Predicting value at risk for cryptocurrencies with generalized random forests ☆

Rebekka Buse [a], Konstantin Görgen [b], Melanie Schienle [a,*]

[a] *Karlsruhe Institute of Technology, Germany*
[b] *Allianz Zürich, Switzerland*

ARTICLE INFO

ABSTRACT

We study the prediction of value at risk (VaR) for cryptocurrencies. In contrast to classic assets, returns of cryptocurrencies are often highly volatile and characterized by large fluctuations around single events. Analyzing a comprehensive set of 105 major cryptocurrencies, we show that generalized random forests (GRF) adapted to quantile prediction have superior performance over other established methods such as quantile regression, GARCH-type models, and CAViaR models. This advantage is especially pronounced in unstable times and for classes of highly volatile cryptocurrencies. Furthermore, we identify important predictors during such times and show their influence on forecasting over time. Moreover, a comprehensive simulation study indicates that the GRF methodology is at least on par with existing methods in VaR predictions for standard types of financial returns, and clearly superior in the cryptocurrency setup.

1. Introduction

Cryptocurrencies are an important and rising part of today's digital economy. Currently, the market capitalization of the top 10 cryptocurrencies in the world is close to $2 trillion and growing.[1] The use of cryptocurrencies in terms of daily volume exploded from 2016 to 2018, attracting individuals and business users such as hedge funds or merchants as well as long-term investors such as crypto-focused and traditional investment funds (Vigliotti & Jones, 2020). The crypto asset market, however, remains highly volatile. For example, an investment in Bitcoin in 2013 would have seen a return of roughly 20, 000% in 2017, but an investment in 2017 would have led to a performance of -75% in 2019.[1] Consequently, there is

a need to predict and monitor the risks associated with cryptocurrencies. To address this, we find that classic approaches such as historical simulation, GARCH-type, or CAViaR methods are too restrictive. More general non-linear methods provide more flexibility to account for such non-standard time series behavior that might be attributed to a large extent to speculators.

In this paper, we propose a novel flexible way for out-of-sample prediction of the value at risk for cryptocurrencies. We use a quantile version of generalized random forests (GRF, see Athey, Tibshirani, & Wager, 2019), which builds upon mean random forests (Breiman, 2001) now tailored to quantiles. This framework proves to be especially promising when dealing with more volatile classes of cryptocurrencies, due to the non-linear structure of their returns. In a comprehensive out-of-sample scenario using more than 100 of the largest cryptocurrencies, GRF outperforms other established methods such as CAViaR (Engle & Manganelli, 2004), quantile regression (Koenker & Hallock, 2001), and GARCH models (Bollerslev, 1986; Glosten, Jagannathan, & Runkle, 1993) over a rolling window, particularly in unstable times.

This can be attributed to the non-parametric approach of random forests that is flexible and adaptable, considering important factors and non-linearity. We further analyze performance in different important subperiods, consider different classes of cryptocurrencies, and employ different sets of covariates with the forest-based methods and the benchmark procedures.

Previous studies have confirmed that there exist speculative bubbles (Cheah & Fry, 2015; Hafner, 2020), and we find that our approach assesses risks especially well during such times. Moreover, we account for a large number of covariates that describe volatility, liquidity, and supply (Liu & Tsyvinski, 2020). It can be seen that variable importance differs substantially depending on time, where long-term measures of standard deviation, an important predictor in stable times, are not relevant predictors for VaR in unstable, volatile times. Furthermore, only a few of the additional covariates besides lagged standard deviations and lagged returns are relevant. We find that especially for other, less volatile classes of cryptocurrencies such as stablecoins, GJR-GARCH models and quantile regression can compete with GRF. A simulation study also highlights that the proposed GRF methodology is at least on par in the prediction of VaR for standard-type financial returns, with clear advantages in the cryptocurrency-type case.

Our paper contributes to the growing literature on cryptocurrencies. Analyses performed in the past include GARCH models (Chu, Chan, Nadarajah, and Osterrieder, 2017) as well as ARMA-GARCH models (Platanakis & Urquhart, 2019), approaches using RiskMetrics (Pafka & Kondor, 2001) and GAS models (Liu, Semeyutin, Lau, & Gozgor, 2020), the application of extreme value theory (Gkillas & Katsiampa, 2018), vine copula-based approaches (Trucios, Tiwari, & Alqahtani, 2020), Markov-switching GARCH models (Maciel, 2020), non-causal autoregressive models (Hencic & Gouriéroux, 2015), and also some machine learning-based approaches (see e.g. Takeda & Sugiyama, 2008). Additionally, cryptocurrencies can be used for diversification in investment strategies with other, traditional assets (see e.g. Petukhina, Trimborn, Härdle, & Elendner, 2021; Trimborn, Li, & Härdle, 2020), as the correlation between them and more established assets tends to be low (Elendner, Trimborn, Ong, & Lee, 2017; Platanakis & Urquhart, 2019). This again poses the question of assessing the risks of cryptocurrencies, where new methods of addressing the abovementioned challenges need to be explored.

The paper is structured as follows. Section 2 presents the underlying data and cryptocurrencies with descriptive statistics and standard Box–Jenkins time series checks. In Section 3, we introduce the main random forest-type techniques for conditional quantiles and present the employed evaluation tests and framework. The comprehensive simulation study in Section 4 demonstrates the performance of the different methods under various data generating processes. The empirical results are contained in Section 5, where we present aggregate prediction results for all currencies in coverage performance (Section 5.1.1) as well as in pairwise comparison tests (Section 5.1.2). In Section 5.2 we provide extensions to the

main results. The section comprises a detailed analysis of important representative currencies, a study of the importance of specific factors, and results for the most recent but shorter post-pandemic period. Finally, we conclude in Section 6. All data and replication materials can be found in the Github repository https://github.com/KITmetricslab/crypto-VaR-predictions.

2. Data

We use daily log returns from Coin Metrics of 105 of the largest cryptocurrencies by market capitalization in US dollars (USD) at the time of retrieval in the period from 07/2010 to 04/2024.[2] The time to define a day is based on Coordinated Universal Time (UTC). The Coin Metrics data include spot-market information from 30 different exchanges, such as Binance, ZB.COM, FTX, OKX, Coinbase, KuCoin, and Kraken. Depending on the currency, the number of available observations varies between 261 and 5012.

The distribution of the data over time and assets is summarized in Table 1. In the cross-section, cryptocurrency returns are characterized by large positive and negative spikes and substantial excess kurtosis in over 75% of all assets. Furthermore, they show substantial skewness, indicating asymmetry in the distribution of log returns and heterogeneity of in the cross-section of considered assets. Remarkably, both positive and negative skewness occurs in the cross-section, and skewness can be classified in more than 10% of the considered assets as high.

We do not detect any stochastic non-stationarities in the data, which is supported by augmented Dickey–Fuller (ADF) tests and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). With Alpha Finance Lab (alpha), Polymath (poly), and Synthetix (snx), KPSS tests against level stationarity seem slightly significant, while trend KPSS tests and ADF tests suggest stationarity. With Algorand (algo), Binance Coin (bnb), Curve DAO Token (crv), FTX Token (ftt), Internet Computer (icp), Aave (lend), OMG Network (omg), SushiSwap (sushi), and Monero (xmr), KPSS tests against trend stationarity are slightly significant, while ADF tests and level KPSS tests again suggest stationarity. All results of the stationarity tests can be found in Table A.14 in the appendix.

Fig. 1 illustrates the median returns (black) over all cryptocurrencies by date. We can see that in the beginning of the time period, only one currency, namely Bitcoin, was present in the data set. From 2014, we see an incremental increase (red line), while there is a jump up in 2017 and a consecutively faster increase in available cryptocurrencies.

---

[2] All presented results are based on data downloaded on 8 April 2024 using the Coin Metrics community data set, which can be downloaded from a public Github repository at https://github.com/coinmetrics/data/. See https://docs.coinmetrics.io/exchanges/all-exchanges for an overview of all exchanges included. All currencies have a maximum market capitalization of more than $15 million USD each. We mark the start of the availability period by the first recorded trade date on any of the exchanges.

Table 1
Descriptive statistics of log returns of cryptocurrencies.

	Min	1%	5%	Median	95%	99%	Max	Skewness	Excess kurtosis	Std. Dev.	Observations
Min%	−1.402	−0.437	−0.169	−0.006	0.001	0.002	0.008	−4.030	3.214	0.001	260.00
1%	−1.378	−0.306	−0.158	−0.004	0.001	0.002	0.015	−3.852	3.412	0.001	352.40
5%	−1.104	−0.246	−0.128	−0.003	0.002	0.006	0.044	−1.214	4.950	0.003	1044.00
25%	−0.635	−0.189	−0.105	−0.001	0.078	0.160	0.398	−0.207	7.968	0.056	1397.75
50%	−0.506	−0.168	−0.094	−0.000	0.094	0.196	0.505	0.364	14.195	0.066	2024.00
75%	−0.378	−0.146	−0.081	0.001	0.106	0.215	0.738	1.094	26.293	0.074	2416.25
95%	−0.041	−0.007	−0.002	0.001	0.137	0.316	1.374	2.349	120.236	0.094	3591.25
99%	−0.019	−0.002	−0.001	0.002	0.196	0.466	1.431	3.345	320.802	0.135	4009.15
Max%	−0.006	−0.002	−0.001	0.002	0.202	0.541	1.462	3.619	326.429	0.141	5012.00

For each cryptocurrency in the sample, we calculate the sample statistics indicated by the column headers. The rows show the quantiles of each of these statistics in the cross-section, thus marking key points of the cross-sectional distribution of the respective statistics in each column. For example, over the cross-section of cryptocurrencies, the median number of available time series observations is 2024.
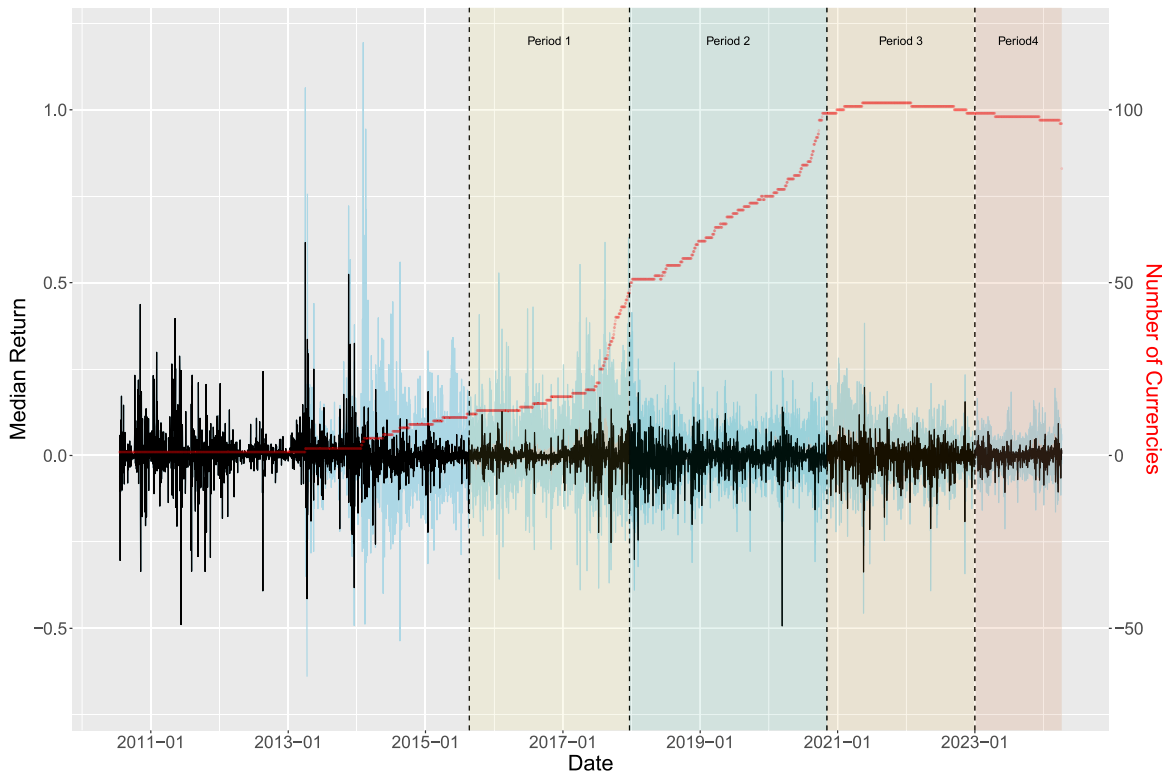

Fig. 1. For each point in time, the figure displays pointwise median returns over all currencies in black, with 5% and 95% sample quantiles in blue. The number of available currencies with non-zero returns is indicated in red. The shaded areas in yellow, green, and orange correspond to the three periods of the main study in Section 5. The substantially shorter red shaded time period is analyzed in Section 5.2.

Given the hugely dynamic evolution of the cryptocurrency market, we distinguish between three main systematically different periods in terms of actively traded currencies and market capitalization and treat them separately. The first period (Period 1) ranges from August 2015 (2015-08-22), where the crypto market was still rising with a relatively small number of available currencies, to the end of 2017 (2017-12-21). It is characterized by a few rapidly changing currencies and extreme returns and volatilities. The second period (Period 2) is less extreme in returns and shows a strong increase in median

market caps with a few dozens of market components. It lasts until November 2020 (2020-11-05) with an increasing number of currencies capturing the beginning of the Covid-19 pandemic. The third period (Period 3) ranges from November 2020 until December 2022 (2022-12-31), with a large number of cryptocurrencies containing more than four times the number of currencies in comparison to Period 2. Period 3 is also marked by the endemic stage of the Covid-19 pandemic. In an extension in Section 5.2.2 we also study the substantially shorter Period 4 (2023-01-01 until 2024-04-06) that is characterized by lower

Table 2
Summary of covariates for different time periods.

Quantile	Ret	Active_Users	Users	Users >$100	Users >$10	SER	Transactions	Velocity	sd_3	sd_7	sd_30	sd_60	Volume
Period 1: Five currencies													
5%	−0.090	9 059	266 499	18 640	40 353	0.002	3 781	6.898	0.007	0.014	0.022	0.025	21
Median	−0.000	156 216	4 934 148	670 847	1 482 211	0.019	56 929	32.161	0.034	0.041	0.049	0.051	9155
95%	0.097	333 644	12 730 286	3 867 047	6 670 343	0.036	179 828	110.209	0.130	0.125	0.118	0.117	188 918
Period 2: 14 currencies													
5%	−0.085	5 147	137 498	11 914	25 839	0.001	19 971	3.823	0.007	0.014	0.022	0.025	109
Median	−0.000	96 620	5 585 463	483 593	1 113 547	0.013	200 836	17.823	0.034	0.040	0.047	0.049	9458
95%	0.092	209 115	15 533 203	2 831 599	6 082 908	0.055	745 090	83.331	0.123	0.118	0.112	0.111	107 684
Period 3: 64 currencies													
5%	−0.084	3 103	731 930	20 990	57 769	0.003	31 838	4.014	0.008	0.015	0.023	0.026	255
Median	−0.000	48 931	2 302 405	201 705	481 099	0.008	158 681	16.833	0.034	0.040	0.046	0.049	4368
95%	0.085	118 964	5 377 667	870 485	1 911 853	0.020	461 557	68.267	0.115	0.110	0.105	0.101	30 618
Period 4: 83 currencies													
5%	−0.085	2 497	559 538	18 741	48 212	0.002	24 963	5.183	0.008	0.016	0.023	0.027	395
Median	−0.000	31 607	1 795 653	132 859	293 588	0.006	120 609	15.492	0.035	0.041	0.047	0.050	3408
95%	0.086	79 546	3 793 313	588 142	1 245 011	0.014	355 595	61.036	0.114	0.110	0.105	0.100	25 221
Full data: 106 currencies													
5%	−0.083	1 980	434 899	15 289	38 671	0.002	19 505	4.926	0.008	0.015	0.022	0.026	406
Median	−0.000	30 350	1 525 012	129 358	304 196	0.005	105 990	15.432	0.034	0.040	0.046	0.049	3538
95%	0.085	75 129	3 452 320	522 625	1 145 034	0.013	311 205	79.758	0.114	0.111	0.105	0.100	21 640

The table displays summary statistics for all included covariates in the different time periods. Please see Table A.15 for the definitions. Displayed values are means of quantiles over all assets contained in the specific time period. "User" corresponds to the variable Total_Users, "Users >$100" marks Total_Users_USD100, and "Users >$10" displays the variable Total_Users_USD10. Additionally in the last column ("Volume"), we provide information on market caps in millions of USD.

cryptocurrency volatility in a post-pandemic setting influenced by the political and economic consequences of the war in Ukraine.[3]

To capture the autoregressive and time series structure of the asset processes, we include the classic one-period lagged return in our analysis and the lagged three-, seven-, 30-, and 60-day return standard deviations. Additionally, we employ information specific to each cryptocurrency in seven external covariates. The external covariates are the number of unique active daily addresses (Active_Users), the number of unique addresses that hold any amount of native units of that currency or at least $10 or $100 USD equivalent (Total_Users, Total_Users_USD10, Total_Users_USD100), the supply equality ratio (SER), i.e. the ratio of supply held by addresses with less than $1/10^7$ of the current supply to the top 1% of addresses with the highest current supply, the number of initiated transactions (Transactions), and the velocity of supply in the current year (Velocity), which describes the ratio of current supply to the sum of the value transferred in the last year. See also Table A.15 in the appendix for details on the covariates.

Table 2 gives an overview of the employed covariates and their values in the different time periods. We can see that in Period 1, we have the most extreme returns on average, as well as the most extreme lagged standard deviations. This is not surprising looking at the first period (853 days), which arguably marks the most volatile period, with many new currencies being created, as well as the second-longest period. In the following period (1050 days), the average median market cap reaches a high as well as the number of users invested in the currencies,

indicating that the market is growing while stabilizing more. This is followed by a sharp drop in the market cap for the last, shortest period (500 days), which starts at the beginning of the Covid-19 pandemic. There, the number of active users, as well as the SER, decreases, indicating that more smaller addresses are pushed off the market, while it is the period with the most currencies.

3. Methodology

To predict the VaR of cryptocurrencies, we advocate the use of tailored non-linear machine learning-based techniques. In this way, we intend to accommodate the large documented share of speculation (Baur, Hong, & Lee, 2018; Ghysels & Nguyen, 2019; Glaser, Zimmermann, Haferkorn, Weber, & Siering, 2014; Selmi, Tiwari, & Hammoudeh, 2018) and the resulting frequent changes in unconditional volatility, which make predictions in this market peculiar. In particular, we focus on generalized random forest methods that are tailored for conditional quantiles of returns and can thus forecast the VaR. The flexible but interpretable non-linearity of the approach allows for a direct comparison to standard linear and (G)ARCH-type models. We also argue that the difference in forecasting performance can be employed to detect periods of bubbles and extensive speculation.

Recall that for daily log returns r_t the VaR_t at level $\alpha \in (0, 1)$ conditional on some covariates x_{t-1} is defined as

$$VaR_t^\alpha(x_{t-1}) = \sup_{r_t} (F(r_t|x_{t-1}) < \alpha),$$ (1)

where F marks the distribution of r_t conditional on x_{t-1}. Generally, the conditioning variables could consist of past lagged returns and standard deviations, but also of external (market) information or other assets. We employ these as covariates that are explained in Section 2.

---

[3] The specific cutoff dates for the periods account for the required training time in model fitting for each currency and the emergence of new assets to ensure that we capture a maximum number of cryptocurrencies in each time period.

We propose a specific random forest-based technique which directly models the conditional VaR in (1), and we evaluate its prediction performance relative to a different random forest-type benchmark and standard parametric time series models. We closely build on existing techniques by Athey et al. (2019) and Meinshausen (2006) that were suggested and have so far only been used in a cross-sectional setting, where performance of both technologies has been similar. For forecasting financial time series and in particular time series of cryptocurrencies, however, we show by comprehensive simulations and empirically that a tailored specification of one approach can yield advantages over existing models, while the other procedure yields only less favorable results. Both considered random forest methodologies rely on the classic random forest idea (Breiman, 2001) that produces an ensemble of (decorrelated) decision trees (see e.g. Hastie, Tibshirani, & Friedman, 2009) for the mean of r_t. In a decision tree, each outcome r_t is sorted into leafs of the tree by binary splits. These splits are performed based on different x_{t-1} components falling above or below specific adaptive threshold values that need to be calculated, for example by the Gini impurity or MSE splitting in classification and regression trees (CART) (Breiman, Friedman, Stone, & Olshen, 1984), or using other criteria. Finally, the prediction for a new r_t is a weighted version of each tree prediction.

In our suggested methodology we rely on the generalized random forests (GRF) from Athey et al. (2019). In this technique, the random forest split criterion is adapted to mimic the task of quantile regression, rather than minimizing a standard mean-squared loss criterion for mean regression tasks. Intuitively, the splits in each leaf are conducted by minimizing the Gini loss, which separates the returns r_t as best as possible at different quantiles. To transform the minimizing problem in the splits into a classification task, the response variable r_t is transformed in each split to obtain pseudo-outcomes $\rho_t = \sum_{m=1}^{M} 1\{r_t > \theta_{qm}\}$, where $1\{r_t > \theta_{qm}\}$ is 1 for all t with $r_t > \theta_{qm}$, and zero otherwise, and $\Theta = (\theta_{q1}, \ldots, \theta_{qM})$ describes a set of M pilot quantiles of r_t in the parent node. These quantiles with levels $\tau = q_1, \ldots, q_M$ are then used to calibrate the split. In Athey et al. (2019), this is formally motivated by moment conditions and gradient approximations, but practically, r_t is relabeled to a nominal scale depending on the largest quantile it does not exceed. In a final step, the optimal split on a variable component p of x_{t-1} and $j = 1, \ldots, J$ observations in the parent node is then based on minimizing the above-mentioned Gini impurity criterion for classification. For a separation into two possible leaf sets $v \in \{l, r\}$, the Gini impurity for one leaf v is $G_p^v = 1 - \sum_{k=1}^{K} p_{k,v}^2$, where $p_{k,v} = \sum_{j=1}^{J} 1\{\rho_j = k \text{ and } \rho_j \in v\}/|v|$ is the proportion of ρ_j in group v with value $k = 1, \ldots, K$. The full loss is then an average weighted by leaf size, yielding

$$G_p = (|l|G_p^l + |r|G_p^r)/(|l| + |r|). \tag{2}$$

In our tailoring of the GRF for cryptocurrencies we feature the Gini loss, since it is fast and in general produces purer nodes than, for example, using entropy as a splitting criterion (see e.g Breiman, 1996). This can be particularly helpful when dealing with changing variance (and thus time-varying quantiles) of returns, where we would like to detect single extreme events. For our specific case of $\alpha = 0.05$, this implies that values larger than $\theta_{0.05}$ in the parent node are given the value 1, while others are 0. Algorithm 1 briefly summarizes the tree building algorithm from Athey et al. (2019) for the quantile version of GRF, where the main differences with regard to the splitting regime in comparison to a classic CART occur in every step the tree is grown.

Algorithm 1 Generalized random forests – Tree building.

Input: Set of "honest", subsampled observations X_T and R_T; minimum node size n_m; quantile levels $\tau = (\tau_1, \ldots, \tau_M)$

1: **Growing the tree** Create root node P_0
2: Initialize queue Q with P_0
3: **while** Queue is not empty **do**
4:   Take the oldest element from Q (parent node P) and remove it from Q
5:   Take a random subsample of p variables index by set $P_{sub} = \{\tilde{1}, \ldots, \tilde{p}\}$ from X_T on which to potentially split and take observations $x_i^{(P_{sub})} = (x_i^{(\tilde{1})}, \ldots, x_i^{(\tilde{p})})$ from P.
6:   Set $loss = \infty$
7:   **for** h in 1 to p **do**
8:     Compute quantiles θ_m of r_t from parent node P at $\tau_1, \ldots, \tau_M$ and compute pseudo-outcomes $\rho_t = \sum_{m=1}^{M} 1\{r_t > \theta_m\}$ for each $r_t \in P$.
9:     For each possible split point in $x^{(h)}$, compute the criterion from Equation (2)
10:     Save loss s_h that minimizes this splitting criterion
11:     Save optimal split point $split_h$
12:     **if** $s_h < loss$ **then**
13:       $loss \leftarrow s_h$
14:       $ind_h \leftarrow h$
15:     **end if**
16:   **end for**
17:   **if** Split on variable h with $split_{ind_h}$ succeeded (based on hyperparameters) **then**
18:     Determine children C_1, C_2 according to optimal split
19:     Add both children C_1, C_2 to a new daughter node each with corresponding observations left and add these to Q
20:   **end if**
21: **end while**
Output: One tree of the forest

As a benchmark, we employ the quantile regression forest (QRF) based on Meinshausen (2006). This random forest, however, obtains the target α-quantile by a weighted average of the empirical CDF $\hat{F}(r_t|x_{t-1}) = E[1_{\{r_t\}|x_{t-1}}]$, rather than from averaging minimiziers of the direct moment condition, which can suffer from robustness problems. Intuitively, log returns r_t that have similar x_{t-1} in comparison to a new observation x_v receive higher weight in the empirical CDF. Similarity weights $w_t(x_v)$ are measured as the relative frequency of how often x_v falls in the same terminal leaf as x_{t-1}, for $t = 1, \ldots, T$, and averaged over all trees for each x_{t-1}. This last step was originally introduced by Meinshausen (2006) for random forests. QRF employs the same splitting regime as the original CART random forest, and therefore does not account explicitly for situations where the variance, and therefore the quantile, changes, as splits are conducted based on a mean-squared error criterion. Since such volatility changes are a feature of our cryptocurrency data, we expect GRF to perform better than QRF, but still include both in the analysis to see potential differences in

predictions. Furthermore, GRF uses so-called honest trees, meaning that different data (usually, the subsampled data for each tree are split again in half) are used for building and filling each of the trees with values.

For both forest-type techniques, we use the automatic simple cross-validation choice for the hyperparameters based on a pre-training set with $K = 1$ and $\tau = \alpha$ in the GRF case of Algorithm 1, and the automatic build-in procedure for QRF in the respective R package QRF. In both cases, the minimum node size is set to 20.

As benchmarks we further include two types of standard time series methods. We use the CAViaR (CAV) methodology by Engle and Manganelli (2004) and standard quantile regression (Koenker & Hallock, 2001). Both make use of quantile regression (QR) techniques (Koenker & Bassett, 1978) that do not minimize the squared error as in ordinary regression, but use the check function $\rho_\alpha(u) = u(\alpha - 1\{u \leq 0\})$ to minimize $L_\alpha(f_\alpha(\cdot), x_t) = \sum_{t=1}^{T} \rho_\alpha(r_t - f_\alpha(r_t, x_t))$. For CAV, we use a symmetric absolute value (SAV) component for $f_\alpha(\cdot)$, i.e. $f_\alpha(r_t, x_t) = \beta_1 + \beta_2 f_\alpha(x_{t-1}, r_{t-1}) + \beta_3|r_{t-1}|$, and $f_\alpha(r_t, x_t) = \beta_4 x_{t-1} + \beta_5 r_{t-1}$ for the quantile regression. We furthermore use an asymmetric CAViaR specification with $f_\alpha(r_t, x_t) = \gamma_1 + \gamma_2 f_\alpha(x_{t-1}, r_{t-1}) + \gamma_3 \max(r_{t-1}, 0) - \gamma_4 \min(r_{t-1}, 0)$ (CAV_ASY). In contrast to the former methods, they can only capture parametric (non-) linear effects, which limits their flexibility. For comparison, we use the asymmetric GJR-GARCH(1,1) model (Glosten et al., 1993); a GARCH(1,1) (Bollerslev, 1986) model; a simple historical simulation (Hist), meaning that we predict VaR_{t+1}^α at level α as the sample α-quantile of the preceding returns in a window of length K, i.e. $(r_{t-K+1}, \ldots, r_t)$; and one that fits a normal distribution to the sample data and uses the theoretical fitted α-quantile as the prediction for VaR_{t+1}^α (NormFit). We do not expect the latter to perform well, as we have high skewness and excess kurtosis in the data (see Table 1 in Section 2).

For the proposed random forest-type procedure and all benchmark procedures that allow for additional co-variates (GRF, QRF, and QR), we include lagged standard deviations (SDs) in addition to the lagged level r_{t-1} in the model. We expect these to capture the strongly varying levels of unconditional volatility, in particular for the cryptocurrencies in the non-linear structure. Additionally, we also employ the above methods and the GARCH(1,1) model using both the latter covariates, as well as the seven external covariates described in Section 2. These methods are GRF-X, QRF-X, QR-X, and GARCH-X. To establish a fair common ground in model complexity for QR, QRF, GARCH-X, and GRF, we select a common set of different SD lags as covariates from an additional Monte Carlo study. For this, we use the simple SAV model from Section 4 as a baseline where unconditional volatilities vary over time. The model is essentially linear autoregressive of order 1 in the VaR and thus directly yields the VaR as outputs. With this, it is possible to select the variables which maximize the p-value of the DQ test for the subsequent simulations and data analysis. For a realistic specification of the SAV model, we allow for regime changes. These regime changes are obtained by testing for regime shifts in the realized variances of all three considered currencies in Section 5.2 with a Chow test, and then

using the empirical distribution of all detected change points as points for regime shifts in the model initializing a new draw of σ_t. Table 3 summarizes the results of this short simulation study and motivates the use of three-, seven-, 30-, and 60-day lagged SDs as covariates.

For the highly speculative cryptocurrency market, we focus on one-step-ahead predictions that are practically the most meaningful. Note that an extension to multistep-ahead predictions with the considered random forest-type techniques would not be straightforward. Since the non-linear GRF and QRF are trained for and therefore tailored to one-step-ahead predictions, further forecast horizons would require changing the fitted models for each forecast horizon and retraining them. Otherwise, the required volatility and other inputs would not be available for producing forecasts from the estimated model. Plugging in the last available observation for these quantities might not be valid in the highly non-linear RF structure. For a fair comparison, retraining RFs would also require adapting the parametric benchmark models accordingly.

To compare the performance of the above methods, we use two types of evaluation approaches. First, we test how well each model predicts the conditional α-VaR over the entire out-of-sample horizon using three different sets of evaluation techniques. The simplest way of checking whether a model predicts VaR^α correctly over a time horizon is to look at its coverage, meaning the number of times r_t is smaller than the predicted VaR_t^α. Ideally, this should be exactly αT times. This measure is called the actual over expected exceedance ratio (AoE) and is computed as $AoE_\alpha = (\alpha T)^{-1} \sum_{t=1}^{T} 1\{r_t < VaR_t^\alpha\}$. To test this intuition formally, we employ three tests: the DQ test[4] from Engle and Manganelli (2004), the Christoffersen test (Christoffersen, 1998), and the Kupiec test (Kupiec, 1995). All three tests assume that the forecasts have correct coverage under the null hypothesis. The Kupiec test is simply the formalization of the above intuition, the Christoffersen test is robust against serial correlation by assuming that $g_t = 1\{r_t < VaR_t^\alpha\} \sim Bern(\alpha)$, and the DQ test additionally accounts for problems with conditional coverage due to clustering of the hits exceedance sequences g_t with a regression-based approach. In the empirical analysis, we only report the values of the DQ test, which is the strictest of the tests, for reasons of clarity. The results for the other tests did not differ substantially and are available upon request from the authors.

Secondly, for comparing the forecast performance of two models 1 and 2 directly, we implement the one-step-ahead test for conditional predictive ability (CPA) from Giacomini and White (2006), Theorem 1, that assumes under the null hypothesis that the forecasts of model 1 and model 2 have on average equal predictive ability conditional on previous information. As suggested by Giacomini and Komunjer (2005), we use the quantile loss function L_α for the test. This tests assumes

---

[4] We use the implementation from the GAS package in R (Ardia, Boudt, & Catania, 2019) including four lagged hit values, a constant, the VaR forecast, and the squared lagged (log) return in the regression model.

Table 3
DQ test p-values for different covariate combinations.

Lagged SD (in days)	3	7	30	3 and 7	3 and 30	7 and 30	3, 7, and 30	3, 7, 30, and 60
DQ test *p*-value	0.205	0.160	0.174	0.249	0.254	0.244	0.298	**0.306**

DQ test p-values for a simulated SAV model, as in Section 4, for GRF. The depicted values are averages over 100 iterations. The maximum *p*-value is marked in bold.

under the null hypothesis that $H_0 : E[\Delta L_t | \mathcal{F}_{t-1}] = E\left[L_\alpha(f_\alpha^{(1)}, r_t) - L_\alpha(f_\alpha^{(2)}, r_t) | \mathcal{F}_{t-1}\right] \equiv E[h_{t-1}\Delta L_t] = 0$ and that this loss difference is a martingale difference sequence, where $\mathcal{F}_{t-1}$ contains all information up to time $t-1$, and $f_\alpha^{(1)}$ and $f_\alpha^{(2)}$ are two competing forecasts. The test statistic is computed using a Wald-type test with a set of factors h_{t-1} that can possibly predict the loss difference ΔL_t and is χ_q^2-distributed under H_0. More specifically, we choose $h_{t-1} = (1, \Delta L_{t-1})$ (i.e. $q = 2$), that is, using the lagged loss difference and an intercept as predictors in a linear regression with parameter β_0 for the simulation and application.

4. Simulation

In this section, we study the finite-sample forecast performance in quantiles for true data generating processes of standard stock-type dynamics as well as for setups of cryptocurrency types. Overall, we find that, as expected, GRF performs well even in simple settings where the oracle parametric models should actually have an advantage. For cases which are designed to mimic the cryptocurrency behavior over time with changing volatilities, GRF outperforms the competitor models. For QRF, the performance is entirely different, as it is outperformed by most other models. These results are also strongly confirmed by pairwise CPA tests.

We study four different types of data generating processes. The first one is a standard GARCH(1,1) process:

$$r_t = z_t \sigma_t \tag{3}$$

$$\sigma_t^2 = \omega + \beta_0 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{4}$$

where the parameters are estimated on the full Bitcoin data to mimic the behavior of established cryptocurrencies (with z_t follows a normal distribution or an asymmetric, skewed-t distribution). We denote these settings as *GARCH Bitcoin Normal* and *GARCH Bitcoin Asym-t*. Moreover, we also consider the specification with $\beta_0 = 0.1$, $\beta_1 = 0.8$, $\omega = 10^{-4}$ and $z_t \sim N(0, 1)$ (*GARCH*) and with z_t as t_5-distributed (*GARCH-t*), which corresponds to standard stock index data as a simple benchmark setting.

Thirdly, for the *SAV Model* setting, we fit a symmetric absolute value (SAV) model to normal returns, i.e.

$$VaR_{t+1} = \gamma_0 + \gamma_1 VaR_t + \gamma_2 |r_t^{(init)} - \gamma_3|, \tag{5}$$

with $r_t^{(init)} \sim N(0, \sigma_t^2)$ and $\frac{\sigma_t}{65} \sim \chi_2^2$, where new draws of σ_t are only taken every 100 observations, keeping σ_t constant meanwhile. We then generate the final return as $r_t \sim N\left(0, \frac{\widehat{VaR_t}}{\Phi(\alpha)^{-1}}\right)$ from the fitted SAV model, where $\Phi(\alpha)^{-1}$ is the quantile function of a standard normal variable. We do this to obtain returns that have exactly the VaR that we obtained from the SAV model before.

Lastly, we use a simple stochastic volatility model that draws the parameters in the conditional volatility specification from a normal innovation and uses pre-estimates of the two lag-one parameters obtained from an average of the mostly emerging cryptocurrencies contained in Group_high of Table 9. We denote this setting as *Varying Vola GARCH*.

For all settings, we generate 2000 return observations and forecast the one-step-ahead VaR over the different rolling window lengths $l = 500, 1000$. We repeat this generation process 200 times $\alpha = 0.05$. For comparison of the different methods described in Section 3, we use the DQ test, the Kupiec test, the Christoffersen test, and the AoE. Note that for all tests, we present aggregate results from two-sided t-tests of the empirical versus the nominal coverage. The results are therefore rejection rates of t-tests against the nominal level of 5%. Therefore, a lower rejection rate and higher mean p-values (in parentheses) indicate better performance. For GRF, QRF, and QR, we use a common set of lagged covariates, as described at the end of Section 3.

Table 4 summarizes the results of the simulation for the 5% *VaR*. According to the more advanced DQ and Christoffersen tests for evaluation, GRF is close to on par with models that directly use the simulated model specification, and outperforms all other methods in cases where the underlying model deviates from standard GARCH-type data generating processes as is the case for cryptocurrencies (*Varying Vola GARCH*). This is in sharp contrast to the QRF model, which cannot compete with the other models in all cases. We only display results for the *GARCH Normal* and *GARCH Bitcoin Asym-t* cases, as *GARCH-t* and *GARCH Bitcoin Normal* yield similar results. A bit of a challenge for GRF is the very simplistic *SAV Model*, where the simple models dominate but the GRF performance still improves on advanced parametric models such as CAV, CAV_asy, and GJR-GARCH. QR is dominated completely in both the SAV and standard GARCH-type settings. In the GARCH Bitcoin models, the CAV procedures perform particularly well, in both the small and large sample cases, while the asymmetric specification cannot outperform its symmetric counterpart in most of the cases, specifically for the small window length. This is not surprising, as the asymmetric specification is more complex and therefore needs more data to perform well. In the third time-varying setting *Varying Vola GARCH*, GRF substantially outperforms all other methods when calibrated in the short time interval – but loses this advantage for the larger time period in particular compared to CAV and CAV_asy. These results suggest that, depending on the underlying data generating process, either GRF or CAV procedures perform best.

Additionally, we conduct direct pairwise comparison tests between the superior random forest-type method

Table 4
Simulation: 5% VaR.

Rolling window	l = 500				l = 1000			
	DQ	Kupiec	Christoffersen	AoE	DQ	Kupiec	Christoffersen	AoE
GARCH Normal								
QRF	0.960 (0.006)	0.330 (0.196)	0.210 (0.270)	1.184	0.710 (0.068)	0.190 (0.336)	0.110 (0.349)	1.159
GRF	0.470 (0.184)	0.000 (0.568)	0.000 (0.544)	1.043	0.270 (0.316)	0.030 (0.537)	0.050 (0.491)	1.031
QR	0.730 (0.047)	0.080 (0.451)	0.100 (0.463)	1.094	0.310 (0.261)	0.060 (0.555)	0.020 (0.518)	1.038
Hist	0.860 (0.048)	0.020 (0.539)	0.190 (0.361)	1.045	0.630 (0.114)	0.080 (0.478)	0.220 (0.358)	1.031
NormFit	0.760 (0.069)	0.080 (0.566)	0.190 (0.322)	1.012	0.590 (0.124)	0.100 (0.449)	0.210 (0.334)	1.009
CAV	0.610 (0.134)	0.010 (0.557)	0.050 (0.541)	1.044	0.200 (0.317)	0.050 (0.492)	0.040 (0.457)	1.034
CAV_ASY	0.760 (0.049)	0.010 (0.484)	0.030 (0.496)	1.076	0.270 (0.249)	0.090 (0.475)	0.070 (0.464)	1.047
GARCH(1,1)	0.470 (0.182)	0.040 (0.521)	0.190 (0.385)	1.057	0.250 (0.320)	0.080 (0.517)	0.060 (0.433)	1.029
GJR-GARCH	0.500 (0.151)	0.040 (0.497)	0.140 (0.383)	1.070	0.280 (0.288)	0.070 (0.506)	0.080 (0.417)	1.040
GARCH Bitcoin Asym-t								
QRF	0.640 (0.086)	0.300 (0.236)	0.170 (0.317)	1.174	0.350 (0.243)	0.140 (0.386)	0.110 (0.390)	1.150
GRF	0.350 (0.303)	0.030 (0.609)	0.060 (0.559)	1.005	0.170 (0.443)	0.050 (0.518)	0.030 (0.490)	1.004
QR	0.470 (0.209)	0.080 (0.427)	0.080 (0.397)	1.102	0.200 (0.419)	0.060 (0.489)	0.090 (0.475)	1.064
Hist	0.990 (0.003)	0.160 (0.425)	0.690 (0.111)	1.081	0.920 (0.021)	0.290 (0.302)	0.590 (0.134)	1.074
NormFit	0.960 (0.008)	0.500 (0.166)	0.780 (0.054)	0.788	0.900 (0.024)	0.550 (0.168)	0.760 (0.056)	0.731
CAV	0.270 (0.305)	0.040 (0.517)	0.030 (0.508)	1.043	0.110 (0.507)	0.070 (0.498)	0.080 (0.454)	1.028
CAV_ASY	0.350 (0.263)	0.080 (0.500)	0.050 (0.495)	1.080	0.160 (0.459)	0.110 (0.500)	0.060 (0.453)	1.047
GARCH(1,1)	0.420 (0.256)	0.220 (0.314)	0.350 (0.219)	0.859	0.270 (0.313)	0.240 (0.315)	0.250 (0.246)	0.826
GJR-GARCH	0.430 (0.247)	0.160 (0.352)	0.310 (0.253)	0.890	0.250 (0.292)	0.220 (0.324)	0.250 (0.252)	0.836
SAV Model								
QRF	0.970 (0.005)	0.300 (0.192)	0.160 (0.270)	1.186	0.710 (0.074)	0.170 (0.328)	0.130 (0.378)	1.175
GRF	0.470 (0.157)	0.010 (0.568)	0.020 (0.589)	1.047	0.230 (0.321)	0.040 (0.511)	0.050 (0.533)	1.061
QR	0.930 (0.014)	0.050 (0.431)	0.050 (0.455)	1.102	0.360 (0.228)	0.050 (0.485)	0.060 (0.516)	1.076
Hist	0.320 (0.227)	0.030 (0.564)	0.070 (0.513)	1.052	0.190 (0.368)	0.060 (0.515)	0.070 (0.480)	1.070
NormFit	0.280 (0.344)	0.030 (0.546)	0.110 (0.501)	1.004	0.150 (0.435)	0.060 (0.484)	0.070 (0.451)	1.022
CAV	0.540 (0.086)	0.010 (0.539)	0.020 (0.534)	1.059	0.160 (0.295)	0.030 (0.547)	0.030 (0.503)	1.053
CAV_ASY	0.750 (0.046)	0.020 (0.481)	0.020 (0.535)	1.081	0.250 (0.275)	0.050 (0.494)	0.070 (0.501)	1.064
GARCH(1,1)	0.210 (0.313)	0.050 (0.582)	0.050 (0.523)	1.035	0.080 (0.448)	0.030 (0.513)	0.030 (0.480)	1.043
GJR-GARCH	0.350 (0.214)	0.060 (0.540)	0.060 (0.479)	1.048	0.060 (0.442)	0.020 (0.547)	0.030 (0.484)	1.050
Varying Vola GARCH								
QRF	0.524 (0.124)	0.270 (0.215)	0.111 (0.297)	1.175	0.410 (0.218)	0.190 (0.335)	0.190 (0.370)	1.166
GRF	0.381 (0.269)	0.000 (0.509)	0.079 (0.393)	1.016	0.240 (0.323)	0.110 (0.436)	0.160 (0.373)	0.959
QR	0.381 (0.195)	0.000 (0.538)	0.143 (0.474)	1.013	0.290 (0.296)	0.060 (0.440)	0.110 (0.381)	1.065
Hist	0.984 (0.004)	0.476 (0.218)	0.810 (0.069)	1.281	0.980 (0.005)	0.620 (0.163)	0.910 (0.040)	1.140
NormFit	0.984 (0.003)	0.476 (0.212)	0.810 (0.053)	1.212	0.980 (0.003)	0.580 (0.162)	0.850 (0.041)	0.991
CAV	0.270 (0.315)	0.079 (0.467)	0.048 (0.472)	1.051	0.110 (0.427)	0.120 (0.406)	0.140 (0.407)	1.012
CAV_ASY	0.397 (0.212)	0.127 (0.436)	0.016 (0.449)	1.084	0.110 (0.420)	0.140 (0.449)	0.150 (0.427)	1.026
GARCH(1,1)	0.302 (0.279)	0.048 (0.458)	0.254 (0.310)	1.060	0.480 (0.205)	0.140 (0.397)	0.340 (0.237)	1.084
GJR-GARCH	0.317 (0.309)	0.095 (0.459)	0.254 (0.348)	1.074	0.500 (0.192)	0.170 (0.419)	0.380 (0.235)	1.093

The table displays rejection rates of t-tests of empirical quantile levels against the nominal level of 5% for DQ, Kupiec, and Christoffersen tests and mean p-values in parentheses. Thus, lower rejection rates and higher p-values indicate better model performance. The $AoE_\alpha = (\alpha T)^{-1} \sum_{t=1}^{T} 1\{r_t < VaR_t^\alpha\}$ should be compared to its optimal value 1.

GRF against the best-performing non-oracle parametric methods (and a simple baseline) via CPA tests for each scenario. The results are reported in Table 5. For the standard GARCH processes, GRF outperforms its competitors on average. However, mean p-values are mostly not significant. As before, we only display the results for the *GARCH Normal* case, since *GARCH-t* and *GARCH Bitcoin Normal* yield similar results. In the other three scenarios, GRF is substantially superior with more significant p-values. In the *GARCH Bitcoin Asym-t* and *SAV Model* cases, the variance in the p-values over different simulation runs is quite high, marked by many significant test results, but with a performance that indicates that both procedures are equally good. For *Varying Vola GARCH*, the performance gains of GRF are substantial, reaching highly significant test results. Overall, the results indicate that even for complex data structures, the power of CPA tests

in small samples appears sufficient if compared to the larger-sample results.

5. Results

In this section, we highlight the advantages from using specific non-linear machine learning-based methods for forecasting the VaR of cryptocurrencies. In particular, we show for a large cross-section of more than 100 cryptocurrencies that the proposed GRF method yields superior performance across a wide range of different types of cryptocurrencies and different time periods. Investigating the underlying drivers, we illustrate that the non-linear model predictions excel especially for assets that are frequently traded by a large number of different users, and for more volatile assets and times.

We predict the 5% VaR as a key quantity in risk management for our comprehensive set of cryptocurrencies.

Table 5

CPA tests on predictions of 5% VaR for different window lengths and different models.

Rolling window	$l = 500$			$l = 1000$		
GRF vs.:	QR	Hist	CAV	QR	Hist	CAV
GARCH Normal						
Mean *p*-value	0.235	0.347	0.383	0.468	0.39	0.504
No. p-values < 0.1	52	23	14	14	23	11
GRF performance	0.877	0.794	0.613	0.539	0.799	0.482
GARCH Bitcoin Asym-t						
Mean *p*-value	0.163	0.06	0.412	0.155	0.098	0.387
No. p-values < 0.1	71	85	12	70	76	10
GRF performance	0.799	0.971	0.598	0.748	0.962	0.356
SAV Model						
Mean *p*-value	0.149	0.294	0.494	0.123	0.337	0.445
No. p-values < 0.1	66	31	12	77	29	13
GRF performance	0.964	0.296	0.567	0.914	0.335	0.495
Varying Vola GARCH						
Mean *p*-value	0.001	0.005	0.464	0.4	0.016	0.363
No. p-values < 0.1	63	62	8	19	97	23
GRF performance	0.987	0.989	0.569	0.363	0.985	0.329

The table displays CPA tests for 5% one-day-ahead VaR forecasts of the best-performing random forest-type techniques in Table 4 versus the best parametric time series models. We report mean p-values, the number of significant p-values over 200 iterations, and the rate at which GRF outperforms the competing method (e.g. a value of 0.8 means that GRF has a smaller error loss than the competing method in 80% of the rolling window forecasts over all runs). Low p-values paired with performance rates larger than 0.5 indicate that GRF outperforms the competing methods.

In an extensive out-of-sample forecasting study, we compare the random forest-based machine learning methods to standard linear time series and GARCH-type models, including approaches with exogenous asset information in covariates. The prediction performance is assessed with the DQ test to obtain an overall aggregate picture on the realized coverages, as well as pairwise CPA tests across different time periods and types of cryptocurrencies.

Based on these findings, we extend our analysis, focusing on three important selected currencies: Bitcoin (btc), the largest currency by far in terms of market cap; Tether (usdt_omni), a stablecoin with lower volatility; and Cardano (ada), a currency specifically allowing for smart contracts. For these examples, we also consider predicted loss series by CPA tests and variable importance measures to uncover important drivers. Furthermore, we extend the aggregate forecast performance study of Periods 1–3 to the shorter post-pandemic Period 4.

5.1. Aggregated forecasting performance

In this section, we provide results on aggregate forecast performance of the different modeling approaches over all cryptocurrencies.

5.1.1. Backtesting

We analyze the three fundamentally distinct periods 2015/08/22–2017/12/21, 2017/12/22–2020/11/05, and 2020/11/06–2022/12/31 of the cryptocurrency market,

which differ according to the number of actively traded currencies, the market capitalization, and the overall market situation (see Section 2). We thus study each period separately, with a focus on Period 3 in the aggregate. Given the largely different market situations, we refrain from reporting results for the full time period. For the fundamentally different post-Covid-19/Ukrainian war situation, we provide results up until today in Section 5.2.2.

We assess the performance of VaR forecasts by their conditional coverage and the dependence structure of exceedances, which we test with the DQ test (see Section 3 for details). This test is the strictest of all backtests, conditioning on past information, and therefore most accurately captures the coverage abilities of the procedures.[5] The results over the different time periods for the 5% VaR predictions are shown in Table 6. We can see that depending on the time period, the median p-values vary strongly, which is not surprising giving the different characteristics of each period and the increasing number of cryptocurrencies in the later periods. In general, GRF is the only method with median values consistently over the 8% level, indicating that it is the most consistently calibrated forecasting method. QRF performs extremely well in Period 1, but rejects the test often in Period 3. Adding external covariates as listed in Table 2 does not improve performance and in general leads to substantially lower p-values, especially for the QR. Only in Period 3, adding external covariates increases the p-values slightly for GRF and substantially for QRF (GRF-X, QRF-X). This indicates that those extra covariates are not necessarily predictive for extreme returns, or rather that the existing measures, such as lagged returns and SD, already comprise the information quite well.

Compared to its non-forest counterparts, only GJR-GARCH can partly keep up with GRF. In particular, the GJR-GARCH has higher median p-values than GRF in Period 3, which is the endemic Covid-19 period and contains the most currencies. It is also marked by less extreme returns and a large reduction in active users (see Table 2), which could indicate that the forest methods excel particularly in highly volatile periods when large shifts in the market are present. In all periods, CAV and the asymmetric CAV are outperformed by GRF in DQ test results. This highlights the inability of the parametric methods to adapt to rapidly changing situations such as in Periods 1 and 2. QR, Hist, and GARCH-X are all fully dominated by GRF throughout the three time periods as expected, as QR can only incorporate changes linearly, and GARCH-X and Hist serve as simple baselines. Additionally, when checking the unconditional (median) coverage represented as the ratio of exceedances relative to the expected ones, random forest-type methods also show superior performance over standard GARCH-type methods, which have quite low coverage in all periods (GARCH-X) or Period 1 (GJR-GARCH).[6] With an LM-type test (see Silvennoinen and Teräsvirta (2016)) we formally confirm time variation of the unconditional volatility that is not captured

---

[5] Detailed results for the other tests are omitted here for reasons of clarity. As with the simulation results, they did not differ qualitatively, and are available upon request from the authors.

[6] Detailed results are available on request.

Table 6

Medians of p-values for DQ tests in different time periods.

Period	GRF	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	GRF-X	QRF-X	QR-X	GARCH-X
1	0.14	0.24	0.04	0.05	0.06	0.03	0.00	0.01	0.00	0.00	0.00
2	0.29	0.04	0.14	0.13	0.31	0.19	0.11	0.16	0.07	0.00	0.00
3	0.08	0.02	0.05	0.01	0.01	0.16	0.00	0.12	0.05	0.00	0.00

The table displays medians of p-values for DQ tests in different time periods. Higher median p-values indicate correct conditional coverage according to the DQ test. Periods correspond to the times indicated in Fig. 1.
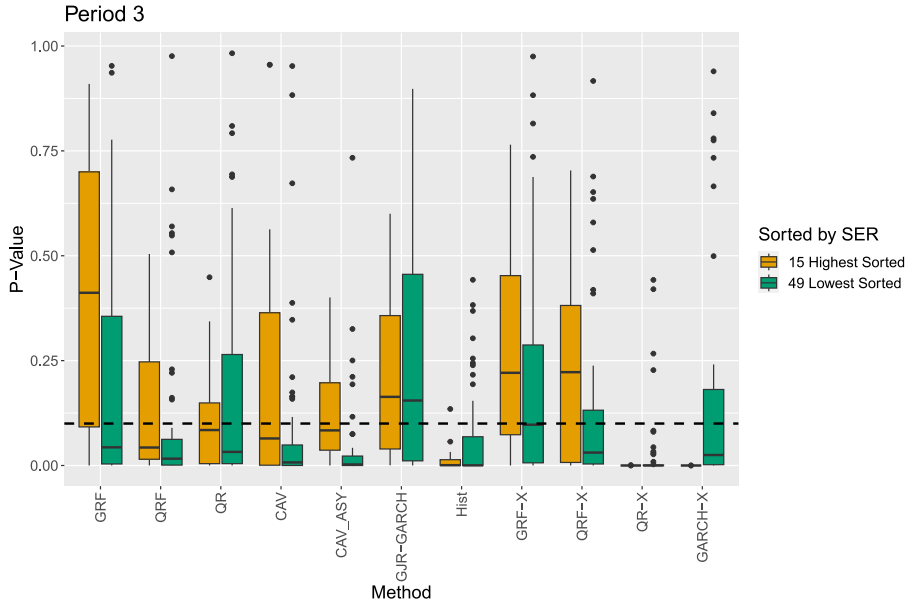

Fig. 2. Boxplots of p-values of DQ tests over all cryptocurrencies contained in Period 3 separated into groups with high vs. low SER for every method. For details on the groups, see Table 7. The dashed horizontal line depicts a level of 0.1.

by standard static GARCH-type specifications but leads to advantages of the forest-type methods. Across most of the cryptocurrencies, we reject the null hypothesis of constant unconditional variance at a significance level of 5% and at even lower levels for the robust version of the test, in particular for Periods 1 and 2.

We further investigate Period 3, which contains 64 cryptocurrencies, especially with regard to how conditional coverage is characterized by external information. For this, we split the currencies into two groups[7] depending on the value of the SER, which characterizes the concentration of supply to users. A low SER consequently implies that most supply is concentrated on a few, large users, often indicating a more stable currency. Fig. 2 depicts boxplots over p-values of DQ tests for each group and procedure in Period 3. In Period 3, GRF(-X) performs particularly well for the groups with a high SER, which are arguably more prone to speculations during hype/bubble periods, due to the large amount of supply held by many small addresses. For the (arguably) more stable addresses with a smaller SER, GJR-GARCH

has higher p-values, followed by GRF-X, during Period 3. This is in contrast to CAV(_ASY), QR, and QRF, which are always outperformed by GRF(-X), as well as the simple baselines Hist and GARCH-X. In Period 2, the general market situation is so volatile that GRF is superior in DQ performance for the low-SER group where in particular the additional regressors in the GRF-X make a difference. For details on all time periods and groups according to the different observable factors, active users, and lagged standard deviations, see Table B.16 in the appendix.

5.1.2. Direct pairwise forecast comparisons

In addition to assessing the coverage performance, we conduct pairwise CPA tests for all considered methods in the three different specified time periods over all cryptocurrencies. The results of the tests are contained in Table 8. Note that the CPA tests require the rolling window length to be smaller than the out-of-sample forecast window to produce valid results. By construction, for the large share of newly introduced cryptocurrencies in Period 3, the out-of-sample size is too low for the CPA tests to have high power. Therefore, we additionally look at direct comparisons of predicted losses (as suggested by Giacomini and White (2006), Section 4), where we compare in the loss series how often GRF is better, i.e. has

---

[7] Split points correspond to obtaining two groups with mean SER values as homogeneous as possible. They are therefore separated around their steepest decay. The high group always comprises the largest SER-sorted quintile and not more than the largest sextile in each period.

Table 7

Medians of p-values for DQ tests sorted by SER in different time periods.

Group	GRF	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	GRF-X	QRF-X	QR-X	GARCH-X
Period 1											
High	0.17	0.81	0.00	0.00	0.00	0.12	0.00	0.01	0.00	0.00	0.00
Low	0.11	0.20	0.08	0.09	0.10	0.02	0.00	0.02	0.01	0.00	0.00
Period 2											
High	0.15	0.03	0.20	0.05	0.41	0.10	0.09	0.13	0.08	0.00	0.00
Low	0.31	0.04	0.08	0.17	0.29	0.27	0.17	0.45	0.07	0.00	0.11
Period 3											
High	0.41	0.04	0.08	0.06	0.08	0.16	0.00	0.22	0.22	0.00	0.00
Low	0.04	0.02	0.03	0.01	0.00	0.15	0.00	0.10	0.03	0.00	0.02

The currencies are divided into two groups based on specific covariate values for each time period. The groups are constructed to best separate the sorted covariate values of the currencies. They are therefore separated at their steepest decay to obtain two groups with homogeneous covariate values, where the top group in each period contains at least the top quintile but not more than the top sextile of the data. The remaining details correspond to Table 6.

a smaller loss, than its competitors.[8] Note that a value of one thus indicates that GRF has a smaller predicted loss over the full loss series.

In general, GRF performs better than its competitors for a majority of cryptocurrencies over all time periods. Table 8 summarizes the results. We can see that QRF is almost always outperformed, and for around 50% of cryptocurrencies, losses are even significantly smaller. This is not surprising, as QRF has a similar structure to GRF while not being tuned to predict the quantiles directly. Thus, we expect it to be less sensitive to changes that only affect the quantile of the return distribution, for example large shock events. The same holds for GARCH-X and Hist, which are clearly outperformed by GRF, as well as QR-X and QRF-X. Adding exogenous information in covariates as part of the non-linear GRF (i.e. GRF-X) is better, especially in later periods (see bottom part of Table B.17 in the appendix). This is interesting because the other methods cannot benefit as much as GRF from additional covariates. Generally, for cryptocurrencies, the non-parametric form of GRF helps to extract information from exogenous covariates, in contrast to standard parametric methods such as QR and GARCH. As GRF accounts specifically for the quantiles in the random forest splitting function, this also helps to favorably integrate additional covariates X, in contrast to QRF. Overall, however, both GRF-procedures seem to perform very similarly, especially in pairwise comparisons. For the full results of GRF-X, see Table B.17 in the appendix. While CAV(_ASY), QR, and GJR-GARCH are outperformed over the majority of cryptocurrencies, only 14% to 24% of these out-performances reach significance. In Section 5.2, we focus on specific cryptocurrencies for a more in-depth understanding.

When considering the single time periods, it is notable how for Periods 1 and 2 (the bottom part of Table 8), GRF(-X) is constantly outperforming the other methods for most cryptocurrencies, and only has somewhat worse performance for doge and the stablecoins, although these are non-significant (see below). For Periods 1 and 2, QR is the most competitive of the other methods, while there is a general tendency for the classic methods to perform

worse with higher volatility of returns, which can be seen in Table 8 for Period 2 where the currencies are ordered from highest 30-day lagged SDs on the left to the lowest on the right. This becomes more apparent in Period 3, where we deal with many more cryptocurrencies (64). Here, methods such as GJR-GARCH and QR are on par with GRF, or even better when looking at low-volatility (partly regulated) stablecoins such as pax, gusd, tusd, and usdt_omni with its derivatives (usdt_eth and usdt_trx). CAV and CAV_asy, on the other hand, are only rarely better here (as indicated by CPA tests), while being significantly outperformed for important and large assets such as btc, ada, xlm, and most stablecoins. The asymmetric version of CAV performs similarly or worse in comparison to CAV most of the time, which is why we only focus on the symmetric version in the following sub-analysis for ease of presentation.

To highlight the specific properties of currencies where GRF outperforms the other methods, we split the assets into two groups. The first group (Group_low) contains assets where GRF performance is low in comparison to the three other methods that were able to compete in some cases with GRF, namely QR, CAV, and GJR-GARCH. We add an asset to that group when at least two of the methods outperform GRF (in terms of the loss difference) for that asset, separately for each time period. All other assets are sorted into the second group (Group_high), indicating high performance of GRF. Note that 40% of currencies belong to Group_high in Period 1, 36% in Period 2, and 33% in Period 3. The corresponding mean daily trading volumes in USD in Group_high are $34,961,093 in Period 1, $419,414,899 in Period 2, and $1,629,301,034 in Period 3.

Table 9 summarizes the results over these groups for each time period and covariate. In each group, we take the mean over all cryptocurrencies of median values for each covariate. We then divide Group_low by Group_high. For example, an SER of 0.52 in Period 3 indicates that cryptocurrencies in Group_low have, on average, a median SER that is 47% to that of Group_high; in other words, the median SER for Group_high is around $2.13 = \frac{1}{0.47}$ times higher than that of Group_low on average. We see that covariates of cryptocurrencies for which GRF performs better have much higher volatility (especially for Periods 2 and 3), a much higher SER,[9] indicating a

---

[8] We use the lagged loss difference and an intercept for loss prediction in an autoregressive setup, since these are the main drivers of the test statistic in the CPA test.

[9] Apart from Period 1, where the only currency belonging to Group_low is doge.

Table 8

Performance and significance of CPA tests over different time periods for GRF.

GRF vs.:	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	GRF-X	QRF-X	QR-X	GARCH-X
Share of GRF with better performance										
Period 1	1.00	0.80	0.60	0.80	1.00	1.00	0.60	1.00	1.00	1.00
Period 2	1.00	0.71	0.79	0.93	0.86	1.00	0.36	1.00	1.00	1.00
Period 3	0.94	0.75	0.75	0.80	0.70	0.97	0.39	0.97	0.98	0.80
Full.Data	0.91	0.70	0.70	0.73	0.58	0.94	0.33	0.92	0.98	0.75
Share of GRF with significantly better performance										
Period 1	0.60	0.40	0.00	0.20	0.00	1.00	0.00	1.00	1.00	0.60
Period 2	0.79	0.29	0.21	0.50	0.29	0.79	0.00	0.93	0.93	0.71
Period 3	0.61	0.20	0.19	0.28	0.14	0.78	0.06	0.77	0.84	0.55
Full Data	0.50	0.14	0.19	0.24	0.12	0.71	0.05	0.66	0.74	0.53
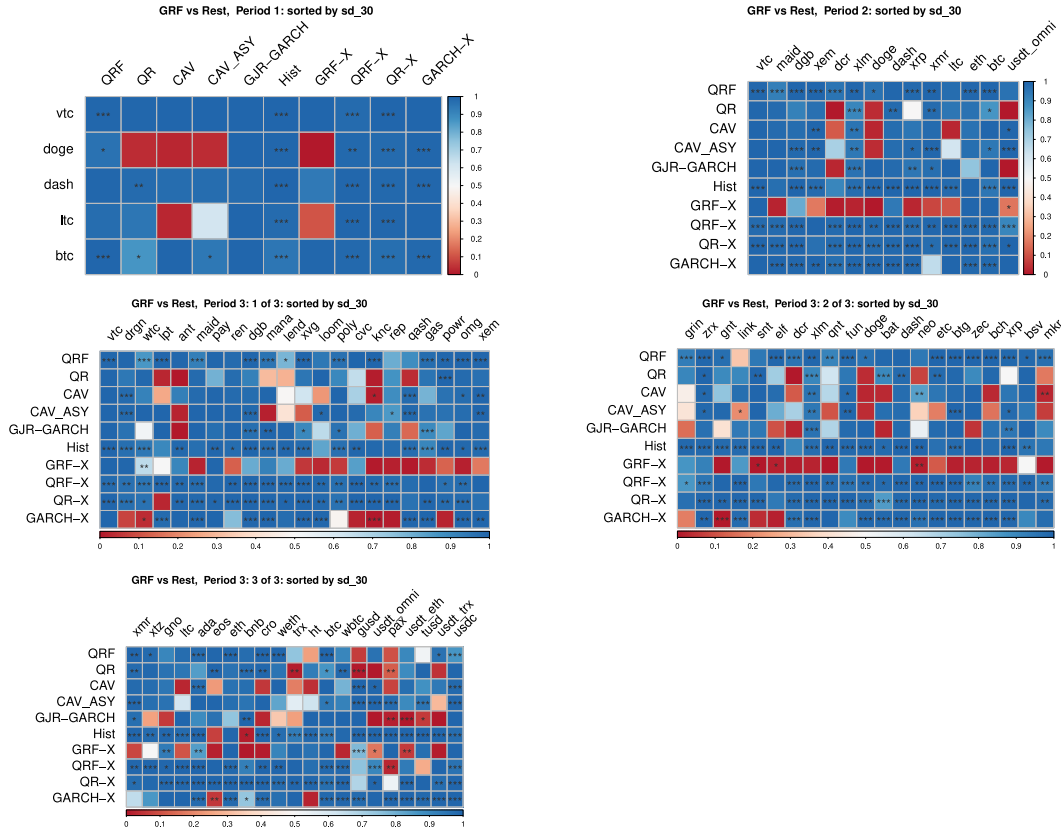
The top part shows summary values that are shares over all cryptocurrencies in the respective time period. It describes the number of times that GRF had a better performance (i.e. more than 50% of predicted losses by the CPA test were smaller for GRF) relative to all cryptocurrencies (in that period), and the number of times that GRF was significantly better (at least at a 10% level) as judged by the CPA test over all cryptocurrencies (in that period). The bottom part shows the detailed results of CPA tests, with the color of each box indicating the performance of GRF. Blue signifies a performance of 1, meaning that GRF has a smaller predicted loss in 100% of cases. *, **, and *** shows significance at levels of 10%, 5%, and 1%, respectively. The values are ordered by 30-day lagged standard deviations from highest to lowest (top-left to bottom-right).

larger concentration of supply at a lot of small addresses, a higher market capitalization, a lower rate of turnover (Velocity), and more active and total users. To summarize, this confirms the observation that GRF performs better for assets with highly varying returns that are traded by a large number of users, which could thus also be prone to speculation. On the other hand, methods such as QR or GJR-GARCH are better with more stable currencies that are used more as a hedging device (e.g. stablecoins). This confirms our findings from the backtests in Section 5.1.1,

where GRF excels in particular for currencies with high SER values, high volatility, and a large number of active users.

5.2. Extensions

5.2.1. In-depth analysis of specific classes of assets

To identify the dynamics of the obtained results and the effect of specific events, we analyze the predicted

Table 9
Difference between covariates of cryptos where GRF is better vs. worse.

	Period 1	Period 2	Period 3	Full data
Ret	0.83	0.96	4.24	1.83
Active_Users	0.30	0.23	3.00	2.40
Total_Users	0.48	0.18	0.32	0.27
Total_Users_USD100	0.12	0.13	0.50	0.50
Total_Users_USD_10	0.18	0.14	0.79	0.71
SER	0.74	0.49	0.47	0.48
Transactions	0.38	0.04	5.78	3.79
Velocity	4.11	3.46	0.93	0.76
sd_3	0.91	0.64	0.88	0.97
sd_7	0.92	0.64	0.88	0.95
sd_30	0.93	0.66	0.87	0.95
sd_60	0.92	0.63	0.88	0.96

The table shows shares of groups of cryptocurrencies where at least two of CAV, QR, and GJR-GARCH have better CPA performance than GRF divided by the rest. Raw values before division are mean values over all cryptocurrencies for the median of each covariate in the respective time period. A value smaller than 1 indicates that currencies where GRF performs better have a higher average median values of the respective covariate than the rest of the currencies.

loss series of CPA tests over the full horizon of availability for the three cryptocurrencies, Bitcoin, Cardano, and Tether, separately. We furthermore show how the impact of covariates on forecast performance changes dynamically over time using variable importance measures of GRF-X.

We choose Bitcoin because it is the largest currency by market cap, with the longest data availability; Tether because it is the largest stablecoin by daily volume and market cap; and Cardano as a fairly new (i.e. fewer observations) yet large currency (again by market cap), which can be used for smart contracts, identity verification, or supply chain tracking.[10] Since we deal with VaR predictions, the initial loss function is the quantile loss with a quantile $\alpha = 0.05$.

First, we look at Bitcoin (btc), the largest and most popular currency, where GRF largely outperforms QR, CAV, and GJR-GARCH in the CPA tests. Fig. 3 shows the predicted loss difference for each of the different methods. GRF outperforms the other methods consistently for most time frames. Secondly, as summarized on the left in Fig. 4, we look at Cardano (ada), a large and fairly new currency offering e.g. smart contracts or supply chain tracking. There, GRF significantly outperforms GJR-GARCH and CAV, while being slightly better than QR although not reaching a significant level. Finally, we also look more closely at Tether (usdt_omni) as the largest stablecoin that is roughly bound to the USD.[11] The right part of Fig. 4 shows the 30-day rolling means of the predicted loss differences. Notably, the rolling loss difference and rolling mean return are already around 10 times smaller than those of btc and ada, indicating that usdt_omni substantially differs from the other two currencies. While the losses of QR are very similar to those of GRF, CAV is significantly outperformed. Only GJR-GARCH has consistently lower predicted losses than GRF, although they are deemed not significant by the CPA tests.

To further understand the drivers of the GRF performance, we obtain variable importance measures that depict the frequency of inclusion in splits of the forest. The variable importance of a covariate x_p is measured as the proportion of splits on x_p relative to all splits in a respective layer l (over all trees in a trained forest), weighted by layer l.[12] In Fig. 5, the importance difference of certain covariates over time for the three currencies is clearly visible. Overall, the lagged return is very important for predicting VaR when returns are quite extreme relative to all returns in a specific asset, in the case of btc in times of hypes and crashes. Intuitively, this finding seems reasonable because in times of bubbles, when the volatility is driven by some short, bubble-like events and returns are highly variable, volatility lagged over a longer time horizon is less predictive for VaR and predictions are driven by events happening shortly before the prediction. In rather unstable times, but not in extreme cases, the lagged SD measures gain importance, while the extra covariates only play a role for assets with relatively small volumes, e.g. when new currencies are created. This also explains why GRF-X performs much better for new, low-market-cap assets in Period 3 (see e.g. Fig. B.6 in the appendix).

Starting with ada, we see that it is the only asset of the three where the number of active addresses plays an important role, where for the other two assets, the 60-day lagged standard deviation is more important. The importance of variables can be split into two periods. The period until the beginning of 2021 is largely dominated by measures that somehow account for trading activity (Active_Users, Total_Users/_USD100/_USD10, Transactions). For reasons of clarity, we only plot the most important one of these measures, Active_Users. The spike of the latter in importance at the beginning of 2021 is likely caused by the massive increase in price and market cap during that time, representing a period of hype with many actively trading users.[13] For the rest of the time period, lagged SDs, mostly three-day lagged SDs, followed by 30-day and 60-day SDs, dominate the predictions of GRF. This change of importance seems reasonable, as the structure of the asset fundamentally changes, with the price increasing tenfold and the volume increasing strongly at the same time.

For btc, we have much more data, covering 10 years, which is why the important variables change frequently in different periods. The lagged return is naturally important in phases of extreme hype and crashes that are characterized by large positive and negative returns: e.g. in the very beginning (when the price was still quite low), at the end of 2013 (the first time btc had a price of $1000 USD), at the end of 2017 (with a price over $19,000 USD),

---

[10] See e.g. https://cardano.org/enterprise/; accessed 19/05/2022.

[11] Tether is backed by USD cash reserves. See https://tether.to/en/.

[12] We use a maximum depth of $d_{max} = 5$ corresponding to the number of covariates and a weight decay of 2, meaning a split further down in each tree receives less weight w_l in the final frequency, as it is less important for the three specific currencies analyzed in the previous section. Specifically, for layer $l = 1, \ldots, 5$, $w_l = \dfrac{l^{-2}}{\sum_{l=1}^{5} l^{-2}}$.

[13] See Fig. B.7 in the appendix for an overview of the log returns of the three currencies.
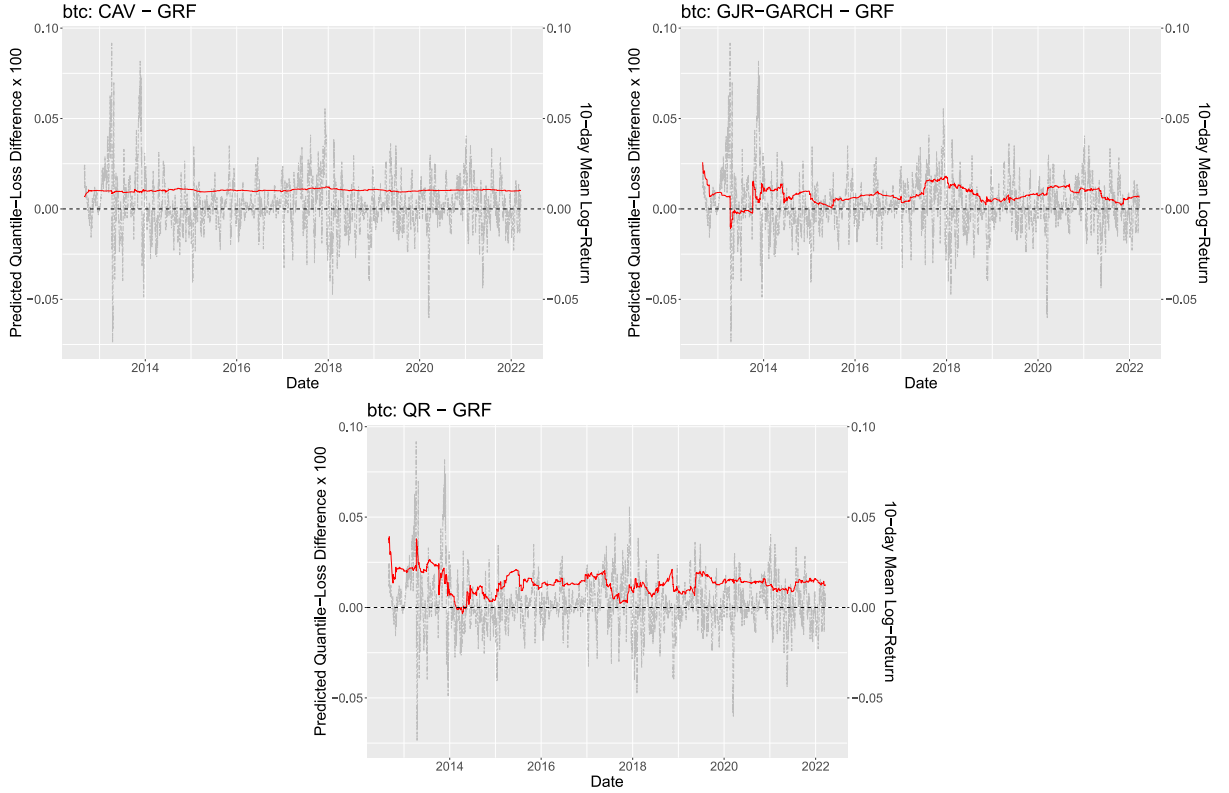
Fig. 3. Rolling 180-day mean of predicted loss difference series (red) of $h_t\hat{\beta}_0$ of CPA tests on Bitcoin (btc) predicted 5% VaR with $l = 500$ for GRF vs. CAV (left), GJR-GARCH (right), and QR (bottom center) with rolling mean 10-day log-returns in grey. A positive predicted loss difference indicates that the prediction error of GRF is smaller than of the compared method.
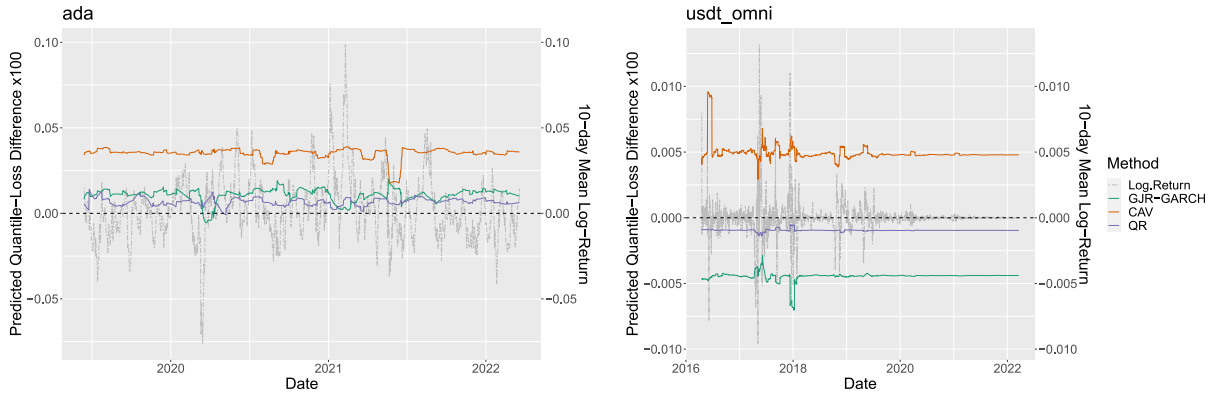
Fig. 4. Rolling 30-day mean of predicted loss difference series of $h_t\hat{\beta}_0$ of CPA tests on Cardano/Tether (ada/usdt_omni) predicted 5% VaR with $l = 500$ for GRF vs. CAV (orange), GJR-GARCH (green), and QR (purple) with rolling mean 10-day log returns in grey. A positive predicted loss difference indicates that the prediction error of GRF is smaller than that of the compared method.

and from mid-2020 to mid-2021 (when there were multiple hypes and crashes during the Covid-19 pandemic). Between these hype periods, the lagged SDs are most important. In 2014 and 2015 and from the end of 2019 to mid-2021, 30-day SDs contribute most to the GRF predictions, followed by seven-day SDs in 2016–2018 and 60-day SDs from 2018 to the end of 2019. This changing scheme is interesting, as 30-day SDs seem to be a good predictor, especially in very unstable times (return-wise),

while seven-day and 60-day SDs are more important in relatively stable times.

Finally, the stablecoin usdt_omni is an exception, being largely dominated by the lagged return. Given that by construction, usdt_omni is essentially bound to the US dollar, its dynamic properties also largely correspond to those of standard currencies. From mid-2019, the prices and returns are rather stable and the volume increases strongly, and the influence of lagged seven-day SDs increases slightly, while still being less important than

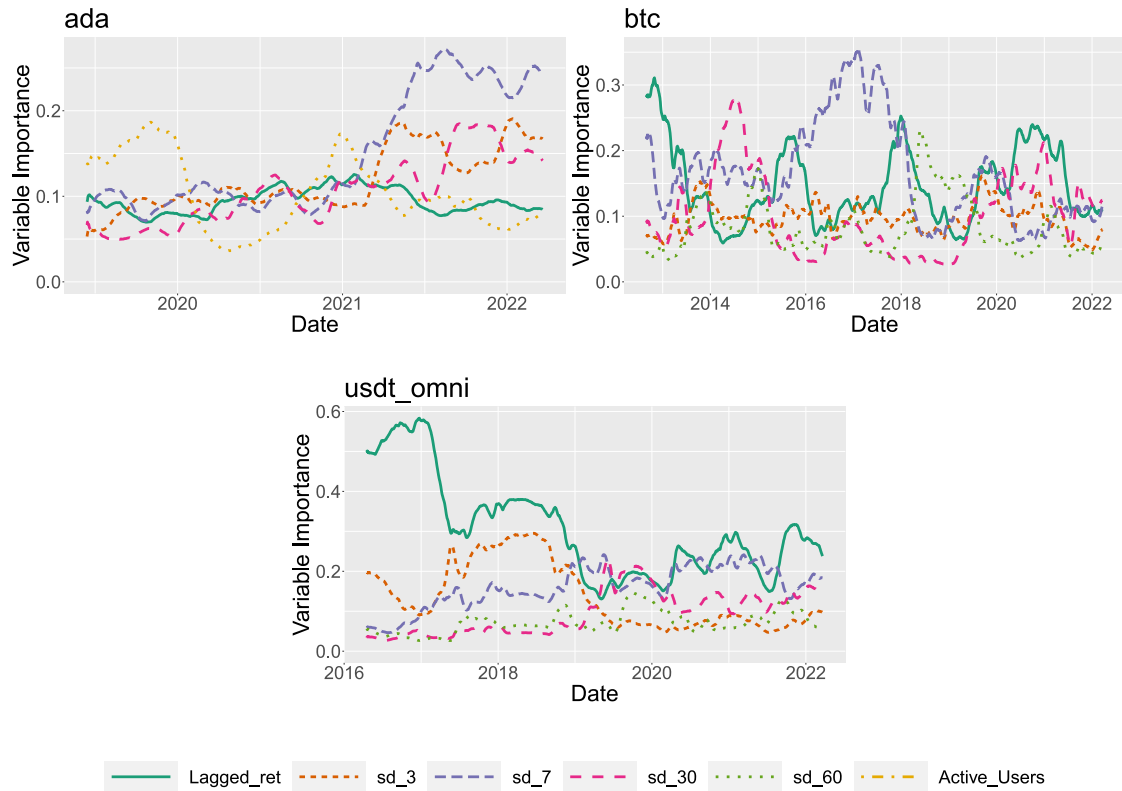**Fig. 5.** Rolling 30-day mean of the GRF variable importance in the out-of-sample period for the full data of ada, btc, and usdt_omni, for predicted one-day-ahead 5% VaR with $l = 500$. Plotted are the five most important variables for each cryptocurrency. The variable importance of covariate x_p is measured as the proportion of splits on x_p relative to all splits in a respective layer l (over all trees in a trained forest), weighted by layer l. The variables are described in Section 2.

Table 10
Period 4: Medians of p-values for DQ tests.

Group	GRF	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	GRF-X	QRF-X	QR-X	GARCH-X
All	0.24	0.12	0.14	0.10	0.11	0.48	0.01	0.33	0.16	0.00	0.03
High	0.22	0.07	0.23	0.12	0.15	0.33	0.00	0.32	0.09	0.00	0.00
Low	0.25	0.14	0.11	0.06	0.10	0.49	0.01	0.34	0.25	0.00	0.09

The table shows medians of p-values for DQ tests. Higher median p-values indicate correct conditional coverage according to the DQ test. The currencies are divided into two high and low groups based on the SER covariate, as in Table 7. The remaining details correspond to Table 6.

lagged returns. This is not surprising, as usdt_omni is quite stable in comparison to btc and ada.

5.2.2. Post-pandemic period

We additionally study the post-pandemic period from 2023/01/01–2024/04/06, which is similar in terms of the number of traded currencies (83) to Period 3 but has different exogenous constraints, such as the war in Ukraine. Nevertheless, it is characterized by an overall lower level of volatility of all cryptocurrencies in the market.

Here in the DQ test, GRF is outperformed by GJR-GARCH over all currencies, but for important subgroups, GRF still dominates. In particular, in Period 4, the impact of characteristics such as SER and number of traded assets is much more pronounced.

This can be seen from the excellent DQ performance results of GRF-X in Table 10 for currencies in the high-SER

group and supported by the impact results of Table 11 where the two characteristics stand out in particular in relation to the periods before. Note that 24% of currencies belong to Group_high in Period 4. The mean daily trading volume in USD in Group_high is $334,296,265 in Period 4. In the pairwise CPA tests, GRF again dominates overall, but in terms of significant differences cannot be distinguished from GRF-X (see Table 12).

6. Conclusion

In this paper, we showed that random forests can significantly improve the forecast performance for VaR predictions when tailored to conditional quantiles. In simulations and in analyzing return data of the 105 largest cryptocurrencies, the proposed generalized random forests

Table 11
Difference between covariates of cryptos where GRF is better vs. worse.

	Period 1	Period 2	Period 3	Period 4	Full data
Ret	0.83	0.96	4.24	0.82	1.83
Active_Users	0.30	0.23	3.00	1.84	2.40
Total_Users	0.48	0.18	0.32	0.10	0.27
Total_Users_USD100	0.12	0.13	0.50	0.14	0.50
Total_Users_USD_10	0.18	0.14	0.79	0.15	0.71
SER	0.74	0.49	0.47	0.24	0.48
Transactions	0.38	0.04	5.78	3.62	3.79
Velocity	4.11	3.46	0.93	0.76	0.76
sd_3	0.91	0.64	0.88	0.99	0.97
sd_7	0.92	0.64	0.88	0.98	0.95
sd_30	0.93	0.66	0.87	0.98	0.95
sd_60	0.92	0.63	0.88	0.99	0.96

The table shows shares of groups of cryptocurrencies where at least two of CAV, QR, and GJR-GARCH have better CPA performance than GRF divided by the rest, calculated as in Table 9. The numerators of the respective quotients are mean values over all cryptocurrencies for the median of each covariate in the respective time period. A value smaller than 1 indicates that currencies where GRF performs better have a higher average median value of the respective covariate than the rest of the currencies.

Table 12
Period 4: Performance and significance of CPA tests for GRF without additional covariates.

GRF vs.:	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	GRF-X	QRF-X	QR-X	GARCH-X
Share of GRF with better performance										
Period 4	0.90	0.70	0.70	0.75	0.65	0.94	0.33	0.94	0.98	0.73
Share of GRF with significantly better performance										
Period 4	0.51	0.17	0.17	0.23	0.13	0.67	0.05	0.70	0.76	0.53
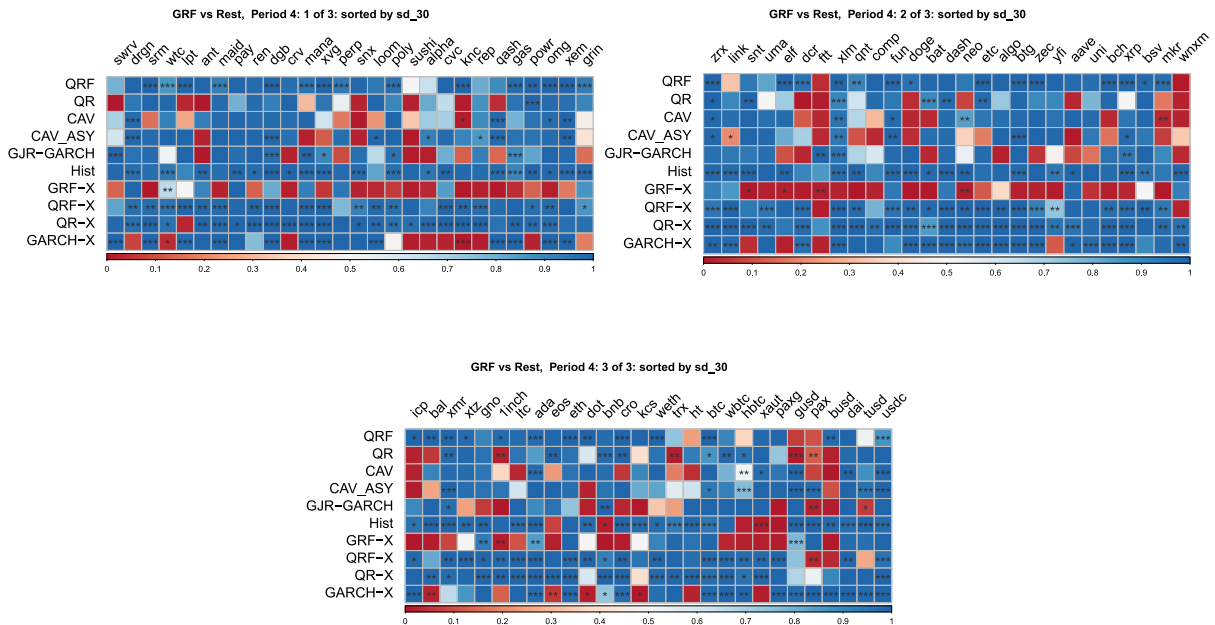
The top part shows summary values that are shares over all cryptocurrencies in Period 4. It describes the number of times that GRF had a better performance (i.e. more than 50% of predicted losses by the CPA test were smaller for GRF) relative to all cryptocurrencies (in that period), and the number of times that GRF was significantly better (at least at a 10% level) as judged by the CPA test over all cryptocurrencies (Period 4). THe bottom grpahics provide an overview of results for CPA tests of GRF vs. all other methods for cryptocurrencies in Period 4 by 30-day lagged standard deviations from highest to lowest. The color of each box indicates the performance of GRF, with 1 indicating that GRF has a smaller predicted loss in 100% of cases. *, **, and *** indicate significance on a level of 10%, 5%, and 1%, respectively.

showed superior prediction performance. In particular, the adaptive non-linear form of GRF appeared to capture time variations of volatility and spike behavior in cryptocurrency returns especially well in contrast to more conventional financial econometric methods. We further showed that GRF is superior at assessing the tail risk of cryptocurrencies in times where volatility in returns is high, which often coincides with increased speculation in the market. In such periods, standard time-series-based procedures, even when augmented with external factors,

R. Buse, K. Görgen and M. Schienle

Table A.13

Overview of all employed crypto assets.

ID	Start date	End date	Obs.	X	Time periods
1inch	2020–12–26	2024–04–07	1199	8	4
aave	2020–10–10	2024–04–07	1276	8	4
ada	2017–12–01	2024–04–07	2320	8	3, 4
algo	2019–06–22	2024–04–07	1752	8	3, 4
alpha	2020–10–11	2024–04–07	1275	8	4
ant	2017–08–29	2024–04–07	2414	8	3, 4
avaxc	2020–09–23	2024–04–06	1292	8	4
avaxp	2020–09–23	2024–04–06	1292	8	4
avaxx	2020–09–23	2024–04–06	1292	7	4
bal	2020–06–25	2024–04–07	1383	8	4
bat	2017–10–06	2024–04–07	2376	8	3, 4
bch	2017–08–01	2024–04–07	2442	8	3, 4
bnb	2017–07–15	2024–04–07	2459	8	3, 4
bnb_eth	2017–07–15	2019–04–21	646	8	
bsv	2018–11–15	2024–04–07	1971	8	3, 4
btc	2010–07–18	2024–04–07	5013	8	1, 2, 3, 4
btg	2017–10–25	2024–04–07	2357	8	3, 4
busd	2019–09–20	2024–04–07	1662	8	4
comp	2020–06–18	2024–04–07	1390	8	4
cro	2019–03–20	2024–04–07	1846	8	3, 4
crv	2020–08–15	2024–04–07	1332	8	4
cvc	2017–09–11	2024–04–07	2401	8	3, 4
dai	2019–11–20	2024–04–07	1601	8	4
dash	2014–02–08	2024–04–07	3712	8	1, 2, 3, 4
dcr	2016–05–17	2024–04–07	2883	8	2, 3, 4
dgb	2015–02–10	2024–04–07	3345	8	2, 3, 4
doge	2014–01–23	2024–04–07	3728	8	1, 2, 3, 4
dot	2020–08–20	2024–04–07	1327	8	4
drgn	2018–01–03	2024–04–07	2287	8	3, 4

are substantially inferior at VaR predictions, suggesting that a comparison of predicted losses could serve as an easy, empirical pre-screening device to detect such speculative bubbles.

Our findings are highly relevant for the risk assessment of cryptocurrencies where we showed that classic methods might lead to a miss-quantification of associated risks. Such results directly impact investment and hedging decisions for respective assets. Specifically, we detected the superior performance of GRF especially for highly volatile cryptocurrencies that have a high number of active users and might thus be prone to speculation and hype. On the other hand, for the class of stablecoins that are usually tied to some large, classic currency such as the USD, and which are usually dominated by a smaller number of large accounts, other classic methods such as GJR-GARCH or quantile regression were more on par with GRF. This also corresponded to the simulation results mimicking the behavior of conventional assets such as stock returns where the advantage of GRF decreased. Major gains of GRF, however, occurred the more cryptocurrency-like the setup became. The proposed random forests methodology can identify important external factors despite the non-linearity of the methodology. We showed that such factors are time varying and changing, particularly in unstable times. For future research, this set of covariates could even be further augmented with other potentially driving real-time factors, such as for example properly extracted and filtered social media information. The relevance of such factors might also

provide additional guidance for relevant exogenous information for portfolio formation and the design of trading strategies, for instance (cp. e.g. Huang, Han, Newton, Platanakis, Stafylas, & Sutcliffe, 2023; Platanakis & Urquhart, 2020).

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.

Acknowledgments

Appendix A. Additional empirical stylized facts

See Tables A.13–A.15.

Appendix B. Tables and figures with additional results

See Figs. B.6–B.7 and Tables B.16–B.18.

Table A.13 (*continued*).

ID	Start date	End date	Obs.	X	Time periods
elf	2017–12–22	2024-04-07	2299	8	3, 4
eos	2018–06–09	2024-04-07	2130	2	3, 4
eos_eth	2017–06–29	2018-06-01	338	8	
etc	2016–07–25	2024-04-07	2814	8	2, 3, 4
eth	2015–08–08	2024-04-07	3166	8	2, 3, 4
ftt	2019–08–20	2024-04-07	1693	8	4
fun	2017–09–02	2024-04-07	2410	8	3, 4
gas	2017–08–08	2024-04-07	2435	8	3, 4
gno	2017–05–02	2024-04-07	2533	8	3, 4
gnt	2017–02–19	2024-04-06	2604	8	3, 4
grin	2019–01–29	2024-04-07	1896	2	3, 4
gusd	2018–09–16	2024-04-07	2031	8	3, 4
hbtc	2019–12–09	2024-04-07	1582	8	4
hedg	2019–11–02	2022-01-27	818	8	
ht	2019–03–06	2024-04-07	1860	8	3, 4
husd	2019–07–18	2022-11-17	1219	8	
icp	2021–05–11	2024-04-07	1063	8	4
kcs	2020–04–04	2024-04-07	1465	5	4
knc	2017–09–27	2024-04-07	2385	8	3, 4
lend	2017–12–09	2024-04-06	2311	8	3, 4
leo_eos	2019–05–21	2024-04-06	1783	4	3, 4
leo_eth	2019–05–21	2024-04-06	1783	8	3, 4
link	2017–09–29	2024-04-07	2383	8	3, 4
loom	2018–05–03	2024-04-07	2167	8	3, 4
lpt	2018–12–20	2024-04-07	1936	8	3, 4
ltc	2013–04–01	2024-04-07	4025	8	1, 2, 3, 4
maid	2014–07–10	2024-04-07	3560	8	2, 3, 4
mana	2017–08–25	2024-04-07	2418	8	3, 4
matic_eth	2019–04–27	2024-04-06	1807	8	3, 4
mkr	2017–12–26	2024-04-07	2295	8	3, 4
neo	2017–07–15	2024-04-07	2459	8	3, 4
nxm	2020–08–26	2024-04-06	1320	8	4
omg	2017–07–15	2024-04-07	2459	8	3, 4
pax	2018–11–30	2024-04-07	1956	8	3, 4
paxg	2020–02–15	2024-04-07	1514	8	4
pay	2017–10–03	2024-04-07	2379	8	3, 4
perp	2021–02–04	2024-04-07	1159	8	4
poly	2018–06–15	2024-04-07	2124	8	3, 4
powr	2017–11–02	2024-04-07	2349	8	3, 4
ppt	2017–09–20	2022-09-12	1819	8	
qash	2017–11–06	2024-04-07	2345	8	3, 4
qnt	2019–03–16	2024-04-07	1850	8	3, 4
ren	2018–12–07	2024-04-07	1949	8	3, 4
renbtc	2020–05–13	2024-04-06	1425	8	4
rep	2016–10–04	2024-04-07	2743	8	3, 4
rev_eth	2020–03–26	2024-03-27	1463	8	
sai	2017–12–23	2019-11-30	708	8	
snt	2017–06–19	2024-04-07	2485	8	3, 4
snx	2020–04–09	2024-04-07	1460	8	4
srm	2020–08–11	2024-04-07	1336	8	4
sushi	2020–09–01	2024-04-07	1315	8	4
swrv	2020–09–22	2024-04-07	1294	8	4
trx	2018–06–25	2024-04-07	2114	3	3, 4
trx_eth	2017–10–07	2018-06-24	261	8	
tusd	2018–07–06	2024-04-07	2103	7	3, 4
uma	2020–09–08	2024-04-07	1308	8	4
uni	2020–09–18	2024-04-07	1298	8	4
usdc	2018–09–28	2024-04-07	2019	8	3, 4
usdk	2020–06–13	2023-04-17	1039	8	
usdt_eth	2017–11–28	2024-04-06	2322	8	3, 4
usdt_omni	2014–10–06	2024-04-06	3471	8	2, 3, 4

Table A.13 (*continued*).

ID	Start date	End date	Obs.	X	Time periods
usdt_trx	2019–04–16	2024-04-06	1818	8	3, 4
vtc	2014–01–29	2023-12-10	3603	8	1, 2, 3
wbtc	2018–11–24	2024-04-07	1962	8	3, 4
weth	2017–12–12	2024-04-07	2309	8	3, 4
wnxm	2020–08–26	2024-04-07	1321	8	4
wtc	2017–08–28	2024-04-07	2415	8	3, 4
xaut	2020–01–25	2024-04-07	1535	8	4
xem	2015–04–01	2024-04-07	3295	3	2, 3, 4
xlm	2015–09–30	2024-04-07	3113	8	2, 3, 4
xmr	2014–05–20	2024-04-07	3611	2	2, 3, 4
xrp	2014–08–15	2024-04-07	3524	8	2, 3, 4
xtz	2018–06–30	2024-04-07	2109	8	3, 4
xvg	2017–09–30	2024-04-07	2382	8	3, 4
yfi	2020–07–25	2024-04-07	1353	8	4
zec	2016–10–29	2024-04-07	2718	8	3, 4
zrx	2017–08–11	2024-04-07	2432	8	3, 4

All employed cryptocurrencies where *X* denotes the number of available covariates out of eight. Numbers in "Time periods" mark periods where the respective asset is fully observed. All cryptocurrencies in this table, with the exception of vtc, exist until the end of Period 4.

Table A.14
Overview of non-stationarity tests.

	KPSS_level	KPSS_trend	ADF
1inch	0.10	0.10	0.01
ada	0.10	0.10	0.01
algo	0.10	0.03	0.01
alpha	0.10	0.01	0.01
avaxc	0.10	0.04	0.01
avaxp	0.10	0.04	0.01
avaxx	0.10	0.04	0.01
bnb	0.06	0.07	0.01
bnb_eth	0.06	0.05	0.01
btc	0.02	0.10	0.01
dash	0.02	0.10	0.01
dot	0.10	0.01	0.01
drgn	0.08	0.10	0.01
hbtc	0.10	0.02	0.01
ht	0.05	0.10	0.01
icp	0.01	0.10	0.01
lend	0.10	0.04	0.01
omg	0.10	0.10	0.01
perp	0.10	0.08	0.01
poly	0.10	0.03	0.01
ppt	0.10	0.06	0.01
renbtc	0.10	0.01	0.01
snx	0.10	0.02	0.01
uni	0.10	0.07	0.01
wbtc	0.10	0.10	0.01
weth	0.10	0.05	0.01
xem	0.03	0.10	0.01
xmr	0.10	0.09	0.01
yfi	0.10	0.04	0.01

Values in columns show p-values for KPSS and ADF tests, as described in Section 2. We only show values for assets that have values smaller than 0.1 for the KPSS tests. No asset had p-values larger than 0.01 for the ADF tests.

Table A.15
External covariates and descriptions.

Variable name	Coin Metrics code	Description
Active_Users	AdrActCnt	The number of unique active daily addresses
Total_Users	AdrBalCnt	The number of unique addresses that hold any amount of native units of that currency
Total_Users_USD100	AdrBalUSD100Cnt	The number of unique addresses that hold at least $100 USD of native units of that currency
Total_Users_USD10	AdrBalUSD10Cnt	The number of unique addresses that hold at least $10 USD of native units of that currency
SER	SER	The supply equality ratio; i.e. the ratio of supply held by addresses with less than one ten-millionth ($1/10^7$) of the current supply to the top one percent of addresses with the highest current supply
Transactions	TxCnt	The number of daily initiated transactions
Velocity	VelCur1yr	The velocity of supply in the current year, which describes the ratio of current supply to the sum of the value transferred in the last year

The table contains all used external covariates. The variable coding corresponds to https://docs.coinmetrics.io/. Detailed variable descriptions are available at https://docs.coinmetrics.io/info/metrics.

Table B.16
Medians of p-values for DQ tests sorted by covariate values in different time periods.

Period	Sorting	GRF	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	GRF-X	QRF-X	QR-X	GARCH-X
High group												
SER												
1	1 Highest	0.17	0.81	0.00	0.00	0.00	0.12	0.00	0.01	0.00	0.00	0.00
2	5 Highest	0.15	0.03	0.20	0.05	0.41	0.10	0.09	0.13	0.08	0.00	0.00
3	15 Highest	0.41	0.04	0.08	0.06	0.08	0.16	0.00	0.22	0.22	0.00	0.00
Number of active users												
1	1 Highest	0.17	0.81	0.00	0.00	0.00	0.12	0.00	0.01	0.00	0.00	0.00
2	5 Highest	0.15	0.06	0.40	0.10	0.41	0.11	0.09	0.14	0.06	0.00	0.00
3	15 Highest	0.37	0.05	0.06	0.06	0.08	0.16	0.00	0.21	0.22	0.00	0.00
Seven-day lagged standard deviation												
1	1 Highest	0.14	0.02	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
2	5 Highest	0.53	0.01	0.27	0.22	0.10	0.23	0.20	0.45	0.17	0.00	0.33
3	15 Highest	0.04	0.00	0.01	0.00	0.00	0.05	0.00	0.01	0.00	0.00	0.02
Low group												
SER												
1	4 Lowest	0.11	0.20	0.08	0.09	0.10	0.02	0.00	0.02	0.01	0.00	0.00
2	9 Lowest	0.31	0.04	0.08	0.17	0.29	0.27	0.17	0.45	0.07	0.00	0.11
3	49 Lowest	0.04	0.02	0.03	0.01	0.00	0.15	0.00	0.10	0.03	0.00	0.02
Number of active users												
1	4 Lowest	0.11	0.20	0.08	0.09	0.10	0.02	0.00	0.02	0.01	0.00	0.00
2	9 Lowest	0.53	0.01	0.08	0.15	0.10	0.23	0.17	0.45	0.08	0.00	0.11
3	49 Lowest	0.06	0.02	0.04	0.01	0.00	0.15	0.00	0.10	0.03	0.00	0.02
Seven-day lagged standard deviation												
1	4 Lowest	0.12	0.43	0.08	0.09	0.10	0.07	0.00	0.02	0.01	0.00	0.00
2	9 Lowest	0.28	0.04	0.06	0.10	0.33	0.11	0.09	0.14	0.06	0.00	0.00
3	49 Lowest	0.21	0.02	0.06	0.02	0.01	0.26	0.00	0.21	0.11	0.00	0.00

The table shows the medians p-values of DQ tests sorted by covariate values in different time periods. The currencies are divided into two groups based on specific covariate values for each time period. The groups are constructed to best separate the sorted covariate values of the currencies. They are therefore separated at their steepest decay to obtain two groups with homogeneous covariate values, where the top group in each period contains at least the top quintile but not more than the top sextile of the data. The remaining details correspond to Table 6.

Table B.17

Performance and significance of CPA tests with GRF-X over different time periods.

GRF-X vs.:	GRF	QRF	QR	CAV	CAV_ASY	GJR-GARCH	Hist	QRF-X	QR-X	GARCH-X
Share of GRF with better performance										
Period 1	0.40	1.00	0.80	0.60	0.80	1.00	1.00	1.00	1.00	1.00
Period 2	0.64	1.00	0.86	0.86	0.93	0.93	1.00	1.00	1.00	1.00
Period 3	0.56	0.92	0.73	0.81	0.80	0.77	0.98	0.97	0.95	0.81
Period 4	0.61	0.93	0.71	0.80	0.80	0.70	0.95	0.94	0.95	0.75
Full data	0.61	0.92	0.70	0.80	0.78	0.63	0.96	0.92	0.96	0.75
Share of GRF with significantly better performance										
Period 1	0.00	0.40	0.20	0.00	0.20	0.00	1.00	1.00	1.00	0.60
Period 2	0.07	0.71	0.21	0.29	0.50	0.29	0.79	0.93	0.93	0.71
Period 3	0.08	0.61	0.17	0.23	0.34	0.20	0.83	0.80	0.88	0.55
Period 4	0.06	0.54	0.16	0.23	0.29	0.19	0.75	0.76	0.80	0.53
Full data	0.08	0.53	0.13	0.24	0.28	0.17	0.76	0.72	0.76	0.53
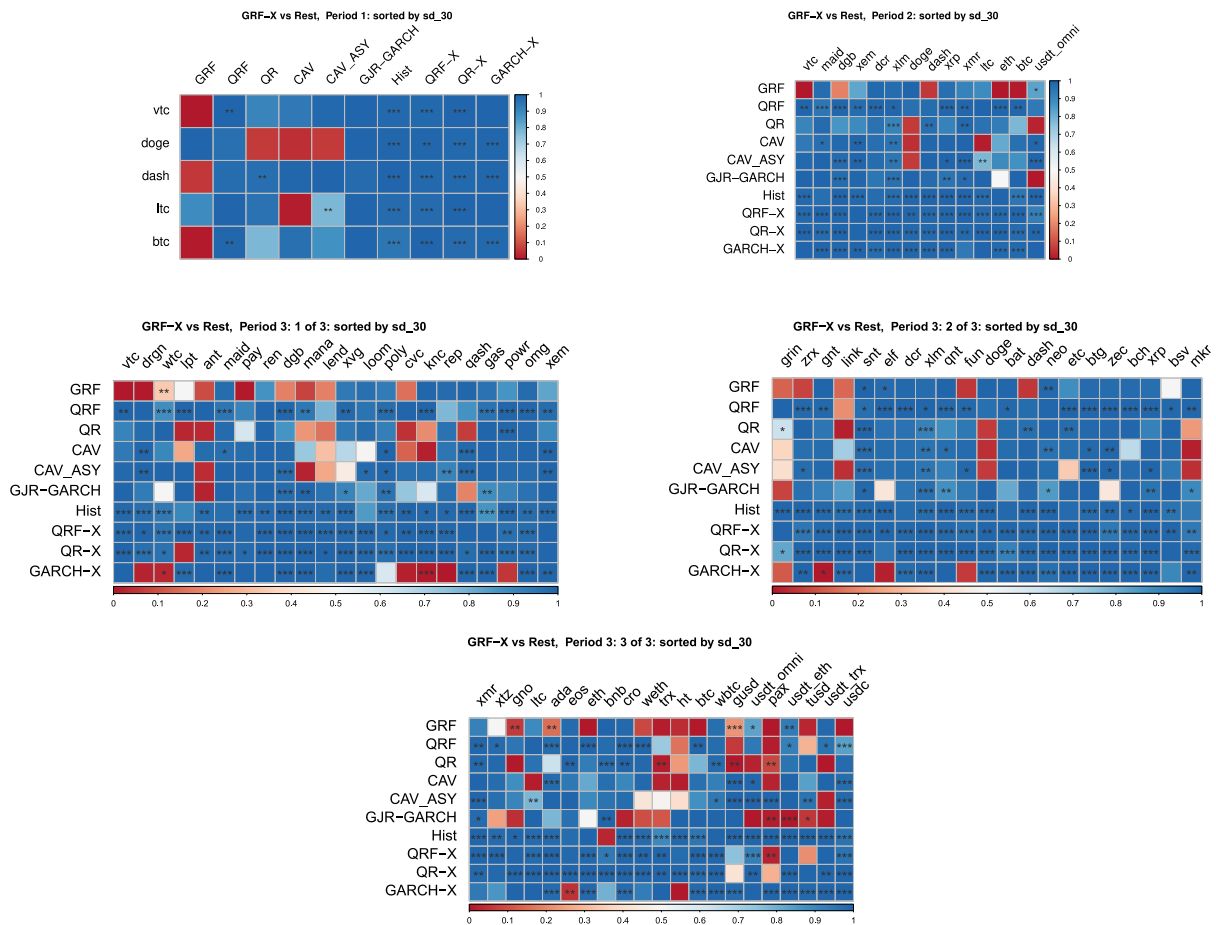
GRF–X vs Rest, Period 1: sorted by sd_30

GRF–X vs Rest, Period 2: sorted by sd_30

GRF–X vs Rest, Period 3: 1 of 3: sorted by sd_30

GRF–X vs Rest, Period 3: 2 of 3: sorted by sd_30

GRF–X vs Rest, Period 3: 3 of 3: sorted by sd_30

The top part shows summary values that are shares over all cryptocurrencies in the respective time period. It describes the number of times that GRF-X had a better performance (i.e. more than 50% of predicted losses by the CPA test were smaller for the GRF-X) relative to all cryptocurrencies (in that period), and the number of times that GRF-X was significantly better (at least at a 10% level) as judged by the CPA test over all cryptocurrencies (in that period). The bottom part shows the detailed results of CPA tests with the color of each box indicating the performance of GRF-X. Blue signifies a performance of 1, meaning that GRF-X has a smaller predicted loss in 100% of cases. *, **, and *** show significance on a level of 10%, 5%, and 1%, respectively. The values are ordered by 30-day lagged standard deviations from highest to lowest.

Table B.18

Difference between covariates of cryptos where GRF-X is better vs. worse.

	Period 1	Period 2	Period 3	Period 4	Full data
Ret	0.83	1.64	−1.16	−0.34	0.16
Active_Users	0.30	0.58	2.68	2.99	2.18
Total_Users	0.48	0.54	0.14	0.12	0.14
Total_Users_USD100	0.12	0.21	0.08	0.11	0.18
Total_Users_USD_10	0.18	0.32	0.12	0.15	0.20
SER	0.74	1.45	0.38	0.29	0.42
Transactions	0.38	0.12	2.67	2.33	1.97
Velocity	4.11	9.26	1.40	1.32	1.02
sd_3	0.91	0.94	0.86	0.81	0.99
sd_7	0.92	0.96	0.87	0.82	0.98
sd_30	0.93	0.98	0.87	0.82	0.99
sd_60	0.92	0.98	0.87	0.84	1.00

The table shows shares of groups of cryptocurrencies where at least two of CAV, QR, and GJR-GARCH have better CPA performance than GRF-X divided by the rest. Raw values before division are mean values over all cryptocurrencies for the median of each covariate in the respective time period.
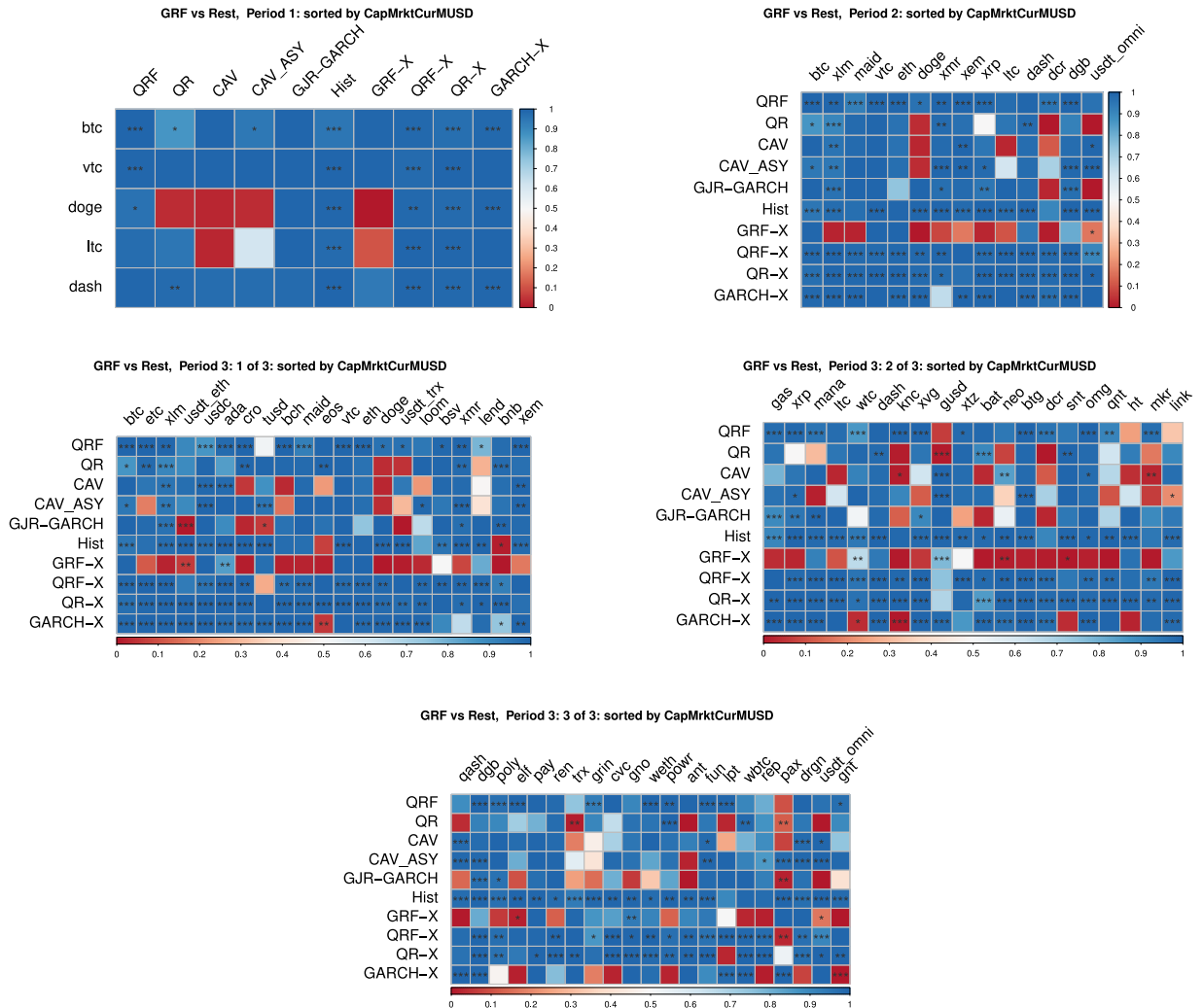

Fig. B.6. Overview of results for CPA tests of GRF vs. all other methods for cryptocurrencies ordered by market cap from highest to lowest, left to right. The color of each box indicates the performance of GRF, with 1 indicating that GRF has a smaller predicted loss in 100% of cases. *, **, and *** indicate significance on a level of 10%, 5%, and 1%, respectively.
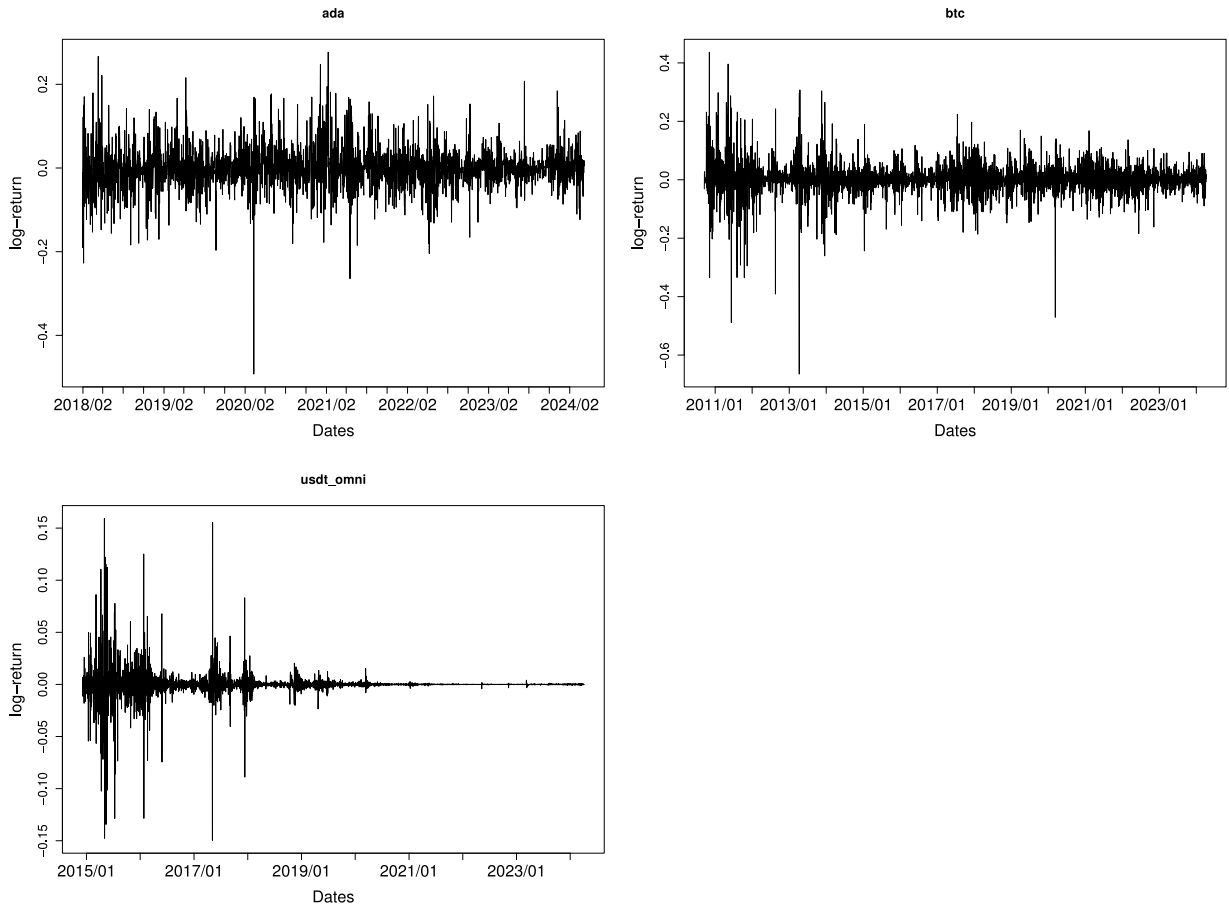
Fig. B.7. Log returns of the specific cryptocurrencies analyzed in Section 5.2.

Data and code availability

All data and replication materials can be found in the Github repository https://github.com/KITmetricslab/crypto-VaR-predictions.

References

Ardia, D., Boudt, K., & Catania, L. (2019). Generalized autoregressive score models in R: The GAS package. *Journal of Statistical Software*, *88*.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*, 1179–1203.

Baur, D. G., Hong, K. H., & Lee, A. D. (2018). Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets Institutions and Money*, *54*, 177–189.

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

Breiman, L. (1996). Some properties of splitting criteria. *Machine Learning*, *24*, 41–47.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Cheah, E. T., & Fry, J. (2015). Speculative bubbles in bitcoin markets? An empirical investigation into the fundamental value of bitcoin. *Economics Letters*, *130*, 32–36.

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

Chu, J., Chan, S., Nadarajah, S., & Osterrieder, J. (2017). GARCH modelling of cryptocurrencies. *Journal of Risk and Financial Management*, *10*, 17.

Elendner, H., Trimborn, S., Ong, B., & Lee, T. M. (2017). The cross-section of crypto-currencies as financial assets: Investing in cryptocurrencies beyond bitcoin. In *Handbook of blockchain, digital finance, and inclusion, volume 1: cryptocurrency, FinTech, InsurTech, and regulation* (pp. 145–173). Elsevier.

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

Ghysels, E., & Nguyen, G. (2019). Price discovery of a speculative asset: Evidence from a bitcoin exchange. *Journal of Risk and Financial Management*, *12*, 164.

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

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

Gkillas, K., & Katsiampa, P. (2018). An application of extreme value theory to cryptocurrencies. *Economics Letters*, *164*, 109–111.

Glaser, F., Zimmermann, K., Haferkorn, M., Weber, M. C., & Siering, M. (2014). Bitcoin – asset or currency? Revealing users' hidden intentions. In *ECIS 2014 proceedings – 22nd European conference on information systems* (pp. 1–14).

Glosten, L. R., Jagannathan, R., & 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*, 1779–1801.

Hafner, C. M. (2020). Testing for bubbles in cryptocurrencies with time-varying volatility. *Journal of Financial Econometrics*, *18*, 233–249.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hencic, A., & Gouriéroux, C. (2015). Noncausal autoregressive model in application to bitcoin/USD exchange rates. In *Studies in Computational Intelligence*: vol. 583, (pp. 17–40). Springer.

Huang, X., Han, W., Newton, D., Platanakis, E., Stafylas, D., & Sutcliffe, C. (2023). The diversification benefits of cryptocurrency asset categories and estimation risk: Pre and post COVID-19. *The European Journal of Finance*, 29, 800–825.

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

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15, 143–156.

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

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root, how sure are we that economic time series have a unit root? *Journal of Econometrics*, 54, 159–178.

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

Liu, Y., & Tsyvinski, A. (2020). Risks and returns of cryptocurrency. *The Review of Financial Studies*, 34, 2689–2727.

Maciel, L. (2020). Cryptocurrencies value-at-risk and expected shortfall: Do regime-switching volatility models improve forecasting? *International Journal of Finance & Economics*, 26, 4840–4855.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7, 983–999.

Pafka, S., & Kondor, I. (2001). Evaluating the RiskMetrics methodology in measuring volatility and value-at-risk in financial markets. *Physica A: Statistical Mechanics and its Applications*, 299, 305–310.

Petukhina, A., Trimborn, S., Härdle, W. K., & Elendner, H. (2021). Investing with cryptocurrencies – Evaluating their potential for portfolio allocation strategies. *Quantitative Finance*, 1–29.

Platanakis, E., & Urquhart, A. (2019). Portfolio management with cryptocurrencies: The role of estimation risk. *Economics Letters*, 177, 76–80.

Platanakis, E., & Urquhart, A. (2020). Should investors include bitcoin in their portfolios? a portfolio theory approach. *The British Accounting Review*, 52, Article 100837.

Selmi, R., Tiwari, A., & Hammoudeh, S. (2018). Efficiency or speculation? A dynamic analysis of the bitcoin market. *Economics Bulletin*, 38, 2037–2046.

Silvennoinen, A., & Teräsvirta, T. (2016). Testing constancy of unconditional variance in volatility models by misspecification and specification tests. *Studies in Nonlinear Dynamics & Econometrics*, 20.

Takeda, A., & Sugiyama, M. (2008). ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on machine learning* (pp. 1056–1063).

Trimborn, S., Li, M., & Härdle, W. K. (2020). Investing with cryptocurrencies – A liquidity constrained investment approach. *Journal of Financial Econometrics*, 18, 280–306.

Trucios, C., Tiwari, A. K., & Alqahtani, F. (2020). Value-at-risk and expected shortfall in cryptocurrencies' portfolio: A vine copula–based approach. *Applied Economics*, 52, 2580–2593.

Vigliotti, M. G., & Jones, H. (2020). The rise and rise of cryptocurrencies. In *The executive guide to blockchain* (pp. 71–91). Springer.