International Journal of Forecasting 36 (2020) 1057–1072

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting risk measures using intraday data in a generalized autoregressive score framework☆

Emese Lazar *, Xiaohan Xue

ICMA Centre, Henley Business School, University of Reading, Reading, RG6 6BA, UK

ARTICLE INFO

Keywords:
Value at risk
Expected shortfall
Generalized autoregressive score dynamics
Realized measures
Intraday data
Risk forecasting

ABSTRACT

A new framework for the joint estimation and forecasting of dynamic value at risk (VaR) and expected shortfall (ES) is proposed by our incorporating intraday information into a generalized autoregressive score (GAS) model introduced by Patton et al., 2019 to estimate risk measures in a quantile regression set-up. We consider four intraday measures: the realized volatility at 5-min and 10-min sampling frequencies, and the overnight return incorporated into these two realized volatilities. In a forecasting study, the set of newly proposed semiparametric models are applied to four international stock market indices (S&P 500, Dow Jones Industrial Average, Nikkei 225 and FTSE 100) and are compared with a range of parametric, nonparametric and semiparametric models, including historical simulations, generalized autoregressive conditional heteroscedasticity (GARCH) models and the original GAS models. VaR and ES forecasts are backtested individually, and the joint loss function is used for comparisons. Our results show that GAS models, enhanced with the realized volatility measures, outperform the benchmark models consistently across all indices and various probability levels.

1. Introduction

From the perspective of financial risk managers, a risk measure can be considered a map from the space of probability distributions to real numbers. Risk measures can provide banks and financial institutions with specific values of potential losses so that risk managers can adjust their capital reserves against the downside risk. Value at risk (VaR) and expected shortfall (ES) are two prevailing measures of financial risk that dominate contemporary financial regulation. VaR provides banks and investment institutions with a loss level that occurs in the worst situation at a given confidence level, and it can be defined as

$$VaR_t^{\alpha} \equiv \inf\{y_t \in \mathbb{R} | F_Y(y_t|\mathscr{F}_{t-1}) \geq \alpha\},$$

where $F_Y(\cdot|\mathscr{F}_{t-1})$ is the cumulative distribution function of asset returns y_t over a horizon given the information set $\mathscr{F}_{t-1}$, and $\alpha \in (0, 1)$ is a given significance level. As a quantile, VaR can be expressed directly in terms of the inverse cumulative distribution function, $VaR_t^{\alpha} = F_Y^{-1}(\alpha|\mathscr{F}_{t-1})$, and as a risk measure, it has the advantage of being intuitive and easily understood.

However, VaR has inherent deficiencies as it ignores the shape and structure of the tail and is not a coherent risk measure in the sense of Artzner, Delbaen, Eber, and Heath (1999). Thus, after the financial crisis of 2007–2008, the Basel Committee on Banking Supervision (2013) proposed a transition from VaR with a confidence level of 99% to ES with a confidence level of 97.5%. ES is the

☆ For insightful and constructive comments, we thank Tim Bollerslev, Isabel Casas, Mike Clements, Zhonghao Fu, Xiaochun Meng, James Taylor, Shixuan Wang, Cheng Yan and seminar participants at the 12th Annual SoFiE Conference, Shanghai, the 2019 Asian Meeting of the Econometrics Society, Xiamen, and EEA-ESEM 2019, Manchester.

* Corresponding author.
E-mail addresses: e.lazar@icmacentre.ac.uk (E. Lazar), x.xue@pgr.reading.ac.uk (X. Xue).

https://doi.org/10.1016/j.ijforecast.2019.10.007

expectation of returns, conditional on its realization lying below VaR, and it can be defined as

$$ES_t^\alpha \equiv \mathbb{E}[y_t | y_t \leq VaR_t^\alpha, \mathscr{F}_{t-1}].$$

ES is a coherent risk measure (Roccioletti, 2015), and it has been suggested as an alternative to VaR in risk management applications because of its superior mathematical properties.

Normally, ES is estimated via a two-stage approach based on VaR estimation. Although ES is itself not elicitable, Fissler, Ziegel, and Gneiting (2016) have shown that the pair $(VaR_t^\alpha, ES_t^\alpha)$ is elicitable (see also Acerbi & Székely, 2014). This means that ES can be estimated jointly with VaR by minimizing a loss function (Fissler & Ziegel, 2016; Ziegel, 2016).

Following the classification of Engle and Manganelli (2004), models in the current literature on estimating and forecasting risk measures can be divided into three main categories: parametric, nonparametric and semiparametric models. Previous studies using parametric models to predict VaR and ES assumed that financial returns follow a certain distribution, such as the standard normal (Gaussian) distribution. In reality, however, it is hardly reasonable to make such strong assumptions. Nonparametric models do not make assumptions about the distribution of financial returns, and have the advantage of being model-free. Although it is not necessary for such models to make a distributional assumption, an inherent problem is the difficulty in finding the optimal size of the estimation window (Engle & Manganelli, 2004). Semiparametric models impose a parametric structure on the dynamics of VaR and ES through their relationship with lagged information, but require no assumptions about the conditional distribution of financial returns (Patton, Ziegel, & Chen, 2019).

Quantile regression, as an approach for estimating risk measures, has only recently been considered: Engle and Manganelli (2004) extended the basic quantile regression model to conditional autoregressive VaR (CAViaR) models; these models focus solely on the estimation of VaR, and it is not obvious how they can be used for ES estimation. To estimate ES jointly with VaR in a semiparametric framework, Taylor (2008) proposed conditional autoregressive expectile (CARE) models, which are based on a simple function of expectiles.[1] Following this, Taylor (2019) synthesized the quantile regression with the maximum likelihood estimation based on an asymmetric Laplace density proposed by Koenker and Machado (1999) and estimated VaR and ES jointly. A growing literature documents a significant improvement in VaR and ES estimation in a quantile regression framework (Bayer, 2018; Halbleib & Pohlmeier, 2012; Wang & Zhao, 2016; Žikeš & Baruník, 2014).

Following the results of Fissler and Ziegel (2016), Patton et al. (2019) presented several novel dynamic models for the joint estimation of VaR and ES. Specifically, they proposed four dynamic semiparametric models for VaR and ES based on the generalized autoregressive score (GAS) framework introduced by Creal, Koopman, and Lucas (2013). This model has been successfully applied in risk measure estimation (Patton, Ziegel, & Chen, 2019), credit default swap spread modelling (Lange, Lucas, & Siegmann, 2017; Oh & Patton, 2018), systemic risk modelling (Bernardi & Catania, 2019; Cerrato, Crosby, Kim, & Zhao, 2017; Eckernkemper, 2017) and high-frequency data modelling (Gorgi, Hansen, Janus, & Koopman, 2018; Lucas & Opschoor, 2018).[2] However, no studies on risk measures incorporating realized volatilities into the GAS framework have been considered so far.[3] This prompted the research question of this article, namely whether adding intraday measures of volatility into the GAS framework increases the accuracy of joint VaR and ES forecasts.

The question whether intraday data can increase the predictive accuracy of risk measures has already been addressed by academics.[4] Several studies extended quantile regression methods and other semiparametric models by using information variables generated from high-frequency data.[5] Many realized volatility measures have been confirmed to perform efficiently. The realized volatility proposed by Andersen and Bollerslev (1998) and Alizadeh, Brandt, and Diebold (2002) is one of the most widely used intraday volatility measures. Inspired by Engle and Manganelli (2004), Fuertes and Olmo (2013) proposed a conditional quantile forecast method combining an effective device to deal with the interday/intraday information. Meng and Taylor (2018) extended the CAViaR model and the quantile regression heterogeneous autoregressive model (HAR) model with realized volatility, overnight return and intraday range. In terms of ES estimation, the CARE models of Taylor (2008) have been extended to allow intraday measures as explanatory variables (Gerlach & Chen, 2014, 2017; Gerlach & Wang, 2016a; Wang, Gerlach, & Chen, 2018).

Although the improvement from adding intraday variables into a semiparametric framework has been widely documented, evidence for using the score-driven model as the framework to estimate risk measures still remains hard to come by. Therefore, in our study, the first contribution is that we extend the set of semiparametric GAS models of Patton et al. (2019) (the two-factor GAS (GAS-2F) model, the one-factor GAS (GAS-1F) model, the GARCH-FZ model and the hybrid GAS/GARCH model) to investigate whether realized measures can increase the predictive accuracy of GAS models. This study is the first one to estimate and forecast VaR and ES jointly by using intraday data in a GAS framework. We shed light on the potential improvement in risk forecasting from adding

[1] The connection between quantiles, expectiles and ES is originally found in Aigner, Amemiya, and Poirier (1976), and was considered further by Newey and Powell (1987).

[2] More studies related to the GAS model can be found at http://www.gasmodel.com/.

[3] Salvatierra and Patton (2015) use measures of realized covariance to build forecasts for copula models.

[4] Both parametric models (see Giot & Laurent, 2004; Hansen, Huang, & Shek, 2012; Louzis, Xanthopoulos-Sisinis, & Refenes, 2014) and semiparametric models (see Clements, Galvão, & Kim, 2008; Fuertes & Olmo, 2013; Gerlach & Wang, 2016b; Žikeš & Baruník, 2014).

[5] See Clements et al. (2008), Fuertes and Olmo (2013), Žikeš and Baruník (2014), Gerlach and Chen (2014, 2017), Gerlach and Wang (2016a).

intraday information in the GAS framework for four stock indices using a long forecasting period (which includes the financial crisis period). Then we perform a thorough analysis to compare our forecasts with those generated from prevailing benchmarks in the current literature. Our results show that incorporating intraday data into the GAS framework results in the forecasts outperforming other (VaR, ES) forecasts in most cases.

Our second contribution to the literature is that we provide empirical evidence that semiparametric models enhanced with realized volatility measures outperform other benchmark models via various backtesting methods. Our proposed models, especially the GAS-2F model, extended with realized volatilities dominate other benchmarks consistently. Thirdly, we compare four different types of realized measures with regard to their forecasting ability for risk measures when added to GAS models.

This article is structured as follows: Section 2 briefly introduces the new GAS models that incorporate intraday information; the data used in our empirical study and the in-sample estimation results are presented in Section 3; Section 4 presents the forecasting study and backtesting results; and Section 5 concludes the article.

2. Models

2.1. GAS models for VaR and ES

Several extensions of the GAS models introduced by Creal et al. (2013) are proposed in Patton et al. (2019), and can be estimated by minimizing the loss function of Fissler and Ziegel (2016) called FZ0:

$$L_{FZ0}(Y, v, e; \alpha) = -\frac{1}{\alpha e}\mathbf{1}\{Y \le v\}(v - Y) + \frac{v}{e} + \log(-e) - 1, \tag{1}$$

where Y denotes the daily return, v and e represent the values of VaR and ES, respectively, and $\mathbf{1}$ is an indicator function which returns 1 when $Y \le v$ (i.e. the VaR is exceeded) and otherwise returns zero. Patton et al. (2019) propose four models (the GAS-2F model, the GAS-1F model, the GARCH-FZ model and the hybrid GAS/GARCH model) to estimate VaR and ES jointly by minimizing the loss function FZ0. The key novelty in their framework is the use of the scaled score (which can be computed as the first-order derivative of the objective function[6]) to drive the time variation in the target parameter. Patton et al. (2019) present a "news impact curve" to show the impact of past observations on current forecasts of VaR and ES through the score variable. When $Y > v$, the realized returns do not affect the estimation. But when $Y \le v$, forecasts of ES and VaR react to realized returns through the score variable. The GAS-FZ models are specified as below:

- GAS-1F model:

$$\begin{aligned} v_t &= a\exp\{\kappa_t\}, \\ e_t &= b\exp\{\kappa_t\}, \quad b < a < 0, \\ \kappa_t &= \omega + \beta\kappa_{t-1} + \gamma H_{t-1}^{-1} s_{t-1}, \end{aligned} \tag{2}$$

where the score variable s_t is defined as

$$\begin{aligned} s_t &\equiv \frac{\partial L_{FZ0}(Y_t, a\exp\{\kappa_t\}, b\exp\{\kappa_t\}; \alpha)}{\partial \kappa} \\ &= -\frac{1}{e_t}\left(\frac{1}{\alpha}\mathbf{1}\{Y_t \le v_t\}Y_t - e_t\right), \end{aligned} \tag{3}$$

and the Hessian factor H_t is set to 1 for simplicity.

- GAS-2F model:

$$\begin{bmatrix} v_t \\ e_t \end{bmatrix} = \mathbf{w} + \mathbf{B}\begin{bmatrix} v_{t-1} \\ e_{t-1} \end{bmatrix} + \mathbf{A}\begin{bmatrix} \lambda_{v,t-1} \\ \lambda_{e,t-1} \end{bmatrix}, \tag{4}$$

where $\mathbf{w}$ is a (2×1) vector, $\mathbf{A}$ is a (2×2) matrix, $\mathbf{B}$ is defined as a diagonal matrix for parsimony and

$$\lambda_{v,t} \equiv -v_t(\mathbf{1}\{Y_t \le v_t\} - \alpha), \tag{5}$$

$$\lambda_{e,t} \equiv \frac{1}{\alpha}\mathbf{1}\{Y_t \le v_t\}Y_t - e_t. \tag{6}$$

- GARCH-FZ model:

$$\begin{aligned} v_t &= a \cdot \sigma_t, \\ e_t &= b \cdot \sigma_t, \quad b < a < 0, \\ \sigma_t^2 &= \omega + \beta\sigma_{t-1}^2 + \gamma Y_{t-1}^2, \end{aligned} \tag{7}$$

where σ_t^2 is the conditional variance and is assumed to follow a GARCH(1,1) process. The parameters of this model are estimated by minimizing the loss function FZ0 in (1) instead of using Quasi Maximum Likelihood Estimation.

- Hybrid GAS/GARCH model (hybrid):

$$\begin{aligned} v_t &= a\exp\{\kappa_t\}, \\ e_t &= b\exp\{\kappa_t\}, \quad b < a < 0, \\ \kappa_t &= \omega + \beta\kappa_{t-1} \\ &\quad + \gamma\left(-\frac{1}{e_{t-1}}\left(\frac{1}{\alpha}\mathbf{1}\{Y_t \le v_t\}Y_{t-1} - e_{t-1}\right)\right) \\ &\quad + \delta\log|Y_{t-1}|, \end{aligned} \tag{8}$$

where the variable κ_t is the log volatility, described by the one-day-lagged log volatility, score factor and the logarithm of absolute return.

2.2. Realized measures

This section provides a brief introduction to various intraday realized measures used in this study. The most popular measure is the realized volatility, defined as

$$\begin{aligned} RV\Delta_t &= \sqrt{\sum_{i=1}^{M}(P_{t,i\cdot\Delta} - P_{t,(i-1)\cdot\Delta})^2}, \\ \Delta &= \frac{S}{M}, \end{aligned} \tag{9}$$

where $RV\Delta_t$ denotes the realized volatility calculated from the sum of M intraday squared returns at frequency Δ within day t. Here the intraday frequency Δ divides the whole span of market opening hours S into M equal intervals, and $P_{t,i\cdot\Delta}$ denotes the log price at time $i\cdot\Delta$ of day

[6] Normally, the objective function is a probability density function, but here the loss function FZ0 acts as the objective function.

t. However, the realized volatility ignores the information from the market overnight return, which is defined as

$$overnight_t = \log(P_{t,0}) - \log(P_{t-1,S}), \tag{10}$$

where $P_{t,0}$ and $P_{t-1,S}$ denote the opening price on day t and the closing price on the previous day, respectively. Several studies have proven that incorporating the overnight return can lead to a more accurate realized measure. In this article, we consider the approach of incorporating the overnight return in the realized volatility of Blair, Poon, and Taylor (2001), Hua and Manzan (2013) and Meng and Taylor (2018) as follows:

$$RN\Delta_t = \sqrt{RV\Delta_t^2 + (overnight_t)^2}. \tag{11}$$

In the following, we will use frequencies of $\Delta = 5$ min and $\Delta = 10$ min. As such, in the next section, RM can signify any of the four realized measures of volatility $RV5_t$, $RV10_t$, $RN5_t$ and $RN10_t$, and we extend the models with these measures.

2.3. GAS models for VaR and ES with realized measures

Salvatierra and Patton (2015) propose a GAS model enhanced with high-frequency measures to obtain a Generalized Realized Autoregressive Score (GRAS) model, which has the equation for the dependence parameter, similar to the last row of (2), replaced with

$$\kappa_t = \omega + \beta\kappa_{t-1} + \gamma H_{t-1}^{-1} s_{t-1} + c\log(RM_{t-1}). \tag{12}$$

They use the realized covariance as RM_t, computed from the intraday prices $P_{t,i\cdot\Delta}$ of a set of assets. They find that the inclusion of 5-min realized covariance significantly improves the in-sample fit and out-of-sample forecasts of the copula models.

Motivated by the set of GAS models and the GRAS model, our new models are proposed as follows:

- GAS-1F model with realized measures (GAS-1F-Re):

$$\begin{aligned} v_t &= a\exp\{\kappa_t\}, \\ e_t &= b\exp\{\kappa_t\}, \quad b < a < 0, \end{aligned} \tag{13}$$

  where κ_t is defined in (12), and the score variable s_t is defined in (3). Here the Hessian factor H_t is set to 1 for simplicity; $\log(RM_t)$ is the logarithm of a realized measure, which can be the realized volatility at 5-min and 10-min sampling frequencies ($RV5$ and $RV10$), and these two realized volatilities with the overnight return incorporated into them ($RN5$ and $RN10$), as defined in Section 2.2.
- GAS-2F model with realized measures (GAS-2F-Re):

$$\begin{bmatrix} v_t \\ e_t \end{bmatrix} = \mathbf{w} + \mathbf{B}\begin{bmatrix} v_{t-1} \\ e_{t-1} \end{bmatrix} + \mathbf{A}\begin{bmatrix} \lambda_{v,t-1} \\ \lambda_{e,t-1} \end{bmatrix} + \mathbf{C}\,RM_{t-1}, \tag{14}$$

  where $\mathbf{w}$ and $\mathbf{C}$ are (2×1) vectors, $\mathbf{A}$ and $\mathbf{B}$ are both (2×2) matrices and $\mathbf{B}$ is defined as a diagonal matrix to simplify computation. Following Patton et al. (2019), we also define the forcing variables $\lambda_{v,t}$ and $\lambda_{e,t}$ as the partial derivatives of the given loss function L_{FZ0} with respect to v_t and e_t, as in (5) and (6).

  Hansen et al. (2012) and Hansen, Lunde, and Voev (2014) introduced a new framework, realized (beta) GARCH, where the variance follows a GARCH(1,1) process, with the squared returns replaced with a realized measure of volatility. Following this model, we propose the following model:
- GARCH-FZ model with realized measures (GARCH-FZ-Re):

$$\begin{aligned} v_t &= a \cdot \sigma_t, \\ e_t &= b \cdot \sigma_t, \quad b < a < 0, \\ \sigma_t^2 &= \omega + \beta\sigma_{t-1}^2 + cRM_{t-1}^2, \end{aligned} \tag{15}$$

  where the daily return Y_{t-1} in the GARCH(1,1) variance equation in (7) is replaced with the realized measure RM_{t-1}. This model is estimated by minimizing the FZ0 loss function.
- Hybrid GAS/GARCH model with realized measures (hybrid-Re):

$$\begin{aligned} v_t &= a\exp\{\kappa_t\}, \\ e_t &= b\exp\{\kappa_t\}, \quad b < a < 0, \\ \kappa_t &= \omega + \beta\kappa_{t-1} + \gamma\left(-\frac{1}{e_{t-1}}\left(\frac{1}{\alpha}\mathbf{1}\{Y_t \le v_t\}Y_{t-1} - e_{t-1}\right)\right) \\ &\quad + \delta\log|Y_{t-1}| + c\log(RM_{t-1}), \end{aligned} \tag{16}$$

  where the log volatility κ_t follows the hybrid GARCH model with one-day-lagged log volatility, score factor, realized measures and absolute daily return.

3. Data and empirical study

3.1. Data description

To evaluate the forecasting performance of the new models and to compare them with benchmark models, we collected daily opening and closing prices of four international stock market indices (S&P 500, Dow Jones Industrial Average (DJIA), Nikkei 225 and FTSE 100) from January 2000 to June 2019 from DataStream. To ensure the applicability of every model, we removed market-specific nontrading days and exactly zero returns from each index series. Part A of Table 1 presents the summary statistics on the four daily equity return series over the full sample period. From the top part of part A of Table 1, the average annualized returns range from 0.544% for the Nikkei 225 to 4.377% for the DJIA, and the annualized standard deviation ranges from 18% for the DJIA to about 24% for the Nikkei 225. All daily return series exhibit substantial kurtosis of around 10. The second and third parts in part A of Table 1 show the sample VaR and ES for four different α values: 1%, 2.5%, 5% and 10%. The Nikkei 225 proves to be different from the other indices since its quantile and ES are lower than the sample risk measures of the other three indices.

Table 1
Summary statistics and marginal distribution estimates.

	S&P 500	DJIA	Nikkei 225	FTSE 100
Part A: Summary statistics				
Mean (annualized)	3.685	4.377	0.544	0.606
Std dev (annualized)	18.900	17.821	23.748	18.105
Skewness	−0.208	−0.125	−0.429	−0.170
Kurtosis	11.176	10.980	9.341	9.487
VaR-0.01	−3.427	−3.294	−4.111	−3.264
VaR-0.025	−2.525	−2.361	−3.051	−2.409
VaR-0.05	−1.885	−1.777	−2.360	−1.788
VaR-0.10	−1.284	−1.182	−1.682	−1.233
ES-0.01	−4.849	−4.568	−6.021	−4.546
ES-0.025	−3.678	−3.453	−4.492	−3.457
ES-0.05	−2.922	−2.750	−3.576	−2.764
ES-0.10	−2.236	−2.096	−2.788	−2.120
Part B: Conditional mean				
Constant	−0.001	0.007	−0.021	−0.003
AR(1)	–	–	–	–
MA(1)	−0.039	–	–	–
Part C: Conditional variance				
Constant	0.010	0.010	0.025	0.014
ARCH	0.065	0.069	0.082	0.116
GARCH	0.926	0.922	0.910	0.874
Part D: Skew-t density				
DoF	9.020	8.130	12.204	22.177
Skewness	−0.092	−0.089	−0.089	−0.162

This table presents the summary statistics of the four daily equity return series studied over the full sample period from January 2000 to June 2019 and marginal distribution estimates over the in-sample period. Part A reports the annualized mean returns, standard deviation of the returns as percentages, skewness, kurtosis and the sample VaR and ES estimates for four choices of α. Part B presents the parameter estimates for AR(m) models of the conditional means of these returns. Part C shows parameter estimates for GARCH–skew-t(1,1) models of the conditional variance. Part D presents parameter estimates for the skew-t density for the standardized residuals.

Part B of Table 1 presents the estimated parameters of the ARMA(p, q) models (where "ARMA" means "autoregressive moving average") where the lags (p, q) are optimally selected via the Bayesian information criterion (BIC) method. The ARMA models for the indices include only a constant except for the S&P 500, which contains a moving average (MA) term with one lag. Part C of Table 1 shows the estimated parameters of the GARCH(1,1) model, where the residuals are assumed to follow the skew-t distribution. Part D of Table 1 presents the degrees of freedom and skewness in the skew-t distribution.

The percentage log overnight returns are generated as in (10). For the realized volatility, the data are obtained at 5-min and 10-min sampling frequencies from the Oxford-Man Institute of Quantitative Finance's realized library[7](see Heber, Lunde, & Shephard, 2009). To generate the new realized measure incorporating the overnight return in realized volatility, we use (11).

The entire sample is divided into an in-sample for estimation and an out-of-sample to backtest the estimated results. We use a rolling window approach, where each model is re-estimated every five trading days using a rolling window of 2000 observations. Then the rest of the period until June 2019 of approximately 2900 days is the out-of-sample period to evaluate one-day-ahead VaR and ES estimates.

[7] This realized library can be accessed at https://realized.oxford-man.ox.ac.uk/.

3.2. Forecasting models

VaR and ES are predicted via the score forecast for one trading day ahead in the out-of-sample period for each series with use of the proposed GAS-realized models and the GARCH-FZ-Re model, as well as nonparametric models and parametric models as benchmarks. For nonparametric models, historical simulations are widely used because of their advantages of being model-free and easy to implement. In our study, we select three commonly used rolling window sizes to forecast VaR and ES: 125, 250 and 500 days. Two popular GARCH models are used in this study, the Gaussian (GARCH-G) and skew-t (GARCH-Skt) models, as parametric model benchmarks. We also consider other established models that use high-frequency data that are considered to be well suited to forecast VaR and ES: the HAR model of Corsi, Mittnik, Pigorsch, and Pigorsch (2008) and the HEAVY model of Shephard and Sheppard (2010). In each model, we estimate VaR and ES with Gaussian and skew-t distributions of the errors in the second step, after the conditional volatility estimation. We also take the semiparametric model of Taylor (2019) based on the asymmetric Laplace distribution into our benchmark set.

To evaluate the performance of the GAS models enhanced with realized measures, we also implement the four models proposed by Patton et al. (2019) as benchmarks. Differently from Patton et al. (2019), who used certain parameters estimated from a fixed in-sample period,

we use a rolling window approach, where each model is re-estimated every five trading days using a window of size 2000 trading days. In this study, we consider four sets of GAS models extended with different realized measures ($RV5$, $RV10$, $RN5$ and $RN10$) as in Section 2.2. In the following section, we provide estimation results obtained with these proposed models.

3.3. In-sample estimation

The parameters of the GAS models and the proposed four sets of GAS-realized models are estimated by our minimizing the loss function in (1). It is hard to estimate these models with a nonsmooth objective function, and this algorithm is sensitive to the starting values used in the search. We optimize the proposed models using the following procedure: for each model, we first generate 10^5 vectors of parameters from predetermined intervals randomly for the parameters of the GAS models. For example, for the parameters (a and b) used to generate VaR and ES in the GAS-1F, GARCH-FZ, and hybrid models, we set the intervals as $[-2, -3]$ and $[-3, -4]$, respectively, to ensure that ES is always less than VaR.[8] We compute the average loss for each vector, and then select the 10 vectors that generate the lowest average loss as initial values for the optimization routine. The vectors are selected as the initial values of the search algorithm for all windows so as to shorten the computational time. We compute the optimal parameters by using a quasi-Newton method and the function *fminunc* as optimization algorithms, which are routines similar to the routine used by Engle and Manganelli (2004).

Table 2 presents the estimated parameters together with their standard errors of the GAS models for the S&P 500, estimated with an estimation period of 2000 days from the beginning of January 2000 for $\alpha = 5\%$. The parameters of the three two-factor GAS models (GAS-2F, GAS-2F-RV5, and GAS-2F-RN5 models) are presented in the upper part of Table 2; we separate the parameters of VaR and ES. The b parameters are statistically significantly different from zero at both the 1% significance level and the 5% significance level for both VaR and ES,[9] which can be explained by the volatility clustering effect. Columns 4–7 in the upper part of Table 1 show the parameters of the GAS-2F model extended with the 5-min realized measures. Because of our adding 5-min realized measures, the degree of clustering decreases for VaR and ES. Also, the parameters a_v and a_e experience a significant decrease after addition of the realized measures. The parameters of the one-day-lagged realized measures RM_{t-1}, c are statistically significantly negative at the 5% significance level for both VaR and ES, indicating that larger values of these realized variables will result in a lower estimated quantile or ES, which is intuitive. The average loss generated by the GAS-2F model is 0.756, which is larger than the loss of the GAS-2F models extended with realized measures (0.735 and 0.734).

The lower part of Table 2 shows the estimated parameters of the other GAS models extended with the 5-min realized measures with an estimation period of 2000 days from the beginning of January 2000 for the S&P 500 for $\alpha = 5\%$. Similarly to the b parameters of the GAS-2F models, the β parameters of the other models are also statistically significantly different from zero at both the 1% significance level and the 5% significance level, which means that the current estimated risk measures rely heavily on the previous estimation. Also, we find that the parameters of realized measures (c for the GAS-1F model, the GARCH-FZ model and the hybrid model) are all statistically significantly positive at both the 1% significance level and the 5% significance level. Intuitively, a large realized volatility will lead to a low quantile through the score variable in these models. We find that the inclusion of realized measures in the updating models results in smaller coefficients of the GAS shocks (γ), which is intuitive. Later, we will see the role that the score variable plays in forecasting VaR and ES. In the following sections we compare the forecasting performance of these four sets of extended models, which gives a total of 16 models, with the 13 benchmark models listed above.

4. Out-of sample forecasting and backtesting

We evaluate one-day-ahead VaR and ES forecasts for the four international stock indices and for the following probability levels: 1%, 2.5%, 5% and 10%. One-day-ahead VaR and ES forecasts are made with parameter values estimated every 5 days for each model and probability level with rolling windows of size 2000 (except for historical simulations). The forecasting sample period for each index is approximately 2900 days. In this section, we backtest the VaR and ES forecasts of the proposed models and compare their performance with that of benchmark models. First, we backtest VaR and ES individually via the dynamic quantile (DQ) regression and the dynamic ES (DES). Following these tests, we use a method based on the FZ0 loss function to backtest VaR and ES jointly.

4.1. Backtesting VaR

The most popular procedures evaluating the performance of VaR forecasts are based mainly on VaR failures; that is,

$$I_t = \mathbf{1}\{Y_t \leq VaR_t^{\alpha}\}.$$

The commonly used VaR backtesting method, known as the unconditional coverage (UC) test, was proposed by Kupiec (1995) and uses the proportion of failures as its main tool. In this test, the hit percentage is defined as the proportion of the returns below the estimated VaR, and then the difference between the hit percentage and its theoretical value of α is examined. Thus, the VaR model is rejected or not rejected according to the null hypothesis

[8] For parameters in the GAS-2F models, the predetermined intervals for w, b, a_v, a_e and c are $[-0.1, 0.1]$, $[0.8, 1]$, $[-0.1, 0.1]$, $[-0.1, 0.1]$ and $[-1, 0]$, respectively. For parameters in the GAS-1F, GARCH-FZ and hybrid models, the predetermined intervals for β, γ, δ, c, a and b are $[0.8, 1]$, $[0, 0.1]$, $[0, 0.1]$, $[0, 0.5]$, $[-2, -3]$ and $[-3, -4]$, respectively.

[9] We use Student's t test for significance testing.

Table 2
Estimated parameters of the GAS models for the S&P 500 for $\alpha = 5\%$.

	GAS-2F		GAS-2F-RV5		GAS-2F-RN5	
	VaR	ES	VaR	ES	VaR	ES
w	−0.009	−0.012	−0.009	−0.016	−0.011	−0.023
(s.e.)	(0.002)	(0.003)	(0.030)	(0.053)	(0.033)	(0.045)
b	0.995	0.995	0.833	0.810	0.814	0.849
(s.e.)	(0.105)	(0.108)	(0.084)	(0.092)	(0.098)	(0.072)
a_v	−0.129	−0.140	−0.125	−0.066	−0.114	−0.118
(s.e.)	(0.070)	(0.103)	(0.304)	(0.629)	(0.416)	(0.466)
a_e	0.002	0.003	0.002	0.001	0.001	0.001
(s.e.)	(0.003)	(0.004)	(0.011)	(0.024)	(0.015)	(0.017)
c	–	–	−0.323	−0.477	−0.353	−0.360
(s.e.)	–	–	(0.148)	(0.208)	(0.190)	(0.158)
Average loss	0.756		0.735		0.733	

	GAS-1F	GCH-FZ	Hybrid	GAS-1F-RV5	GARCH-FZ-RV5	Hybrid-RV5	GAS-1F-RN5	GARCH-FZ-RN5	Hybrid-RN5
β	0.993	0.922	0.993	0.857	0.857	0.875	0.851	0.761	0.872
(s.e.)	(0.002)	(0.088)	(0.002)	(0.116)	(0.081)	(0.072)	(0.143)	(0.077)	(0.096)
γ	0.008	0.032	0.008	0.004	–	0.004	0.004	–	0.004
(s.e.)	(0.001)	(0.007)	(0.001)	(0.009)	–	(0.007)	(0.013)	–	(0.011)
δ	–	–	4.393×10^{-8}	–	–	0.010	–	–	0.009
(s.e.)	–	–	(1.552×10^{-9})	–	–	(0.016)	–	–	(0.018)
c	–	–	–	0.127	0.095	0.141	0.133	0.084	0.142
(s.e.)	–	–	–	(0.013)	(0.012)	(0.056)	(0.016)	(0.009)	(0.051)
a	−1.774	−2.269	−1.752	−1.973	−2.818	−2.150	−1.962	−2.987	−2.053
(s.e.)	(4.451)	(0.393)	(5.726)	(2.529)	(0.410)	(2.160)	(3.422)	(0.430)	(2.294)
b	−2.401	−3.043	−2.355	−2.599	−3.610	−2.779	−2.601	−3.822	−2.709
(s.e.)	(5.987)	(0.765)	(7.709)	(3.310)	(0.670)	(2.819)	(4.467)	(0.672)	(3.029)
Average loss	0.761	0.780	0.761	0.737	0.727	0.753	0.734	0.722	0.749

This table presents the parameter estimates and standard errors of the four GAS models proposed in Patton et al. (2019) and eight GAS models enhanced with 5-min realized volatility (and overnight returns) for VaR and ES for the S&P 500 using the first rolling window of 2000 days starting with January 2000. The upper part presents the estimated parameters of the two-factor GAS models. The lower part presents the parameters of the GAS-1F model, the GARCH model and the hybrid-factor GAS model estimated with the FZ0 loss minimization. The bottom row of each part presents the average (in-sample) losses from these models.

of the UC test below, on the basis of which the likelihood ratio test is performed:

$$H_{UC}^{VaR} : \mathbb{E}_{t-1}[I_t] = \alpha.$$

Table 3 presents the number of model rejections of the above null hypothesis for four daily equity return series over the out-of-sample period for the 29 different forecasting models at significance levels of 1% and 5%, respectively, and for different probability levels. To obtain the data, we perform the UC test above for all indices, and count the number of rejections for each model.

The third and fourth columns in Table 3 show that the proposed GAS models extended with realized measures generally tend to have a lower number of UC test rejections as compared with the number of rejections of the GAS-FZ models of Patton et al. (2019) for $\alpha = 1\%$. The GARCH-Skt model and the HEAVY model with a skew-t distribution (HEAVY-Skt) also tend to have a lower number of rejections at the 1% significance level. At the 5% significance level, several GAS-FZ models with overnight returns incorporated in the realized volatility have zero rejections in the UC test. In general, adding realized measures into GAS models for predicting VaR achieves a lower number of test rejections on the basis of our results obtained with the hit percentage test.

However, the UC test is statistically weak for a small sample size, and has been criticized in several studies (see Nieto & Ruiz, 2016) because it ignores the clustering of failures. To address these drawbacks, the conditional coverage (CC) test is considered, in which the null hypothesis is as follows:

$$H_{CC}^{VaR} : \mathbb{E}_{t-1}[I_t | I_{t-1}] = \alpha.$$

We use the DQ test proposed by Engle and Manganelli (2004) to implement the CC test. The DQ test has power against the misspecification of ignoring conditionally correlated probabilities and can be extended to examine other explanatory variables. The DQ test examines whether the hit variable defined as $Hit_{v,t} = \mathbf{1}\{Y_t \leq VaR_t\} - \alpha$ follows an independent and identically distributed Bernoulli distribution with probability level α and whether it is independent of the VaR estimator; the expected value of $Hit_{v,t}$ is 0. Furthermore, from the definition of the quantile function, the conditional expectation of VaR_t given any information known at $t-1$ must also be 0, which means that the hit function cannot be correlated with other lagged variables. Also, $Hit_{v,t}$ must not be autocorrelated. If $Hit_{v,t}$ satisfies the conditions stated above, then there will be no autocorrelation in the hits, and no measurement error. We include one lag of $Hit_{v,t}$ in the regression of the test. Consider the following DQ regression:

$$Hit_{v,t} = a_0 + a_1 Hit_{v,t-1} + a_2 VaR_{t-1} + u_{v,t}, \tag{17}$$

where $\mathbf{a} = [a_0, a_1, a_2]$ is the set of parameters of the regression. On the basis of the null hypothesis, we test whether all parameters in the set $\mathbf{a}$ are zero. Performing

Table 3
Number of model rejections based on hit percentages of VaR forecasts (UC test) for the four indices for different α values.

Number	Model	1% VaR		2.5% VaR		5% VaR		10% VaR	
		1%	5%	1%	5%	1%	5%	1%	5%
1	RW-125	3	3	**0**	**0**	**0**	**0**	**0**	**0**
2	RW-250	1	2	**0**	1	**0**	**0**	**0**	**0**
3	RW-500	**0**	2	1	1	**0**	1	**0**	**0**
4	GARCH-G	4	4	3	3	1	1	**0**	1
5	GARCH-Skt	**0**	1	**0**	3	**0**	**0**	**0**	**0**
6	HAR-Skt-RV5	4	4	4	4	4	4	4	4
7	HEAVY-N-RV5	4	4	4	4	**0**	3	**0**	**0**
8	HEAVY-Skt-RV5	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**
9	AL-CAViaR-Sym	2	3	1	3	**0**	**0**	**0**	**0**
10	GAS-2F	3	3	2	2	**0**	**0**	1	2
11	GAS-1F	**0**	3	**0**	**0**	**0**	**0**	1	1
12	GARCH-FZ	1	2	1	3	**0**	**0**	**0**	1
13	Hybrid	2	2	**0**	1	**0**	**0**	1	1
14	GAS-2F-RV5	**0**	1	1	1	1	1	1	1
15	GAS-1F-RV5	**0**	1	**0**	1	**0**	1	**0**	**0**
16	GARCH-FZ-RV5	**0**	1	**0**	1	**0**	**0**	**0**	**0**
17	Hybrid-RV5	2	3	**0**	1	**0**	**0**	**0**	**0**
18	GAS-2F-RV10	1	1	1	1	1	1	1	1
19	GAS-1F-RV10	**0**	2	1	1	**0**	1	**0**	**0**
20	GARCH-FZ-RV10	1	1	1	1	**0**	**0**	**0**	**0**
21	Hybrid-RV10	2	3	1	1	**0**	**0**	**0**	1
22	GAS-2F-RN5	2	3	**0**	1	**0**	**0**	**0**	**0**
23	GAS-1F-RN5	**0**	1	**0**	**0**	**0**	**0**	**0**	1
24	GARCH-FZ-RN5	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
25	Hybrid-RN5	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1
26	GAS-2F-RN10	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**
27	GAS-1F-RN10	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1
28	GARCH-FZ-RN10	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
29	Hybrid-RN10	**0**	1	**0**	**0**	**0**	**0**	1	1

This table presents the number of model rejections based on hit percentages of VaR forecasts (UC test) for the four daily equity return series over the out-of-sample period for 29 different forecasting models. The first three rows (models 1–3) correspond to rolling window historical forecasts, the next two rows (models 4 and 5) correspond to GARCH forecasts based on different distributions for the standardized residuals, the next four rows (models 6–9) correspond to forecasts using high-frequency data and the CAViaR model based on the asymmetric Laplace distribution, the next four rows (models 10–13) correspond to GAS models proposed by Patton et al. (2019) and the last 16 rows (models 14–29) correspond to the GAS models extended with the 5-min and 10-min realized measures.

this DQ test gives a test statistic which is distributed $\mathcal{X}^2(3)$ asymptotically.

Columns 6–9 in Table 4 show the p values of the DQ test of VaR forecasts for $\alpha = 1\%$ for the four stock indices. Values of p greater than 5% indicate no evidence against optimality at the 5% significance level (in bold), and values between 1% and 5% are in italics. For the S&P 500, all of our newly proposed models pass the DQ test at the 1% significance level. When we consider the Nikkei 225 and the FTSE 100, we see significant improvements after adding realized measures in the GAS models. For the DJIA, using realized measures, we find that fewer models fail the DQ test, whereas the historical simulations pass the test, and the GARCH-Skt model performs well. However, for this index, all of the GAS-1F models extended with realized measures are able to pass the DQ test for all four indices. Overall, adding realized measures enables GAS-FZ models to reduce the number of rejections in the DQ test for $\alpha = 1\%$.

For $\alpha = 2.5\%$ (see Table 5), we obtain similar results, namely that adding realized measures generally reduces the number of rejections in the DQ test. For the DJIA, the GAS-2F model can pass the test after addition of realized measures $RN5$ and $RN10$. For $\alpha = 5\%$, in Table 6, we can see that all original GAS-FZ models can pass the DQ test across the four indices except the hybrid model for the S&P 500. After addition of realized measures in the GAS models, it can be seen that the p values increase and the DQ test is generally passed. Table 7 presents the number of model rejections at the 1% and 5% significance levels for quantile regression VaR backtests across the four indices for different probability levels. It can be concluded that the set of GAS models extended with realized measures tend to have a lower number of rejections than the original GAS models and several other benchmarks. It should be noted that the four GAS-1F models extended with different realized measures have the least number of rejections in the DQ test, especially for low values of α.

4.2. Backtesting ES

All models that we consider produce both VaR and ES forecasts. From an economic point of view, for example, when we compare the 2.5% ES forecasts of the GAS-1F-RV5 model and the 2.5% ES forecasts of the GAS-1F model, the first one has, on average, an ES forecast lower by 13.29% for the S&P 500, 17.49% for the DJIA, 8.40% for the Nikkei 225 and 5.31% for the FTSE 100. The results

Table 4
Out-of-sample average losses and dynamic regression tests ($\alpha = 1\%$) for the VaR and ES forecasts.

	Average loss				DQ test (VaR) p values				DES test (ES) p values			
	S&P 500	DJIA	Nikkei 225	FTSE 100	S&P 500	DJIA	Nikkei 225	FTSE 100	S&P 500	DJIA	Nikkei 225	FTSE 100
RW-125	1.479	1.400	1.864	1.298	**0.063**	**0.109**	*0.017*	**0.087**	*0.032*	**0.056**	0.008	**0.082**
RW-250	1.522	1.473	1.928	1.377	**0.350**	**0.302**	*0.042*	*0.043*	**0.255**	**0.204**	*0.024*	**0.075**
RW-500	1.633	1.550	1.998	1.464	**0.128**	**0.159**	*0.017*	*0.028*	**0.170**	**0.162**	*0.049*	**0.058**
GARCH-G	1.380	1.246	1.636	1.190	0.001	0.004	*0.031*	0.000	0.000	0.001	*0.012*	0.000
GARCH-Skt	1.246	1.128	1.565	1.105	*0.043*	**0.114**	**0.550**	**0.265**	*0.036*	*0.049*	**0.433**	**0.268**
HAR-Skt-RV5	1.306	1.118	2.735	1.132	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001
HEAVY-N-RV5	1.233	1.164	1.609	1.137	0.000	0.000	0.003	0.000	0.000	0.000	0.001	0.000
HEAVY-Skt-RV5	1.117	1.047	**1.507**	1.065	**0.063**	*0.021*	**0.414**	**0.145**	**0.053**	*0.028*	**0.310**	**0.166**
AL-CAViaR-Sym	1.306	1.158	1.529	1.102	0.004	**0.095**	**0.255**	**0.296**	0.007	**0.131**	**0.182**	**0.314**
GAS-2F	1.244	1.260	1.670	1.217	*0.035*	**0.075**	**0.493**	0.001	**0.059**	*0.044*	**0.409**	*0.011*
GAS-1F	1.222	1.184	1.650	1.184	**0.209**	**0.219**	*0.019*	**0.629**	**0.371**	**0.274**	*0.025*	**0.832**
GARCH-FZ	1.241	1.147	1.521	1.088	*0.044*	**0.230**	**0.260**	0.000	**0.055**	**0.331**	**0.288**	0.000
Hybrid	1.242	1.140	1.533	1.180	**0.457**	**0.146**	**0.235**	**0.087**	**0.367**	**0.360**	**0.206**	**0.098**
GAS-2F-RV5	*1.106*	1.016	1.562	1.080	**0.687**	0.000	**0.121**	0.000	**0.675**	0.000	**0.198**	0.000
GAS-1F-RV5	1.109	1.008	1.521	1.060	**0.304**	**0.252**	**0.403**	**0.706**	**0.242**	**0.283**	**0.328**	**0.610**
GARCH-FZ-RV5	1.118	1.031	1.518	1.113	**0.349**	**0.525**	**0.296**	**0.112**	**0.253**	**0.268**	**0.251**	**0.089**
Hybrid-RV5	1.141	1.087	1.575	1.069	**0.253**	0.000	*0.033*	**0.324**	**0.205**	0.000	*0.036*	**0.364**
GAS-2F-RV10	1.121	1.015	1.610	1.066	**0.685**	0.000	0.000	**0.671**	**0.812**	0.000	0.005	**0.720**
GAS-1F-RV10	1.117	1.024	1.557	1.071	**0.239**	**0.450**	**0.327**	**0.538**	**0.212**	**0.391**	**0.248**	**0.510**
GARCH-FZ-RV10	1.116	1.052	1.534	1.104	**0.496**	**0.830**	**0.140**	**0.078**	**0.391**	**0.692**	**0.137**	**0.064**
Hybrid-RV10	1.131	1.097	1.617	**1.054**	**0.126**	0.000	0.000	**0.868**	**0.129**	0.000	0.001	**0.836**
GAS-2F-RN5	1.165	*1.001*	1.553	1.076	*0.028*	0.000	0.006	**0.499**	*0.019*	0.000	*0.023*	**0.703**
GAS-1F-RN5	1.109	**0.995**	*1.518*	1.063	**0.295**	**0.429**	**0.164**	**0.243**	**0.262**	**0.445**	**0.190**	**0.223**
GARCH-FZ-RN5	1.123	1.012	1.598	1.109	**0.250**	0.000	**0.659**	**0.058**	**0.157**	0.000	**0.435**	*0.050*
Hybrid-RN5	1.118	1.026	1.582	1.071	**0.319**	0.000	**0.286**	**0.286**	**0.237**	0.000	**0.262**	**0.269**
GAS-2F-RN10	1.133	1.005	1.565	1.065	**0.193**	**0.502**	**0.258**	**0.377**	**0.225**	**0.555**	**0.308**	**0.486**
GAS-1F-RN10	**1.102**	1.014	1.586	*1.060*	**0.790**	**0.696**	**0.548**	**0.340**	**0.717**	**0.702**	**0.371**	**0.334**
GARCH-FZ-RN10	1.123	1.021	1.620	1.113	**0.697**	0.000	**0.192**	**0.115**	**0.570**	0.000	**0.093**	**0.080**
Hybrid-RN10	1.118	1.031	1.549	1.062	**0.261**	0.000	**0.382**	**0.818**	**0.274**	0.000	**0.354**	**0.784**

Columns 2–5 present the average losses, obtained with the FZ0 loss function, for the four daily equity return series over the out-of-sample period for α=1%. The lowest average loss in each column is highlighted in bold, and the second lowest is highlighted in italics. Columns 6–9 and columns 10–13 present p values from dynamic regression tests for the VaR and ES forecasts, respectively. Values greater than 0.05 (indicating no evidence against optimality at the 0.05 level) are in bold, and values between 0.01 and 0.05 are in italics.

indicate that ignoring realized measures overestimates risk on average. Looking at the significance of these values, we follow the backtesting method of Patton et al. (2019) to evaluate the ES estimates individually using a DES regression test:

$$\lambda^s_{e,t} = b_0 + b_1\lambda^s_{e,t-1} + b_2 ES_{t-1} + u_{e,t}, \tag{18}$$

where $\lambda^s_{e,t}$ is the standardized version of $\lambda_{e,t}$ defined in (6) ($\lambda^s_{e,t} = \frac{\lambda_{e,t}}{e_t} = \frac{1}{\alpha}\mathbf{1}\{Y_t \le VaR_t\}\frac{Y_t}{e_t} - 1$) and $\mathbf{b} = [b_0, b_1, b_2]$ is the set of parameters of the regression. On the basis of the null hypothesis, we test whether all parameters in set $\mathbf{b}$ are zero.

Columns 10–13 in Table 4 show the p values from the DES test of the ES forecasts for $\alpha = 1\%$ for the four stock indices. Similarly to the results of the DQ test, incorporating the realized measure $RN10$ in GAS models seems to reduce the number of backtest rejections for the Nikkei 225 and the FTSE 100. GAS-1F models with realized measures can pass the DES test at the 5% significance level for all indices, which is consistent with the results of the DQ test. The GAS-2F model, after addition of the risk measure $RN10$, passes the DES test for all indices. Almost all of our new models pass the DES test across the four indices for $\alpha = 2.5\%$, except the GAS-2F model for the Nikkei 225, as can be seen in columns 10–13 in Table 5. Table 6 presents similar results across the four indices for $\alpha = 5\%$; some benchmarks also have p values higher than 5% (e.g. the HEAVY-Skt model). Table 7 summarizes the total number of model rejections at the 1% and 5% significance levels for the DES regression backtests across the four indices for different probability levels. The GAS-1F models enhanced with realized measures have the smallest number of backtest rejections.

4.3. Joint backtesting of the (VaR, ES) risk measures

To compare jointly the VaR and ES forecasts generated by different models, in this section, a loss function proposed in Fissler and Ziegel (2016) is used. Fissler and Ziegel (2016) discuss how VaR and ES are jointly elicitable and present a group of loss functions for risk measure estimation and backtesting. We follow the choice of Patton et al. (2019) for the loss function FZ0, as defined in (1). To compare the performance of each model using the FZ0 loss function, we calculate the average loss $L_{FZ0} = \frac{1}{T}\sum_{t=1}^{T} L_{FZ0,t}$ for different α values across the four indices.

Columns 2–5 in Table 4 present the average losses for the four equity return series over the out-of-sample period for 13 different benchmark forecasting models and 16 newly proposed models that use the 5-min and 10-min realized measures. The lowest average loss in each column is highlighted in bold, and the second lowest is

Table 5
Out-of-sample average losses and dynamic regression tests ($\alpha = 2.5\%$) for the VaR and ES forecasts.

	Average loss				DQ test (VaR) p values				DES test (ES) p values			
	S&P 500	DJIA	Nikkei 225	FTSE 100	S&P 500	DJIA	Nikkei 225	FTSE 100	S&P 500	DJIA	Nikkei 225	FTSE 100
RW-125	1.198	1.120	1.522	1.063	**0.147**	*0.014*	**0.067**	*0.013*	**0.113**	*0.025*	**0.069**	*0.036*
RW-250	1.238	1.167	1.550	1.128	*0.025*	**0.059**	*0.024*	*0.030*	**0.145**	**0.212**	**0.054**	**0.110**
RW-500	1.347	1.281	1.623	1.235	0.001	0.005	0.006	0.000	*0.018*	*0.018*	*0.025*	*0.020*
GARCH-G	1.080	0.982	1.341	0.989	*0.026*	*0.028*	**0.305**	0.000	0.003	0.005	**0.086**	0.000
GARCH-Skt	1.034	0.942	1.320	0.950	**0.179**	**0.215**	**0.794**	**0.095**	**0.128**	**0.234**	**0.551**	**0.088**
HAR-Skt-RV5	1.044	0.925	2.053	0.959	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.002
HEAVY-N-RV5	0.964	0.898	1.327	0.946	*0.018*	*0.016*	*0.016*	*0.021*	0.002	0.003	0.003	0.006
HEAVY-Skt-RV5	0.926	0.863	**1.291**	0.918	**0.253**	**0.274**	**0.235**	**0.095**	**0.127**	**0.158**	**0.164**	**0.091**
AL-CAViaR-Sym	1.064	0.957	1.311	0.945	**0.075**	**0.172**	**0.530**	**0.067**	*0.035*	**0.235**	**0.434**	**0.072**
GAS-2F	1.057	1.001	1.414	0.979	**0.383**	0.004	**0.103**	0.004	**0.178**	*0.032*	*0.017*	*0.014*
GAS-1F	1.041	0.971	1.356	0.970	**0.765**	**0.316**	**0.073**	**0.873**	**0.654**	**0.284**	**0.083**	**0.899**
GARCH-FZ	1.033	0.929	1.308	0.956	**0.091**	**0.201**	**0.468**	**0.084**	**0.067**	**0.264**	**0.387**	**0.070**
Hybrid	1.020	0.943	1.300	0.954	**0.651**	**0.305**	**0.348**	**0.166**	**0.478**	**0.533**	**0.358**	**0.138**
GAS-2F-RV5	0.947	0.851	1.331	0.909	**0.378**	**0.356**	0.001	**0.886**	**0.740**	**0.360**	0.000	**0.977**
GAS-1F-RV5	0.936	0.844	1.319	0.913	**0.517**	**0.519**	**0.207**	**0.152**	**0.434**	**0.543**	**0.139**	**0.194**
GARCH-FZ-RV5	**0.924**	0.855	*1.297*	0.919	**0.785**	**0.642**	**0.185**	**0.193**	**0.605**	**0.407**	**0.123**	**0.185**
Hybrid-RV5	0.950	0.871	1.315	0.912	**0.419**	**0.338**	**0.185**	**0.862**	**0.367**	**0.343**	**0.147**	**0.896**
GAS-2F-RV10	0.934	0.846	1.338	*0.908*	**0.528**	**0.772**	0.000	**0.792**	**0.800**	**0.722**	0.000	**0.789**
GAS-1F-RV10	0.934	0.869	1.305	0.914	**0.174**	**0.491**	**0.154**	**0.733**	**0.237**	**0.489**	**0.101**	**0.621**
GARCH-FZ-RV10	0.931	0.856	1.311	0.915	**0.795**	**0.695**	**0.103**	**0.130**	**0.631**	**0.465**	**0.075**	**0.120**
Hybrid-RV10	0.946	0.883	1.306	0.913	**0.339**	**0.714**	**0.137**	**0.775**	**0.416**	**0.704**	**0.112**	**0.756**
GAS-2F-RN5	0.942	0.845	1.311	0.910	**0.085**	**0.362**	**0.411**	**0.751**	**0.206**	**0.335**	**0.171**	**0.787**
GAS-1F-RN5	0.939	0.843	1.320	0.914	**0.419**	**0.536**	**0.588**	**0.717**	**0.449**	**0.559**	**0.395**	**0.650**
GARCH-FZ-RN5	*0.925*	0.844	1.330	0.917	**0.816**	**0.876**	**0.738**	**0.224**	**0.696**	**0.659**	**0.516**	**0.220**
Hybrid-RN5	0.942	0.870	1.305	0.912	**0.229**	**0.679**	**0.571**	**0.814**	**0.306**	**0.654**	**0.406**	**0.804**
GAS-2F-RN10	0.937	**0.831**	1.305	**0.907**	**0.029**	**0.804**	**0.429**	**0.855**	**0.116**	**0.839**	**0.210**	**0.824**
GAS-1F-RN10	0.929	0.845	1.318	0.911	**0.391**	**0.493**	**0.730**	**0.233**	**0.402**	**0.508**	**0.506**	**0.245**
GARCH-FZ-RN10	0.930	*0.840*	1.330	0.913	**0.810**	**0.860**	**0.793**	**0.120**	**0.737**	**0.721**	**0.542**	**0.118**
Hybrid-RN10	0.938	0.881	1.305	0.914	**0.286**	**0.401**	**0.644**	**0.545**	**0.381**	**0.457**	**0.452**	**0.438**

Columns 2–5 present the average losses, obtained with the FZ0 loss function, for the four daily equity return series over the out-of-sample period for $\alpha = 2.5\%$. The lowest average loss in each column is highlighted in bold, and the second lowest is highlighted in italics. Columns 6–9 and columns 10–13 present p values from dynamic regression tests for the VaR and ES forecasts, respectively. Values greater than 0.05 (indicating no evidence against optimality at the 0.05 level) are in bold, and values between 0.01 and 0.05 are in italics.

highlighted in italics. For $\alpha = 1\%$, the GAS-FZ models enhanced with the realized volatility using overnight returns and the HEAVY-Skt model perform well overall.

For $\alpha = 2.5\%$ (see Table 5), the GAS-2F model using the 10-min realized volatility and overnight returns (GAS-2F-RN10) outperforms the other models, with lower loss than most of the other models for most series, and is consistently ranked well, being the best model for the DJIA and the FTSE 100. In Table 6 ($\alpha = 5\%$), the GAS-2F-RN5 and GAS-2F-RN10 models outperform the other models, with the lowest loss for the DJIA and the FTSE 100, respectively. The HEAVY-Skt model has the lowest loss for the S&P 500.

Table 8 presents the rankings (with the best-performing model ranked 1 and the worst ranked 29) based on average losses using the FZ0 loss function for the four index return series over the out-of-sample period for the 29 different forecasting models. Columns 6 and 12 give the average rank across the four series and columns 7 and 13 give the rank of the average. For $\alpha = 1\%$, the best-performing model is the GAS-1F model with the 5-min realized volatility and overnight returns, followed by the GAS-1F models extended with the other two realized measures. For $\alpha = 2.5\%$, the GAS-2F-RN10, GARCH-FZ-RV5 and GAS-1F-RN10 models are the three models having the lowest average losses. For $\alpha = 5\%$ and $\alpha = 10\%$, our proposed models have a relatively higher rank than the benchmarks, except for the HEAVY-Skt model, which is ranked second for $\alpha = 5\%$.

Another observation here is that the losses generated from the GAS-FZ models with realized measures are generally lower than the loss generated from most benchmark approaches. However, the HEAVY-Skt model is always one of best five models considered in the overall ranking for all four probability levels. This suggests that the variables extracted from intraday data provide useful information for risk measure forecasting.

To analyse the relative performance of each model, we use the Diebold–Mariano (DM) test to compare any two models using differences in average losses. In this study, t statistics from the DM test compare the average losses, using the FZ0 loss function, for the indices and for different probability levels over the out-of-sample period. A negative t statistic indicates that the row model outperforms the column model with a significant loss difference. Absolute values greater than 1.96 (2.575 or 1.64) indicate that the average loss difference is significantly different from zero at the 95% (99% or 90%) confidence level. In Fig. 1, we present the results for the S&P 500 with the null hypothesis that the row model and the column model have equal values for the loss function. The numbering of the models used in Fig. 1 is given in the first

Table 6
Out-of-sample average losses and dynamic regression tests ($\alpha = 5\%$) for the VaR and ES forecasts.

	Average loss				DQ test (VaR) p values				DES test (ES) p values			
	S&P 500	DJIA	Nikkei 225	FTSE 100	S&P 500	DJIA	Nikkei 225	FTSE 100	S&P 500	DJIA	Nikkei 225	FTSE 100
RW-125	0.977	0.894	1.282	0.876	0.008	**0.075**	0.000	0.002	**0.074**	**0.200**	*0.045*	*0.014*
RW-250	1.011	0.950	1.288	0.931	0.008	**0.065**	**0.072**	0.001	**0.093**	**0.213**	**0.113**	0.007
RW-500	1.104	1.058	1.348	0.993	0.003	0.001	0.000	0.000	0.006	0.003	0.004	0.001
GARCH-G	0.849	0.775	1.142	0.808	**0.715**	**0.840**	**0.949**	*0.024*	**0.243**	**0.273**	**0.448**	0.004
GARCH-Skt	0.836	0.764	1.135	0.794	**0.857**	**0.968**	**0.979**	**0.255**	**0.722**	**0.738**	**0.830**	**0.201**
HAR-Skt-RV5	0.826	0.733	1.613	0.791	0.000	0.000	0.000	0.007	0.000	0.000	0.000	0.002
HEAVY-N-RV5	0.755	0.698	1.123	0.779	**0.477**	**0.265**	**0.299**	**0.144**	**0.079**	*0.046*	*0.048*	*0.032*
HEAVY-Skt-RV5	**0.743**	0.686	1.110	0.768	**0.624**	**0.548**	**0.582**	**0.432**	**0.497**	**0.490**	**0.344**	**0.364**
AL-CAViaR-Sym	0.854	0.770	1.133	0.794	**0.573**	**0.452**	**0.959**	**0.125**	**0.221**	**0.367**	**0.692**	**0.090**
GAS-2F	0.861	0.801	1.151	0.796	**0.624**	**0.455**	**0.703**	**0.264**	**0.206**	**0.200**	**0.562**	**0.456**
GAS-1F	0.848	0.782	1.140	0.786	**0.059**	**0.280**	**0.131**	**0.560**	*0.028*	**0.159**	**0.121**	**0.659**
GARCH-FZ	0.839	0.762	1.134	0.793	**0.441**	**0.456**	**0.973**	**0.255**	**0.388**	**0.417**	**0.732**	**0.218**
Hybrid	0.853	0.770	1.113	0.794	*0.046*	**0.108**	**0.986**	**0.507**	*0.016*	**0.155**	**0.897**	**0.319**
GAS-2F-RV5	0.748	0.679	1.120	*0.764*	**0.237**	**0.503**	0.002	**0.782**	**0.136**	**0.580**	0.000	**0.934**
GAS-1F-RV5	*0.744*	0.684	1.113	0.769	**0.779**	**0.945**	**0.286**	**0.177**	**0.822**	**0.908**	**0.174**	**0.275**
GARCH-FZ-RV5	0.746	0.689	**1.109**	0.771	**0.931**	**0.914**	**0.617**	**0.538**	**0.957**	**0.718**	**0.322**	**0.417**
Hybrid-RV5	0.765	0.693	1.118	0.767	**0.677**	**0.801**	**0.424**	**0.474**	**0.937**	**0.784**	**0.258**	**0.577**
GAS-2F-RV10	0.747	0.676	1.119	0.766	**0.078**	**0.520**	0.004	**0.410**	**0.063**	**0.593**	0.001	**0.666**
GAS-1F-RV10	0.751	0.683	*1.109*	0.774	**0.589**	**0.909**	**0.230**	**0.374**	**0.786**	**0.863**	**0.156**	**0.314**
GARCH-FZ-RV10	0.750	0.689	1.123	0.773	**0.808**	**0.821**	**0.655**	**0.386**	**0.857**	**0.743**	**0.294**	**0.265**
Hybrid-RV10	0.754	0.696	1.108	0.767	**0.561**	**0.578**	**0.435**	**0.349**	**0.750**	**0.887**	**0.274**	**0.563**
GAS-2F-RN5	0.749	**0.671**	1.116	0.767	**0.317**	**0.755**	**0.921**	**0.648**	**0.221**	**0.770**	**0.660**	**0.573**
GAS-1F-RN5	0.747	0.681	1.123	0.766	**0.641**	**0.838**	**0.983**	**0.424**	**0.904**	**0.824**	**0.866**	**0.454**
GARCH-FZ-RN5	0.745	0.679	1.137	0.770	**0.788**	**0.779**	**0.972**	**0.492**	**0.887**	**0.822**	**0.780**	**0.394**
Hybrid-RN5	0.753	0.702	1.117	0.771	**0.668**	**0.820**	**0.998**	**0.445**	**0.923**	**0.813**	**0.815**	**0.599**
GAS-2F-RN10	0.748	*0.674*	1.118	**0.763**	**0.082**	**0.261**	**0.904**	**0.611**	**0.087**	**0.370**	**0.644**	**0.862**
GAS-1F-RN10	0.751	0.682	1.115	0.769	**0.596**	**0.775**	**0.856**	**0.384**	**0.923**	**0.830**	**0.966**	**0.379**
GARCH-FZ-RN10	0.746	0.676	1.134	0.771	**0.729**	**0.621**	**0.990**	**0.378**	**0.923**	**0.826**	**0.900**	**0.277**
Hybrid-RN10	0.765	0.709	1.112	0.768	**0.272**	**0.563**	**0.546**	**0.428**	**0.777**	**0.837**	**0.975**	**0.626**

Columns 2–5 panel present the average losses, obtained with the FZ0 loss function, for the four daily equity return series over the out-of-sample period for $\alpha = 5\%$. The lowest average loss in each column is highlighted in bold, and the second lowest is highlighted in italics. Columns 6–9 and columns 10–13 present p values from dynamic regression tests for the VaR and ES forecasts, respectively. Values greater than 0.05 (indicating no evidence against optimality at the 0.05 level) are in bold, and values between 0.01 and 0.05 are in italics.

column in Table 3. Positive test statistics corresponding to darker colours mean that the row model has larger losses than the column model. The white blocks mean that the row model dominates the column model in loss comparison at the 95% significance level; the light-green (below white in the colour bar) blocks mean that the row model has lower average loss than the column model, but not significantly so; and the dark-red blocks mean that the row model has higher loss than the column model at the 95% significance level. In Fig. 1, at the 1% level, the rows for model 8 (HEAVY-Skt-RV5), model 23 (GAS-1F-RN5) and model 27 (GAS-1F-RN10) have lighter blocks than the other rows; therefore, these are the three best-performing models for the S&P 500 at the 1% level. For the 2.5% level, model 8, model 24 (GARCH-FZ-RN5) and model 27 outperform the others. At the 5% and 10% levels, model 3, model 24 and model 28 (GARCH-FZ-RN10) are the three best-performing models for the S&P 500.

Following Wang et al. (2018) and Taylor (2019), we use the model confidence set (MCS) test introduced by Hansen, Lunde, and Nason (2011) to compare the forecasting models via the FZ0 loss function. This approach builds MCSs using one-sided elimination based on the DM test. In this study, we consider the 75% confidence level[10] and use two methods: the R method using sums of absolute values for calculating the test statistic for MCS, and the SQ method, which uses the summed squares.[11] Table 9 presents the number of models within the MCS test using the block bootstrap with a block length of 12 and 10,000 replications based on the losses generated from the FZ0 loss function. The GAS-2F-RN10 model is the best-performing model overall, and the GAS models extended with realized measures perform better than most of the benchmark models. The main finding from the MCS test echoes the results from the other backtesting methods. The result that some GAS models enhanced with realized measures end up more often in the MCS than the HAR and HEAVY models highlights the usefulness of the score function that the GAS models build on, and we also provide evidence that the use of realized measures enhances the risk forecasts of GAS models.

[10] The 95% confidence level was considered as well with similar results (results are available on request).

[11] Details can be found on page 465 in Hansen et al. (2011); and the MATLAB code for MCS testing can be downloaded from https://github.com/bashtage/mfe-toolbox/.

Table 7
Rejections at the 1% and 5% significance levels for DQ and DES regression backtests across the four indices.

	$\alpha = 1\%$				$\alpha = 2.5\%$				$\alpha = 5\%$				$\alpha = 10\%$			
	VaR		ES		VaR		ES		VaR		ES		VaR		ES	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
RW-125	**0**	1	1	2	**0**	2	**0**	2	3	3	**0**	2	3	4	1	2
RW-250	**0**	2	**0**	1	**0**	3	**0**	**0**	2	2	1	1	1	3	1	2
RW-500	**0**	2	**0**	1	4	4	**0**	4	4	4	4	4	3	4	4	4
GARCH-G	3	4	3	4	1	3	3	3	**0**	1	1	1	**0**	2	**0**	**0**
GARCH-Skt	**0**	1	**0**	2	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**
HAR-Skt-RV5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
HEAVY-N-RV5	4	4	4	4	**0**	4	4	4	**0**	**0**	**0**	3	**0**	**0**	**0**	**0**
HEAVY-Skt-RV5	**0**	1	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	2	**0**	1
AL-CAViaR-Sym	1	1	1	1	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**
GAS-2F	1	2	0	2	2	2	**0**	3	**0**	**0**	**0**	**0**	**0**	1	1	1
GAS-1F	**0**	1	0	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	1	1	1	1
GARCH-FZ	1	2	1	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	2	**0**	**0**
Hybrid	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	1	2	2	**0**	1
GAS-2F-RV5	2	2	2	2	1	1	1	1	1	1	1	1	2	3	1	3
GAS-1F-RV5	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	1
GARCH-FZ-RV5	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1
Hybrid-RV5	1	2	1	2	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
GAS-2F-RV10	2	2	2	2	1	1	1	1	1	1	1	1	1	3	3	3
GAS-1F-RV10	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	1
GARCH-FZ-RV10	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	1	1	1
Hybrid-RV10	2	2	2	2	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
GAS-2F-RN5	2	3	1	3	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	2	2	**0**	2
GAS-1F-RN5	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	1	**0**	1
GARCH-FZ-RN5	1	1	1	2	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1
Hybrid-RN5	1	1	1	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
GAS-2F-RN10	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**	**0**	2	2	1	2
GAS-1F-RN10	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	2	**0**	1
GARCH-FZ-RN10	1	1	1	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	1
Hybrid-RN10	1	1	1	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**0**	**0**

This table presents the number of data series (out of four) with model rejections at the 1% and 5% significance levels for the DQ (VaR) and DES (ES) backtests across the four indices for four probability levels. Smaller numbers of model rejections are preferable (value of 0 in bold).

Table 8
Out-of-sample performance rankings for various values of α.

	$\alpha = 1\%$						$\alpha = 2.5\%$					
	S&P 500	DJIA	Nikkei 225	FTSE 100	Average	Rank	S&P 500	DJIA	Nikkei 225	FTSE 100	Average	Rank
RW-125	27	27	26	27	26.8	27	27	27	26	27	26.8	27
RW-250	28	28	27	28	27.8	28	28	28	27	28	27.8	28
RW-500	29	29	28	29	28.8	29	29	29	28	29	28.8	29
GARCH-G	26	25	23	25	24.8	25	26	25	23	26	25.0	25
GARCH-Skt	23	19	14	17	18.3	21	21	21	17	20	19.8	22
HAR-Skt-RV5	24	18	29	21	23.0	24	23	19	29	23	23.5	23
HEAVY-N-RV5	19	23	19	22	20.8	22	18	18	18	19	18.3	20
HEAVY-Skt-RV5	7	14	**1**	7	7.3	4	**3**	12	**1**	16	8.0	4
AL-CAViaR-Sym	25	22	6	15	17.0	19	25	23	11	18	19.3	21
GAS-2F	22	26	25	26	24.8	26	24	26	25	25	25.0	26
GAS-1F	18	24	24	24	22.5	23	22	24	24	24	23.5	24
GARCH-FZ	20	21	4	14	14.8	17	20	20	9	22	17.8	19
Hybrid	21	20	7	23	17.8	20	19	22	**3**	21	16.3	18
GAS-2F-RV5	**2**	8	12	13	8.8	5	16	9	21	**3**	12.3	15
GAS-1F-RV5	**3**	4	5	**3**	3.8	**2**	9	4	15	9	9.3	8
GARCH-FZ-RV5	8	12	**3**	19	10.5	10	**1**	10	**2**	17	7.5	**2**
Hybrid-RV5	16	16	15	9	14.0	16	17	15	13	7	13.0	17
GAS-2F-RV10	11	7	20	8	11.5	12	8	8	22	**2**	10.0	10
GAS-1F-RV10	6	10	11	11	9.5	8	7	13	4	12	9.0	7
GARCH-FZ-RV10	5	15	8	16	11.0	11	6	11	12	14	10.8	12
Hybrid-RV10	14	17	21	**1**	13.3	14	15	17	8	10	12.5	16

(*continued on next page*)

Table 8 (*continued*).

GAS-2F-RN5	17	**2**	10	12	10.3	9	14	6	10	4	8.5	5
GAS-1F-RN5	4	**1**	**2**	5	3.0	**1**	12	**3**	16	13	11.0	13
GARCH-FZ-RN5	12	5	18	18	13.3	15	**2**	5	19	15	10.3	11
Hybrid-RN5	10	11	16	10	11.8	13	13	14	6	6	9.8	9
GAS-2F-RN10	15	**3**	13	6	9.3	7	10	**1**	5	**1**	4.3	**1**
GAS-1F-RN10	**1**	6	17	**2**	6.5	**3**	4	7	14	5	7.5	**3**
GARCH-FZ-RN10	13	9	22	20	16.0	18	5	**2**	20	8	8.8	6
Hybrid-RN10	9	13	9	4	8.8	6	11	16	7	11	11.3	14
	$\alpha = 5\%$						$\alpha = 10\%$					
	S&P 500	DJIA	Nikkei 225	FTSE 100	Average	Rank	S&P 500	DJIA	Nikkei 225	FTSE 100	Average	Rank
RW-125	27	27	26	27	26.8	27	27	27	27	27	27.0	27
RW-250	28	28	27	28	27.8	28	28	28	26	28	27.5	28
RW-500	29	29	28	29	28.8	29	29	29	28	29	28.8	29
GARCH-G	23	24	24	26	24.3	25	23	26	25	25	24.8	26
GARCH-Skt	20	21	21	24	21.5	21	20	23	23	22	22.0	23
HAR-Skt-RV5	19	19	29	20	21.8	22	15	15	29	16	18.8	18
HEAVY-N-RV5	16	16	16	18	16.5	18	5	14	14	10	10.8	12
HEAVY-Skt-RV5	**1**	11	4	8	6.0	**2**	**1**	10	9	7	6.8	4
AL-CAViaR-Sym	25	23	18	22	22.0	23	24	21	22	24	22.8	25
GAS-2F	26	26	25	25	25.5	26	16	18	16	26	19.0	20
GAS-1F	22	25	23	19	22.3	24	26	20	20	18	21.0	22
GARCH-FZ	21	20	19	21	20.3	20	22	22	24	21	22.3	24
Hybrid	24	22	7	23	19.0	19	21	16	10	17	16.0	16
GAS-2F-RV5	9	5	14	**2**	7.5	6	6	7	5	4	5.5	**2**
GAS-1F-RV5	**2**	10	6	11	7.3	5	7	**1**	**2**	5	3.8	**1**
GARCH-RV5	4	12	**2**	15	8.3	7	**3**	11	6	12	8.0	8
Hybrid-RV5	18	14	12	6	12.5	15	18	24	**3**	23	17.0	17
GAS-2F-RV10	6	4	13	**3**	6.5	**3**	14	9	11	**3**	9.3	10
GAS-1F-RV10	12	9	**3**	17	10.3	11	11	8	**1**	8	7.0	5
GARCH-FZ-RV10	11	13	17	16	14.3	17	9	12	17	14	13.0	14
Hybrid-RV10	15	15	**1**	5	9.0	9	19	19	4	19	15.3	15
GAS-2F-RN5	10	**1**	9	7	6.8	4	8	**3**	12	**2**	6.3	**3**
GAS-1F-RN5	7	7	15	4	8.3	8	13	13	8	9	10.8	13
GARCH-FZ-RN5	**3**	6	22	12	10.8	13	**2**	**2**	19	11	8.5	9
Hybrid-RN5	14	17	10	13	13.5	16	17	25	18	15	18.8	19
GAS-2F-RN10	8	**2**	11	**1**	5.5	**1**	10	6	13	**1**	7.5	6
GAS-1F-RN10	13	8	8	10	9.8	10	12	5	7	6	7.5	7
GARCH-FZ-RN10	5	**3**	20	14	10.5	12	4	4	21	13	10.5	11
Hybrid-RN10	17	18	5	9	12.3	14	25	17	15	20	19.3	21

This table presents the rankings (with the best-performing model ranked 1 and the worst ranked 29) based on the average losses obtained with the FZ0 loss function for four daily equity return series over the out-of-sample period for 29 different forecasting models. Columns 7 and 13 present the average rank across the four equity return series.

5. Conclusions

Patton et al. (2019) proposed a set of semiparametric models (GAS-FZ) in a GAS framework to estimate risk measures. This study provides an extension of this, using exogenous information from high-frequency data to improve on the prediction of VaR and ES. This provides a new semiparametric framework named GAS-FZ-realized, which is proposed for estimating and forecasting VaR and ES jointly. Through incorporation of four realized measures (5-min and 10-min realized volatility with or without the overnight return) into the GAS-FZ models, we observe an improvement in forecasting risk measures over both in-sample and out-of-sample periods.

We use the newly proposed models to estimate the VaR and ES of four international stock indices empirically over the period from 2000 to 2019. The parameters of the models are estimated by our minimizing the FZ loss function of Fissler and Ziegel (2016). Then VaR and ES forecasts are generated and individually backtested with use of the UC test and DQ (and DES) regression tests, and the joint loss function is computed. The main finding is that forecasts generated from the GAS-FZ-realized models outperform forecasts based on GARCH models or historical simulations, even those based on the original GAS-FZ models. The only exception is the HEAVY-Skt-RV5 model, which is difficult to beat.

To conclude, the GAS-FZ-realized models, especially the GAS-2F model combined with the 10-min realized volatility and the overnight return, can provide more accurate risk measures for risk management across different stock indices and probability levels when compared with the other models. This work could be potentially extended by improving the ES component, as the dynamics of VaR may not change simultaneously with ES, for example by modelling an autoregressive relationship between VaR and ES or by assuming a dynamic omega

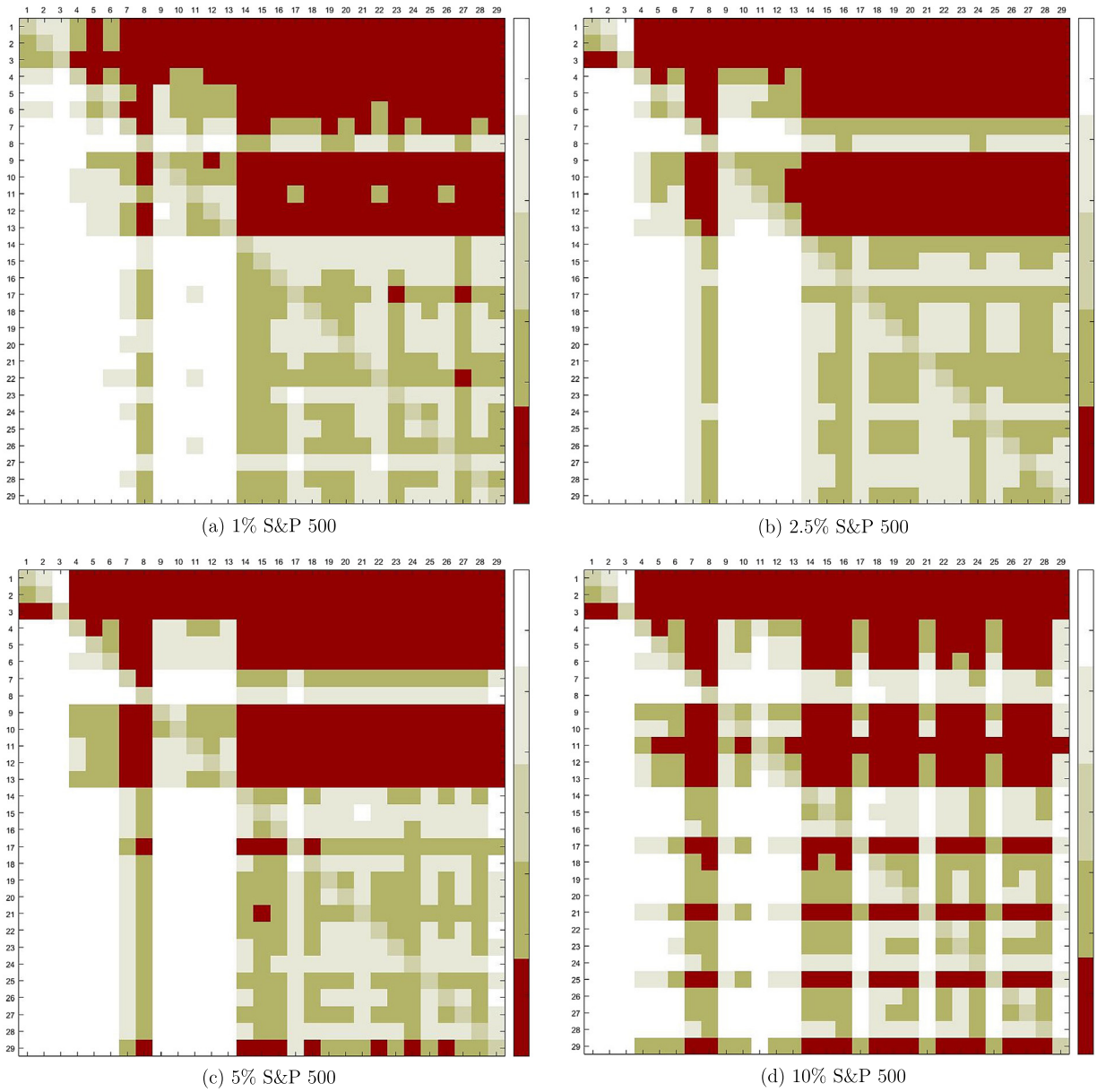

Fig. 1. Colour map based on the DM test comparing the average losses obtained with the FZ0 loss function over the out-of-sample period for 29 different models for the S&P 500: (a) 1% level; (b) 2.5% level; (c) 5% level; (d) 10% level. White blocks mean that the row model has lower average loss than the column model at the 5% significance level; light-green (below white in the colour bar) blocks mean that the row model has lower average loss than the column model, but is not significantly different from it, and so on. Darker-colour blocks mean that the row model has higher average loss than the column model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9
The 75% model confidence set for the R and SQ methods across the four stock indices.

	Summed absolute values (R method)					Summed squares (SQ method)				
	1%	2.5%	5%	10%	Total	1%	2.5%	5%	10%	Total
RW-125	0	0	0	0	0	0	0	0	0	0
RW-250	0	0	0	0	0	0	0	0	0	0
RW-500	0	0	0	0	0	0	0	0	0	0

(*continued on next page*)

Table 9 (*continued*).

	Summed absolute values (R method)					Summed squares (SQ method)				
	1%	2.5%	5%	10%	Total	1%	2.5%	5%	10%	Total
GARCH-N	0	0	0	0	0	0	0	0	0	0
GARCH-Skt	0	0	0	0	0	0	1	0	0	1
HAR-Skt-RV5	0	0	0	0	0	0	0	0	0	0
HEAVY-N-RV5	0	0	0	0	0	0	1	2	1	4
HEAVY-Skt-RV5	3	**3**	3	2	**11**	3	**3**	3	**2**	11
AL-CAViaR-Sym	2	1	0	0	3	2	2	0	0	4
GAS-2F	0	0	0	0	0	0	0	0	0	0
GAS-1F	0	0	0	0	0	0	0	0	0	0
GARCH-FZ	2	1	0	0	3	2	1	0	0	3
Hybrid	1	1	1	0	3	1	1	1	0	3
GAS-2F-RV5	**4**	1	3	**3**	**11**	**4**	2	3	**2**	11
GAS-1F-RV5	**4**	2	3	2	**11**	**4**	3	3	**2**	**12**
GARCH-FZ-RV5	2	3	3	**3**	**11**	3	3	3	**2**	11
Hybrid-RV5	1	1	2	1	5	2	2	2	1	7
GAS-2F-RV10	3	2	**4**	2	**11**	3	2	**4**	**2**	11
GAS-1F-RV10	2	3	2	2	9	2	3	2	**2**	9
GARCH-FZ-RV10	2	2	1	1	6	3	3	3	1	10
Hybrid-RV10	2	2	2	1	7	2	2	3	1	8
GAS-2F-RN5	2	2	**4**	2	10	2	3	**4**	**2**	11
GAS-1F-RN5	**4**	2	2	0	8	**4**	3	3	1	11
GARCH-FZ-RN5	2	2	2	2	8	2	2	2	**2**	8
Hybrid-RN5	2	3	3	0	8	2	3	3	0	8
GAS-2F-RN10	2	**4**	**4**	2	**12**	3	**4**	**4**	**2**	**13**
GAS-1F-RN10	2	2	3	1	8	3	3	3	1	10
GARCH-FZ-RN10	1	2	3	1	7	2	2	3	**2**	9
Hybrid-RN10	3	3	2	0	8	3	3	2	0	8

This table presents the number of indices for which each method is within the MCS at the 75% confidence level based on the FZ0 loss function. The highest value (in bold) means that the model is the most favoured one across the four stock indices and for different probability levels.

ratio to describe the relationship between the two measures (Taylor, 2019). Moreover, this study can be extended by using realized volatility at different frequencies or via other proposed realized measures, for example those found in Meng and Taylor (2018).

References

Acerbi, C., & Székely, B. (2014). Back-testing expected shortfall. *Risk*, *27*, 76–81.

Aigner, D. J., Amemiya, T., & Poirier, D. J. (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.

Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, *57*(3), 1047–1091.

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885–905.

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

Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: A revised market risk framework. *Basel Committee on Banking Supervision, Basel*.

Bayer, S. (2018). Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and statistics*, *8*, 56–77.

Bernardi, M., & Catania, L. (2019). Switching generalized autoregressive score copula models with application to systemic risk. *Journal of Applied Econometrics*, *34*(1), 43–65.

Blair, B. J., Poon, S.-H., & Taylor, S. J. (2001). Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. *Journal of Econometrics*, *105*(1), 5–26.

Cerrato, M., Crosby, J., Kim, M., & Zhao, Y. (2017). The joint credit risk of UK global-systemically important banks. *Journal of Futures Markets*, *37*(10), 964–988.

Clements, M. P., Galvão, A. B., & Kim, J. H. (2008). Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, *15*(4), 729–750.

Corsi, F., Mittnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, *27*(1–3), 46–78.

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

Eckernkemper, T. (2017). Modeling systemic risk: Time-varying tail dependence when forecasting marginal expected shortfall. *Journal of Financial Econometrics*, *16*(1), 63–117.

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

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

Fissler, T., Ziegel, J. F., & Gneiting, T. (2016). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *Risk*, *29*(1), 58—61.

Fuertes, A.-M., & Olmo, J. (2013). Optimally harnessing inter-day and intra-day information for daily value-at-risk prediction. *International Journal of Forecasting*, *29*(1), 28–42.

Gerlach, R., & Chen, C. W. (2014). Bayesian expected shortfall forecasting incorporating the intraday range. *Journal of Financial Econometrics*, *14*(1), 128–158.

Gerlach, R., & Chen, C. W. (2017). Semi-parametric expected shortfall forecasting in financial markets. *Journal of Statistical Computation and Simulation*, *87*(6), 1084–1106.

Gerlach, R., & Wang, C. (2016a). Bayesian semi-parametric realized-care models for tail risk forecasting incorporating realized measures. arXiv preprint arXiv:1612.08488.

Gerlach, R., & Wang, C. (2016b). Forecasting risk via realized GARCH, incorporating the realized range. *Quantitative Finance*, *16*(4), 501–511.

Giot, P., & Laurent, S. (2004). Modelling daily value-at-risk using realized volatility and arch type models. *Journal of Empirical Finance*, *11*(3), 379–398.

Gorgi, P., Hansen, P., Janus, P., & Koopman, S. (2018). Realized Wishart-GARCH: a score-driven multi-asset volatility model. *Journal of Financial Econometrics*, *17*(1), 1–32.

Halbleib, R., & Pohlmeier, W. (2012). Improving the value at risk forecasts: Theory and evidence from the financial crisis. *Journal of Economic Dynamics and Control*, *36*(8), 1212–1228.

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*(6), 877–906.

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

Hansen, P. R., Lunde, A., & Voev, V. (2014). Realized beta GARCH: A multivariate GARCH model with realized measures of volatility. *Journal of Applied Econometrics*, *29*(5), 774–799.

Heber, G., Lunde, A., & Shephard, N. (2009). *Oxford-Mann institute's realized library version 0.1*.

Hua, J., & Manzan, S. (2013). Forecasting the return distribution using high-frequency volatility measures. *Journal of Banking & Finance*, *37*(11), 4381–4403.

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

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, *3*(2), 73–84.

Lange, R.-J., Lucas, A., & Siegmann, A. (2017). Score-driven systemic risk signaling for European sovereign bond yields and CDS spreads. In M. Billio, L. Pelizzon, & R. Savona (Eds.), *Systemic Risk Tomography* (pp. 129–150). Elsevier.

Louzis, D. P., Xanthopoulos-Sisinis, S., & Refenes, A. P. (2014). Realized volatility models and alternative value-at-risk prediction strategies. *Economic Modelling*, *40*, 101–116.

Lucas, A., & Opschoor, A. (2018). Fractional integration and fat tails for realized covariance kernels. *Journal of Financial Econometrics*, *17*(1), 66–90.

Meng, X., & Taylor, J. W. (2018). An approximate long-memory range-based approach for value at risk estimation. *International Journal of Forecasting*, *34*(3), 377–388.

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

Nieto, M. R., & Ruiz, E. (2016). Frontiers in var forecasting and backtesting. *International Journal of Forecasting*, *32*(2), 475–501.

Oh, D. H., & Patton, A. J. (2018). Time-varying systemic risk: Evidence from a dynamic copula model of CDS spreads. *Journal of Business & Economic Statistics*, *36*(2), 181–195.

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

Roccioletti, S. (2015). Backtesting value at risk and expected shortfall. Springer.

Salvatierra, I. D. L., & Patton, A. J. (2015). Dynamic copula models and high frequency data. *Journal of Empirical Finance*, *30*, 120–135.

Shephard, N., & Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics*, *25*(2), 197–231.

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.

Wang, C., Gerlach, R., & Chen, Q. (2018). A semi-parametric realized joint value-at-risk and expected shortfall regression framework. arXiv preprint arXiv:1807.02422.

Wang, C.-S., & Zhao, Z. (2016). Conditional value-at-risk: Semiparametric estimation and inference. *Journal of Econometrics*, *195*(1), 86–103.

Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance*, *26*(4), 901–918.

Žikeš, F., & Baruník, J. (2014). Semi-parametric conditional quantile models for financial returns and realized volatility. *Journal of Financial Econometrics*, *14*(1), 185–226.