Received: 29 May 2015 | Revised: 23 December 2016

DOI: 10.1002/jae.2575

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Combining density forecasts using focused scoring rules

Anne Opschoor[1,2] | Dick van Dijk[2,3,4] | Michel van der Wel[2,3,4,5]

[1]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2]Tinbergen Institute, The Netherlands

[3]Erasmus University Rotterdam, Rotterdam, The Netherlands

[4]ERIM, Rotterdam, The Netherlands

[5]CREATES, Aarhus University, Aarhus, Denmark

Correspondence
Anne Opschoor, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands.
Email: a.opschoor@vu.nl

Summary

We investigate the added value of combining density forecasts focused on a specific region of support. We develop forecast combination schemes that assign weights to individual predictive densities based on the censored likelihood scoring rule and the continuous ranked probability scoring rule (CRPS) and compare these to weighting schemes based on the log score and the equally weighted scheme. We apply this approach in the context of measuring downside risk in equity markets using recently developed volatility models, including HEAVY, realized GARCH and GAS models, applied to daily returns on the S&P 500, DJIA, FTSE and Nikkei indexes from 2000 until 2013. The results show that combined density forecasts based on optimizing the censored likelihood scoring rule significantly outperform pooling based on equal weights, optimizing the CRPS or log scoring rule. In addition, 99% Value-at-Risk estimates improve when weights are based on the censored likelihood scoring rule.

1 | INTRODUCTION

Value-at-Risk (VaR) is a commonly used measure of downside risk for investments. Financial institutions are allowed by regulation (i.e., the Basel accords) to report VaR estimates for their asset portfolios obtained from their own "internal" model (subject to approval by the supervisory authorities). This luxury also creates a challenge. Given the availability of an abundant number of different methods for measuring (and managing) downside risk, financial institutions face the difficult task of choosing the "best" model for their purposes. This creates uncertainty because the true data-generating process is unknown. Boucher, Daníelsson, Kouontchou, and Maillet (2014) and Daníelsson, James, Valenzuela, and Zer (2016) analyze this notion of so-called "model risk" in the context of risk management, and provide tools to quantify the extent to which measures such as VaR are subject to the uncertainty involved in choosing a particular model specification. The point of departure for this paper is the conventional wisdom that each model is an incomplete description of reality. Hence relying upon a single model is dangerous when constructing a VaR estimate, because any model is "wrong" in some sense. For that reason, we examine whether combining different models may result in superior VaR estimates.

The contribution of this paper is to assess the merits of density forecast combination schemes that assign weights to individual density forecasts based on scoring rules that allow us to focus on a specific region of the densities' support that is of particular interest. This is motivated by the empirical context of measuring downside risk as described above, where VaR and related measures are characteristics of the left tail of the distribution of asset returns. Hence the accuracy of density forecasts in that region is of crucial importance. Combining density forecasts based on their behavior in the left tail may then offer better performance than combinations based on the complete density. We examine two scoring rules that meet this requirement of "focus," namely the censored likelihood (csl) scoring rule advocated by Diks, Panchenko, and van Dijk (2011), and (a weighted version of) the continuous ranked probability score (CRPS) of Gneiting and Ranjan (2011). We benchmark both scoring rules

against the conventional log score function (see, e.g., Mitchell & Hall, 2005), which considers the complete density. In addition, we include an equally weighted density forecast as a natural benchmark.

We investigate two weighting schemes for combining density forecasts based on these scoring rules. The first scheme, put forward by Hall and Mitchell (2007) and Geweke and Amisano (2011), aims to find "optimal" weights, in the sense that they maximize the average score of the combined density forecast with respect to the realization of the variable. This corresponds to minimizing the distance, as measured by the Kullback–Leibler information criterion, between the combined density forecasts and the true but unknown density (see Hall & Mitchell, 2007). We extend this approach by substituting the log scoring rule by the censored likelihood scoring rule and the continuous ranked probability score.

The second scheme, adopted from Jore, Mitchell, and Vahey (2010), assigns weights to the various individual density forecasts based on their relative average scores. Garratt, Mitchell, Vahey, and Wakerly (2011) and Aastveit, Gerdrup, Jore, and Thorsrud (2014) show that these recursive weighting schemes perform well when combining density forecasts of inflation and GDP respectively. The latter study also finds that this scheme performs better in terms of point forecast evaluation than standard point forecast combinations. Pettenuzzo and Ravazzolo (2016) show the added value of this weighting scheme in their empirical application of forecasting stock returns.

We use our novel methodology in an empirical application involving several recently developed univariate volatility models. In particular, beyond the traditional (exponential) generalized autoregressive conditional heteroskedasticity ((E)GARCH) model (Bollerslev, 1986; Nelson, 1991), we consider the HEAVY model (Shephard & Sheppard, 2010) and the realized GARCH model (Hansen, Huang, & Shek, 2012) that include realized measures, as well as the GAS model (Creal, Koopman, & Lucas, 2013). In addition to these different specifications for the volatility dynamics, motivated by Bao, Lee, and Saltoğlu (2007), among others, we consider a variety of conditional distributions, including the normal, Student-t, Laplace and skewed-t distributions. All models are applied to daily returns on the S&P 500, DJIA, FTSE and Nikkei stock market indexes from 2000 until 2013.

We evaluate the various combined density forecasts in both statistical and economical terms. Statistically, we compare predictive accuracy in the left tail by means of the censored likelihood scoring rule. Because we consider multiple weighting schemes, we use the model confidence set (MCS) approach of Hansen, Lunde, and Nason (2011) to account for the possible dependence between the csl scores of the combined density forecasts associated with these schemes. We assess the economic value of the combined density forecasts by examining the quality of the resulting 1-day 99% VaR estimates. For this purpose we use the conditional coverage (CC) test of Christoffersen (1998) and the dynamic quantile (DQ) test of Engle and Manganelli (2004).

Our results show statistically that density forecasts in the tail are more accurate if we pool density forecasts using the "optimal" weights based on the csl scoring rule compared to the optimal weights related to the whole density (log score) and equal weights. In addition, the optimal csl weights compete with the relative Jore weighting scheme with the log scoring rule. Further, pooling density forecasts based on the weighted CRPS function does not lead to superior density forecasts in the left tail. Economically, our results indicate that 99% 1-day VaR estimates based on pooling density forecasts with the optimal csl weights are superior against any other considered weighting scheme. That is, (i) the actual number of violations matches more closely the nominal value and VaR violations do not occur in clusters, and (ii) the exceedances are unpredictable.

Closely related to our approach is the work of Kapetanios, Mitchell, Price, and Fawcett (2015), who generalize the approach of Hall and Mitchell (2007) and Geweke and Amisano (2011) by proposing nonlinear opinion pools. More specifically, the weights assigned to each density forecast may vary by region of the density. Gneiting and Ranjan (2013) also develop nonlinear weighting schemes such as spread-adjusted and beta-transformed linear pools. A notable difference with our approach is that both studies still consider the full support of the densities in forming the combination, while the focus here is specifically on a particular region of the densities' support.

Another strand of literature related to our work is the focus on density forecast combinations from a Bayesian point of view (see, e.g., Aastveit, Ravazzolo, & Van Dijk, 2016; Billio, Casarin, Ravazzolo, & van Dijk, 2013; Del Negro, Hasegawa, & Schorfheide, 2016; Waggoner & Zha, 2012). These papers treat the combination weights as (time-varying) random variables. In the current paper we limit ourselves to static weights, although re-estimating the weights using a rolling window does allow the weights to vary over time. We leave the specification of a dynamic process for the combination weights as a topic for further research.

The remainder of this paper is organized as follows. Section 2 puts forward our methodology of combining density forecasts using focused scoring rules such as the csl and CRPS. Section 3 provides an overview of the univariate volatility models and the related assumed conditional density functions, which are used in the empirical application in Section 4. Section 5 concludes.

2 | COMBINING DENSITY FORECASTS

We consider the situation of a decision maker having n different forecast methods for a time series variable of interest y at his disposal. Each of these methods provides a predictive distribution for y in period t conditional on information available at time

$t-1$. The predictive density corresponding to a particular forecast method A_i, $i = 1, \ldots, n$, is denoted as $p_t(y_t|\mathcal{I}_{t-1}, A_i)$, where $\mathcal{I}_{t-1}$ indicates the information set up to and including time $t-1$. Suppose that at time T the decision maker aims to find the "best" predictive density for y_{T+1}, given an available history of density forecasts for the most recent τ periods $t = T-\tau+1, \ldots, T$ for all the n methods and the corresponding realizations $\mathcal{Y}_{T,\tau} = \{y_{T-\tau+1}, \ldots, y_T\}$. An often used approach for this purpose is to make use of scoring rules, which measure the quality of density forecasts by assigning a numerical score. Typically, a scoring rule is a function that depends on the density forecast and the actually observed value, such that a higher score is associated with a "better" density forecast. According to Gneiting and Raftery (2007), a scoring rule is *proper* if it satisfies the condition that incorrect density forecasts do not receive a higher score on average than the true density. This property is important and a natural requirement for any rational decision maker; see also Diebold, Gunther, and Tay (1998) for a discussion.

A well-founded proper scoring rule is the log score function (see, Amisano & Giacomini, 2007; Mitchell & Hall, 2005, among others). This function simply takes the logarithm of the predictive density evaluated at the realization y_t. Thus, for a particular method A_i at a specific time t, the log score S^l is given by

$$S^l(y_t; A_i) = \log p_t(y_t|\mathcal{I}_{t-1}, A_i). \tag{1}$$

As discussed in detail in Hall and Mitchell (2007), the log scoring rule is closely related to information-theoretic goodness-of-fit measures such as the Kullback–Leibler information criterion (KLIC). A higher value of the log score associated with the density forecast $p_t(y_t|\mathcal{I}_{t-1}, A_i)$ coincides with a lower value of the KLIC or, put differently, maximizing the log score is equivalent to minimizing the KLIC.

Scoring rules such as the log score in Equation 1 can be used for choosing a single, "best" density forecast from the n available candidates. Obviously, this boils down to selecting the density $p_{T+1}(y_{T+1}|\mathcal{I}_T, A_i)$ from the forecast method A_i that has achieved the highest log score over the history $\mathcal{Y}_{T,\tau}$. It is, however, highly unlikely that a particular forecast method renders the "true" predictive density. As a result, it might be beneficial to combine various density forecasts. The literature on this topic dates back at least to Bacharach (1974), who considered linear combinations of (subjective) probability distributions, known as *linear opinion pools*. Hall and Mitchell (2007) and Geweke and Amisano (2011) also studied linear opinion pools and introduced this idea into the econometric forecasting literature. In terms of our context and notation, a linear opinion pool is a predictive density of the form

$$\sum_{i=1}^{n} w_i p_t(y_t|\mathcal{I}_{t-1}, A_i), \quad \text{with} \quad w_i \geqslant 0 \quad \text{for} \quad i = 1, \ldots, n \quad \text{and} \quad \sum_{i=1}^{n} w_i = 1, \tag{2}$$

where the restrictions on the weights w_i are imposed to ensure that Equation 2 is a valid probability density function.

Obviously, a crucial issue determining the success (or failure) of a linear opinion pool is the choice of the combination weights w_i in Equation 2. We examine and compare the three approaches that have been most popular in practice so far. First, following Geweke and Amisano (2011) an "optimal" prediction pool is obtained when the weights at time T are chosen so as to optimize past performance in terms of the log scoring rule, that is, by maximizing

$$S^l(\mathcal{Y}_{T,\tau}, C) = \sum_{t=T-\tau+1}^{T} \log \left[\sum_{i=1}^{n} w_i p_t(y_t|\mathcal{I}_{t-1}, A_i) \right], \tag{3}$$

with C indicating that a combination of density forecasts is evaluated instead of an individual predictive density from a particular method A_i. Second, Jore et al. (2010) put forward a weighting scheme in which the relative past performance of each individual predictive density in terms of the log score determines its weight in the opinion pool, that is,

$$w_i = \frac{\exp\left(\sum_{t=T-\tau+1}^{T} \log p_t(y_t|\mathcal{I}_{t-1}, A_i) \right)}{\sum_{i=1}^{n} \exp\left(\sum_{t=T-\tau+1}^{T} \log p_t(y_t|\mathcal{I}_{t-1}, A_i) \right)}. \tag{4}$$

It is useful to note that, in contrast to the optimal opinion pool based on maximizing Equation 3, in this case the log score of the combined predictive density is not necessarily maximized. Third, as a benchmark, we form an equally weighted opinion pool with $w_i = 1/n$, $i = 1, \ldots, n$. In practice, the equally weighted combination often turns out to be difficult to beat.[1]

[1] An important reason for this finding is that in the first two approaches the combination weights are treated as unknown parameters. The associated estimation uncertainty, which can be quite large especially when τ, the number of historical density forecasts used in the estimation, is small, deteriorates forecast performance. Combined with the fact that often the optimal weights are not substantially different from $1/n$, it is often found that even an optimal opinion pool may not improve upon an equally weighted combination; see Timmermann (2006) for a discussion in the context of combining point forecasts.

The main aim of this paper is to extend the idea of linear opinion pools but to combine predictive densities in the best possible way when the focus is on a particular region of interest. For this purpose, we consider two alternative scoring rules. First, we make use of the censored likelihood (csl) scoring rule, advocated by Diks et al. (2011). They prove that this scoring rule is proper and demonstrate its usefulness in case one is interested in the accuracy of density forecasts in a specific region. The csl score function for a specific region B_t for method A_i at time t reads

$$S^{\text{csl}}(y_t|A_i) = I[y_t \in B_t] \log p_t(y_t|\mathcal{I}_{t-1}, A_i) + I[y_t \in B_t^c] \log \left(\int_{B_t^c} p_t(y|\mathcal{I}_{t-1}, A_i) \text{d}y \right), \tag{5}$$

with B_t^c the complement of B_t and $I[\cdot]$ an indicator function that takes the value 1 if the argument is true. The first term on the right-hand side in Equation 5 essentially uses the familiar log scoring rule to measure the quality of the density forecast $p_t(y_t|\mathcal{I}_{t-1}, A_i)$ in the region of interest B_t. The second term computes the value of the cumulative distribution function (CDF) of the density forecast in the region outside B_t. Hence any observation outside B_t ignores the shape of $p_t(y_t|\mathcal{I}_{t-1}, A_i)$ outside B_t. As discussed in Diks et al. (2011), this second term is necessary, however, to obtain a proper scoring rule. It is not difficult to understand that omitting it would result in a score function that favors (possibly incorrect) predictive densities with (relatively) more probability mass in the region of interest B_t (see also Amisano & Giacomini, 2007). Note that Equation 5 simplifies to the log scoring rule of Equation 1 if B_t represents the full sample space.

The second alternative scoring rule we consider in this paper is a weighted version of the continuous ranked probability score (CRPS) of Gneiting and Ranjan (2011), which is defined as

$$S^{\text{CRPS}}(y_t; A_i) = \int_{-\infty}^{\infty} \text{PS}(F(z|\mathcal{I}_{t-1}, A_i), I[y_t \leqslant z]) u(z) \text{d}z. \tag{6}$$

Here, $u(z)$ is a non-negative weight function on the real line and $F(z|\mathcal{I}_{t-1}, A_i)$ is the CDF of the predictive density $p_t(z|\mathcal{I}_{t-1}, A_i)$. Further, $\text{PS}(F(z), I[y < z])$ is the quadratic or Brier probability score (Gneiting & Raftery, 2007):

$$\text{PS}(F(z), I[y \leqslant z]) = (F(z) - I[y < z])^2 \tag{7}$$

for the probability forecast $F(z)$ of the binary event $[y \leqslant z]$. We refer to Gneiting and Ranjan (2011) for a decision-theoretic interpretation of this scoring rule. In line with the csl scoring rule, we choose $u(z)$ in Equation 6 as an indicator function equal to one in the region B_t and zero elsewhere. Note that the (weighted) CRPS actually is a loss function, in the sense that "better" density forecasts lead to a smaller CRPS value.

In order to combine the n available predictive densities $p_t(y_t|\mathcal{I}_{t-1}, A_i)$, $i = 1, \dots, n$, based on the csl score or the CRPS, we consider the same approaches as discussed before for the log scoring rule. We continue to use linear opinion pools as defined in Equation 2, but now obtain "optimal" weights by either maximizing the corresponding censored likelihood score function over the history $\mathcal{Y}_{T,\tau}$, that is,

$$S^{\text{csl}}(\mathcal{Y}_{T,\tau}, C) = \sum_{t=T-\tau+1}^{T} \log \left[\sum_{i=1}^{n} w_i \left(I[y_t \in B_t] p_t(y_t|\mathcal{I}_{t-1}, A_i) + I[y_t \in B_t^c] \int_{B_t^c} p_t(y|\mathcal{I}_{t-1}, A_i) \text{d}y \right) \right] \tag{8}$$

or by minimizing the CRPS function, given by

$$S^{\text{CRPS}}(\mathcal{Y}_{T,\tau}, C) = \sum_{t=T-\tau+1}^{T} \left[\int_{-\infty}^{\infty} \text{PS} \left(\sum_{i=1}^{n} w_i F(z|\mathcal{I}_{t-1}, A_i), I[y_t \leqslant z] \right) u(z) \text{d}z \right]. \tag{9}$$

Similarly, we adapt the weighting scheme of Jore et al. (2010) by replacing the log score in Equation 4 with the csl score given by Equation 5 or the (inverse) CRPS given by Equation 6.[2]

[2]As discussed before, the CRPS is a loss function, and we therefore invert its values when adapting the Jore weighting scheme.

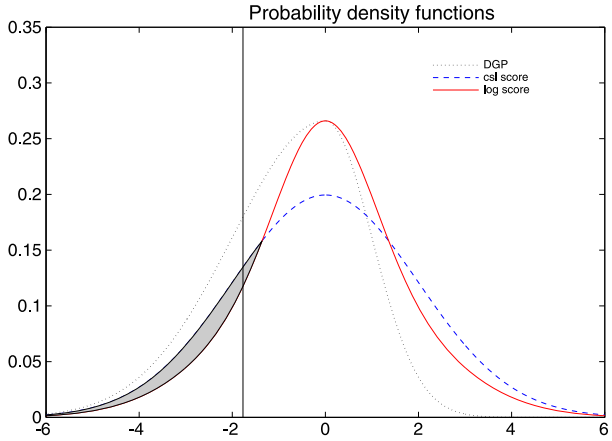

FIGURE 1 Probability density functions (pdfs). This figure depicts pdfs corresponding to Example 1. The black dashed line represents the true density (DGP); the solid line and the dashed line denote the pdf of the combined density resulting from optimizing the log score function and the censored likelihood score function, respectively. Further, the grey shaded area represents the difference in the pdfs associated with both scoring rules. In case of the csl score function, B_t represents the left tail $y_t < \hat{y}^{0.15}$, with $\hat{y}^{0.15}$ being the 0.15th quantile of the distribution of y_t, which is represented by the vertical line [Colour figure can be viewed at wileyonlinelibrary.com]

In our empirical application using volatility models and downside risk measures for stock index returns, our focus is on the left tail of the distribution, rather than on the whole density. This exactly provides the motivation for using the csl scoring rule and a weighted version of the CRPS. Of course, if one of the forecast methods corresponds to the true data-generating process (DGP), the weight should be maximized on the corresponding predictive density, and we would not require the csl scoring rule or the CRPS. As pointed out by Geweke and Amisano (2011), the concept of Bayesian model averaging (BMA) is related to the crucial assumption that the model set contains the true model. This implies that after many data points the assigned weight goes to unity for one of the considered models. However, it is highly unlikely that any of the proposed volatility models is perfectly true. Moreover, Aastveit et al. (2016) show that the concept of "model incompleteness", that is, the true model is not in the pool, plays a large role for nowcasting when the data uncertainty is large. Example 1 shows the importance of the csl scoring rule compared to the log scoring rule when all forecast methods are misspecified.

Example 1. *Suppose the DGP of a time series y_t is given by the two-piece Normal distribution with density*

$$f(y_t;\mu,\sigma_1,\sigma_2) = \begin{cases} G\exp(-\frac{1}{2\cdot\sigma_1^2}(y_t-\mu)^2) & \text{if} \quad y_t \leqslant \mu \\ G\exp(-\frac{1}{2\cdot\sigma_2^2}(y_t-\mu)^2) & \text{if} \quad y_t > \mu, \end{cases} \tag{10}$$

with $G = \sqrt{\frac{2}{\pi}}(\sigma_2+\sigma_2)^{-1}$. For this particular example, suppose $\mu = 0, \sigma_1 = 2$ and $\sigma_2 = 1$.

We consider a combination of two predictive densities that are both misspecified. Suppose we take two normal distributions, both with mean zero and standard deviations 2 and 1, respectively. Then the combined density reads $\hat{p}(y_t) = w_1 N_1(0,2) + w_2 N_2(0,1)$ with $w_1, w_2 \geqslant 0$ and $w_1 + w_2 = 1$. Assume further that our focus is on the left tail $y_t < \hat{y}^{0.15}$ with $\hat{y}^{0.15}$ the 0.15th quantile of the two-piece Normal distribution of y_t.

The optimal weights from the log scoring rule are equal to $w_1 = 2/3$ and $w_2 = 1/3$ respectively.[3] *The weights associated with the censored likelihood scoring rule are equal to 1 and 0 respectively, which follows directly from the log scoring rule with B_t the left tail $y_t < \hat{y}^{0.15}$.*

Figure 1 shows the related density functions. The shaded grey area represents the difference between both functions associated with the combined predictive density resulting from optimizing both scoring rules. Clearly, this area shows that using the csl approach approximates the left tail more accurately than the log score approach.

[3]These values can be understood intuitively by noting that for the two-piece normal distribution it holds that $\Pr[y_t \leqslant \mu] = \frac{\sigma_1}{\sigma_1+\sigma_2}$. Hence, in our specific example, $y_t \leqslant 0$ with probability 2/3 and this area is up to a scalar perfectly fitted by the $N_1(0,2)$ distribution. Likewise, $y_t \geqslant 0$ with probability 1/3, where the corresponding area almost exactly matches the $N_2(0,1)$ distribution.

3 | FORECAST METHODS

In this section we outline the forecast methods that we consider in our empirical analysis of density forecasts for daily stock index returns. Empirically, the distribution of daily asset returns is characterized by time-varying (conditional) volatility and non-normal features such as peakedness, fat-tailedness and skewness. Our forecast methods essentially combine two elements that attempt to capture these characteristics: a (parametric) specification of the volatility dynamics and a specific (parametric) type of conditional distribution for the returns. All forecast methods can be framed in the following general set-up for the asset return y_t on day t:

$$y_t = \mu + \sqrt{h_t} z_t, \quad \text{with} \quad z_t | \mathcal{I}_{t-1} \sim D(0,1), \tag{11}$$

where μ denotes the conditional mean of the returns,[4] $h_t = \mathsf{V}[y_t|\mathcal{I}_{t-1}]$ is the conditional variance, and z_t is the standardized unexpected return following a certain distribution $D(\cdot)$ with mean zero and unit variance. Note that $D(\cdot)$ corresponds to the conditional distribution of the returns, that is, $y_t|\mathcal{I}_{t-1} \sim D(\mu, h_t)$. We will consider the normal, Student-t, Laplace and the skewed-t distribution of Hansen (1994) as possible choices for $D(\cdot)$. These distributions are characterized by its widespread use (Normal), fat-tailedness (Student-t, Laplace) and the allowance of being asymmetric (skewed-t).

We propose four classes of models for the conditional variance h_t. The first class is the popular GARCH type of model. The traditional GARCH(1,1) model (Bollerslev, 1986) for h_t is given by

$$h_t = \omega + \alpha(y_{t-1} - \mu)^2 + \beta h_{t-1}, \tag{12}$$

with $\omega > 0, \alpha > 0$ and $\beta > 0$ to ensure a positive variance. Note that here the lagged squared demeaned return is the innovation for the conditional variance. A well-known extension to this model is made by the EGARCH model of Nelson (1991), which specifies the log of the conditional variance as

$$\log h_t = \omega + \gamma z_{t-1} + \alpha(|z_{t-1}| - E[|z_{t-1}|]) + \beta \log h_{t-1}. \tag{13}$$

This model allows for asymmetric effects of return shocks on volatility, which are often found to be empirically relevant for stock (index) returns in the sense that negative unexpected returns trigger a much stronger increase in volatility than positive unexpected returns of the same magnitude.

Although many other extensions of the GARCH model have been proposed, here we only consider the EGARCH model because of its popularity. We restrict also the other considered model classes in this study to their basic specification, although several variants/extensions are possible. We make this choice as our main goal is to compare different volatility model *classes* (combined with different conditional return distributions), and not subtly different models *within* a specific class. We refer to Hansen and Lunde (2005) and Bao et al. (2007) for such comparisons.

Creal et al. (2013) develop the broader class of generalized autoregressive score (GAS) models, which includes the GARCH model in Equation 12 as a special case. The key property of the GAS models is that innovations for time-varying parameters are based on the score of the probability density function at time t.[5] In our context of volatility dynamics, the time-varying parameters are the conditional variances h_t. This interpretation becomes clear from the structure for h_t in the GAS(1,1) model, given by

$$\begin{aligned} h_t &= \omega + \alpha s_{t-1} + \beta h_{t-1}, \\ s_t &= Q_t \nabla_t, \\ \nabla_t &= \frac{\partial \log D(y_t|h_t, \mathcal{I}_{t-1}, \theta)}{\partial h_t}, \end{aligned} \tag{14}$$

with $D(y_t|h_t, \mathcal{I}_{t-1}, \theta)$ the conditional return density, θ the parameter vector, ∇_t the score and Q_t a scale factor. We follow Creal et al. (2013) and define the scale factor as $1/\mathsf{E}_t[\nabla_t^2]$, where E_t denotes the expectation with respect to the return density $D(y_t|h_t, \mathcal{I}_{t-1}, \theta)$. In the case of a fat-tailed Student-t distribution with ν degrees of freedom, the score-based volatility model reads

[4] For ease of exposition, we assume the conditional mean is constant over time, although this could easily be relaxed to allow for a time-varying mean μ_t.
[5] Note that we use the term "score" with two different meanings: (i) a number that is assigned to measure the quality of density forecasts, as in Equations 1 and 5; and (ii) here in the GAS models to indicate the derivative of the log density with respect to a certain parameter.

$$h_t = \omega + \alpha(1 + 3/\nu)\frac{\nu + 1}{(\nu - 2) + \frac{(y_{t-1}-\mu)^2}{h_{t-1}}}(y_{t-1} - \mu)^2 + \beta h_{t-1}, \tag{15}$$

and will be labeled as the GAS-t model. This illustrates one of the key differences between a standard GARCH approach as in Equation 12: the second term on the right-hand side in Equation 15 is such that more extreme unexpected returns are downweighted and have a more moderate effect on the variance. Intuitively, if the distribution is more heavy tailed, it is less likely that an extreme observation corresponds with an increase in volatility. Note that this is a function of the degrees of freedom ν; when $\nu \to \infty$, Equation 15 again converges to the GARCH specification in Equation 12, as the Student-t distribution converges to a normal in that case. We again impose $\omega > 0$, $\alpha > 0$ and $\beta > 0$ in the estimation of the parameters.

The third and fourth model classes that we consider employ realized measures to describe the dynamics of daily volatility. A realized measure is a nonparametric or "model-free" estimator of the variance of an asset return based on return observations at a higher frequency than the interval that the variance refers to. Realized measures provide a more accurate estimate of daily volatility than the squared daily return, as used in the GARCH models (see Andersen, Bollerslev, Diebold, & Labys, 2003).

A recently developed model that explicitly incorporates a realized measure in daily volatility models is the HEAVY model of Shephard and Sheppard (2010). In particular, this model assumes the following structure for the conditional variance h_t and the conditional expectation of the realized measure $\xi_t = \mathsf{E}[\mathrm{RM}_t|\mathcal{I}_{t-1}]$:

$$h_t = \omega + \alpha \mathrm{RM}_{t-1} + \beta h_{t-1}, \tag{16}$$

$$\xi_t = \omega_R + \alpha_R \mathrm{RM}_{t-1} + \beta_R \xi_{t-1}. \tag{17}$$

All parameters should be positive to avoid negative values of h_t and ξ_t. The HEAVY model is seen to consist of a GARCH structure for both h_t and ξ_t, with the lagged realized measure RM_{t-1} as innovation term.

A second model that relates conditional volatility to realized measures is the realized GARCH model (RGARCH) of Hansen et al. (2012). The basic specification is given by

$$h_t = \omega + \alpha \mathrm{RM}_{t-1} + \beta h_{t-1}, \tag{18}$$

$$\mathrm{RM}_t = \delta + \phi h_t + \tau(z_t) + u_t, \tag{19}$$

with u_t an additional random variable with mean zero and variance σ_u^2. Further, $\tau(z_t)$ is the leverage function, defined in the basic form as $\tau_1 z_t + \tau_2(z_t^2 - 1)$. This function allows for the empirical finding that negative and positive shocks may have a different impact on the volatility, comparable to the EGARCH specification in Equation 13. Except for τ_1, which is typically negative, all parameters are restricted to be positive. The dynamics for h_t are similar for both the HEAVY and RGARCH models; however, the difference arises in the specification of (the expectation of) RM_t. The HEAVY model proposes a GARCH structure for $\mathsf{E}[\mathrm{RM}_t|\mathcal{I}_{t-1}]$, while the RGARCH model explicitly relates RM_t to the conditional variance at time t and additionally introduces a leverage component.

We estimate the parameters in all models by maximum likelihood. This is not computationally involved or time consuming, since we are dealing with univariate models with a maximum of 11 parameters (in the case of the RGARCH specification combined with the skewed-t distribution) to be estimated. In addition, we can estimate the HEAVY parameters of Equations 16 and 17 separately (see Shephard & Sheppard, 2010, for details).

4 | EMPIRICAL ANALYSIS

4.1 | Data

We analyze the benefits of using the scoring rules based on the left tail to combine density forecasts obtained from the various forecast methods discussed in Section 3 for daily returns on four major stock market indexes: S&P 500, DJIA, FTSE and Nikkei. Our sample covers the period from January 3, 2000 to June 28, 2013. Daily returns as well as the corresponding realized measures are obtained from the Oxford-Man Institute's "Realized Library".[6] We follow Shephard and Sheppard (2010) and

[6]See http://realized.oxford-man.ox.ac.uk/.

use the realized kernel (see Barndorff-Nielsen, Hansen, Lunde, & Shephard, 2008) as the realized measure in the HEAVY and RGARCH specifications in Equations 16 and 18.[7] Days on which the exchange is closed are deleted from the sample.[8]

4.2 | Implementation details

We apply a rolling window scheme to estimate the model parameters and construct density forecasts. Specifically, we use a window of approximately 3 years ($T_{\text{est}} = 750$ observations) to estimate the model parameters and construct one-step-ahead forecasts of h_t and the corresponding density forecasts for each time t (t $= T_{\text{est}} + 1, \ldots, T$). We choose $T_{est} = 750$ such that there is a sufficient number of observations for parameter estimation of the models. After T_w subsequent density forecasts have been obtained for each model, we first construct the "optimal" weighting schemes. This involves optimizing Equations 3, 8 and 9 to obtain w_t. Optimizing the first two score functions is done by applying the algorithm of Conflitti, De Mol and Giannone (2015). This iterative algorithm is easy to implement and works well even when the number of forecasts to combine gets large. We refer to the Appendix for more details. Following Gneiting and Ranjan (2011), we proxy the integral of Equation 6 by the sum of evaluated CRPS values using a fine grid. Optimizing Equation 9 is now fairly easy, as this function is a quadratic problem. We repeat the optimization of Equations 3, 8 and 9 by means of a rolling scheme with a window of $T_w = 750$ density forecast evaluations at each time t(t $= T_{\text{est}} + T_w, T_{\text{est}} + T_w + 1, \ldots, T - 1$). To summarize the above approach, we repeat the following steps:

For $t_s = 750, \ldots, T - 1$, do

Step 1: Estimate the parameters of the individual volatility models $A_i (i = 1, \ldots 18)$, using 750 observations: $t = t_s - 749, \ldots, t_s$.

Step 2: Construct a one-step-ahead density forecast for each model A_i at time $t = t_s$.

Step 3: If $t_s \geqslant 1,499$, optimize the log score function, the csl score function and the CRPS function of Equations 3, 8 and 9 based on the past 750 predictive densities made in Step 2 at time $t = t_s - 749, \ldots, t_s$ to obtain w_{t_s+1}.

Beyond estimating weights by means of optimization, we estimate the Jore weights of Equation 4 recursively, applied to the log score, csl score and the CRPS. We set τ equal to 250 as a training sample to initialize the weights. As the first weight of the optimized weighting schemes is computed at $t = 1,500$ (since we need the first 750 observations to estimate the parameters and then 750 one-step-ahead predictions to optimize the particular score function), the first recursive Jore weights are also computed at $t = 1,500$. It uses the particular score of the predictive densities made at $t = 1,250, 1,251, \ldots, 1,499$.

For the csl score function, we define the region B_t as the left tail $y_t < \hat{r}_t^{\kappa}$ with $\hat{r}_t^{\kappa}$ the κth quantile of the empirical distribution of the 750 returns in the corresponding estimation window. Similarly, we define the weight function $u(z)$ in Equation 6 as a step function, which takes the value one in the region B_t and zero elsewhere.

When choosing a particular value of T_w (τ) and κ, we first emphasize the trade-off in the choice of κ. Given our interest in the left tail, we should take a small value of κ. However, the corresponding number of observations in the region of interest becomes very small, such that the variation in the csl scores across the different models declines.[9] Similarly, there is a trade-off in the choice of $T_w(\tau)$. On the one hand, choosing $T_w(\tau)$ as large as possible is desired in order to use more information to compute the weights w_t. On the other hand, if the relative performance of different models varies over time, a smaller value of $T_w(\tau)$ might be a better choice. In addition, $T_w(\tau)$ and κ are interrelated as a low value of κ combined with a small window results in a small number of observations within the region B_t. Given these trade-offs and the relation between those two variables, we choose $T_w = 750$, $\tau = 250$, and κ equal to 0.15 and 0.25, respectively. This corresponds to 112 (0.15) or 187 (0.25) observations in the left tail to optimize the weights. In addition, there are around 38 (0.15) and 62 (0.25) observations in the left tail for the Jore weighting scheme with the csl scoring rule.

4.3 | Evaluation

We assess the accuracy of our (combined) density forecasts in two ways. First, we focus purely on the predictive density in the left tail and investigate statistically whether combining densities based on the left tail adds any value.

Since we have quite some different combination schemes, we do not conduct a bivariate test of equal performance of two different schemes as done in Diks et al. (2011) and Gneiting and Ranjan (2011). Instead, we account for the interdependence

[7] The realized kernel is a robust high-frequency estimator of the variance, correcting for market microstructure noise.
[8] We delete 1–1.5% of the days on the S&P 500, DJIA and FTSE indexes and 3% in the case of the Nikkei index.
[9] Recall that if y_t is outside the region B_t with B_t the left tail $y_t < \hat{r}_t^{\kappa}$, the csl score is the CDF of y_t in the complement of the region.

between all schemes and use the model confidence set (MCS) of Hansen et al. (2011), applied on the (minus) csl and CRPS loss functions.

The second way we explore the additional value of using censored densities is based on 1-day VaR estimates. For the individual models considered in this study, the 1-day VaR estimate reads

$$\text{VaR}_t^{1-q} = \mu + z_q\sqrt{h_t}, \tag{20}$$

with μ the estimated conditional mean return, h_t the (forecast) conditional variance, and z_q represents the qth quantile of the assumed conditional distribution. However, we cannot apply Equation 20 when our predictive distribution is a combination of individual distributions since the VaR of a mixture of densities is not necessarily equal to the weighted average of the individual VaRs. We use simulation techniques to overcome this issue. That is, we simulate daily returns from each individual model/distribution according to the assigned weight (and conditional variance) to obtain the required quantile of the total distribution to compute the $(1-q)\%$ VaR.

We evaluate the accuracy of the VaR estimates both in terms of their individual and relative performance. The former is done by two tests. First, we use the conditional coverage test of Christoffersen (1998), which combines two features of a well-specified VaR measure: the actual frequency of violations should be in line with the expected number of violations, and the violations should not occur in clusters. The second test is the dynamic quantile (DQ) test of Engle and Manganelli (2004), which investigates whether the VaR exceedances are predictable. More specifically, the exceedance of a VaR of a method at time t should be unrelated to any information available at time $t-1$.

We compare the relative performance of VaR estimates of all combination methods using the following asymmetric linear tick loss function, which is also used in the conditional predictive ability (CPA) test of Giacomini and White (2006):

$$\text{L}_{A_i}^{q}(e_t) = (q - I[e_t < 0])e_t, \tag{21}$$

where $e_t = y_t - \text{VaR}_t^{1-q}$. The loss function is asymmetric in the sense that if a VaR violation occurs (i.e., $e_t < 0$)) the negative number $q-1$ is multiplied by the magnitude of the violation e_t, resulting in a penalization of $(1-q)\times e_t$. In contrast to this, if there is no violation, the loss is equal to $q \times e_t$, which is considerably lower.[10] Hence a model A_i is more penalized when a VaR violation is observed. The larger the magnitude of this violation, the larger the penalization. Similar to the density forecasts, we compute the losses of all combination methods and use the MCS for comparison.

4.4 | Results

Before assessing the merits of combining density forecasts based on various scoring rules in statistical and economic terms, we first present the weights obtained by our six different weighting schemes. Three schemes are based on optimizing the log score function in Equation 3, the csl score function in Equation 8 and the CRPS function in Equation 9, respectively. The other three schemes are based on the Jore weighting scheme (see Equation 4) with the aforementioned scoring rules as input.

Figure 2 shows the result of the iterative process of optimizing weights according to the log, csl and CRPS score functions in the left-hand panel, as well as the recursive Jore weights in the right-hand panel. The csl scoring rule and CRPS are both based on the 0.15th quantile of the in-sample returns. Each subgraph depicts the weights of the 18 models obtained for the daily S&P 500 returns.[11]

The figure shows in the first place that the resulting optimized weights differ considerably across the used scoring rule. For example, optimizing the csl score function results in dominance by the HEAVY skewed-t model, the EGARCH skewed-t model and the HEAVY Laplace model, respectively, while optimizing the CRPS function allocates a huge weight to the Gaussian HEAVY model during the period 2009–2011.

An interesting second result from the left-hand panel is the strong appetite for models that incorporate realized measures since the HEAVY model gets in general a large weight in all weighting schemes. This does not surprise us, as incorporating realized measures will improve the measurements and forecasts of volatility. In addition, the subgraphs also show the preference for fat-tailed distributions, such as the skewed-t and the Laplace distribution.

Finally, the Jore weights deviate substantially from the optimal weights with the differences between all model weights being much smaller. Put differently, the Jore weighting schemes discriminate only at a marginal level between all models. An exception

[10] Suppose the 95% 1-day VaR of model A and B are equal to -5% and -8% respectively, while the actual return is -6%. The loss associated with model A is equal to $(0.05-1)(-1) = 0.95$, whereas the loss of model B is equal to $(0.05-0)(-2) = 0.10$.

[11] Weighting schemes corresponding to the other three stock market indexes have a similar interpretation and are available upon request.

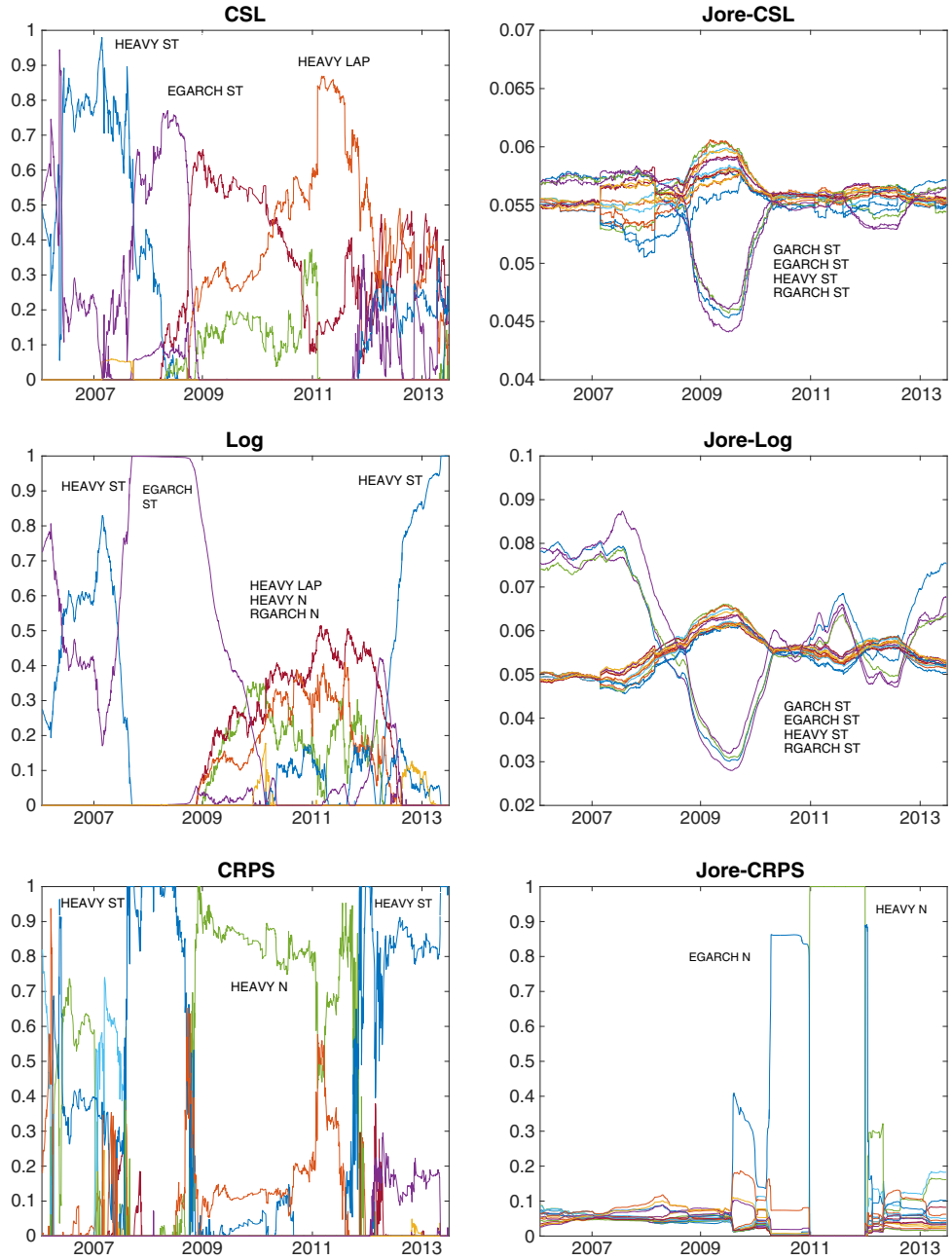

FIGURE 2 Pooling weights, S&P 500 index. This figure depicts the evolution of weights based on optimizing the logarithmic score function of Equation 3, the csl score function of Equation 8 and the CRPS function of Equation 9 with a moving window of $T_w = 750$ one-step-ahead evaluated density forecasts, using daily returns of the S&P 500 index. In addition, recursive Jore weights of Equation 4 are depicted for the three scoring rules of above, with a training sample of 250 observations. In the case of the csl (CRPS) score function, B_t $(u(z))$ represents the left tail $y_t < \hat{r}^{0.15}$ $(u < \hat{r}^{0.15})$, with $\hat{r}^{0.15}$ the 0.15th quantile of the empirical distribution of the moving estimation window of 750 returns. The labels refer to the models that have the highest weight at a given period. The abbreviations "ST," "Lap" and "N" stand for skewed-t, Laplace and normal, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

is given by the bottom right panel of the figure: although the Jore-CRPS weights are in the first part of the sample similar to the Jore-csl and Jore-log weights, a large weight is given to the EGARCH N model in 2010 and the HEAVY N model in 2011.

4.4.1 | Statistical results

Table 1 examines the usefulness of pooling with weights while focusing on the left tail by showing results of the MCS, which at the start contains seven different weighting schemes: three schemes based on optimizing the log score, csl score and CRPS

TABLE 1 Evaluation of 1-day-ahead combined density forecasts

	$\kappa = 0.15$				$\kappa = 0.25$			
	S&P 500	**DJIA**	**FTSE**	**Nikkei**	**S&P 500**	**DJIA**	**FTSE**	**Nikkei**
Panel A: csl score loss function								
csl	**1**	**1**	**1**	**1**	**1**	**1**	**0.717**	**1**
CRPS	0.012	**0.959**	0.001	**0.867**	**0.873**	**0.901**	0.000	0.018
Jore-csl	0.012	0.026	0.008	**0.213**	0.029	0.017	0.000	0.045
Jore-CRPS	0.012	0.017	0.001	**0.213**	**0.334**	0.011	0.000	**0.464**
log	0.012	0.017	**0.942**	**0.127**	0.018	0.011	**1**	**0.464**
Jore-log	**0.642**	**0.959**	**0.452**	**0.764**	**0.873**	**0.901**	**0.106**	**0.464**
EQW	0.012	0.017	0.001	**0.127**	0.018	0.011	0.000	0.018
Panel B: Log score loss function								
csl	**0.978**	**0.980**	**0.123**	**0.963**	**0.816**	**1**	**0.163**	**0.919**
CRPS	0.065	0.000	0.000	0.000	**0.169**	0.000	0.000	0.000
Jore-csl	0.003	0.001	0.000	0.001	0.006	0.000	0.000	0.000
Jore-CRPS	0.000	0.000	0.000	0.051	0.001	0.000	0.000	0.095
log	**1**	**1**	**1**	**1**	**1**	**0.813**	**1**	**1**
Jore-log	**0.123**	0.077	0.000	0.069	**0.187**	0.080	0.000	0.094
EQW	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000

Note. This table reports results of testing equal predictive accuracy of combined density forecasts, with weighing schemes based on the csl scoring rule of Equation 5, the log scoring rule of 1 or the CRPS of Equation 6. In the case of the csl score (CRPS) function, B_t ($u(z)$) represents the left tail $y_t < \hat{r}^{\kappa}$ ($u < \hat{r}^{\kappa}$), with $\hat{r}^{\kappa}$ the κth quantile of the empirical distribution of the in-sample returns. We set κ equal to 0.15 and 0.25, respectively. The weights are repeatedly optimized based on a moving window of 750 evaluated density forecasts. In addition, recursive Jore weights of Equation 4 are computed for the three scoring rules above, with a training sample of 250 observations. We report p-values associated with the MCS. Bold values represent models that are in the MCS, using a significance level of 10%. Panel A compares combined density forecasts schemes using the csl score function as "loss function," whereas panel B uses the log score function to compare the various combination methods. All models underlying the weighting schemes are estimated with a moving window of 750 daily returns from the S&P 500, DJIA, FTSE and Nikkei index through the period January 2000 to June 2013. The MCS results are based on 1864 (S&P 500), 1866 (DJIA), 1882 (FTSE) and 1766 (Nikkei) out-of-sample observations, respectively.

function, respectively, three Jore weighting schemes using the same three scoring rules and finally the equal weighting scheme as a benchmark. The table reports the p-value associated with each weighting scheme. Bold values represent schemes that are in the MCS, using a significance level of 10%. Panel A shows results where the csl score function serves as the loss function with κ equal to 0.15 and 0.25, respectively. Panel B shows MCS results with the log score as loss function. We do not provide results with the weighted version of the CRPS as loss function, because in that case all weighting schemes stay in the MCS. Put differently, the CRPS loss function is not able to discriminate significantly between the various schemes.

Panel A shows three interesting results. First, in the case of optimized weighting schemes, combining density forecasts with the focus on the left tail adds gains compared to combined density forecasts combined based on the whole distribution. However, this result holds only for the optimized weights based on the csl score function, not for the optimized weights based on the CRPS function. The optimized csl weights are namely always in the MCS, irrespective of κ. Moreover, the associated p-values equal 1 in seven out of eight cases. This means that the left tail of the return indexes is more accurately fitted by combining density forecasts based on the csl score function, instead of focusing on the whole distribution (i.e., log score function). Second, when comparing the optimized weighting schemes with the recursive Jore weighting schemes, the Jore-log weights are able to compete with the optimized csl weights. However, in the case of the FTSE index and $\kappa = 0.25$, the Jore-log weights are borderline significant in the MCS as the p-value equals 0.106. Third, our benchmark of equal weights is statistically outperformed in almost all cases, except for the Nikkei returns combined with $\kappa = 0.15$.

The second part of the table, Panel B, shows results when the loss function contains the whole density, that is, the log score function. We see that the performance of the optimized log score weights is superior. This result is not surprising, as using only the left tail of the density to estimate the weights ignores information about the remaining part of the density, while the log scoring rule does use this information. Nevertheless, the optimized csl weights are still always in the MCS. Surprisingly, the Jore-log weights fall outside the MCS for the DJIA, FTSE and Nikkei indexes.

All in all, this table provides a first statistical evidence that combined density forecasts in the left tail significantly improve when using weights based on the csl score function compared to using weights based on the log score function or using equal

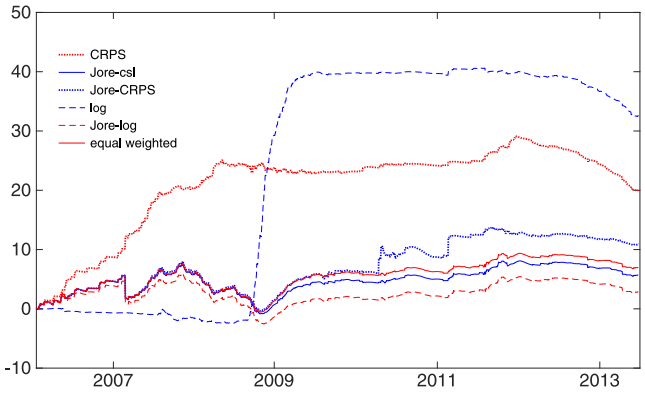

FIGURE 3 Censored likelihood scores over time. This figure depicts the cumulative sum of the difference between censored likelihood scores of combined one-step-ahead density forecasts of S&P 500 returns based on various weighting schemes. In particular, we plot the difference in cumulative optimized csl score and five competing weighting schemes: weights based on the log score or CRPS function (see Equations 8, 3 and 9 with a moving window of 750 evaluated density forecasts), the Jore weighting scheme using the three aforementioned scoring rules and finally the equal weighing scheme. For the csl scoring rule and the CRPS, we choose B_t and $u(z)$ as the left tail, that is, $y_t < \hat{r}^{0.15}$ and $u < \hat{r}^{0.15}$ with $\hat{r}^{0.15}$ the κ-th quantile of the empirical distribution of the in-sample returns [Colour figure can be viewed at wileyonlinelibrary.com]

weights. Finally, using only the left tail to forecast the whole density does not always imply less accurate density forecasts overall.

Figure 3 illustrates the evolution of the cumulative gain in the csl scores associated with our various weighting schemes over the out-of-sample period $t = \tau, \tau + 1, \ldots, M$, which is defined as

$$\sum_{t=\tau}^{M} \log \left[\sum_{i=1}^{n} w^*_{i,t-1} \left(I[y_t \in B_t] \log p_t(y_t; \mathcal{I}_{t-1}, A_i) + I[y_t \in B_t^c] \log \int_{B_t^c} p_t(y; \mathcal{I}_{t-1}, A_i) \mathrm{d}y \right) \right], \tag{22}$$

where $w^*_{i,t-1}$ is the optimized weight for model A_i at the end of trading day $t-1$, based on the evaluated density forecasts at time $t - T_w$ through $t - 1$. The graph shows the cumulative difference of the csl scores corresponding to the optimized csl weights relative to weights obtained by optimizing the log score function and the CRPS function of Equations 3 and 9, as well as the difference in csl score with respect to Jore weights with the log, csl and CRPS score function and equal weights.

The figure suggests that the gain of the optimized csl weighting scheme relative to most of the competing schemes occurs mainly at the start and end of the Global Financial Crisis (GFC). The gain relative to the optimized log score and the CRPS weights are substantial. Our main competing weighting scheme is the Jore-log scheme, as also noted by Table 1. We outperform this scheme mainly during the years 2007 and 2009.

4.4.2 | Economic results

Table 2 sheds light on the economic impact of combining density forecasts with the focus on the left tail in the context of VaR estimates. We consider 99% VaR estimates, that is, $q = 0.01$ in Equation 20. For each stock index, we compare the frequency and independence of the VaR violations corresponding to the specific weighting scheme. In addition, we report the p-value that corresponds to the dynamic quantile test.[12]

Table 2 leads first to the main conclusion that VaR estimates based on the optimized csl scoring are superior against VaR estimates based on any other considered optimized, recursive or equal weighting scheme. Optimizing the csl score function leads to superior VaR estimates, as it passes the conditional coverage test as well as the DQ test for all indexes, irrespective of κ. This means a correct number of violations (1%), which in turn do not come in clusters. In addition, the one-step-ahead VaR estimates at time t are not predictable by any other information at the same time period. Most of the competitive weighting schemes perform well in the VaR estimates of US and UK indexes, but they fail in the case of the NIKKEI index returns.

Second, we also compute asymmetric tick losses of Equation 21 for all weighting schemes and compare them by means of the MCS. The results are not shown in the table, as all p-values are above 10%. Hence there is no statistical difference between the

[12]We follow Engle and Manganelli (2004) and statistically test whether there is correlation between the hit series, defined as $(y_t < VaR_t^{1-q}) - q$, and the VaR estimates or four lags of the hit series.

TABLE 2 Evaluation of 1-day-ahead 99% VaR estimates

Weight scheme	$V(\%)$	p_{cc}	p_{DQ}	$V(\%)$	p_{cc}	p_{DQ}
	S&P 500			DJIA		
csl(0.15)	25 (1.34)	**0.261**	**0.175**	22 (1.18)	**0.580**	**0.339**
csl(0.25)	25 (1.34)	**0.261**	**0.093**	21 (1.13)	**0.687**	**0.343**
CRPS(0.15)	31 (1.67)	0.018	0.037	28 (1.50)	**0.084**	**0.078**
CRPS(0.25)	28 (1.51)	**0.082**	**0.190**	26 (1.40)	**0.188**	**0.286**
Jore-csl(15)	20 (1.08)	**0.764**	**0.225**	15 (0.81)	**0.610**	**0.690**
Jore-csl(25)	20 (1.08)	**0.764**	**0.227**	16 (0.86)	**0.722**	**0.657**
Jore-CRPS(15)	28 (1.51)	**0.082**	**0.096**	23 (1.24)	**0.468**	**0.152**
Jore-CRPS(25)	22 (1.18)	**0.571**	**0.267**	15 (0.81)	**0.610**	**0.680**
Jore-log	20 (1.08)	**0.764**	**0.227**	15 (0.81)	**0.610**	**0.690**
log	31 (1.67)	0.018	0.001	25 (1.34)	**0.266**	0.007
EQW	21 (1.13)	**0.677**	**0.230**	15 (0.81)	**0.610**	**0.690**
	FTSE			Nikkei		
csl(0.15)	28 (1.49)	**0.090**	**0.091**	17 (0.96)	**0.359**	**0.187**
csl(0.25)	23 (1.22)	**0.481**	**0.290**	18 (1.02)	**0.399**	**0.236**
CRPS(0.15)	23 (1.22)	**0.481**	**0.334**	11 (0.62)	0.038	0.006
CRPS(0.25)	25 (1.33)	**0.278**	**0.195**	14 (0.79)	**0.170**	0.007
Jore-csl(15)	26 (1.38)	**0.198**	**0.103**	11 (0.62)	0.038	0.005
Jore-csl(25)	24 (1.28)	**0.374**	**0.203**	11 (0.62)	0.038	0.005
Jore-CRPS(15)	27 (1.44)	**0.136**	0.038	16 (0.91)	**0.302**	**0.108**
Jore-CRPS(25)	23 (1.22)	**0.481**	**0.238**	12 (0.68)	**0.069**	0.013
Jore-log	26 (1.38)	**0.198**	**0.106**	11 (0.62)	0.038	0.005
log	20 (1.06)	**0.775**	**0.829**	19 (1.08)	**0.416**	**0.053**
EQW	24 (1.28)	**0.374**	**0.204**	11 (0.62)	0.038	0.005

Note. This table provides the accuracy of 1-day-ahead 99% VaR estimates of the daily return of the S&P 500, DJIA, FTSE and Nikkei index, obtained by combining density forecasts by various weighting schemes. These schemes are based on optimizing functions based on the csl scoring rule of Equation 5 (with $\kappa = 0.15$ and 0.25), the log scoring rule of Equation 1 or the CRPS function of Equation 6. In addition, we benchmark against equal weights and compute also recursive Jore weights of Equation 4 or the three scoring rules above, with a training sample of 250 observations. The columns represent the number of violations and corresponding percentage in parentheses and the p-values of the conditional coverage (CC) test and the dynamic quantile (DQ) test. Bold numbers represent those models which have a p-value associated with the CC and DQ test above 5%. The number of estimated VaRs for each series is equal to 1,864 (S&P 500), 1,866 (DJIA), 1,882(FTSE) and 1,766 (Nikkei), respectively.

various weighting schemes with respect to the asymmetric tick loss function. Note that this loss function should be evaluated with care, as it punishes VaR exceedances more than non-exceedances. This implies that if a particular weighting scheme produces in the limit no violations at all, its associated asymmetric tick loss is on average lower than a weighting scheme that produces the correct 1% number of violations.

Third, and finally, let us relate the economic results to the statistical results of Table 1: We see a coherence in the sense that the optimized csl weights perform well, both statistically and economically. In contrast to this, the recursive Jore weights with the log score function lead statistically to good density forecasts (in the left tail); however, it does not produce superior VaR estimates for all the stock indexes. Another remarkable result is that the benchmark of equal weights does not perform well statistically; however, the VaR estimates are correct except for the NIKKEI index. This indicates that there could be a difference between performing well on predicting (the shape of) the whole left tail as tested statistically in the previous subsection, and predicting one specific point of the left tail, as done in estimating the 99% VaR.

To summarize, one-step-ahead VaR estimates based on combining density forecasts by optimizing the csl scoring rule improve compared to using equal weights, the log score function and the CRPS function, either optimized or adopted to the recursive Jore weighting scheme. This improvement is with respect to the nominal size and independence of the VaR violations and to the unpredictability of the VaR exceedances.

5 | CONCLUSION

We investigate the benefits of combining density forecasts with weights based on a specific region of interest. We develop a new density forecast combination scheme based on the censored likelihood scoring rule (Diks et al., 2011) and the CRPS function (Gneiting & Ranjan, 2011). Using daily returns on the S&P 500, DJIA, FTSE and Nikkei stock market indexes from 2000 until 2013, we apply our technique on recently developed univariate volatility models, including the GAS, HEAVY and realized GARCH models.

Our results show that density forecasts in the tail are statistically more accurate if one pools density forecasts based on optimizing the censored likelihood scoring rule rather than optimizing the log score rule, the CRPS or using equal weights. Further, the optimized csl weights compete with the recursive weighting scheme of Jore et al. (2010) combined with the log scoring rule.

Second, we show that 1-day VaR estimates based on the censored likelihood scoring rule are superior compared to VaR estimates based on combined density forecasts using equal weights, the CRPS function or the log score function. This improvement is with respect to the nominal frequency and independence of VaR violations and the unpredictability of the VaR exceedances by past information. Our results imply that risk managers and portfolio managers might benefit from combining density forecasts with the focus on the left tail using the csl scoring rule.

ACKNOWLEDGMENTS

An earlier version of this paper was titled "Improving density forecasts and Value-at-Risk estimates by combining densities." We appreciate the comments of participants at the 34th International Symposium on Forecasting (Rotterdam, July 2014), the 22nd Annual Society for Nonlinear Dynamics and Econometrics (SNDE) Symposium (New York, April 2014), the SoFiE Workshop on Skewness, Heavy Tails, Market Crashes and Dynamics (Cambridge, April 2014) and the 4th Amsterdam–Bonn Workshop in Econometrics (Amsterdam, October 2013), as well as seminar participants at the University of Nottingham. Michel van der Wel is grateful to the Netherlands Organisation for Scientific Research (NWO) for a Veni grant; and for support from CREATES, funded by the Danish National Research Foundation. We are responsible for all errors.

REFERENCES

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., & Thorsrud, L. A. (2014). Nowcasting GDP in real-time: A density combination approach. *Journal of Business and Economic Statistics*, *32*, 48–68.

Aastveit, K. A., Ravazzolo, F., & Van Dijk, H. K. (2016). Combined density nowcasting in an uncertain economic environment. *Journal of Business and Economic Statistics*. Advance online publication. https://doi.org/10.1080/07350015.2015.1137760

Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, *25*, 177–190.

Andersen, T., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*, 579–625.

Bacharach, J. (1974). *Bayesian Dialogues (Working Paper)*. Oxford, UK: Oxford University.

Bao, Y., Lee, T.-H., & Saltoğlu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, *26*, 203–225.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, *76*, 1481–1536.

Billio, M., Casarin, R., Ravazzolo, F., & van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, *177*, 213–232.

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

Boucher, C. M., Daníelsson, J., Kouontchou, P. S., & Maillet, B. B. (2014). Risk models-at-risk. *Journal of Banking and Finance*, *44*, 72–92.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, *39*, 841–862.

Conflitti, C., De Mol, C., & Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, *31*, 1096–1103.

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

Daníelsson, J., James, K. R., Valenzuela, M., & Zer, I. (2016). Model risk of risk models. *Journal of Financial Stability*, *23*, 79–91.

Del Negro, M., Hasegawa, R. B., & Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, *192*, 391–405.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*, 863–883.

Diks, C., Panchenko, V., & van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*, 215–230.

Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive Value-at-Risk by regression quantiles. *Journal of Business and Economic Statistics*, *22*, 367–381.

Garratt, A., Mitchell, J., Vahey, S. P., & Wakerly, E. C. (2011). Real-time inflation forecast densities from ensemble Phillips curves. *North American Journal of Economics and Finance*, *22*, 77–87.

Geweke, J., & Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, *164*, 130–141.

Giacomini, R., & White, H. L. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.

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

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, *29*, 411–422.

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782.

Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *Journal of Forecasting*, *23*, 1–13.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, *35*, 705–730.

Hansen, P. R., Huang, Z., & Shek, H. H. (2012). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, *27*, 877–906.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, *20*, 873–889.

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

Jore, A. S., Mitchell, J., & Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, *25*, 621–634.

Kapetanios, G., Mitchell, J., Price, S., & Fawcett, N. (2015). Generalised density forecast combinations. *Journal of Econometrics*, *188*, 150–165.

Mitchell, J., & Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR fan charts of inflation. *Oxford Bulletin of Economics and Statistics*, *67*, 995–1033.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*, 347–370.

Pettenuzzo, D., & Ravazzolo, F. (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, *31*, 1312–1332.

Shephard, N., & Sheppard, K. (2010). Realising the future: Forecasting with high- frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics*, *25*, 197–231.

Timmermann, A. (2006). Forecast Combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Amsterdam, Netherlands: Elsevier.

Waggoner, D. F., & Zha, T. (2012). Confronting model misspecification in macroeconomics. *Journal of Econometrics*, *171*, 167–184.

How to cite this article: Opschoor A, van Dijk D, van der Wel M. Combining density forecasts using focused scoring Rules. *J Appl Econ*. 2017;0:1–16. https://doi.org/10.1002/jae.2575

APPENDIX: OPTIMIZING WEIGHTS

We follow Conflitti et al. (2015) to optimize the weights according to the log or csl score function of Equations 3 and 8, respectively. We provide here only an outline of the algorithm.

Define $\mathbf{p}(y_{t+1})$ as the vector of n density forecasts $p_i(y_{t+1}) = p_{t+1}(y_{t+1}; \mathcal{I}_t, A_i)$ $(i = 1, \ldots n)$ of the variable y_{t+1} at time t. The combined density is then equal to

$$p(y_{t+1}) = \mathbf{w}'\mathbf{p}(y_{t+1}) = \sum_{i=1}^{n} w_i p_i(y_{t+1}), \tag{A1}$$

with the assumption that the weights are non-negative and add up to one. For both scoring rules, we have to maximize the logarithm of the combined (censored) density over a given time period:

$$\Phi(\mathbf{w}) = \frac{1}{T-1} \sum_{t=1}^{T-1} \log p(y_{t+1}). \tag{A2}$$

Note that we omitted the factor $\frac{1}{T-1}$ in Equations 5 and 1. This does not change the result as it is a constant. Define the $(T-1)\times n$ matrix P with non-negative elements $P_{ti} = p_i(y_{t+1})$. Now, Equation A2 can be rewritten as $\frac{1}{T-1}\sum_{t=1}^{T-1} \log(P\mathbf{w}_t)$. Denote $\mathbf{w}_{\text{opt}}$

as the maximum of $\Phi(\mathbf{w})$ subject to the weight constraints. Further, the Lagrange multiplier is introduced to take into account these constraints:

$$\Phi_\lambda(\mathbf{w}) = \frac{1}{T-1}\sum_{t=1}^{T-1}\log(P\mathbf{w}_t) - \lambda\sum_{i=1}^{N} w_i. \tag{A3}$$

Instead of optimizing Equation A3, Conflitti et al. (2015) consider the following "surrogate" function, which depends on a vector $\mathbf{a}$ of arbitrary weights:

$$\Psi_\lambda(\mathbf{w};\mathbf{a}) = \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{i=1}^{n} b_{ti}\log\left(\frac{w_i}{a_i}\sum_{l=1}^{n}\log P_{tl}a_l\right) - \lambda\sum_{i=1}^{n} w_i, \tag{A4}$$

with $b_{ti} = \frac{P_{ti}a_i}{\sum_{l=1}^{n} P_{tl}a_l}$. Further, the function has the properties $\Psi_\lambda(\mathbf{a};\mathbf{a}) = \Psi_\lambda(\mathbf{a})$ for any $\mathbf{a}$ and $\Psi_\lambda(\mathbf{w};\mathbf{a}) \leqslant \Psi_\lambda(\mathbf{w})$ for any $\mathbf{a}$ and $\mathbf{w}$.

The iterative algorithm is defined as

$$\mathbf{w}_\lambda^{(k+1)} = \arg\max_{w} \Psi_\lambda(\mathbf{w};\mathbf{w}_\lambda^{(k)}), \tag{A5}$$

which yields a monotonic increase of Ψ_λ, according to the two aforementioned properties. Setting the derivatives of $\Psi_\lambda(\mathbf{w};\mathbf{w}_\lambda^{(k)})$ with respect to w_i equal to zero leads to the maximum $w_{\lambda,i} = (1/\lambda)\sum_{t=1}^{T-1} b_{ti}$. Using the constraint that the weights should sum to one, it holds that $\lambda = T-1$. This changes Equation A5 into

$$w_i^{(k+1)} = w_i^{(k)}\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{P_{ti}}{\sum_{l=1}^{n} P_{tl}w_l^{(k)}}, \tag{A6}$$

where we replace a_i by $w_i^{(k)}$ in the expression b_{ti}. We start the algorithm with equal weights, that is, $w_i^0 = 1/n$, and use as a stopping criterion a tolerance of 10^{-6} of the sum of the absolute deviation of two successive iterations.