Modelling uncertainty in financial tail risk: a forecasting combination and weighted quantile approach

Giuseppe Storti[1], Chao Wang[2]

[1] Department of Economics and Statistics, University of Salerno

[2] Discipline of Business Analytics, The University of Sydney

Abstract

A novel forecasting combination and weighted quantile based tail risk forecasting framework is proposed, aiming to reduce the impact of modelling uncertainty in financial tail risk forecasting. The proposed approach is based on a two-step estimation procedure. The first step involves the combination of Value-at-Risk (VaR) forecasts at a grid of different quantile levels. A range of parametric and semi-parametric models is selected as the model universe which is incorporated in the forecasting combination procedure. The quantile forecasting combination weights are estimated by optimizing the quantile loss. In the second step, the Expected Shortfall (ES) is computed as a weighted average of combined quantiles. The quantiles weighting structure used to generate the ES forecast is determined by minimizing a strictly consistent joint VaR and ES loss function of the Fissler-Ziegel class. The proposed framework is applied to six stock market indices and its forecasting performance is compared to each individual model in the model universe and a simple average approach. The forecasting results based on a number of evaluations support the proposed framework.

Keywords: Value-at-Risk, Expected Shortfall, forecasting combination, weighted quantile, quantile loss, joint loss.

1 Introduction

Since the introduction by J.P. Morgan in the RiskMetrics model at 1993, Value-at-Risk (VaR) has been widely employed by financial institutions and corporations around the world to assist their decision making in relation to capital allocation and risk management. VaR is a quantitative tool to measure and control financial risk and represents the market risk as one number. VaR has become a standard measurement for capital allocation and risk management. Let $\mathcal{I}_t$ be the information available at time t and

$$F_t(r) = Pr(r_t \leq r|\mathcal{I}_{t-1})$$

be the Cumulative Distribution Function (CDF) of return r_t conditional on $\mathcal{I}_{t-1}$. We assume that $F_t(.)$ is strictly increasing and continuous on the real line $\Re$. Under this assumption, the α level VaR (quantile) at time t can be defined as:

$$Q_{t,\alpha} = F_t^{-1}(\alpha), \qquad 0 < \alpha < 1.$$

However, VaR has been subject to criticism because it cannot measure the expected loss for violations and is not mathematically coherent, in that it can favor non-diversification. Expected Shortfall (ES), proposed by Artzner (1997) and Artzner et al. (1999), gives the expected loss, conditional on returns exceeding a VaR threshold, and is a coherent measure; thus, in recent years it has become more widely employed for tail risk measurement and is now favored by the Basel Committee on Banking Supervision. Within the same framework as above, the α level ES can be shown to be equal to the tail conditional expectation of r_t (see Acerbi and Tasche, 2002, among others):

$$ES_{t,\alpha} = E(r_t|r_t \leq Q_{t,\alpha}, \mathcal{I}_{t-1}). \tag{1}$$

The Basel III Accord, which was implemented in 2019, places new emphasis on ES. Its recommendations for market risk management are illustrated in the 2019 document *Minimum Capital Requirements for Market Risk* that says: "ES must be computed on a daily basis for the bank-wide internal models to determine market risk capital requirements. ES must also be computed on a daily basis for each trading desk that uses the internal models approach (IMA).”; “In calculating ES, a bank must use a 97.5th percentile, one-tailed confidence level” (Basel Committee on Banking Supervision (2019), p. 89). Therefore, in the empirical application of our paper, we focus on one-step-ahead tail risk forecasting at the $\alpha = 2.5\%$ quantile level. In order to simplify notation, in the remainder, unless differently specified, the following notational conventions are adopted: $\mathrm{ES}_{t,\alpha} \equiv ES_t$ and $Q_{t,\alpha} \equiv Q_t$, where α denotes the target 2.5% level for the estimation of VaR and ES.

Forecasts of VaR and ES can be generated through a variety of different models. Some of these, such as completely specified GARCH models, are fully parametric since they rely on the exact specification of the conditional distribution of returns and of the volatility dynamics. Differently, semi-parametric approaches require specific assumptions on the risk dynamics but without a return distribution assumption. Semi-parametric models can be applied to generate forecasts of VaR alone, that is the case of the conditional autoregressive VaR (CAViaR) models proposed by Engle and Manganelli (2004), or joint forecasts of the pair (VaR, ES). A joint semi-parametric model that directly estimates both VaR and ES, referred to here as the ES-CAViaR model, is proposed by Taylor (2019). Through incorporating an Asymmetric Laplace (AL) distribution with a time-varying scale, a quasi-likelihood can be built to enable the joint estimation of the conditional VaR and conditional ES in this framework.

Fissler and Ziegel (2016) develop a family of joint loss functions (or “scoring rules”) for the associated VaR and ES series that are strictly consistent for the pair (VaR, ES), that is, they are uniquely minimized by the true VaR and ES series. Applying specific choices of functions in the class of joint loss functions of Fissler and Ziegel (2016), it can be shown that such a loss function is exactly the same as the negative of the AL log-likelihood function presented in Taylor (2019). Patton et al. (2019) propose new dynamic models for VaR and ES, through adopting the generalized autoregressive score (GAS) framework (Creal et al. 2013 and Harvey 2013) and utilizing the loss functions in Fissler and Ziegel (2016).

Alternatively, a variety of semi-parametric approaches to the prediction of VaR and

ES can be obtained by combining Quasi Maximum Likelihood (QML) estimation of the volatility coefficients with some non-parametric estimator of the error quantiles. Widely diffused and effective solutions rely on extreme value theory results, such as in the peaks-over-threshold approach (Gilli et al., 2006).

Storti and Wang (2021) have recently proposed a new ES estimation and forecasting framework, referred to as the Weighted Quantile (WQ) approach, where the ES is modelled as weighted average of tail quantiles. The quantiles are produced from the CAViaR model of Engle and Manganelli (2004) by grid search of a range of equally spaced quantile levels below the target VaR level, i.e., 2.5%. An advantage of this approach is that it sensibly reduces the impact of model uncertainty in the prediction of ES, that is modelled according to its natural definition as an average of tail quantiles. However, the specification of the optimal dynamic model for each quantile level is still subject to uncertainty. In order to limit the impact of the overall model uncertainty on the generation of joint (VaR, ES) forecasts, the WQ framework could actually be extended by replacing forecasts of tail quantiles from a single CAViaR model, as in Storti and Wang (2021), with forecasts combinations from an ensemble of different models of a possibly heterogeneous nature that includes parametric as well as semi-parametric models. Then the ES forecasts could be generated as weighted averages of "combined" VaR predictors at different levels, employing the WQ framework as in Storti and Wang (2021).

The main motivation for this extension of the WQ framework relies on the consideration that, given the values of conditional tail quantiles, the ES is theoretically defined as the expectation of these quantiles. So, it can be immediately recognized that most of the modelling uncertainty is related to the modelling of tail quantiles, which will be addressed in this paper. The only residual uncertainty affecting ES estimation is potentially related to the identification of the grid of tail quantiles. However, as extensively discussed in Storti and Wang (2021), the ES estimates obtained through the WQ approach are not particularly sensitive to the selection of these hyper-parameters.

Aim of this paper is then to propose a novel approach to forecast VaR and ES based on Forecasting Combination and Weighted Quantile (FC-WQ) techniques. The proposed framework can be treated as a generalization of the Weighted Quantile framework proposed by Storti and Wang (2021), and its effectiveness is investigated through applications to stock market data.

The paper is structured as follows. Section 2 introduces the notation used in the paper. A review of strictly consistent scoring functions for joint estimation of VaR and ES is conducted in Section 3. Section 4 presents the the proposed approach and discusses its technical implementation details. The selected model universe for forecast combination is shown in Section 5. The results of an empirical application to real stock market data are presented and discussed in Section 6. Finally, Section 7 concludes.

2 Notation

In this section, to facilitate the understanding of technical details of the proposed framework, we report a summary of the mathematical notation for the main quantities/variables used in the paper.

- r_t: day t to day $t-1$ close-to-close log-returns.
- $\mathcal{I}_t$: information set available at time t.
- $F_t(r)$: CDF of returns conditional on $\mathcal{I}_{t-1}$.
- M: number of quantile levels α_j ($j = 1, \ldots, M$) included in the grid used for WQ estimation. α is used to denote the target 2.5% quantile level for VaR and ES.
- n_{mod}: number of individual models included in the model universe for forecasting combination.
- T: length of the full-sample returns series. N is the in-sample size and H is the out-of-sample size, so $T = N + H$.
- $\widehat{\mathbf{Q}}^{(\alpha_j)}_{1:N,i}$: $(N \times 1)$ time series of in-sample quantile estimates with in-sample data from $t = 1$ to N, generated by model i on quantile level α_j, where $i = 1, \ldots, n_{mod}$ and $j = 1, \ldots, M$.

- $\widehat{\mathbf{Q}}^{(\alpha_j)}_{N+1:N+H,i}$: $(H \times 1)$ time series of out-of-sample one-step-ahead quantile forecasts generated by model i with out-of-sample data from $t = N+1$ to $N+H$ on quantile level α_j, where $i = 1, \ldots, n_{mod}$, $j = 1, \ldots, M$.

- $\widehat{Q}^{(\alpha_j)}_{N+h,i}$: (1×1) h-th one-step-ahead quantile forecast generated by model i on quantile level α_j, where $i = 1, \ldots, n_{mod}$, $j = 1, \ldots, M$, $h = 1, \ldots, H$.

- $\widehat{Q}^{(\alpha_j)}_{t,i}$: (1×1) quantile estimate/forecast at time t generated by model i on level α_j, $t = 1, \ldots, N+H$, $i = 1, \ldots, n_{mod}$, $j = 1, \ldots, M$.

- $\widehat{\mathbf{Q}}^{(U)}_{1:(N+H)}$: $(T \times (M \times n_{mod}))$, *quantile universe*, matrix of quantile estimates/forecasts generated by all n_{mod} models included in the model universe, from $t = 1$ to $N+H$ on all M quantile levels.

- $\widehat{\mathbf{Q}}^{(U,\alpha_j)}_{1:N+H}$: $(T \times n_{mod})$ matrix of quantile estimates/forecasts generated by all n_{mod} models included in the model universe, from $t = 1$ to $N+H$ on quantile level α_j, where $j = 1, \ldots, M$.

- $\widehat{\mathbf{Q}}^{(C,\alpha_j)}_{1:N}$: $(N \times 1)$ time series of *combined* in-sample quantile estimates generated with in-sample data from $t = 1$ to N on quantile level α_j, where $j = 1, \ldots, M$.

- $\widehat{\mathbf{Q}}^{(C,\alpha_j)}_{N+1:N+H}$: $(H \times 1)$ time series of *combined* out-of-sample one-step-ahead quantile forecasts generated with out-of-sample data from $t = N+1$ to $N+H$ on quantile level α_j, where $j = 1, \ldots, M$.

- $\widehat{Q}^{(C,\alpha_j)}_{N+h}$: (1×1) h-th one-step-ahead *combined* quantile forecast on quantile level α_j. $\widehat{Q}^{(C,\alpha_M)}_{N+h} = \widehat{Q}^{(C,\alpha)}_{N+h} = \widehat{Q}^{(C)}_{N+h}$ is the target $\alpha = 2.5\%$ quantile level VaR forecast.

- $\widehat{Q}^{(C,\alpha_j)}_{t}$: (1×1) *combined* quantile estimate/forecast at time t on quantile level α_j, $t = 1, \ldots, N+H$, $j = 1, \ldots, M$.

- $\widehat{ES}^{(\text{FC-WQ})}_{N+h}$: (1×1) h-th one-step-ahead ES forecast at the target $\alpha = 2.5\%$ quantile level employing the FC-WQ approach, $h = 1, \ldots, H$.

3 Strictly consistent scoring functions for joint estimation of VaR and ES

From the main definition of ES provided in Equation (1), it follows that that ES_t is related to $F_t(.)$ by the following integral

$$ES_t = \frac{1}{F_t(Q_t)} \int_{-\infty}^{Q_t} r dF_t(r) = \frac{1}{\alpha} \int_{-\infty}^{Q_t} r dF_t(r), \tag{2}$$

that, after a simple change of variable, can be rewritten as

$$ES_t = \frac{1}{\alpha} \int_0^{\alpha} Q_{t,p} dp. \tag{3}$$

In the literature, several alternative parameterizations of ES_t and VaR_t have been proposed. The involved parameters can be consistently estimated from real data by minimizing appropriately chosen strictly consistent scoring functions. If the interest is solely in the estimation of VaR, the unknown coefficients in the dynamic specification of Q_t can be estimated by quantile regression mimizing the expected quantile loss function

$$QL(r_t, Q_t; \alpha) = (\alpha - I_t)(r_t - Q_t)$$

with $I_t = I(r_t < Q_t)$, where $I(A)$ is the indicator function taking value 1 if event A occurs and 0 otherwise, for $t = 1, \ldots, N$.

Koenker and Machado (1999) show that the quantile regression estimator is equivalent to a maximum likelihood estimator when assuming that the data are conditionally distributed as an Asymmetric Laplace (AL) with a mode at the quantile of interest. If r_t is the return on day t and $Pr(r_t < Q_t|\mathcal{I}_{t-1}) = \alpha$, then the parameters in the model for Q_t can be estimated by maximizing a quasi-likelihood based on:

$$p(r_t|\mathcal{I}_{t-1}) = \frac{\alpha(1-\alpha)}{\sigma} \exp\left(\frac{-(r_t - Q_t)(\alpha - I(r_t < Q_t))}{\sigma} \right),$$

for $t = 1, \ldots, N$ and where σ is a scale parameter.

Taylor (2019) extends this result to incorporate the associated ES quantity into the likelihood expression, noting a link between ES_t and a dynamic σ_t, resulting in the conditional density function:

$$p(r_t|\mathcal{I}_{t-1}) = \frac{(\alpha - 1)}{ES_t} \exp\left(\frac{(r_t - Q_t)(\alpha - I(r_t < Q_t))}{\alpha ES_t} \right). \tag{4}$$

This allows a likelihood function to be built and maximised, given model expressions for (Q_t, ES_t). In Equation (4), r_t is the daily return, Q_t and ES_t denote the α target level VaR and ES on day t. Taylor (2019) notes that the negative logarithm of the resulting likelihood function is strictly consistent for (Q_t, ES_t) considered jointly, i.e., it fits into the class of jointly consistent scoring functions for VaR and ES developed by Fissler and Ziegel (2016).

Members of this family are strictly consistent for (Q_t, ES_t), i.e., their expectations are uniquely minimized by the true VaR and ES series. The general form of this functional family is:

$$\begin{aligned} S_t(r_t, Q_t, ES_t) &= (I_t - \alpha)G_1(Q_t) - I_t G_1(r_t) + G_2(ES_t)\left(ES_t - Q_t + \frac{I_t}{\alpha}(Q_t - r_t)\right) \\ &- H(ES_t) + a(r_t)\,, \end{aligned} \tag{5}$$

where $G_1(.)$ is increasing, $G_2(.)$ is strictly increasing and strictly convex, $G_2 = H^{'}$ and $\lim_{x\to-\infty} G_2(x) = 0$ and $a(\cdot)$ is a real-valued integrable function.

As discussed in Taylor (2019), assuming r_t to have zero mean, making the choices: $G_1(x) = 0$, $G_2(x) = -1/x$, $H(x) = -\log(-x)$ and $a = 1 - \log(1-\alpha)$, which satisfy the required criteria, returns the scoring function:

$$S_t(r_t, Q_t, ES_t) = -\log\left(\frac{\alpha - 1}{\text{ES}_t}\right) - \frac{(r_t - Q_t)(\alpha - I(r_t < Q_t))}{\alpha \text{ES}_t}. \tag{6}$$

Taylor (2019) refers to Equation (6) as the AL log score which is a strictly consistent scoring function whose expectation is jointly minimized by the true VaR and ES series. The negative of Equation (6) equals to the log of Equation (4) and can be treated as the AL log-likelihood.

4 Proposed framework

4.1 The weighted quantile framework

In the Weighted Quantile (WQ) framework, α level ES forecasts from a given model are generated in two steps. In step 1, given a grid of quantile levels $\alpha_j \leq \alpha$, $j = 1, \ldots, M$, with

$0 < \alpha_1 < \alpha_2 < \ldots < \alpha_M = \alpha$, an ensemble of M VaR forecasts ($\widehat{Q}_t^{(\alpha_j)}$) is produced. In their empirical application Storti and Wang (2021) use CAViaR models for the modelling of tail quantiles. However, it is worth remarking that any model, parametric or semi-parametric, could be used to generate VaR forecasts for any $\alpha_j \leq \alpha$. In principle, even different models could be used to fit quantiles at different levels. This feature makes the WQ framework highly flexible and adaptive.

In step 2, ES forecasts are then generated as an affine function of the tail quantile forecasts at levels $\alpha_j \leq \alpha$ as in Equation (7):

$$\widehat{ES}_t^{(\mathrm{WQ})} = w_0 + \sum_{j=1}^{M} w_j \widehat{Q}_t^{(\alpha_j)}, \tag{7}$$

where the weights w_j, $j = 1, \ldots, M$, are generated by some flexible and parsimonious function, such as the Beta function. Namely, for $j = 1, \ldots, M$, we have $w_j = w\left(\frac{j}{M}; a, b\right)$ with

$$w(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}. \tag{8}$$

The main reasons for adopting the Beta specification to model the weights behaviour in Equation (7) are its parsimony, since it only depends on two parameters, and flexibility. However, as discussed in Storti and Wang (2021), other parameterizations are feasible and, in particular, for sufficiently low values of M, the weights w_j can be easily estimated individually as "free" parameters.

The intercept w_0 is estimated along with the other parameters and allows to correct biases potentially arising from the left truncation in the chosen grid of tail quantiles, thus further increasing the flexibility of the WQ approach.

Given first stage VaR forecasts, the unknown coefficients in Equation (7) are estimated by optimizing some strictly consistent scoring function such as the AL (Taylor, 2019) or some other strictly consistent scoring function in the Fissler-Ziegel class.

4.2 A unified framework encompassing Forecast Combination and Weighted Quantile estimation

It is important to note that model uncertainty mainly affects step 1 of the WQ procedure, since the estimation formula used in step 2 naturally stems from the mathematical definition of ES as a function of the tail VaRs. In step 1 of the procedure, different models could result in being optimal for different quantile levels. On the other hand, in step 2, the ES is estimated applying its natural definition as expectation of the tail quantiles. The only residual source of modelling uncertainty is related to the selection of the grid of tail quantiles used for estimation in Equation (7) and, in particular, to the choice of the lower bound α_1 and of the value of M. The impact of the choice of the lower bound is, by construction, controlled by the intercept w_0 and by the data driven weighting structure of the ES estimator, as also documented empirically by Storti and Wang (2021). Regarding the choice of M, given the high computing power routinely available even on standard personal computers, implementing the WQ approach with a high value of M, that virtually eliminates the discretization error, is not an issue and could be easily done while still keeping the computing time at reasonable levels. On the other hand, Storti and Wang (2021) show, by simulations and applications to real data, that the WQ gives remarkably good performances even for values of M as low as 3 and that negligible improvements, in terms of forecasting accuracy, are expected from an increase in M.

These considerations motivate a two-step approach based on a Forecast Combination and Weighted Quantile (FC-WQ) strategy, that is the main contribution of this paper. The technical details on the implementation of the two steps of the FC-WQ procedure are presented as below.

Step 1: Assume that n_{mod} different VaR forecasting methods are available and that each of them is fitted to generate series of VaR forecasts for a set of strictly increasing quantile orders $\boldsymbol{\alpha_M} = [0 < \alpha_1, \ldots, \alpha_M = \alpha]$. To be consistent with the standard regulatory prescription, we will focus on the case $\alpha = 2.5\%$.

Letting N be the length of the in-sample window and $T >> N$ the length of the full-sample returns series, for each available model in our *model universe*, M VaR forecasts series will be generated from each model

$$\widehat{Q}^{(\alpha_j)}_{N+1,i}, \ldots, \widehat{Q}^{(\alpha_j)}_{N+H,i} \tag{9}$$

for $i = 1, \ldots, n_{mod}$, $j = 1, \ldots, M$ and where $H = T - N$ is the length of the out-of-sample period to be used for forecast evaluation. Overall, this will yield a total of $M \times n_{mod}$ series of VaR forecasts. For the i-th model and j-th quantile level, the generic h-th one-step-ahead forecast $\widehat{Q}^{(\alpha_j)}_{N+h,i}$ $(h = 1, \ldots, H)$ will be based on the model fitted to observations from h to $N + h - 1$. Forecast combination is then used as a technique for reducing model uncertainty in VaR forecasting, yielding the *combined quantile predictor* that takes the general form

$$\widehat{Q}^{(C,\alpha_j)}_t = c_{0,j} + \sum_{i=1}^{n_{mod}} c_{i,j} \widehat{Q}^{(\alpha_j)}_{t,i}. \tag{10}$$

Therefore, in step 1 VaR forecasts from different models are combined to generate a set of "combined" VaR predictors at different α_j levels. In order to overcome the potential quantile crossing problem, the monotonization method proposed by Chernozhukov et al. (2010) is employed.

The framework in (10) can be in principle extended to consider non-linear combination schemes. This possibility is however not investigated in this paper. The proposed framework is highly flexible and can employ any models that could produce VaR estimates and forecasts, such as GARCH (Bollerslev, 1986), CAViaR, etc. Details of the model universe will be presented in Section 5.

Step 2: In step 2, employing the WQ approach, the conditional ES at time t is modelled as the weighted average of the combined quantiles from step 1:

$$\widehat{ES}^{(\text{FC-WQ})}_t = w_0 + \sum_{j=1}^{M} w_j \widehat{Q}^{(C,\alpha_j)}_t, \tag{11}$$

where the weights w_i, $i = 1, \ldots, M$, are generated by a Beta weight function as in Equation (8)[1].

[1]Following the implementation of WQ in (Storti and Wang, 2021), we set the number of grid points equal to $M + 1$, so that the weight of the α_M-quantile is not 0 by construction when using the Beta weight function to parameterize the weights pattern.

The combination of joint forecasts of VaR and ES has so far received scarce attention in the literature. The only contribution in this field is, to the extent of our knowledge, given by the paper of Taylor (2020) whose approach is however structurally different from the one that is taken in this paper. First, Taylor (2020) optimizes an AL loss function to combine joint (VaR,ES) models at a given target level α, while our approach uses richer information on quantile levels falling in the tail of the distribution below α. Second, our model universe is composed of VaR models rather than of joint (VaR, ES) models.

The next section focuses on the estimation strategy followed to estimate the c_i, for $i = 1, ..., n_{mod}$, and the intercept term c_0 in step 1, and the (w_0,a,b) coefficients in step 2.

4.3 Implementation of the FC-WQ predictor: estimation procedure

Next, we provide a detailed description of the estimation procedures implemented in the two steps of the proposed FC-WQ framework. Although, for ease of explanation, we focus on the standard risk level $\alpha = 2.5\%$, the method can be immediately extended to other values of α.

Estimation Step 1:

Under step 1, given the target quantile level $\alpha = 2.5\%$, an equally spaced grid of quantile levels of size M is selected,

$$\boldsymbol{\alpha_M} = [\alpha_1, \alpha_2, \ldots, \alpha_M],$$

where $\alpha_j = \alpha_{j-1} + \eta$, with $\alpha_M = \alpha$ and $\eta = (\alpha_M - \alpha_1)/(M-1)$, for $j = 2, \ldots, M$. The value of the lower bound α_1 is selected as 0.005 and M is selected as 3 and 5 in this paper, according to the findings in Storti and Wang (2021). For example, with $M = 3$ and $\alpha = 2.5\%$, fixing $\alpha_1 = 0.005$ we have $\eta = 0.01$ and the grid of quantile levels as $\boldsymbol{\alpha_M} = [0.005, 0.015, 0.025]$.

Now, for each trial quantile level $\alpha_j \in \boldsymbol{\alpha_M}$, n_{mod} individual models are employed to produce the $(N \times 1)$ time series of conditional in-sample quantiles $\widehat{\mathbf{Q}}^{(\alpha_j)}_{1:N,i}$ and the 1st one-step-ahead quantile forecasts $\widehat{Q}^{(\alpha_j)}_{N+1,i}$, for $i = 1, \ldots, n_{mod}$ and $j = 1, \ldots, M$. The set of in-sample quantiles for n_{mod} models and at all trial quantile levels, from α_1 to α_M, is collected in the $N \times (M \times n_{mod})$ array $\widehat{\mathbf{Q}}_{1:N}$.

In our paper, a rolling window forecasting scheme is adopted, with T as the length of the full-sample returns series. Therefore, for each step in the $H = T - N$ out-of-sample periods (still for each trial quantile level $\alpha_j \in \boldsymbol{\alpha_M}$), n_{mod} individual models are employed to produce the time series of one-step-ahead quantile forecasts $\widehat{Q}^{(\alpha_j)}_{N+1,i}, \ldots, \widehat{Q}^{(\alpha_j)}_{N+H,i}$, for $i = 1, \ldots, n_{mod}$ and $j = 1, \ldots, M$. This set of out-of-sample quantiles for n_{mod} models at all trial quantile levels, from α_1 to α_M, is collected in the $H \times (M \times n_{mod})$ array $\widehat{\mathbf{Q}}_{(N+1):(N+H)}$, here $N + H = T$.

Concatenating the two matrices $\widehat{\mathbf{Q}}_{1:N}$ (including in-sample quantile estimates) and $\widehat{\mathbf{Q}}_{(N+1):(N+H)}$ (including one-step-ahead out-of-sample quantile forecasts) produces our *quantile universe* $\widehat{\mathbf{Q}}^{(U)}_{1:(N+H)}$ which is a matrix of size $T \times (M \times n_{mod})$. The component of $\widehat{\mathbf{Q}}^{(U)}_{1:(N+H)}$ including in-sample VaR estimates and out-of-sample VaR forecasts at level α_j will be denoted as $\widehat{\mathbf{Q}}^{(U,\alpha_j)}_{1:(N+H)}$, that is a $T \times n_{mod}$ matrix.

The $\widehat{\mathbf{Q}}^{(U,\alpha_j)}_{1:(N+H)}$ matrix and the series of returns $\mathbf{r}_{1:(N+H)}$ are then given as input to the estimation procedure for VaR "combination" weights. Namely, as in Giacomini and Komunjer (2005), for a given VaR level α_j and a forecast origin $t \geq N$, the coefficients $c_{i,j,t}$ used for combining VaR forecasts at time t, for $i = 0, \ldots, n_{mod}$ (to include the intercept term c_0) can be estimated by minimizing the quantile loss function over a rolling window of fixed size N

$$\begin{aligned} \overline{QL}_{t,N}(\alpha_j, \mathbf{c}_{j,t}) &= \frac{1}{N} \sum_{k=1}^{N} QL(r_{t-k}, \widehat{\mathbf{Q}}^{(U,\alpha_j)}_{t-k}; \alpha_j) \\ &= \frac{1}{N} \sum_{k=1}^{N} (\alpha_j - I_{t-k,j}) \left(r_{t-k} - \widehat{\mathbf{X}}^{(U,\alpha_j)}_{t-k} \mathbf{c}_{j,t} \right), \end{aligned} \tag{12}$$

where $\mathbf{c}_{j,t} = (c_{0,j,t}, c_{1,j,t}, \ldots, c_{n_{mod},j,t})'$, $\widehat{\mathbf{Q}}^{(U,\alpha_j)}_{t} \equiv \widehat{\mathbf{Q}}^{(U,\alpha_j)}_{t:t}$ (a vector of size $1 \times n_{mod}$, from row t in matrix $\widehat{\mathbf{Q}}^{(U,\alpha_j)}_{1:(N+H)}$), $\widehat{\mathbf{X}}^{(U,\alpha_j)}_{t} = \left[1 \ \widehat{\mathbf{Q}}^{(U,\alpha_j)}_{t}\right]$ (a vector of size $1 \times (n_{mod} + 1)$), $I_{t,j} = I(r_t < \widehat{\mathbf{X}}^{(U,\alpha_j)}_{t} \mathbf{c}_{j,t})$, for each quantile level α_j, $j = 1, \ldots, M$. Analytically, the estimated coefficients for combining one-step-ahead VaR forecasts with origin at time $N + h$ are given by

$$\widehat{\mathbf{c}}_{j,N+h} = \underset{\mathbf{c}_{j,N+h}}{\arg\min}\ \overline{QL}_{N+h,N}(\alpha_j, \mathbf{c}_{j,N+h}). \tag{13}$$

Combined VaR forecasts are then finally computed through substituting the fitted $\widehat{\mathbf{c}}_{j,N+h}$ coefficients in Equation (10)

$$\widehat{Q}^{(C,\alpha_j)}_{N+h} = \widehat{\mathbf{X}}^{(U,\alpha_j)}_{N+h}\widehat{\mathbf{c}}_{j,N+h}, \qquad \text{for } h = 1, \ldots, H. \tag{14}$$

Step 1 of the FC-WQ generates an output matrix of combined quantile predictors $\widehat{\mathbf{Q}}^{(C)}_{1:T}$ of size $T \times M$. Each column in this matrix is produced from combining conditional quantile forecasts from n_{mod} models. Rows from 1 to N are computed by estimating the combination weights on the in-sample quantile estimates from the n_{mod} candidate models and then using these weights to combine the time series of in-sample quantile estimates. The same weights, based on in-sample VaR estimates from time 1 to N, are used to generate row $N+1$ of the matrix. Row $N+2$ is then generated using weights estimated on the $N \times n_{mod}$ time series composed of in-sample quantile estimates from time 2 to N (for the first $N-1$ rows), and by the formerly generated vector of quantile forecasts at time $N+1$ (for the last row). This procedure is iterated, in a rolling window fashion, until the end of the available sample T. A detailed step by step implementation description is shown in the Algorithm 1 presented after the estimation step 2.

The values in column M and rows from $N+1$ to $N+H$ in the matrix $\widehat{\mathbf{Q}}^{(C)}_{1:T}$ contain the 2.5% target level one-step-ahead quantile forecasts $\widehat{\mathbf{Q}}^{(C,\alpha_M)}_{(N+1):(N+H)}$, noting $\alpha_M = \alpha = 2.5\%$.

Estimation step 2:

In the second stage of our approach, we predict the conditional ES at time $N+h$ as an affine function of the elements of $\widehat{\mathbf{Q}}^{(C)}_{N+h}$ (row $N+h$ in matrix $\widehat{\mathbf{Q}}^{(C)}_{1:T}$). The only unknown parameters in Equation (11) for calculating $\text{ES}^{(\text{FC-WQ})}$ estimator are (w_0, a, b). Conditioning on first stage produced VaR series $\widehat{\mathbf{Q}}^{(C)}_{t}$ (a vector of size $1 \times M$, row t in matrix $\widehat{\mathbf{Q}}^{(C)}_{1:T}$) and letting $\boldsymbol{\theta_t} = (w_{0,t}, a_t, b_t)'$, the values of the coefficients used for generating the ES forecast at time t can be estimated by minimizing wrt $\boldsymbol{\theta}_t$ the strictly consistent scoring function:

$$\bar{S}_{t,N}\left(\alpha, \boldsymbol{\theta}_t|\mathbf{c}_t\right) = \frac{1}{N}\sum_{k=1}^{N} S_{t-k}\left(\alpha, \boldsymbol{\theta}_t|\mathbf{c}_t\right),$$

where

$$S_t\left(\alpha, \boldsymbol{\theta}_t|\mathbf{c}_t\right) = -\log\left(\frac{\alpha-1}{\widehat{ES}_t}\right) - \frac{\left(r_t - \widehat{Q}_t^{(C,\alpha)}\right)\left(\alpha - I\left(r_t \leq \widehat{Q}_t^{(C,\alpha)}\right)\right)}{\alpha\,\widehat{ES}_t},$$

and

$$\widehat{ES}_t = \widehat{\mathbf{X}}_t^{*}\mathbf{w}_t, \tag{15}$$

with $\widehat{\mathbf{X}}_t^{*} = [1\ \widehat{\mathbf{Q}}_t^{(C)}]$ (a vector of size $1 \times (M+1)$) and $\mathbf{w}_t = (w_{0,t}, w_{1,t}, \ldots, w_{M,t})'$. $\widehat{Q}_t^{(C,\alpha)} \equiv \widehat{Q}_t^{(C,\alpha_M)}$ is the quantile input at the target $\alpha = \alpha_M = 2.5\%$ quantile level and produced from the estimation step 1 (row t column M in matrix $\widehat{\mathbf{Q}}_{1:T}^{(C)}$).

The estimated coefficients for combining M one-step-ahead VaR forecasts for time $N+h$ are obtained as

$$\widehat{\boldsymbol{\theta}}_{N+h} = \underset{\boldsymbol{\theta}_{N+h}}{\arg\min}\ \bar{S}_{N+h,N}\left(\alpha, \boldsymbol{\theta}_{N+h}|\mathbf{c}_{N+h}\right), \tag{16}$$

where $\widehat{\boldsymbol{\theta}}_{N+h} = (\widehat{w}_{0,N+h}, \widehat{a}_{N+h}, \widehat{b}_{N+h})$. Minimization of the above AL log-score was implemented using the Quasi-Newton optimizer implemented in the Matlab *fminunc* function.

The estimated parameters a_{N+H} and b_{N+H} are plugged in the Beta lag function (8) to produce $\widehat{w}_{1,N+h}, \ldots, \widehat{w}_{M,N+h}$. Therefore, the fitted coefficients

$$\widehat{\mathbf{w}}_{N+h} = (\widehat{w}_{0,N+h}, \widehat{w}_{1,N+h}, \ldots, \widehat{w}_{M,N+h})' \tag{17}$$

are then employed in the weighted quantile framework as below to produce ES forecast

$$\widehat{ES}_{N+h}^{(\text{FC-WQ})} = \widehat{\mathbf{X}}_{N+h}^{*}\widehat{\mathbf{w}}_{N+h}, \qquad \text{for } h = 1, \ldots, H. \tag{18}$$

For the sake of clarity, the detailed step by step description of the outlined forecasting algorithm is presented in Algorithm 1.

Algorithm 1 Forecast Combination and Weighted Quantile Algorithm

Input: In-sample quantile estimates $\widehat{\mathbf{Q}}^{(\alpha_j)}_{1:N,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M)$; one-step-ahead quantile forecasts $\widehat{\mathbf{Q}}^{(\alpha_j)}_{N+1:N+h,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M,\ h = 1, \ldots, H)$.

Output: One-step-ahead combined VaR forecast $\widehat{Q}^{(C,\alpha_j)}_{N+h}$ $(j = 1, \ldots, M)$; $\alpha = 2.5\%$ ES one-step-ahead forecast $\widehat{ES}^{(\text{FC-WQ})}_{N+h}$, $h = 1, \ldots, H$.

1: **for** $h = 1, \ldots, H$ **do**

2: Employ the in-sample quantile estimates $\widehat{\mathbf{Q}}^{(\alpha_j)}_{h:N,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M)$, one-step-ahead quantile forecasts $\widehat{\mathbf{Q}}^{(\alpha_j)}_{N+1:N+h-1,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M)$ and Equation (13) to estimate the quantile combination weights $\widehat{\mathbf{c}}_{j,N+h}$ {In iteration 1, only in-sample quantile estimates $\widehat{\mathbf{Q}}^{(\alpha_j)}_{1:N,i}$ are used};

3: Incorporate Equation (14), estimated weight $\widehat{\mathbf{c}}_{j,N+h}$, and h-th one-step-ahead quantile forecasts $\widehat{Q}^{(\alpha_j)}_{N+h,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M)$ to produce the combined h-th one-step-ahead quantile (VaR) forecast $\widehat{Q}^{(C,\alpha_j)}_{N+h}$. Here, $\widehat{Q}^{(C,\alpha_M)}_{N+h} = \widehat{Q}^{(C,\alpha)}_{N+h}$ is the target 2.5% quantile level h-th one-step-ahead combined VaR forecast;

4: Use the combined quantile estimators $\widehat{\mathbf{Q}}^{(C)}_{h:N+h-1}$ (calculated from the in-sample quantiles $\widehat{\mathbf{Q}}^{(\alpha_j)}_{h:N,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M)$, one-step-ahead quantile forecasts $\widehat{\mathbf{Q}}^{(\alpha_j)}_{N+1:N+h-1,i}$ $(i = 1, \ldots, n_{mod},\ j = 1, \ldots, M)$ and estimated combination weights $\hat{\mathbf{c}}_{j,N+h}$), to estimate the WQ Beta weights $\widehat{\mathbf{w}}_{t+h}$ using Equation (17); {In our study out-of-sample size H is greater than the in-sample size N, thus when $h > N$ in-sample quantiles are all replaced by one-step-ahead quantile forecasts when calculating $\widehat{\mathbf{Q}}^{(C)}_{h:N+h-1}$};

5: Employ Equation (18), estimated WQ Beta weights $\widehat{\mathbf{w}}_{t+h}$ and combined one-step-ahead quantile forecasts $\widehat{\mathbf{Q}}^{(C)}_{N+h}$, to produce the target 2.5% ES one-step-ahead forecast $\widehat{ES}^{(\text{FC-WQ})}_{N+h}$.

6: **end for**

5 The model universe

Various types of models, such as parametric and semi-parametric VaR models are selected as candidate models in the model universe. All these models, except for the CAViaR (Engle and Manganelli, 2004), can be also used to generate ES forecasts. These ES forecasts will not be of interest for the implementation of the FC-WQ procedure but will be later used as benchmarks for ES forecasts comparison. In total, $n_{mod} = 8$ different models are selected with details shown below.

GJR-GARCH-t (parametric): the GJR-GARCH model, proposed by Glosten et al. (1993), extends the parametric GARCH model to capture the well-known leverage effect (negative returns at time $t-1$ have a larger impact on the volatility at time t than positive returns). In addition, to capture the fat-tail property of financial returns, the Student-t distribution is employed.

EGARCH-t (parametric): as another commonly used parametric model, the EGARCH model (Nelson, 1991) is also included as a candidate model. The EGARCH model does not require any positivity restriction on the parameters, since its volatility equation is on log-variance instead of variance itself. Thus the positivity of the variance is automatically satisfied, which is an important advantage of the framework. Again, the Student-t distribution is used to describe the potential leptokurtosis of the conditional return distribution.

POT-GJR-GARCH-t (semi-parametric): we also consider some semi-parametric models as part of the model universe. The peaks-over-threshold (POT)-GJR-GARCH combines the GJR-GARCH model with an extreme value theory (McNeil and Frey, 2000) approach to the fitting of the tail properties of the error distribution. Namely, this is accomplished by applying the POT approach to the returns standardized by the GJR-GARCH-t estimated volatility; see Gilli et al. (2006) for details.

POT-EGARCH-t (semi-parametric): this approach is similar to the above described POT-GJR-GARCH-t with the difference that the POT method is applied to returns standardized by the EGARCH-t estimated volatility.

GJR-GARCH-t-HS (semi-parametric): a semi-parametric filtered historical simulation approach, which could potentially produce improved tail risk forecasting results, is also included. The series of in-sample conditional variance ($\hat{\sigma}_t$) is estimated based on the fitted GJR-GARCH-t model. The error quantiles and tail expectations are then estimated by computing the relevant sample quantiles ($\hat{q}^{(\alpha)}$) and tail averages ($\hat{c}^{(\alpha)}$) of standardized returns $r_t/\hat{\sigma}_t$. Finally, level-α VaR and ES forecasts are obtained by multiplying $\hat{q}^{(\alpha)}$ and $\hat{c}^{(\alpha)}$, respectively, by the forecast $\hat{\sigma}_{N+1}$ from the fitted GJR-GARCH model.

EGARCH-t-HS (semi-parametric): this approach employs a similar procedure as the GJR-GARCH-t-HS, by replacing GJR-GARCH-t with EGARCH-t for volatility estimation and forecasting.

We would like to emphasise that, for the above models which rely on parametric volatility estimates and forecasts, we do not need to re-estimate the model for each trial quantile level α_j; $j = 1, \ldots, M$, regarding the VaR and ES calculation. Differently, the remaining two models rely on direct modelling of VaR dynamics. Therefore, they need to be re-estimated for each trial quantile level α_j.

CAViaR-AS (semi-parametric): the CAViaR models proposed by Engle and Manganelli (2004) are estimated using quantile regression. Although the CAViaR framework does not directly produce ES estimates and forecasts, it can be used without any issues in our FC-WQ approach which only requires the quantile estimates and forecasts from individual models.

The CAViaR with asymmetric slope (CAViaR-AS) framework which aims to capture the leverage effect is employed:

$$Q_t = \beta_0 + \beta_1 Q_{t-1} + (\beta_2 I_{[r_{t-1} \geq 0]} + \beta_3 I_{[r_{t-1} < 0]})|r_{t-1}|. \tag{19}$$

As documented by Engle and Manganelli (2004), solutions to the optimization of the quantile loss objective function can be heavily dependent on the chosen initial values. To account for this issue, we adopt a multi-start optimization procedure inspired by that suggested in Engle and Manganelli (2004).

CARE-AS (semi-parametric): a different approach to joint estimation of VaR and ES is based on the theory of expectiles. The concept of expectile is closely related to the concept of quantile. The τ level expectile μ_τ, as defined by Aigner et al. (1976), can be estimated through minimizing the following Asymmetric Least Squares (ALS) criterion (Newey and Powell, 1987):

$$\sum_{t=1}^{N} |\tau - I(r_t < \mu_\tau)|(r_t - \mu_\tau)^2 \, , \tag{20}$$

no distributional assumption is required to estimate μ_τ here.

Taylor (2008) proposes a class of semi-parametric models for VaR and ES forecasting, called Conditional Autoregressive Expectile (CARE) models, with a similar form to the CAViaR model. Under the CARE framework, the lagged returns drive the expectiles and model parameters are estimated via minimizing an ALS criterion.

To select the appropriate expectile levels for VaR and ES estimation, implementation of CARE type models requires a grid search process between 0 and the target quantile level 2.5%, based on the optimization of the violation rate (VRate, the percentage of returns exceeding VaR estimates). The size of the expectile level grid search is selected as 100 in our paper.

The CARE with asymmetric slope (CARE-AS) specification as below is included in the model universe, where the expectile responds asymmetrically to positive and negative returns:

$$\mu_{t;\tau} = \beta_{0;\tau} + \beta_{1;\tau}\mu_{t-1;\tau} + (\beta_{2;\tau} I_{[r_{t-1}\geq 0]} + \beta_{3;\tau} I_{[r_{t-1}<0]})|r_{t-1}| \, . \tag{21}$$

6 Empirical study

6.1 Data and empirical study design

The daily data, including open, high, low and closing prices, are downloaded from Thomson Reuters Tick History and cover the period from the beginning of 2000 to the end of 2015. The closing price is employed to calculate the daily return r_t. Data are collected for six market indices: S&P500 (US), Hang Seng (Hong Kong), FTSE 100 (UK), DAX (Germany), SMI (Swiss) and ASX200 (Australia).

As described in Section 4.3, a rolling window with fixed in-sample size is employed for estimation and to produce each one-step-ahead forecast in the forecasting period. Table 1 reports the in-sample size for each series, which differs due to different non-trading days occurring in each market.

The forecasting study incorporates a 8 year out-of-sample period, with the start date of the out-of-sample chosen as January 2008 (to include the 2008 GFC as part of the out-of-sample period) and out-of-sample size H as 2000. Therefore, the end of the forecasting period is around the end of 2015, with small differences among different markets due to calendar effects.

Employing the proposed FC-WQ framework, both daily one-step-ahead VaR and ES forecasts are produced for the returns on the six indices. VaR forecasts are produced for the whole range of selected trial quantile values defined in Section 4.3 while, for ES forecasting, $\alpha = 2.5\%$ is chosen as target level, as recommended by Basel Committee on Banking Supervision (2019).

For comparison, VaR and ES forecasts are also generated from each individual model included in the model universe, as presented in Section 5. Since the CAViaR-AS model cannot directly produce the ES forecasts, the ES-CAViaR models of Taylor (2019) are also included in the ES study, again employing the CAViaR-AS models as the specification of the quantile regression component. Two VaR to ES relationships, additive and multiplicative, are employed for the ES-CAViaR framework. We name the models as

ES-CAViaR-Add-AS and ES-CAViaR-Mult-AS, respectively. Then, to assist the optimization in the estimation of ES-CAViaR models, the initial values of the parameters of the ES component are also selected by means of an additional random sampling procedure, following Taylor (2019).

6.2 VaR forecasts evaluation

One-step-ahead forecasts of VaR and ES are generated for each day in the forecasting period for each data series. This section focuses on the evaluation of VaR forecasts. For brevity, we only report results for the 2.5% quantile level. However, we would like to emphasize that the adopted quantile forecast combination approach allows to reduce the impact of model uncertainty, potentially improving the quantile estimation and forecasting accuracy, for each trial quantile level, i.e., $\alpha_1, \alpha_2, \ldots, \alpha_M$. This is expected to positively affect the second step ES estimation and forecast (details to be shown in the following section).

First, the VaR violation rate (VRate) is employed to initially assess VaR forecasting accuracy. VRate is simply the proportion of returns that exceed the forecasted VaR in the forecasting period, as in Equation (22)

$$\text{VRate} = \frac{1}{H} \sum_{t=N+1}^{N+H} I(r_t < \text{VaR}_t)\,, \tag{22}$$

where N is the in-sample size and $H = 2000$ is the out-of-sample size. Models with a VRate closest to the nominal quantile level $\alpha = 2.5\%$ are preferred, or equivalently $\frac{\text{VRate}}{\alpha}$ closest to 1.

Table 1 summarizes the $\frac{\text{VRate}}{\alpha}$ (the closer to 1 the better) at the 2.5% quantiles over the six indices for all competing models. The "MAD" column shows the Mean Absolute Deviation, employing 2.5% as the target VRate, across the six indices. The "Avg Rank" column is the average of the ranks, across indices, based on the absolute value of the deviation of each VRate from the 2.5% target level. Box indicates the best model, while dashed box indicates the 2nd best model.

Overall, the proposed FC-WQ frameworks produce favourable VRate results, compared with the other 8 competing individual models in the model universe. The "MAD" value from FC-WQ is 0.0028 which is the smallest. The EGARCH-t-HS is ranked the best, followed by FC-WQ, GJR-GARCH-t-HS and CARE-AS.

Here, we would like to mention that only the first step (quantile forecast combination) in the proposed FC-WQ framework would affect the VaR forecasting performance. The second step (weighted combined quantile of each trial quantile level) will determine the ES forecasting performance which will be presented in the following section.

Table 1: $\frac{\text{VRate}}{\alpha}$ across the six markets.

Model	S&P500	HangSeng	FTSE	DAX	SMI	ASX200	Avg Rank	MAD
GJR-GARCH-t	1.68	1.26	1.50	1.68	1.36	1.58	8.67	0.0128
EGARCH-t	1.62	1.24	1.50	1.64	1.48	1.50	8.17	0.0124
POT-GJR-GARCH-t	1.30	1.08	1.06	1.26	1.12	1.14	4.67	0.0040
POT-EGARCH-t	1.26	1.10	1.08	1.18	1.18	1.10	3.83	0.0038
GJR-GARCH-t-HS	1.30	1.06	1.04	1.26	1.10	1.10	3.33	0.0036
EGARCH-t-HS	1.26	1.10	1.06	1.16	1.16	1.10	3.17	0.0035
CAViaR-AS	1.18	1.00	1.08	1.28	1.28	1.10	4.00	0.0038
CARE-AS	1.08	1.00	1.16	1.24	1.24	0.98	3.33	0.0031
FC-WQ	1.04	0.98	1.00	1.18	1.32	0.90	3.33	0.0028
Out-of-sample H	2000	2000	2000	2000	2000	2000		
In-sample N	1905	1890	1943	1936	1930	1871		

Note: Box indicates the favoured model and dashed box indicates the 2nd ranked model based on the average rank and MAD.

The average value of the quantile loss over the out-of-sample period is then used to compare the VaR forecast accuracy of competing models. This choice is motivated considering that the standard quantile loss function is strictly consistent, i.e., the expected loss is a minimum at the true quantile series. The quantile loss function is the one that is employed to optimize the quantile forecast combination weights, as described in Section 4.3. Thus, the most accurate VaR forecasting model is expected produce the minimized aggregated quantile loss function, given as in Equation (23):

$$\sum_{t=N+1}^{N+H} (\alpha - I(r_t < Q_t))(r_t - Q_t) \ , \tag{23}$$

where N is the in-sample size and $H = 2000$ is the out-of-sample size. $\widehat{Q}_{N+1}, \ldots, \widehat{Q}_{N+H}$ is a series of quantile forecasts at level $\alpha = 2.5\%$ for the observations $r_{N+1}, \ldots, r_{N+H}$.

The values of the out-of-sample quantile loss are presented in Table 2. The average loss is included in the "Avg Loss" column. The average rank based on ranks of quantile loss across six markets is calculated and shown in the "Avg Rank" column. Box indicates the favoured model and dashed box indicates the 2nd ranked model based on the average loss and rank.

Based on the quantile loss results, we can see that the proposed FC-WQ framework is characterized by very competitive performances, with the smallest average quantile loss value 164.2. With respect to the average ranking, the EGARCH-t-HS ranks the best, closely followed by the FC-WQ and POT-EGARCH-t approaches. The GJR-GARCH-t and EGARCH-t are in general least preferred, with the average rank as 7.83 and 6.83 respectively. Although GJR-GARCH-t and EGARCH-t are included in the forecasting combination process, the FC-WQ is still capable of producing competitive quantile forecasting results, which lends evidence on the combination weights estimation scheme, as described in Section 4.3.

Table 2: 2.5% quantile loss function values across the markets.

Model	S&P500	HangSeng	FTSE	DAX	SMI	ASX200	Avg Loss	Avg Rank
GJR-GARCH-t	162.9	196.4	156.4	182.9	159.4	141.7	166.6	7.83
EGARCH-t	166.9	194.9	155.2	181.6	159.3	140.3	166.4	6.83
POT-GJR-GARCH-t	161.1	195.0	154.5	180.7	159.3	139.8	165.1	5.50
POT-EGARCH-t	163.8	193.7	153.0	179.5	157.5	138.4	164.3	3.17
GJR-GARCH-t-HS	161.0	194.9	154.5	180.7	159.3	139.8	165.0	4.83
EGARCH-t-HS	163.7	193.6	153.0	179.5	157.5	138.4	164.3	2.50
CAViaR-AS	167.7	190.6	153.1	179.2	159.0	139.9	164.9	3.83
CARE-AS	168.3	189.7	154.7	180.8	163.6	142.4	166.6	6.83
FC-WQ	160.7	191.4	155.1	180.2	159.3	138.5	164.2	3.67

Note:Box indicates the favoured model and dashed box indicates the 2nd ranked model based on the average loss and rank.

In addition, for S&P 500 the quantile loss values for each time step across the whole forecasting period are visualised in Figure 1. Namely, EGARCH-t, EGARCH-t-HS, CAViaR-AS and FC-WQ are compared. The forecast combination is evidently characterized by more "stabilized" quantile loss values. For example, between 2009 and 2012, the quantile loss from the forecast combination approach is consistently smaller than that of the competing models, including the EGARCH-t-HS and CAViaR-AS which have good quantile loss performance as shown in Table 2. Therefore, the proposed forecast combination not only allows obtaining an improved predictor via combining different functional forms, i.e., parametric and semi-parametric models, but also produces more robust and stabilized quantile forecasts through the quantile combination process. Such time stability argument is consistent with the aim of forecasting combinations that is to account for model uncertainty and to provide a forecasting performance that is optimal (or close to being optimal) and stable across time, which is supported by the results in Table 1, 2 and Figure 1.

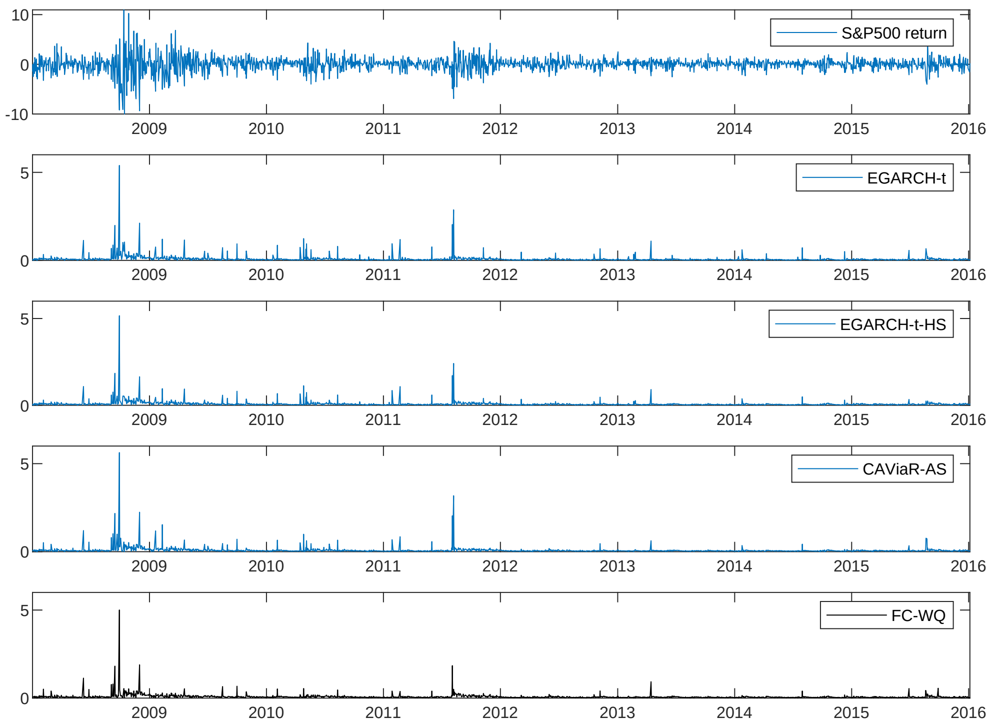

Figure 1: S&P500 out-of-sample quantile loss value from EGARCH-t, EGARCH-t-HS, CAViaR-AS and FC-WQ.

Lastly, to further assess the validity of the quantile forecast combination, we have employed the VaR calibration tests defined in Patton et al. (2019), with the code developed by the authors. The test employs a MZ type regression of generalized VaR residuals on fitted VaR and lagged generalized residuals. Regression coefficients are fitted by Ordinary Least Squares (OLS) and standard errors are computed by a Newey-West estimator with

20 lags. The p-values of each data set and model produced from the test are presented in Table 3. In addition, on the 10% significance level, the column "Total" shows the total number of rejections (p-value less than 10%) for each model. Regarding the results, CAViaR-AS receives the least number of rejections and is closely followed by several models, including the proposed FC-WQ, which are rejected twice. The GJR-GARCH-t and EGARCH-t models are the most frequently rejected by the test.

Table 3: 2.5% VaR calibration test at the 10% significance level.

Model	S&P500	HangSeng	FTSE	DAX	SMI	ASX200	Total
GJR-GARCH-t	0.003	0.494	0.010	0.001	0.025	0.002	5
EGARCH-t	0.056	0.576	0.018	0.006	0.002	0.024	5
POT-GJR-GARCH-t	0.255	0.948	0.750	0.037	0.000	0.000	3
POT-EGARCH-t	0.740	0.936	0.757	0.639	0.059	0.000	2
GJR-GARCH-t-HS	0.255	0.953	0.807	0.037	0.000	0.000	3
EGARCH-t-HS	0.740	0.936	0.838	0.706	0.074	0.000	2
CAViaR-AS	0.816	0.849	0.796	0.501	0.106	0	1
CARE-AS	0.910	0.610	0.734	0.605	0.097	0.000	2
FC-WQ	0.932	0.682	0.873	0.738	0.069	0.000	2

Note: Box indicates the favoured model and dashed box indicates the 2nd ranked model based on the total number of rejections on the 10% significance level. p-values are presented for each index and each model.

6.3 ES forecasts evaluation

We remind that a key feature of the proposed framework is that, for each trial quantile level, the combined VaR predictor can potentially have a different structure in terms of included models and assigned weights. This flexibility can only expected to be beneficial for the ES forecasting performance. Also, no specific assumptions are formulated on relationship linking VaR and ES but, consistently with its theoretical definition, ES is computed as a weighted average of combined VaR forecasts. This design naturally yields a combined ES predictor that has been purged of the impact of model uncertainty. Due to these features, it is expected that, compared to single forecasting models, the FC-WQ framework could be characterized by an improved and more stable ES forecasting performance. Aim of this section is to provide empirical evidence supporting this hypothesis.

In the FC-WQ framework, we consider $M = 3$ and $M = 5$ quantile trial levels for the weighted quantile process, with the corresponding framework named as FC-WQ-M, $M = 3, 5$. Storti and Wang (2021) show that the weighted quantile framework is not sensitive to the choice of M, meaning a small value of M, i.e., 3 or 5, can be chosen in real data applications to reduce the computation requirement.

To evaluate the FC-WQ framework more comprehensively, we assess the ability of the different models under comparison to forecast VaR and ES jointly, employing the joint loss values in Equation (6). We use this to jointly compare the VaR and ES forecasts from all models, because the AL log-score in Equation (6) is a strictly consistent scoring function that is jointly minimized by the true VaR and ES series.

First, Figure 2 shows the S&P500 ES forecasts from EGARCH-t, EGARCH-t-HS, ES-CAViaR-Mult-AS and FC-WQ-3. To make a more in-depth comparison of these models, Figure 3 presents the S&P 500 AL joint loss (log-score) values for each time step across the out-of-sample period. In general, we have a consistent story as in the quantile loss plot in Figure 1. The ES forecasts are again characterized by more stabilized and smaller joint loss values than the ones from the competing individual models, such as EGARCH-t, EGARCH-t-HS and ES-CAViaR-Mult-AS. These regularities are also consistently observed across different data sets.

In addition, we also test the ES forecasting performance via incorporating the Simple Average of the combined quantile forecasts (FC-SA), as in below Equation (24). The value of M is also selected as 3 and 5 respectively.

$$ES_t^{(\text{FC-SA})} = \frac{1}{M} \sum_{j=1}^{M} \widehat{Q}_t^{(C,\alpha_j)} \,. \tag{24}$$

Table 4 reports, for each model and data series, the value of the loss function in Equation (6) aggregated over the out-of-sample period: $\mathbf{S} = \sum_{t=N+1}^{N+H} S_t$, with $H = 2000$. In general, the proposed FC-WQ models produce on average the smallest joint loss, i.e, 4257.6 for FC-WQ-3 and 4257.8 for FC-WQ-5, and are best ranked (together with ES-CAViaR-Mult-AS). The results of FC-WQ-3 and FC-WQ-5 are quite close to each other, which is consistent with the observations in Storti and Wang (2021) and means the choice

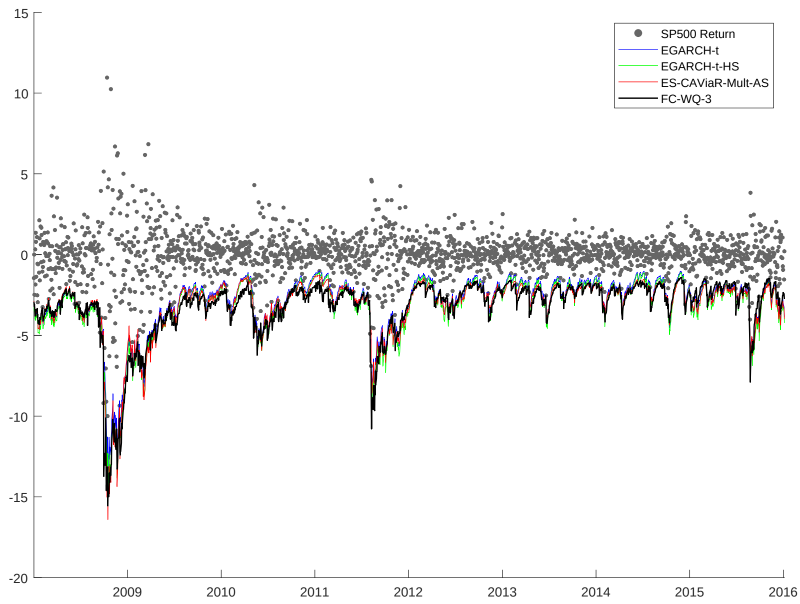

Figure 2: S&P500 ES forecasts from EGARCH-t, EGARCH-t-HS, ES-CAViaR-Mult-AS and FC-WQ-3.

of $M = 3$ can already produce ES forecast with good accuracy. The FC-SA framework produces joint loss values which are consistently larger than that from the corresponding FC-WQ, i.e., comparing FC-SA-M to FC-WQ-M ($M = 3, 5$). This lends evidence on the effectiveness of employing the Beta weighting scheme on the combined quantile forecasts. Lastly, GJR-GARCH-t and EGARCH-t, which are included in the model universe of the FC-WQ framework, are least preferred, with average rank as 12 and 11.5 respectively.

Lastly, similar to the VaR calibration test, following Patton et al. (2019) an ES regression-based calibration test is also conducted with results shown in Table 5. Overall, the observations are similar to that of the VaR calibration test. On the 10% significance level, the proposed framework FC-WQ together with CARE-AS are least likely to be rejected by the test, compared with other competing models. The ES-CAViaR-Mult-AS model produces top ranked joint loss results, while it is rejected on 4 markets via the calibration test. The GJR-GARCH-t and EGARCH-t are rejected for all six data sets. However, the proposed FC-WQ framework which includes the GJR-GARCH-t

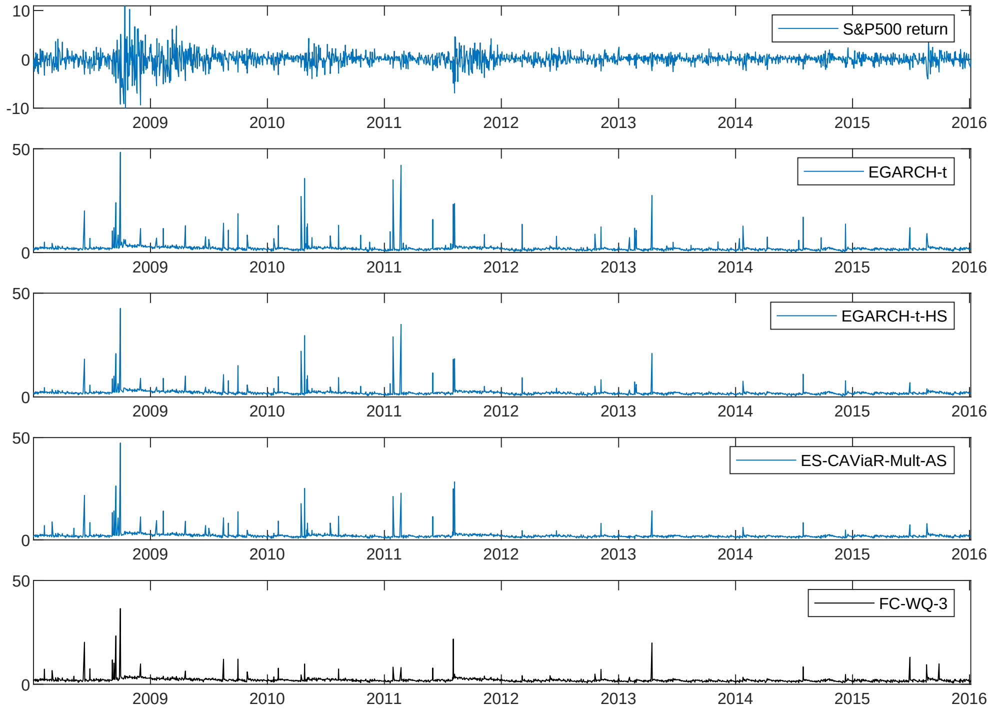

Figure 3: S&P500 out-of-sample VaR and ES AL joint loss from EGARCH-t, EGARCH-t-HS, ES-CAViaR-Mult-AS and FC-WQ-3.

Table 4: 2.5% VaR and 2.5% ES joint loss function values across the markets.

Model	S&P500	HangSeng	FTSE	DAX	SMI	ASX200	Avg Loss	Avg Rank
GJR-GARCH-t	4239.2	4601.7	4183.8	4601.4	4253.6	4009.1	4314.8	12.00
EGARCH-t	4290.6	4586.9	4192.1	4585.2	4245.2	3990.4	4315.1	11.50
POT-GJR-GARCH-t	4175.3	4587.6	4131.4	4557.7	4215.5	3963.5	4271.8	7.83
POT-EGARCH-t	4226.2	4575.6	4146.7	4543.2	4185.6	3954.6	4272.0	6.33
GJR-GARCH-t-HS	4174.1	4587.7	4131.3	4556.6	4216.0	3962.5	4271.3	7.50
EGARCH-t-HS	4225.1	4575.0	4146.9	4542.8	4185.5	3953.9	4271.5	5.67
CARE-AS	4276.8	4550.7	4160.6	4514.8	4252.5	4024.9	4296.7	8.33
ES-CAViaR-Add-AS	4242.3	4551.9	4131.8	4506.8	4192.1	3992.9	4269.6	5.33
ES-CAViaR-Mult-AS	4242.3	4564.2	4117.8	4509.1	4188.3	3977.4	4266.5	4.83
FC-SA-3	4157.7	4572.8	4138.7	4551.7	4211.5	3966.8	4266.5	6.83
FC-SA-5	4149.7	4568.1	4182.3	4534.7	4213.1	3946.6	4265.8	5.67
FC-WQ-3	4152.9	4568.1	4136.5	4537.1	4193.7	3957.5	4257.6	4.83
FC-WQ-5	4141.9	4565.6	4174.6	4520.8	4198.6	3945.6	4257.8	4.33

Note:Box indicates the favoured model and dashed box indicates the 2nd ranked model based on the average loss and rank.

and EGARCH-t in its model universe can still produce competitive ES forecasts via the quantile forecast combination and weighting scheme, which again lends support on its effectiveness.

Table 5: 2.5% ES calibration test at the 10% significance level.

Model	S&P500	HangSeng	FTSE	DAX	SMI	ASX200	Total
GJR-GARCH-t	0.000	0.037	0.000	0.000	0.000	0.000	6
EGARCH-t	0.003	0.066	0.000	0.000	0.000	0.001	6
POT-GJR-GARCH-t	0.067	0.368	0.135	0.003	0.000	0.000	4
POT-EGARCH-t	0.201	0.338	0.161	0.079	0.008	0.000	3
GJR-GARCH-t-HS	0.069	0.420	0.175	0.003	0.000	0.000	4
EGARCH-t-HS	0.205	0.345	0.207	0.091	0.011	0.000	3
CARE-AS	0.508	0.473	0.122	0.180	0.009	0.000	2
ES-CAViaR-Add-AS	0.241	0.725	0.123	0.059	0.014	0.000	3
ES-CAViaR-Mult-AS	0.298	0.696	0.086	0.055	0.018	0.000	4
FC-SA-3	0.458	0.541	0.311	0.131	0.007	0	2
FC-SA-5	0.729	0.506	0.343	0.170	0.004	0	2
FC-WQ-3	0.459	0.553	0.306	0.119	0.009	0.000	2
FC-WQ-5	0.712	0.450	0.344	0.166	0.006	0.000	2

Note: Box indicates the favoured model and dashed box indicates the 2nd ranked model based on the total number of rejections on the 10% significance level. p-values are presented for each index and each model.

7 Conclusion

In this paper, in order to reduce the impact of model uncertainty in tail risk forecasting, we propose an innovative framework based on a forecast combination and weighted quantile (FC-WQ) approach that extends the WQ approach in Storti and Wang (2021). The first step forecasting combination procedure combines quantile forecasts from all VaR models included in the model universe, on a grid of quantile levels. Then, the combined quantiles forecasts are employed as input to a quantile weighting scheme which is used to produce ES forecasts. The coefficients involved in the two steps of the procedure are estimated via optimizing the quantile loss and a strictly consistent joint VaR and ES loss, respectively. The selected model universe consists of parametric and semi-parametric models.

Compared to the VaR & ES forecasting combination approach in Taylor (2020), which combines "pairs" of VaR & ES forecasting from various models, our approach breaks the tie between the VaR and ES model, and only requires VaR forecasts from individual VaR models. Therefore, we implicitly consider a greater variety of functional forms without making the number of parameters explode.

In a comprehensive empirical study, improvements in the out-of-sample forecasting of tail risks, especially ES, are observed, compared to each individual model in the model universe and a simple average approach. A further advantage of the proposed forecasting combination and weighted quantile framework is its attitude to return "stabilized" quantile and joint loss values.

The proposed framework can be extended in a number of directions. First, other quantile forecasting combination schemes, i.e., considering a multivariate quantile framework in the spirit of White et al. (2010), could be considered as alternatives. Second, the first step quantile forecasting combination approach could be replaced by a cross validation approach which selects one model for each quantile level in the chosen grid. Third, a greater range of models, such as the ones incorporating high frequency based realized measures, could be included in the model universe. All these extensions are currently left for future investigation.

References

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

Aigner, D. J., T. Amemiya, and D. J. Poirier (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review 17*(2), 377–396.

Artzner, P. (1997). Thinking coherently. *Risk*, 68–71.

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

Basel Committee on Banking Supervision (2019). *Minimum capital requirements for market risk*. Bank for International Settlements.

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

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

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

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

Fissler, T. and J. F. Ziegel (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics 44*(4), 1680–1707.

Giacomini, R. and I. Komunjer (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics 23*(4), 416–431.

Gilli, M. et al. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics 27*(2), 207–228.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance 48*(5), 1779–1801.

Harvey, A. (2013). *Dynamic Models for Volatility and Heavy Tails.* Econometric Society Monographs.

Koenker, R. and J. A. F. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association 94*(448), 1296–1310.

McNeil, A. J. and R. Frey (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.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica 55*(4), 819–847.

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

Storti, G. and C. Wang (2021). Nonparametric expected shortfall forecasting incorporating weighted quantiles. *International Journal of Forecasting. In 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 & Economic Statistics 37*(1), 121–133.

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

White, H., T.-H. Kim, and S. Manganelli (2010). *Modeling Autoregressive Conditional Skewness and Kurtosis with Multi-Quantile CAViaR*. United Kingdom: Oxford University Press.