University of Reading

VaR and ES forecasting via recurrent neural network-based stateful models

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Qiu, Z., Lazar, E. ORCID: https://orcid.org/0000-0002-8761-0754 and Nakata, K. ORCID: https://orcid.org/0000-0002-7986-6012 (2024) VaR and ES forecasting via recurrent neural network-based stateful models. International Review of Financial Analysis, 92. 103102. ISSN 1873-8079 doi: 10.1016/j.irfa.2024.103102 Available at https://centaur.reading.ac.uk/114808/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1016/j.irfa.2024.103102

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

University of Reading

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

International Review of Financial Analysis 92 (2024) 103102

Contents lists available at ScienceDirect

International Review of Financial Analysis

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

VaR and ES forecasting via recurrent neural network-based stateful models

Zhiguo Qiu [a], Emese Lazar [a,*], Keiichi Nakata [b]

[a] ICMA Centre, Henley Business School, University of Reading, Reading, United Kingdom
[b] Informatics Research Centre, Henley Business School, University of Reading, Reading, United Kingdom

ARTICLE INFO

JEL classification:
C32
C53
G17
C45

Keywords:
Risk models
Value-at-Risk
Expected shortfall
Machine learning
Neural networks

ABSTRACT

Due to the widespread and quickly escalating effects of large negative returns, as well as due to the increase in the importance of regulatory framework for financial institutions, the accurate measurement of financial risks has become a relevant question in the academia and industry. This paper proposes three novel models based on stateful Recurrent Neural Networks (RNN) and Feed-Forward Neural Networks (FNN) to build forecasts for Value-at-Risk (VaR) and Expected Shortfall (ES). We apply the models to six asset return time series spanning over more than 20 years. Our results reveal that the RNN-based stateful models generally outperform the non-stateful RNN models and econometric benchmark models including rolling window models, Generalized AutoRegressive Conditional Heteroskedasticity (GARCH)-type models, and Generalized Autoregressive Score (GAS) models, in terms of VaR and ES forecasting.

1. Introduction

In recent years, risk measurement has increased in importance in finance due to the overarching damages in the economy that can be caused by shocks related to market crashes. The Value-at-Risk (VaR) has been widely used by financial institutions to capture market risk since its introduction in the RiskMetrics model by J.P. Morgan in 1989 (McNeil et al., 2015). VaR refers to an asset's worst return over a predetermined period given a significance level (α). However, VaR has faced criticism due to not being a coherent measure. In contrast to VaR, Expected Shortfall (ES), proposed by Artzner (1997) and Artzner et al. (1999), is coherent which is a desirable property of risk measures, where the α-level ES denotes the expectation of returns below the α-level VaR. After the financial crisis of 2007–2008, the third Basel Accord (Basel Committee on Banking Supervision, 2010) recommends the ES to be used as the main measure of risk replacing VaR. These measures have been widely implemented in the financial industry, as, among other uses, a risk management tool to help estimate the loss on an asset for a defined risk level and allocate the risk more efficiently.

It is a well-know fact that financial asset returns follow fat-tailed distributions (Mandelbrot & Mandelbrot, 1997). Additionally, distributions in the financial markets have asymmetric features, see Alberg et al. (2008), Almeida and Hotta (2014) and Aliyev et al. (2020). Furthermore, regime changes can occur, as highlighted by Ardia et al. (2018) and BenSaïda et al. (2018), whilst time-varying volatility, kurtosis, and skewness also characterize financial returns and should be considered when estimating risk — see Chan and Gray (2006) and Guermat and Harris (2002) and Lucas and Zhang (2016), among others. In recent years, models based on machine learning have gained considerable popularity in finance. One contributing factor is the lack of reliance on simplifying assumptions of the models that use machine learning technologies. Neural networks is a popular branch of machine learning techniques that have gained popularity due to their ability to learn and forecast complex patterns in data. By employing generative neural networks to capture the statistical characteristics of input data, these models can subsequently generate more accurate outputs. As such, generative neural networks are able to depict the dependency structure inherent in asset returns (Arian et al., 2022).

This paper proposes three noval applications based on stateful Recurrent Neural Networks (RNN) and Feed-Forward Neural Networks (FNN) in forecasting VaR and ES. To illustrate the performance of the models in forecasting VaR and ES, we first implement the proposed models on simulated daily return series. To provide an empirical application, we implement the models on daily returns of six assets, using ten different econometric models as benchmarks. Additionally, we use several backtests to test the performance of the proposed models in predicting VaR and ES.

This paper makes three main contributions. Firstly, we propose three novel models for predicting VaR and ES using RNN structures.

* Corresponding author.
E-mail addresses: zhiguo.qiu@pgr.reading.ac.uk (Z. Qiu), e.lazar@icmacentre.ac.uk (E. Lazar), k.nakata@henley.ac.uk (K. Nakata).

[1] Training each machine learning model takes several minutes in the empirical study, see Table 3.

https://doi.org/10.1016/j.irfa.2024.103102
Received 28 September 2023; Received in revised form 10 December 2023; Accepted 19 January 2024
Available online 20 January 2024
1057-5219/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

The proposed structures predict the VaR and ES via learning the serial dependence in the historical data. The structures of the three models are simple and easy to implement and model estimation is time efficient.[1] To avoid overfitting, we use the early-stopping method which stops the training process when the loss on the validation set no longer decreases, as well as the dropout technique which randomly drops out samples from previous layers to reduce overfitting. Compared to the traditional econometric VaR and ES models such as GARCH-type models which rely on certain assumptions about the distribution of the underlying asset returns, the proposed machine learning models take a non-parametric approach, meaning that they do not make any assumptions about the underlying distribution of the data. This gives more flexibility than traditional models to capture complex patterns in the data.

Secondly, the simulation study illustrates that the proposed models successfully capture the underlying structure of VaR and ES. We generate 60 daily returns series via the GARCH(1,1) model with different parameter values. The results show a high correlation between the forecasted VaR and ES obtained by the proposed models and the true VaR and true ES, and the low loss score approximating the true loss score indicates the learning ability of the proposed models.

Thirdly, we undertake a comprehensive comparison between stateful RNN approaches, non-stateful RNN models, and ten benchmark econometric models including rolling window models, GARCH-type models, and GAS models, in an out-of-sample analysis of forecasting VaR and ES over the period from January 2010 to May 2022. The empirical results provide evidence to support the use of the proposed models and identify the best-performing stateful RNN model. Moreover, the results show the beneficial impact of enabling the RNN structure to be stateful, as it effectively enhances the models' ability to capture the underlying structure of tail risk in financial data.

This paper is structured as follows: Section 2 presents the literature review; Section 3 discusses popular econometric VaR and ES, as well as the machine learning models for VaR and ES prediction; Section 4 presents the simulation study; Section 5 presents our empirical analysis, with conclusions in Section 6.

2. Literature review

Quantile regression is a well-known approach for estimating risk measures. Engle and Manganelli (2004) propose a novel approach known as the basic quantile regression model to estimate the conditional autoregressive VaR (CAViaR), which is a well established risk measure. However, this measure is for VaR, so it does not consider the value of the losses in the tail of the distribution. The asymmetric CAViaR model is preferable due to its ability to capture the distinct effects of positive and negative returns on VaR. However, this model can be affected by potential estimation errors (see, for example Huang et al. (2009)). There is a vast amount of literature built around Generalized Autoregressive Conditional Heteroskedasticity (GARCH)-type models that have been widely implemented to obtain risk measures, see So and Philip (2006), Hartz et al. (2006), Degiannakis et al. (2013), and Bucevska (2013) for examples of GARCH-based risk measures. However, these models require specific assumptions about the distribution of underlying asset returns, which might not fully capture the stylized facts of the financial returns.

After the financial crisis of 2007–2008, ES have gained considerable popularity in finance. There is extensive literature on estimating VaR and ES jointly. Taylor (2008) proposes the conditional autoregressive expectile (CARE) models that obtain joint VaR and ES estimates. However, as indicated by Xu et al. (2016), the challenge in parametric CARE modeling is associated with specifying a particular parametric form. Based on the asymmetric Laplace distribution, Taylor (2019) proposes a semiparametric approach with a new scoring function for jointly modeling VaR and ES. Following this, Gerlach and Wang (2020) further extend the CAViaR models by incorporating a realized measure in the dynamics models. However, the CAViaR models imply the same dynamics for VaR and ES. Similarly, the additive autoregressive structure in Taylor (2019) appears inefficient, as both VaR and ES are influenced by changes in volatility. To address this, Taylor (2022) develops a model with a time-varying multiplicative factor Omega ratio to jointly forecast VaR and ES. Patton et al. (2019) proposes several dynamic semiparametric models to jointly forecast VaR and ES, based on the generalized autoregressive score (GAS) framework, which demonstrates good performance overall. However, these semiparametric models can be sensitive to the choice of initial values, potentially resulting in a lack of robustness. Recently, Zhang et al. (2023) proposes a semiparametric methodology for forecasting multiperiod tail risk.

The estimation of these models often requires the use of loss functions. One popular score function that can be used to estimate the parameters of the quantile regression model is the quantile loss function proposed by Koenker and Bassett, Jr. (1978), which is widely implemented for VaR estimation such as for the CAViaR model. Following the CAViaR model, which solely estimates the VaR fail risk, Taylor (2008) proposes the conditional auto-regressive expectile model which involves the asymmetric least square (ALS) regression to forecast the VaR and ES jointly, estimated using the expectile loss function. Based on the FZ loss function proposed by Fissler and Ziegel (2016), Taylor (2019) proposes the asymmetric Laplace (AL) log score function for parameter estimation to forecast the VaR and ES jointly.

Machine learning models have shown great potential in time series prediction. A large part of the literature has focused on the application of machine learning techniques in financial returns forecasts and has shown that these methods can increase the accuracy of forecasts. Patel et al. (2015) forecasts the direction of stocks and stock index movements using machine learning techniques. Gu et al. (2020) compares different linear and nonlinear machine learning models to forecast asset returns and shows that machine learning methods can improve the empirical performance of asset pricing, and concludes that neural networks are the best-performing methods among the machine learning models considered. Saha et al. (2021) uses deep neural networks and event-specific features to predict stock price movements. However, machine learning has been used for other areas of applications such as derivatives pricing and volatility modeling, see Hutchinson et al. (1994), Ye and Zhang (2019), Zhang (2020), Iva¸scu (2021), Vrontos et al. (2021) and Lu et al. (2022).

Machine learning techniques have also been considered for VaR and ES forecasting. Khan (2011) combines the Support Vector Machine model with the heterogeneous autoregressive model to improve VaR prediction. Shim et al. (2012) proposes the semiparametric Support Vector Quantile regression model to estimate VaR. It combines long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) neural networks with GARCH-type models to build one-day ahead VaR estimates using the Parametric and Filtered Historical Simulation (FHS) method. Wu and Yan (2019) develop a conditional quantile model with LSTM neural networks for VaR prediction. Ormaniec et al. (2022) proposes a novel VaR estimator by using the LSTM neural network and concludes that the proposed machine learning model has a better performance when compared to GARCH models on real market data. Cont et al. (2022) proposes a novel data-driven approach using Generative Adversarial Network (GAN) architecture for tail risk estimation. Chronopoulos et al. (2023) uses a deep quantile estimator, based on neural networks, to forecast VaR. However, a substantial gap exists in forecasting VaR and ES jointly using neural networks.

3. Methodology

In the following, we present two popular risk measures in Section 3.1, the FZ0 score function in Section 3.2 and the frameworks of FNN and RNN in Section 3.3. Furthermore, we propose three novel applications of stateful machine learning models in VaR and ES forecasting in Section 3.4, and the benchmark models are presented in Section 3.5.

3.1. Risk measures

This section provides a concise presentation of two prevalent risk measures in finance, namely VaR and ES. Although both measures quantify portfolio risk, they differ in the type of risk they measure. The concept of VaR dates back to as early as 1922 when the New York Stock Exchange imposed capital requirements on firms (Holton, 2003). VaR is defined as the maximum potential loss that will not be exceeded at a specific significance level over a given time horizon:

$$VaR_t^{\alpha} = sup\{x : F_r(x) \leq \alpha\} \tag{1}$$

where $F_r(x)$ is the corresponding cumulative distribution function of assets return, r is the asset return, and $\alpha \in (0,1)$ is a given quantile. Thus, the VaR can be rewritten as the inverse of the cumulative distribution function: $VaR_t^{\alpha} = F_r^{-1}(\alpha)$.

Also, VaR can be considered as the value that gives a weight of α in the cumulative density of returns:

$$\alpha = \int_{-\infty}^{VaR_t^{\alpha}} f_r(x)dx \tag{2}$$

where $f_r(x)$ denotes the probability density function of the returns.

VaR has emerged as the dominant risk measure in both industry and academia, having been adopted by Basel II in 1996 (Duffie & Pan, 1997). It is the primary metric employed by banks and investment institutions to estimate the level of losses that may occur in the event of worst-case scenarios at a given confidence level (Sollis, 2009). Until recently, VaR served as the benchmark for most banks and investment institutions for optimizing capital allocation to manage risk (Philippe, 2001).

However, VaR has an inherent limitation as a risk measurement method in that it fails to account for the shape and structure of the distribution of returns in the tail of the return distribution, rendering it incapable of capturing the expected losses (Roccioletti, 2015). According to Artzner et al. (1999), Delbaen (2002) and Acerbi and Tasche (2002), a risk measure $\phi(X)$ is said to be coherent, if it satisfies the following four conditions:

i. Sub-additivity: $\phi(X+Y) \leq \phi(X) + \phi(Y)$, for any $X, Y, X+Y \in V$

ii. Monotonous: $\phi(X) \leq 0$, for any $X \geq 0, X \in V$

iii. Homogeneity: $\phi(aX) = a\phi(X)$, for any $X, aX \in V, a > 0$

iv. Translational invariance: $\phi(X+a) = \phi(X) - a$, for any $X \in V, a \in R$

In our notation, V is a set of real-valued random variables on some probability space (Ω, A, P). The property of coherence ensures that the risk measure behaves consistently and intuitively, providing meaningful and reliable assessments of risk in financial and other domains. Moreover, as pointed out by Artzner et al. (1999), VaR lacks the property of subadditivity for a portfolio. In response to this deficiency, the Basel Committee on Banking Supervision proposed the transition from VaR to ES in the aftermath of the 2008 financial crisis, as ES measures risk by considering both the amount and frequency of losses at a given level of significance (Basel Committee on Banking Supervision, 2013). ES, a coherent risk measure, is based on the expected loss that surpasses the VaR, which is calculated by taking the expected value of the loss that goes beyond VaR. Specifically, ES (at a significance level α) can be expressed as follows:

$$ES_t^{\alpha} = E[r_t | r_t \leq VaR_t^{\alpha}] \tag{3}$$

While both VaR and ES are extensively used in finance for market risk estimation and management, ES is deemed a more reliable and prudent risk measure than VaR, for it incorporates the severity of losses beyond the VaR level. Nevertheless, the calculation of ES necessitates the estimation of the entire tail of the distribution, rather than a solitary percentile, thus making the process more intricate.

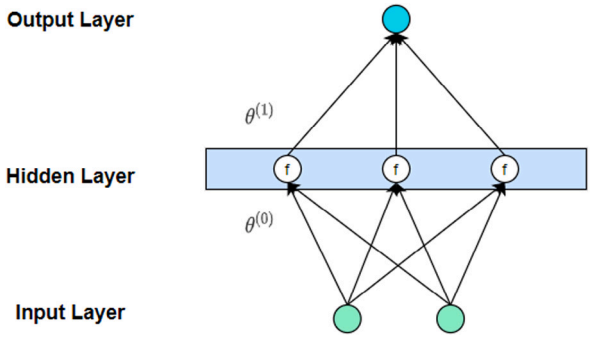

Fig. 1. Simple FNN framework.

3.2. Score function

In this section, we present the FZ0 loss score function that has been widely adopted for model estimation in the literature. Fissler and Ziegel (2016) propose a novel FZ loss function for the joint estimation of VaR and ES, which can be expressed as follows:

$$\begin{aligned} L_{FZ}(y,v,e;\alpha,G_1,G_2) = (1\{y \leq v\} - \alpha)(G_1(v) - G_1(y) + \frac{1}{\alpha}G_2(e)v) \\ - G_2(e)(\frac{1}{\alpha}1\{y \leq v\}y - e) - g_2(e) \end{aligned} \tag{4}$$

where G_1 is weekly increasing, G_2 is strictly increasing and strictly positive, and $g_2'(e) = G_2$, with e denoting the ES and v denoting VaR. The VaR and ES estimates are obtained by minimizing the FZ loss function as follows:

$$(VaR_t^{\alpha}, ES_t^{\alpha}) = \arg\min \mathbb{E}_{t-1}[L_{FZ}(y_t,v,e;\alpha,G_1,G_2)] \tag{5}$$

Patton et al. (2019) considers a special case of this loss function, referred to as the FZ0 loss function, which is obtained by restricting $G_1(x)$ to 0 and $G_2(x)$ to $-1/x$:

$$L_{FZ0}(y,v,e;\alpha) = -\frac{1}{\alpha e}1\{y \leq v\}(v-y) + \frac{v}{e} + \log(-e) - 1 \tag{6}$$

According to the findings of Patton et al. (2019), the GAS models estimated by minimizing the FZ0 loss function exhibit superior performance compared to benchmark models when forecasting VaR and ES.

3.3. The framework of FNN and RNN

Artificial neural networks, or simply neural networks, involve the technique of discovering significant patterns in data by using a high-dimensional approach, inclusive of drop-out layers or regularization methods to tackle the problem of overfitting. In this section, we explore two well known neural network frameworks, namely, the feed-forward neural network (FNN) and the recurrent neural network (RNN), both of which serve as fundamental frameworks within the proposed models in Section 3.4.

3.3.1. FNN structures

The artificial neural network is a nonlinear method with theoretical support as "universal approximators" for smooth predictive schemes (Hornik et al., 1989). Among the system of neural networks, "feed-forward" networks are composed of three parts: the "input layer" including raw predictors, "hidden layers" which interact and transform predictors, and the "output layer" which aggregates all hidden layers to predict outcomes. Hidden layers consist of nodes (neurons) that connect each layer and transmit signals among nodes of different hidden layers. Fig. 1 illustrates the structure of the simplest neural network with one hidden layer.

Table 1
Parameters (true value) used to calculate true VaR and true ES.

α	DoF = 3, Skewness = −0.8		DoF = 5, Skewness = −0.5		DoF = 10, Skewness = −1	
	a_α	b_α	a_α	b_α	a_α	b_α
1%	−3.518	−5.767	−3.289	−4.506	−3.252	−4.118
2.5%	−2.297	−3.980	−2.408	−3.465	−2.496	−3.337
5%	−1.566	−2.929	−1.801	−2.767	−1.924	−2.757

Notes: This table presents the true parameters used to calculate the true VaR and true ES for the daily return series simulated via the GARCH(1,1) model with three different skewed t-distributions.

As shown in Fig. 1, there is one hidden layer with three nodes. Each node captures the information linearly from all the inputs from the input layer. Afterwards, each node implements a so-called "activation function" $f(.)$ on the aggregated inputs signals. The outputs from the ith node in the hidden layer can be expressed as:

$$x_i^{(1)} = f(\theta_{i,0}^{(0)} + \sum_{j=1}^{2} z_j \theta_{i,j}^{(0)}) \tag{7}$$

where z_j denotes the raw inputs, $\theta_{i,j}^{(0)}$, $j = 1, 2$ are the two parameters used for transmitting the raw inputs signals to the ith node in the hidden layer. Finally, the outputs from each node are linearly aggregated into the output layer to make a prediction expressed as: $\theta_0^{(1)} + \sum_{j=1}^{3} x_j^{(1)} \theta_j^{(1)}$.

3.3.2. RNN structures

Linear regression-type methods, including traditional models such as AR, MA, and ARMA, are commonly employed in statistical time series models to model the target variable. However, compared to these conventional linear regression-type models, RNN models proposed by Elman (1990) exhibit superior efficiency in modeling complicated non-linear dynamics and long-term serial dependence. Eq. (8) describes the structure of RNN:

$$\begin{aligned} &h_t = \Delta(\mu x_t + w h_{t-1} + b), \quad h_0 = 0 \\ &y_t | h_t \sim p(y_t | h_t), \quad t = 1, 2, \ldots \end{aligned} \tag{8}$$

where μ, w, b are the model parameters; the recurrently-updated hidden unit h_t stores previous timestep memories and employs the activation function Δ. Such a structural design facilitates the capture of serial dependence within the underlying data. The learning objective pertains to the estimation of the optimal conditional distribution $p(y_t|h_t)$. Furthermore, when the activation function Δ in Eq. (8) is chosen as a linear function and the input x_t represents the square of returns, the RNN process is governed by an equation identical to the conditional volatility process expressed in Eq. (20).

Compared to conventional FNNs, RNNs possess the ability to utilize their internal state, commonly referred to as memory, to process input sequences. Consequently, RNNs are well-suited for capturing significant and efficacious past information, thereby enhancing their decision-making abilities.

3.4. RNN-based models

This section proposes three novel applications of RNN-based stateful models and their corresponding non-stateful models designed for VaR and ES prediction. To preserve the hidden state and memory across input data batches during training, we enable the models to be stateful[2] by setting $stateful = 1$ in Eqs. (10), (13), and (18). This approach enables the network to assimilate information from previous batches and better capture long-term dependencies, without relying on assumptions about the probability density function (PDF) of the distribution $F(r_t|I_{t-1})$, unlike GARCH-type models. We assume there exists an unknown relationship between the target variables (VaR, ES) and the covariate, the square of return (Y_{t-1}^2), for a given α. To learn and capture the unknown relationship, we use the following three stateful RNN models,[3] which also incorporate an FNN structure, as previously discussed in Section 3.3, to achieve one-day ahead VaR and ES forecasts.

3.4.1. SRNN-VE-1 model

To estimate the relationship between the target variables and the covariate, the SRNN-VE-1 model is:

$$[v_t, e_t] = FNN(h_t) \tag{9}$$

$$h_t = RNN(h_{t-1}, Y_{t-1}^2, stateful) \tag{10}$$

where the Y_{t-1}^2 denotes the square of return at time point $t-1$; h_t denotes the hidden variable in the RNN structure; $stateful$ is a binary variable which is 1 if the model is stateful, and 0 otherwise. The RNN layer is set to be stateful to prevent the hidden variable from resetting after each batch.

3.4.2. SRNN-VE-2 model

In order to enhance the capacity for capturing non-linear dependencies, an additional layer, referred to as Eq. (12), is introduced between the FNN and RNN. The SRNN-VE-2 model is defined as:

$$[v_t, e_t] = FNN(k_t) \tag{11}$$

$$k_t = \sqrt{(abs(h_t))} \tag{12}$$

$$h_t = RNN(h_{t-1}, Y_{t-1}^2, stateful) \tag{13}$$

where the $abs(.)$ denotes absolute value. When a linear activation function Δ is employed in Eq. (8), the RNN architecture exhibits similar characteristics to the GARCH(1,1) model, specifically when the input is represented by the square of returns. Moreover, Eq. (12) can also enhance the similarity of the SRNN-VE-2 model to the GARCH-FZ models, thereby facilitating the interpretation that the stateful models are able to capture VaR and ES.

3.4.3. SRNN-VE-3 model

Finally, we consider a hybrid model by combining the model SRNN-VE-1 with the model SRNN-VE-2, combining the two approaches to measure VaR and ES.

$$[v_t, e_t] = -abs([v_t^1, e_t^1] \oplus [v_t^2, e_t^2]) \tag{14}$$

$$[v_t^2, e_t^2] = FNN(k_t) \tag{15}$$

$$[v_t^1, e_t^1] = FNN(h_t) \tag{16}$$

$$k_t = \sqrt{(abs(h_t))} \tag{17}$$

$$h_t = RNN(h_{t-1}, Y_{t-1}^2, stateful) \tag{18}$$

where the operation $\oplus$ represents element-by-element addition.

[2] The details about the 'stateful' setting can be found in TensorFlow, at https://www.tensorflow.org/. The TensorFlow version that we use for empirical study is v2.9.2.

[3] We use the notation SRNN-VE for these stateful RNN models that measure VaR and ES. These models also include the FNN structure. The notation $FNN()$ and $RNN()$ correspond to the FNN structure and RNN structure, respectively in Sections 3.3.1 and 3.3.2.

Table 2
Average correlations and average FZ0 losses over the simulated out-of-sample time period.

	DoF = 3, Skewness = −0.8				DoF = 5, Skewness = −0.5				DoF = 10, Skewness = −1			
Panel A: $\alpha = 1\%$												
Average correlations between true and predicted VaR												
	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3
Pearson	1	0.753	0.806	0.938	1	0.860	0.922	0.944	1	0.964	0.875	0.944
Spearman	1	0.888	0.752	0.969	1	0.884	0.908	0.957	1	0.980	0.873	0.952
Average correlations between true value and predicted ES												
	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3
Pearson	1	0.753	0.806	0.915	1	0.860	0.922	0.931	1	0.964	0.875	0.937
Spearman	1	0.787	0.752	0.967	1	0.884	0.908	0.957	1	0.980	0.873	0.952
Average FZ0 Loss	1.640	1.695	**1.679**	1.697	1.473	1.514	**1.494**	1.516	1.394	1.420	**1.415**	1.423
Panel B: $\alpha = 2.5\%$												
Average correlations between true and predicted VaR												
	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3
Pearson	1	0.909	0.856	0.895	1	0.933	0.893	0.913	1	0.946	0.980	0.918
Spearman	1	0.972	0.842	0.916	1	0.965	0.892	0.918	1	0.964	0.977	0.923
Average correlations between true value and predicted ES												
	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3
Pearson	1	0.909	0.856	0.850	1	0.933	0.893	0.892	1	0.946	0.980	0.907
Spearman	1	0.972	0.842	0.916	1	0.965	0.892	0.918	1	0.964	0.977	0.923
Average FZ0 Loss	1.270	1.298	**1.290**	1.309	1.209	1.227	**1.223**	1.232	1.182	1.197	**1.192**	1.200
Panel C: $\alpha = 5\%$												
Average correlations between true and predicted VaR												
	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3
Pearson	1	0.867	0.885	0.866	1	0.907	0.982	0.888	1	0.922	0.984	0.896
Spearman	1	0.942	0.879	0.862	1	0.942	0.985	0.882	1	0.944	0.982	0.895
Average correlations between true and predicted ES												
	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3	True Value	SRNN-VE-1	SRNN-VE-2	SRNN-VE-3
Pearson	1	0.867	0.885	0.793	1	0.907	0.982	0.858	1	0.922	0.984	0.881
Spearman	1	0.942	0.879	0.862	1	0.942	0.979	0.883	1	0.944	0.982	0.895
Average FZ0 Loss	0.968	0.990	**0.982**	1.000	0.987	0.999	**0.995**	1.005	0.995	1.005	**1.002**	1.007

Notes: This table presents the average correlation coefficients between the true value and predicted values of VaR and ES for each model, and average FZ0 loss scores on the out-of-sample period for three data-generating processes with different DoF and Skewness ((DoF = 3, Skewness = −0.8), (DoF = 5, Skewness = −0.5), and (DoF = 10, Skewness = −1). The average correlation coefficients and the average FZ0 loss scores are calculated by taking the average value of the 20 simulated series. The top part of Panel A reports the correlations between the true VaR and the predicted VaR of the three SRNN-VE models when $\alpha = 1\%$. The middle part of Panel A reports the correlation coefficients between the true ES and the predicted ES from three SRNN-VE models. The lower part of Panel A presents the average FZ0 loss scores of the models. The true FZ0 loss scores are calculated by Eq. (28). Panel B and Panel C present the average correlation coefficients and average FZ0 loss scores obtained for $\alpha = 2.5\%$ and $\alpha = 5\%$, respectively. The lowest average loss for each DoF and Skewness combination is shown in bold.

Table 3
Time cost of training.

Model name	Time cost
RNN-VE-1	4 min, 22 s
RNN-VE-2	6 min, 29 s
RNN-VE-3	4 min, 52 s
SRNN-VE-1	7 min, 23 s
SRNN-VE-2	4 min, 26 s
SRNN-VE-3	8 min, 11 s

Notes: This table presents the time required to train the stateful RNN models and their corresponding non-stateful RNN models on S&P 500 in order to obtain 1%-level VaR and ES predictions, based on a desktop PC comprising four 3.30 GHz quad-core CPUs (specifically, i5-4590 CPUs).

In the three RNN-based stateful models mentioned above, the number of nodes is set to 1, and the activation function is 'linear' both in the FNN layer[4] and RNN layer.[5] The early stopping[6] technique and dropout[7] are applied to mitigate the problem of overfitting. To facilitate a comparison between the stateful and non-stateful models, in addition to the above three models we also implement three non-stateful models (denoted RNN-VE-1, RNN-VE-2, and RNN-VE-3) in our empirical study. These models are defined similarly to Eq. (9)–Eq. (18) except that $stateful = 0$ in Eq. (10), Eq. (13), and Eq. (18) for the three models, respectively.

3.5. Popular VaR and ES forecasting models

This section introduces three distinct categories of econometric models that have been used to forecast VaR and ES. Specifically, we present rolling window-based models, GARCH-type models, and the GAS models (detailed in Sections 3.5.1, 3.5.2, and 3.5.3, respectively). In the following, we select ten benchmark models for model comparison in Section 5.2, including three rolling window-based models with different window sizes, three GARCH-type models with different distributional assumptions, and four GAS models introduced by Patton et al. (2019).

3.5.1. Rolling window-based models

A rolling window approach for estimating VaR and ES can be succinctly described as follows:

$$\begin{aligned} \hat{VaR}_t &= quantile\{Y_s\}_{s=t-m}^{t-1} \\ \hat{ES}_t &= \frac{1}{\alpha m}\sum_{s=t-m}^{t-1} Y_s \mathbb{1}\{Y_s \le \hat{VaR}_s\} \end{aligned} \tag{19}$$

where the $quantile\{Y_s\}_{s=t-m}^{t-1}$ represents the sample quantile of Y_s during the period from time point $(t-m)$ to $(t-1)$, and $\mathbb{1}\{.\}$ denotes the indicator function. This paper employs window sizes of $m = 125$, 250 and 500 with the models referred to as RW-125, RW-250, and RW-500, respectively.

The historical simulation models, as non-parametric specifications, are conceptually straightforward and easy to understand. They rely on observed values from historical data without the need for complex model assumptions, and no parameter estimation is required. However, the produced estimates depend entirely on the selected estimation period. The historical simulation model tends to underestimate risk when the data exhibit no large negative shocks over the estimation period. Moreover, some historical simulation models may exhibit tardiness in incorporating the impact of market crashes, resulting in a delayed reaction of the risk estimates (Abad et al., 2014).

3.5.2. GARCH-type models

The GARCH models proposed by Engle (1982) and Bollerslev (1986) have gained significant popularity in the realm of finance. Among the GARCH models, the basic GARCH(1,1) model stands out as the simplest and most commonly employed, and is defined as follows:

$$\begin{aligned} r_t &= \mu + u_t, \qquad t = 1, 2, \dots, T \\ u_t &= \sigma_t \epsilon_t, \qquad \epsilon_t \sim i.i.d(0,1) \\ \sigma_t^2 &= \beta_0 + \beta_1 u_{t-1}^2 + \beta_2 \sigma_{t-1}^2 \end{aligned} \tag{20}$$

where σ_t^2 denotes the daily variance of the returns r_t.

For parametric GARCH models, the distribution of ϵ_t is an input of the model and is used to estimate the parameters based on the likelihood function. The Standard Normal and Student's t-distributions are the most commonly employed distributions for ϵ_t. In empirical studies applied to financial returns, the Student's t-distribution often outperforms the normal distribution in effectively capturing heavy-tailed returns, as reported by Billio and Pelizzon (2000).

Based on the aforementioned GARCH models, the daily VaR and ES are estimated via the following equations:

$$\begin{aligned} VaR_{t+1}^{\alpha} &= \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{q}_{t+1}(\alpha) \\ ES_{t+1}^{\alpha} &= \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{S}(\alpha) \end{aligned} \tag{21}$$

Here $\hat{\mu}_{t+1}$ and $\hat{\sigma}_{t+1}$ represent the conditional value of the predicted mean and standard deviation of r_{t+1}, respectively, based on the available information up to time t; $\hat{S}(\alpha)$ is defined as $\mathbb{E}[\epsilon_{t+1}|\epsilon_{t+1} < \hat{q}_{t+1}(\alpha)]$; and the quantile $\hat{q_{t+1}}(\alpha) = inf\{\epsilon_{t+1} : \alpha \le F(\epsilon_{t+1})\}$, where F denotes the cumulative distribution function of the standardized residuals ϵ.

The forecast of the conditional volatility ($\hat{\sigma}_{t+1}$) is derived from the GARCH model, while the mean of returns ($\hat{\mu}_{t+1}$) is computed based on past returns. To obtain the α-level VaR estimate, a linear function of the forecast of the conditional $\alpha - quantile$ of the standardized residuals, $\hat{q_{t+1}}(\alpha)$ is computed. It is evident that both the VaR and ES estimators are contingent on the presumed distribution of the standardized residuals.

GARCH-type models are part of the parametric models' family, and they often assume a specific distribution, such as the standard normal distribution, for the error term. However, such strong assumptions often present challenges in fully capturing the stylized facts of the financial returns.

This paper employs three GARCH-type models as benchmarks, namely the GARCH model with normally distributed residuals (GCH-N), the GARCH model with the skewed t-distribution of Hansen (1994) (GCH-Skt), and the GARCH model with the distribution estimated by the Empirical Distribution Function as a nonparametric alternative (GCH-EDF).

3.5.3. GAS models

This paper also uses the GAS models (namely, the FZ2F model, FZ1F model, GCH-FZ model, and the Hybrid model) proposed by Patton et al. (2019) as benchmarks for model comparison. The models are given by:

- FZ2F model:

$$\begin{bmatrix} v_t \\ e_t \end{bmatrix} = \mathbf{W} + \mathbf{B}\begin{bmatrix} v_{t-1} \\ e_{t-1} \end{bmatrix} + \mathbf{A}\begin{bmatrix} \lambda_{v,t-1} \\ \lambda_{e,t-1} \end{bmatrix} \tag{22}$$

Here $\lambda_{v,t-1} \equiv -v_t(\mathbb{1}\{Y_{t-1} \le v_{t-1}\} - \alpha)$ and $\lambda_{e,t-1} \equiv \frac{1}{\alpha}\mathbb{1}\{Y_{t-1} \le v_{t-1}\}Y_{t-1} - e_{t-1}$; $\mathbf{W}$ is a vector of size (2×1); $\mathbf{B}$ is a diagonal matrix, and $\mathbf{A}$ is a (2×2) matrix.

[4] We use the Application Programming Interface (API) of TensorFlow: 'tensorflow.keras.layers.Dense()' from https://www.tensorflow.org.

[5] We use the API of Tensorflow 'tensorflow.keras.layers.SimpleRNN()' from https://www.tensorflow.org.

[6] We monitor the loss on the validation set, see the API 'tensorflow.keras.callbacks.EarlyStopping()' from https://www.tensorflow.org.

[7] To implement the dropout technique, we use the API 'tensorflow.keras.layers.SimpleRNN(dropout=0.2)' from https://www.tensorflow.org.

Table 4
Summary statistics.

	S&P 500	FTSE	DJIA	Oil Spot	Gold Spot	USDJPY
Total count	5614	5645	5610	5595	5595	5803
In-sample	2502	2513	2503	2485	2485	2589
Out-of-sample	3112	3132	3107	3110	3110	3214
Mean	4.688	0.582	4.822	9.464	8.039	1.032
StdDev	19.669	18.449	18.819	42.423	16.94	9.596
Skew	−0.391	−0.301	−0.398	0.031	−0.259	−0.041
Kurt	13.327	10.138	15.81	16.222	9.252	7.388
VaR ($\alpha = 0.01$)	−3.55	−3.438	−3.533	−7.561	−2.995	−1.643
VaR ($\alpha = 0.025$)	−2.589	−2.48	−2.412	−5.166	−2.202	−1.213
VaR ($\alpha = 0.05$)	−1.916	−1.814	−1.824	−3.936	−1.669	−0.955
VaR ($\alpha = 0.10$)	−1.296	−1.238	−1.201	−2.76	−1.162	−0.688
ES ($\alpha = 0.01$)	−5.218	−4.71	−5.075	−10.394	−4.115	−2.22
ES ($\alpha = 0.025$)	−3.894	−3.597	−3.722	−7.804	−3.167	−1.724
ES ($\alpha = 0.05$)	−3.06	−2.862	−2.904	−6.153	−2.537	−1.395
ES ($\alpha = 0.10$)	−2.317	−2.175	−2.189	−4.722	−1.96	−1.1

Notes: This table presents summary statistics on the 6 daily asset return series, over the full sample period from 1 January 2000 to 31 May 2022. The first three rows report the total number of observations over the full sample period, the number of observations over the in-sample period, and the number of observations over the out-of-sample period. Also, the annualized mean, standard deviation, skewness and kurtosis of these daily return series are reported. The last eight rows present the sample Value-at-Risk for four different values of α and the corresponding sample Expected Shortfall estimates.

Table 5
Parameter estimates of ARMA and GARCH(1,1) models.

	S&P 500	FTSE	DJIA	Oil Spot	Gold Spot	USDJPY
ARMA model						
Constant	−0.011	−0.008	−0.003	0.046	0.052	−0.003
AR(1)	−0.087		−0.086			
MA(1)		−0.040				−0.057
Order(p,q)	(2, 0)	(0, 4)	(2, 0)	(0, 0)	(0, 0)	(0, 1)
GARCH(1,1) with skewed t-distribution model						
ω	0.010	0.000	0.010	0.159	0.010	0.008
β	0.073	0.169	0.076	0.069	0.036	0.048
γ	0.920	0.831	0.918	0.907	0.957	0.934
DoF	9.042	15.212	8.556	8.764	5.718	7.750
Skewness	−0.105	−0.105	−0.095	−0.078	−0.022	−0.057

Notes: This table presents parameter estimates for the six assets' daily return series, over the in-sample period from January 2000 to December 2009. The first panel reports the parameters and order of the optimal ARMA model, with the selection made based on the BIC method. The second panel presents parameter estimates for GARCH(1,1) with the skewed t-distribution.

Table 6
Parameter estimates of GAS models for VaR and ES for S&P 500 (α = 1%).

	FZ2F			FZ1F	GCH-FZ	Hybrid
	VaR	ES				
ω	−0.060	−0.094	β	0.991	0.923	0.977
(s.e.)	(0.177)	(0.514)	(s.e.)	(0.002)	(0.189)	(0.007)
b	0.979	0.978	γ	0.003	0.053	0.003
(s.e.)	(0.062)	(0.121)	(s.e.)	(0.000)	(0.017)	(0.001)
a_v	0.000	0.001	δ			0.013
(s.e.)	(0.582)	(0.004)	(s.e.)			(0.002)
a_e	0.000	0.001	a	−2.279	−2.777	−3.906
(s.e.)	(1.652)	(0.013)	(s.e.)	(1.213)	(0.406)	(9.113)
			b	−3.283	−3.529	−4.931
			(s.e.)	(1.796)	(1.295)	(11.688)
Average loss	0.592			0.603	0.637	0.590

Notes: This table presents the estimated parameters as well as their standard errors (s.e.) of the GAS models for S&P 500 daily return series for 1%-level VaR and ES forecasting, over the in-sample period from January 2000 to December 2009. The left panel shows the results for the two-factor GAS model (FZ2F). The right panel reports the results for the three models: the one-factor GAS model (FZ1F), the GARCH model estimated by FZ loss minimization (GCH-FZ), and the hybrid-factor GAS model (Hybrid). The average (in-sample) FZ0 losses for these models are shown in the bottom row.

Table 7
Numbers of rejections at the 1% and 5% significance levels.

Significance level	α = 1%				α = 2.5%			
	VaR		ES		VaR		ES	
	1%	5%	1%	5%	1%	5%	1%	5%
RW-125	5	6	4	6	6	6	4	6
RW-250	4	5	**2**	5	2	6	1	4
RW-500	6	6	**2**	6	5	5	5	6
GCH-N	3	4	4	4	1	3	3	4
GCH-Skt	2	**2**	**2**	**2**	1	2	1	**2**
GCH-EDF	2	3	**2**	3	1	**1**	1	**2**
FZ2F	**1**	**2**	**2**	3	2	3	3	3
FZ1F	4	4	4	4	1	**1**	1	**2**
GCH-FZ	2	4	**2**	4	2	2	2	**2**
Hybrid	3	4	3	4	1	2	1	**2**
RNN-VE-1	5	5	5	5	4	4	4	4
RNN-VE-2	4	4	4	4	2	3	4	4
RNN-VE-3	3	3	**2**	**2**	2	2	2	**2**
SRNN-VE-1	2	4	**2**	4	**0**	4	1	5
SRNN-VE-2	2	4	**2**	4	**0**	2	1	**2**
SRNN-VE-3	**1**	**2**	**2**	3	**0**	**1**	**0**	**2**

Notes: This table presents the number of assets that obtain rejections for DQ and DES regression backtests over the out-of-sample period, at 1% and 5% significance levels, across the six daily returns. The lowest number of rejections in each column is shown in bold.

Table 8
Out-of-sample performance rankings.

α =1%									
ID	Model name	S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY	Average	Rank
1	RW-125	13	12	14	10	10	15	12.33	14
2	RW-250	15	14	15	15	15	14	14.67	15
3	RW-500	16	15	16	16	14	16	15.50	16
4	GCH-N	9	16	10	9	13	3	10.00	10
5	GCH-Skt	4	13	3	5	3	1	4.83	3
6	GCH-EDF	8	8	4	4	5	4	5.50	4
7	FZ2F	10	4	9	11	6	13	8.83	9
8	FZ1F	5	5	7	8	4	12	6.83	7
9	GCH-FZ	7	7	5	1	7	11	6.33	6
10	Hybrid	3	6	6	7	16	7	7.50	8
11	RNN-VE-1	14	10	11	14	9	10	11.33	13
12	RNN-VE-2	12	9	12	12	12	9	11.00	11
13	RNN-VE-3	11	11	13	13	11	8	11.17	12
14	SRNN-VE-1	2	2	2	3	1	6	2.67	2
15	SRNN-VE-2	6	3	8	6	8	2	5.50	5
16	SRNN-VE-3	1	1	1	2	2	5	2.00	1
α =2.5%									
ID	Model name	S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY	Average	Rank
1	RW-125	13	9	14	11	11	15	12.17	12
2	RW-250	15	14	15	15	15	11	14.17	15
3	RW-500	16	16	16	16	14	16	15.67	16
4	GCH-N	10	15	10	6	7	1	8.17	9
5	GCH-Skt	6	12	5	5	4	4	6.00	6
6	GCH-EDF	5	7	6	3	5	2	4.67	4
7	FZ2F	8	8	9	10	10	10	9.17	10
8	FZ1F	7	4	7	7	9	9	7.17	8
9	GCH-FZ	4	6	3	1	6	7	4.50	3
10	Hybrid	3	2	4	9	8	8	5.67	5
11	RNN-VE-1	14	11	12	14	12	13	12.67	13
12	RNN-VE-2	11	10	11	12	16	12	12.00	11
13	RNN-VE-3	12	13	13	13	13	14	13.00	14
14	SRNN-VE-1	2	3	2	2	1	6	2.67	2
15	SRNN-VE-2	9	5	8	8	3	3	6.00	7
16	SRNN-VE-3	1	1	1	4	2	5	2.33	1

Notes: This table presents model rankings (with the best-performing model ranked 1 and the worst ranked 16) based on the average losses obtained with the FZ0 loss function for 6 daily return series over the out-of-sample period for 16 different forecasting models. Columns 10 present the average rank across the six return series, for 1% and 2.5% α values, respectively.

- FZ1F model:

$$\begin{aligned} v_t &= a\exp(\kappa_t) \\ e_t &= b\exp(\kappa_t), \quad b < a < 0 \\ \kappa_t &= \omega + \beta\kappa_{t-1} + \gamma H_{t-1}^{-1} s_{t-1} \end{aligned} \tag{23}$$

where the scaling matrix H_{t-1} is set to 1, and the score variable s_{t-1} is:

$$s_{t-1} = -\frac{1}{e_{t-1}}\left(\frac{1}{\alpha}\mathbb{1}\{Y_{t-1} \le v_{t-1}\}Y_{t-1} - e_{t-1}\right) \tag{24}$$

- GCH-FZ model:

$$\begin{aligned} v_t &= a\sigma_t \\ e_t &= b\sigma_t, \quad b < a < 0 \\ \sigma_t^2 &= w + \beta\sigma_{t-1}^2 + \gamma Y_{t-1}^2 \end{aligned} \tag{25}$$

The conditional variance σ_t^2 is assumed to follow the GARCH(1,1) process. Instead of using Quasi Maximum Likelihood Estimation, the GCH-FZ model estimates the parameters via the FZ0 loss score (Eq. (6)) minimization.

- Hybrid model:

$$\begin{aligned} v_t &= a\exp(\kappa_t) \\ e_t &= b\exp(\kappa_t), \quad b < a < 0 \\ \kappa_t &= \omega + \beta\kappa_{t-1} + \gamma\frac{1}{e_{t-1}}\left(\frac{1}{\alpha}\mathbb{1}\{Y_{t-1} \le v_{t-1}\}Y_{t-1} - e_{t-1}\right) + \delta\mathbf{log}|Y_{t-1}| \end{aligned} \tag{26}$$

GAS models are semiparametric; they introduce a parametric structure to capture the dynamics of VaR and ES without making assumptions about the conditional distribution of the error term. The models rely on an autoregressive term that drives the risk measures. The estimation of these models is, however, challenging, lacking closed-form analytic solutions. The use of numerical methods increases the time and computational costs, and finding a global solution poses a challenge. Additionally, these models are sensitive to initial values in the estimation. In the following we will use all of the above models for model comparison purposes.

4. Simulation study

In this section, we employ simulations to demonstrate the efficacy of the RNN-based stateful models in capturing the sequential dependence of VaR and ES. The data generating process is given by the following GARCH process:

$$\begin{aligned} Y_t &= \sigma_t\eta_t \\ \sigma_t^2 &= w + \beta\sigma_{t-1}^2 + \gamma Y_{t-1}^2 \\ \eta_t &\sim iid\, F_\eta(0,1) \end{aligned} \tag{27}$$

The GARCH simulation uses parameter values ω, β, and γ as (0.05, 0.9, 0.05). We use Hansen's (1994) skewed t-distribution. For the simulation, we implement three sets of parameters for DoF and skewness, namely (DoF = 3, Skewness = −0.8), (DoF = 5, Skewness = −0.5), and (DoF = 10, Skewness = −1). The values of parameters (b_α, a_α) used to compute VaR and ES estimates using the equation below, as in Patton

Table 9
Out-of-sample average losses and results for the dynamic regression tests for the VaR and ES forecasts.

		α = 1%																	
ID	Model name	Average loss						DQ Test (VaR) P Values						DES Test (ES) P Values					
		S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY	S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY	S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY
1	RW-125	1.499	1.311	1.434	2.178	1.277	0.736	0.006	0.002	0.013	0.001	0.000	0.002	0.007	0.002	0.015	0.014	0.000	0.001
2	RW-250	1.558	1.372	1.533	2.322	1.339	0.691	0.002	0.001	0.011	0.000	0.003	**0.210**	0.040	0.001	0.044	0.038	0.004	**0.202**
3	RW-500	1.645	1.440	1.644	2.420	1.331	0.788	0.000	0.000	0.000	0.006	0.004	0.008	0.011	0.000	0.012	0.040	0.000	0.011
4	GCH-N	1.406	1.582	1.315	2.118	1.322	0.575	0.000	0.000	0.000	**0.234**	0.030	**0.569**	0.000	0.000	0.000	**0.066**	0.005	**0.261**
5	GCH-Skt	1.241	1.363	1.153	2.062	1.222	**0.570**	**0.117**	0.000	**0.254**	**0.556**	**0.447**	0.000	**0.086**	0.000	**0.206**	**0.474**	**0.453**	0.000
6	GCH-EDF	1.294	1.239	1.159	2.058	1.227	0.577	0.004	0.020	**0.069**	**0.558**	**0.593**	0.000	0.004	0.013	**0.085**	**0.433**	**0.469**	0.000
7	FZ2F	1.414	1.187	1.213	2.221	1.258	0.665	0.033	**0.196**	**0.229**	**0.057**	**0.229**	0.000	0.021	**0.297**	**0.157**	0.004	**0.176**	0.000
8	FZ1F	1.251	1.203	1.192	2.085	1.227	0.660	**0.107**	0.000	0.000	**0.873**	0.008	0.000	**0.211**	0.000	0.000	**0.623**	0.001	0.000
9	GCH-FZ	1.276	1.231	1.183	**2.029**	1.261	0.659	0.012	0.027	**0.096**	**0.344**	0.000	0.000	0.011	0.018	**0.103**	**0.391**	0.000	0.000
10	Hybrid	1.235	1.207	1.184	2.074	1.401	0.621	**0.162**	0.029	0.000	**0.382**	0.000	0.000	**0.166**	0.041	0.000	**0.649**	0.000	0.000
11	RNN-VE-1	1.523	1.286	1.344	2.257	1.273	0.650	0.000	0.000	0.000	**0.738**	0.000	0.000	0.000	0.000	0.000	**0.407**	0.000	0.000
12	RNN-VE-2	1.464	1.286	1.362	2.245	1.286	0.648	0.000	0.000	0.000	**0.903**	**0.308**	0.000	0.000	0.000	0.000	**0.812**	**0.107**	0.000
13	RNN-VE-3	1.437	1.307	1.422	2.248	1.277	0.642	0.007	0.000	**0.753**	**0.685**	**0.804**	0.000	**0.767**	0.000	**0.821**	**0.549**	**0.306**	0.000
14	SRNN-VE-1	*1.229*	*1.148*	*1.150*	2.049	**1.205**	0.604	0.014	**0.081**	0.014	0.000	**0.610**	0.000	0.015	**0.094**	0.025	0.000	**0.406**	0.000
15	SRNN-VE-2	1.271	1.182	1.203	2.074	1.269	**0.570**	0.020	**0.196**	0.037	**0.094**	0.000	0.000	0.018	**0.234**	**0.060**	0.038	0.000	0.000
16	SRNN-VE-3	**1.205**	**1.130**	**1.134**	*2.039*	*1.207*	0.591	**0.074**	**0.335**	0.044	**0.091**	**0.131**	0.000	**0.059**	**0.280**	**0.062**	0.008	0.045	0.000
		α = 2.5%																	
ID	Model name	Average loss						DQ Test (VaR) P Values						DES Test (ES) P Values					
		S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY	S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY	S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY
1	RW-125	1.209	1.058	1.141	1.907	1.032	0.436	0.001	0.001	0.003	0.002	0.000	0.001	0.005	0.001	0.010	0.011	0.003	0.001
2	RW-250	1.237	1.102	1.193	1.964	1.051	0.408	0.017	0.017	0.007	0.044	0.008	0.032	0.043	0.010	**0.071**	**0.055**	0.030	0.035
3	RW-500	1.313	1.207	1.291	2.086	1.043	0.482	0.000	0.000	0.000	0.002	**0.084**	0.009	0.000	0.000	0.001	0.008	0.047	0.001
4	GCH-N	1.059	1.159	0.996	1.813	0.999	**0.332**	0.044	0.000	0.012	**0.977**	**0.173**	**0.163**	0.003	0.000	0.001	**0.552**	0.038	**0.415**
5	GCH-Skt	1.010	1.080	0.949	1.802	0.974	0.339	**0.333**	0.000	**0.280**	**0.703**	**0.336**	0.012	**0.167**	0.000	**0.172**	**0.514**	**0.443**	0.011
6	GCH-EDF	1.008	1.008	0.950	1.799	0.975	*0.336*	**0.528**	0.001	**0.261**	**0.981**	**0.526**	**0.054**	**0.218**	0.000	**0.154**	**0.802**	**0.496**	0.022
7	FZ2F	1.030	1.040	0.977	1.895	1.026	0.389	0.026	0.000	**0.098**	**0.494**	**0.155**	0.000	**0.064**	0.000	**0.086**	**0.441**	0.003	0.000
8	FZ1F	1.024	0.968	0.950	1.822	1.015	0.375	**0.713**	**0.492**	**0.508**	**0.717**	**0.133**	0.001	**0.499**	**0.545**	**0.346**	**0.624**	0.030	0.006
9	GCH-FZ	1.007	1.001	0.940	**1.795**	0.987	0.358	**0.636**	0.002	**0.394**	**0.293**	**0.181**	0.001	**0.339**	0.001	**0.268**	**0.273**	**0.063**	0.000
10	Hybrid	1.005	*0.955*	0.947	1.854	1.014	0.370	**0.844**	**0.293**	**0.702**	**0.376**	0.044	0.000	**0.664**	**0.392**	**0.609**	**0.170**	0.045	0.001
11	RNN-VE-1	1.218	1.080	1.091	1.950	1.032	0.411	0.000	0.000	0.000	**0.493**	**0.984**	0.000	0.000	0.000	0.000	**0.800**	**0.400**	0.000
12	RNN-VE-2	1.148	1.078	1.079	1.946	1.056	0.410	**0.783**	0.033	0.000	**0.371**	**0.066**	0.000	0.104	0.004	0.000	**0.784**	0.000	0.000
13	RNN-VE-3	1.177	1.091	1.117	1.947	1.036	0.411	**0.209**	0.000	**0.360**	**0.400**	**0.854**	0.000	**0.584**	0.000	**0.226**	**0.859**	**0.123**	0.000
14	SRNN-VE-1	*0.986*	0.960	*0.930*	*1.797*	**0.968**	0.354	0.044	0.047	0.016	**0.286**	**0.723**	0.048	0.036	0.032	0.012	0.001	**0.814**	0.044
15	SRNN-VE-2	1.048	0.973	0.955	1.830	0.973	0.337	**0.151**	0.047	**0.143**	0.018	**0.432**	**0.139**	0.039	**0.070**	**0.131**	0.003	**0.421**	**0.205**
16	SRNN-VE-3	**0.975**	**0.948**	**0.918**	1.799	*0.969*	0.351	**0.212**	**0.140**	**0.150**	**0.088**	**0.633**	0.041	**0.081**	**0.089**	**0.147**	0.012	**0.775**	0.027

Notes: Columns 3–8 present the average FZ0 losses for 6 different assets daily return series. The lowest average loss in each column is shown in bold, and the second lowest is shown in italics. Columns 9–14 and columns 15–20 present p-values of the dynamic regression tests DQ and DES, respectively, for the VaR and ES forecasts. Values greater than 0.05 (indicating no evidence against optimality at the 5% level) are in bold.

Table 10
Average losses and ranks of the models over the period 1 Jan 2010 to 31 Dec 2018.

α = 1%									
ID	Model name	Average loss						Average rank	Rank
		S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY		
1	RW-125	1.392	1.167	1.320	1.912	1.268	0.775	12	14
2	RW-250	1.359	1.221	1.344	2.022	1.329	0.731	14.167	15
3	RW-500	1.426	1.303	1.377	2.200	1.248	0.843	14.5	16
4	GCH-N	1.291	1.288	1.151	1.876	1.317	0.662	9.833	9
5	GCH-Skt	1.157	1.115	1.038	**1.864**	1.219	**0.642**	3.667	2
6	GCH-EDF	1.197	1.038	1.040	1.867	1.223	0.647	4.667	4
7	FZ2F	1.273	1.074	1.093	1.962	1.249	0.720	10	10
8	FZ1F	1.166	1.075	1.050	1.872	**1.205**	0.712	5.667	5
9	GCH-FZ	1.179	1.031	1.059	1.877	1.273	0.713	7.833	8
10	Hybrid	1.154	1.040	1.040	1.890	1.425	0.682	7.167	7
11	RNN-VE-1	1.455	1.178	1.220	1.959	1.239	0.706	11	12
12	RNN-VE-2	1.379	1.171	1.290	1.965	1.255	0.701	11.667	13
13	RNN-VE-3	1.314	1.186	1.237	1.959	1.245	0.687	10.5	11
14	SRNN-VE-1	1.119	0.977	1.036	1.918	1.206	0.673	4	3
15	SRNN-VE-2	1.182	1.013	1.078	1.873	1.281	0.647	6.167	6
16	SRNN-VE-3	**1.106**	**0.965**	**1.019**	1.913	1.212	0.662	**3.167**	**1**
α = 2.5%									
ID	Model name	Average loss						Average rank	Rank
		S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY		
1	RW-125	1.087	0.945	1.020	1.692	1.025	0.499	12.5	11
2	RW-250	1.079	0.986	1.020	1.757	1.043	0.462	13.833	15
3	RW-500	1.153	1.072	1.103	1.898	1.012	0.530	14.833	16
4	GCH-N	0.957	1.001	0.868	1.663	0.998	0.413	7.833	9
5	GCH-Skt	0.920	0.933	0.835	1.664	**0.975**	0.413	4.167	3
6	GCH-EDF	0.920	0.872	0.835	1.661	0.976	**0.411**	4	1
7	FZ2F	0.917	0.952	0.869	1.661	1.031	0.440	8.167	10
8	FZ1F	0.935	0.853	0.858	**1.645**	1.011	0.429	6.5	7
9	GCH-FZ	0.918	0.867	0.827	1.673	0.997	0.423	5	5
10	Hybrid	0.921	0.834	0.852	1.674	1.021	0.428	6.833	8
11	RNN-VE-1	1.143	0.975	0.998	1.748	1.024	0.464	12.667	13
12	RNN-VE-2	1.068	0.976	0.999	1.751	1.042	0.463	12.833	14
13	RNN-VE-3	1.078	0.988	0.992	1.748	1.028	0.464	12.5	11
14	SRNN-VE-1	0.896	0.833	0.823	1.680	0.977	0.425	**4**	**1**
15	SRNN-VE-2	0.967	0.848	0.840	1.677	0.981	0.420	6.167	6
16	SRNN-VE-3	**0.893**	**0.825**	**0.818**	1.684	0.978	0.429	4.167	3

Notes: This table presents average FZ0 losses and model rankings over the out-of-sample period from 1 Jan 2010 to 31 Dec 2018, for 1% and 2.5% α values, respectively. The in-sample period is from 1 January 2000 to 31 December 2009. The lowest values for columns 3–10 is shown in bold. Column 9 present the average ranks across 6 different assets. Column 10 presents model rankings (with the best-performing model ranked 1 and the worst ranked 16) based on the column 9.

et al. (2019), can be found in Table 1.

$$\begin{aligned} v_t^{\alpha} &= a_{\alpha}\sigma_t \\ e_t^{\alpha} &= b_{\alpha}\sigma_t \\ b_{\alpha} &< a_{\alpha} < 0 \end{aligned} \tag{28}$$

Twenty return series are generated for each pair of parameters (DoF, Skewness), each consisting of 10,000 returns. The simulated data is divided into three segments: a training set consisting of the first 3750 data points, a validation set made up of the subsequent 1250 data points, and an out-of-sample set comprising the last 5000 data points. The validation set is used to implement early stopping to terminate the training process when the loss on the validation set reaches a plateau.

After training, we forecast one-day-ahead VaR and ES for each series for the out-of-sample dataset. Fig. 2 displays the actual 1%-VaR, calculated using Eq. (28), as well as the 1%-VaR estimated using the SRNN-VE-1, SRNN-VE-2, and SRNN-VE-3 models, respectively, for the first simulated return series. Fig. 3 presents the true 1%-ES, obtained using Eq. (28), alongside the predicted 1%-ES. We calculate the average out-of-sample FZ0 loss score and average linear correlation coefficient over twenty simulated return series for each (DoF, Skewness) parameter combination. The results are summarized in Table 2; these show that the VaR and ES series predicted by all three RNN-based stateful models exhibit high correlations (mostly above 85%) with the true values, and the average FZ0 loss values are close to the true loss values, indicating that the models are able to capture the tail behavior of Y_t. As anticipated, in Table 2, the losses of the RNN-based stateful estimated models are higher than, but not far from, the true loss of the GARCH with skewed t-distribution that was used as data generating process (DGP), so the estimated models slightly underperform the DGP. Also, as anticipated, the SRNN-VE-2 model, owing to its resemblance to the GARCH models, outperforms the other stateful models across all levels of α.

5. Empirical study

In this section, we evaluate the performance of the three RNN-based stateful models and three non-stateful RNN models discussed in Section 3.4 on daily financial returns. Additionally, we conduct an evaluation of ten popular risk models which include the rolling window models with window length 125, 250 and 500 days, denoted by RW-125, RW-250, and RW-500, GARCH-type models based on the Normal, Skewed t and EDF distributions, denoted by GCH-N, GCH-Skt, and GCH-EDF, and four GAS models proposed by Patton et al. (2019), specifically the FZ2F, FZ1F, GCH-FZ, and Hybrid model, which serve as benchmark models (so we consider a total of 13 benchmark models).

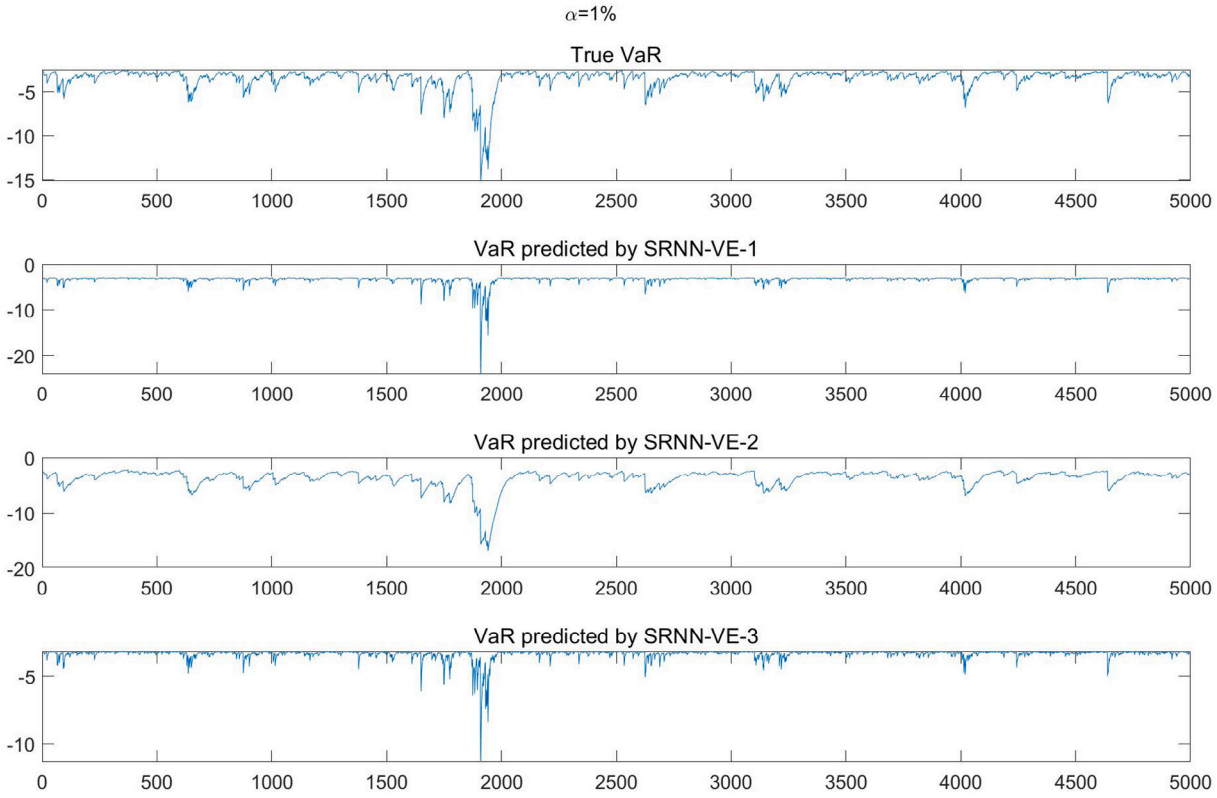

Fig. 2. The true and forecasted 1%-VaR of the daily return series simulated via GARCH(1,1) with skewed t-distribution (DoF = 5, Skewness = −0.5) for the three stateful models on the out-of-sample set.

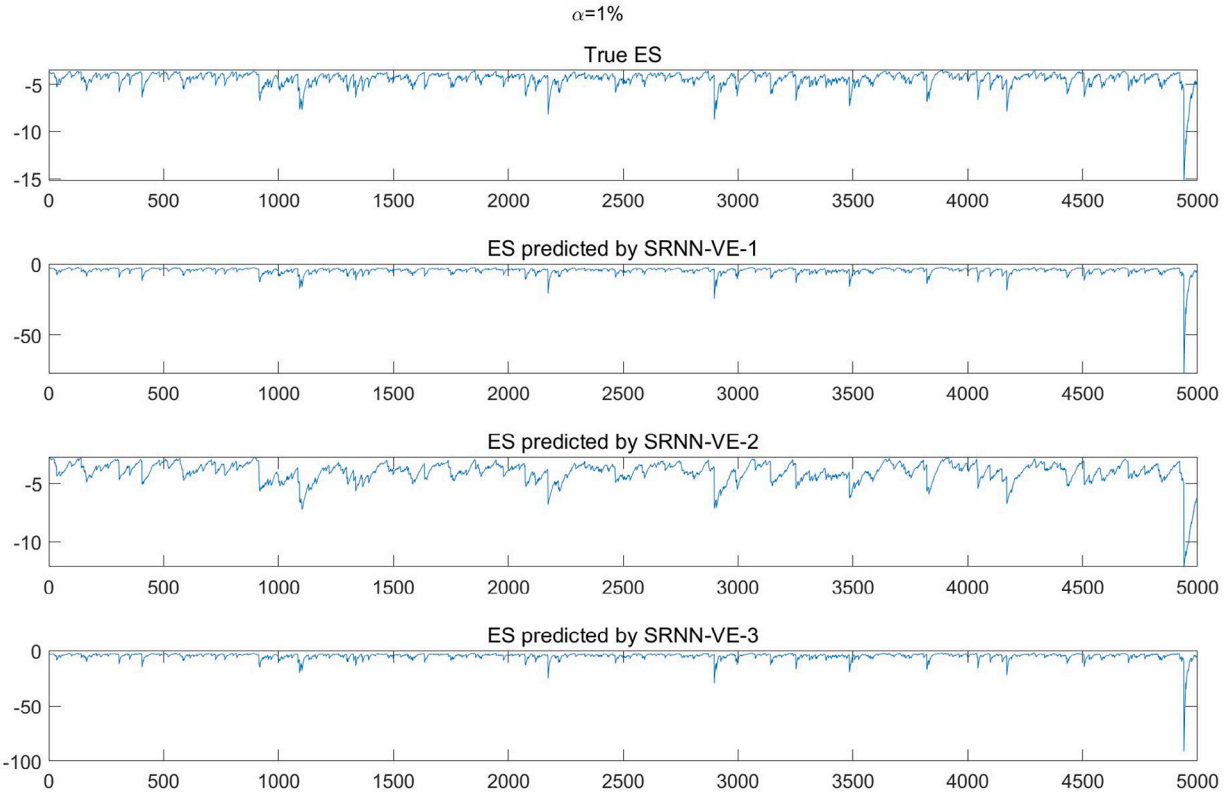

Fig. 3. The true and forecasted 1%-ES of the daily return series simulated via GARCH(1,1) with skewed t-distribution (DoF = 5, Skewness = −0.5) for the three stateful models on the out-of-sample set.

Table 11
Average losses and ranks of the models over the period 1 Jan 2021 to 31 May 2022.

α = 1%									
ID	Model name	Average loss						Average rank	Rank
		S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY		
1	RW-125	1.313	**1.412**	1.136	2.584	1.102	0.370	6.833	6
2	RW-250	1.660	1.460	1.462	2.631	1.151	0.426	10.167	12
3	RW-500	1.825	1.417	1.881	2.530	1.280	0.438	11.5	15
4	GCH-N	1.565	2.108	1.401	2.488	1.308	**0.172**	11.333	13
5	GCH-Skt	1.288	1.645	1.160	2.334	1.119	0.280	5.833	2
6	GCH-EDF	1.289	1.648	1.141	2.330	1.174	0.314	6.667	5
7	FZ2F	1.160	1.597	1.226	2.974	1.104	0.355	6.833	6
8	FZ1F	**1.116**	1.493	1.359	3.076	**1.078**	0.347	6.333	4
9	GCH-FZ	1.211	1.725	1.125	2.413	1.187	0.292	6.167	3
10	Hybrid	1.207	1.766	1.344	3.037	1.080	0.351	8.5	8
11	RNN-VE-1	1.285	1.704	1.322	3.464	1.208	0.495	12.333	16
12	RNN-VE-2	1.184	1.495	**1.039**	3.137	1.229	0.487	8.5	8
13	RNN-VE-3	1.278	1.769	1.260	3.338	1.164	0.482	11.333	13
14	SRNN-VE-1	1.220	1.913	1.152	2.436	1.294	0.366	9.833	11
15	SRNN-VE-2	1.203	1.656	1.118	**2.330**	1.191	0.300	**5.167**	**1**
16	SRNN-VE-3	1.154	1.845	1.163	2.430	1.243	0.368	8.667	10
α = 2.5%									
ID	Model name	Average loss						Average rank	Rank
		S&P 500	FTSE	DJIA	OilSpot	GoldSpot	USDJPY		
1	RW-125	1.141	1.216	0.956	2.208	**0.928**	−0.017	5.333	2
2	RW-250	1.359	1.191	1.212	2.262	0.934	0.069	8.667	9
3	RW-500	1.484	1.260	1.446	2.263	0.953	0.193	11.5	14
4	GCH-N	1.216	1.408	1.051	2.035	1.046	**−0.042**	10.167	11
5	GCH-Skt	1.102	1.272	0.970	**1.993**	0.978	0.002	6.167	4
6	GCH-EDF	1.082	1.258	0.957	1.993	1.002	0.041	5.833	3
7	FZ2F	1.169	1.230	0.936	2.141	1.015	−0.030	6.667	6
8	FZ1F	1.238	1.235	1.074	2.274	0.980	0.088	10.167	11
9	GCH-FZ	1.080	1.266	0.924	2.028	1.003	0.038	6.167	4
10	Hybrid	1.118	1.311	1.065	2.278	1.045	0.075	11.5	14
11	RNN-VE-1	1.172	1.232	0.999	2.736	1.025	0.247	11.5	14
12	RNN-VE-2	1.103	**1.183**	0.929	2.437	1.115	0.247	9.333	10
13	RNN-VE-3	1.149	1.259	0.977	2.529	1.129	0.250	12.667	16
14	SRNN-VE-1	1.059	1.285	0.918	2.023	1.053	0.100	8	7
15	SRNN-VE-2	**1.034**	1.254	**0.896**	1.998	1.001	0.066	**4.167**	**1**
16	SRNN-VE-3	1.049	1.379	0.912	2.070	1.025	0.106	8.167	8

Notes: This table presents average FZ0 losses and model rankings over the out-of-sample period from 1 Jan 2021 to 31 May 2022, for 1% and 2.5% α values, respectively. The in-sample period is from 1 January 2015 to 31 December 2020. The lowest values for columns 3–10 is shown in bold. Column 9 present the average ranks across 6 different assets. Column 10 presents model rankings (with the best-performing model ranked 1 and the worst ranked 16) based on the column 9.

5.1. Data description

Our analysis is based on six assets, which consist of three equity indices (S&P 500, FTSE, and DJIA), two commodities (oil spot price and gold spot price), and one exchange rate USDJPY, spanning from 1 January 2000 to 31 May 2022.[8] We estimate the model parameters using the first ten years (1 January 2000 to 31 Dec 2009) and reserve the remaining 13 years of the data for evaluation and model comparison. In our empirical study, we use the square of the demeaned log return as the input of the RNN-based models, which is calculated using Eq. (29). As shown in Table 3, training each machine learning model is not very time consuming. This demonstrates the time efficiency of the six RNN-based models.

$$r_t = log(P_t) - log(P_{t-1}) \tag{29}$$

Table 4 displays the summary statistics of the six daily asset return series for the full sample period. The first three rows report the total number of observations for the full sample period, the number of observations for the in-sample period, and the number of observations for the out-of-sample period. The table also presents the sample Value-at-Risk estimates for four distinct choices of α and the corresponding sample Expected Shortfall estimates.

5.2. Model comparison

This section compares the performance of the RNN-based stateful models discussed in Section 3.4 with the thirteen benchmark models, including a comparison based on backtesting criteria.

Table 5 displays the parameter estimates for the autoregressive moving average ARMA(p,q) model and the GARCH(1,1) model combined with the skewed t-distribution. The ARMA order (p,q) is determined based on the Bayesian information criterion (BIC). In the second panel, the estimated parameters of the GARCH(1,1) model, as well as the degrees of freedom (DoF) and skewness, are presented.

Table 6 presents the parameter estimates as well as their corresponding standard errors for the GAS models proposed by Patton et al. (2019) used to forecast VaR and ES for the S&P 500 daily return series at 1% α level. The analysis was conducted over the in-sample period ranging from January 2000 to December 2009. The left panel of the table outlines the results of the two-factor GAS model (FZ2F)

[8] The indices daily close prices are obtained from https://realized.oxfordman.ox.ac.uk/ accessed in 2022; the two commodity prices and the exchange rate USDJPY are obtained from Bloomberg; the ticker name for oil spot price and gold spot price are XAU Curncy and USCRWTIC index, respectively.

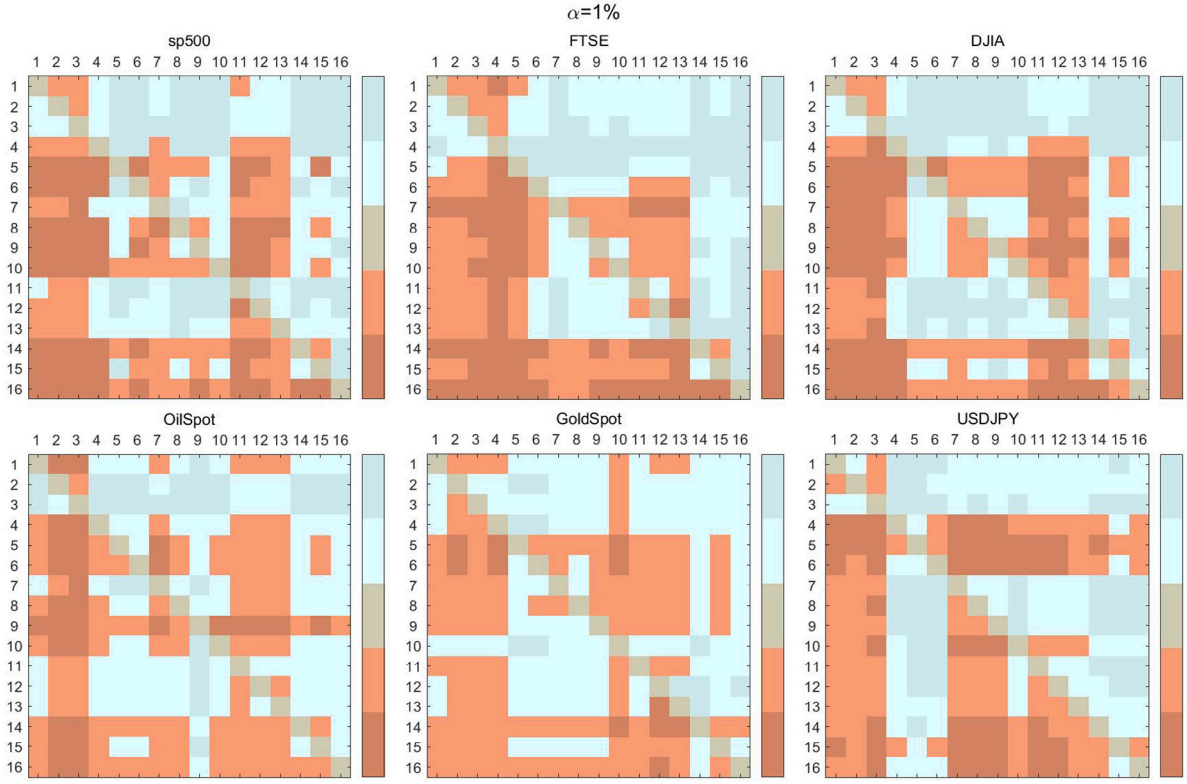

Fig. 4. Color map based on the DM test comparing the average losses over the out-of-sample period for 16 different models estimated on the six different assets, S&P 500, FTSE, DJIA, Oil Spot prices, Gold Spot prices, and USDJPY exchange rate at 1% level of risk. Dark orange blocks mean that the row model has a lower average loss than the column model at 5% significance level; light orange blocks mean that the row model has a lower average loss than the column model, but the difference is not significant at 5% level. Blue blocks mean that the row model has a higher average loss than the column model, with the darkest shade denoting a difference that is significant at 5% level. The numbering of the models is based on the ID numbers given in Table 8.

whereas the right panel contains the outcomes of three models, the one-factor GAS model (FZ1F), the GARCH model estimated using FZ loss minimization (GCH-FZ), and the hybrid-factor GAS model (Hybrid). The bottom row of the table shows the average (in-sample) FZ0 losses of all the aforementioned models.

Table 7 displays the number of rejections at significance levels of 1% and 5% for both the dynamic quantile (DQ) test and the dynamic ES (DES) regression test proposed by Engle and Manganelli (2004) which evaluate the performances of VaR and ES forecasts for the six assets. The DQ regression test with one lag is based on the equation presented in Eq. (30), with the test providing valuable insights into the reliability of the analyzed models for risk estimation.

$$Hit_{\alpha,t} = w_0 + w_1 Hit_{\alpha,t-1} + w_2 v_{t-1} + u_t \tag{30}$$

where the $Hit_{\alpha,t}$ is defined as $Hit_{\alpha,t} = \mathbb{1}\{Y_t \leq v_t\} - \alpha$, and u_t is the regression residual. The DES regression test is based on the following regression:

$$\lambda^S_{e,t} = b_0 + b_1 \lambda^S_{e,t-1} + b_2 e_{t-1} + u_{e,t} \tag{31}$$

where the $\lambda^S_{e,t}$ is defined as $\lambda^S_{e,t} = \mathbb{1}\{Y_t \leq v_t\}\frac{Y_t}{e_t} - 1$, and $u_{e,t}$ is the regression residual. The results show that the SRNN-VE-3 model exhibits the fewest rejections in the DQ test, at both 1% and 5% significance levels, for both 1% and 2.5% α levels. In the DES test, the SRNN-VE-3 model demonstrates a performance that is comparable to that observed in the DQ test. Moreover, the stateful models demonstrate a higher likelihood of passing both the DQ and DES tests compared to their corresponding non-stateful counterparts.

Table 8 presents rankings based on average FZ0 loss values obtained for six daily return series over the out-of-sample period for 16 distinct forecasting models, with the best-performing model ranked 1 and the worst is ranked 16. Column 10 indicates the average rank across the six return series for both $\alpha = 1\%$ and $\alpha = 2.5\%$. Notably, the SRNN-VE-3 model is consistently ranked first, and the SRNN-VE-1 model ranked second for α values of 1% and 2.5%. However, the corresponding non-stateful models (RNN-VE-3, and RNN-VE-1) exhibit low rankings. Similarly, the SRNN-VE-2 model consistently outperforms the RNN-VE-2 model.

Table 9 presents a summary of the average FZ0 loss scores, along with the corresponding p-values of the DQ test and DES test.[9] The table displays the results on six different assets' daily return series for risk levels of $\alpha = 1\%$ and $\alpha = 2.5\%$. The column with the lowest average loss is highlighted in bold, while the second lowest is in italics. The p-values of the dynamic regression tests of the VaR forecasts are presented in columns 9 to 14, and columns 15 to 20 display the p-values for ES. Any values greater than 0.05 are highlighted in bold to indicate absence of evidence against optimality at 5% level. The table shows that the RNN-based stateful model exhibits the best performance among all models for the S&P 500, FTSE, DJIA, Gold Spot price, and USDJPY return series for $\alpha = 1\%$. For risk level $\alpha = 2.5\%$, except for the Oil Spot price and USDJPY assets, at least one out of the three stateful models has the lowest average loss among all models. Especially, the SRNN-VE-3 model consistently exhibits the lowest average loss across all three indices, for both 1% and 2.5% α levels.

The Diebold–Mariano (DM) test, initially introduced by Diebold and Mariano (1995), constitutes a statistical test employed to compare the forecast accuracy of two competing forecasting models. There is a large amount of literature demonstrating how the DM test helps identifying statistically different forecasting performances of two models,

[9] More details on the DQ test and DES test can be found in Engle and Manganelli (2004) and Patton et al. (2019).

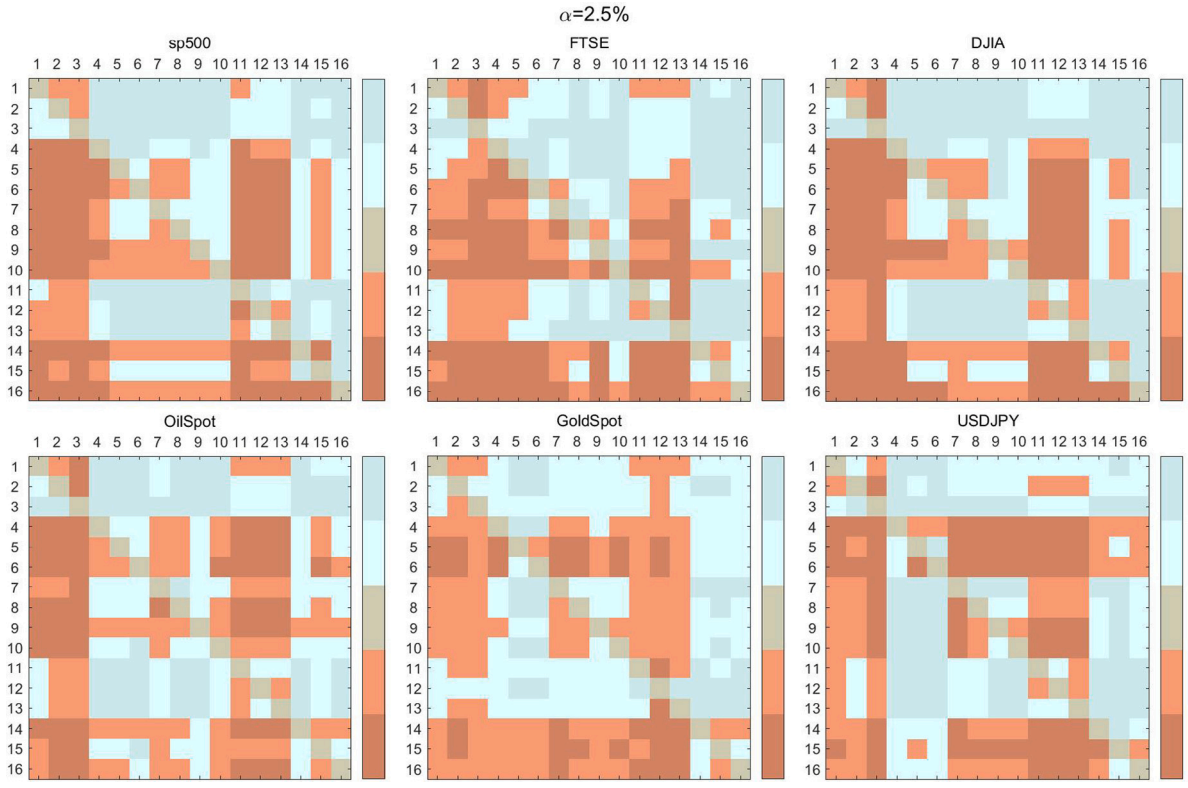

Fig. 5. Color map based on the DM test comparing the average FZ0 losses over the out-of-sample period for 16 different models estimated on the six different assets, S&P 500, FTSE, DJIA, Oil Spot prices, Gold Spot prices, and USDJPY exchange rate at 2.5% level of risk. Dark orange blocks mean that the row model has a lower average loss than the column model at 5% significance level; light orange blocks mean that the row model has a lower average loss than the column model, but the difference is not significant at 5% level. Blue blocks mean that the row model has a higher average loss than the column model, with the darkest shade denoting a difference that is significant at 5% level. The numbering of the models is based on the ID numbers given in Table 8.

see Diebold (1998), Chen et al. (2014), Mariano and Preve (2012) and Patton et al. (2019). In the paper, the DM test is used to compare the average FZ0 losses of two models, with the results displayed in Figs. 4 and 5. A negative t-statistic of this test indicates that the row model exhibits lower average FZ0 loss than the column model. The critical value of 1.96 is used to identify differences that are significant at the 95% confidence level. The light orange shading in Figs. 4 and 5 signifies that the row models outperform the column model, but the loss difference is not statistically significant. The dark orange shading indicates that the row model exhibits significantly lower average FZ0 loss compared to the column model, at 95% confidence level. The light blue color indicates that the row model exhibits a marginally higher average FZ0 loss compared to the column model; however, the difference is not found to be statistically significant. Conversely, the dark blue color signifies a statistically significant difference, indicating that the row model presents a markedly higher average FZ0 loss than the column model, at a confidence level of 95%. The first ten rows denote the ten benchmark models, with the model numbering in Figs. 4 and 5 following the numbering in column 'ID' in Tables 8 and 9. The results show evidence that the stateful models generally outperform the benchmark models.

In conclusion, a comprehensive analysis combining loss ranking results and DM test results indicates that the SRNN-VE-3 model and SRNN-VE-1 model exhibit outstanding performance, ranked first and second, respectively, compared to the benchmark models. Conversely, the SRNN-VE-2 model displays superior performance for specific series, such as the USDJPY currency for risk level $\alpha = 1\%$, but fails to outperform the SRNN-VE-1 and SRNN-VE-3 models for the majority of the analyzed series. Furthermore, it is consistently observed that the stateful RNN models outperform their corresponding non-stateful RNN models.

5.3. Robustness test

We conduct two robustness tests based on shorter sample periods. The first robustness test uses the same in-sample period with length 10 years, and a shorter out-of-sample period, from 1 January 2010 to 31 December 2018. Taking into consideration that potential regime shifts can occur during this period which might affect the performance of the models, our second robustness test employs a shorter in-sample period, from 1 January 2015 to 31 December 2020, as well as a shorter out-of-sample period, 1 January 2021 to 31 May 2022.

The main results of these two robustness tests described above are presented in Tables 10 and 11, for 1% and 2.5% α values, respectively. In these tables, columns 3–8 present the average FZ0 losses. Column 9 presents the average ranks across 6 different assets, which are based on the average FZ0 losses for 6 daily return series over the out-of-sample period for 16 different forecasting models. Column 10 presents model rankings (with the best-performing model ranked 1 and the worst ranked 16) based on Column 9. From these results it can be concluded that the stateful models consistently outperform their non-stateful counterparts, and they also demonstrate superior performance compared to the other benchmarks models. The GARCH-based models (GCH-Skt, GCH-EDF, and GCH-FZ) have also performed well. However, the SRNN-VE-2 model stands out as the top performer in the last out-of-sample data period (2021–22) considered.

6. Conclusion

This paper proposes three novel models for the joint estimation of Value-at-Risk and Expected Shortfall, employing the stateful Recurrent Neural Network and Feedforward Neural Network Machine Learning frameworks. Competing methods mostly employ econometric models

such as GARCH-type models and GAS models to build joint forecasts of VaR and ES. In contrast, the proposed stateful models do not rely on specific model assumptions, in particular there is no need to make distribution assumptions for the error term, thereby exhibiting conceptual advantages over conventional econometric approaches.

The simulation study demonstrates that the stateful models predict risk levels that are close to the actual values, and the average FZ0 loss values are close to the true loss values. This indicates that these models are able to successfully capture the distributional tail behavior of returns.

We put forward an empirical application in which we forecast VaR and ES jointly on daily return series of six assets (including three international stock indices, two commodities, and one currency), based on a dataset ranging 22 years from January 2000 to May 2022. Comparing the performance of the RNN-based stateful models with thirteen alternative models, this paper provides evidence to support that the RNN-based stateful models outperform the benchmark models in terms of VaR and ES forecasting. The empirical results indicate that enabling the RNN structure to be stateful enhances the ability of the models to learn the tail risk structure of financial data.

The risk models presented in this article hold practical and theoretical significance for financial practitioners, regulatory authorities, and the academic community, providing improvements in risk management and prediction. Practitioners and risk managers can benefit from the proposed methodology, obtaining more accurate and timely risk estimates to enhance their ability to formulate effective risk management strategies. This is particularly crucial for professionals making rapid decisions in an ever-changing market environment. Additionally, regulatory authorities can derive benefits from this study as more accurate risk assessments contribute to the formulation of precise regulatory policies, ensuring stability and transparency in financial markets. Furthermore, the insights of this study can generate research directions for academics to consider other potential applications of RNN-based stateful models in finance.

The RNN-based stateful models, consisting of one stateful RNN layer and one FNN layer, carry low computational costs. As such, it could be worthwhile evaluating the performance of the models for stock returns (which are known to be more volatile) or other financial returns. Also, the potential performance enhancements associated with deeper networks remain an area for further investigation. Furthermore, a limitation of our study is the lack of a multivariate approach which would be worthwhile to consider for future research, with potential applications for fast portfolio risk estimation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abad, P., Benito, S., & López, C. (2014). A comprehensive review of value at risk methodologies. *The Spanish Review of Financial Economics*, *12*(1), 15–32.

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, *26*(7), 1487–1503.

Alberg, D., Shalit, H., & Yosef, R. (2008). Estimating stock market volatility using asymmetric GARCH models. *Applied Financial Economics*, *18*(15), 1201–1208.

Aliyev, F., Ajayi, R., & Gasim, N. (2020). Modelling asymmetric market volatility with univariate GARCH models: Evidence from nasdaq-100. *The Journal of Economic Asymmetries*, *22*, Article e00167.

Almeida, D. d., & Hotta, L. K. (2014). The leverage effect and the asymmetry of the error distribution in GARCH-based models: The case of Brazilian market related series. *Pesquisa Operacional*, *34*, 237–250.

Ardia, D., Bluteau, K., Boudt, K., & Catania, L. (2018). Forecasting risk with Markov-switching GARCH models: A large-scale performance study. *International Journal of Forecasting*, *34*(4), 733–747.

Arian, H., Moghimi, M., Tabatabaei, E., & Zamani, S. (2022). Encoded value-at-risk: A machine learning approach for portfolio risk measurement. *Mathematics and Computers in Simulation*, *202*, 500–525.

Artzner, P. (1997). Thinking coherently. *Risk*, *10*, 68–71.

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 (2010). Basel III: A global regulatory framework for more resilient banks and banking systems, bank for international settlements. http://www.bis.org/publ/bcbs189.pdf.

Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: A revised market risk framework. https://www.bis.org/publ/bcbs265.pdf,

BenSaïda, A., Boubaker, S., Nguyen, D. K., & Slim, S. (2018). Value-at-risk under market shifts through highly flexible models. *Journal of Forecasting*, *37*(8), 790–804.

Billio, M., & Pelizzon, L. (2000). Value-at-risk: A multivariate switching regime approach. *Journal of Empirical Finance*, *7*(5), 531–554.

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

Bucevska, V. (2013). An empirical evaluation of GARCH models in value-at-risk estimation: Evidence from the Macedonian stock exchange. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, *4*(1), 49–64.

Chan, K. F., & Gray, P. (2006). Using extreme value theory to measure value-at-risk for daily electricity spot prices. *International Journal of Forecasting*, *22*(2), 283–300.

Chen, H., Wan, Q., & Wang, Y. (2014). Refined diebold-mariano test methods for the evaluation of wind power forecasting models. *Energies*, *7*(7), 4185–4198.

Chronopoulos, I., Raftapostolos, A., & Kapetanios, G. (2023). Forecasting value-at-risk using deep neural network quantile regression. *Journal of Financial Econometrics*, [forthcoming].

Cont, R., Cucuringu, M., Xu, R., & Zhang, C. (2022). TAIL-GAN: Nonparametric scenario generation for tail risk estimation. arXiv preprint arXiv:2203.01664.

Degiannakis, S., Floros, C., & Dent, P. (2013). Forecasting value-at-risk and expected shortfall using fractionally integrated models of conditional volatility: International evidence. *International Review of Financial Analysis*, *27*, 21–33.

Delbaen, F. (2002). Coherent risk measures on general probability spaces. In *Advances in finance and stochastics: Essays in honour of dieter sondermann* (pp. 1–37). Springer.

Diebold, F. X. (1998). *Elements of forecasting*. Citeseer.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–263.

Duffie, D., & Pan, J. (1997). An overview of value-at-risk. *Journal of Derivatives*, *4*(3), 7–49.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007.

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.

Gerlach, R., & Wang, C. (2020). Semi-parametric dynamic asymmetric Laplace models for tail risk forecasting, incorporating realized measures. *International Journal of Forecasting*, *36*(2), 489–506.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Guermat, C., & Harris, R. D. (2002). Forecasting value-at-risk allowing for time variation in the variance and kurtosis of portfolio returns. *International Journal of Forecasting*, *18*(3), 409–419.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, *35*(3), 705–730.

Hartz, C., Mittnik, S., & Paolella, M. (2006). Accurate value-at-risk forecasting based on the normal-GARCH model. *Computational Statistics & Data Analysis*, *51*(4), 2295–2312.

Holton, G. (2003). *Academic press advanced finance series*, *Value-at-risk: Theory and practice*. Academic Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Huang, D., Yu, B., Fabozzi, F. J., & Fukushima, M. (2009). CAViaR-based forecast for oil price risk. *Energy Economics*, *31*(4), 511–518.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, *49*(3), 851–889.

Iva¸scu, C.-F. (2021). Option pricing using machine learning. *Expert Systems with Applications*, *163*, Article 113799.

Khan, M. A. I. (2011). Modelling daily value-at-risk using realized volatility, non-linear support vector machine and ARCH type models. *Journal of Economics and International Finance*, *3*(5), 305.

Koenker, R., & Bassett, Jr., G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Lu, X., Ma, F., Xu, J., & Zhang, Z. (2022). Oil futures volatility predictability: New evidence based on machine learning models. *International Review of Financial Analysis*, *83*, Article 102299.

Lucas, A., & Zhang, X. (2016). Score-driven exponentially weighted moving averages and value-at-risk forecasting. *International Journal of Forecasting*, *32*(2), 293–302.

Mandelbrot, B. B., & Mandelbrot, B. B. (1997). *The variation of certain speculative prices*. Springer.

Mariano, R. S., & Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics*, *169*(1), 123–130.

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: Toncepts, techniques and tools revised edition*. Princeton University Press.

Ormaniec, W., Pitera, M., Safarveisi, S., & Schmidt, T. (2022). Estimating value-at-risk: LSTM vs. GARCH. arXiv preprint arXiv:2207.10539.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259–268.

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.

Philippe, J. (2001). *MacGraw-Hill international editions: Finance series*, *Value at risk: The new benchmark for managing financial risk*. McGraw-Hill.

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

Saha, S., Gao, J., & Gerlach, R. (2021). Stock movement prediction on ex-dividend day using event specific features and machine learning techniques. In *2021 International joint conference on neural networks* (pp. 1–10). IEEE.

Shim, J., Kim, Y., Lee, J., & Hwang, C. (2012). Estimating value-at-risk with semiparametric support vector quantile regression. *Computational Statistics*, *27*(4), 685–700.

So, M. K., & Philip, L. (2006). Empirical analysis of GARCH models in value-at-risk estimation. *Journal of International Financial Markets, Institutions and Money*, *16*(2), 180–197.

Sollis, R. (2009). Value-at-risk: A critical overview. *Journal of Financial Regulation and Compliance*, *17*(4), 398–414.

Taylor, J. W. (2008). Estimating value-at-risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, *6*(2), 231–252.

Taylor, J. W. (2019). Forecasting value-at-risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, *37*(1), 121–133.

Taylor, J. W. (2022). Forecasting value-at-risk and expected shortfall using a model with a dynamic omega ratio. *Journal of Banking & Finance*, *140*, Article 106519.

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Implied volatility directional forecasting: A machine learning approach. *Quantitative Finance*, *21*(10), 1687–1706.

Wu, Q., & Yan, X. (2019). Capturing deep tail risk via sequential learning of quantile dynamics. *Journal of Economic Dynamics & Control*, *109*, Article 103771.

Xu, Q., Liu, X., Jiang, C., & Yu, K. (2016). Nonparametric conditional autoregressive expectile model via neural network with applications to estimating financial risk. *Applied Stochastic Models in Business and Industry*, *32*(6), 882–908.

Ye, T., & Zhang, L. (2019). Derivatives pricing via machine learning. *Boston University Questrom School of Business Research Paper*, *3352688*.

Zhang, L. (2020). A general framework of derivatives pricing. *Journal of Mathematical Finance*, *10*(2), 255–266.

Zhang, N., Su, X., & Qi, S. (2023). An empirical investigation of multiperiod tail risk forecasting models. *International Review of Financial Analysis*, *86*, Article 102498.