A Comparison of Methods for Forecasting Value-at-Risk and Expected Shortfall of Cryptocurrencies

Carlos Trucíos*[1] and James W. Taylor[2]

[1]Department of Statistics. University of Campinas, Brazil.

[2]Saïd Business School. University of Oxford, United Kingdom.

September 22, 2022

Abstract

Several procedures to forecast daily risk measures in cryptocurrency markets have been recently implemented in the literature. Among them, long-memory processes, procedures taking into account the presence of extreme observations, procedures that include more than a single regime, as well as quantile regression-based models have performed substantially better than standard methods in terms of forecasting risk measures. Those procedures are revisited in this paper, and their Value-at-Risk and Expected Shortfall forecasting performance are evaluated using recent Bitcoin and Ethereum data that includes periods of turbulence due to the COVID-19 pandemic, the third halving of Bitcoin and the Lexia class action. Additionally, in order to mitigate the influence of model misspecification and enhance the forecasting performance obtained by individual models, we evaluate the use of several forecast combining strategies. Our results, based on a comprehensive backtesting exercise, reveal that, for Bitcoin, there is no single procedure outperforming all other models, but for Ethereum, there is evidence showing that the GAS model is a suitable alternative for forecasting both risk measures. We found that the combining methods were not able to outperform the better of the individual models.

Keywords: digital assets, forecast combining, model misspecification, outliers, risk measures, structural breaks.

JEL classification: C10, C22, C53, G17, G32

*Corresponding author: ctrucios@unicamp.br

The authors thank the computational support of the Centre for Applied Research on Econometrics, Finance and Statistics (CAREFS), Brazil.

1 Introduction

Bitcoin, the first and major cryptocurrency, was introduced in 2008 by Nakamoto (2008) as a way to facilitate electronic payments between individuals without going through a third party. Since its inception, the cryptocurrency market has increased considerably and up to now, there are more than 21,000 cryptocurrencies, summing up to a total market capitalisation of more than nine hundred billion dollars[1].

The development and expansion of digital currencies around the world are attributed, among other reasons, to their decentralised nature, their low transaction cost and the loss of trust by individuals in the monetary system. See; for instance Dyhrberg (2016), Bouri et al. (2017) and Luther and Salter (2017) for detailed discussions.

The impressive growth in cryptocurrency markets in recent years has attracted the attention of investors, financial regulators, policy-makers, companies, central banks and country governments. For instance; a number of markets, including the Chicago Mercantile Exchange, NASDAQ and the Tokyo Financial Exchange started to trade cryptocurrencies some months ago; the number of crypto-exchanges (platforms where cryptocurrencies are traded) is rising worldwide; El Salvador adopted Bitcoin as a legal tender and the crypto ATMs industry is expanding worldwide[2]. These pieces of evidence reflect the increasing interest in cryptocurrency markets as well as the important role they are playing nowadays.

Whilst traditional markets are regulated and risk measures are widely used in financial institutions and also required by the Basel II and Basel III accords, cryptocurrency markets are not regulated yet and the formal use of risk measures is not required. However, the study of risk measures in cryptocurrency markets is important from the point of view of investors, hedge funds, market makers, and traders, since they are useful for placing better order limits, devising option pricing strategies, and developing trading systems. Furthermore, consideration of risk measurement for cryptocurrencies is crucial for defining future regulatory policies.

Unlike traditional markets where several procedures to estimate risk measures have been successfully applied, only a few procedures have been shown to be useful in cryptocurrency markets. The lack of fit obtained by several volatility models is attributed

[1]Source: coinmarketcap.com, 20 September 2022.

[2]According to coinatmradar.com, on September 20 of 2022, there are more than 38500 crypto ATMs spread over 78 countries.

to the presence of extreme observations and regime changes in the volatility dynamics, two characteristics observed in the cryptocurrency data. See; for instance, Troster et al. (2019), Ardia et al. (2019b), Alexander and Dakos (2020), Trucíos (2019), Caporale and Zekokh (2019), Liu et al. (2020) and Maciel (2020).

The fact that extreme observations and structural breaks can badly affect volatility model forecasting performance is discussed, in a broader context, by Hillebrand (2005), Bauwens et al. (2010), Carnero et al. (2012), Boudt et al. (2013), Trucíos and Hotta (2016), Ardia et al. (2018), Hotta and Trucíos (2018) among others. This poor performance is explained by the fact that after crisis periods or large shocks (where outliers and/or structural breaks arise) risk tends to be overestimated, because the influence of those observations remains in the estimation period for a long time, affecting drastically the volatility estimation and consequently the risk forecasts. See; Danielsson (2011) and Harvey (2013) for interesting related discussions.

As mentioned previously, several procedures to forecast risk measures in cryptocurrency data have been implemented in recent years, but only a few have been shown to be useful. For instance, in a Bitcoin volatility context, Troster et al. (2019) and Trucíos (2019) make a comprehensive comparison of several volatility models and conclude that robust-to-outliers procedures outperform non-robust ones. Both provide evidence in favour of generalised autoregressive score (GAS) models (Harvey, 2013; Creal et al., 2013) while Trucíos (2019) also provides evidence supporting the use of a robust bootstrap GARCH-based model. On the other hand, Ardia et al. (2019b) compares regime-switching models against single-regime ones and concludes that the former outperform the latter. The findings of Ardia et al. (2019b) are also supported by Alexander and Dakos (2020) who reach the same conclusion considering prices of Bitcoin and Ethereum in several crypto-exchanges such as, Bitfinex, Coinbase, Gemini, Kraken and Poloniex.

In terms of Value-at-Risk (VaR) forecasting accuracy for cryptocurrency data, Ardia et al. (2019b), Caporale and Zekokh (2019) and Maciel (2020) conclude that multiple-regime models are better than single-regime for cryptocurrency data. Troster et al. (2019) reports evidence in favour of GAS against GARCH-type models when forecasting the VaR of Bitcoin. Trucíos (2019) shows that the robust bootstrap procedure of Trucíos et al. (2017) forecasts the VaR of Bitcoin more accurately than GAS and GARCH-type models. Liu et al. (2020) concludes that IGAS models (a GAS analogue of IGARCH models) are

good alternatives to predict the VaR of Bitcoin, Ethereum and Litecoin. Soylu et al. (2020) finds evidence favourable to long-memory volatility processes over short-memory ones. Additionally, Li et al. (2021) find benefit in forecasting the VaR of Bitcoin using quantile regression-based models, namely the conditional autoregressive value at risk (CAViaR) models of Engle and Manganelli (2004).

In an Expected Shortfall (ES) context, the cryptocurrency literature is even more scarce, with only a small number of studies having been performed. Acereda et al. (2020) uses GARCH-type models under different innovation distributions and concludes that NAGARCH and CGARCH models with heavy tailed distribution are good alternatives to forecast the ES. Soylu et al. (2020) finds evidence favourable to long-memory volatility processes, and Caporale and Zekokh (2019) and Maciel (2020) conclude that multiple-regime models outperform single-regime ones for cryptocurrency data.

Although the literature provides some evidence in support of certain individual models for forecasting risk measures for cryptocurrency data, there is no clear consensus about which model is best. In terms of building a better model based on the more successful individual models, it is not obvious how to proceed. In light of the good results obtained in several fields when applying forecast combination strategies (for example, see Atiya, 2020; Thomson et al., 2019) and due to its ability to reduce the misspecification influence of individual models as well as synthesise the diverse sources of information (Timmermann, 2006), we evaluate the use of forecast combining to improve the VaR and ES forecasting performance obtained by individual models. In our cryptocurrency case, where it is not clear which of a diverse set of models is best for forecasting risk measures for daily data, forecast combination is an interesting approach to deal with model uncertainty and exploit the information provided by each individual model.

The contribution of this paper is threefold. First, in the context of ES forecasting, we evaluate the accuracy of GAS, CAViaR and robust bootstrap GARCH-based models, which have been found useful for forecasting the VaR. Second, we revisit some methods advocated for forecasting risk measures for cryptocurrency data, and compare their VaR and ES forecasting performance under periods of turbulence that include the COVID-19 period, the third halving of Bitcoin[3] and the Lexia class action[4]. Third, we evaluate

[3]The halving of Bitcoin is an event where the reward given to Bitcoin miners for processing transactions is halved. The third halving occurred on May 11, 2020

[4]On May 19, 2021, the law firm Lexia in collaboration with the Swiss Blockchain Consortium, filed a class action against Binance.

the use of forecast combining as a way to deal with model misspecification and improve forecast accuracy.

The rest of the paper is organised as follows. In Section 2, we introduce the concepts of VaR and ES, and describe the individual forecasting methods that we apply to our cryptocurrency data. Section 3 describes the benefits of forecast combination as well as forecast combining strategies that we use. In Section 4, the data description and results are reported. Finally, Section 5 concludes.

2 VaR and ES forecasting

For a given cryptocurrency, let P_t be the daily closing price at time t and let $r_t = 100\times\log(P_t/P_{t-1})$ be its corresponding percentage log-return (hereafter called the return). Assuming that returns follow a continuous distribution, the one-step-ahead VaR and ES for a given risk level α are defined as:

$$\begin{aligned}\mathrm{VaR}^{\alpha}_{\mathrm{T}+1} &\coloneqq \mathrm{Sup}\{\mathrm{x}\in\mathbb{R}:\mathrm{F}(\mathrm{x}|\mathcal{F}_{\mathrm{T}})\leq\alpha\}\\ \mathrm{ES}^{\alpha}_{\mathrm{T}+1} &\coloneqq \mathbb{E}[r_{T+1}|r_{T+1}\leq\mathrm{VaR}^{\alpha}_{\mathrm{T}+1},\mathcal{F}_{\mathrm{T}}]\end{aligned}$$

where F is the conditional returns distribution and $\mathcal{F}_T$ stands for the information available up to time T. For a chosen horizon, which in our case is a day-ahead, the probability of observing a return less than or equal to the VaR is α, while the ES is the expected value of exceedances beyond the VaR.

Although there are several methods to forecast the VaR and ES in the literature (see, Righi and Ceretta, 2015; Nieto and Ruiz, 2016; Bayer and Dimitriadis, 2020; Calmon et al., 2020, for interesting reviews), only a few procedures have shown good performance in cryptocurrency markets. Those procedures are briefly described next.

Without loss of generality, we assume zero-mean and serially uncorrelated returns,

$$r_t = \sigma_t\epsilon_t,$$

where $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a zero-mean unit-variance *iid* random variable (hereafter called the innovations) and σ_t stands for volatility. In cases where the returns are not zero-mean, returns are first centred or an ARMA(p,q) filter is previously fitted on the returns de-

pending whether the mean is constant over time or not.

2.1 NAGARCH model

Acereda et al. (2020) use the nonlinear asymmetric GARCH (NAGARCH) model of Engle and Ng (1993) to forecast the ES for Bitcoin and three other cryptocurrencies. The NAGARCH(1,1) volatility equation is given by

$$\sigma_{t+1}^2 = \omega + \delta(r_t - c\sigma_t)^2 + \beta\sigma_t^2 \tag{1}$$

with parameters $\omega,\ \delta, \beta, c$ satisfying stationarity conditions and where the parameter c accounts for the leverage effect.

We estimate the parameters by maximum likelihood under the assumption that the innovations (ϵ_t) follow a skew-t distribution. The NAGARCH model with the same distributional assumption was previously used by Acereda et al. (2020) to forecast the ES of Bitcoin and other cryptocurrencies and the results were encouraging.

We use this model to forecast not only the ES as in Acereda et al. (2020) but also to forecast the VaR. The day-ahead VaR and ES forecasts are given by

$$\begin{aligned} \widehat{\mathrm{VaR}^{\alpha}}_{T+1} &= \hat{F}_{T+1}^{-1}(\alpha) \\ \widehat{\mathrm{ES}^{\alpha}}_{T+1} &= \frac{1}{\alpha}\int_{-\infty}^{\widehat{\mathrm{VaR}^{\alpha}}_{T+1}} x f_{T+1}(x)dx, \end{aligned} \tag{2}$$

where $\hat{F}_{T+1}^{-1}(\alpha)$ is the α-quantile of the estimated conditional return distribution and f_{T+1} is the conditional return density function. Expression (2) holds regardless of the continuous innovation distribution used. For our skew-t distribution case, those values are obtained by numerical approximation and adaptive quadrature numerical integration using the algorithms of Hill (1970) and Piessens et al. (2012), respectively.

2.2 FIGARCH model

The fractionally integrated GARCH (FIGARCH) model with heavy tailed innovation distribution was previously used by Soylu et al. (2020) to forecast the VaR and ES of Bitcoin, Ethereum and Ripple and the results suggest that this model is useful to forecast both risk measures. The FIGARCH model (Baillie et al., 1996) is a long memory volatility

model, which means that volatility exhibits long-range dependence, a feature not captured by classical GARCH models. The FIGARCH(1,d,1) volatility equation is given by

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + [1 - \beta L - \delta L(1-L)^d]r_t^2,$$

where L is the lag operator and the parameters ω, β and δ satisfy stationarity conditions. We estimate the parameters by maximum likelihood assuming a skew-t innovation distribution as in Soylu et al. (2020) and the one-step-ahead VaR and ES forecasts are obtained as in (2).

2.3 Robust bootstrap GARCH-based model

A robust bootstrap GARCH-based model has been proposed by Trucíos et al. (2017) and used to forecast the VaR of Bitcoin by Trucíos (2019). The promising results obtained motivate us in this paper to extend it to the ES context. The procedure mitigates the effects of additive outliers in the estimation of volatilities and densities of returns. A key aspect of the approach relies on the following volatility equation of Boudt et al. (2013)

$$\sigma_t^2 = \omega + \delta\gamma_c r_c\left(\frac{r_{t-1}^2}{\sigma_{t-1}^2}\right)\sigma_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{3}$$

where ω, δ and β are parameters satisfying stationarity conditions, γ_c is a constant to guarantee Fisher consistency and $r_c(\cdot)$ is a robust filter (with a tuning parameter c) given by[5]

$$r_c(x) = \begin{cases} 1, & \text{if } x > c, \\ x, & \text{if } x \leq c. \end{cases}$$

The robust bootstrap procedure can be summarised in the following four steps:

- *Step 1:* Estimate the GARCH model in a robust way using Equation (3) and the M-estimator of Boudt et al. (2013). Then, obtain the standardised residuals $\hat{\epsilon}_t = r_t/\hat{\sigma}_t$ and denote by $\hat{F}_\epsilon$ their empirical distribution.

[5]The robust filter used by Boudt et al. (2013) is slightly different, but results in Trucíos et al. (2015, 2017) reveal that better forecasting performance is achieved using the filter defined here. As in Trucíos et al. (2017) and Trucíos (2019), we use $c = 9$.

- *Step 2:* Generate bootstrap series through the following recursion:

$$r_t^* = \sigma_t^* \epsilon_t^* \text{ and } \sigma_{t+1}^{*2} = \hat{\omega} + \hat{\delta}\gamma_c r_c \left(\frac{r_t^{*2}}{\sigma_t^{*2}} \right) \sigma_t^{*2} + \hat{\beta}\sigma_t^{*2},$$

  where $\sigma_1^{*2} = \hat{\sigma}_1^2$, $r_c(\cdot)$ is a filter similar to the one used in Equation (3) but replacing large values by new squared bootstrap extractions from $\hat{F}_\epsilon$ and the estimated parameters $(\hat{\omega}, \hat{\delta}, \hat{\beta})$ are those obtained in Step 1.

- *Step 3*: Obtain one-step-ahead forecasts as

$$\hat{r}_{T+1}^* = \hat{\sigma}_{T+1}^* \epsilon_{T+1}^* \quad \text{and} \quad \hat{\sigma}_{T+1}^{*2} = \hat{\omega}^* + \hat{\delta}^* \gamma_c r_c \left(\frac{r_T^{*2}}{\hat{\sigma}_T^{*2}} \right) \hat{\sigma}_T^{*2} + \hat{\beta}^* \hat{\sigma}_T^{*2},$$

  where $\hat{r}_T^* = r_T$, ϵ_{T+1}^* are bootstrap extractions from $\hat{F}_\epsilon$ and for $t = 2, ..., T$

$$\hat{\sigma}_t^{*2} = \hat{\omega}^* + \hat{\delta}^* \gamma_c r_c \left(\frac{r_{t-1}^2}{\hat{\sigma}_{t-1}^{*2}} \right) \hat{\sigma}_{t-1}^{*2} + \hat{\beta}^* \hat{\sigma}_{t-1}^{*2},$$

  with $\hat{\sigma}_1^{*2} = \hat{\sigma}_1^2$ and $(\hat{\omega}^*, \hat{\delta}^*, \hat{\beta}^*)$ being the estimates obtained from the bootstrap series using the same estimation procedure in Step 1.

- *Step 4:* Repeat steps 2 and 3 B times to obtain B bootstrap replicates, then estimate the $\text{VaR}_{\text{T+1}}^{\alpha}$ and $\text{ES}_{\text{T+1}}^{\alpha}$ as the α-quantile of the bootstrap replicates and the average returns smaller than $\text{VaR}_{\text{T+1}}^{\alpha}$, respectively.

2.4 GAS models

Generalised autoregressive score (GAS) models have previously been used to forecast the VaR in cryptocurrency data by Troster et al. (2019), and Liu et al. (2020), among others with interesting and promising results. In this paper, we extend their use to ES forecasting. GAS models (Harvey, 2013; Creal et al., 2013) use the score of the conditional distribution instead of the squared returns in the volatility equation (as commonly used in GARCH-type models). Its volatility equation is given by

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \delta s_t \left[\frac{\partial \log f(r_t|\sigma_t^2)}{\partial \sigma_t^2} \right],$$

where s_t is a scaling function for the score and $f(\cdot)$ is the density function of the assumed distribution. Typically, $s_t = \mathcal{I}_t^{-1}$, with $\mathcal{I}_t$ being the Fisher information. This choice is

very natural in a volatility modelling context and encompasses, for instance, the popular GARCH model. The GAS structure allows information in the whole distribution to be taken into account, rather than only the second-order moments.

The parameters are estimated by maximum likelihood assuming a skew-t innovation distribution. This choice is particularly useful since it provides robustness against extreme observations. The one-step-ahead VaR and ES forecasts are obtained as in (2).

2.5 MSGARCH

The Markov-Switching GARCH models (MSGARCH) of Haas et al. (2004) have previously been used to forecast risk measures in cryptocurrency data by Ardia et al. (2019b), Caporale and Zekokh (2019), Alexander and Dakos (2020) and Maciel (2020). They deal with structural breaks in the volatility process, which cause quick changes in the volatility. Such breaks can lead to high persistence in volatility observed empirically with single-regime models.

The procedure proposed by Haas et al. (2004) allows the volatility equation in each regime to evolve independently. Assuming that all regimes follow a GARCH(1,1) process, the variance equation on regime $\pi_t = k$ is given by

$$\sigma^2_{k,t} = \omega_k + \delta_k r^2_{t-1} + \beta_k \sigma^2_{k,t-1},$$

with parameters $\omega_k, \delta_k, \beta_k$ satisfying stationarity conditions. The hidden sequence $\pi_t = 1, \ldots, k$ evolves according to a first-order ergodic homogeneous Markov chain with transition probability matrix $P = \{p_{ij}\}_{j=1}^{k}$, with elements $p_{ij} = P(\pi_t = j | \pi_{t-1} = i)$.

As in Ardia et al. (2019b), Caporale and Zekokh (2019) and Maciel (2020), the estimation is made in a Bayesian framework via MCMC using the adaptive procedure of Hoogerheide and van Dijk (2010). The MSGARCH model, under different GARCH-type specifications, has previously been used to forecast the VaR and/or ES in cryptocurrency data by Ardia et al. (2019b) and Caporale and Zekokh (2019), Alexander and Dakos (2020) and Maciel (2020). Here, we follow Alexander and Dakos (2020) and consider two regimes with each one following a GARCH(1,1) process under a skew-t innovation distribution.

2.6 CAViaR-RegressionForES

The conditional autoregressive VaR model (CAViaR) proposed by Engle and Manganelli (2004) have recently been used to forecast the VaR of Bitcoin by Li et al. (2021). The promising results obtained motivate us to use this methodology not only to forecast the VaR but also the ES. CAViaR models use a quantile regression framework to forecast the VaR with no need to estimate volatility as in Sections 2.1 – 2.5. Using the so-called symmetric absolute value CAViaR specification[6], the VaR for a given risk level α is modelled by

$$\text{VaR}_t^\alpha = \beta_0 + \beta_1 |r_{t-1}| + \beta_2 \text{VaR}_{t-1}^\alpha, \tag{4}$$

where the parameters β_0, β_1 and β_2 are estimated by minimising

$$\frac{1}{T}\sum_{t=1}^{T} \left[\alpha - I(r_t \leq \text{VaR}_t^\alpha)\right]\left[r_t - \text{VaR}_t^\alpha\right]. \tag{5}$$

CAViaR has only previously been used to forecast the VaR of Bitcoin, and not other cryptocurrencies, and it has not been used as the basis for predicting the ES of any cryptocurrencies. Therefore, we extend the study of Li et al. (2021) and forecast both VaR and ES of Bitcoin and Ethereum. Notwithstanding CAViaR does not provide a direct way to estimate the ES, we follow Manganelli and Engle (2004) and estimate the ES through

$$\widehat{\text{ES}^\alpha}_t = \hat{\phi}\widehat{\text{VaR}^\alpha}_t, \quad \text{for } t = 1, \ldots, T+1,$$

where $\hat{\phi}$ is the estimated parameter obtained by regressing the returns exceedance beyond the estimated VaR against its corresponding estimated VaR. For simplicity, this method is referred as CAViaR in the empirical application.

2.7 CAViaR-EVT

Extending the results in Li et al. (2021), we evaluate whether incorporating extreme value theory (EVT) into the CAViaR framework, as proposed by Manganelli and Engle (2004), is useful for forecasting the VaR and ES for Bitcoin and Ethereum

[6]There are other specifications that could be used in a CAViaR context, but we prefer to use the simplest one since Li et al. (2021) does not report evidence that complex specifications outperform the simplest one for the Bitcoin case.

The procedure proposed by Manganelli and Engle (2004) is based on the EVT method of McNeil and Frey (2000) and can be summarised in the following three steps:

- *Step 1:* Use the CAViaR model to estimate the VaR for a risk level θ not as extreme as the desired risk level α. Then, calculate the standardised quantile residuals through $r_t/\widehat{\text{VaR}^{\theta}}_t - 1$.

- *Step 2:* Fit the generalised Pareto distribution to the positive standardised quantile residuals and obtain $\hat{\vartheta}$ and $\hat{\varepsilon}$, the respective estimated scale and shape parameters.

- *Step 3:* For a given risk level α and for $t = 1, \ldots, T+1$, obtain the VaR and ES estimates by

$$\widehat{\text{VaR}^{\alpha}}_t = \widehat{\text{VaR}^{\theta}}_t \Big(1 + \frac{\hat{\vartheta}}{\hat{\varepsilon}}\Big[\Big(\frac{\theta}{\alpha}\Big)^{\hat{\varepsilon}} - 1\Big]\Big) \quad \text{and}$$

$$\widehat{\text{ES}^{\alpha}}_t = \widehat{\text{VaR}^{\theta}}_t \Big(1 + \frac{\frac{\hat{\vartheta}}{\hat{\varepsilon}}\Big[\Big(\frac{\theta}{\alpha}\Big)^{\hat{\varepsilon}} - 1\Big] + \hat{\vartheta}}{1 - \hat{\varepsilon}}\Big).$$

Following Manganelli and Engle (2004) and Taylor (2019), we use θ equal to 7.5%.

2.8 CAViaR-ALD

Extending the study of Li et al. (2021) and motivated by the recently proposed quantile regression approach of Taylor (2019) that estimate both VaR and ES jointly, we evaluate whether this method is useful to forecast both risk measures for Bitcoin and Ethereum.

Based on the results of Koenker and Machado (1999) and using the fact that the asymmetric Laplace density (ALD) can be written as

$$f(r_t) = \frac{\alpha - 1}{\text{ES}_\text{t}^{\alpha}} \exp\Big(\frac{\big(r_t - \text{VaR}_t^{\alpha}\big)\big(\alpha - I(r_t \leq \text{VaR}_t^{\alpha})\big)}{\alpha \text{ES}_t^{\alpha}}\Big), \tag{6}$$

Taylor (2019) proposes to maximise the likelihood of the aforementioned distribution considering that the VaR component follows a CAViaR model (hence the name CAViaR-ALD) and the ES, in its simplest form, is given by

$$\text{ES}_\text{t}^{\alpha} = (1 + \exp(\gamma_0))\text{VaR}_\text{t}^{\alpha}. \tag{7}$$

The parametrisation used in the density function (6) as well as the relationship between

ES and VaR given in Equation (7) allow us to estimate the VaR and ES jointly avoiding the ES crossing the VaR. Additionally, it is worth mentioning that this method does not assume that returns follow an asymmetric Laplace distribution, since α is not estimated but a chosen fixed value equal to the risk level desired.

3 Forecast Combinations

Forecast combining provides a pragmatic way to deal with model misspecification and to synthesise the information extracted from the data by individual models. Its use in a risk measurement context has been less exploited than in other fields, but there are several studies providing empirical evidence in support of combining. See; for instance, Giacomini and Komunjer (2005), Halbleib and Pohlmeier (2012), Bayer (2018), Taylor (2020) and Happersberger et al. (2020). Those approaches can be divided into basic and scoring function minimisation strategies, which are briefly described next.

3.1 Basic combining strategies

The first group consists of basic strategies, which are fast and easy to implement. They rely on the computation of the average, median, maximum and minimum values. The average and median are standard approaches in the forecast combining literature, while the maximum and minimum are potentially of interest in the risk context, as they provide more and less conservative choices from a set of forecasts.

Let M be the number of individual forecasting methods and let $\widehat{\mathrm{VaR}^{\mathrm{i}}}_{\mathrm{T+1}}$ and $\widehat{\mathrm{ES}^{\mathrm{i}}}_{\mathrm{T+1}}$ $(i = 1, ..., \mathrm{M})$ be the corresponding day-ahead VaR and ES forecasts obtained by the i-th method. The basic combining strategies are defined as follows

(i) Simple average (AVG)

$$\widehat{\mathrm{VaR}^{\mathrm{c}}}_{\mathrm{T+1}} = \sum_{\mathrm{i=1}}^{\mathrm{M}} \frac{\widehat{\mathrm{VaR}^{\mathrm{i}}}_{\mathrm{T+1}}}{\mathrm{M}} \quad \text{and} \quad \widehat{\mathrm{ES}^{\mathrm{c}}}_{\mathrm{T+1}} = \sum_{\mathrm{i=1}}^{\mathrm{M}} \frac{\widehat{\mathrm{ES}^{\mathrm{i}}}_{\mathrm{T+1}}}{\mathrm{M}}.$$

(ii) Median value (MED)

$$\widehat{\mathrm{VaR}^{\mathrm{c}}}_{\mathrm{T+1}} = \mathrm{Med}\{\widehat{\mathrm{VaR}^{1}}_{\mathrm{T+1}}, \ldots, \widehat{\mathrm{VaR}^{\mathrm{M}}}_{\mathrm{T+1}}\} \text{ and}$$
$$\widehat{\mathrm{ES}^{\mathrm{c}}}_{\mathrm{T+1}} = \mathrm{Med}\{\widehat{\mathrm{ES}^{1}}_{\mathrm{T+1}}, \ldots, \widehat{\mathrm{ES}^{\mathrm{M}}}_{\mathrm{T+1}}\}.$$

(iii) Maximum value (MAX)

$$\widehat{\mathrm{VaR}^{\mathrm{c}}}_{\mathrm{T+1}} = \mathrm{Max}\{\widehat{\mathrm{VaR}^{1}}_{\mathrm{T+1}}, \ldots, \widehat{\mathrm{VaR}^{\mathrm{M}}}_{\mathrm{T+1}}\} \text{ and}$$
$$\widehat{\mathrm{ES}^{\mathrm{c}}}_{\mathrm{T+1}} = \mathrm{Max}\{\widehat{\mathrm{ES}^{1}}_{\mathrm{T+1}}, \ldots, \widehat{\mathrm{ES}^{\mathrm{M}}}_{\mathrm{T+1}}\}.$$

(iv) Minimum value (MIN)

$$\widehat{\mathrm{VaR}^{\mathrm{c}}}_{\mathrm{T+1}} = \mathrm{Min}\{\widehat{\mathrm{VaR}^{1}}_{\mathrm{T+1}}, \ldots, \widehat{\mathrm{VaR}^{\mathrm{M}}}_{\mathrm{T+1}}\} \text{ and}$$
$$\widehat{\mathrm{ES}^{\mathrm{c}}}_{\mathrm{T+1}} = \mathrm{Min}\{\widehat{\mathrm{ES}^{1}}_{\mathrm{T+1}}, \ldots, \widehat{\mathrm{ES}^{\mathrm{M}}}_{\mathrm{T+1}}\}.$$

While the AVG strategy has been used by Taylor (2020) for both VaR and ES forecasting, the other three strategies have been used only for VaR forecasting (McAleer et al., 2013a,b; Bayer, 2018; Buczyński and Chlebus, 2019). Their applicability for ES forecast combining is assessed here.

3.2 Combining strategies based on scoring function minimisation

Based on the results of Fissler and Ziegel (2016), Taylor (2020) proposes two forecast combination strategies to deal with VaR and ES, namely, minimum score combining (MSC) and relative score combining (RSC). Those strategies rely on the combining weights being estimated by the minimisation of a scoring function of the form

$$\begin{aligned} \mathrm{S}(\mathrm{VaR_t}, \mathrm{ES_t}, \mathrm{r_r}) = {} & (\mathrm{I}(\mathrm{r_t} \leq \mathrm{VaR_t}) - \alpha)\mathrm{G_1}(\mathrm{VaR_t}) - \mathrm{I}(\mathrm{r_t} \leq \mathrm{VaR_t})\mathrm{G_1}(\mathrm{r_t}) \\ & + \mathrm{G_2}(\mathrm{ES_t})(\mathrm{ES_t} - \mathrm{VaR_t} + \mathrm{I}(\mathrm{r_t} \leq \mathrm{VaR_t})(\mathrm{VaR_t} - \mathrm{r_t})/\alpha) \\ & + \zeta_2(\mathrm{ES_t}) + \mathrm{a}(\mathrm{r_t}), \end{aligned} \tag{8}$$

where α is the risk level; and the functions $\mathrm{G_1}$, $\mathrm{G_2}$, ζ_2 and a satisfy certain conditions such as $\mathrm{G_1}$ and ζ_2 are increasing, ζ_2 is convex, and $\mathrm{G_2} = \zeta_2'$. See; Fissler and Ziegel (2016) and Dimitriadis and Bayer (2019) for further details.

In the MSC strategy, the combined estimators are given by

$$\begin{aligned}\widehat{\mathrm{VaR^c}}_{\mathrm{T+1}} &= \sum_{i=1}^{M} \tau_i \widehat{\mathrm{VaR^i}}_{T+1} \text{ and} \\ \widehat{\mathrm{ES^c}}_{\mathrm{T+1}} &= \widehat{\mathrm{VaR^c}}_{\mathrm{T+1}} + \sum_{\mathrm{i=1}}^{\mathrm{M}} \kappa_{\mathrm{i}} (\widehat{\mathrm{ES^i}}_{\mathrm{T+1}} - \widehat{\mathrm{VaR^i}}_{\mathrm{T+1}}),\end{aligned} \tag{9}$$

where the two sets of combining weights τ_i and κ_i $(i = 1, ..., M)$ are obtained in a single step by minimising the sum of the in-sample values of the chosen scoring function belonging to the class presented in expression (8) subject to the constraints τ_i, $\kappa_i \geq 0$, $\sum_{i=1}^{M} \tau_i = 1$ and $\sum_{i=1}^{M} \kappa_i = 1$. The structure of the expression for the ES ensures that the ES combined forecast exceeds the VaR combined forecast, which would not be guaranteed if the ES combined forecast was simply a linear combination of the individual ES forecasts.

The other strategy proposed by Taylor (2020) is the RSC strategy, which is computationally lighter than the MSC and leads to a single set of weights for both VaR and ES. In this case, the combined estimators are obtained by

$$\begin{aligned}\widehat{\mathrm{VaR^c}}_{\mathrm{T+1}} &= \sum_{i=1}^{M} \eta_i \widehat{\mathrm{VaR^i}}_{T+1} \quad \text{and} \\ \widehat{\mathrm{ES^c}}_{\mathrm{T+1}} &= \sum_{i=1}^{M} \eta_i \widehat{\mathrm{ES^i}}_{T+1},\end{aligned} \tag{10}$$

where

$$\eta_i = \frac{\exp\left(-\lambda \sum_{j=1}^{T} \mathrm{S}\left(\widehat{\mathrm{VaR^i_j}}, \widehat{\mathrm{ES^i_j}}, \mathrm{r_j}\right)\right)}{\sum_{i=1}^{M} \exp\left(-\lambda \sum_{j=1}^{T} \mathrm{S}\left(\widehat{\mathrm{VaR^i_j}}, \widehat{\mathrm{ES^i_j}}, \mathrm{r_j}\right)\right)},$$

with the tuning parameter $\lambda > 0$ being the value that minimises the sum of the in-sample values of the scoring function.

Note that a value of λ close to zero leads to a simple average where all forecasting methods have the same weight, while a large value of λ leads to the selection of a single method with the best historical performance. As with the MSC method, the structure of the RCS combining expressions ensures that the ES combined forecast exceeds the VaR combined forecast.

In our empirical study, we consider three different versions of both the MSC and

RSC combining strategies. Those versions correspond to the parameters estimated using three different forms of the scoring function of expression (8). We considered the scoring functions proposed by Fissler and Ziegel (2016), Nolde et al. (2017) and Taylor (2019), which we denote here as FZG, NZ and AL, respectively. The AL scoring function is equal to the negative of the log of the ALD likelihood of expression (8). We refer to the different versions of the combining strategies as MSC_{FZG}, MSC_{NZ}, MSC_{AL}, RSC_{FZG}, RSC_{NZ} and RSC_{AL}.

4 Empirical Study

4.1 Data

We analyse daily closing prices[7] of Bitcoin and Ethereum, the two major cryptocurrencies, which represent more than 50% of the cryptocurrency market capitalisation. Prices spanning from August 17, 2017, to July 22, 2022, were obtained from Binance, one of the largest crypto-exchanges around the world. Binance started its operations in August 2017 with Bitcoin and Ethereum being the firsts cryptocurrencies to be traded. Although there are several other cryptocurrencies being traded on Binance nowadays, they were not considered in our analysis because there is so little historical data available.

All analyses were performed using R software (R Core Team, 2021) and the volatility models were implemented using the R packages rugarch (Ghalanos, 2020), GAS (Ardia et al., 2019c), MSGARCH (Ardia et al., 2019a), RobGARCHBoot (Trucios, 2020) and self implemented routines. For reproducibility purposes, the codes as well as the data used in the paper are available on the Github repository `github.com/ctruciosm/CryptoForeComb`.

We used a rolling window scheme with window size of 1000 days, which led to an out-of-sample period of 800 days. For each window, we estimated parameters for the individual models and the combining methods, and then produced day-ahead VaR and ES forecasts. We then moved the window forward one day, and repeated the estimation and forecasting procedure. The same procedure is done until no more data is available.

Table 1 reports the descriptive statistics of Bitcoin and Ethereum daily returns and Figure 1 displays the daily returns (left panel), the sample auto-correlation function of

[7]Cryptocurrencies are traded 24/7 and the last price traded on the day is considered as closing price.

returns (middle panel) and the sample auto-correlation function of squared returns (right panel). The autocorrelations are reported with their respective 95% confidence bands, computed using the generalised non-parametric Bartlett's formula (Francq and Zakoïan, 2009) and the Bartlett's formula (Bartlett, 1946) for the returns and squared returns, respectively. Classical $\pm z_{\alpha/2}/\sqrt{n}$ bands are also reported, in dotted lines, for return's autocorrelations. The vertical dashed lines in the left panel of Figure 1 indicate the beginning of the out-of-sample period (May 14, 2020).

Both series report extreme observations with maximum returns larger than 20% and minimum returns less than -50%. The minimum values all occur on March 12, 2020, one day after the World Health Organisation announced that COVID-19 can be characterised as a pandemic.

Bitcoin reports an annualised volatility of 80.8% ($4.23\% \times \sqrt{365}$) while Ethereum an annualised volatility of 102.4% ($5.36\% \times \sqrt{365}$). Both cryptocurrencies report high kurtosis, which can be partially explained by the presence of extreme observations. Additionally, both series report positive mean and median values as well as negative skewness.

	Min	Q_1	Median	Mean	Q_3	Max	S.D	Skew	Kurtosis
Bitcoin	-50.26	-1.71	0.15	0.09	1.94	20.30	4.23	-1.03	16.33
Ethereum	-59.05	-2.25	0.13	0.09	2.84	23.38	5.36	-1.03	13.65

Table 1: Descriptive statistics of daily returns. Q_1 and Q_3 stand for the first and third quartiles while S.D stands for the standard deviation.

From Figure 1, we can observe that both series of returns exhibit a non negligible serial correlation. So, in order to forecast the VaR and ES we first apply an appropriate AR(p) filter and then the procedures described in Section 2 are applied on the residuals. For each window, the autoregressive order p was selected as $p = 0$, 1 or 2, using the Akaike information criterion. Additionally to the models described in Section 2, we included the widely known GARCH (Bollerslev, 1986) and GJR (Glosten et al., 1993) models with a skew-t innovation distribution as benchmarks. We then applied the combining methods from Section 3.

As pointed out by Atiya (2020) and Taylor (2019), forecast combination is a powerful tool when different information is used by the models or when the same information is used by the models in different ways. In those cases, we say that the models are diverse, and a good example of this is when we use models based on different assumptions. In

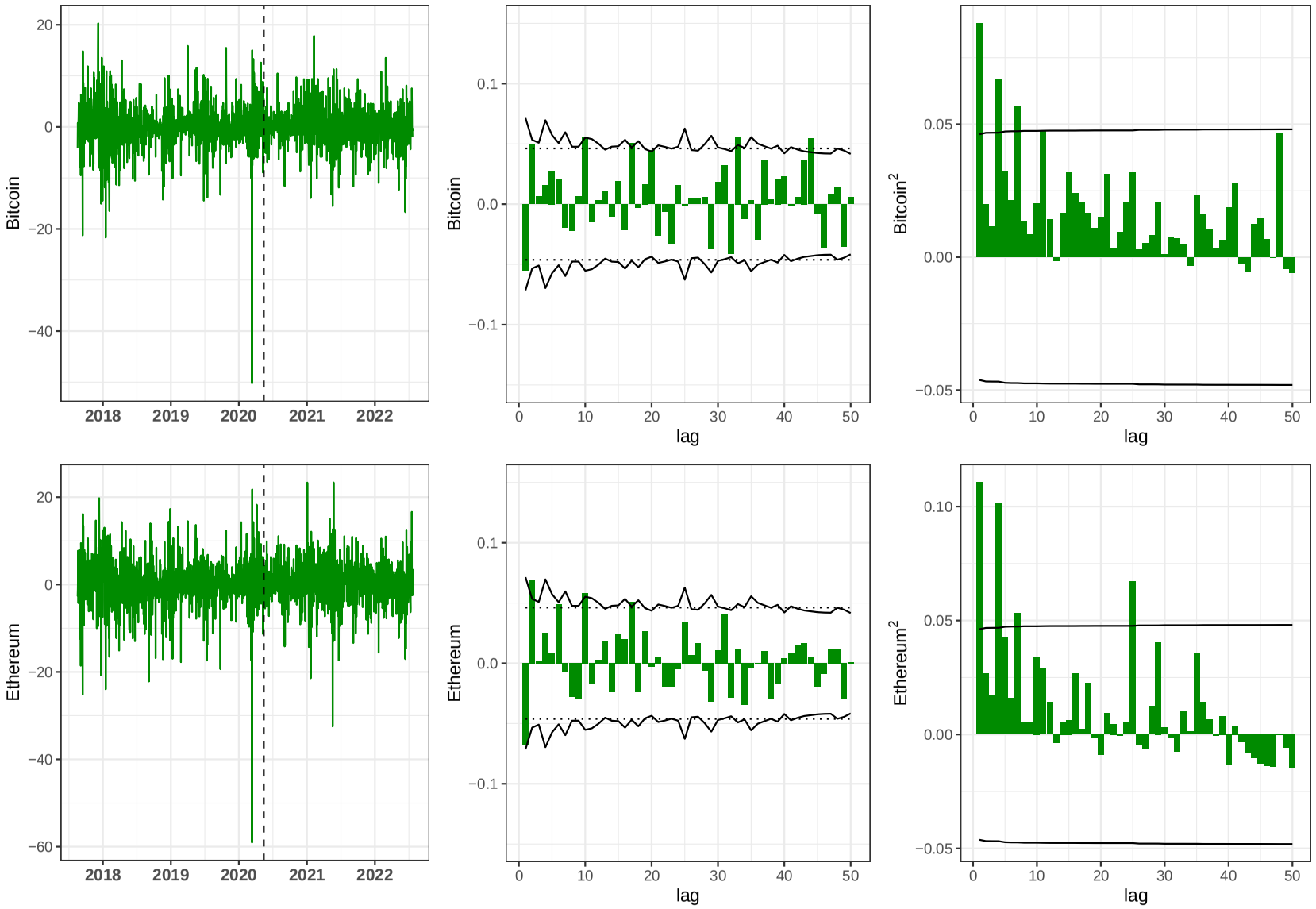

Figure 1: Daily returns (left panel), sample auto-correlation function of returns (middle panel) and sample auto-correlation function of squared returns (right panel) for Bitcoin and Ethereum. Full sample period. Vertical dashed lines (May 14, 2020) split the data into in-sample and out-of-sample periods.

our case, the models implemented are built using the same information but are based on different assumptions, and are therefore a reasonably diverse set of models to combine.

4.2 Out-of-sample results

In our empirical study, we evaluate the one-step-ahead forecast of the VaR and ES for the 2.5% and 5% risk levels. Although the 1% risk level is quite often considered, we do not include it here because the small size of the out-of-sample period (800 trading days) renders the statistical tests inconclusive.

To evaluate the VaR and ES forecasting performance in the out-of-sample period, we use a comprehensive back-testing exercise based on both calibration tests (to evaluate whether the forecasts are valid or not) and scoring functions (to evaluate the precision of the forecasts). In addition to the widely-used quantile loss (QL) function (González-Rivera et al., 2004), we also use the scoring functions proposed by Fissler et al. (2016), Nolde et al. (2017) and Taylor (2019), which are all particular cases of Equation (8). These scoring functions are denoted here as FZG, NZ and AL, respectively. The QL scoring function considers only the VaR, while the other scoring functions consider VaR and ES jointly. The implemented calibration tests are commonly used in the financial econometrics literature and are listed in Table 2. Roughly speaking, the null hypothesis in all calibration tests can be interpreted as 'the risk measure (namely, VaR and/or ES) is correctly specified', where different definitions for the term *correct* are used in each hypothesis test. For a brief explanation of those tests we refer to Nieto and Ruiz (2016), Righi and Ceretta (2015) and Hallin and Trucíos (2021).

Tables 3 and 4 report the percentage of hits (returns smaller than VaR), the p-values of the calibrations tests and the average scoring function of the individual and combining methods. Shaded rows indicate procedures delivering satisfactory calibration test results (i.e., fail to reject the null hypothesis at 5% significance level in all calibration tests). To compare the precision of the forecasts, we use the model confidence set approach of Hansen et al. (2011) at 5% significance level on the four scoring functions previously mentioned (QL, FZG, NZ and AL). For Bitcoin and Ethereum, the model confidence set at 5% significance does not distinguish between the global performance obtained by the various procedures implemented. Almost all of them belong to the model confidence set, and so we do not report the results in the tables.

As pointed out by Giacomini and Rossi (2010), in periods of instability, the global forecasting performance can hide important information about the performance of the competing models over time. Bearing in mind that our out-of-sample period is characterised by instabilities in the cryptocurrency market such as the COVID-19 pandemic period, the third halving of Bitcoin and the Lexia class action, we go one step further than previously done in the cryptocurrency literature and perform the fluctuation test of Giacomini and Rossi (2010) to compare, between those methods with satisfactory calibration test results, the relative superiority over time of their scoring function values.

It is worth mentioning that procedures that do not deliver satisfactory calibration test results are not appropriate for forecasting the risk measures, even if they produce the smallest scoring functions. On the other hand, for procedures with satisfactory calibration test results, the one with the smallest scoring functions is preferred. Therefore, better methods are those for which the percentage of hits is close to the nominal risk level; the null hypotheses in the calibration tests are not rejected; and, between those methods, the scoring functions have lower values.

Test	Proposed by	Used to evaluate
Conditional coverage (CC)	Christoffersen (1998)	VaR
Dynamic quantile (DQ)	Engle and Manganelli (2004)	VaR
VaR quantile regression (VQ)	Gaglianone et al. (2011)	VaR
Exceedance residuals (ER)	McNeil and Frey (2000)	ES and VaR
Conditional calibration (CoC)	Nolde et al. (2017)	ES and VaR
Exceedance shortfall regression (ESR)	Bayer and Dimitriadis (2020)	ES

Table 2: Calibration tests used to evaluate VaR and ES accuracy.

Out-of-sample results for Bitcoin

Results for Bitcoin are reported in Table 3. For the 2.5% risk level, four individual models (GARCH, GAS, FIGARCH and CAViaRALD) and three combining strategies (MAX, MSC_{AL}, RSC_{AL}) deliver satisfactory calibration test results. For the 5% risk level, three individual models (GARCH, MSGARCH, FIGARCH) and three combining strategies (MED, RSC_{FZG}, RSC_{AL}) deliver satisfactory calibration test results.

Although some combining strategies deliver satisfactory calibration tests results for each risk level, the model confidence set at 5% significance level does not distinguish between the global performance obtained between the individual models and combining

Table 3: One-step-ahead VaR and ES backtesting for Bitcoin at 2.5% and 5% risk levels. Shaded rows indicate procedures with p-values larger than 0.05 in all calibration tests. Smallest average scoring functions values are underlined.

				Calibration test p-values						Average scoring functions			
			Hits	CC	DQ	VQ	ER	CoC	ESR	QL	FZG	NZ	AL
2.5%	Indiv.	GARCH	2.4	0.613	0.839	0.175	0.856	0.121	0.920	0.273	0.964	3.297	3.413
		GJR	2.5	0.598	0.871	0.000	0.855	0.163	0.882	0.274	0.965	3.308	3.422
		GAS	3.1	0.246	0.495	0.312	0.636	0.475	0.476	0.275	0.965	3.297	3.407
		MSGARCH	2.0	0.464	0.666	0.032	0.714	0.454	0.906	0.279	0.969	3.328	3.432
		Boot.	2.8	0.486	0.597	0.327	0.994	0.000	0.962	0.279	0.970	3.343	3.443
		FIGARCH	2.4	0.613	0.828	0.222	0.865	0.107	0.916	0.274	0.964	3.299	3.415
		NGARCH	2.1	0.542	0.706	0.014	0.666	0.253	0.895	0.276	0.966	3.321	3.431
		CAViaR	2.9	0.406	0.262	0.038	0.856	0.243	0.734	0.279	0.969	3.332	3.436
		CAViaREVT	2.8	0.486	0.768	0.000	0.984	0.001	0.929	0.277	0.968	3.329	3.436
		CAViaRALD	2.5	0.598	0.127	0.097	0.842	0.151	0.929	0.281	0.971	3.357	3.459
		AVG	2.2	0.594	0.836	0.013	0.822	0.190	0.835	0.275	0.965	3.305	3.418
		MED	2.5	0.598	0.885	0.002	0.884	0.093	0.830	0.273	0.963	3.294	3.411
		MAX	3.5	0.084	0.076	0.117	0.704	0.254	0.388	0.280	0.970	3.334	3.434
	Comb.	MIN	1.6	0.193	0.801	0.000	0.957	0.000	0.991	0.281	0.972	3.369	3.466
		MSC_{FZG}	2.8	0.486	0.530	0.001	0.942	0.035	0.903	0.278	0.969	3.331	3.437
		MSC_{NZ}	3.0	0.323	0.235	0.053	0.933	0.035	0.882	0.278	0.969	3.338	3.447
		MSC_{AL}	2.6	0.553	0.322	0.090	0.844	0.118	0.891	0.279	0.969	3.341	3.448
		RSC_{FZG}	2.6	0.553	0.349	0.034	0.778	0.222	0.853	0.282	0.972	3.355	3.452
		RSC_{NZ}	2.6	0.553	0.086	0.045	0.718	0.406	0.846	0.284	0.974	3.372	3.468
		RSC_{AL}	2.6	0.553	0.082	0.065	0.787	0.245	0.888	0.283	0.974	3.370	3.467
5%	Indiv.	GARCH	5.4	0.869	0.574	0.186	0.853	0.202	0.838	0.440	1.127	2.952	3.217
		GJR	6.4	0.210	0.216	0.039	0.958	0.021	0.820	0.443	1.130	2.964	3.227
		GAS	6.0	0.362	0.123	0.041	0.540	0.453	0.414	0.447	1.134	2.971	3.227
		MSGARCH	5.6	0.681	0.215	0.069	0.945	0.065	0.897	0.444	1.131	2.964	3.226
		Boot.	5.4	0.501	0.077	0.025	0.974	0.002	0.990	0.448	1.135	2.985	3.240
		FIGARCH	5.5	0.759	0.512	0.104	0.877	0.142	0.862	0.441	1.128	2.957	3.221
		NGARCH	6.1	0.194	0.070	0.146	0.949	0.022	0.876	0.443	1.130	2.968	3.230
		CAViaR	5.8	0.452	0.199	0.008	0.742	0.389	0.630	0.446	1.133	2.969	3.227
		CAViaREVT	6.1	0.311	0.154	0.067	0.978	0.005	0.935	0.447	1.134	2.980	3.239
		CAViaRALD	5.8	0.452	0.223	0.068	0.949	0.033	0.921	0.447	1.135	2.979	3.238
		AVG	5.8	0.452	0.227	0.047	0.926	0.069	0.858	0.443	1.130	2.960	3.222
		MED	5.8	0.452	0.255	0.119	0.889	0.114	0.813	0.442	1.129	2.957	3.219
		MAX	7.1	0.030	0.019	0.039	0.787	0.047	0.251	0.450	1.137	2.985	3.240
	Comb.	MIN	4.4	0.661	0.265	0.000	0.970	0.000	1.000	0.446	1.133	2.985	3.246
		MSC_{FZG}	6.0	0.362	0.109	0.016	0.937	0.027	0.875	0.448	1.135	2.986	3.246
		MSC_{NZ}	6.1	0.311	0.130	0.033	0.963	0.023	0.867	0.448	1.136	2.983	3.241
		MSC_{AL}	6.0	0.362	0.180	0.048	0.962	0.027	0.882	0.448	1.135	2.981	3.239
		RSC_{FZG}	6.0	0.362	0.206	0.057	0.899	0.077	0.804	0.448	1.136	2.982	3.241
		RSC_{NZ}	6.0	0.362	0.192	0.040	0.910	0.073	0.811	0.448	1.136	2.983	3.242
		RSC_{AL}	5.9	0.411	0.184	0.061	0.917	0.065	0.857	0.448	1.136	2.983	3.242

strategies implemented. Actually, the model confidence set does not distinguish between the global performance obtained between all procedures implemented, pointing them out as statistically equal (as mentioned in the previous Section, almost all procedures are in the model confidence set).

The fluctuation test of Giacomini and Rossi (2010) was applied to investigate whether there was statistical difference over time in the scoring functions of the methods that had the best calibration test results. For the 2.5% risk level, Figure 2 compares the GAS procedure and the MSC_{AL} combining strategy. The figure shows the fluctuation statistic (solid green line) with their corresponding two-sided 5% critical values (dashed

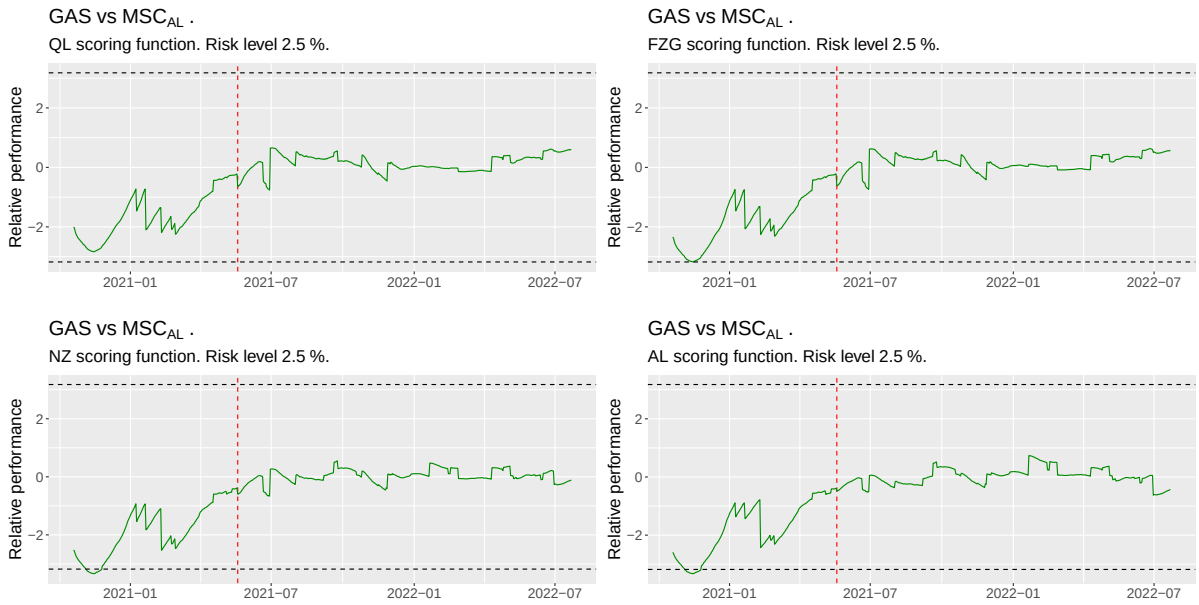

Figure 2: For Bitcoin, fluctuation test comparing the 2.5% VaR and ES forecasting performance of the GAS model and MSC_{AL} combining strategy according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

horizontal black lines). A fluctuation statistic above/below the upper/lower critical value line indicates that the first model (GAS) is outperformed by/outperforms the second one (i.e, the MSC_{AL} combining strategy) for that period of time. The results in Figure 2 indicate that, in two out of the four scoring functions, the GAS procedure outperforms the MSC_{AL} combining strategy for a short period of time (the solid line crosses the inferior dashed line). For the other pairwise comparisons, the fluctuation test indicates no superiority of one model against another over time[8] (those results are not showed here to save space but are available in Figures S1–S20 of the supplementary material). For the 5% risk level, the fluctuation test also reveals no statistically significant difference over time in the scoring function results for all pairwise comparisons (see Figures S21–S35 of the supplementary material).

In general, for any risk level, there is no single approach outperforming all other procedures neither globally nor over time. Nevertheless, some individual and combining strategies have a good performance for forecasting both risk measures, with three of them suitable for both risk levels, namely, GARCH, FIGARCH and RSC_{AL} procedures.

[8]For the GAS and CAViaRALD models, the fluctuation statistic is on the borderline (in favour of the GAS model) for one out of the four scoring functions.

Table 4: One-step-ahead VaR and ES backtesting for Ethereum at 2.5% and 5% risk levels. Shaded rows indicate procedures with p-values larger than 0.05 in all calibration tests. Smallest average scoring functions values are underlined.

				Calibration test p-values						Average scoring functions			
			Hits	CC	DQ	VQ	ER	CoC	ESR	QL	FZG	NZ	AL
2.5%	Indiv.	GARCH	1.8	0.278	0.774	0.192	0.636	0.214	0.893	0.356	1.043	3.727	3.650
		GJR	2.0	0.464	0.163	0.350	0.810	0.200	0.857	0.354	1.041	3.717	3.644
		GAS	2.4	0.749	0.274	0.837	0.486	0.967	0.538	0.354	1.041	3.702	3.627
		MSGARCH	1.5	0.123	0.551	0.210	0.554	0.068	0.959	0.366	1.053	3.795	3.693
		Boot.	2.1	0.542	0.187	0.213	0.918	0.080	0.911	0.371	1.058	3.813	3.698
		FIGARCH	1.8	0.278	0.778	0.245	0.805	0.078	0.946	0.356	1.042	3.728	3.652
		NGARCH	1.8	0.278	0.778	0.246	0.671	0.191	0.885	0.355	1.042	3.724	3.649
		CAViaR	2.5	0.814	0.002	0.007	0.753	0.393	0.873	0.374	1.061	3.851	3.732
		CAViaREVT	2.4	0.749	0.037	0.088	0.910	0.077	0.954	0.370	1.057	3.808	3.698
		CAViaRALD	2.5	0.814	0.001	0.003	0.729	0.440	0.918	0.385	1.073	3.935	3.795
		AVG	2.0	0.464	0.133	0.332	0.772	0.296	0.829	0.362	1.048	3.758	3.667
		MED	2.0	0.464	0.149	0.311	0.799	0.222	0.897	0.358	1.045	3.742	3.659
		MAX	3.2	0.068	0.000	0.340	0.554	0.488	0.328	0.371	1.058	3.837	3.733
		MIN	1.5	0.123	0.539	0.019	0.931	0.002	0.921	0.376	1.063	3.850	3.722
	Comb.	MSC_{FZG}	2.1	0.524	0.140	0.425	0.658	0.631	0.817	0.362	1.049	3.762	3.669
		MSC_{NZ}	2.1	0.524	0.140	0.466	0.840	0.315	0.819	0.363	1.050	3.763	3.669
		MSC_{AL}	2.1	0.542	0.201	0.350	0.860	0.248	0.869	0.359	1.046	3.742	3.657
		RSC_{FZG}	2.2	0.594	0.001	0.050	0.720	0.511	0.834	0.374	1.060	3.830	3.710
		RSC_{NZ}	2.4	0.749	0.001	0.035	0.753	0.435	0.866	0.375	1.062	3.844	3.720
		RSC_{AL}	2.2	0.647	0.084	0.058	0.737	0.403	0.917	0.372	1.059	3.829	3.712
5%	Indiv.	GARCH	5.5	0.759	0.832	0.465	0.934	0.052	0.933	0.568	1.248	3.323	3.447
		GJR	5.4	0.802	0.548	0.191	0.919	0.062	0.939	0.569	1.249	3.325	3.448
		GAS	6.2	0.259	0.101	0.193	0.822	0.194	0.565	0.573	1.254	3.331	3.448
		MSGARCH	5.0	1.000	0.862	0.629	0.934	0.062	0.974	0.576	1.257	3.359	3.475
		Boot.	4.5	0.476	0.264	0.045	0.854	0.086	0.973	0.587	1.267	3.387	3.490
		FIGARCH	5.2	0.822	0.821	0.300	0.949	0.024	0.972	0.570	1.250	3.330	3.452
		NGARCH	5.4	0.802	0.605	0.504	0.926	0.060	0.935	0.568	1.249	3.325	3.449
		CAViaR	5.2	0.486	0.144	0.222	0.912	0.118	0.960	0.581	1.261	3.371	3.482
		CAViaREVT	5.9	0.233	0.019	0.011	0.958	0.016	0.973	0.586	1.267	3.392	3.498
		CAViaRALD	5.2	0.822	0.077	0.112	0.947	0.044	0.986	0.583	1.264	3.380	3.488
		AVG	5.1	0.816	0.630	0.315	0.913	0.097	0.948	0.573	1.253	3.341	3.460
		MED	5.4	0.802	0.713	0.325	0.938	0.059	0.954	0.571	1.251	3.333	3.454
		MAX	7.0	0.042	0.025	0.016	0.823	0.051	0.407	0.584	1.264	3.372	3.479
		MIN	3.6	0.173	0.214	0.006	0.894	0.002	1.000	0.586	1.266	3.393	3.497
	Comb.	MSC_{FZG}	5.4	0.501	0.257	0.221	0.854	0.276	0.881	0.581	1.261	3.368	3.479
		MSC_{NZ}	5.5	0.499	0.106	0.193	0.959	0.071	0.942	0.582	1.262	3.362	3.470
		MSC_{AL}	5.6	0.247	0.072	0.134	0.971	0.044	0.962	0.580	1.261	3.360	3.469
		RSC_{FZG}	5.4	0.501	0.078	0.112	0.861	0.202	0.914	0.581	1.262	3.368	3.478
		RSC_{NZ}	5.2	0.486	0.084	0.128	0.872	0.151	0.947	0.582	1.262	3.371	3.480
		RSC_{AL}	5.2	0.822	0.145	0.144	0.914	0.088	0.959	0.580	1.260	3.365	3.476

Out-of-sample results for Ethereum

Results for Ethereum are reported in Table 4 and reveal that a number of individual models, useful for forecasting both risk measures in previous studies, remain useful for forecasting the risk measures for this recent set of data.

For the 2.5% risk level, all individual procedures, except the quantile regression-based ones, deliver satisfactory calibration test results. Additionally, six out of the ten combining strategies also deliver satisfactory calibration test results. The model confidence set at 5% significance level, as mentioned previously, does not distinguish between the global performance obtained between implemented methods. However, the fluctuation test re-

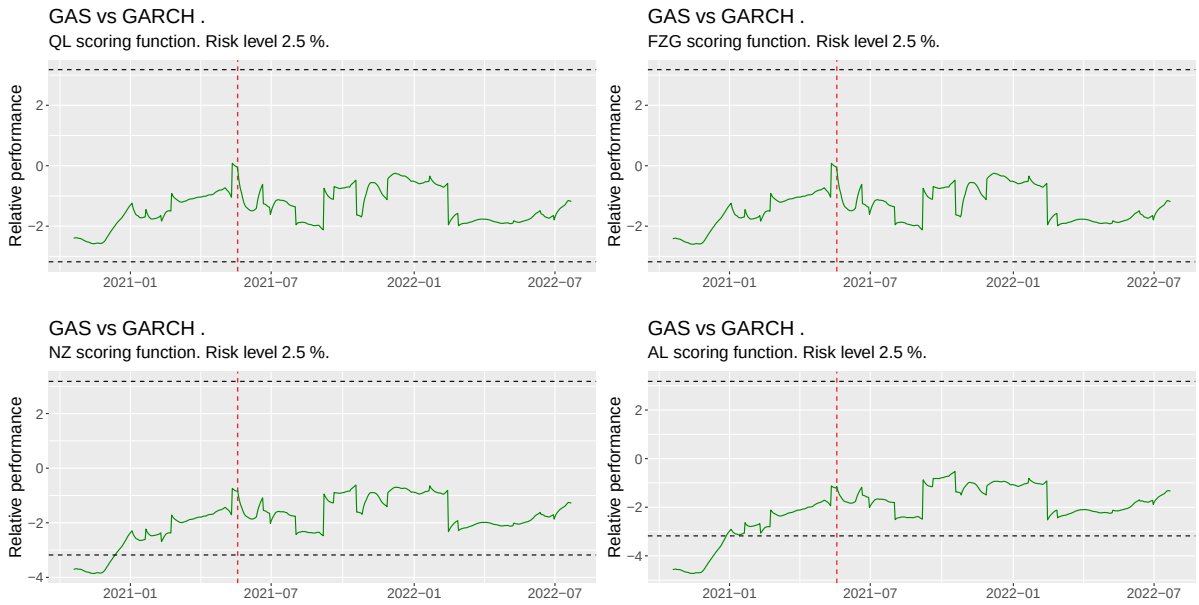

Figure 3: For Ethereum, fluctuation test comparing the 2.5% VaR and ES forecasting performance of the GAS model and GARCH model according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

veals that the GAS model outperforms all other procedures with satisfactory calibration test results in at least one period of time, and is never outperformed by the competitor models. Figures 3 - 6 report the fluctuation test for the GAS model against GARCH, MSGARCH, AVG and MSC_{AL} procedures. The other pairwise comparisons are reported in the Figures S36–S43 of the supplementary material.

For the 5% risk level, six individual models (GARCH, GJR, GAS, MSGARCH, NGARCH and CAViaR) and seven combining strategies (AVG, MED, MSC_{FZG}, MSC_{NZ}, RSC_{FZG}, RSC_{NZ} and RSC_{AL}) deliver satisfactory calibration test results with none of them being preferable according to the model confidence set at 5% of significance level (almost all procedures belong to the model confidence set). Nevertheless, the fluctuation test reveals that, at least for one scoring function considered, the GAS model outperforms all other procedures with satisfactory calibration test results, and is never outperformed by the competitor models. Figures 7 - 9 report the fluctuation test for the GAS model against MSGARCH, CAViaR and RSC_{FZG} procedures. The other pairwise comparisons are reported in Figures S44–S52 of the supplementary material.

In summary, for Ethereum, the GAS model performed particularly well. It delivers satisfactory calibration test results for both risk levels and, at least for a period of time,

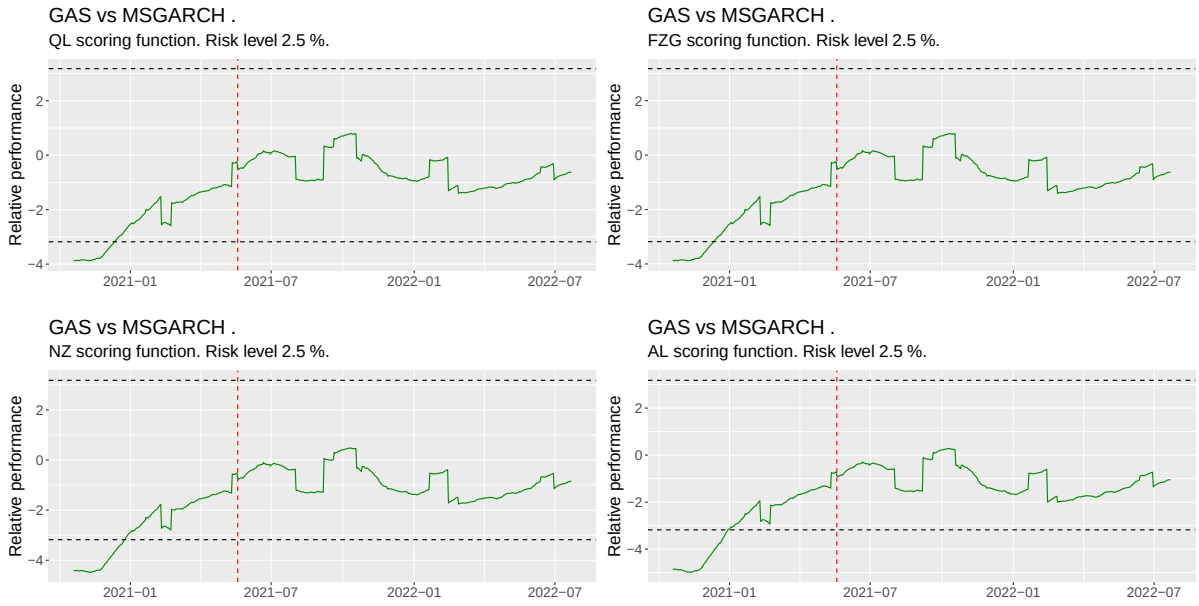

Figure 4: For Ethereum, fluctuation test comparing the 2.5% VaR and ES forecasting performance of the GAS model and MSGARCH model according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

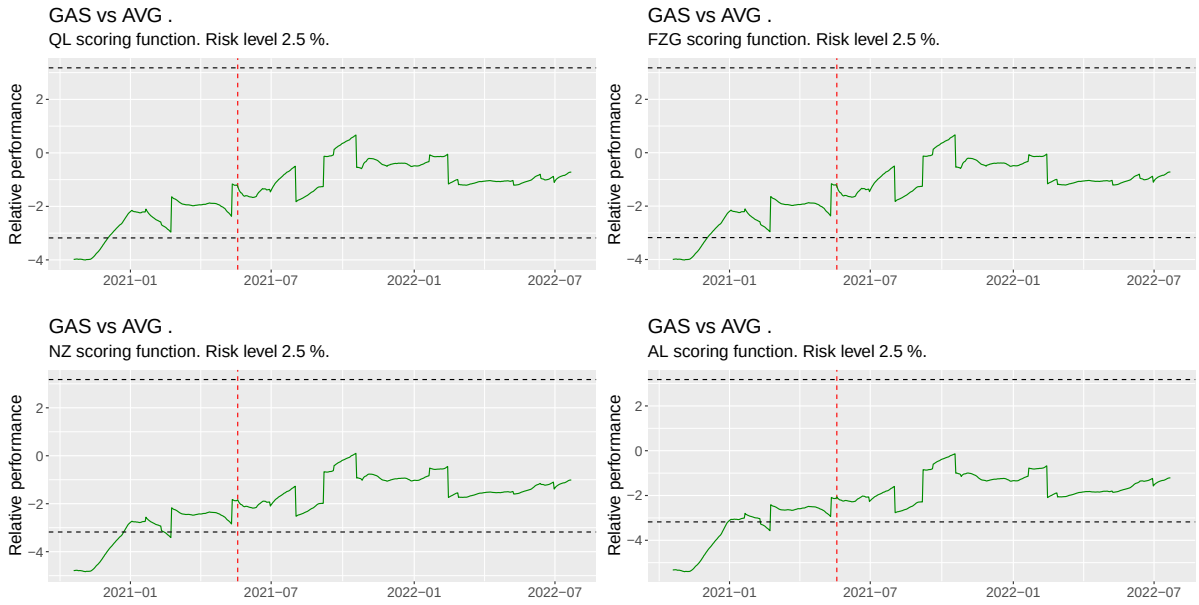

Figure 5: For Ethereum, fluctuation test comparing the 2.5% VaR and ES forecasting performance of the GAS model and AVG combining strategy according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

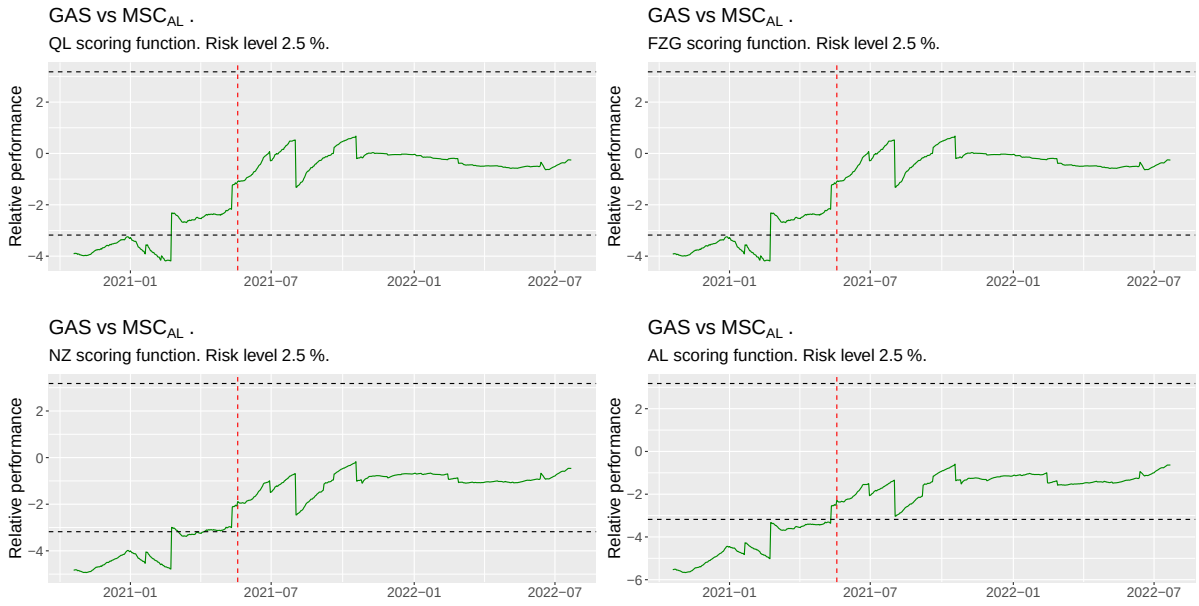

Figure 6: For Ethereum, fluctuation test comparing the 2.5% VaR and ES forecasting performance of the GAS model and MSC_{AL} combining strategy according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

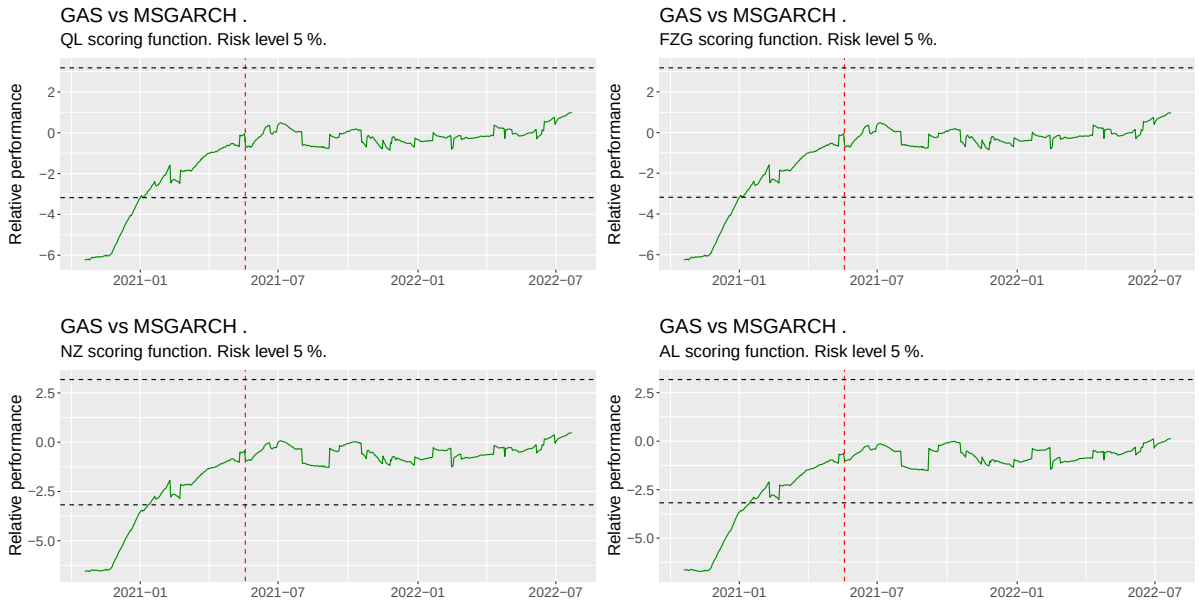

Figure 7: For Ethereum, fluctuation test comparing the 5% VaR and ES forecasting performance of the GAS model and MSGARCH model according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

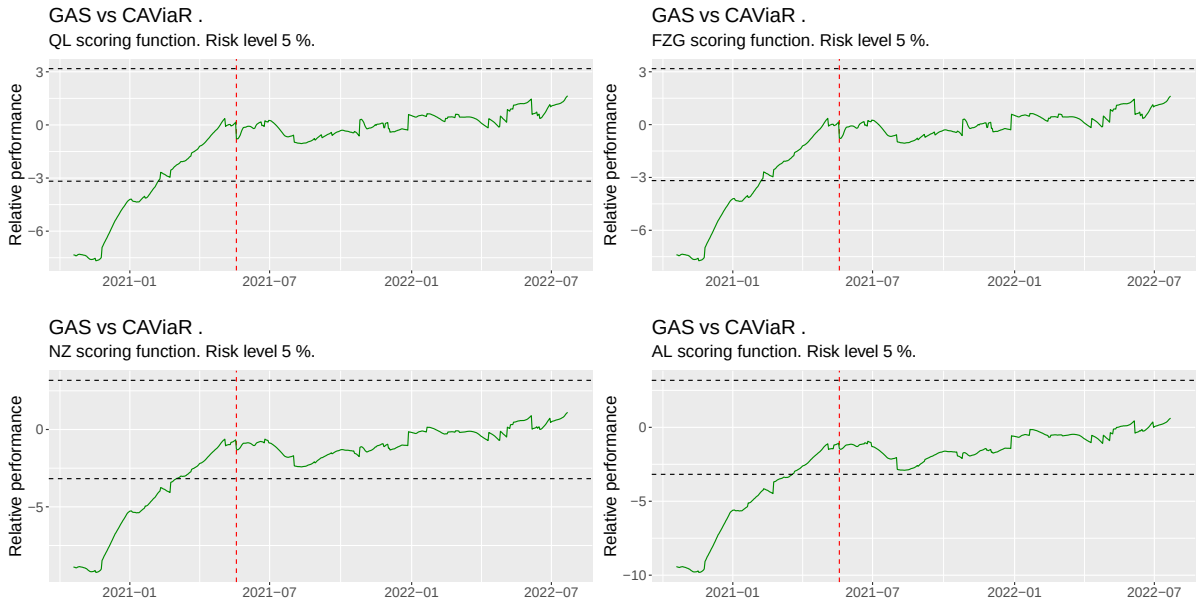

Figure 8: For Ethereum, fluctuation test comparing the 5% VaR and ES forecasting performance of the GAS model and CAViaR model according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

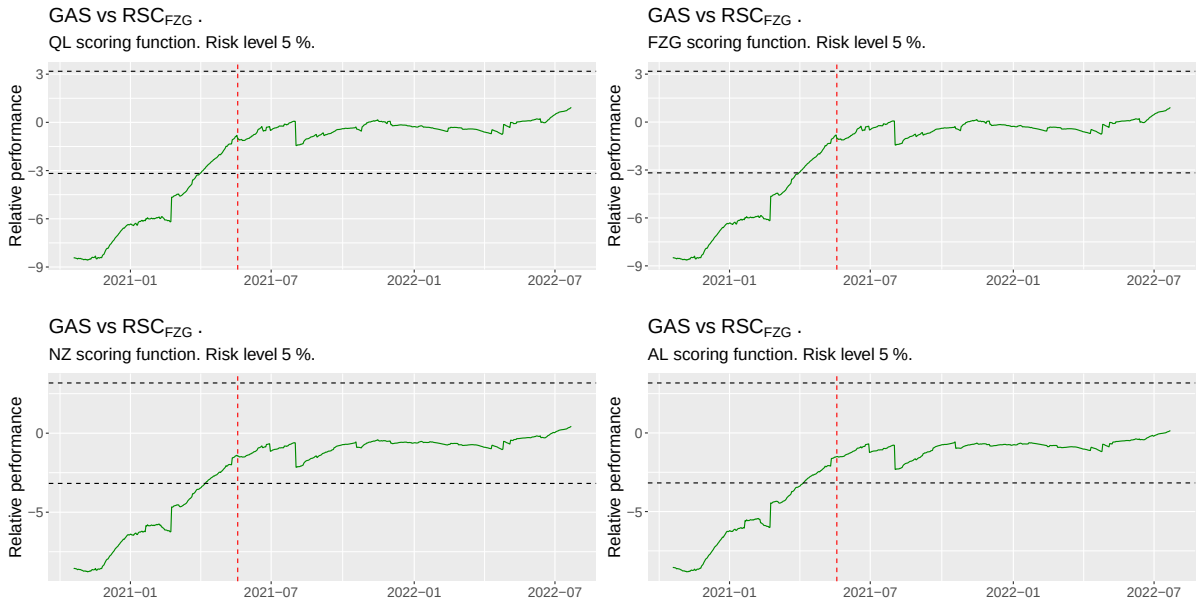

Figure 9: For Ethereum, fluctuation test comparing the 5% VaR and ES forecasting performance of the GAS model and RSC_{FZG} combining strategy according to four scoring functions. Vertical dashed lines indicate the Lexia class action. Dashed horizontal black lines indicate the two-sided 5% fluctuation test critical values.

outperforms other procedures with satisfactory calibration test results.

5 Conclusions and further research

We use recent cryptocurrency data that includes periods of turbulence due to the COVID-19 pandemic, the third halving of Bitcoin and the Lexia class action, to evaluate whether the VaR and ES forecasting performance of models, shown to be useful for forecasting risk measures in the past, remain useful in more recent times for cryptocurrencies. We also evaluate whether the robust bootstrap approach of Trucíos et al. (2017), the GAS model (Harvey, 2013; Creal et al., 2013), as well as the CAViaR model (Engle and Manganelli, 2004), found to be useful for forecasting the VaR of cryptocurrencies in previous studies, are also useful for forecasting the ES of cryptocurrencies. Finally, we evaluate whether forecast combination strategies can improve the VaR and ES forecasts obtained by individual models.

Regarding the individual models, the best performing models differed for the two cryptocurrencies. For Bitcoin, a full set of satisfactory calibration test results is produced by some individual and combining strategies with none of them being superior to the others according to the calibration and fluctuation tests on the scoring functions. For Ethereum, several procedures deliver satisfactory calibration test results for both risk levels, with the GAS model the only one outperforming all its competitor in at least one period of time.

Combining strategies performed reasonably, but they were unable to outperform all individual models, neither globally nor over time. It was perhaps a little surprising that we did not obtain better results with the more sophisticated combining strategies. One explanation for this is the presence of outliers in the cryptocurrency returns series. In addition to having a detrimental effect on some of the individual models in the combination, it can also have an unhelpful influence on the estimation of the combining weights. A further explanation for the combining methods not outperforming all individual models is that the context may not be as fertile for forecast combination as other applications. We must acknowledge that, although there were notable differences in the specifications of the individual models, they were all univariate time series models, which limits the diversity of information provided by each model, and hence limits the potential for improved

accuracy from forecast combining.

In terms of future research, it would be interesting to see research into the impact of outliers on VaR and ES models. A model needs to be robust to outliers, but not so robust that the VaR and ES forecasts do not capture the possibility of such extremes. Our reflections on the limitations of combining models in this application motivates future consideration of models that capture relevant information provided by other sources of data when modelling the cryptocurrency returns. Another interesting research direction would be consideration of other forms of combining models, such as the recent paper by Lu et al. (2021), which applies machine learning to combine forecasts of VaR and ES. A final comment on future research is that, as time passes, longer time series will obviously become available, which will potentially enable researchers to develop a revised understanding of which individual models and combining strategies are most useful.

References

Acereda, B., Leon, A., and Mora, J. (2020). Estimating the expected shortfall of cryptocurrencies: An evaluation based on backtesting. *Finance Research Letters*, 33:101181.

Alexander, C. and Dakos, M. (2020). A critical investigation of cryptocurrency data and analysis. *Quantitative Finance*, 20(2):173–188.

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

Ardia, D., Bluteau, K., Boudt, K., Catania, L., and Trottier, D.-A. (2019a). Markov-switching GARCH models in R: The MSGARCH package. *Journal of Statistical Software*, 91(4):1–38.

Ardia, D., Bluteau, K., and Rüede, M. (2019b). Regime changes in Bitcoin GARCH volatility dynamics. *Finance Research Letters*, 29:266–271.

Ardia, D., Boudt, K., and Catania, L. (2019c). Generalized autoregressive score models in R: The GAS package. *Journal of Statistical Software*, 88(6):1–28.

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, 36(1):197–200.

Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1):3–30.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41.

Bauwens, L., Preminger, A., and Rombouts, J. V. (2010). Theory and inference for a Markov switching GARCH model. *The Econometrics Journal*, 13(2):218–244.

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

Bayer, S. and Dimitriadis, T. (2020). Regression based expected shortfall backtesting. *Journal of Financial Econometrics*.

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

Boudt, K., Danielsson, J., and Laurent, S. (2013). Robust forecasting of dynamic conditional correlation GARCH models. *International Journal of Forecasting*, 29(2):244–257.

Bouri, E., Jalkh, N., Molnár, P., and Roubaud, D. (2017). Bitcoin for energy commodities before and after the december 2013 crash: diversifier, hedge or safe haven? *Applied Economics*, 49(50):5063–5073.

Buczyński, M. and Chlebus, M. (2019). Old-fashioned parametric models are still the best: a comparison of value-at-risk approaches in several volatility states. *Journal of Risk Model Validation*, 14(2).

Calmon, W., Ferioli, E., Lettieri, D., Soares, J., and Pizzinga, A. (2020). An extensive comparison of some well-established value at risk methods. *International Statistical Review*, 89(1):148–166.

Caporale, G. M. and Zekokh, T. (2019). Modelling volatility of cryptocurrencies using Markov-Switching GARCH models. *Research in International Business and Finance*, 48:143–155.

Carnero, M. A., Peña, D., and Ruiz, E. (2012). Estimating GARCH volatility in the presence of outliers. *Economics Letters*, 114(1):86–90.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

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

Danielsson, J. (2011). Risk and crises. voxeu.org url:. *http://voxeu.org/ article/risk-and-crises-how-models-failed-and-are-failing.*

Dimitriadis, T. and Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, 13(1):1823–1871.

Dyhrberg, A. H. (2016). Bitcoin, gold and the dollar – A GARCH volatility analysis. *Finance Research Letters*, 16:85–92.

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

Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5):1749–1778.

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

Fissler, T., Ziegel, J. F., and Gneiting, T. (2016). Expected Shortfall is jointly elicitable with Value at Risk - implications for backtesting. *Risk*, 5(1):8–16.

Francq, C. and Zakoïan, J.-M. (2009). Bartlett's formula for a general class of nonlinear processes. *Journal of Time Series Analysis*, 30(4):449–465.

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011). Evaluating value-at-risk models via quantile regression. *Journal of Business & Economic Statistics*, 29(1):150–160.

Ghalanos, A. (2020). *rugarch: Univariate GARCH models.* R package version 1.4-4.

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

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

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

González-Rivera, G., Lee, T.-H., and Mishra, S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting*, 20(4):629–645.

Haas, M., Mittnik, S., and Paolella, M. S. (2004). A new approach to markov-switching GARCH models. *Journal of financial Econometrics*, 2(4):493–530.

Halbleib, R. and 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.

Hallin, M. and Trucíos, C. (2021). Forecasting value-at-risk and expected shortfall in large portfolios: A general dynamic factor model approach. *Econometrics and Statistics*.

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

Happersberger, D., Lohre, H., and Nolte, I. (2020). Estimating portfolio risk for tail risk protection strategies. *European Financial Management*.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series*, volume 52. Cambridge University Press.

Hill, G. W. (1970). Algorithm 396: Student's t-quantiles. *Communications of the ACM*, 13(10):619–620.

Hillebrand, E. (2005). Neglecting parameter changes in GARCH models. *Journal of Econometrics*, 129(1-2):121–138.

Hoogerheide, L. and van Dijk, H. K. (2010). Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling. *International Journal of Forecasting*, 26(2):231–247.

Hotta, L. K. and Trucíos, C. (2018). Inference in (M)GARCH Models in the Presence of Additive Outliers: Specification, Estimation and Prediction. In Lavor, C. and Neto, F. A. M. G., editors, *Advances in Mathematics and Applications*, pages 179–202. Springer.

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

Li, Z., Dong, H., Floros, C., Charemis, A., and Failler, P. (2021). Re-examining bitcoin volatility: A CAViaR-based approach. *Emerging Markets Finance and Trade*, pages 1–19.

Liu, W., Semeyutin, A., Lau, C. K. M., and Gozgor, G. (2020). Forecasting value-at-risk of cryptocurrencies with riskmetrics type models. *Research in International Business and Finance*, 54:101259.

Lu, X., Liu, C., Lai, K. K., and Cui, H. (2021). Risk measurement in Bitcoin market by fusing LSTM with the joint-regression-combined forecasting model. *Kybernetes*.

Luther, W. J. and Salter, A. W. (2017). Bitcoin and the bailout. *The Quarterly Review of Economics and Finance*, 66:50–56.

Maciel, L. (2020). Cryptocurrencies value-at-risk and expected shortfall: Do regime-switching volatility models improve forecasting? *International Journal of Finance & Economics*.

Manganelli, S. and Engle, R. F. (2004). A comparison of value-at-risk models in finance. In *Risk measures for the 21st century*, pages 123–144. Wiley, Chichester.

McAleer, M., Jimenez-Martin, J.-A., and Perez-Amaral, T. (2013a). GFC- risk management strategies under the Basel accord. *International Review of Economics & Finance*, 27:97–111.

McAleer, M., Jiménez-Martín, J.-Á., and Pérez-Amaral, T. (2013b). International evidence on GFC- forecasts for risk management under the Basel accord. *Journal of Forecasting*, 32(3):267–288.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3-4):271–300.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Technical report.

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

Nolde, N., Ziegel, J. F., et al. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4):1833–1874.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W., and Kahaner, D. K. (2012). *Quadpack: a subroutine package for automatic integration*, volume 1. Springer Science & Business Media.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Righi, M. B. and Ceretta, P. S. (2015). A comparison of expected shortfall estimation models. *Journal of Economics and Business*, 78:14–47.

Soylu, P. K., Okur, M., Çatıkkaş, Ö., and Altintig, Z. A. (2020). Long memory in the volatility of selected cryptocurrencies: Bitcoin, Ethereum and Ripple. *Journal of Risk and Financial Management*, 13(6):107.

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

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

Thomson, M. E., Pollock, A. C., Önkal, D., and Gönül, M. S. (2019). Combining forecasts: Performance and coherence. *International Journal of Forecasting*, 35(2):474–484.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196.

Troster, V., Tiwari, A. K., Shahbaz, M., and Macedo, D. N. (2019). Bitcoin returns and risk: A general GARCH and GAS analysis. *Finance Research Letters*, 30:187–193.

Trucíos, C. (2019). Forecasting Bitcoin risk measures: A approach. *International Journal of Forecasting*, 35(3):836–847.

Trucios, C. (2020). *RobGARCHBoot: Robust Bootstrap Forecast Densities for GARCH Models*. R package version 1.1.0.

Trucíos, C. and Hotta, L. K. (2016). Bootstrap prediction in univariate volatility models with leverage effect. *Mathematics and Computers in Simulation*, 120:91–103.

Trucíos, C., Hotta, L. K., and Ruiz, E. (2015). Robust bootstrap forecast densities for GARCH models: returns, volatilities and value-at-risk. *UC3M Working Papers Statistics and Econometrics*, 15(22).

Trucíos, C., Hotta, L. K., and Ruiz, E. (2017). bootstrap forecast densities for GARCH returns and volatilities. *Journal of Statistical Computation and Simulation*, 87(16):3152–3174.