms of a class of joint scoring functions.

Table 5

VaR evaluated using the quantile skill score (%), and VaR and ES evaluated jointly using the AL, NZ, FZG and AS skill scores (%).

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
<i>Individual methods</i>										
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	20.0	8.7	11.4	20.0	33.0	11.1	4.2	6.6	11.2	19.2
HAR-Range	20.8	9.8	12.4	20.8	33.4	10.5	4.2	6.4	10.6	18.2
CARE-AS	20.9	9.5	12.2	21.0	34.1	11.6	4.4	6.9	11.7	19.8
CAViaR-AS-EVT	20.1	9.0	11.7	20.1	31.9	11.6	4.4	6.8	11.7	19.8
<i>Combining all</i>										
Simple average	21.4	10.0	12.6	21.4	34.2	11.7	4.6	7.0	11.8	20.0
Relative score	21.9	10.2	12.9	21.9	35.1	11.9	4.6	7.1	12.0	20.5
Minimum score	21.8	10.2	12.9	21.8	34.7	11.9	4.6	7.1	12.0	20.4
<i>Combining all except historical simulation</i>										
Simple average	22.2	10.3	13.1	22.2	35.3	12.1	4.7	7.2	12.2	20.6
Relative score	22.1	10.3	13.0	22.1	35.3	11.9	4.6	7.1	12.0	20.4
Minimum score	22.4	10.5	13.2	22.4	35.6	12.0	4.7	7.2	12.1	20.5

Notes: The quantile score is presented in Eq. (1). Table 1 defines the AL, NZ, FZG and AS scores, which are versions of the joint VaR and ES score of Eq. (2). The values presented here are the result of averaging skill scores across the five indices. Bold indicates the best method(s) in each column.

We have the following comments regarding Tables 3 to 5:

- (i) The results are reasonably consistent across the five indices.
- (ii) The results are reasonably consistent across the quantile score and the four joint VaR and ES scores.
- (iii) For the 1% probability level, the best of the five individual methods were the HAR-Range approach and the CARE method.
- (iv) For the 5% probability level, the best of the five individual methods were the CARE and CAViaR-based methods.
- (v) For both the 1% and 5% probability levels, all of the combining methods outperformed all of the individual methods.
- (vi) For the 1% probability level, minimum score combining of the four competitive individual methods was the most accurate approach.

- (vii) For the 5% probability level, the simple average of the four competitive individual methods was the most accurate approach, just slightly outperforming minimum score combining. To try to provide some explanation for why the simple average is more competitive for the 5% probability level than for the 1% probability level, we first note that the other two combining methods were optimised by minimising the AL score. Looking at the columns of Table 5 that correspond to the AL score, we see that the results for the four sophisticated individual methods differ notably for the 1% probability level, but are quite similar for the 5% probability level. Thus, it is not too surprising that the simple average was a reasonable approach for the 5% probability level, while a weighted average combination was preferable for the 1% probability level.
- (viii) Perhaps unsurprisingly, the simple average was affected more by the inclusion of the historical simulation approach in the combination than the other combining methods. However, interestingly the other two combining methods were also affected to a certain degree, with the minimum score approach producing notably better results for the 1% probability level when historical simulation was excluded. This supports the view that it is unwise to include a poor method in a combination.

The model confidence set (MCS) testing framework of Hansen, Lunde, and Nason (2011) enables one to obtain a set of models for which there is a pre-specified probability that the set contains the best model, when judged by a chosen loss function. If a model is not contained in the MCS, it is considered to be less likely to be the best model than those that are included in the MCS. We implemented MCS testing separately based on the quantile score and the four joint scoring functions, which evaluate the VaR and ES forecast accuracies jointly. In each MCS test, we used the *equivalence* test based on the Diebold-Mariano test and the one-sided *elimination* rule described as $T_{max,M}$ by Hansen et al. (2011). We followed Hansen et al. (2011) by considering 75% and 90% confidence levels, and we report the

results of the tests in Tables 6 and 7, respectively. For each scoring function, the tables report the numbers of indices for which each method was included in the MCS. As we have five indices in our study, the best possible value in each table is 5. With five indices, five scoring functions, two probability levels and two confidence levels, we applied the MCS test 100 times. Historical simulation was clearly the worst method, while the other four individual methods were included in most of the sets. This is perhaps not surprising, as we chose these methods because we felt that they would be competitive, meaning that they might be useful in a combination. In the rows of Tables 6 and 7 that corresponding to the combining methods, only one entry is not 5, indicating that the combining methods were excluded from just one MCS out of the 100 that we constructed.

Table 6

VaR evaluated using model confidence sets based on the quantile score, and VaR and ES evaluated jointly using model confidence sets based on the AL, NZ, FZG and AS scores. The values presented are the numbers of indices for which each method is within the model confidence set for a 75% confidence level.

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
<i>Individual methods</i>										
Historical simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	4	3	3	4	4	4	4	4	4	3
HAR-Range	5	5	5	5	5	4	5	4	4	3
CARE-AS	5	3	4	5	5	5	5	5	5	4
CAViaR-AS-EVT	4	3	4	4	3	5	5	5	5	5
<i>Combining all</i>										
Simple average	5	5	5	5	5	5	5	5	5	5
Relative score	5	5	5	5	5	5	5	5	5	5
Minimum score	5	5	5	5	4	5	5	5	5	5
<i>Combining all except historical simulation</i>										
Simple average	5	5	5	5	5	5	5	5	5	5
Relative score	5	5	5	5	5	5	5	5	5	5
Minimum score	5	5	5	5	5	5	5	5	5	5

Notes: The quantile score is presented in Eq. (1). Table 1 defines the AL, NZ, FZG and AS scores, which are versions of the joint VaR and ES score in Eq. (2). Higher values in this table are better, with 5 being the highest possible.

Table 7

VaR evaluated using model confidence sets based on the quantile score, and VaR and ES evaluated jointly using model confidence sets based on the AL, NZ, FZG and AS scores. The values presented are the numbers of indices for which each method is within the model confidence set for a 90% confidence level.

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
<i>Individual methods</i>										
Historical simulation	0	0	0	0	0	0	0	0	0	0
GJR-GARCH	4	4	4	4	4	4	4	4	5	5
HAR-Range	5	5	5	5	5	5	5	5	5	5
CARE-AS	5	5	5	5	5	5	5	5	5	5
CAViaR-AS-EVT	4	5	5	4	4	5	5	5	5	5
<i>Combining all</i>										
Simple average	5	5	5	5	5	5	5	5	5	5
Relative score	5	5	5	5	5	5	5	5	5	5
Minimum score	5	5	5	5	5	5	5	5	5	5
<i>Combining all except historical simulation</i>										
Simple average	5	5	5	5	5	5	5	5	5	5
Relative score	5	5	5	5	5	5	5	5	5	5
Minimum score	5	5	5	5	5	5	5	5	5	5

Notes: The quantile score is presented in Eq. (1). Table 1 defines the AL, NZ, FZG and AS scores, which are versions of the joint VaR and ES score in Eq. (2). Higher values in this table are better, with 5 being the highest possible.

Our comparison of methods has been based on the final 2,000 observations in each series. With the financial crisis beginning not long before the start of this period, Figure 1 shows that the first half of this period is perhaps more volatile than the second half. Table 8 compares the results for the two halves of the out-of-sample period, focusing on the AL skill score, which evaluates both VaR and ES forecast accuracies. The results in this table confirm that the rankings of methods were broadly similar for both halves of our out-of-sample period.

Table 8

VaR and ES evaluated using the AL skill score (%) for different out-of-sample periods.

	1% probability level			5% probability level		
	Penultimate 1,000 days	Final 1,000 days	All 2,000 days	Penultimate 1,000 days	Final 1,000 days	All 2,000 days
<i>Individual methods</i>						
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	12.8	4.9	8.7	5.7	2.8	4.2
HAR-Range	13.6	6.4	9.8	5.8	2.7	4.2
CARE-AS	14.3	5.2	9.5	6.5	2.6	4.4
CAViaR-AS-EVT	12.8	5.6	9.0	6.1	2.8	4.4
<i>Combining all</i>						
Simple average	13.9	6.6	10.0	6.4	2.9	4.6
Relative score	14.2	6.5	10.2	6.4	3.0	4.6
Minimum score	14.2	6.7	10.2	6.5	3.0	4.6
<i>Combining all except historical simulation</i>						
Simple average	14.5	6.5	10.3	6.5	3.1	4.7
Relative score	14.4	6.6	10.3	6.3	3.0	4.6
Minimum score	14.7	6.7	10.5	6.5	3.0	4.7

Notes: Table 1 defines the AL score, which is a version of the joint VaR and ES score in Eq. (2). The values presented here are the result of averaging skill scores across the five indices. Higher skill score values are better. Bold indicates the best method(s) in each column.

5. Simulation study

We investigated the combining methods further by implementing an empirical study using data simulated from the following three data generating processes (DGP):

DGP1: This is a GJR-GARCH(1,1) process with a Student t distribution. We chose the parameters to be the average of the 2,000 sets of parameter values that we had estimated for the filtered FTSE 100 returns using the 2,000 rolling windows.

DGP2: We generated data from a GJR-GARCH(1,1) process with the error term drawn alternately from a Student t distribution with three degrees of freedom and a Gamma distribution with its shape and scale parameters equal to 2 and 0.5, respectively. The values drawn from the Gamma distribution were standardised so that they had a zero mean and unit variance. We used the same GJR-GARCH model parameters as in

DGP1. Note that this process and DGP3 were used by Manganelli and Engle (2004) in their study of VaR and ES.

DGP3: This process was based on an asymmetric slope CAViaR process with a probability level of 15%. We chose this model because it is used in one of the individual methods that is included in our empirical analysis. We used parameters that were the average of those estimated using the rolling windows for the filtered FTSE 100 returns. Starting with initial values q_0 and y_0 for the quantile and the observation, we used the CAViaR process to generate q_1 , the value of the quantile for the first period. The simulated value y_1 for this period was then generated from a distribution with a quantile equal to q_1 . This was achieved by multiplying a randomly-generated value by a standard deviation equal to q_1 divided by the quantile of the distribution that is used to generate the random value. We generated the random values by sequentially using Student t distributions with three and four degrees of freedom, and a Gamma distribution with its shape and scale parameters equal to 2 and 0.5, respectively. We then repeated this procedure using q_{t-1} and y_{t-1} in the CAViaR model for generating q_t , which is then used with a randomly-sampled value to produce y_t .

We generated one series of 6,000 observations from each of the three DGPs. The structure of our study matched our analysis of the stock indices, with rolling windows of 2,000 periods being used for repeated re-estimation of parameters. The final 2,000 periods were then used for comparing the out-of-sample forecast accuracy. We implemented the same individual methods that we considered for the stock indices, with the exception of the HAR-Range method, which relies on intraday data. We considered combinations of all four individual methods, as well as combinations that excluded historical simulation. Tables 9 to 11 present out-of-sample skill scores for the simulated series. We have the following comments regarding these results:

- (i) For DGP1, the GJR-GARCH model is optimal, and so it was to be expected that this method would perform well. However, it is noticeable that its results were matched by those from the relative score and minimum score combining methods.
- (ii) For DGP1, the simple average benefitted from the removal of historical simulation. For DGP2 and DGP3, the same was true only for the 5% probability level.
- (iii) For all three processes, removing historical simulation from the combination failed to improve the results of the relative score and minimum score combining methods noticeably.
- (iv) For DGP2, simple average combining was the best method for the 1% probability level. For the 5% probability level, all combining methods performed well, with the best results being achieved by the simple average with historical simulation excluded.
- (v) Given the nature of DGP3, it is not surprising to see that the CAViaR-AS-EVT method performed well for this process. However, comparable results were achieved by simple average combining for the 1% probability level. For the 5% probability level, all of the combining methods performed very well, with the exception of the simple average with historical simulation included.

Table 9

For simulated data from DGP1, VaR evaluated using the quantile skill score (%), and VaR and ES evaluated jointly using the AL, NZ, FZG and AS skill scores (%).

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
<i>Individual methods</i>										
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	25.1	8.9	14.1	24.9	41.1	14.2	4.7	8.1	14.1	23.0
CARE-AS	24.2	8.6	13.5	24.0	40.5	13.8	4.5	7.9	13.7	22.2
CAViaR-AS-EVT	23.3	8.3	13.1	23.1	37.7	14.0	4.6	8.0	13.9	22.5
<i>Combining all</i>										
Simple average	21.9	7.7	12.3	21.7	36.4	12.9	4.1	7.2	12.8	21.5
Relative score	25.2	8.9	14.1	25.0	41.1	14.2	4.7	8.1	14.1	23.0
Minimum score	24.4	8.7	13.7	24.2	39.5	14.1	4.6	8.1	14.0	22.9
<i>Combining all except historical simulation</i>										
Simple average	25.1	8.8	14.0	24.8	41.8	14.1	4.6	8.1	14.0	22.7
Relative score	25.2	8.9	14.1	25.0	41.1	14.2	4.7	8.1	14.1	23.0
Minimum score	25.2	8.9	14.1	24.9	41.2	14.2	4.7	8.1	14.1	23.0

Notes: The quantile score is presented in Eq. (1). Table 1 defines the AL, NZ, FZG and AS scores, which are versions of the joint VaR and ES score in Eq. (2). Higher skill score values are better. Bold indicates the best method(s) in each column.

Table 10

For simulated data from DGP2, VaR evaluated using the quantile skill score (%), and VaR and ES evaluated jointly using the AL, NZ, FZG and AS skill scores (%).

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
<i>Individual methods</i>										
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	14.4	8.7	10.0	14.5	19.8	10.5	5.4	7.4	10.6	12.6
CARE-AS	14.8	8.0	9.7	14.9	18.2	10.6	5.3	7.3	10.8	12.9
CAViaR-AS-EVT	14.8	8.9	10.4	15.0	17.5	10.2	5.3	7.2	10.3	11.6
<i>Combining all</i>										
Simple average	16.8	9.4	11.2	16.9	22.2	9.8	5.1	7.0	9.9	12.1
Relative score	14.7	8.8	10.2	14.8	20.1	10.7	5.5	7.6	10.8	12.9
Minimum score	14.9	8.7	10.2	15.0	20.1	10.5	5.4	7.4	10.7	12.8
<i>Combining all except historical simulation</i>										
Simple average	16.2	9.5	11.1	16.4	20.4	10.8	5.6	7.7	11.0	13.0
Relative score	14.6	8.7	10.2	14.8	20.1	10.7	5.5	7.6	10.8	12.9
Minimum score	14.6	8.6	10.0	14.7	19.4	10.7	5.5	7.5	10.8	12.9

Notes: The quantile score is presented in Eq. (1). Table 1 defines the AL, NZ, FZG and AS scores, which are versions of the joint VaR and ES score in Eq. (2). Higher skill score values are better. Bold indicates the best method in each column.

Table 11

For simulated data from DGP3, VaR evaluated using the quantile skill score (%), and VaR and ES evaluated jointly using the AL, NZ, FZG and AS skill scores (%).

	1% probability level					5% probability level				
	Quantile score	AL	NZ	FZG	AS	Quantile score	AL	NZ	FZG	AS
<i>Individual methods</i>										
Historical simulation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GJR-GARCH	17.6	13.1	12.7	17.8	23.1	13.4	6.8	9.0	13.5	16.8
CARE-AS	18.6	13.8	13.4	18.7	22.9	12.8	6.5	8.6	12.9	16.1
CAViaR-AS-EVT	19.4	14.8	14.2	19.6	23.8	13.5	7.2	9.3	13.6	16.5
<i>Combining all</i>										
Simple average	19.4	13.6	13.5	19.5	24.9	12.8	6.3	8.5	12.9	17.0
Relative score	18.3	13.5	13.2	18.5	22.5	13.4	6.9	9.1	13.6	17.0
Minimum score	18.6	13.5	13.2	18.7	23.5	13.5	6.9	9.1	13.6	17.1
<i>Combining all except historical simulation</i>										
Simple average	19.7	14.6	14.2	19.8	24.7	13.7	7.1	9.3	13.8	17.2
Relative score	18.5	13.9	13.5	18.7	23.1	13.5	7.1	9.2	13.6	16.9
Minimum score	17.8	13.5	13.0	18.0	21.6	13.6	7.1	9.3	13.7	17.0

Notes: The quantile score is presented in Eq. (1). Table 1 defines the AL, NZ, FZG and AS scores, which are versions of the joint VaR and ES score in Eq. (2). Higher skill score values are better. Bold indicates the best method(s) in each column.

6. Summary and concluding comments

This paper has introduced forecast combination to ES prediction. As ES is not elicitable, we estimate combining weights for VaR and ES simultaneously using recently-proposed joint scoring functions. Our minimum score combining approach allows convex combining weights to differ for VaR and ES prediction. We also considered a relative score combining approach that has the appeal of requiring the estimation of only one parameter.

Our empirical study of stock indices focused on the scoring functions for comparing the methods. We combined parametric, nonparametric and semiparametric time series methods, as well as a method based on the intraday range. Looking at the average performances across our set of stock indices, we found that all of the individual methods were outperformed by all of the combining methods. For the 1% probability level, the best results were provided by weighted combining approaches. The simple average also performed well, particularly for the 5% probability level, provided that the poorly-performing historical

simulation method was omitted from the combination. A simulation study provided support for the combining methods.

In future work, it would be interesting to consider alternative combining methods, and different sets of individual methods. Multi-step-ahead prediction is another potential area for future work. Combining weights for multiple lead times could be estimated by minimising a joint score summed over different lead times, or the weights could be estimated separately for each lead time, which seems appealing, as the relative performances of methods can vary across lead times. It would also be interesting to compare forecast accuracies using the Murphy diagrams of Ziegel et al. (2017), and perhaps also measures of economic significance.

Acknowledgements

We are very grateful to an associate editor and two referees for providing comments that helped to improve the paper greatly.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.ijforecast.2019.5.05.014>. It contains out-of-sample results for the combining methods with weights estimated using the quantile score, and the NZ and FZG scoring functions.

References

- Acerbi, C., & Székeley, B. (2014). Backtesting expected shortfall. *Risk*, December, 76-81.
- Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *Journal of Finance*, 57(3), 1047-1091.

- Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203-228.
- Basel Committee (2016). *Minimum capital requirements for market risk*. Technical report, Basel Committee on Banking Supervision.
- Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4), 451-468.
- Brownlees, C. T., & Gallo, G. M. (2010). Comparison of volatility measures: A risk management perspective. *Journal of Financial Econometrics*, 8(1), 29-56.
- Chen, C. W. S., Gerlach, R., Hwang, B. B. K., & McAleer, M. (2012). Forecasting Value-at-Risk using nonlinear regression quantiles and the intra-day range. *International Journal of Forecasting*, 28(3), 557-574.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174-196.
- Dimitriadis, T., & Bayer, S. (2017). *A joint quantile and expected shortfall regression framework*. Working paper, arXiv:1704.02213v3.
- Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1, 93-125.
- Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(3), 505-562.
- Elliott, G., & Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, 122(1), 47-80.
- Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, 22(4), 367-381.
- Escanciano, J. C., & Olmo, J. (2010). Backtesting parametric value-at-risk with estimation risk. *Journal of Business and Economic Statistics*, 28(1), 36-51.

- Fissler, T., & Ziegel, J. A. (2016). Higher order elicibility and Osband's Principle. *Annals of Statistics*, 44(4), 1680-1707.
- Fissler, T., Ziegel, J. A., & Gneiting, T. (2016). Expected shortfall is jointly elicitable with value at risk – implications for backtesting. *Risk*, January, 58-61.
- 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.
- Gerlach, R., & Chen, C.W. (2014). Bayesian expected shortfall forecasting incorporating the intraday range. *Journal of Financial Econometrics*, 14(1), 128-158.
- Giacomini, R., & Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business and Economic Statistics*, 23(4), 416-431.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746-762.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378.
- Granger, C. W. J. (1989). Combining forecasts – twenty years later. *Journal of Forecasting*, 8(3), 167-173.
- Granger, C. W. J., White, H., & Kamstra, M. J. (1989). Interval forecasting: An analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, 40(1), 87-96.
- Halbleib, R., & Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from financial crisis. *Journal of Economic Dynamics and Control*, 36(8), 1212-1228.
- Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *Journal of Forecasting*, 23(1), 1-13.
- Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453-497.

- 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.
- Jeon, J., & Taylor, J. W. (2013). Using implied volatility with CAViaR models for value at risk estimation. *Journal of Forecasting*, 32(1), 62-74.
- Jore, A. S., Mitchell, J., & Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25(4), 621–634.
- Manganelli, S., & Engle, R. F. (2004). A comparison of value-at-risk models in finance. In G. Szegö (ed.), *Risk measures for the 21st century* (pp. 123-144). Chichester, Wiley.
- McAleer, M., Jiménez-Martín, J.-A., & Pérez-Amaral, T. (2013a). Has the Basel Accord improved risk management during the global financial crisis? *The North American Journal of Economics and Finance*, 26(C), 250-265.
- McAleer, M., Jiménez-Martín, J.-A., & Pérez-Amaral, T. (2013b). International evidence on GFC-robust forecasts for risk management under the Basel Accords. *Journal of Forecasting*, 32(3), 267-288.
- McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7(3-4), 271-300.
- Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819-847.
- Nieto, M. R., & Ruiz, E. (2016). Frontiers in VaR forecasting and backtesting. *International Journal of Forecasting*, 32(2), 475-501.
- Nolde, N., & Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), 1833-1874.
- Opschoor, A., Van Dijk, D., & van der Wel, M. (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics*, 32(7), 1298-1313.

- Patton, A. J., Ziegel, J. F., & Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2), 388-413.
- Shan, K., & Yang, Y. (2009). Combining regression quantile estimators. *Statistica Sinica*, 19(3), 1171-1191.
- Stock, J. H., & Watson, M. (2001). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R.F. Engle & H. White (eds.), *Festschrift in honour of Clive Granger* (pp. 1-44). Cambridge, Cambridge University Press.
- 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 and Economic Statistics*, 37, 121-133.
- 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.
- Ziegel, J. F., Krüger, F., Jordan, A., & Fernando, F. (2017). *Murphy diagrams: Forecast evaluation of expected shortfall*. Working paper, arXiv:1705.04537.